Risk-based Renewal Prioritization Models (RPM) for Potable Water Pipeline Infrastructure Systems

Anmol Vishwakarma

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Civil Engineering

Sunil K. Sinha, Committee Chair
Marc A. Edwards
Jason K. Deane
Naren Ramakrishnan

December 8, 2025

Blacksburg, VA

Keywords: water pipeline infrastructure asset management, likelihood of failure, consequence of failure, risk analysis, model validation, renewal prioritization, expert systems, machine learning

Risk based Renewal Prioritization Models (RPM) for Potable Water Pipeline Infrastructure Systems

Anmol Vishwakarma

Abstract

Water pipelines are critical infrastructure assets buried across the United States, responsible for delivering safe drinking water at adequate pressures from source to customers. A majority of these pipelines were installed in the mid-twentieth century without adequate financial planning for future renewal, creating a growing renewal backlog under tight budget and operational constraints. Decades of utility data and practice-based knowledge, combined with advances in Artificial Intelligence (AI) and computational resources, now make it possible to revisit how renewal decisions are made. A review of current water pipeline renewal methods reveals major gaps, including weak integration of risk with decision criteria, ad hoc selection of modeling algorithms without strategic foresight, and limited, often internal-only, real-world validation.

This dissertation addresses these gaps by developing and testing an AI-enabled framework for risk-based renewal prioritization of water pipelines. The work has four main goals: (1) developing an AI model to predict the performance and Likelihood of Failure (LOF) of any water pipeline segment on a 0–5 scale, (2) creating an AI model to predict

the Consequence of Failure (COF) of any segment on a 0–5 scale, spanning economic, environmental, and social/service impacts, (3) building a multi-criteria optimization model to generate prioritized renewal portfolios that incorporate risk, cost, equity, and delivery constraints within budget limits, and (4) establishing experimental protocols to evaluate, verify, and validate model results against field inspections, retrospective failures, and expert judgement across multiple utilities. Applied to several U.S. utilities, the integrated LOF, COF, and portfolio models outperform age-based and heuristic baselines on predictive accuracy, calibration, and risk-reduction-per-dollar, while producing more spatially coherent and operationally feasible renewal programs in retrospective tests.

Finally, this research evaluates whether the additional effort required for data collection, model interpretation, and governance is justified relative to current utility practices, with tradeoffs assessed in terms of reduced emergency failures and costs, enhanced transparency and accountability in decision-making, and improved public trust. In the short term, the proposed framework supports more cost-effective and defensible capital improvement planning; in the long term, it provides a template for shifting water utilities from reactive, break-driven repairs to proactive, data-informed management of buried pipeline infrastructure using explainable AI models with characterized uncertainties.

Risk based Renewal Prioritization Models (RPM) for Potable Water Pipeline Infrastructure Systems

Anmol Vishwakarma

General Audience Abstract

Water pipelines are the hidden backbone of modern life, carrying clean water from treatment plants to homes and businesses. Many of these pipes in the United States were installed more than 50 years ago and are now aging, often without sufficient planning for their renewal. As these systems deteriorate, unexpected pipe breaks can flood streets, disrupt traffic, waste treated water, and create costly emergencies that are difficult for utilities and communities to absorb.

Advances in data, Artificial Intelligence (AI), and computing power now offer a chance to help utilities make more proactive and informed decisions about which pipes to renew and when. However, many current renewal practices still oversimplify risk, ignore real-world construction and budget constraints, and rarely undergo rigorous testing against observed failures and field inspections. This research develops and tests AI-based tools that (1) estimate how likely each pipe is to fail, (2) estimate what would happen if it fails, including economic, environmental, and social impacts, (3) create a decision model

that balances cost, risk, equity, and practical construction constraints within budget limits, and (4) establish practical procedures to scientifically test these models against real failures, inspection data, and expert assessments from multiple water utilities. When applied to several U.S. systems, these tools perform better than existing methods at identifying which pipes should be renewed and assembling renewal plans that achieve greater risk reduction for each dollar spent, with fewer construction conflicts and neighborhood disruptions in planning scenarios.

The study also examines whether the added effort of using AI models is justified by benefits such as fewer emergencies, reduced costs, and greater transparency for customers, regulators, and decision-makers. In the short term, the methods support more cost-effective and accountable planning; in the long term, they aim to help water utilities move away from crisis-driven repairs toward proactive, data-driven management of the buried infrastructure that supports everyday life.

To the child in Delhi who learned that water and air could not be trusted,

and to all who still live with that uncertainty

Acknowledgments

This dissertation began with a simple, slightly reckless decision to leave home, learn a new country, and try to do meaningful science in the middle of a pandemic. Nothing about the past few years has been controlled or tidy. It has mostly been one slightly bewildered, over-caffeinated graduate student held up by a lot of generous people. Whatever credit now appears next to my name really belongs to those who gave me their time, patience, and belief. I am deeply grateful to everyone who made it possible for me to move from New Delhi to Blacksburg, get through COVID and the lonely stretches in between, and still find the energy to finish this work.

First and foremost, I thank my advisor, Dr. Sunil K. Sinha, for many years of guidance and for repeatedly choosing to invest in a student whose ideas often arrived faster than his proof. His trust, steady support, and willingness to give me time and freedom have shaped every chapter in this dissertation. I am equally grateful to my committee members, Dr. Marc Edwards, Dr. Jason Deane, Dr. Naren Ramakrishnan, and Dr. Kathleen Hancock (former committee member), whose professionalism, intellectual honesty, and sharp questions forced this work to grow from interesting ideas into a defensible

scientific contribution. Each of them, in different ways, asked me to aim higher than I thought I could, while still allowing my sentences the occasional long walk.

I am thankful to Dr. Lee Sears and Dr. Jessica Torrey (then at the United States Bureau of Reclamation, USBR); Dr. Rich Niswonger, Dr. Jaime Painter, and Dr. Galen Gorski (then at the United States Geological Survey, USGS); Dr. Rick Archibald, Dr. Valentine Anantharaj and Dr. Mallikarjun Shankar (then at the Oak Ridge National Laboratory, ORNL); Darcy Male, Emmanuel Benyella, Rohit Dixit, Jimma Blen, Kishia Powell (then at Washington Suburban Sanitary Commission, WSSC Water); Nirmala Mahadevan (then at California Department of Water Resources); Eric Zúñiga (then at California State Water Resources Control Board); Christine Voudy (then at Georgia Environmental Protection Division); Ken Thompson and Chris Dermody (then at Jacobs Engineering); and Dr. Jian Zhang, Dr. Harry Zhang, and Walter Graf (then at the Water Research Foundation, WRF) for championing the projects that framed this dissertation. Their confidence in our research group, and their commitment to asking harder, deeper questions about water infrastructure, created space for a graduate student like me to explore these problems in more detail than I ever could have imagined. The opportunities they opened like access to real systems, real uncertainties, and real stakes are what allowed

this work to move beyond an academic exercise and toward contributions that, I hope, will eventually help our field grow and solve practical problems for utilities and the communities they serve. I also extend my sincere gratitude to the many engineers at water utilities across the United States whose collaboration, data contributions, and technical insights made this research possible. They generously shared their systems, time, and hard-won field experience with a graduate student they had never met in person, often while dealing with real pipe failures rather than the simulated ones on my screen.

I am grateful to the Sustainable Water Infrastructure Management (SWIM) Center, the Charles E. Via, Jr. Department of Civil and Environmental Engineering at Virginia Tech, USGS, USBR, ORNL, WRF, AWWA, NASSCO, ASCE, Microsoft, and BlackRock for the grants, scholarships, fellowships and invaluable support that sustained this work and bought the time and headspace needed to think, to fail, and to try again. Their support paid for far more than conference travel and tuition; it funded the luxury of working on long-term questions instead of short-term survival.

I extend heartfelt thanks to former SWIM graduate students Dr. Berk Uslu, Dr. Stephen Welling, Dr. Shaoqing Ge, Dr. Mayank Khurana, Pruthvi Patel, Pururaj Singh Shekhawat, and Darshan Vekaria, for their friendships and willingness to share advice

and stories of their own struggles. Their willingness to normalize failure, confusion, and course corrections made my own setbacks feel survivable and occasionally even routine. I would like to thank Dr. John Little for delivering the most inspiring lectures I have ever attended. His classes quietly shifted my trajectory toward mathematical modeling and systems thinking, without which this dissertation would not exist, and certainly would not have these many equations. My sincere appreciation also goes to James Carolan, whose grounded, real-world feedback pushed me to connect my modeling work to the practical realities faced by utilities. I am also grateful to Dr. Bahareh Behkam and Dr. Amrinder Nain, who opened their doors to me, especially around Thanksgiving, when most people are with their families and international students like me cannot easily travel home. Their example made it clear that research life means very little if it is not paired with genuine care for the people doing the work. To their students specially, Dr. Naimat Bari, Atharva Agashe, Hajar Chokhmane, and Ridi Barua, thank you for your friendship, for pulling me into your celebrations, and for making sure at least some evenings were about good food and bad jokes.

I am also deeply grateful to Dr. Pierre Glynn, whose example has been one of the most enduring academic inspirations in my life. Even in retirement and in the later stages

of his career, he brought an almost youthful energy, curiosity, and generosity to every conversation. His willingness to keep learning, to share his knowledge and wisdom, and to introduce me to remarkable books on philosophy and the natural world reminded me that rigorous science and deep reflection can, and should coexist. I wish him the very best of health and continuing fulfillment in everything he chooses to pursue.

Finally, I would like to thank Kathy Laskowski, who quietly brought sanity and steadiness to an otherwise chaotic research life. In the swirl of deadlines, shifting projects, and self-doubt, she was consistently kind, patient, and honest. Her reassurance that I was, in fact, ready to graduate did more than calm my nerves; it helped me see myself as someone who could finish this work and move forward. If this dissertation has a well-conditioned trajectory instead of a numerical blow-up, it is partly because she kept nudging me in the right direction.

Behind all of this professional support stands a quieter, deeper layer of care. My move from New Delhi to the United States, the years of distance from home, and the disruption of COVID demanded more than academic help. For seeing me through that, I owe more than I can say to my family and friends.

To my parents, Veena and Arun Vishwakarma, thank you for accepting the absurdity of your child digging into buried pipes (pun fully intended) halfway across the world and still backing that choice without conditions. You absorbed the uncertainty, the late-night phone calls, and the many years when I was physically absent and emotionally preoccupied, while still insisting that this journey mattered. I will always be grateful for everything you managed to provide from far away and for making trips to the United States when I could not come home because of lost passports, pandemics and research duties. I would not wish that kind of anxiety on anyone, and yet you handled it with more grace than I did.

My deepest thanks go to Dr. Binita Saha (my soon to be wife). You entered this journey when it was already long, tangled, and uncertain, and chose to stay. Thank you for enduring the busy weekends, the daily 5 p.m. meetings, the half-present conversations, and the emotional whiplash of funding decisions, paper reviews, and visa and job anxieties. You attended more practice talks, slide run-throughs, and "quick updates" than anyone should reasonably have to, and still managed to cheer for each small step forward. You carried more than your share of the emotional load so that I could carry this dissertation

to the end. Your faith in a future beyond the PhD, and your quiet insistence that my worth is not equal to my output, have been the most important safeguards of all.

To my sisters, Smriti and Sakriti Vishwakarma, and their families including Shashank, Nitesh, Mysha and little Nirvaan. Thank you for keeping "home" alive through messages, shared jokes, and small updates that cut through the gradient of time zones and work. A special thanks also to Sunil Vishwakarma for treating me like your own son and playing such a big part in my growing-up years. You all reminded me that in this family I am a son, brother, and uncle first and everything else a distant second, and that in the grand scheme of things, being part of this slightly noisy, slightly chaotic, very loving family is the most reliable system I know.

A special thanks to Krishna Saha and Biman Saha for treating me like a son from the very beginning, for opening up the world of Bengali food, language, and festivals, and for showing up exactly when we needed it most. In the final month, when our Blacksburg apartment had basically turned into a command center for two dissertations, journal papers, and defense presentations, you stepped in, took charge, and quietly brought back a sense of home.

To my friends here, thank you for providing a home away from home. Special thanks go to Saurabh Pant, Shivam Goel, Saurabh Gupta, and Saloni Gupta for tolerating my strange work schedule during holidays and for pretending not to notice that "I'll be there in ten minutes" was always quoted in PhD minutes, not real minutes. Thank you to Jishna Ganguly and Arun Noel Victor for always being understanding, inviting me to your special events, and giving a semblance of normalcy to an otherwise chaotic life.

To all my friends from home including Shubhankar Mishra, Unmukt Deswal, Amit Panda, Aditya Rajan Tigga, Surabhi Seth, Nipun Gupta, Aman Kumar, Saurabh Dubey, Neha Sood, Shreyas Raman, Lavanya Singh, Shivendra Tandon, Salman Mujtaba, Soumyashree, Saurabh Siddhartha, Lakshay Ahuja, Aseem Saxena, Mohan Sharma, Malvika Banerji, Romit Nath, Shantanu Bhide, Divyang Baldota, Dipika Dinesh, and Tariq Mudassir, thank you for staying in my corner across continents and years. You did not demand regular updates or perfect availability; you simply made it clear that when I had the capacity to show up, you would be there. The fact that I was never removed from any group chat, despite my erratic participation, is a kindness I do not take lightly. Your patience and your ability to treat me as the same person who left Delhi, not just as a

perpetually busy graduate student, gave me a continuity of self that was essential to finishing. You all were, and will remain, my family.

Beyond the people I have been fortunate enough to know in person, this journey has also been guided by voices I have only met on the page. I am deeply grateful to thinkers such as Jiddu Krishnamurti, Dr. Elinor Ostrom, and Dr. Daniel Kahneman, among many others, whose writings I stumbled upon along the way. They gave me language for doubt and curiosity, helped me make sense of suffering and uncertainty, and slowly turned abstract words like *love* and *empathy* into concrete, everyday practices. Their ideas arrived in the quiet hours when experiments failed, visas were uncertain, or the future felt narrow, and reminded me that the task is not only to finish a dissertation but to learn how to be fully human, preferably without turning every feeling into a five-point scale.

I am not religious, but I remain indebted to the culture that raised me, to the stories, songs, and practices that quietly taught me to value learning, to respect work done well, and to try to be a loving, creative, and productive person. Those roots have shaped how I see the world and why I care about water, infrastructure, and the communities that depend on them. This dissertation is one small attempt to honor a

responsibility that was recognized long before any climate models, asset management plans or risk analyses, in a simple Vedic prayer:

"आपो हि ष्ठा मयोभुवस् ता न ऊर्जे दधातन । महे रणाय चक्षसे ॥" (ऋग्वेद १०.९.१, आपः सूक्तम्)

["You are the source of happiness, O waters; grant us strength and nourishment, so we may perceive the highest truth." (Rig Veda 10.9.1, Āpaḥ Sūktam)]

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AADT: Average Annual Daily Traffic

AC: Asbestos Cement

ADT: Average Daily Traffic

AHP: Analytic Hierarchy Process

AMP: Asset Management Plan

ASCE: American Society of Civil Engineers

ASTM: American Society for Testing and Materials

AWWA: American Water Works Association

BIL: Bipartisan Infrastructure Law

BN: Batch Normalization

BW: Bar Wrapped

C300: AWWA C300 standard for RCCP

C909: AWWA C909 standard for PVC-O pipe

CI: Cast Iron; Confidence Interval (context-dependent)

CIP: Capital Improvement Plan

CISA: Infrastructure Security Agency

COF: Consequence of Failure

CP: Cathodic Protection

DDM: Design Decision Model

DI: Ductile Iron

DOT: Department of Transportation

DWINSA: Drinking Water Infrastructure Needs Survey and Assessment

DWSRF: Drinking Water State Revolving Fund

EC: Embedded-Cylinder

ECE: Expected Calibration Error

EM: Electromagnetic

EPA: Environmental Protection Agency

EVV: Evaluation, Verification, and Validation

F1216: Resin-Impregnated Tube (ASTM F1216)

FIS: Fuzzy Inference Systems

FMEA: Failure Modes and Effects Analysis

FRP: Fiber Reinforced Plastic

GA: Genetic Algorithm

GHG: Greenhouse gas

GIS: Geographic Information System

GPR: Ground Penetrating Radar

GRP: Glass Reinforced Plastic

GWP: Global Warming Potential

HDPE: High-Density Polyethylene

HGL: Hydraulic Grade Line

HMI: Human–Machine Interface

H0: Null hypothesis

HSM: Highway Safety Manual

ICUMAS: Integrated Conference on Urban Management and Safety

IQR: Interquartile Range

J100: AWWA J100 risk and resilience standard

KPI: Key performance indicator

LC: Lined-Cylinder

LCA: Life-Cycle Assessment

LCC: Asset life-cycle cost

LCCA: Life-Cycle Cost Analysis

LOF: Likelihood of Failure

LOFGT: LOF ground truth

LOM: Logistic Ordinal Model

LOS: Level of service

LR: Logistic Regression

LSLR: Lead Service Line Replacement

M28: AWWA Manual M28 (Rehabilitation of Water Mains)

MAUT: Multi-Attribute Utility Theory

MCDA: Multi-Criteria Decision Analysis

MCO: Multi-Criteria Optimization

ML: Machine Learning

MLP: Multi-Layer Perceptron

MSE: Mean Squared Error

NASSCO: National Association of Sewer Service Companies

NCA5: Fifth National Climate Assessment

NPV: Net Present Value

NRC: National Research Council

NRW: Non-revenue water

O&M: Operations and Maintenance

PCCP: Prestressed Concrete Cylinder Pipe

PE: Polyethylene

PE4710: PE4710 high-performance HDPE resin designation

PI: Performance Index

PRV: Pressure Reducing Valve

PSR: Pipe Survival Ratio

PVC: Polyvinyl Chloride

PVCO: Molecularly Oriented Polyvinyl Chloride

QA/QC: Quality Assurance / Quality Control

RBF: Radial Basis Function

RCC: Reinforced Concrete Cylinder

RCCP: Reinforced Concrete Cylinder Pipe

RF: Random Forest

RFT: Remote Field Technologies

RM: Risk Management

RPM: Risk-based Renewal Prioritization Models or Reinforced Plastic Mortar (context-dependent)

RTRP: Reinforced Thermosetting Resin Pipeline

RUL: Remaining useful life

RWT: Remaining wall thickness

SDR: Standard Dimension Ratio

SDWA: Safe Drinking Water Act

SETS: Socio-Ecological-Technical Systems

SGD: Stochastic Gradient Descent

SOM: Smallest of Maxima

SOS: System-of-Systems

SRF: State Revolving Fund (Clean Water and Drinking Water)

SSURGO: Soil Survey Geographic Database

ST: Steel

SVM: Support Vector Machine

SWIM: Sustainable Water Infrastructure Management (SWIM) Center

TBL: Triple-bottom-line

TCP: Traffic control plan

UQ: Uncertainty Quantification

USBR: United States Bureau of Reclamation

USGCRP: Global Change Research Program

USGS: United States Geological Survey

WRF: Water Research Foundation

XAI: Explainable AI

XGB: XGBoost

Chapter 1

Introduction

This chapter presents the problem, the evidence base, and the proposed solution pathway for this research. It first frames urban drinking-water networks as Socio-Ecological-Technical Systems (SETS) and summarizes why buried assets and institutional arrangements complicate planning under uncertainty. It then presents a brief overview of the current practice and literature to surface gaps in likelihood-of-failure, consequence-of-failure, and decision integration. A short policy and financing context explains the mandates and funding channels that shape water pipeline infrastructure renewal planning. The chapter then states the problem, motivation, goals, objectives, and hypotheses that guide the research. It concludes with the specific contributions, the scope and a brief discussion of the limitations of the proposed research.

1.1 Water Infrastructure System-of-Systems (SoS)

Urban drinking-water networks are large, spatially distributed SETS in which reliable service emerges from coupled interactions among engineered assets, institutions and human behavior, and biophysical processes. As complex adaptive systems, they exhibit nonlinearity, feedback, path dependence, and threshold effects, such that small perturbations can propagate across infrastructures and institutions (Ostrom, 2009; Meadows, 2008; Rinaldi et al., 2001; Chester & Allenby, 2019; Scheffer et al., 2009). These interactions are illustrated in Figure 1-1.

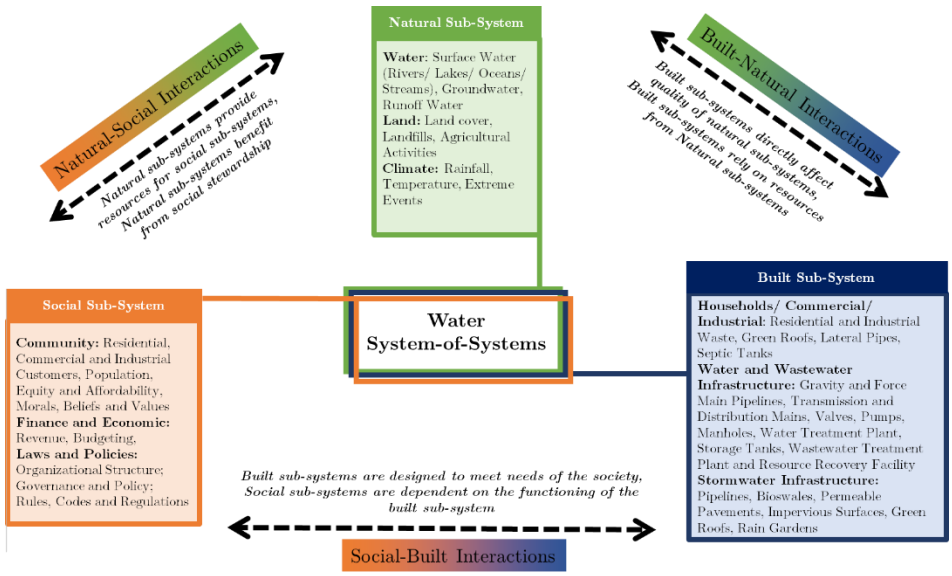

Figure 1-1: Water infrastructure System-of-Systems (SOS) perform under complex interactions between the natural, built and social systems

Since most assets are buried and designed to be unobtrusive, the system is largely invisible to the customers, which creates planning and accountability challenges. Sustaining such an invisible system requires a substantial institutional apparatus. A workforce of more than 1.7 million professionals across utilities, engineering firms, contractors, suppliers, and regulators plan, operate, renew, finance, and oversee source protection, treatment, transmission, distribution, and appurtenances such as hydrants, valves, storage tanks, reservoirs, and pumps (Tomer & Kane, 2018). Figure 1-2 provides a typical schematic showing the arrangements of these components along the source-to-consumer chain.

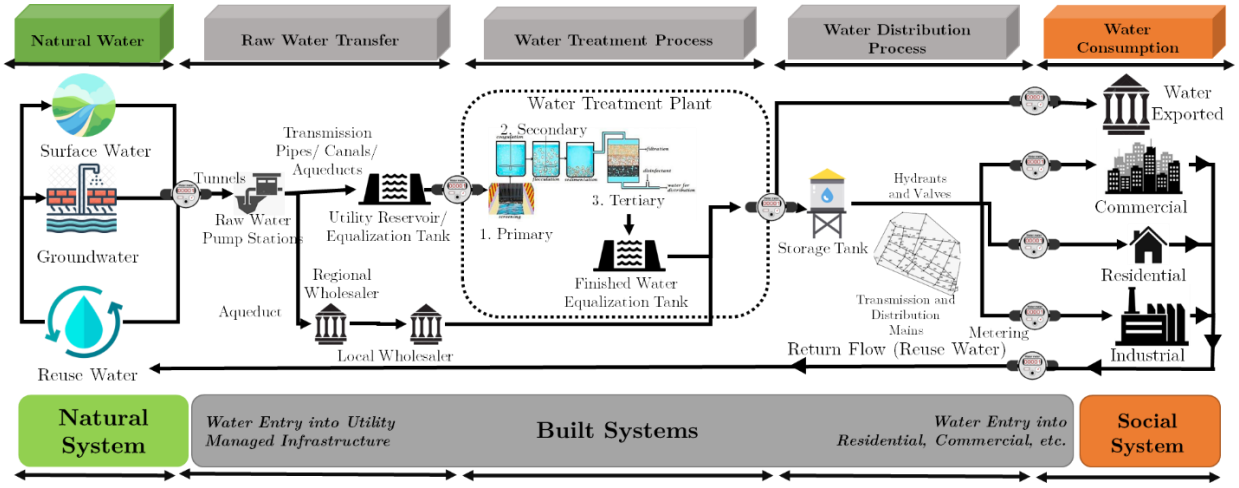

Figure 1-2: Drinking water service is made possible by complex chain of interactions between components across the natural, built and social systems

Within these SETS, buried pipelines transmission and distribution mains are the dominant and least observable asset class, and they are where support for renewal

decisions can reap the maximum benefit. At the national scale, the drinking water pipeline inventory exceeds two million miles, with renewal needs on the order of hundreds of billions of dollars over the next two decades (ASCE, 2025; USEPA, 2021). Utilities consistently rank aging infrastructure and financing among their top concerns; these pressures are compounded by intensifying hydrometeorological hazards like heat, drought, flooding that stress hydraulics and soils and raise disruption costs in densifying urban areas (AWWA, 2025; Bordreau et. al., 2022). Although sensing and data systems are improving, coverage and quality remain uneven across materials, geologies, vintages, and utilities, complicating benchmarking and model transferability (Sinha, 2021). Planning under uncertainty is therefore the norm, not the exception.

1.2 Policy Landscape

Building on these realities, U.S. renewal planning is shaped by a federal policy architecture that sets enforceable water-quality standards, mandates risk-informed resilience planning and directs substantial capital toward distribution-system upgrades. Under the Safe Drinking Water Act (SDWA), EPA promulgates and enforces National Primary Drinking Water Regulations for public water systems, anchoring compliance and

investment decisions. America's Water Infrastructure Act of 2018 amended SDWA §1433 to require community water systems serving >3,300 people to complete risk-and-resilience assessments and emergency response plans, explicitly tying asset condition, hazard exposure, and continuity of operations to capital planning. The 2024 National Security Memorandum on Critical Infrastructure Security and Resilience (NSM-22) further elevates risk-informed, consequence-mitigation planning across sectors and affirms EPA's water-sector risk-management role, reinforcing expectations for auditable, resilience-focused investments.

Although these programs set clear mandates and channels, only a subset directly funds the core problem of distribution-main renewal and prioritization. The DWSRF general supplemental ($11.7 B, FY2022–2026) can finance main replacement/rehab and planning/design, and states may use DWSRF set-asides for capacity development, asset management, and pre-construction support; however, set-asides rarely fund O&M activities like renewal (repair, rehabilitation or replacement) or condition assessment and are usually spent on new constructions. By contrast, the $15 B Lead Service Line Replacement (LSLR) and $4 B Emerging Contaminants streams are targeted (lead services and contaminant mitigation) rather than mains. AWIA §1433 requires risk-and-resilience

assessments and emergency response plans but does not appropriate new implementation dollars, and NSM-22 is a policy directive without funding. Against need, EPA's 7th Drinking Water Infrastructure Needs Survey and Assessment (DWINSA) estimates $625 B over 20 years, including $420.8 B (67%) for distribution/transmission. Even counting the full $30.7 B Bipartisan Infrastructure Law (BIL) appropriations to DWSRF, that is ~5% of total need, and the flexible portion for mains ($11.7 B) is ~3% of the distribution/transmission need are ample evidence of a persistent funding gap that heightens the value of better pipe renewal decision support methods. Finally, the Fifth National Climate Assessment documents intensifying hydrometeorological stressors on water infrastructure, underscoring the need for transparent, evidence-based renewal portfolios. In this policy context, the following section formulates the renewal portfolio problem that relates to selecting which pipelines to renew and when under regulatory, budgetary, and constructability constraints and motivates the explainable risk based portfolio framework developed in this research.

1.3 Problem Statement

Within this context, this dissertation addresses the pipeline renewal portfolio problem: determining which pipeline assets to renew and when, under budgetary, regulatory, and constructability constraints, and demonstrating using field and operational evidence that the recommended projects are warranted. The relevance of this problem is well established. Studies show that credibility and adoption depend on aligning renewal recommendations with asset-management principles and making them auditable against independent evidence; model fit alone is insufficient for real-world uptake (ISO 55000, 2014; Oreskes et al., 1994). To make the renewal portfolio problem concrete, we adopt standard asset-management units and terms (AWWA M28, 2014; AWWA M77, 2018; ISO 55000, 2014). These are listed in Appendix A.

1.4 Motivation

Armed with the common terminologies, we can map the methodological landscape more fully. Literature on pipeline risk and renewal fall into 5 major categories: (1) descriptive/heuristic and expert-rule approaches (e.g., break-rate indicators, age rules); (2)

statistical reliability/deterioration models count and point-process, survival/renewal, spatio-temporal and hierarchical Bayesian formulations; (3) mechanistic/physics-based models of corrosion, remaining strength, and transient loading; (4) machine-learning/AI and physics-informed hybrids (tree ensembles, boosting, neural nets) typically coupled with explainability; and (5) decision analytics and prescriptive optimization, which combine valuation frameworks (MCDA, benefit–cost) with integer/multi-objective, stochastic/robust portfolio design for clustering and scheduling under budgets, moratoria, and resource constraints. Across these categories, four gaps occur. First, LOF, COF and decision criteria are often characterized in an oversimplified way. Second, LOF, COF and delivery constraints are usually studied by different research groups and their integration under a common modeling framework has been explored in limited ways. Third, modeling studies are often subject to insufficient ground truth validation in operational settings and limited explainability of the results, which impedes real world adoption and continuous improvement (Kleiner & Rajani, 2001; Le Gat, 2008; Sinha, 2021; Belton & Stewart, 2012; Nemhauser & Wolsey, 1988; Deb, 2001; Birge & Louveaux, 2011; Breiman, 2001; Doshi-Velez & Kim, 2017). Chapter 2 elaborates these categories and gaps further and provides specific evidence for the proposed integrated modeling framework and validation protocols.

In practice, we see many gaps that necessitate this research. Many water utilities have limited in-house analytics capacity, so COF scoring is often heuristic and manually applied (e.g., "large diameter near a major road") during the manual project selection process, while LOF models frequently default to age/material proxies despite a richer scientific literature, largely because parameterization and curation of external covariates (soils, hydrology, land use) are time-consuming. Multi-criteria decision methods and portfolio selection are commonly outsourced to private vendors. These tools are proprietary and water utilities cannot audit how weights, trade-offs, or constraints produced a recommendation, which conflicts with best practice for decision transparency (Belton & Stewart, 2002; ISO 55000, 2014) and with the engineering expectation of explainability (Doshi-Velez & Kim, 2017). Interest in machine learning has surged, but single-utility datasets are typically small, sparse, and biased by local practices. Models trained on such data risk overfitting and weak transfer across systems unless supported by representative data, uncertainty quantification, and field validation (Domingos, 2012; Hastie, Tibshirani & Friedman, 2009; Oreskes, Shrader-Frechette & Belitz, 1994). In short, today's practice often under-specifies COF, over-weights simple LOF surrogates, and relies on opaque decision criteria. These conditions motivate the comprehensive, explainable, constraint-aware and field validated methods developed in this dissertation.

These gaps persist for several reasons. Buried-condition data are sparse, and time-resolved observations are rarer still. Historical records are biased across materials and installation vintages. Governance is siloed where multiple departments within a water utility work on different pieces of the puzzle. For example, in a typical water utility, asset management, operations, IT/data science, and design/planning department often work with limited sharing of preferences and decision rationales. Budgetary allocations are influenced by leadership preferences and short political terms. An aging workforce amplifies risk aversion, so visible short-term fixes are favored over novel but potentially better methods. Finally, model opacity further hinders adoption. Utilities need explainable methods with quantified uncertainty to build trust in capital improvement planning (Doshi-Velez & Kim, 2017; Ribeiro, Singh & Guestrin, 2016).

Against this backdrop, the outstanding need is a validated, explainable, and constraint-aware renewal portfolio framework that integrates LOF, COF, and renewal decision constraints, quantifies and communicates uncertainty and demonstrates performance against independent field and operational evidence. The timing is favorable as sensing and utility data for condition and context are expanding (e.g., standardized condition assessment and external covariates) (AWWA M77, 2018); modern optimization readily encodes

budgets, clustering, moratoria, and resource limits; and explainable machine learning offers tools to render recommendations auditable to engineers and managers (Doshi-Velez & Kim, 2017; Ribeiro, Singh & Guestrin, 2016). Anchoring such a framework in asset-management principles strengthens institutional fit and governance (ISO 55000, 2014). If successful, the proposed approach advances both scholarship and practice. It provides a mathematically structured basis for risk characterization and portfolio design, increases risk reduced per dollar while lowering avoidable disruption, supports public safety and continuity of economic activities in a service area, improves transparency and trust in CIP and yields repeatable methods that can generalize to other linear infrastructure systems.

1.5 Goals, Objectives and Hypotheses

Building on the identified gap, this study pursues an overall goal to develop robust risk models capable of accurately categorizing water pipeline segments into risk zones for accurate proactive renewal prioritization decision support. The goals and objectives outlined in this section will guide this research to identify, synthesize and configure

measurable risk signals into explainable and scalable computational models for water pipeline infrastructure systems.

Goal 1: Development of an AI model that predicts the performance for Cast Iron (CI), Ductile Iron (DI), Steel (ST), Prestressed Concrete Cylinder (PCC), Reinforced Concrete Cylinder (RCC), Bar Wrapped (BW), Polyvinyl Chloride (PVC), High Density Polyethylene (HDPE) and Asbestos Cement (AC) pipe material types based on structural and functional factors on a 0-5 scale.

- **Objective 1**: Develop a verified list of all parameters required to model pipe performance for pipe node segments in a geodatabase.
- **Objective 2**: Create an interpretable knowledgebase mapping input parameters to output performance index.
- **Objective 3**: Perform robust evaluation, verification and validation of the model based on artificial data, secondary data from utilities and ground truth data.

Goal 2: Development of an AI COF index prediction model for mixed criticality water pipeline systems.

- **Objective 1**: Develop a verified list of all parameters required to predict the impacts of any water pipeline failure.

- **Objective 2**: Create an interpretable knowledgebase mapping input parameters to output COF index.
- **Objective 3**: Perform robust evaluation, verification and validation of the model based on artificial data, secondary data from utilities and ground truth data.

Goal 3: Development of a risk model that can accurately prioritize water pipe candidates for renewal based on the optimization of multiple renewal decision criteria.

- **Objective 1**: Create a risk matrix screening tool with iso-risk-contours to categorize all water pipe node assets in any geodatabase into different risk zones for different types of risk management decisions.
- **Objective 2**: Develop a multicriteria decision support algorithm considering generalizable and practical criteria to create the most efficient priority list of water pipeline renewal projects under a given budget constraint
- **Objective 3**: Perform robust evaluation, verification and validation of the model based on artificial data, secondary data from utilities and ground truth data.

Achieving these goals requires formulation of specific questions and hypotheses that will guide the data collection and analyses. These are presented in detail in Chapter 3.

1.6 Approach Overview

This research presents an integrated approach to renewal prioritization that combines mechanism-aware LOF, modular COF with quantified uncertainty, and delivery constraints (budget, crew capacity, work-zone proximity, moratoria, outage windows) in a unified risk formulation optimized across segment → project → portfolio scales (ISO 55000, 2014; Rajani & Kleiner, 2001). Triple-bottom-line (TBL) consequences (social, economic, environmental) are incorporated alongside utility operational practices and renewal-action complexity to support short-, medium-, and long-term planning. A hybrid of possibilistic and probabilistic methods maps inputs to outputs: a fuzzy-logic teacher encodes expert rules to yield interpretable input–output relations, and an artificial-neural-network student is distilled on these relations to improve generalization while preserving traceable explanations (Doshi-Velez & Kim, 2017; Hinton et al., 2015). At recommendation time, the system provides explainability-in-use through an interpretable rule base, feature attributions via sensitivity or perturbation analysis, and concise rule-path summaries for operator review. The Evaluation, Verification and Validation section provides a multi-layered framework with detailed inspection and condition concordance tests and asset sampling methods. Results are supported with supplementary information like

quantification of trade-offs using risk-reduction-per-$1M, customer-hours avoided, reductions in work-zone conflicts, and budget utilization, with sensitivity and ablation studies for robustness. Reusable protocols standardize parameterization and outcome metrics; pattern-stratified synthetic datasets and utility datasets follow schemas that enforce units, ranges, and provenance, with fixed seeds and logs for reproducibility. In real-world tests with representative datasets, the model ranks and selects pipelines for inclusion in 5-year Capital Improvement Plans, reduces reactive renewals, and improves cost recovery by mitigating failure-impact costs through proactive planning.

1.7 Contributions

This research proposed 3 novel contributions to the body of knowledge of water main renewal prioritization:

- Generalizable protocols for input parameterization and outcome-metric selection that transfer across utilities and materials.

- A teacher–student AI methodology in which a fuzzy-logic expert system replaces opaque deep-learning teachers, enabling interpretable mappings for training, evaluation, and verification.
- A multi-layer validation protocol that advances risk-model assessment from single-metric accuracy to decision-relevant evidence aligned with field conditions and expert judgment.

1.8 Scope, Assumptions and Limitations

This study addresses a critical-infrastructure domain where water utilities must balance risk, resilience, and confidentiality under established standards and policy (ISO 31000:2018; PPD-21/CISA; EPA AWIA 2018). The geographic and asset scope is potable water distribution pipelines across representative U.S. utilities. The modeling and analyses cover common materials and diameter classes where data sufficiency permits. The work focuses on Operation and Maintenance (O&M) phase of the water pipeline lifecycle including the renewal decisions (repair, rehabilitation, replacement) rather than new construction, and excludes water-quality mixing dynamics, treatment-plant processes, and wastewater networks. Results depend on the fidelity and completeness of utility records,

audit quality, and local governance constraints; to mitigate over-generalization we use scenario and sensitivity analyses and align data hygiene with established water-audit practice (AWWA M36).

Additionally, ethical safeguards in this research are followed to reflect the sector's public-health mission. First, participating utility data are sanitized, governed by Non-Disclosure Agreements (NDAs), and reported in aggregate to avoid identification. Second, no proprietary technology is endorsed to prevent conflicts of interest. Third, sampling targets sufficient coverage to present findings across risk scenarios rather than to single out any one utility (PPD-21/CISA; EPA AWIA risk-and-resilience requirements).

1.9 Dissertation Outline

Figure 1-3 shows the outline of this dissertation. Chapters 1 & 2 introduce the topic with SETS framing, practice gaps, objectives and scope. Chapter 3 explains the study design, data sources, model stack, uncertainty quantification. Chapters 4, 5 and 6 explain the specifications for all 3 proposed models like I/O Parameterization, teacher-student strategy, algorithm details and model evaluation, verification and validation

results and final Chapter 7 provides a discussion comparing the findings to the objectives of this research along with the detailed limitations and viable future work directions

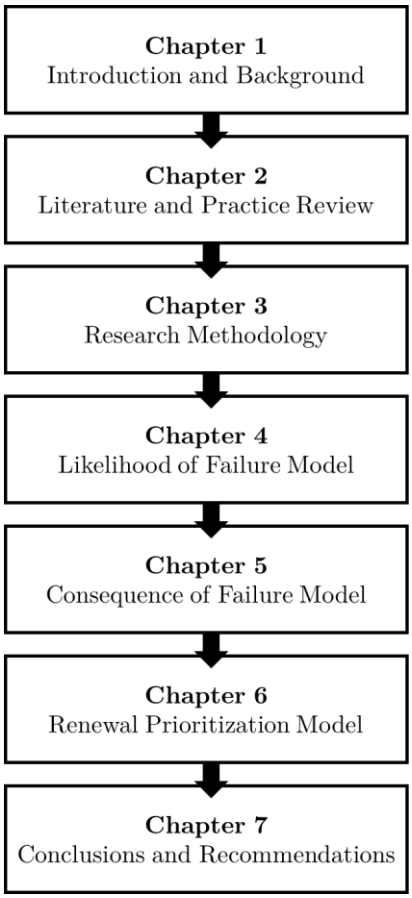

Figure 1-3: Outline of this dissertation showing the main topics discussed in each chapter

The next chapter presents a review of the current state of research and practice in renewal prioritization, detailing where existing methods succeed, where they fall short, and the specific gaps this dissertation addresses.

Chapter 2

Review of Literature and Practice

This chapter maps the state of the art and current practice in water-pipeline renewal decision-making. It first lays out the review protocol, databases searched, time window, screening criteria, and a schema for evaluating study types, data, features, methods, metrics, and the quality of uncertainty quantification, explainability, and validation protocols. It then quantifies the corpus with bibliometrics, and evidence maps across five domains namely, LOF, COF, renewal prioritization, model EVV (evaluation, verification, validation, uncertainty, explainability, openness), and adjacent domains. A narrative synthesis follows, presenting the substantive review and key findings by domain, with representative methods, data regimes, and evidence strength highlighted and interpreted in the context of utility decision needs. Next, a structured matrix maps what the literature provides (method capabilities, data utilization, validation level) to what utilities require in practice (CIP cadence, budget/moratoria rules, crew and permit constraints, auditability). The chapter then assesses how studies report verification, validation, uncertainty,

and explainability, and identifies persistent gaps and meta-issues. Finally, it distills design requirements that feed directly into the modeling choices and the evaluation, verification, and validation protocols framed by Goals 1–3 and shows how these requirements map to later chapters.

2.1 Review Scope and Protocol

This review synthesizes evidence on (i) LOF models for potable water pipelines, (ii) COF modeling across social, economic, environmental, public-health, equity, and mobility pathways, and (iii) prescriptive portfolio and delivery-constraint formulations that translate LOF–COF into implementable renewal plans. The review begins by situating the research at the intersection of several knowledge domains rather than within a single methodological silo. Figure 2-1 synthesizes those domains in a cross-sectional Venn diagram. The left lens ("Water Pipeline Infrastructure Systems") anchors the physical and operational context related to materials and deterioration, hydraulics and pressure, GIS/topology, and model verification and validation. The upper lens ("Decision Science") covers risk-based asset management, risk matrices, and multi-criteria and multi-objective

optimization. The right lens ("AI") spans machine learning, deep learning, and fuzzy logic, with data-science workflows that support uncertainty quantification and explainability.

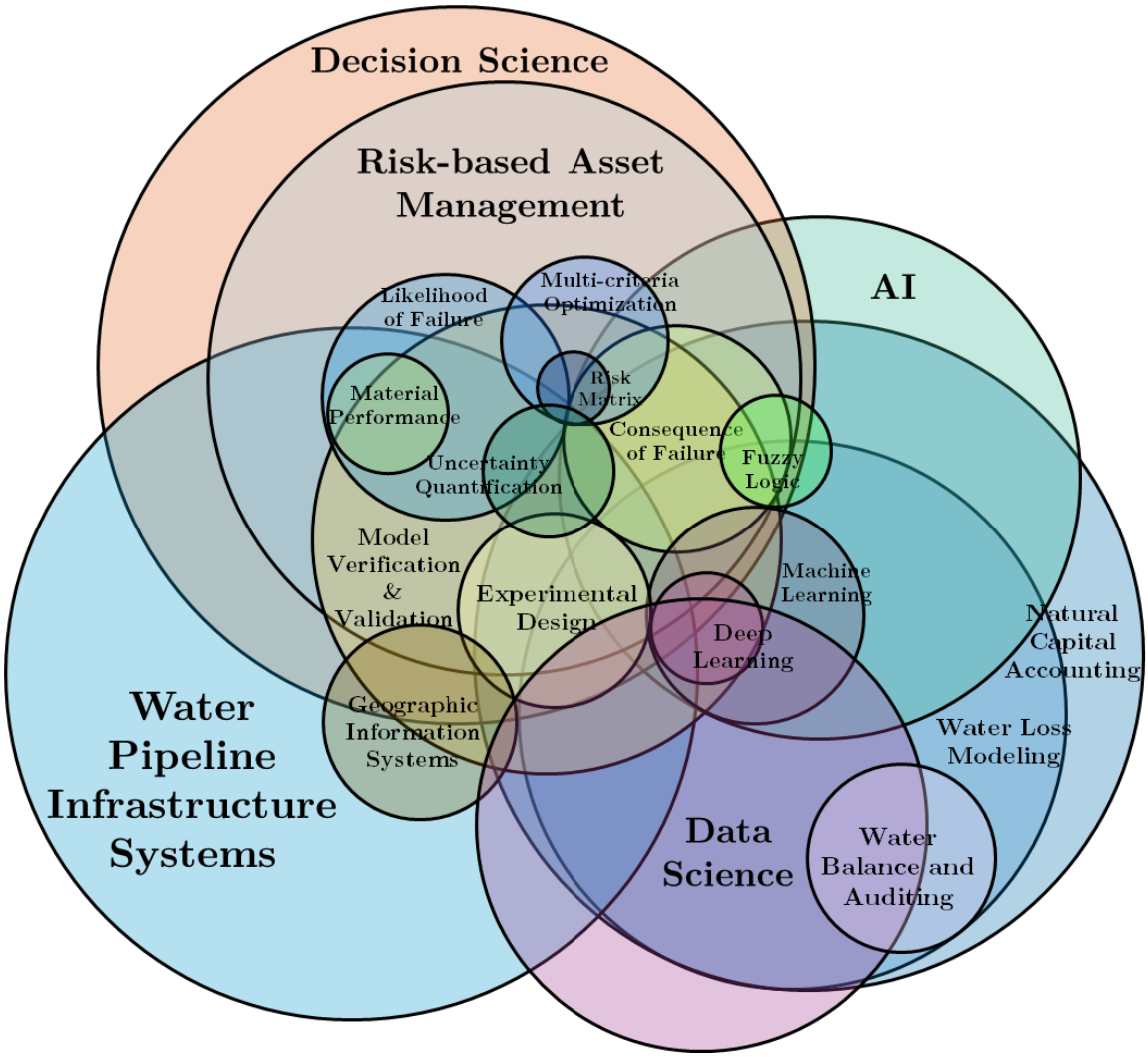

Figure 2-1: Research domains reviewed as part of the literature review in this research

Overlaps emphasize where this dissertation operates: LOF and COF modeling sit at the junction of asset science, decision methods, and AI; experimental design and model

EVV connect these methods to practice; and adjacent practices such as natural-capital accounting, water-loss analysis, and water-balance auditing inform how consequences and benefits are measured. The diagram is schematic (not to scale) and is used to communicate scope and integration: methods are selected and evaluated where these domains intersect to produce decision-ready evidence.

Complementing the literature synthesis, a structured practice review grounds the research in real utility conditions. The accompanying map (Figure 2-2) shows U.S. water utilities that provided information through interviews, data exchanges, and document review. This engagement leverages Virginia Tech's Sustainable Water Infrastructure Management (SWIM) Center partnerships with utilities, technology providers, consultants, manufacturers, and standards bodies, and draws on work conducted under the PIPEiD project ("Collection and Compilation of Water Pipeline Field Performance Data," OMB #1006-0031) funded by the U.S. Bureau of Reclamation and the U.S. Geological Survey. Participants span a wide range of system sizes (serving ~10,000 to ~5,000,000 people), supply roles (retail, wholesale, hybrid), and ownership models (public, private, federal) (Sinha 2021). The practice inputs were used to characterize current analytics capabilities, data limitations, and implementation constraints (budgets, crews, moratoria, outage

windows, work-zone proximity, permitting), and to understand requirements for auditability and explainability in capital planning. Marker locations indicate participating utilities and partners; inclusion denotes engagement rather than endorsement, and the sample is broad and intended to denote the statistical representation of the practice review.

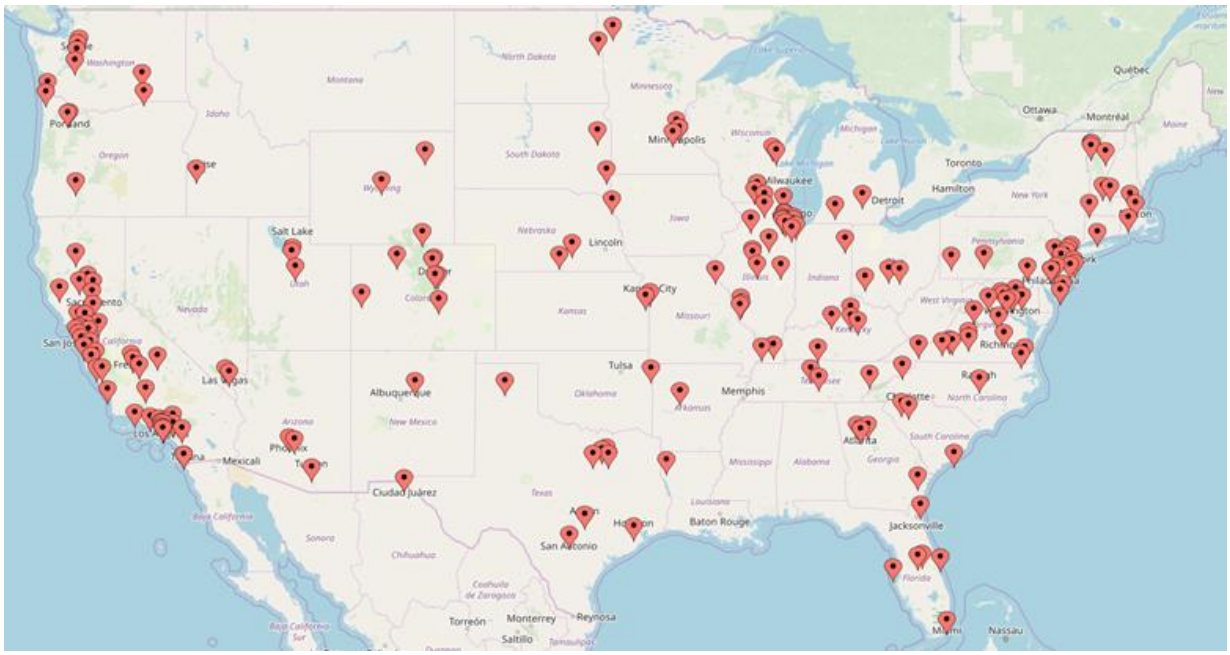

Figure 2-2: Practice review from utilities (of varying sizes and ownership types) across the US for collecting quantitative and qualitative data related to pipeline performance, failure impacts and decision criteria

Together, the cross-disciplinary literature map (Figure 2-1) and the practice landscape (Figure 2-2) define the evidentiary and operational space for this dissertation. They motivate the design choices taken in subsequent chapters and bound the specific research questions summarized in Table 2-1, ensuring that proposed models are not only

methodologically sound but also implementable, auditable, and aligned with utility decision processes.

Table 2-1: Research questions (RQs) to guide the study design choices

Research Goals	**RQ**	**Description**
Goal 1: Pipe Performance	1.	What are the key structural and functional factors that can predict the performance of different water pipe materials?
	2.	How to mathematically map relationships between pipe performance parameters to reliably predict the performance index of a pipe node on a 0-5 scale?
Goal 2: Consequence of Failure	3.	What are the key economic, environmental, social, operational and renewal factors that can predict the consequences of failure of water transmission or distribution pipelines?
	4.	How to mathematically map relationships between different COF parameters to reliably predict the COF index of a pipe node on a 0-5 scale?
Goal 3: Risk Model	5.	How to effectively screen and further optimize risk rankings of water pipeline candidates based on utility specific decision criteria to support proactive renewal programs?
Overall Goal: Model Validation	6.	How to develop and implement protocols for validating the proposed renewal prioritization models that can reliably describe the model uncertainties?

Because adoption depends on practice fit, the scope explicitly includes standards, guidance, and utility reports alongside peer-reviewed research, and it considers adjacent domains (complexity science, systems-of-systems, policy/regulatory, public health, and systematic-review methods) where they inform problem formulation, risk modeling, decision processes and experimental design.

2.2 Search Strategy

The search spans Web of Science, Scopus, Engineering Village (Compendex), ASCE Library, IWA Publishing, IEEE Xplore, and Google Scholar for recall, plus targeted gray-literature sources (EPA, AWWA, ASCE, DHS/CISA, state regulator portals, utility CIP documents, vendor white papers). The time window captures contemporary methods (approximately 2000–present) with backward chaining to seminal work. Query strings combine controlled terms and free text for assets, materials and diameters, modeling families (statistical, mechanistic, ML, optimization), decision analytics, uncertainty, explainability, and validation. Consistent with PRISMA 2020 (Page et al. 2021), the review reports the information sources, date ranges, and full search strings; the two-stage selection process with inclusion/exclusion criteria and recorded reasons; and a flow diagram summarizing records identified, deduplicated, screened, assessed for eligibility, excluded with reasons, and included. Data-extraction items (study type, data regime, materials/diameters, features, methods, performance metrics, uncertainty, explainability, validation, implementation status) and a study-class-specific quality appraisal are specified. This chapter aims to map coverage and performs a structured charting of characteristics and

evidence maps across the five domains with heterogeneity described narratively (Tricco et al. 2018).

2.3 Screening

Database and grey-literature searches identify all relevant literature candidates. A deduplication step is used for screening entries across sources, and keywords in title/abstract retains items on potable water pipeline assets (or methods clearly transferable to potable mains) that address at least one focal domain: LOF, COF, renewal prioritization/constraints, model EVV (evaluation, verification, validation, uncertainty, explainability, openness), or adjacent domains that inform pipeline risk and decision science (complexity science, systems-of-systems, policy/regulatory, public health, or systematic-review methods). Screening excludes opinion/news pieces, non-method items, and studies without an evaluative metric. Full-text review then confirms eligibility against the same content criteria, applies feasibility checks (English language and accessible full text), and records standardized reasons for exclusion (out of scope; insufficient methods; no evaluative metric; non-transferable domain; inaccessible full text). The reported counts measured as records identified, after deduplication, titles/abstracts screened, records excluded at

title/abstract, full texts assessed, full texts excluded with reasons, and studies included match the boxes in the flow diagram shown in Figure 2-3.

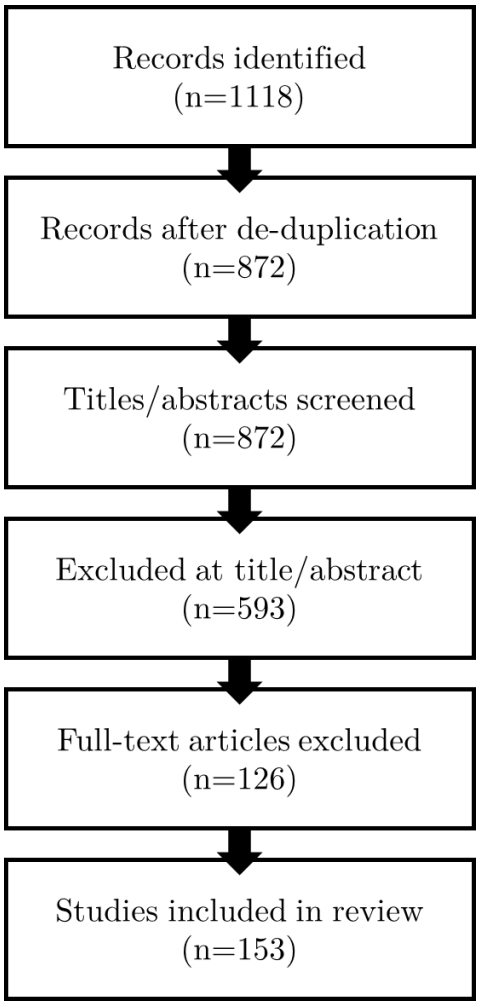

Figure 2-3: Flow of steps in literature review

Figure 2-4 shows annual inclusions to the corpus following the screening workflow. Counts prior to 2000 are sparse because pre-2000 coverage was limited to seminal, field-defining works used to anchor concepts and methods, whereas post-2000 literature was

comprehensively searched across databases and grey sources. Accordingly, the series reflects the review's inclusion protocol rather than a census of all publications; the pre-2000 segment should be read as a curated baseline, not a representative tally.

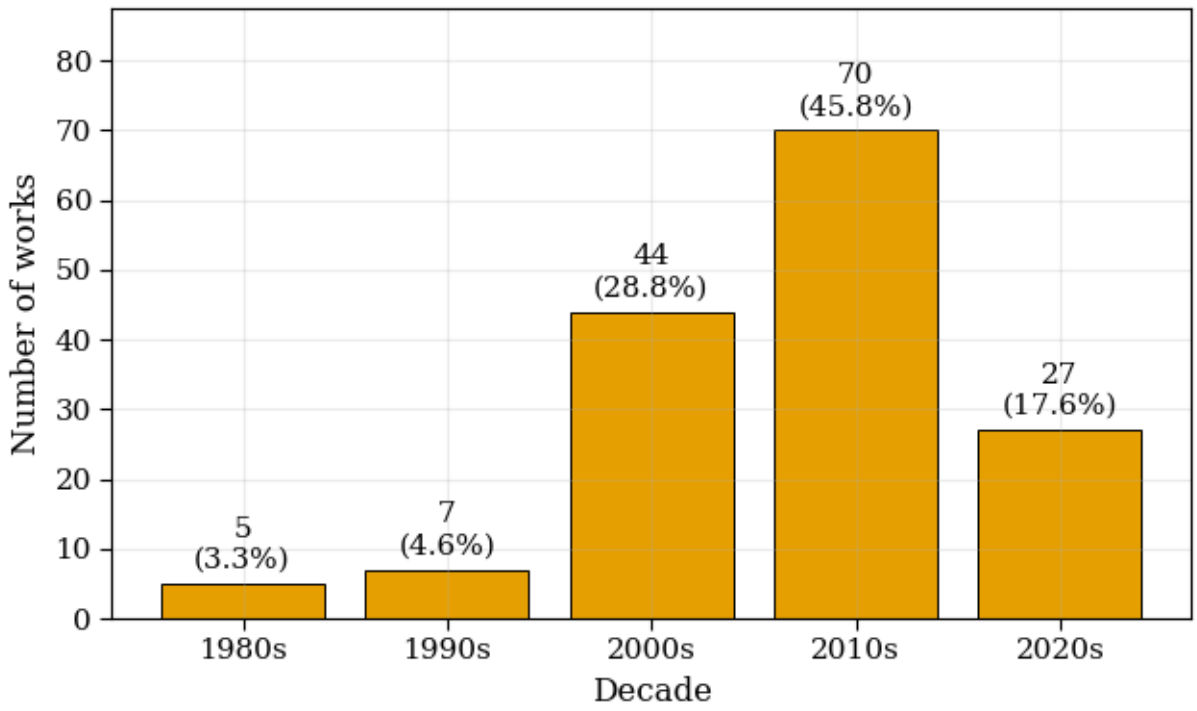

Figure 2-4: Count of studies by each decade included in the literature review after screening and deduplication (Records plotted: 153)

2.4 Coding schema

Each included study is classified using a structured dictionary that mirrors the five evidence domains presented later in the heatmaps for LOF, COF, Renewal Prioritization,

Model EVV, and adjacent domains, together with baseline bibliographic/context fields. The record for each study captures its profile (study type: Statistical, Mechanistic, ML, Hybrid, Optimization, or Practice/Standard; publication year, venue, geography or utility context, and data regime such as single versus multi-utility and approximate scale). LOF coverage records the model family (statistical/mechanistic/ML/hybrid), whether an age-only proxy is used, the degree of mechanism awareness, the training protocol, and the LOF performance metrics reported (for example, AUC, Brier score, and calibration). COF coverage records which consequence pathways are modeled (service disruption, economic, environmental, public-health, equity, and traffic/transport), the units or metrics used (for example, customer-hours, cost, or incident counts), and whether consequences are linked to network and land-use context. Renewal Prioritization and constraints capture the method class (MCDA/benefit–cost, multi-objective, stochastic/robust) and whether delivery constraints are encoded (budget, crew capacity, moratoria/permits, outage windows, and work-zone proximity). Model EVV records validation (internal; external hold-out; inspection/field concordance), uncertainty treatment (calibration, prediction intervals, scenario/sensitivity), explainability (global, local, rule-based/XAI), and openness (code/data availability). Adjacent domains captures explicit use or adaptation of

complexity science, systems-of-systems, policy/regulatory analysis, public-health framing, or systematic-review methods in ways that inform pipeline risk or decision design.

Counting rules and reliability: Because many studies span multiple topics, records are multi-labeled (e.g., LOF + COF + EVV). The Hybrid label is assigned only when multiple LOF model families are first-class parts of the contribution (not merely baseline comparators). Ambiguities are resolved using concise precedence rules (definitions and examples). These encoding rules and features studied in each domain category is presented in Table 2-2. The five heatmaps compute row-wise proportions within study-type strata and annotate each row with n to indicate the base sample size.

Table 2-2: Coding Dictionary used for organizing literature

Domain	Features Studied	Examples / encoding
Study profile	StudyID; Title; Year; Venue; Geography; Utility context; Data regime; Asset class; Materials; Diameters	Data regime: Single Utility Small / Single Utility Large / Multi Utility. Asset class: Distribution / Transmission / Both. Materials/diameters: semicolon-separated.
LOF	LOF model family; Age-only proxy; Mechanism-aware; Training/eval protocol; LOF metrics	Family: Statistical / Mechanistic / ML / Hybrid (multi-select if first-class). Age-only, Mechanism-aware: 0/1. Protocol: Random CV / Temporal Split / Spatial Split / External Holdout / Other. Metrics: AUC; Brier; Calibration; etc.
COF	Pathways; Units/metrics; Network/land-use linkage	Pathways: Service, Economic, Environment, Public-health, Equity, Traffic (0/1 each). Units: Customer Hours; Cost; Incidents. Linkage: 0/1.
Renewal prioritization & constraints	Method class; Constraints encoded	Method: MCDA/Benefit–Cost; Multi-objective; Stochastic/Robust (0/1 each). Constraints: Budget; Crew; Moratoria/Permits; Outage windows; Work-zone proximity (0/1 each).
Model EVV (Evaluation/	Validation; Uncertainty; Explainability; Openness	Validation: Uncertainty quantification, Model Performance Metrics and Ground Truthing (0/1 each). Uncertainty: Calibration; Prediction intervals; Sensitivity (0/1). Explainability:

Verification/ Validation)		Global; Local; Rule-based; XAI (0/1). Openness: Open code; Open data (0/1).
Adjacent domains	Complexity science; Systems-of-systems; Policy/regulatory; Public health; SLR methods	Each 0/1 if explicitly used to inform pipeline risk or decision design.
Implementation maturity	Prototype; Pilot; Production	Pick most advanced demonstrated stage.
Admin	Inclusion decision; Exclusion reason; Comments	Inclusion: Include/Exclude; Exclusion reason: Out of Scope/ Insufficient Method/ No Metric/ Non Transferable/ Inaccessible.

2.5 Quality Evaluation Rubric

Quality evaluation is tailored by study class and recorded with the coding. Prediction studies (statistical/ML/hybrid) are checked for transparent data partitions, handling of missing data, calibration reporting, and risks of overfitting with appropriate validation. Mechanistic studies are assessed for parameter provenance, identifiability, boundary conditions, and validation against lab/field observations, with sensitivity analysis expected. Optimization/ portfolio studies are evaluated for objective/ constraint fidelity to utility realities (budgets, moratoria/permits, crews, outage windows, proximity), solution quality (e.g., optimality gap or convergence diagnostics), robustness (scenario/stochastic treatment), and explainability-in-use (traceable trade-offs/audit trails). Practice/ standards/ gray literature is appraised for governance clarity, auditability, reproducibility, evidence

provenance, and date/currency. This review maps coverage and design requirements; no effect-size pooling is attempted, and heterogeneity is described narratively.

Table 2-3 lists the criteria (C1–C16), what each criterion checks, evaluation anchors and the study classes to which it applies. Three criteria are highlighted as decision-critical across many contexts: external/temporal validation (C6), inspection/field concordance (C7), and constraint fidelity and implementation ability (C10). "NA" is allowed when a criterion does not apply (e.g., optimality gap for a purely predictive study).

Table 2-3: Rubric (evaluation and scoring criteria) for evaluating quality of the selected literature

ID	Criterion	Evaluation Criteria	Absent/ Poor	Partial/ Adequate	Strong/ Best
C1	Problem framing & data transparency	Scope, asset class, cohort, data regime	Vague scope; dataset not described	Basic scope; partial dataset description	Clear scope; dataset, cohort, filters fully described
C2	Outcome definition & labeling	Failure/impact definitions; label quality	Ambiguous/ unstable labels	Defined but noisy; limited checks	Precise, auditable labels; QA/QC described
C3	Predictor handling & missingness	Feature provenance; imputation	Missingness ignored; ad-hoc features	Minimal imputation; limited provenance	Systematic handling; provenance + ranges/units
C4	Model specification & training protocol	Leakage control; partitions	Possible leakage; unclear splits	Clear splits; some leakage risk	Leakage addressed; temporal/utility splits justified
C5	Internal validation	CV/bootstrapping/hold-out	None	Basic CV/hold-out	Robust CV with variance; calibrated reporting
C6	External/ temporal validation	Generalization across time/utility	None	Limited (small or weak)	Clear external/temporal hold-out with rationale
C7	Inspection/field concordance	Agreement with field/inspection	None	Limited spot checks	Systematic concordance or blinded checks
C8	Uncertainty quantification	Calibration; intervals; sensitivity	None	Single method only	Calibration + intervals and/or rigorous sensitivity
C9	Explainability	Global, local, or rule-based	None	One view (e.g., feature ranks)	Complementary views (global + local or rule-base)

ID	Criterion	Evaluation Criteria	Absent/ Poor	Partial/ Adequate	Strong/ Best
C10	Constraint fidelity & implementation ability	Budgets, crews, moratoria, outages, proximity	Abstract or omitted	Some constraints encoded	Realistic encoding + parameterization sources
C11	Optimization quality	Objective clarity; optimality/convergence	Not reported	Heuristic with weak diagnostics	Optimality gap or robust convergence evidence
C12	Parameter provenance & identifiability	Mechanistic parameters & sensitivity	Unclear sources	Mixed provenance; limited identifiability	Traceable sources; identifiability + sensitivity
C13	Experimental/ lab/ field validation (mechanistic)	Empirical corroboration	None	Limited/ surrogate	Targeted tests or field back-checks
C14	Openness (code)	Reproducible code	None	Available upon request	Public repository with instructions
C15	Openness (data)	Data or schema/ protocol	None	Partial sample/metadata	Public data or detailed access protocol

2.6 Literature Corpus Overview

The corpus is first profiled with descriptive bibliometrics to show where the evidence is concentrated and how it has evolved. Because distribution mains dominate exposure in practice and more than four-fifths of U.S. pipe inventory is under 16 inches, asset-class coverage is stratified by distribution vs. transmission and further by diameter bins (<16 in, 16–36 in, >36 in) and material families (cementitious, metallic, plastic) to reveal whether methods validated on small-diameter networks also extend to larger, high-consequence mains. Method-family shares (Statistical, Mechanistic, ML, Hybrid, Optimization, Practice/Standards) are tracked over time to show shifts from age/material

proxies toward mechanism-aware and hybrid approaches and from descriptive screening toward portfolio optimization. Validation practice is summarized across three evidence categories: Uncertainty Quantification (UQ), model-performance metrics, and ground-truth protocols. UQ includes sensitivity analyses, feature-importance summaries, and attribution methods (e.g., SHAP). Performance metrics include RMSE/MAE for regression, ROC/AUC, and Precision/Recall/F1 for classification. Ground-truth protocols cover explicit testing and concordance procedures (e.g., blind testing parameters, experimental setup, and model-agreement protocols).

Figure 2-5 tracks the share of included studies that report each category of evaluation evidence by publication year: Uncertainty Quantification (UQ), model-performance metrics, and ground-truth protocols. Performance metrics are most prevalent across the period, UQ appears slightly less frequently, and ground-truth protocols remain rare. Even when some level of previously unseen data is collected for ground truth testing, the studies lack detailed protocols that explain the validation experimental design in detail to support reproducibility. Each line shows the within-year proportion of studies that include at least one subfeature from the category (e.g., ROC/AUC, RMSE/MAE, or Precision/Recall/F1 for performance; sensitivity analysis, feature importance, or attribution for UQ; blind

testing parameters, experimental setup, or model-agreement protocols for ground truth). The visualization highlights the relative emphasis of reporting practices over time and underscores the persistent scarcity of explicit ground-truthing in water pipeline performance modeling.

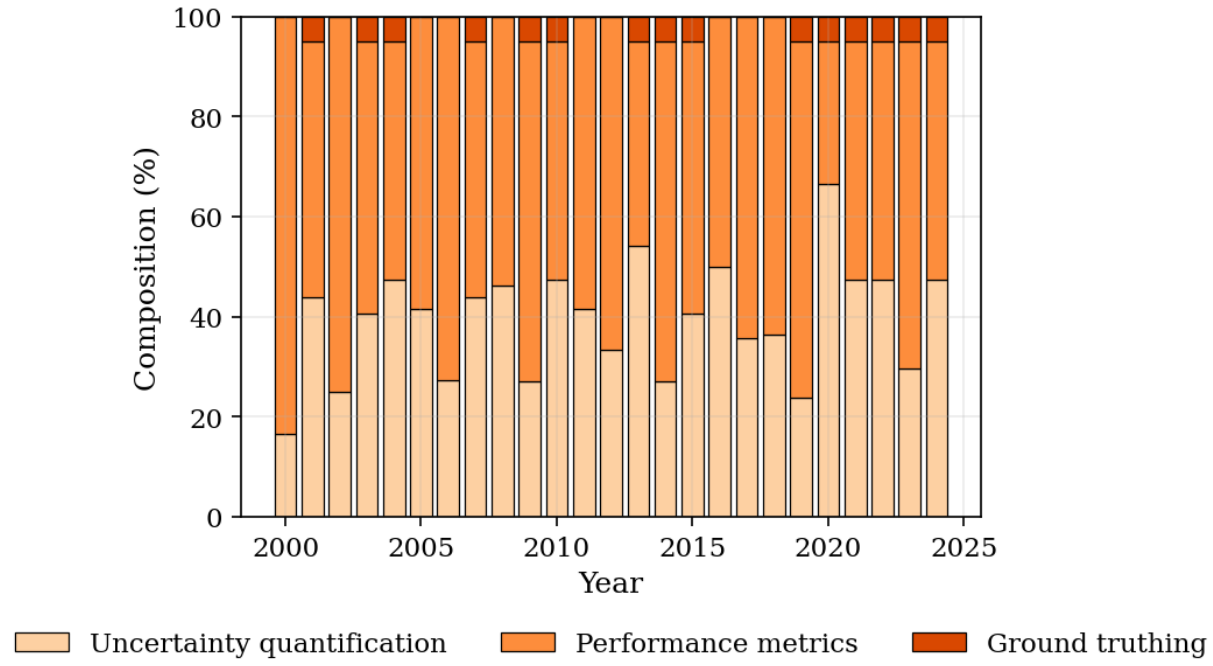

Figure 2-5: Percentage of selected studies in the 3 levels of EVV organized by the year of publishing date

The next subsection is a narrative synthesis of literature and practice organized into recurring themes that help establish point of departure, organize data, setup the renewal prioritization models and guide the methodological choices in this research.

2.7 State-of-the-art Review

This section outlines the significance of water pipeline infrastructure, the challenges posed by these aging systems, and the necessity for improved risk based renewal modeling. It highlights the current knowledge gaps in literature and lays the foundation for the proposed models by demonstrating their relevance and potential impact.

2.7.1 Water Pipeline Infrastructure Systems

The U.S. has over 2 million miles of water pipelines, essential for public health, economic stability, and societal development. Many of these pipelines are reaching the end of their service lives (ASCE 2025), leading to increased failures, service disruptions, and economic strain on utilities. AWWA surveys confirm that renewal and replacement of aging infrastructure have been utilities' top concern for the past five years (AWWA 2025).

Water pipeline systems are broadly classified into transmission mains, distribution mains, service mains, and plumbing systems. Transmission systems, typically large-diameter pipelines, transport water over long distances, while distribution systems consist of smaller-diameter pipes, designed to maintain adequate pressure for consumers. Over 80% of U.S. water pipelines are less than 16 inches in diameter (Sinha 2021), reflecting the

dominance of distribution systems (Figure 2-6). In this chapter LOF is decomposed into three pathways, structural corrosion/deterioration, structural loading and stress from traffic/soil/hoop pressure, and functional hydraulics/capacity to distinguish age-only surrogates from mechanism-aware models and to support targeted validation.

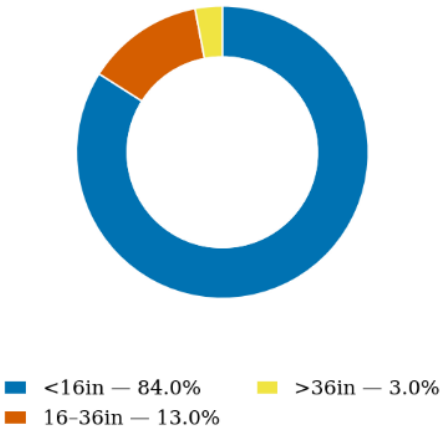

Figure 2-6: Percentage distribution of water pipelines based on diameter categories (Less than 16 inches, 16-36 inches and greater than 36 inches) (Sinha 2021)

The choice of pipeline material varies based on diameter, soil conditions, environmental factors, and lifecycle costs (Figure 2-7), with materials classified into cementitious, metallic, and plastic categories. These materials interact with their surroundings in complex ways, influencing failure modes and service life (Figure 2-9).

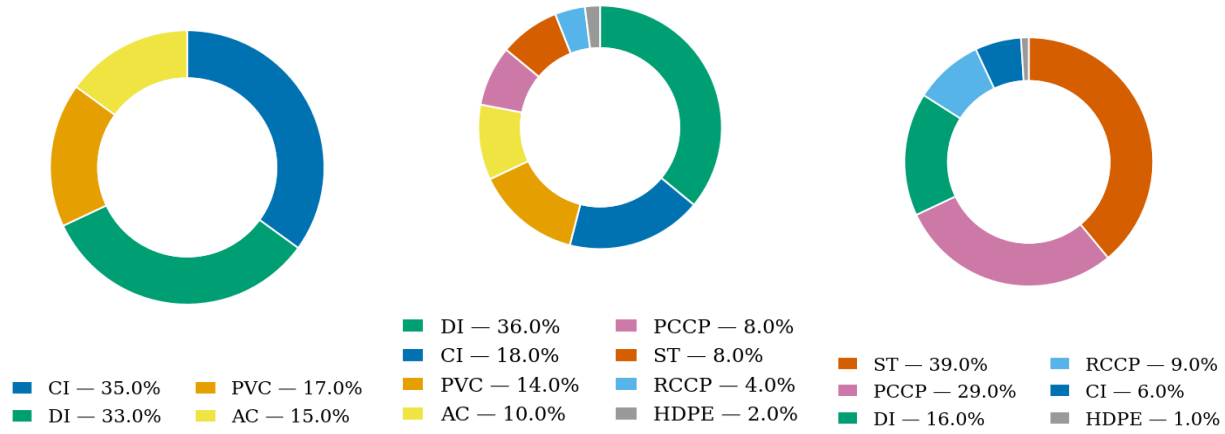

Figure 2-7: Left: Material percentage distribution for <16in diameter pipelines, Center: Material percentage distribution for 16in-36in diameter pipelines, Right: Material percentage distribution for >36in diameter pipelines (Sinha 2021)

2.7.2 Performance Characteristics of Water Pipeline Materials

Selecting pipe materials for commissioning into a distribution and transmission system is a core asset-management decision. Past performance matters, but it must be weighed alongside current manufacturing standards, design guidance, local exposures (soil, traffic, pressure, climate), construction quality assurance and future lifecycle costs. Standards and manuals are periodically revised to reflect improved processes (e.g., casting, linings/coatings, polymer resins), better hydraulic and structural design methods, and new field knowledge. Aspects related to installation oversight like trench support, bedding, jointing, torque, pressure testing, and disinfection are equally critical. Poor construction

can erase the advantages of a well-chosen material (AWWA, 2016; 2017; 2018; 2020/2022).

This section presents lifecycle material properties, deterioration mechanisms, and operational contexts to support LOF modeling. This section traces shifts in the usage of pipe materials and their characteristics over more than 2 centuries. The timelines link those shifts to observed break patterns, failure modes, and consequence profiles (Rajani & Kleiner, 2001; Folkman, 2018; AWWA M-series). The goal is to define failure mechanism-aligned features, units, and ranges; capture typical behaviors as well as outliers; and provide a defensible basis for the fuzzy "teacher" parameters, membership functions and rules and the subsequent student learners.

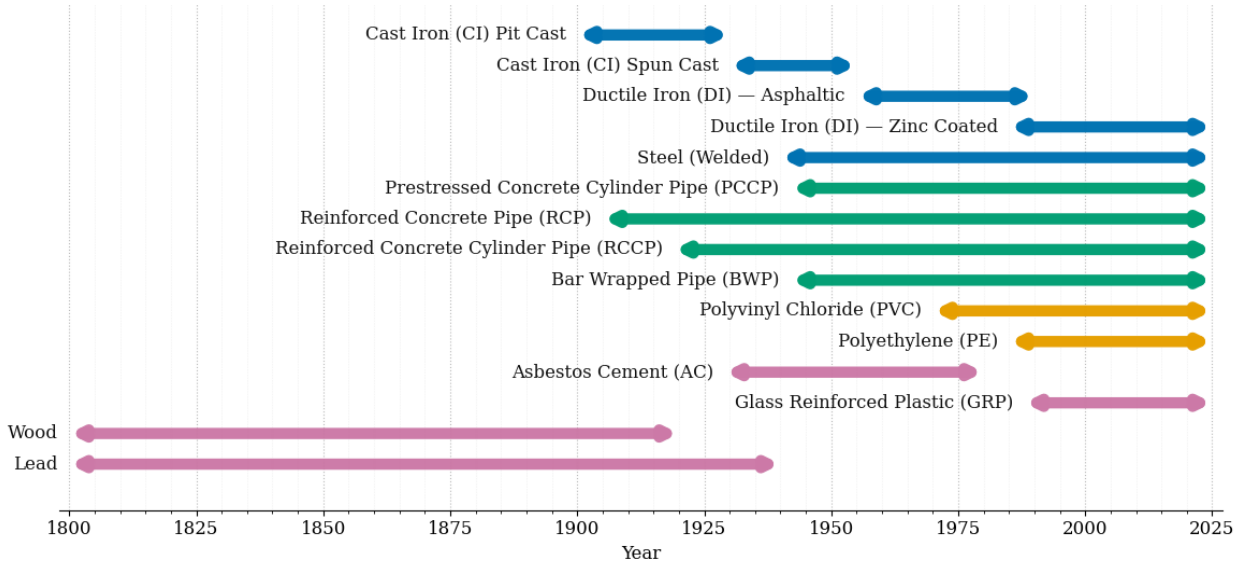

Figure 2-8: Timeline of usage of typical water pipeline materials since 1800s

Past research has explored aspects of pipeline performance, focusing on failure modes, rates, and remaining useful life (Burn 2006; Davis et al. 2008; Hu et al. 2013). However, these studies are subject to epistemic uncertainties due to challenges in directly observing buried pipelines, requiring reliance on proxy parameters and assumptions. Additionally, pipeline criticality assessments face aleatory uncertainties, as failure consequences vary across residential, commercial, industrial, agricultural, and environmental sectors, which dynamically interact (Cromwell 2002; Gaewski and Blaha 2007; Raucher 2005; Raucher 2017).

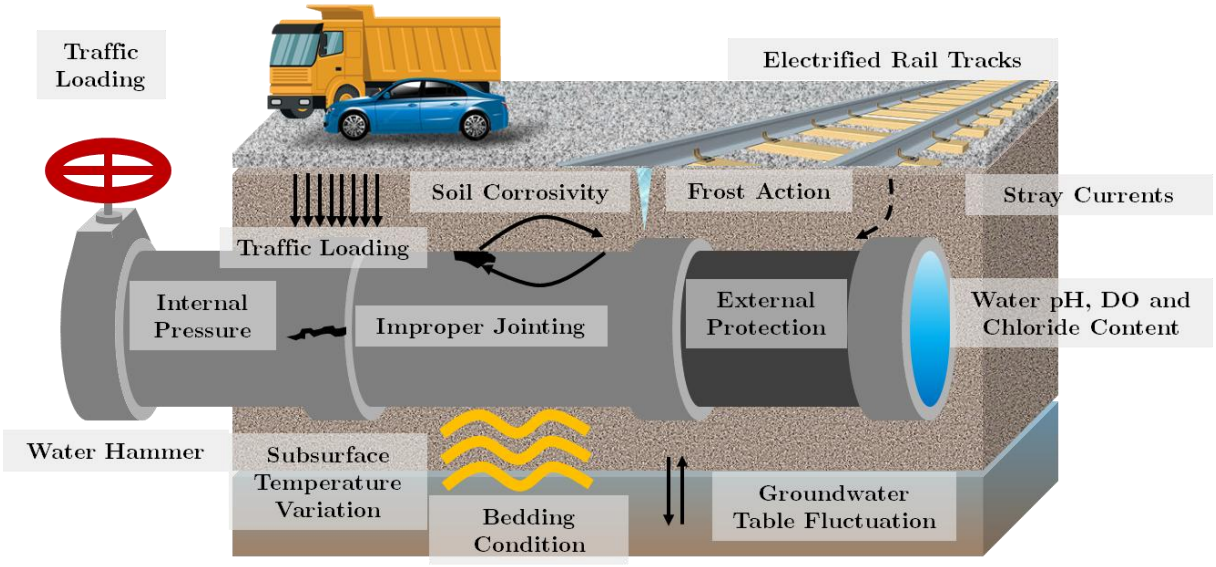

Figure 2-9: A variety of internal, external factors (including the pipe characteristics itself) influence the performance of an operational water pipeline

2.7.2.1 Metallic Pipe Materials' Performance Characteristics

Metallic pipeline materials (cast iron, ductile iron, and steel) still account for a large share of U.S. distribution mileage and span the widest range of installation eras and environments. Their performance varies strongly with the physical and chemical characteristics of soil including moisture, redox potential, pH, among other factors. Understanding material evolution, appurtenances (joints, valves, hydrants), and corrosion control is central to life-cycle management. This section consolidates the manufacturing and

standards history, corrosion prevention approaches, typical failure modes, and what prior studies imply for today's design and renewal choices (Rajani et al., 2011; WRF, 2016).

With improvements in metallurgy and casting, nominal wall thicknesses decreased over time, especially as DI replaced gray cast iron (CI). DI's nodular graphite dramatically increases strength and impact resistance compared with CI, but thinner walls mean less inherent corrosion allowance. Modern designs therefore treat corrosion control as a first-class design variable tied to soil aggressiveness and consequence of failure (Rajani et al., 2011). AWWA C105's 10-point soil evaluation (resistivity, pH, redox, sulfides, moisture) is widely used to screen for aggressive conditions and trigger protections such as polyethylene encasement and/or Cathodic Protection (CP) (AWWA C105). Field experience shows that performance depends on doing the fundamentals well like making sure coatings are intact, encasements undamaged, and CP designed, installed, and monitored correctly (NRC, 2009).

The DIPRA–Corrpro Design Decision Model (DDM) statistically analyzed decades of iron-pipe corrosion data, evaluated the AWWA 10-point scale as a predictor of aggressive soils, and formalized risk-based selection of protections. Updates emphasize chloride effects, groundwater intrusion into the pipe zone, and in the most aggressive soils for

distribution mains, zinc coatings with enhanced polyethylene sleeves (Corrpro & DIPRA, 2005; program updates summarized in NRC, 2009). Debates remain about polyethylene's reliability under damage or fluctuating water tables and recommend careful construction with QA/QC (NRC, 2009).

A U.S. Bureau of Reclamation review concluded that PE plus CP can outperform unprotected iron in highly corrosive soils when designed and installed correctly, but cautioned that limited, heterogeneous data and inconsistent pitting statistics complicate long-life claims. Using gas-pipeline benchmarks with bonded dielectrics and CP, USBR argued that a reliable 50-year life in soils <2,000 Ω-cm is not assured for DI with PE+CP and called for better monitoring (e.g., inline inspection where feasible) and more systematic studies across diameters, pressure classes, coatings, and soils (USBR, 2004; NRC, 2009).

Cast Iron (CI): Early pit-cast CI often had thick walls and, despite manufacturing variability, can persist in benign soils. Later spun-cast CI is thinner. Pre-1955 lead/"leadite" joints typically used for these pipes are a distinct cohort with known issues (thermal mismatch, microbiologically influenced cracking of bells, leakage). CI failure modes include circumferential/longitudinal/spiral cracking from bending and pressure, bell shearing/splitting, and corrosion defects (pitting, graphitization, tuberculation). Soil

pH < 4 or > 8.5, low/negative redox, sulfides, and resistivity below ~2,000 Ω-cm are strongly associated with external corrosion (Singley et al., 1984; AWWA C105; Rajani et al., 2011; Makar et al., 2001).

Ductile iron (DI): Introduced mid-1950s, DI pairs higher strength/ductility with cement-mortar linings and standardized joints (C110, C111, C115). Because DI is less prone to stress-cracking splits than CI, observed failures skew more toward third-party damage, joint defects, and corrosion where protections are absent or compromised. Thickness design follows flexible-pipe principles (C150/C151); service allowance is explicitly separated from casting tolerance, and pressure-class options enabled thinner walls in some diameters, again raising the importance of explicit corrosion control (AWWA C150/C151; Rajani et al., 2011).

Steel: Steel dominates many large-diameter, high-consequence mains. Performance depends heavily on bonded dielectric coatings, linings, CP, weld quality, and joint type. External pitting corrosion is the most common failure mechanism in North American surveys, though rates vary by era, coating, CP practice, and soil (Folkman, 2012; WRF, 2016). Cross-industry evidence (oil/gas) reinforces the roles of disbonded tapes, near-neutral pH stress corrosion cracking, carbonate/bicarbonate chemistry, hydrogen effects, and

stray current, mechanisms relevant when protections are inadequate or degraded (Bolzoni et al., 2000; Chen et al., 2002; 2006; Velázquez et al., 2009; Meresht et al., 2011). Reported "failure events per 100 mi-yr" differ widely across countries and survey methods; for water, higher rates are often tied to older permeable coatings without CP, underscoring that "performance" (condition, leaks avoided) may be a better target metric than crude failure counts for critical steel mains (Gould et al., 2013; Mackellar & Pearson, 2003; USBR, 2009; WRF, 2016).

For large diameter steel pipes, joints and fittings are frequent places where failure initiates. Misaligned or degraded gaskets, corroded bolts/rings, and material transitions can leak, undermine bedding, and precipitate structural failure, especially under transients (valve/hydrant operations). For steel, welded joints tend to perform well with proper QA/QC and coating/lining reinstatement; rubber-ring joints can leak with settlement or pinched gaskets; legacy lead joints are fragile and uncommon today (WRF, 2016; AWWA jointing standards).

Implications for modeling: Soil resistivity/pH/redox/sulfides/moisture; coating/encasement/CP presence, type and quality; joint type/era; surge exposure; vintage cohorts are the main factors used to directly connect to likely deterioration and failure

modes across CI, DI, and steel. Validation should prioritize strata where protections differ (e.g., DI with PE only vs PE with CP; steel with bonded dielectric with CP vs legacy bituminous) and where joint cohorts change.

2.7.2.2 Plastic Pipe Materials' Performance Characteristics

Plastic pipelines (PVC/uPVC, oriented PVC-O, and PE/HDPE) are now common in distribution because they resist corrosion, are light and easy to install, and offer smooth interiors. Their long-term performance, however, is highly sensitive to construction practice (embedment, compaction, point loading), jointing quality, surge/air management, and exposure to disinfectants and organics. This section summarizes material evolution and standards, typical failure modes, and the implications for design, inspection, and renewal planning (Uni-Bell, 2013; AWWA M55; Davis et al., 2007).

PVC (uPVC and PVC-O): PVC entered water service in the 1950s and expanded rapidly after the late-1970s as utilities sought corrosion-resistant, lower-cost alternatives to iron. The 2007 revision of AWWA C900 incorporated WaterRF research, reduced the factor of safety (2.5→2.0), and separated surge allowance from pressure class. Further, in 2016, C900 replaced C905 and extended sizes from 4–60 in. PVC-O (C909) raises design stress, enabling thinner walls (Burn, 2006; Uni-Bell, 2013). In practice,

performance depends on appropriate SDR/pressure class selection for steady and transient pressures, careful bedding to avoid rock impingement, and correct joint assembly.

Typical modes include (i) long-term yield failures at highly stressed fittings (tees, elbows) where stresses exceed long-term capacity; (ii) brittle "blown section" fractures where a crack bifurcates and rejoins; (iii) environmental stress cracking (ESC) associated with localized solvation by certain organics, seen as smooth, glassy surfaces with craze development before catastrophic failure; (iv) impact-initiated cracks from handling or third-party strikes; (v) longitudinal splits exacerbated by trapped air and dynamic loading; and (vi) joint leakage from misalignment, over-deflection, damaged gaskets, or pullout (Knight, 2004; Uni-Bell, 2013). Many of these are installation-centric, so QA/QC at construction and transient control in operations are decisive.

PE/HDPE (including PE4710): Modern PE (PE4710) combines high ductility, toughness, and fused joints that create leak-free, fully restrained systems when procedures are followed. Key references include AWWA C901 ($\leq$3 in), C906 (4–65 in), and M55, with fusion governed by ASTM F2620/F3124 (butt) and F1055/F1290 (electrofusion), plus MAB-01/02 procedures. PE's flexibility suits narrow trenching and trenchless methods

(HDD, sliplining), but design must address temperature-driven expansion/contraction, restraint, and soil-pipe interaction (AWWA M55; PPI, 2008).

HDPE shows three broad leakage/burst regimes in legacy resins: ductile overload, non-ductile slit/pinhole cracking, and later-stage failures influenced by oxidative degradation; newer PE4710 resins were developed to raise resistance to slow crack growth and rapid crack propagation (Davis et al., 2007). Oxidative degradation can progress from antioxidant depletion at the inner wall (under certain disinfectant exposures) to molecular weight loss and finally non-ductile failure if design/operation is unsuitable; disinfectant management and material selection matter (PPI TN-44). Thermal movement is larger than for metals but can be beneficial (stress relief) when restrained and detailed correctly (Najafi, 2015). Permeation by certain organics is possible. AWWA standards include permeation provisions, and risk depends strongly on local exposure scenarios (Ong et al., 2008; PPI position papers).

For PVC, gasketed push-on joints dominate. Joint tightness depends on alignment, deflection limits, and gasket condition (Uni-Bell, 2013). For PE, butt-fusion and electrofusion can deliver joints as strong as the pipe when procedures, tooling, cleanliness, and parameter recording (ASTM F3124) are followed; poor fusion practice is a leading root

cause where issues arise (AWWA M55; MAB-01/02). Across both materials, valves, hydrants, and tapping practices can introduce local stress raisers; transient control (air management, surge) reduces crack initiation and joint distress.

Mechanism-grounded features that matter for LOF include embedment quality and backfill angularity; cover and deflection; SDR/pressure class vs steady/surge pressures; air management and transient history; resin generation (e.g., PE4710 vs earlier), disinfectant exposure, temperature; fusion QA logs (parameters, coupons, F3124 records); joint type/deflection; nearby utility construction (impact risk); and contaminant exposure (permeation potential).

2.7.2.3 Concrete Pipe Materials' Performance Characteristics

Concrete pressure pipeline materials PCCP (AWWA C301), RCCP (AWWA C300/C302), and bar-wrapped concrete cylinder pipe (AWWA C303) are the backbone of many large-diameter transmission systems. Because they carry high volumes, their consequence of failure is inherently high, so understanding how materials, manufacturing eras, and installation/operation interact is essential for risk and renewal planning.

PCCP combines a thin steel cylinder, concrete core(s), high-strength prestressing wire, and a mortar coating. The two main types of PCCP are Lined-Cylinder (LC) and Embedded-Cylinder (EC), which differ in where the steel cylinder sits relative to the concrete core (Romer et al., 2007; Ge & Sinha, 2014). RCCP comes in cylinder (C300) and non-cylinder (C302) variants; C302 relies on the concrete wall to contain pressure. C303 (bar-wrapped) uses a steel cylinder with spiral steel bar and mortar coating and above ~36 in behaves more like a semi-flexible pipe than a rigid one (AWWA C303).

Manufacturing & standards eras in this section help create natural "risk cohorts." Three PCCP eras are especially useful for modeling LOF: (i) pre-1964 (conservative design), (ii) 1964–1984 (reduced conservatism: thinner coatings, higher core stress, smaller wire sizes), and (iii) post-1984/1992 (standards tightened: denser mortar, thicker coatings, slurry under wire, detailed design moved to AWWA C304) (AwwaRF, 2008; Romer et al., 2007). Two cross-cutting issues amplify era effects:

- Mortar coating porosity (≈1970–late-1980s): drier mixes increased internal friction and left interconnected voids, reducing protection against chlorides/sulfates and cyclic wetting/drying; moisture limits and absorption testing were later tightened (C301-07/14).

- Prestressing wire susceptibility (≈1970–1988): early ASTM A648 allowed very high tensile strengths. Dynamically strain-aged wire was more prone to splitting/corrosion and hydrogen embrittlement, especially if cathodic protection potentials exceeded ~1 V or if corrosion pits formed. Subsequent torsion/relaxation and hydrogen-resistance tests (ASTM A1032-04) improved outcomes (Romer et al., 2007). Other flags include brief allowances for #18-gauge cylinders and cold-rolled steel on small diameters (thin, low ductility, weld-sensitive) and tight wire spacing near ends (harder to encapsulate with mortar) (Woodcock, 2008).

Failures typically initiate chemically at the exterior where coating cracks/delamination or high-porosity mortar let groundwater reach the wire. Next, carbonation lowers pH at the wire level. Subsequently, chlorides depassivate steel, especially with cyclic wet/dry concentrating salts leading to wire corrosion and breaks. Enough wire loss raises hoop stresses on the cylinder and with pressure transients or other sudden stresses, cylinder burst can occur (Romer et al., 2007; Ge & Sinha, 2014). Hydrogen embrittlement accelerates wire breaks in susceptible vintages (1970–1988) (Romer et al., 2007). Internally, localized steel exposure can also tuberculate. However, external pathways typically dominate PCCP. For RCCP and C303, structural cracking patterns mirror bedding

support and loading conditions and typically exhibit circumferential cracks from differential bedding/haunch support and longitudinal cracks from overburden/traffic or point loading (C300/C302/C303 standards; ACPPA).

High-value predictors of distress include construction damage to coatings, missing joint coating/diaper, poor bedding (rocks/point loads), poor grouting at bells/spigots, settlement, shifted/looped gaskets, and surge events from pump/valve operations that drive crack initiation at crown/spring line (Ge & Sinha, 2014; Kola, 2010). Pressurization/depressurization cycles can trigger wire snaps in embrittled vintages (Stroebele et al., 2010). Misapplied cathodic protection (>~1 V) can evolve hydrogen at steel, worsening embrittlement, particularly in older high-strength wire classes (Romer et al., 2007).

Chlorides, even in trace amounts at pH ~9–10, can initiate corrosion at the wire if they reach it. CO_2 promotes carbonation of mortar, lowering pH; precipitation-driven groundwater movement can accelerate both ion transport and bedding undermining. Vegetation is rarely a direct cause but can contribute organic acids/CO_2 locally (Price et al., 1998; Woodcock, 2008).

Joint leaks (misfit, out-of-round, cracked core at joint, missing coating), alignment changes, and diaper failures are common distress indicators. For C303 and C302/C300,

inadequate joint protection allows corrosion to local steel, while impact and point loads at joints concentrate stresses (ACPPA; Kola, 2010).

LOF modeling requires prioritization of features that encode era and make (pre-1964 / 1964–84 / post-1984, LC vs EC, C300 vs C302 vs C303), wire class (I/II vs III/IV), coating quality proxies (moisture/absorption specs, slurry-under-wire adoption), cylinder gauge and steel type, bedding/settlement risk, cathodic protection history, surge/transient history, joint condition (coating/diaper/grout), and environmental exposure (chlorides, CO_2, wet/dry cycling). For COF, diameter and hydraulic criticality need to be coupled with outage/customer-hours and repair logistics. Operationally, focus is required on surge control, joint protection, bedding remediation, and cathodic protection setpoints. Renewal options vary by type and failure mode (e.g., clamp repairs for cylinder leaks, liners, carbon fiber wraps, or replacement with cylinder-type concrete pipe) but success hinges on stable support and durable external protection (ACPPA).

2.7.2.4 Other Pipe Materials' Performance Characteristics

Beyond metallic, plastic, and cementitious classes, several "other" materials appear in legacy systems or niche applications. Their failure mechanisms are dominated by pipe and soil chemistry in addition to the trench support and human handling, not just

pressure. That makes water quality, soil exposure, and workmanship central to modeling failure.

Asbestos Cement (AC): Asbestos-cement is Portland cement with silica and asbestos fibers. Utilities adopted it for its smooth interior, low headloss, and immunity to galvanic corrosion, but its low bending strength makes it vulnerable to impact, handling errors, and uneven support. Early Type I (moist-cured) products gave way in the mid-1930s to steam-autoclaved Type II with lower free lime and better resistance to acid and sulfate attacks. Correspondingly, U.S. specifications progressed from SS-P-531 (1940) to AWWA C400 (1953) pressure classes. The dominant deterioration mechanisms for AC pipes are lime leaching in soft or low-alkalinity waters, which removes calcium hydroxide and weakens the matrix, and sulfate attack in water or soils, which forms expansive minerals that swell and crack the cement phase. External cracking is common where expansive clays, settlement, or point loads act on small diameters. Lifecycle risk in AC correlates with water aggressiveness and alkalinity, sulfate or chloride exposure, age, diameter, bedding uniformity, and any history of impact or third-party activity. Design and construction quality matter where aspects like joint fit, proper insertion, bedding, and backfill reduce risk, while storage and transit damage can seed future failures. Joints are slip

couplings with rubber gaskets that tend to fail through displaced gaskets, socket cracking, seal aging, or material degradation when misalignment or over-deflection is present. For modeling, include water chemistry, soil chemistry, settlement or expansive soil flags, vintage (Type I vs II), diameter, impact events, and joint condition. Practical actions range from water conditioning and support remediation to clamp repairs and targeted renewals in high-sulfate corridors, always treating AC handling as a regulated O&M and safety issue.

Glass Reinforced Plastic (GRP)/ Fiber Reinforced Plastic (FRP), including Reinforced Plastic Mortar (RPM), Reinforced Thermosetting Resin Pipeline (RTRP) and Fiberglass Reinforced Epoxy (FRE): GRP is a fiber-reinforced thermoset composite in which glass fibers provide tensile strength and the resin matrix provides shape and compressive strength. RPM adds sand for stiffness in larger diameters. It is attractive for corrosion resistance, trenchless friendliness, and low weight, yet performance hinges on the fiber–matrix system, manufacturing quality, and control of deflection. AWWA C950 governs pressure pipe, with ASTM families such as D3517 and D3262 for pressure and sewer applications, D2992 for long-term pressure ratings, D2412 for stiffness, and D4161 for joints. Municipal use widened from the 1960s onward. Damage

occurs in the matrix (blisters or softening from mechanical, chemical, or thermal loads), in the fibers (fracture under bending or point loads), or at the fiber–matrix interface (debonding), often presenting as creep or creep-rupture when sustained load or deflection exceeds limits. Construction and tapping can introduce local tears or ruptures, particularly where rock contacts or poor tapping practice exist. Likelihood of failure rises with lower-grade resin systems, insufficient stiffness class relative to burial and traffic, excessive deflection from poor bedding and haunch support, point loads, aggressive chemistry, and temperature swings. Also, elastomeric-seal joints can add sensitivity to misfit and over-deflection. Modeling should reflect stiffness and diameter, cover and traffic, measured or estimated deflection, trench geology, trenchless records, temperature range, and chemical exposure. Remedies emphasize enforcing deflection limits, improving bedding and side support, using proper tapping procedures, and confirming chemical compatibility.

Wood: Historic wood staves or bored logs persist in older districts. When buried and kept water-filled, they can last for decades because oxygen is limited, but unknown laterals and easy tampering complicate management. Risk assessment relies on age and records, proximity to structures that alter moisture and oxygen, and redevelopment activity. The common operational posture is to identify, isolate, and replace wood segments

when encountered while acknowledging inventory uncertainty in service-disruption planning.

Lead: Lead persists mainly in service lines and premise plumbing. The primary concern is public-health risk from lead release rather than structural failure. Risk depends on inventory certainty, water chemistry and corrosion control, disturbance history, and the presence of partial replacements. Management belongs to a compliance and public-health program with corrosion control and full lead-service-line replacement, while capital planning focuses on replacement logistics and equity rather than likelihood of failure alone.

Copper: Copper became the norm in services and premise plumbing because it forms protective films and tolerates temperature variation. Types K, L, and M refer to wall thickness, with K the thickest. It can still pit internally or erode-corrode at bends and high velocities, and it can corrode externally in aggressive soils. Risk increases with high velocities and sediment loads, tight bends, temperature cycling, localized chemistries with unfavorable chloride-to-sulfate ratios, and corrosive soils. Mitigation uses velocity control, smoother hydraulics, and targeted replacements where pitting recurs. A summary of all the characteristics discussed in this section is presented in Table 2-4.

Table 2-4: Summary of typical characteristics of various transmission, distribution and service pipe materials currently in use in the US

Material	Typical dia. / uses	Key Strengths	Typical failure root causes	Typical failure modes	Standards snapshot	Renewal/diagnostic cues
Cast Iron (CI)	Legacy distribution/transmission; 4–48 in typical	High stiffness; long legacy track record	Internal/external corrosion; graphitization; bell-and-spigot joint leakage; surge cracking	Circumferential breaks, joint leaks, corrosion pinholes, longitudinal splits (frozen soil/surge)	AWWA C106 (historic), joints per C111 (A21.11)	High break history; tuberculation; low residuals/pressure; poor soil corrosivity indices
Ductile Iron (DI)	Distribution/transmission; 4–64 in typical	Tough; ductile; pressure capable; many joint options	External corrosion in aggressive soils; coating/lining holidays; stray current	Bell/socket cracks, corrosion holes, longitudinal splits under surge or poor bedding	AWWA C151 pipe, C104 cement-mortar lining, C111 joints	Soil corrosivity + CIS/PCM; coating condition; break clustering; wall-loss from coupons/UT
Steel (ML&C)	Large transmission; >24 in common	High pressure; custom fabrications; light for size	Coating/lining damage; weld defects; external corrosion; ground movement	Seam/circumferential weld cracks; corrosion leaks; buckling at low cover; surge fatigue	AWWA C200 steel pipe; C205 mortar lining; C207 flanges; C206 field welding	Coating holidays/CP data; leak history; hydrotest/surge history; weld NDE results
PVC (uPVC / C900; includes PVC-O C909)	Distribution/transmission; 4–60 in (C900)	Corrosion-resistant; light; smooth bore; easy handling	Impact damage; point loading (rocks); ESC with certain organics; UV with long exposure	Brittle fractures, longitudinal splits, joint leaks, environmental stress cracking	AWWA C900 (replaced C905 in 2016); C909 (PVC-O) higher stress rating	Install era/cohorts (pre-2007 vs post); bedding quality; surge history; solvent/organics exposure
HDPE (PE4710)	Services to large mains; ¾–65 in (C906)	High ductility; fused leak-free joints; trenchless-friendly; impact tolerant	Thermal expansion/contraction; oxidative disinfectant attack; UV if low carbon black	Ductile overload, pinholes/slits, oxidative surface embrittlement, fusion/joint defects	AWWA C906; AWWA M55; PPI TR-4/TR-46; ASTM F2620 (fusion)	Disinfectant strength/contact time; fusion QA records; temperature swings; restraint at appurtenances
PCCP (C301 LC/ECP)	Large transmission; 24–252 in (ECP), 16–60 in (LCP)	High strength/rigidity; thin steel cylinder; long spans	Wire corrosion/embrittlement (1970–1988 cohorts); porous mortar eras; hydrogen from over-CP	Wire breaks → loss of prestress; coating spall; cylinder burst; joint leaks	AWWA C301; design C304; key eras: pre-1964 / 1964–84 / post-1984	Wire-break counts (acoustic/EM); mortar absorption; CP levels (<1 V); bedding; chloride/CO_2 soils
RCCP (C300)	Large transmission; 20–144 in+	Rigid; handles high external loads; flexible joint options	Improper bedding/haunch; joint issues; external loads	Circumferential cracking (bending), longitudinal cracking (overload), wall degradation	AWWA C300 (since 1952; updates '57, '64, '74, '89, '93)	Trench stability; haunch support; traffic loads; joint integrity; freeze-thaw exposure

Bar-Wrapped Concrete Cylinder (BWP, C303)	10–72 in; semi-flexible behavior >36 in	Corrosion-resistant mortar lining/coating; lighter than PCCP; weldable cylinder	Mortar impact cracking; point loads; bedding deficiencies; joint protection	Localized bar corrosion under coating damage; circumferential/longitudinal cracks; joint distress	AWWA C303 (1970→2017); moisture/absorption tightened in 1995/2017	Coating condition; bedding quality; sidefill; joint wraps; cylinder thickness (#16 vs #18 gauge)
Asbestos Cement (AC)	Distribution; common mid-20th c.	Corrosion-resistant interior; smooth bore; decent rigidity	Lime leaching in soft/low-alkalinity water; sulfate attack; low bending strength; impact	Internal deterioration; external cracks/holes from soil movement; joint gasket/seal issues	US SS-P-531 (1940); AWWA C400-53T; Type I→II (steam-autoclaved) transition mid-1930s	Aggressiveness index; sulfate/clay soils; age; coupling/gasket condition; handling damage history
GRP / FRP (AWWA C950; ASTM D3517 etc.)	Pressure/gravity; often >12 in with sand core (RPM)	Corrosion-resistant; light; fast install; no CP required	Matrix/fiber/interface damage; chloride/humidity attack on glass if not encapsulated; creep	Blistering; cracks/disbonding; rupture at taps/impact; long-term creep rupture	AWWA C950; ASTM D3517/D3262/D3754; extensive resin/fiber test methods	Resin system; laminate QA; soil/load changes; hot fluids; tapping records
Wood	Historic distribution; 1800s–early 1900s	Low cost (historic); workable	Mechanical damage; tapping theft (historic); limited life if exposed to air	Leakage at joints; rot if air+water present; crushing	Historic municipal specs; largely abandoned in place	If encountered during work: abandonment, localized removals; heritage notes
Lead (service lines/premise)	Services; premise plumbing	Malleable; durable formability	Public health risk (leaching); galvanic coupling; scale disturbance	Leaks at wiped joints; kinks; cracks; corrosion pinholes	EPA LCR governs abatement; most installs pre-1940 (some to 1980s)	LCR exceedance; inventory confirmation; full LSL replacement programs
Copper (Types K/L/M)	Premise/service; K (thick), L (med), M (thin)	Corrosion-resistant; temperature tolerant; durable	Erosion/corrosion at bends with sediment; external soil corrosion; pitting	Circumferential cracks at bends; pinhole leaks; joint failures	ASTM B88, B42, etc.	Water chemistry (pH/alkalinity/chloride); velocity; bend/fixture history

2.7.3 Asset Management and Risk Analysis Frameworks

The coexistence of epistemic and aleatory uncertainties makes accurate pipeline renewal prioritization difficult. This research addresses these challenges by reducing epistemic uncertainties through a transparent, knowledge-driven modeling approach that systematically captures pipeline deterioration and failure impact mechanisms. A white-box

knowledge base, built through an extensive literature review and industry best practices, will enhance model interpretability and accuracy. Additionally, this research will improve aleatory uncertainty quantification by incorporating confidence intervals into risk predictions, ensuring probabilistic estimates that reflect inherent variability. These advancements will be driven by comprehensive utility data collection, heuristic-integrated mathematical modeling, and rigorous scientific validation, ensuring model robustness and real-world applicability.

The integration of asset management in water utilities began with the Government Accounting Standards Bureau (GASB) Statement 34 in 1999 (Kim et al. 2018), prompting utilities and public agencies to formalize asset management practices (Giglio 2018). Since then, various definitions and standards, such as PAS 55 (Argent 2007) and ISO 55000 (Hodkiewicz 2015), have emerged. Several frameworks have been applied to pipeline infrastructure, including Failure Modes and Effects Analysis (FMEA) (Wang et al. 2023), Reliability Centered Maintenance (RCM) (Geisbush and Ariaratnam 2023), Life Cycle Cost Analysis (LCCA) (Thomas et al. 2016), Life Cycle Analysis (LCA) (Du et al. 2013), and Risk Management (RM) (Muhlbauer 2004).

While these frameworks support pipe renewal decisions, they often lack comprehensiveness. FMEA and RCM focus primarily on LOF but overlook socio-economic consequences. LCCA and LCA assess financial and environmental costs but do not explicitly address failure risk and prioritization. In contrast, Risk Management (RM) offers a more holistic approach by integrating LOF-based frameworks (like FMEA and RCM) with renewal prioritization (as in LCCA) while also accounting for COF, including social, environmental, and economic impacts. Its modular nature allows integration with optimization algorithms, ensuring robust decision-making for pipe renewal prioritization.

Risk analysis has been widely adopted across sectors since its introduction in public health in the early 20th century (Wolman 1921), influencing agencies like NASA, U.S. Army Laboratories, and the FDA. The water sector formally adopted proactive risk management in the 1990s, treating drinking water as a regulated product in nations like Iceland (Setty et al. 2019). Following guidelines from WHO and IWA (IWA 2004; WHO 2004), risk management frameworks for drinking water have now been implemented in over 90 countries (WHO 2017), with U.S.-specific standards like J100 Risk Analysis and Management for Critical Asset Protection (RAMCAP) (AWWA 2010) developed for water supply systems.

Despite these advancements, pipeline failures and renewal costs continue to rise, as evidenced by the 2021 USEPA survey, which estimates that $472.6 billion is needed over 20 years to maintain and improve drinking water infrastructure (USEPA 2021). This highlights the limitations of current frameworks and underscores the need for improved decision-support models.

This research is particularly timely for four key reasons. First, advancements in sensor technology now support real-time condition assessment, making predictive modeling more reliable. Second, state-of-the-art AI-based modeling techniques enable the development of more representative and adaptable models, enhancing decision-making. Third, escalating failure rates across the U.S. signal an impending sustained rise in break rates and renewal needs which threatens to overwhelm utilities and demand emergency interventions. Fourth, the increasing emphasis on fairness, trust, and explainability in AI demonstrates the importance of transparent, unbiased, and scientifically valid AI-driven decision-support models for water infrastructure management. Given these urgent challenges, it is critical for utilities to actively engage in the development of scalable, data-driven tools to enhance pipeline renewal prioritization and long-term infrastructure sustainability.

2.7.4 Key Findings from Literature Review

The 5 domains in the scope of the review: (1) structural and functional performance of water pipelines, (2) dependencies and interdependencies influencing COF modeling, (3) decision science for renewal prioritization, (4) model verification and validation techniques, and (5) adjacent domains highlight the gaps in current methodologies and inform the development of improved risk-based renewal models. These are shown in Table 2-5.

Table 2-5: Selected literature classified based on modeling family and evaluated based on coverage of 3 key research themes (LOF, COF and decision constraints)

Family	Purpose	Typical methods/ examples	Captures (LOF / COF / Constraints) *	Strengths	Common limitations
Descriptive / heuristics & expert rules	Quick screening; data-light triage	Break-rate indicators, age rules, material/diameter lookups, expert scoring	LOF: ✓ (coarse); COF: ◐ (proxy); Constraints: ✗	Simple, transparent, low data needs	Unstable across cohorts; weak transferability; ignores implementability
Statistical reliability / deterioration	Estimate failure propensity/ time	Survival/hazard (Cox, Weibull); hierarchical Bayesian; spatio-temporal models	LOF: ✓; COF: ◐ (if linked); Constraints: ✗	Statistically grounded; uncertainty quantifiable	Often excludes COF and delivery constraints; external validation uneven
Mechanistic / physics-based	Model deterioration/ strength explicitly	Corrosion/soil aggressivity models; remaining strength; transient loading/fatigue	LOF: ✓ (mechanism-based); COF: ✗; Constraints: ✗	Physical interpretability; extrapolation to new regimes	Data/calibration intensive; partial coverage of materials; coupling to network context hard
Machine learning / AI (incl. physics-informed hybrids)	Learn nonlinear patterns; improve predictive power	Tree ensembles, boosting, random forests; neural nets; hybrid PINNs	LOF: ✓; COF: ◐ (if features exist); Constraints: ✗	Captures interactions, nonlinearity; scalable	Adoption hinges on explainability; dataset bias/leakage risks; validation-in-use rare
Decision analytics & prescriptive optimization	Turn risk and value into executable plans	MCDA/benefit–cost; integer & multi-objective portfolio; clustering & scheduling; stochastic/robust variants under budgets, moratoria, permits, crews	LOF: via inputs; COF: via valuation; Constraints: ✓	Aligns with budgets & governance; produces buildable projects; supports trade-offs	Quality depends on upstream inputs/weights; can be opaque without explanations/ sensitivity

✓ Complete coverage, ◐ Partial coverage, ✗ Not covered

2.7.5 Structural and Functional Performance of Water Pipelines

LOF studies cluster into (i) age-only or age-plus-material proxies and (ii) mechanism-aware models that resolve corrosion/deterioration, loading/stress, and hydraulics/capacity. The former are simple but transfer poorly across cohorts; the latter require richer covariates (soils, bedding, traffic class, pressures/transients, network redundancy) yet generalize better when validated externally. Reported metrics should include discrimination (e.g., AUC/PR-AUC), calibration (reliability curves/Brier), and temporal validation; very few papers provide all three.

Existing pipeline performance models utilize diverse methodologies, each with strengths and limitations. Deterministic models, including empirical and mechanistic approaches, rely on well-defined relationships between asset attributes and failure rates but require careful classification of pipeline groups (Kleiner and Rajani 2001; St. Clair and Sinha 2012). Probabilistic models apply statistical techniques to historical data (Ge and Sinha 2014; Mazumder Ram et al. 2018), though their utility diminishes with sparse datasets.

Advancements in AI techniques, particularly Artificial Neural Networks (ANNs) and Fuzzy Logic, have improved failure modeling by capturing nonlinear relationships

and incorporating expert reasoning (Senouci et al. 2014; St. Clair 2013). Fuzzy logic models offer a rule-based, interpretable approach, making them suitable when data is uncertain (Angkasuwansiri 2013; Uslu 2017). While fuzzy models rely on expert-driven rules rather than data-driven learning, they can serve as an effective precursor to ML models by structuring reliable input-output mappings. For instance, graphitic corrosion in cast iron pipes, influenced by soil resistivity and chloride content, can be represented through fuzzy if-then rules, enabling accurate predictions even with limited longitudinal data.

2.7.6 Field Data Collection Techniques

This section presents a summary of field data collection techniques critical to collect reliable ground truth data for model training, validation and improvement

2.7.6.1 General Methods (For all pipe materials)

Accurate condition assessment of water pipelines requires a combination of above-ground inspections, excavation-based evaluations, and specialized non-invasive techniques tailored to specific pipe materials.

Above Ground Inspection: Above-ground inspections involve visually assessing pipeline alignments for external signs of damage. For buried pipes, indicators such as

pooling water, collapsed pavement, or stray current sources can suggest pipeline deterioration. For above-ground pipes, inspectors look for cracks, corrosion, misalignments, and leaks. This method is simple, cost-effective, and requires minimal training, but its ability to assess buried pipelines is limited.

Excavation for Visual Inspection: Direct excavation allows inspectors to visually examine external pipe walls for signs of corrosion, cracks, or material degradation (Figure 2-10.). This method provides more precise condition data and is often performed alongside other assessment technologies. However, excavation is costly, offers no insight into internal conditions, and may pose safety risks for fragile, aging pipelines.

Figure 2-10: Example of Direct Assessment (Courtesy WSSC, 2025)

2.7.6.2 Corrosion and Wall Thickness Data in Metallic Pipes

Magnetic Flux Leakage (MFL) is widely used for detecting corrosion pits, wall thinning, and cracks in cast iron, ductile iron, and steel pipes. By magnetizing the pipe wall (see Figure 2-11 left side), the method identifies disruptions in magnetic flux due to defects (see Figure 2-11 right side). While MFL is effective for small and unlined metallic pipes, it struggles to detect short or shallow defects, leading to some uncertainty in the results (Rizzo, 2010).

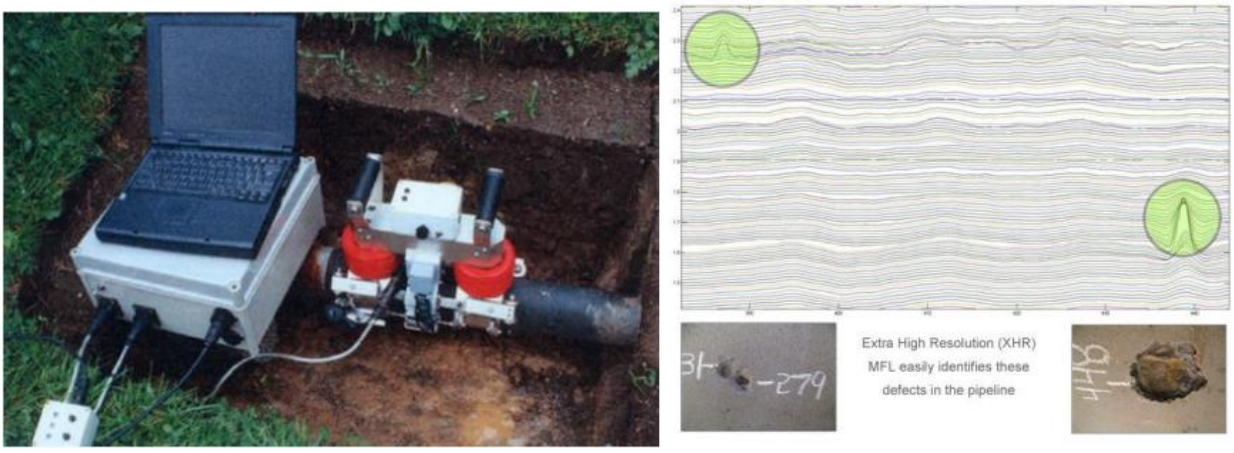

Figure 2-11: Left shows a Full Circumferential Pipe Wall Inspection Tool. Right shows Contour Map Showing the Flux Density from Hole Defect. (Courtesy of PURE Technologies)

Ultrasonic Testing Methods: Ultrasonic testing uses mechanical stress waves to measure pipe wall thickness at specific points (Rizzo, 2010). This technique is highly effective for steel and ductile iron pipes but is less reliable for cast iron due to material

inhomogeneity. While useful for detecting remaining wall thickness, the method requires clean test surfaces (Costello et al. 2007) and is slow due to its point-by-point measurement process. A handheld ultrasonic probe is shown in Figure 2-12.

Figure 2-12: Handheld Ultrasonic Testing Tool. (WRF, 2008)

2.7.6.3 Wall Shape, Loss and Ovality Data in Concrete Pipes and Plastic Pipes

Structured Laser Profiling: Laser profiling generates a continuous line of light along the internal circumference of a pipeline to assess shape, ovality, and vertical deflection (Costello et al. 2007) as shown in Figure 2-13. This method is particularly useful for evaluating structural integrity in concrete and plastic pipes. However, pipelines must be cleaned and dewatered before inspection, and the method does not detect cracks reliably.

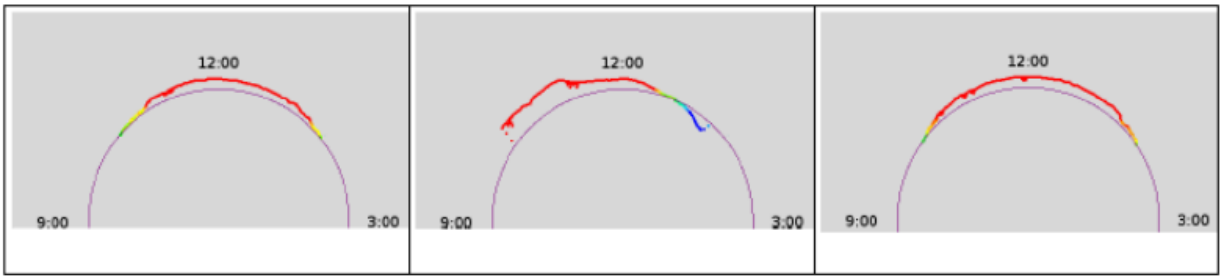

Figure 2-13: Example of Three Laser Data Readings in concrete pipe. (Courtesy of RedZone Robotics)

Assessing the condition of plastic pipes (HDPE, PVC) requires a combination of external and indirect evaluation techniques, as well as forensic analysis of failed sections. While non-invasive internal assessment methods for plastic pipes are still evolving, several complementary approaches can be employed to gather meaningful condition data. Techniques such as laser profiling can be used to assess ovality and deformation. Additionally, forensic examinations of extracted pipe sections provide valuable insights into mechanical damage, including buckling, stress cracking, and ovality. By integrating these techniques, the required information can be gathered to test our model results.

Remote Field Technologies for PCCP Wire Break Detection: Remote field technologies (RFT), including remote field eddy current (RFEC) (illustrated in Figure 2-14) and broadband electromagnetic (BEM) methods, are widely used to detect

prestressing wire breaks in prestressed concrete cylinder pipe (PCCP). These methods overcome the depth penetration limitations of traditional eddy current testing.

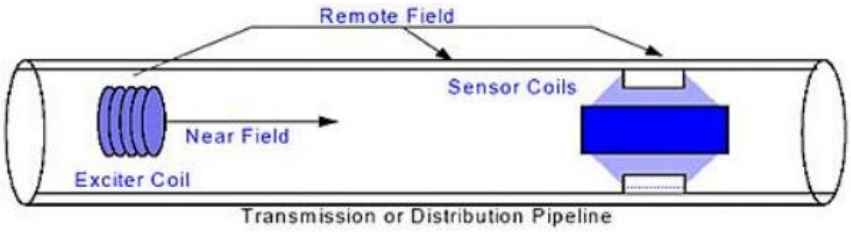

Figure 2-14: The Remote Field Effect. (USDOE)

RFEC deploys a probe with multiple magnetic coils through the pipeline, where an exciter coil generates eddy currents that travel through the pipe wall. A detector coil, placed two to three pipe diameters away, captures variations in the electromagnetic field to identify anomalies and wire breaks. BEM, a frequency-independent variant (application shown in Figure 2-15), adapts to different materials and environmental conditions while minimizing electromagnetic noise interference.

Figure 2-15: BEM Hand-Held Tool Being Used to Scan a Gray Cast Iron Pipe. (WRF, 2008)

RFEC provides electromagnetic signals indicating wire breaks and pipeline anomalies (Figure 2-16 left side), while BEM generates surface contour maps to assess cracks, wall thickness, and material degradation (Figure 2-16 right side).

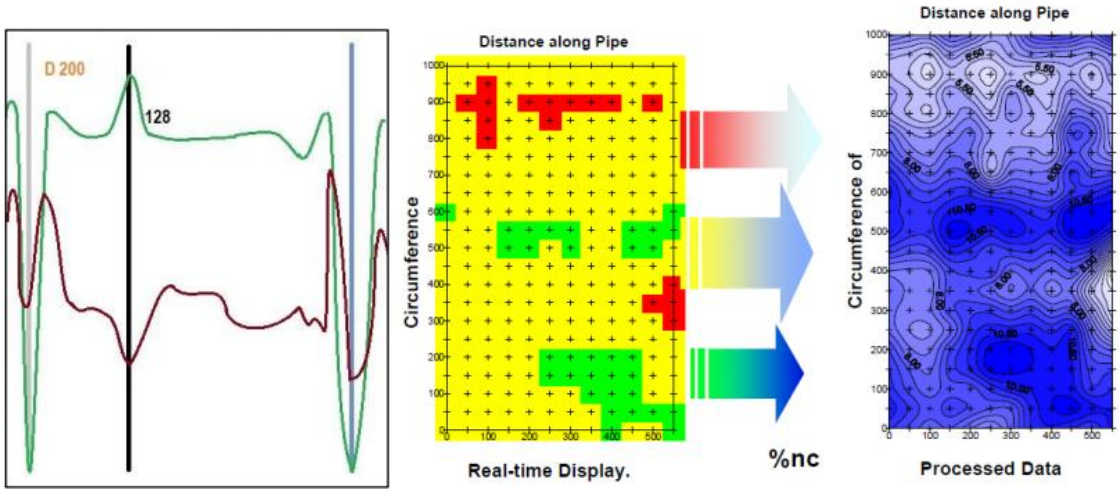

Figure 2-16: Left shows electromagnetic signal obtained from internal inspection of PCCP Using a Robotic RFT Tool. (Sinha 2021). Right shows typical BEM Data (WRF, 2008)

These technologies function in various flow conditions and pipe diameters. RFEC tools are available in manned and unmanned configurations for both pressurized and

dewatered pipelines, while BEM is particularly effective for penetrating thick coatings and detecting wire breaks in PCCP. RFEC is sensitive to environmental interferences, including motion, pipeline joints, and adjacent steel structures. It is unsuitable for non-metallic pipes and requires careful calibration for PCCP assessment to ensure accurate results.

In summary, each condition assessment method has distinct advantages and limitations. While above-ground inspections and excavation provide immediate visual assessments, they lack depth in quantitative analysis. MFL and ultrasonic testing offer high-resolution data for metallic pipes but require specific conditions for accurate results. Laser profiling effectively captures pipe deformation but is costly and requires pre-inspection cleaning. The assessment of plastic pipes remains a challenge due to the lack of real-time monitoring technologies. It is assumed that a combination of these methods will be required to collect the condition data comprehensively.

2.7.7 Consequence of Failure of Water Pipelines

Consistent with the conceptual frame, COF is treated as a five-component vector: economic, environmental, social, operational, and renewal complexity. Much of the literature quantifies only direct economic costs, with fewer studies estimating social outcomes (e.g., customer-hours without service, vehicle-hours of delay), environmental

receptors/impacts, or operational burden (e.g., isolation time, fire-flow deficit). Renewal complexity factors like depth, congestion, permit burden rarely appears explicitly despite its effect on feasibility and cost variance. The literature indicates that many renewal prioritization models focus solely on LOF, neglecting COF considerations. Such models align more with Reliability-Centered Maintenance (RCM) rather than risk-based asset management. However, water pipelines are Mixed Criticality Systems (MCS) where prioritization should factor in both LOF and COF, as some pipes (e.g., transmission mains) have far greater consequences of failure than others (e.g., smaller diameter pipes that are considered "redundant" and allowed to fail) (Burns and Davis 2018).

Most COF studies focus narrowly on direct economic costs, often overlooking broader social, environmental, and reputational impacts (Cromwell 2002; Gaewski and Blaha 2007; Raucher 2005). For example, pipeline failures can cause traffic disruptions, service outages, contamination risks, and environmental damage. The lack of readily available data has limited efforts to incorporate these factors into models.

2.7.8 Risk-based Renewal Prioritization Modeling

The definition of risk varies significantly across studies, leading to inconsistent risk assessments in water pipeline infrastructure (Aven 2016; Dawood et al. 2019; Kombo

Mpindou et al. 2022; Pollard et al. 2004). Many studies define risk as a function of either vulnerability, likelihood, or consequence of failure, without a standardized approach for combining these factors. This inconsistency mirrors similar challenges in supply chain risk management (Heckmann et al. 2015), whereas industries like nuclear energy have adopted common risk frameworks (Kaplan and Garrick 1981), allowing more standardized analyses.

Most risk models in the literature rely on deterministic cost estimation or weighted techniques for modeling COF (Agrawal et al. 2019; Cromwell 2002; Gaewski and Blaha 2007; Raucher 2017). However, deterministic models tend to oversimplify complex infrastructure systems, require precise parameter measurements that may be unavailable, and fail to account for utilities with limited failure datasets. To address these limitations, fuzzy logic-based expert systems can model uncertainty using linguistic variables and supplement ML training datasets with structured input-output mappings, improving their ability to learn realistic failure patterns.

Infrastructure renewal has evolved toward Multi-Criteria Optimization (MCO) methods, moving beyond traditional cost-minimization approaches. Early applications used decision theory-based techniques like Analytic Hierarchy Process (AHP) and Multi-

Attribute Utility Theory (MAUT) (Saaty 1980), but these approaches struggled to handle multi-objective problems effectively. Advances in AI and computational optimization led to the adoption of Genetic Algorithms (GAs) (Holland 1975), with the introduction of the Non-dominated Sorting Genetic Algorithm II (NSGA-II) (Deb et al. 2002) representing a breakthrough. NSGA-II enables simultaneous optimization of cost, failure risk, and environmental impacts (Gebre et al. 2021) and has shown promise for water pipeline renewal by allowing decision-makers to evaluate trade-offs between competing objectives. However, challenges persist in adapting these models to infrastructure-specific criteria and ensuring applicability across utilities of different sizes (Zimmerman & Faris 2010). Portfolio formulations should encode annual budgets, crew capacity, paving moratoria/seasonal no-cut windows, outage windows, and work-zone proximity. Many studies optimize with partial constraint sets; practice input (see practice review map) indicates that proximity and outage feasibility are frequent binding constraints and should be modeled alongside budgets.

Risk perception among utility asset managers varies widely, often leading to inconsistent prioritization decisions influenced by cognitive biases, such as availability heuristics and risk aversion (Slovic 1987; Kahneman & Tversky 1979; Aven 2016). Without a

structured risk-adjusted framework, these biases result in inefficient resource allocation. A formalized decision-making process that incorporates scientifically validated risk assessments is necessary to improve consistency across utilities (Gilboa 2009). This research proposes a hybrid AI approach where a fuzzy logic inference system generates structured input-output mappings to train ML models, following a knowledge distillation framework where an interpretable teacher model (fuzzy logic) transfers structured domain knowledge to a student ML model (Hinton et al. 2015). Prior studies demonstrate that incorporating fuzzy logic-based reasoning into ML training enhances model learning and interpretability (Shi et al. 2019). By integrating these methodologies, this research seeks to develop a transparent, interpretable, and scalable pipeline renewal prioritization framework that improves consistency and reliability in infrastructure decision-making.

2.7.9 Model Verification and Validation Review

Mathematical models provide simplified representations of complex systems, enabling large-scale computational analysis. However, their reliability depends on rigorous validation to ensure they accurately reflect real-world conditions (ASME 2009). Validation should be a continuous process rather than a one-time assessment, requiring iterative improvements. Given that ground truth is never fully known, validation protocols must

account for assumptions and experimental contexts. Additionally, models must be interpretable, explainable, and useful for decision-makers. Existing water pipeline validation approaches often rely on static, pass-fail criteria, failing to adapt to evolving datasets, changing requirements, or model refinements. These shortcomings, along with infrastructure constraints such as inaccessibility of buried pipelines, non-interruptible operational requirements, regulatory restrictions, and high excavation costs, make validation particularly challenging. A pragmatic approach is needed to balance resource limitations while ensuring trustworthy results, especially for smaller utilities with constrained data collection capabilities.

To address these challenges, this research draws from validation frameworks in aerospace, healthcare, transportation, and electric grids, where similar constraints exist. Following the EVV rubric in section 2.5, evidence is categorized as internal validation (resampling/hold-out within the same utility), external validation (cross-utility or temporally independent), and inspection/field concordance (e.g., digs, condition surveys). Uncertainty quantification is recorded as calibration checks, prediction intervals, and scenario/sensitivity analyses; explainability is recorded as global feature effects and local attributions (e.g., SHAP), and openness as availability of code and/or data. Across the

corpus, internal validation is common, external validation is less frequent, ground truthing is rare, and explainability/openness are uneven, key gaps the dissertation addresses.

Across complex engineered systems, validation is hard for shared reasons. Aerospace models depend on extensive simulations because space conditions are difficult to replicate while autonomous transport must contend with weather and human behavior and modern power grids juggle renewables that disrupt legacy assumptions, all characteristic challenges of complex systems modeling (Boccara, 2010). In healthcare, ethical and biological constraints limit data and complicate disease modeling (Marques et al., 2021). These examples highlight the need for adaptable, pragmatic and holistic validation strategies in the water sector.

Water pipeline risk models face similar validation gaps, primarily due to reliance on biased, utility-collected datasets without standardized forensic failure data (Kleiner & Rajani, 2001; Halfawy, 2008). This results in inconsistent risk assessments and limited generalizability. Additionally, many models fail to advance beyond pilot testing due to skepticism from asset managers regarding transparency and reliability for Capital Improvement Plans (Scholten et al., 2013). The absence of standardized validation protocols restricts model adoption, emphasizing the need for robust Verification & Validation

(V&V) frameworks to improve trust, scalability, and integration into asset management strategies (Halfawy, 2008; Scholten et al., 2014).

This research aims to address these deficiencies by standardizing risk analysis methodologies to enhance comparability across utilities, expanding COF modeling to incorporate economic, social, and environmental impacts, and developing lifecycle-based performance assessments. It will evaluate alternative risk modeling techniques, including hybrid fuzzy-ML approaches, establish rigorous validation protocols to ensure the models are robust, reliable, and applicable in real-world settings. By filling these critical gaps, this research seeks to improve model reliability, enhance accuracy, and provide a scientifically grounded framework for pipeline renewal decision-making.

2.7.10Gaps between Literature and Practice

This section explains what the literature currently delivers with what utilities require to implement renewal decisions under real delivery constraints. Table 2-6 presents these gaps by domains relevant to this research. The rightmost columns identify the gap and its likely root causes (data scarcity/heterogeneity, incentive and procurement structures, regulatory expectations, and validation-in-use hurdles).

Table 2-6: Matrix to evaluate key gaps between literature and practice and corresponding root causes based on research focus

Research Focus	Findings from Literature	Evidence strength	Utility Requirements	Gap	Likely root causes
LOF: Structural integrity (corrosion, loading/stress, hydraulics)	Statistical models not helpful for asset level prioritization; mechanistic and ML Models not generalizable across material/ diameter cohorts	Internal statistical evaluation and sensitivity common; ground truthing sparse	Asset level LOF with high explainability and temporal/ material/ diameter generalizability	High	Undefined failure; limited data and mechanism integration with soils/ traffic/ pressure data
LOF: Functional hydraulics/ capacity/ demand	Network simulations used ad hoc; rarely coupled to LOF	Locality level data; limited validation	Integrated hydraulic deficits in LOF and COF, tied to service levels	Moderate	Siloing between modeling groups; compute/ temporal data access and integration ability
COF: Economic	Direct repair costs considered but without detailed cost categories	High	Full operational costs (direct + indirect) and avoided-loss framing	Low	Accounting boundaries differ; data on indirect costs limited
COF: Environmental	Limited identified parameters and receptor/pathway modeling	Low	Receptor-based screening with defensible proxies	High	Monitoring gaps; geospatial linkage effort; unclear standards
COF: Social	Limited consideration of customer hours, business disruption or traffic delay	Low	Customer-hours and mobility impacts at project scale	High	Data silo barriers; need for repeatable estimators
COF: Operational	Limited consideration of isolation times and fire flow deficits	Low–moderate	Isolation/valving burden and fire-flow effects as standard outputs	Moderate–high	Issues with integrating topology and valve data to models; time to compute
COF: Renewal complexity	Seldom explicitly considered	Low	Constructability index (depth, congestion, permits)	High	Constructor input missing from analytic loops
Renewal constraints	Budget commonly included; others partial or ignored	Moderate	Full constraint set encoded with auditable rules	High	Data on moratoria/ outage windows not centralized;

Research Focus	Findings from Literature	Evidence strength	Utility Requirements	Gap	Likely root causes
					objective focused on cost only
Portfolio design	Multi-objective methods present; robustness based on validation experiments uncertain	Moderate	Risk-reduction per $, service KPIs, conflict avoidance	High	Parameterization of optimization algorithm and compute time
EVV	Sensitivity analysis dominates; external limited; ground truthing rare	Low	Cross-utility/ temporal external tests and field concordance	High	Access to test beds; cost to inspect; procurement not rewarding EVV
UQ & EXPL*	Sensitivity prevalent; calibration/intervals patchy; explainability improving	Uneven	Calibrated probabilities; SHAP*/ feature paths for audit	Moderate	Tooling/skills; legacy KPIs ignore calibration
Openness & reproducibility	Limited open code/data	Low–moderate	Re-runnable pipelines with schemas and versioning	Moderate	IP*/vendor concerns; data sharing agreements

*UQ (Uncertainty Quantification): methods to characterize uncertainty (e.g., calibration checks, prediction intervals, scenario/sensitivity analyses); EXPL (Explainability): techniques that make model reasoning auditable and understandable (global feature effects, local attributions); SHAP (Shapley Additive Explanations): a local attribution method that decomposes a model prediction into feature contributions; KPIs (Key Performance Indicators): metrics used to gauge outcomes (e.g., risk reduction per $1M, customer-hours avoided). IP (Intellectual Property): proprietary rights that can limit open release of data, models, or code.

A synthesis of root-causes from Table 2-6 shows recurrent of four causes: (i) data (sparse, nonstandard, siloed; limited condition/inspection ground truth), (ii) incentives (procurement rewards features over EVV and openness; short political cycles favor visible short-term fixes), (iii) process (departmental silos between asset management, operations, and engineering; limited constructor input for constructability), and (iv) regulation/governance (requirements emphasize compliance reporting over calibrated risk and decision

auditability). These root causes guide the design requirements and choices in the following chapters.

2.7.11Implications

The gaps matrix yields a concise set of requirements that shape the modeling and validation architecture. The requirements are written to be testable and to map directly to the following chapters.

- R1. Mechanism-aware LOF. Resolve LOF into structural corrosion/deterioration, structural loading/stress, and functional hydraulics/capacity; report discrimination and calibration with temporal validation.
- R2. Modular COF vector. Quantify five consequence classes i.e. economic, environmental, social, operational, renewal complexity in natural units with clear normalization for aggregation.
- R3. Constraint-aware portfolio design. Encode all binding delivery constraints (annual budgets, crew capacity, paving moratoria/seasonal no-cut windows, outage windows, work-zone proximity) with auditable rules.

- R4. Segment→Project→Portfolio aggregation. Provide explicit formation rules (contiguity, outage feasibility, street class, cost thresholds) and scheduling logic for multi-year plans.
- R5. Uncertainty quantification (UQ). Provide calibration assessment, prediction intervals where applicable, and scenario/sensitivity analyses at both model and portfolio layers.
- R6. Explainability-in-use and audit trails. Deliver global feature effects and local attributions (e.g., SHAP/rule paths) attached to each recommendation, with exportable audit logs.
- R7. Multi-layered EVV. Demonstrate internal validation, external/temporal validation (ideally cross-utility), and inspection/field concordance tests tied to decision thresholds.
- R8. Reproducible protocols and openness. Define schemas, units, and data provenance; provide re-runnable pipelines and parameter logs to enable review and transfer.

2.7.12Summary

This chapter mapped the state of the art and current practice for water-pipeline renewal across five domains namely, LOF, COF, renewal prioritization under delivery

constraints, model Evaluation/ Verification/ Validation (EVV), and adjacent domains that shape decision design. Bibliometrics and domain heatmaps showed that literature remains dominated by studies using statistical, mechanistic and ML techniques and suffer generalizability issues across material/ diameter cohort due to challenges in integrating data from various sources to create a performance metric based on validated failure mechanisms capturing various pathways like corrosion, loading related stresses and functional metrics relating to hydraulics, pipe capacity or service demands. COF modeling is frequently reduced to direct costs, with limited treatment of social, environmental, operational, and renewal-complexity impacts in natural units such as customer-hours, vehicle-hours of delay, isolation time, or constructability indices. On the renewal prioritization decision side, budget constraints are widely modeled but practical delivery constraints like crew capacity, paving moratoria and seasonal no-cut windows, outage windows, and work-zone proximity are inconsistently encoded, leading to portfolios that may not be buildable. Robust and scientific EVV is lacking across literature and practice with reporting limited to sensitivity or correlational metrics. However, cross-utility generalizability, ground truth validation, and explainability are often missing. A comparison against utility needs identified conceptual gaps (like often missing COF modules and over reliance on LOF aspect) and exploration of simpler techniques like weighted models often related to the lack of

technical modeling expertise. Often, water utilities are restricted by their political boundaries and lack data for model training and validation in addition to issues like fragmented efforts in departmental silos as well as political expectations related to term cycles. These findings motivate this dissertation's design requirements aiming at constructing, integrating and testing mechanism-aware LOF model; a modular COF model; and a constraint driven renewal prioritization model. All the models are developed to have high explainability and reproducible protocols to enhance real-world acceptance.

Chapter 3

Research Methodology

This chapter outlines the research strategies, procedures, and techniques used to achieve the development of the knowledge structured supervised models for water pipeline renewal prioritization. This chapter follows a structured adaptation of the research onion framework (Saunders et al. 2016). The methodological choices are expressed as philosophical assumptions, research approaches, strategies, choices, time horizons, techniques, and procedures. This chapter is structured as a study-design pipeline and not a project timeline to enhance reproducibility irrespective of the chronological order. Where timing affects interpretation, the temporal context of each step is described. This chapter first defines the data provenance and preprocessing protocols so that data can be reproduced exactly for future research. Then, this chapter explains the methods to develop the knowledgebase which is further used for structured supervised training of ML models. The results from these methods are presented in Chapters 4, 5 and 6. Additionally, data reliability indices are shown to support uncertainty quantification. Next, this chapter

presents the sampling and stratification strategies to achieve a statistically efficiency design and guide unbiased EVV tests. Finally, this chapter presents research hypotheses to guide the data analytics.

Figure 3-1 shows the layered summary of the research design and the methodological commitments from the outside in. This summary will be explained in detail in the following subsections. Read from outside to inside, the figure shows how philosophical stance and approach discipline the concrete design choices and analyses that follow.

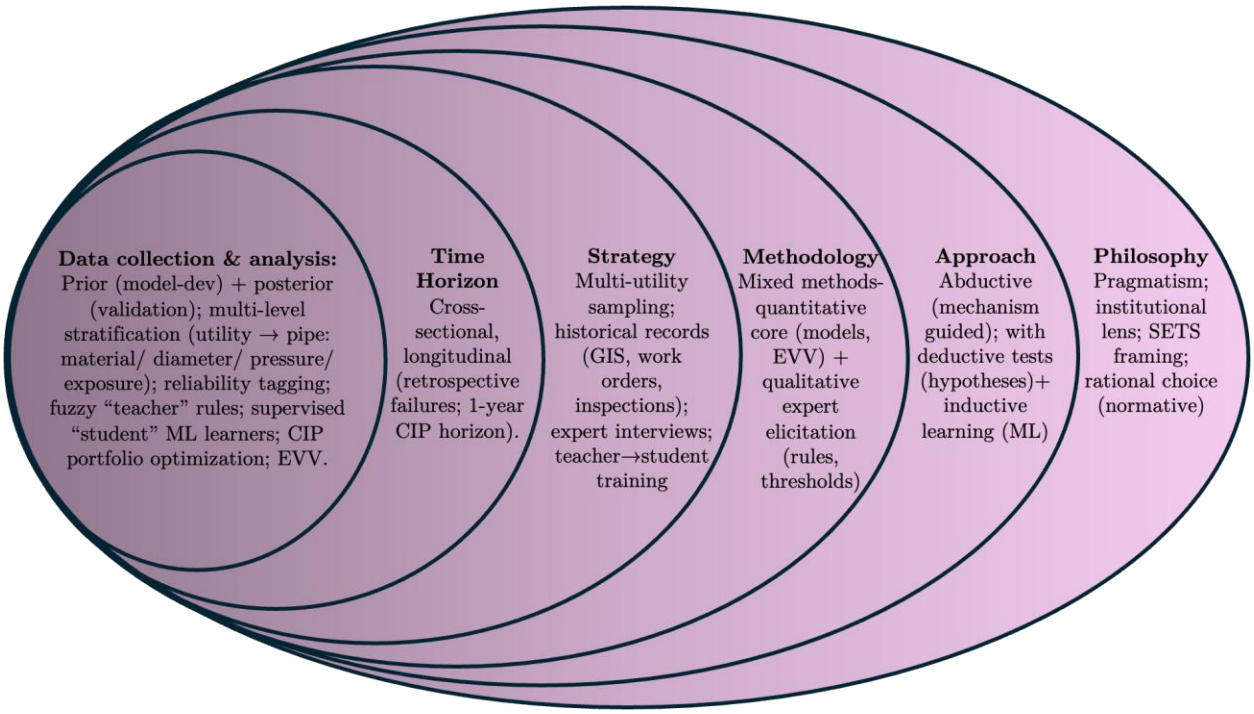

Figure 3-1: Layered summary of research study design

3.1 Research Philosophy

This dissertation adopts a pragmatist ontology within an institutional frame. Water utilities are social institutions whose infrastructure management policies and operational patterns shape decisions about objectively real infrastructure i.e. buried pipes and appurtenances, their materials and environments (soil, pressure, traffic), and the residential, commercial and industrial services they support. These assumptions are consistent with the view that collectively constructed institutions govern tangible systems (Searle, 1995; Scott, 2013). For this research, these entities and constraints (budgets, regulatory requirements, decision policies) are treated as valid objects of inquiry whose inherent patterns can be encoded for analysis and decision support.

Mathematical models are instruments used to encode and operationalize this ontology. The LOF model represents stochastic states of water pipeline assets that evolve in space and time with covariates aligned with failure mechanisms like soil corrosivity, internal pressure, diameter, and material. The COF model is developed to represent multi-dimensional impacts on operations, economics, customers, and the environment. The renewal-portfolio optimizer encodes decision constraints used by asset managers and/or project planners as feasibility sets (e.g., budgets, road moratoria, work-zone logistics,

equity and sequencing rules). The optimizer then searches for viable portfolios that satisfy these constraints while improving risk outcomes. This framing acknowledges that priorities shift as exposures, constraints, and organizational risk tolerance evolve (Scott, 2013).

Epistemology is mixed-methods where quantitative measurements (condition assessment, forensic examinations, failure/work-order histories) and qualitative knowledge (expert interviews, practitioner workshops, literature synthesis) are admitted as warranted evidence when they can be integrated and triangulated (Abowitz & Toole, 2010). Quantitative sources provide measurable anchors while supporting uncertainty estimation and outlier detection. Qualitative sources define construct boundaries, risk scenario narratives, and operational thresholds that guide feature design and the knowledge base used for structured, teacher-guided supervised learning in LOF and COF.

The decision-making processes behind pipeline renewal align with rational choice theory, introduced by von Neumann and Morgenstern (2007), which postulates that decision-makers aim to maximize utility by weighing costs, performance, and risks. This theory integrates well with risk-based modeling concepts, including uncertainty quantification and preference management, which are crucial when selecting among competing renewal alternatives.

3.2 Study-design overview

This section provides an outline of the experimental design pipeline organized into purpose, inputs, methods, and output subsections. All relevant details can be found in each of the subsequent chapters for each of the 3 main models in this work for future reproducibility. Although presented as an experimental design for reproducibility, several activities (e.g., feature selection and engineering, parameter assumptions, uncertainty analysis, verification) were revisited iteratively as evidence accumulated.

3.2.1 Data Collection, Compilation and Processing

Goal: Develop protocols to create ML-ready datasets to train, calibrate, and perform EVV on the LOF, COF, and portfolio models.

The database is collected in 2 parts: *prior* (model-development) data and *posterior* (validation) data. Sources include utility GIS inventories, failure/work-orders, hydraulic/operations layers, field/forensic measurements, expert workshops and interviews, open environmental/socioeconomic datasets, and documented social, environmental, and economic impacts of water-pipeline failures. Each record corresponds to a pipe segment defined between two network nodes (e.g., valves, hydrants, junctions) because most inventories originate from machine-generated geodatabases that connect these nodes, all asset

attributes and environmental covariates are resolved at the segment level, ensuring apples-to-apples comparisons across the dataset. All inputs pass a scripted process—ingest → de-duplicate → schema/units standardization → feature enrichment (spatial joins, vintage bands, exposure flags) → field-level reliability tagging—yielding high-quality tabular datasets for modeling and validation.

Prior model parameter development involves creating datasets, model parameters, and statistical distributions to generate accurate predictions and evaluate model performance. These parameters are derived from an extensive literature review and practice-based insights from multiple water utilities collaborating with the SWIM Center at Virginia Tech for the PIPEiD project (Sinha 2021). These utilities vary in size (small: <100,000; medium: 100,000–250,000; large: >250,000) and ownership models (public and private), providing diverse perspectives on pipeline renewal strategies from reactive to proactive approaches. The participation of utilities with varying management styles enhances the representativeness and reliability of the data. Figure 3-2 illustrates the geographical distribution of utilities that provided expert knowledge.

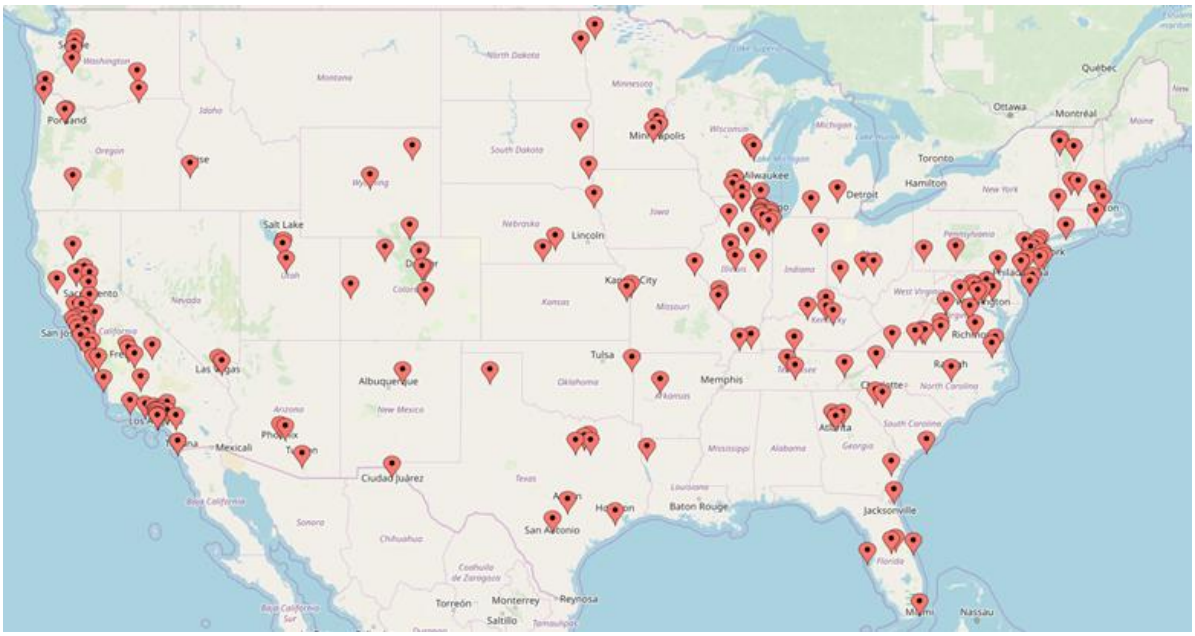

Figure 3-2: Practice review from utilities participating in the PIPEiD project shared useful real-world information related to pipeline performance and decision criteria typically unavailable in secondary datasets

Posterior model validation data consists of high-reliability datasets used as ground truth to validate the proposed models. These datasets are collected in collaboration with multiple water utilities in the U.S. ensuring a dataset with a range of operational and environmental conditions. Both qualitative and quantitative data are collected to capture a comprehensive understanding of pipeline renewal dynamics. Qualitative data provides contextual insights and expert perspectives, while quantitative data offers objectivity, reproducibility, and statistical comparability. Integrating both ensures model validity through triangulation. Data collected from water utilities includes various formats that

contribute to model development and validation. Asset inventory geodatabases provide spatially referenced pipeline system inventories, requiring GIS software such as ArcGIS Pro, QGIS, or AutoCAD for data transformation and analysis. Failure Records Spreadsheets contain historical work orders with unique pipeline segment identifiers, offering insights into past failures. Experiments and observations documents include reports on in-situ forensic evaluations, such as corrosion assessments, thickness measurements, and historical pipeline break analyses. Additionally, communications documents capture transcripts, recordings, and expert interviews, providing qualitative insights into utility decision-making processes. Beyond utility data, published literature serves as a source of secondary datasets, offering validated statistical distributions on failure impacts and renewal decisions from peer-reviewed studies. Online open-access databases supplement utility datasets with environmental and socio-economic parameters obtained from federal, state, and collaborative agencies, available in formats such as geodatabases, flat files, and structured reports. Lastly, direct field measurements are collected for model validation, providing ground truth data to assess the reliability and applicability of predictive models in real-world pipeline systems. This integration process from diverse data sources, categorized as prior (for model execution) and posterior (for validation experiments), are illustrated in Figure 3-3. The structured data collection approach ensures a robust foundation

for model development, calibration, and real-world applicability testing. Outputs from this stage are a metadata dictionary (definitions, units, spatio-temporal resolution, reliability) and tabular datasets. This design enables triangulation, reproducibility, and credible EVV by making data quality and provenance explicit.

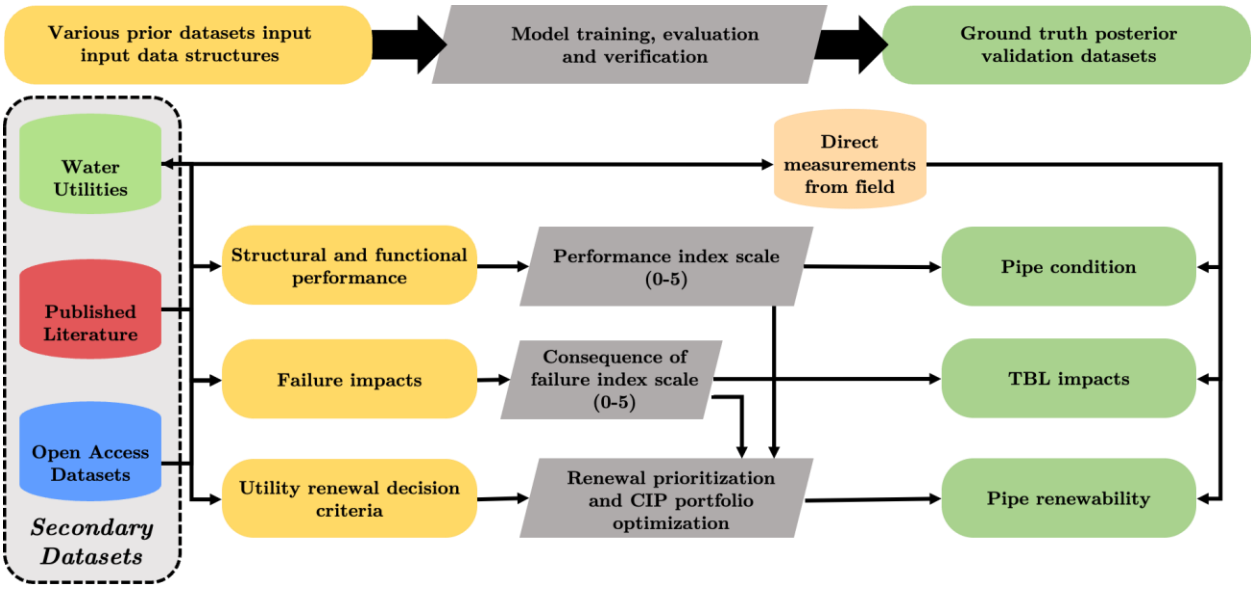

Figure 3-3: Use of prior and posterior datasets at different stages of this research

3.2.1.1 Water Utility Data Management Example

The data collected in this research comes from water utilities of varying sizes and requires structured protocols to translate raw datasets, survey responses, and expert heuristics into the proposed data structures. A representative example from Utility A

illustrates this process. Utility A provided its water pipeline inventory as a personal geodatabase (.mdb format), which was converted into a file geodatabase in ArcGIS Pro for geospatial analysis. The dataset consists of 173,106 pipe segments, with installation records dating back to 1915 and work order data available from 2006 onward. The summary of the data collected from Utility A is provided in Table 3-1.

Table 3-1: Data provided by Utility A for structural and functional performance aspects

Physical/Structural	**Operational/ Functional**
Node Identification Number	Pressure Zone
Node Length	Pipe Renewal Record
Pipe Material	Pipe Failure Record
Pipe Diameter	Node Elevation at vertical datum
Pipe Class	Hydraulic Grade
Installation Date	Pipe Internal Protection Type and Date
	Pipe External Protection Type and Date

The geodatabase includes four key datasets:

- Water Inventory – Contains core pipeline attributes such as material (CI, DI, PCCP, HDPE, PVC, Steel, Asbestos, Copper), diameter, length, pressure zone, pipe class, lining type and date, encasement, and cathodic protection.

- Work Orders History – A flat file listing all renewal activities since 2006, linking failure records to individual pipes based on unique identifiers. Key fields include failure date, cause, repair details, and defect descriptions.
- Hydraulic Grade – Provides high and low hydraulic gradient data by pressure zones, offering insights into theoretical operating pressures, though transient events like water hammers remain stronger determinants of pipeline deterioration.
- Pressure Zones – Defines clusters of pipelines experiencing frequent high-pressure conditions, serving as indicators of areas with increased pipeline stress.

To enhance the modeling dataset, the raw data from Utility A was enriched with additional variables. This involved modifying existing parameters (e.g., converting installation dates to vintage categories), applying educated assumptions (e.g., estimating pipe depth), and integrating external open-source federal datasets to incorporate key predictors such as soil characteristics, slope, land cover, and proximity to critical facilities. Despite being a large utility with structured data collection, significant gaps required assumptions and supplemental datasets, a challenge that is even more pronounced in smaller utilities with limited data availability. Additionally, data resolution varies. Some parameters (e.g., pipe length, shape, and hydraulic capacity) are collected at the pipe level, while others

(e.g., annual capital and operational costs) are recorded at the utility level. To enhance the modeling dataset, the raw data from Utility A was enriched with additional variables (see Table 3-2). This involved modifying existing parameters (e.g., converting installation dates to vintage categories), applying educated assumptions (e.g., estimating pipe depth), and integrating external open-source federal datasets to incorporate key predictors such as soil characteristics, slope, land cover, and proximity to critical facilities.

Table 3-2: Derived Parameters for Utility A

Physical/Structural	**Environmental**	**Social**
Pipe Depth	Soil Characteristics (~20 parameters)	Land Cover (High, Medium and Low Density Residential, Commercial and Industrial Areas)
Pipe Vintage	Climate Characteristics (Mean Annual Precipitation and Temperature)	Hospitals
Pipe Slope	Traffic Volume	Dialysis Centers
Pipe C-Factor	Groundwater Table Depth	Primary Education Institutions
	Extreme Events	Rail Tracks and Roadways
	Sensitive Wetland Areas	

The results of exploratory analysis performed on the pipe inventory and work order datasets, following spatial joining operations in GIS, are summarized in Table 3-3.

Table 3-3: Summary table for the distribution of different pipe materials

Material	Mileage (miles)	Mileage (miles)		Average Diameter (inch)	Average Diameter (inch)		Number of Repair Records		
		Lined	Unlined		Lined	Unlined	All	Lined	Unlined
CI	2,574.7	935.5	1,639.1	8.2	7.8	8.4	15,615	5,445	10,170
DI (Asphaltic)	2,758.8	238.5	2,520.4	8.6	8.5	8.6	3,599	1,185	2,414
DI (Zn-Coated)	5.5	1.1	4.3	11.6	17.0	10.6	12	0	12
DI (PE-Coated)	0.7	0.0	0.7	11.7	8.0	12.9	5	2	3
PCCP	350.0	5.8	344.2	28.7	64.8	28.1	183	0	183
HDPE	0.1	0.0	0.1	8.0	NA	8.0	0	0	0
PVC	7.4	3.0	4.4	8.1	7.9	8.2	5	3	2
Steel	39.5	13.4	26.1	40.0	65.8	32.4	5	1	4
Asbestos	3.0	0.1	2.9	8.7	9.5	8.7	6	0	6
Copper	1.3	0.0	1.3	1.8	NA	1.8	5	0	5
Unknown	18.3	1.4	16.9	12.5	15.1	12.4	18	3	15
All Pipes	5,759.2	1,198.8	4,560.4	9.0	8.3	9.2	19,453	6,639	12,814

Despite being a large utility with structured data collection, significant gaps required assumptions and supplemental datasets, a challenge that is even more pronounced in smaller utilities with limited data availability. Additionally, data resolution varies as some parameters (e.g., pipe length, shape, and hydraulic capacity) are collected at the pipe level, while others (e.g., annual capital and operational costs) are recorded at the utility level.

3.2.2 Sampling and Reliability

Goal: Build statistically efficient, *bias-controlled* training and assessment sets that reflect a realistic water utility dataset.

Data stratification is performed at the following multiple levels:

- Utility level (coverage first): Ensure representation across utility size (small/medium/large), ownership (public/private), and ecological/operating cohorts (climate/soil/urban–rural/coastal/mountainous). This preserves diverse decision contexts based on factors like resource availability (time, expertise, data, finance), risk attitudes (risk-taking/risk-neutral/risk-averse), management styles (reactive or proactive).
- Pipe level (mechanism aware): Stratify by material, diameter, pressure/exposure (including pressure class, vintage, soil class and other informative cohorts) while preserving important distinctions based on deterioration mechanisms (e.g., PCCP vs metallics vs plastics).

Each record carries a field-level reliability index $r_i \in [0,5]$ from preprocessing to improve data quality over exclusion. This index is described in detail in the following chapters. Repairs, cross-checks, and imputations are documented, and only irreparable records

are flagged. During implementation, r_i informs the decision maker about the relative level of confidence to be ascribed to results in addition to encouraging better data collection for future.

3.2.3 Targets and Constructs with a Knowledge Structured "Teacher"

Goal: Precisely define outputs from each model and how those outputs are grounded in observable mechanisms and measurable units. Also, introduce the knowledge-based "teacher" that makes these targets structured and predictable.

LOF: A 0–5 index capturing structural and functional condition over a one-year capital-planning horizon. Each of the 5 index bands have explainable, mechanism-aligned and measurable criteria (e.g., corrosion/wall-loss for metallics, wire-break progression for PCCP, hydraulic degradation/ ovality for PVC) and guardrails (feasible deterioration slopes; material/diameter eligibility).

Teacher FIS (LOF): For each material family, a hand-crafted fuzzy inference system encodes expert/empirical rules (e.g., *IF soil corrosivity high AND pressure transients frequent THEN LOF* $\geq$ *3 (Poor)*), producing: (i) an output LOF score (0–5), (ii) an uncertainty band, and (iii) contracts (selective monotonicity, feasibility) that downstream

learners must respect. These yield structured supervision for ML: the "student" imitates the teacher while calibrating to data.

COF: A modular 0–5 index that aggregates economic, operational/service, environmental, social/equity, and renewal-complexity factors. Each module reports a separate 0-5 index with a corresponding measurement map (e.g., repair costs in USD, customer-hours, gallons lost).

Teacher FIS (COF): Module-specific fuzzy rules link context to consequence (e.g., *IF pipe serves critical facilities AND outage duration long THEN service consequence ≥ 4 (Catastrophic)*), again outputting a 0–5 score, uncertainty, and constraint contracts for the student models.

Risk: Define $\mathcal{R} = f(\mathrm{LOF}, \mathrm{COF}; \boldsymbol{\theta})$ where LOF and COF are dimensionless indices on $[0,5]$ scales. When needed for reporting or portfolio scoring, COF (or its modules) can be mapped to units and combined with a severity function of LOF to produce risk estimates with measurable units. The outputs will be measurable and explainable definitions for index scale measurements.

3.2.4 Modeling stack (LOF, COF, Portfolio)

Goal: Train ML algorithms to learn patterns from structured Input/ Output datasets created using fuzzy teacher systems (for LOF and COF) and integrate with multi-criteria renewal decision algorithm to prepare water pipeline CIP portfolio.

LOF model: Supervised learner with features tied to failure mechanisms (e.g., wall-loss drivers for metallics; wire-break surrogates for PCCP; hydraulic degradation for PVC). Fuzzy teacher datasets impose selective monotonicity where causal direction is known (e.g., higher corrosivity translates to non-decreasing LOF). Before training, class imbalance is handled via stratified reweighting. Outputs include banded LOF (0–5), calibrated scores, and uncertainty.

COF model: Supervised learner combining patterns across economic impacts, operational impacts, environmental impacts, social impacts, and renewal-complexity fuzzy teacher modules.

Portfolio optimization: A multi-objective formulation over portfolio sets S: minimize risk, cost, water losses, and downtime; maximize equity, concurrent projects, economic opportunities. This formulation is subject to hard budget constraints and

neighborhood pipe selection approach and the optimization algorithm computes Pareto frontiers to provide a ranked list of pipe renewal projects.

Outputs are versioned model specifications (data schema, features, constraints, calibration, limits) and configuration hashes (including the code, hyperparameters and seeds) to ensure exact reproducibility.

3.2.5 Training, testing and implementation

Goal: Ensure training on data that represent real utility conditions, enriched with relevant external (critical facilities, soil and traffic characteristics etc.) reliable federal, state and local datasets, and quality-checked end to end.

The training process requires a compiled dataset that (i) reflects the actual asset mix from partnering water utilities (materials, diameters, vintages, pressures, work orders, inspections); (ii) is enriched with external context like soil attributes from USGS SSURGO (e.g., metallic/concrete corrosivity, drainage, texture), traffic exposure from DOT AADT, municipal layers (e.g., hospitals/critical facilities, land use, sensitive areas), and other open reliable datasets; and (iii) includes knowledge-structured targets from fuzzy inference

"teacher" systems that encode material-specific deterioration and consequence mechanisms. Where performance categories are imbalanced, conservative oversampling (and/or class-weighted losses) is applied within training folds only. Quality improvements (unit normalization, plausibility checks, reliability weights, and explicit imputation flags) are applied consistently across models. The entire compiled dataset is partitioned by pipe material (CI, DI, Steel, PVC, PE, PCCP, RCP and AC) and three diameter bands (<8 in, 8–24 in, >24 in), and separate models are trained per segment to reflect mechanism differences. For example, the focus is to capture corrosion/wall-loss mechanisms in metallics, wire-break behavior in PCCP, hydraulic degradation in PVC while diameter shifts operational characteristics (like depth and availability of data), and consequence scales (different risk attitudes for different diameter pipes). Segmentation avoids averaging across incompatible failure patterns, improves calibration for minority cohorts (e.g., large-diameter mains), and simplifies implementation where material-specific models can be versioned, enabled, or retired as a water utility pipe inventory evolves (e.g., decommissioned AC) without disturbing others.

3.2.6 Evaluation, Verification and Validation (EVV)

Goal: Demonstrate that the models work based on tests designed to measure performance on synthetically/heuristically developed, water utility basic level datasets and water utility reliable ground truth datasets.

The proposed "teacher" and "student" models are rigorously tested during the evaluation, verification and validation phases, where aspects like the alignment with the intended behavior (evaluation), training performance (verification) and agreement with the ground truth (validation) is explained. This 3 step rigorous process is shown in Figure 3-4. The evaluation phase ensures that "teacher" fuzzy models perform as intended and align with theoretical expectations. This step involves rigorous internal testing using synthetic and secondary data to assess model design, correctness in the directionality of predictions, sensitivity, and robustness under edge conditions, including extreme (lowest, highest) and average cases including the output class boundaries.

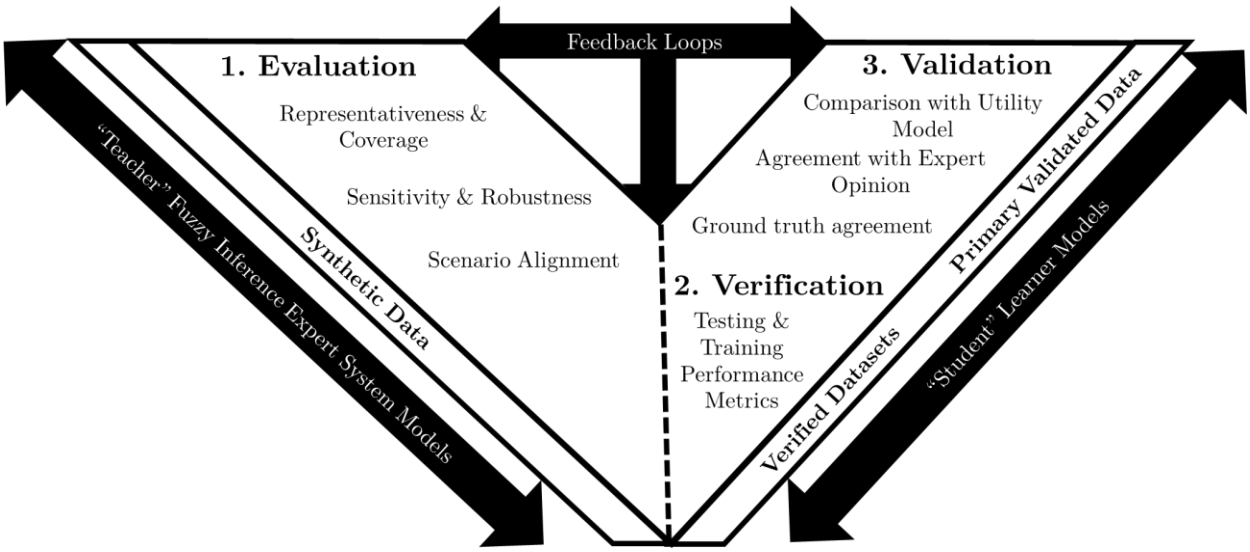

Figure 3-4: Model Evaluation, Verification and Validation (EVV)

By benchmarking against key criteria such as representativeness, consistency, and scenario alignment, the models are refined to ensure they reliably capture real-world dynamics. The model verification stage tests whether the "student" trained ML algorithms can effectively learn and generalize from the fuzzy logic model. Classification metrics are observed using confusion matrices to establish model performance in predicting all 5 output classes. This step serves as an initial verification of whether ML models can replicate the structured decision-making process of the fuzzy logic system while improving scalability and computational efficiency. After internal evaluation and student-model verification, we begin staged collaboration with utilities under NDA to validate the student models on real systems. We start with low-friction comparisons that require easy to share

dataset, overlaying our LOF with each utility's current scores, sharing rank lists within material and diameter cohorts, and reviewing outliers together. When alignment looks promising, we move to selected assets with high-reliability evidence like condition assessments and exhumed-pipe forensics to compute quantitative conformity (e.g., error distributions and concordance, the share of cases within ± 1 LOF band) and to document qualitative limits of applicability. Each cycle produces a tracked set of refinements to student training data where needed, and a brief acceptance note describing where the model is ready for operational use and where caution or added data is required.

3.2.7 Research Hypotheses

The goal of this research is to produce a scientifically validated decision support tool that can accurately create a priority list of pipe geospatial nodes for future condition assessment and renewal. To successfully implement the abovementioned methods, three goals are outlined. Goal 1 (LOF): develop a robust mechanism-aligned likelihood-of-failure model that covers corrosion/deterioration, loading/stress, and hydraulics/capacity. Goal 2 (COF): construct a modular consequence model spanning economic, environmental, social/service, operational, and renewal-complexity components with calibrated uncertainty. Goal 3 (Portfolio): optimize renewal actions under budget and constructability constraints

to produce portfolios that dominate rank-only baselines on risk outcomes and operational feasibility. Corresponding overall hypotheses are shown in Table 3-4 and are tested through a layered Evaluation, Verification and Validation (EVV) program.

Table 3-4: Hypotheses tested in this research categorized by each of the research goals

Goal	H_x	Hypothesis
Goal 1: LOF model	**H_{1a} (Mechanism and context coverage)**	The LOF framework explicitly encodes structural, functional, and environmental drivers across major materials and diameter bands, extending beyond age/diameter practice by covering a broader set of documented deterioration mechanisms.
	H_{1b} (Student learning fidelity)	Student ML LOF models learn the fuzzy-teacher mappings with high accuracy and macro-F1 on held-out and synthetic stress-test data, yielding strongly diagonal confusion matrices with very few multi-band misclassifications.
	H_{1c} (Ground-truth concordance)	In independent validation cohorts with condition measurements (wall-thickness loss, wire-break counts) and retrospective failures, higher LOF bands are associated with worse measured condition and higher failure frequencies, with ordinal agreement statistics significantly above chance and errors dominated by $\pm$1-band deviations.
	H_{1d} (Expert concordance and face validity)	For curated LOF scenarios, asset managers and field staff judge the predicted LOF bands as broadly consistent with operational experience at rates well above chance (including tolerant $\pm$1-band agreement), and observed disagreements are explainable by data lineage or explicit policy choices rather than erratic model behavior.
Goal 2: COF model	**H_{2a} (Dimensional coverage and representativeness)**	The COF framework decomposes consequence into explicit economic, environmental, social/service, and operational sub-indices, and membership-function panels plus best/average/worst scenarios demonstrate coherent, monotone coverage from low- to high-impact combinations in each dimension.
	H_{2b} (Modular behavior and structural verification)	Within each COF dimension, increasing adverse inputs (for example, higher repair costs, more critical customers, tighter access constraints) produces monotone increases in the corresponding sub-index and in the overall COF band, and global sensitivity analysis shows no single parameter or module dominates the index, supporting stable modular substitution.
	H_{2c} (Agreement with existing utility indices and expert judgement)	When compared with incumbent utility consequence indices and expert scenario ratings, COF bands show strong ordinal alignment, with most cases on or near the diagonal of confusion matrices and positive, substantial rank correlations, and divergences trace to scale/scope differences rather than incoherent model behavior.
	H_{2d} (Ground-truth consequence calibration)	For documented main-break events with usable consequence descriptions, higher COF bands align with more severe observed proxies (for example, outage duration, disruption, visible damage), and confusion matrices plus ordinal metrics indicate broad calibration with only a small number of explainable two-band outliers.

Goal	H_x	Hypothesis
Goal 3: Renewal Prioritization model	**H_{3a} (Portfolio effectiveness under constraints)**	Under fixed budget constraints and realistic candidate pre-screening, GA-optimized renewal portfolios built from LOF, COF, and auxiliary scores capture more risk per unit cost than simple rank-only or cost-weighted baselines across multiple utilities.
	H_{3b} (Decision alignment and acceptability)	In scenario-based validation with three utilities, planners' and asset managers' preferred options align with GA-recommended portfolios at rates well above chance, and where they diverge, qualitative comments point to scope or data limitations rather than systematic contradictions.
	H_{3c} (Stability and robustness of portfolio recommendations)	Across changes in scalarization weights, random seeds, and utility datasets, the GA portfolios occupy a compact region of the risk–equity trade-off space and scalar performance metrics vary modestly, indicating that the recommended portfolios are robust to reasonable variations in preferences and initialization.

3.3 Summary

This chapter established the design of this study that underlies all results: (i) data provenance and preprocessing with field-level reliability, (ii) multi-level stratification and sampling, (iii) definition of targets and constructs (LOF, COF, Risk) with measurement maps, (iv) a teacher–student modeling stack, (v) training and implementation protocols for exact reruns, (vi) EVV procedures, and (vii) ethics and reproducibility guardrails. Together, these choices turn heterogeneous utility data into auditable evidence for renewal decisions.

Scope and assumptions: The horizon is one capital-planning year. LoF and CoF are dimensionless 0–5 indices with band definitions tied to observable mechanisms and

consequences; risk uses these indices with unit discipline when reporting decision value. Results are conditioned on the available data, documented imputations, and declared constraints (budget, constructability, equity), all recorded in the metadata and model cards.

Threats to validity and mitigations: Potential risks include measurement error, cohort imbalance, and limited external generalizability. Mitigations include reliability-aware preprocessing instead of wholesale exclusion, stratified design and holdouts across utilities/time, calibration checks, and uncertainty/sensitivity analyses with prespecified operating thresholds.

The next three chapters implement this protocol in turn. Chapter 3 (LOF) develops the likelihood-of-failure index: data and features, teacher rules and student learners, calibration, uncertainty, and EVV. Chapter 4 (COF) presents the modular consequence model, its measurement maps and aggregation, with uncertainty and EVV. Chapter 5 (Portfolio) formulates multi-objective renewal optimization under constraints, reports Pareto analyses, and quantifies decision value (risk reduction per dollar and customer-hours avoided). Each chapter cites the exact data cut (hash), configuration, and split manifests used, and reports results with confidence intervals and calibration diagnostics. The

appendices supply the metadata book, split manifests, environment receipts, and the reproducibility ledger (artifact $\rightarrow$ script $\rightarrow$ hash). Together these elements complete the methodological specification and set up the model development and testing that follow.

Chapter 4

Likelihood of Failure Model

This chapter develops and tests the LOF index models for buried drinking-water pipes over a one-year capital-planning horizon. LOF provides the predictive signal that, together with COF, drives risk analysis and renewal prioritization decision making. The chapter proceeds from explaining failure mechanisms to mathematically formulate into robust LOF prediction models. This chapter derives features and criteria from the literature and practice, assembles multi-utility datasets enriched with federal/state open access datasets, encodes expert knowledge in per-material and diameter fuzzy "teacher" rules, and trains data-driven "student" learners. ML models are stress-tested at nine different levels to challenge the model targeting typical and edge cases under a detailed EVV framework. Outputs are a calibrated, explainable 0–5 LOF score with uncertainty and guardrails, designed to integrate with the risk and portfolio framework. The chapter closes with field-validation protocols so performance can be monitored and improved as new evidence arrives.

4.1 Goal and Scope

The LOF index prediction models quantify how likely a pipe segment is to fail over a one-year capital-planning horizon. It integrates structural conditions (e.g., wall-loss, cracks, wire breaks) and functional performance (e.g., capacity loss from roughness/ovality, pressure stress). LOF is reported on a 0–5 scale with five labeled bands and explicit criteria so it can drive three decisions: (i) triage for inspection and monitoring, (ii) prioritization for renewal projects by integrating with COF and decision criteria within a budget constraint, and (iii) assessing overall health of water pipeline transmission and distribution system. The model covers the following contexts found in U.S. drinking-water systems:

- Material classes: Cast Iron (CI), Ductile Iron (DI), Steel (ST), Polyvinyl Chloride (PVC), Polyethylene (PE), Prestressed Concrete Cylinder Pipe (PCCP) and Asbestos Cement (AC).
- Diameter classes: <8 in, 8–24 in, >24 in.
- Lifecycle contexts: Commissioned pipes in operation influenced by pressure and transients, soil/groundwater corrosivity and bedding, traffic loading, climate/hydrology, demand/criticality, redundancy, among other factors.

In addition, segment-level representation of pipes is pragmatically defined by the length of pipe material segmented by network nodes such as valves or hydrants. This characteristic is defined by the geodatabase as a unit of measurement and ensures that attributes, exposures and visualizations are aligned consistently and enables the utilization of model results for visualization and decision support.

Knowledge-structured "teachers" are built first per-material category using Fuzzy Inference Systems (FIS). Fuzzy inference is used for the teacher models because it is the only practical way to encode failure mechanism-level knowledge, measurement uncertainties, and policy guardrails in one transparent "white-box" knowledgebase. A fuzzy set lets a quantity have graded membership between 0 and 1 (e.g., "somewhat high pressure"), so inputs that are noisy, proxy-based, or sparse do not force brittle thresholds. Linguistic IF–THEN rules map directly to failure mechanisms are linguistic, and therefore interpretable. Membership shapes (triangles, bells, sigmoids/Z) let us tune overlaps to measured ambiguity and data-reliability tiers, while the aggregation and defuzzification step (conversion of the fuzzy result to a 0–5 LOF) yields a continuous, calibrated score together with an uncertainty band derived from rule conflict and coverage. Competing "teacher" options either overfit or obscure the mechanism. Linear models are too rigid for interacting

stressors and unconstrained machine-learning models are efficient but opaque and hard to reconcile with physical priors. The fuzzy teacher therefore provides (i) an interpretable, mechanism-aligned prior, (ii) sample-efficient supervision for the student models, and (iii) an explanation layer that can be reviewed with utility experts and traced back to measurable anchors (e.g., remaining wall thickness, pressure, frost action). Here rules are manually set up to represent failure mechanisms (e.g., "high soil corrosivity + frequent pressure spikes → LOF at least Poor"). These rules produce a LOF score (0–5) and an uncertainty band. Subsequently, data-driven "students" then learn from the input-output structured datasets se FIS to learn the teacher signals reliably, yielding calibrated predictions. The FIS can also serve as a prior knowledgebase containing linguistic explanations for the otherwise "black-box" models. This two-stage design injects domain knowledge, improves sample efficiency, and keeps outputs interpretable.

Next, EVV is performed on the teacher-student models. *Evaluation* checks internal behavior (coverage, sensitivity, calibration) of the teacher models. *Verification* confirms that student models correctly learn the teacher and utility patterns through training and testing performance and stress testing. To stress-test the trained student models, nine deliberately chosen test bands exercise edge cases. The nine bands are used to create a

synthetic input-output dataset that is previously unseen by the student learner models and ensure confidence in predictive accuracy and directionality. *Validation* tests external realism by checking student learner models' performance on real world utility data and unseen ground-truth observations. Detailed, documented field-data collection protocols specify what to collect (e.g., wall-thickness, wire-break counts, ovality, pressure logs), how to score reliability, and what acceptance windows trigger adoption or revision. Results can be tracked over time, and any discrepancies can be fed back into the rule base, features, or training data to ensure continuous improvement. Versioned artifacts (data cut, configuration, split manifests, environment receipts) across the EVV phase makes the protocols auditable and reproducible. The overall process is illustrated in Figure 4-1.

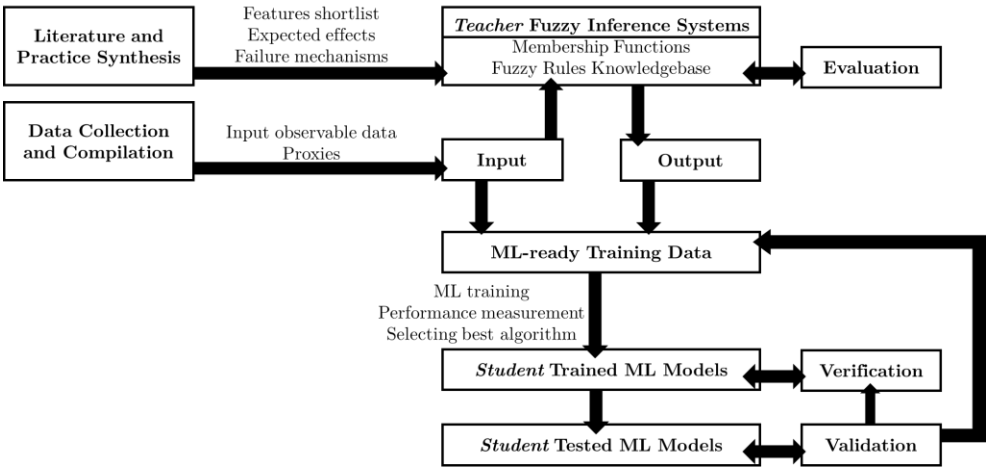

Figure 4-1: LOF modeling workflow illustration

4.2 LOF Grounding in Failure Mechanisms

The LOF model is anchored in how pipes fail. For each material and diameter family, the literature and practice converge on a set of deterioration and loading mechanisms (e.g., soil-side corrosion in metallics; loss of prestress in PCCP; ovality and pressure transients in PVC/PE; chemical attack and poor bedding in concrete). These were previously summarized in Table 2-4. Those mechanisms are used to (i) pick measurable features and proxies, (ii) state the expected direction of effect on LOF (selective monotonicity), and (iii) define low/medium/high bands for fuzzy membership functions. Membership function ranges and cut points are set by mechanism-aligned references: design ratings based on pressure class and Standard Dimension Ratio (SDR), standards and guidance thresholds (e.g., low soil resistivity = high corrosivity based on recommendations from USBR, AWWA and DIPRA), inspection metrics (e.g., ovality %, wall thickness), and empirical quantiles calibrated to failure histories. SDR (Standard Dimension Ratio) is the ratio of a pipe's outside diameter to its minimum wall thickness:

$$SDR = \frac{D_0}{t}$$

A lower SDR means a thicker wall and therefore a higher pressure rating/greater surge tolerance (e.g., SDR 17 is stronger than SDR 21). For thermoplastic pipes (PVC/PE), allowable working pressure is roughly inversely proportional to SDR (per thin-wall hoop stress relationships), so SDR directly encodes pressure capacity.

The "teacher" fuzzy rule base is then hand-crafted from these ingredients (e.g., "IF corrosivity is high AND Cathodic Protection is absent THEN LOF is High or worse"), giving structured targets for student learners and explainable linguistic rules for decision makers and model validators.

4.3 LOF as an Output Metric for Modeling

Literature review showed that the real world application of all the models developed in the past suffered from a lack of coherent, specific and consistent definition of terms like failure, leak and break typically used as the predictive outcomes in many articles. This section explains the definition of the term failure and the specific definitions of the output LOF index so that it is measurable.

4.3.1 Definition of Pipe Operational Failure

In this research, "failure" is defined in two complementary ways that match utility practice. Structural failure is a loss of integrity that creates a physical change in the pipe such as a rupture, through-wall perforation, a loss of wall thickness due to corrosion or leaching or growth rate of broken prestressing wires in PCCP, or ovality beyond design in thin-wall plastics and GRP. Functional failure is the inability to meet level-of-service targets even if the pipe can withstand the same physical stresses and the structure remains intact. For example, sustained headloss at required flows, persistent pressure deficits at customers, unacceptable leakages through joints, recurrent transients that force derating, or repeated water-quality non-compliance traceable to the asset. Recurrent transients that force derating means the system is seeing repeated pressure surges (quick up-and-down swings from pump starts/stops, fast valve moves, hydrant operations, power trips, etc.) often enough and large enough that the utility must permanently operate the pipe below its nominal design limits to stay safe. In practice, that "derating" can be lowering the maximum zone pressure, capping pump speeds, lengthening valve-closure times, or changing setpoints so those spikes stay within safe bounds. It protects the pipe from fatigue, joint damage, wire-break growth (in PCCP), and crack initiation (in plastics/GRP), but

it can also reduce level-of-service for example, fire-flow or peak-hour pressures at high points. If that reduced operating envelope means the asset can't meet the stated level-of-service, we count it as a functional failure. A real life example could look like this. A main designed to run up to ~115 psi repeatedly sees 40–60 psi spikes when a booster cycles, so the utility caps it at 90 psi and slows ramps. As a result, the spikes stop, but some cul-de-sacs now fall under the pressure target at peak demand.

A failure event is counted when a pipe is unable to structurally withstand the operational stresses or functionally unable to match the promised minimum level of service beyond a defined duration. These thresholds should align with the local level-of-service policies set by the water utility.

4.3.2 Development of Target Output LOF Index

The LOF target output is the probability that an asset will cross the above failure threshold within a chosen horizon H. This research assumes $H = 12$ months, which is the usual window for capital improvement or replacement planning for a water utility. The range can be modified from 3 to 60 months to match planning cadence. The model ties this output LOF target output to observable and measurable evidence. LOF output index is expressed with five bands (Very Low, Low, Moderate, High, Very High) matching

output classes used by water utilities in asset management programs as shown in Table 4-1. Moreover, a 0-5 index provides a balanced trade-off between granularity in results and computational simplicity and eases integration with other asset classes. The LOF index can reported as a continuous numerical quantity or a class label for communication. When the model is applied to a real-world distribution system, it is expected to have the majority (typically over 90%) of the pipes falling into the "Moderate" middle rating. To respect the differences in classes as we move towards more extreme values, each level is further subdivided to allow finer differentiation between classes and future expansion of the output rating scale as better data becomes available.

Table 4-1: Output LOF Index (0-5) detailed class definitions

Index		Detailed Definitions
0-1	Very Low	Pipes rated as **Very Low** exhibit pristine structural integrity with no signs of wear or degradation and maintain peak functional performance. **0.0-0.5**: Pipes show no internal or external corrosion, cracks, or coating/lining degradation. Concrete pipes have no wire breaks, no cylinder exposure, and no signs of concrete deterioration. Plastic pipes have no signs of ovality or deformation. Full hydraulic efficiency is maintained with no internal tuberculation. No service interruptions ever. **0.5-1.0**: Pipes have more than 95% remaining wall thickness with internal tuberculation less than 0.2 mm. Concrete pipes exhibit no wire breaks, and lining/coating is intact. Plastic pipes show minimal ovality (<2%). No service interruptions ever, reflecting near-new conditions.
1-2	Low	Pipes rated as **Low** show minor signs of wear and tear but maintain overall good structural integrity and functional performance. **1.0-1.5**: Pipes have more than 95% remaining wall thickness with internal tuberculation up to 0.5 mm. Concrete pipes have no wirebreaks, minor coating/lining wear, and early-stage corrosion. Plastic pipes may have ovality up to 2%. No service interruptions ever. **1.5-2.0**: Pipes have more than 90% remaining wall thickness with internal tuberculation up to 1 mm. Concrete pipes have no wirebreaks, minor concrete deterioration, and slight cylinder exposure. Plastic pipes may exhibit ovality up to 7%. No service interruptions ever.

2-3	Moderate	Pipes rated as **Moderate** have moderate wear and some structural and functional degradation. **2.0-2.5**: Pipes have more than 85% remaining wall thickness with internal tuberculation of 1-2 mm. Concrete pipes do not have any wirebreaks, moderate coating/lining deterioration, and early concrete deterioration. Plastic pipes may have ovality up to 10% and minor surface deformation. Up to 1 service interruption in the past 10 years. **2.5-3.0**: Pipes have more than 85% remaining wall thickness with internal tuberculation of 2-3 mm. Concrete pipes may have up to 2 wirebreaks, moderate cylinder exposure, and signs of early-stage corrosion. Plastic pipes may have ovality up to 15% and surface cracks. Up to 1 service interruption in the past 10 years.
3-4	High	Pipes rated as **High** show significant structural degradation and reduced functional performance, leading to frequent service interruptions. **3.0-3.5**: Pipes have more than 80% remaining wall thickness with internal tuberculation of 3-4 mm. Concrete pipes may have up to 5 wire breaks, significant coating/lining deterioration, and moderate concrete spalling. Plastic pipes may have ovality up to 20% and moderate deformation. Up to 2 service interruptions in the past 10 years. **3.5-4.0**: Pipes have more than 80% remaining wall thickness with internal tuberculation of 4-5 mm. Concrete pipes may have up to 10 wire breaks, severe cylinder exposure, and advanced corrosion. Plastic pipes may show ovality up to 25% and significant deformation. Up to 2 service interruptions in the past 10 years.
4-5	Very High	Pipes rated as **Very High** are severely degraded with major structural issues and extremely poor functional performance, requiring immediate replacement. **4.0-4.5**: Pipes have more than 75% remaining wall thickness with internal tuberculation of 5-6 mm. Concrete pipes may have more than 15 wire breaks per segment, severe coating/lining failure, extensive concrete spalling, and severe cylinder exposure. Plastic pipes may have ovality exceeding 30% and severe deformation or cracks. More than 2 service interruptions in the past 10 years. **4.5-5.0**: Pipes have less than 75% remaining wall thickness with internal tuberculation over 6 mm. Concrete pipes exhibit more than 15 wire breaks, coating/lining failure, concrete collapse, and severe corrosion. Plastic pipes show extreme ovality, deformation, or structural failure. More than 2 service interruptions in the past 10 years.

For metallic pipe, the output is defined based on measured wall-loss and pitting depth from inspection, corrosion potential, leak history, pressure and surge exposure, and soil corrosivity, combined with normalized headloss at a reference flow. For PCCP and bar-wrapped pipe, the output is defined based broken-wire counts and their growth over time, mortar condition including cracking, delamination and carbonation depth, cylinder condition, transient exposure, and bedding quality including rock point contacts. For PVC and PE, the output is defined based ovality and strain versus bedding conditions,

documented impact or installation damage, joint integrity, surge exceedances and, where relevant, permeation or environmental stress cracking. For AC, the output is defined based internal chemistry (aggressiveness and alkalinity), sulfate exposure, evidence of lime leaching, external soil chemistry, cracking or embrittlement indicators, and bedding quality. For GRP, the output is defined based laminate condition (blistering, cracking, disbonding), measured creep or deflection under load, installation damage, and joint performance. Cross-cutting hydraulic observables like headloss, pressure deficits, leak events, and transient metrics add context to all pipe material LOF measurements.

4.4 Input Data and Feature Specifications

This subsection defines exactly what we feed into the LOF models and how. For every candidate predictor we maintain the variable's name, plain-language definition, unit and scaling, primary data source, spatial/temporal resolution, expected effect on LOF (direction or shape), a 1–5 reliability tag, material applicability, and notes on proxies or "can't-observe" fallbacks. All predictors are stored twice, in raw units and in a normalized/fuzzy form aligned to the model's 0–5 output index (so guardrails and explainability rules can be applied consistently).

4.4.1 Spatial Resolution

At the local scale, the pipe segment remains the atomic unit where each segment is uniquely referenced by its upstream–downstream node pair and scored on the 0–5 LOF index with reliability weighting from the data-quality scheme. That segment record is time-stamped and versioned so trend analyses respect feasible deterioration rates and directionality. We then aggregate segments into neighborhood "projects" (the contiguous set you would deliver as one job, such as a block or corridor). Project-level LOF is computed as length-weighted statistics or clusters on GIS using means and percentiles (e.g., 90th) of member segments for metrics like LOF and failure rates. These project summaries support packaging and sequencing and are specially useful in the renewal prioritization model (Chapter 6).

At the network scale (the full utility), LOF reporting can be standardized around a small set of coarse but stable metrics so leadership can track system health and progress over time. These metrics include a length-weighted mean LOF, the distribution across the five output classes, expected failures per 100 mile-years, and the Renewal Shortfall which is calculated as the annual footage in Poor/Bad minus the planned renewal footage.

Ideally, this can be further scaled up at the sector-benchmarking level enabling cross-utility comparisons. We can retain the same 0–5 LOF definition, publish per-mile and per-connection indicators, and include simple adjustments for diameter, material share, soil aggressivity, and climate zone so that a coastal PCCP-heavy utility is not judged by the same baseline as an inland PVC-dominant system. These comparisons can be framed as descriptive benchmarks rather than rankings. They can be useful for learning and policy level target-setting, even though this scale level is beyond the scope of this study.

Geospatial referencing for LOF: To integrate soils, climate, and break history without double-counting, we use a five-level link-node framework.

Level 1 assigns unique directional node-based IDs to pipeline segments, preserving flow directionality for hydraulic modeling, failure impact assessment and future graph theory based applications for better assessment of pipe criticality and hydraulics. Level 2 removes directionality, assigning unique node IDs at intersections and breakpoints, which is useful for material-based analysis, historical assessments, and general network topology evaluations. For enhanced spatial accuracy, Level 3 incorporates geographic coordinates along with elevation data, enabling pressure zone delineation and failure risk estimation using

digital elevation models (DEMs). Level 4 simplifies this by using only latitude and longitude, making it easier to overlay pipelines with external datasets such as soil conditions, land use, and climate zones. Finally, Level 5 provides a basic route-street reference if no other information is available, allowing pipelines to be identified based on their proximity to roadways, which is particularly useful for field inspections, construction planning, and utility coordination. This follows the standard method for georeferencing water pipe nodes following the FHWA's link-node and route-street referencing techniques (see Figure 4-2).

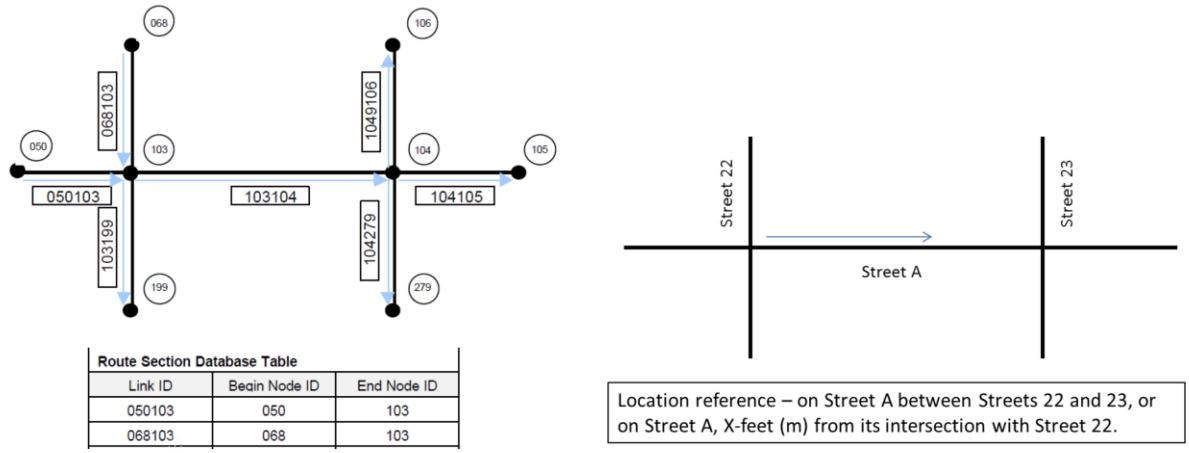

Figure 4-2: Geospatial referencing techniques (FHWA 2001). Figure on the left illustrates the directional link-node technique (Level 1) and the figure on the right illustrates the route-street referencing technique (Level 5) as applied on US highways geospatial data

4.4.2 Temporal Resolution

We anchor the analysis to discrete "analysis years" (calendar or the utility's fiscal year; we adopt the utility's convention by default) and build every feature as-of a cutoff date to avoid look-ahead bias. Renewal interventions and work-order signals like breaks, leaks, emergency repairs, rehabilitation or replacement are compiled in annual windows, and environmental layers that change slowly (soil indices, groundwater regime, corrosion potential, climate normals) are also summarized annually. Where higher-frequency observations exist (e.g., seasonal freeze–thaw or monthly pressure logs), we aggregate them into annual statistics, so the yearly frame remains consistent without discarding the higher resolution data. At the pipe-segment level, each record carries exact date stamps for installation, renewals, inspections, and failures.

4.4.3 Data Assumptions and Reliability Levels

Each pipe node sample needs to be judged for uncertainties to ensure it is usable for modeling and reliable for testing hypotheses in the validation experiments. All data used in this research is categorized into 5 levels as shown in Table 4-2. The basic level data with lowest reliability is based on the professional experience and intuition of Subject

Matter Experts (SMEs) such as a utility asset manager's assessment of bedding quality without inspection evidence. In the middle range, Lab Test Data consists of components like hydrants, pumps or scaled-down versions of water pipeline systems in a lab setting, such as lab testing of a water pump unit or simulating a small-scale water distribution network to study flow rates and pressure conditions. At the highest reliability level, Operational Real-World Data provides the most reliable and representative data from actual system operations using reliable measurement technologies. Examples of such data include real time water quality or flow monitoring data, remaining wall thickness, renewal activity in an intervention, visual condition assessment including forensic data collected after pipeline failure like number of wire breaks (for PCCP), failure mode, etc.

Table 4-2: Data Reliability Score to Quantify Uncertainties

Reliability Level	Category	Definition
1	**Educated Guess**: Domain Knowledge	Data and insights provided by SMEs based on their experience, intuition, and knowledge.
2	**Derived**: Simulated/ Theoretical/ Assumption	Data derived from fundamental scientific principles and physical laws, combined with advanced computational models and simulations to predict system behavior under various scenarios.
3	**Direct Measurement**: Lab Scale	Data collected from testing components or scaled-down versions of pipeline systems in a controlled lab setting.
4	**Direct Measurement:** Pilot Scale	Data obtained from testing the system in a live but controlled environment, under actual operating conditions.
5	**Direct Measurement:** System Scale	Data collected from the system's operation in the real world without any controlled testing environment, providing the most reliable and representative data.

When a preferred measurement is unavailable, we keep LOF estimable by following an explicit, ordered proxy path and tagging each substitute with a reliability score (1–5) and full provenance. For example, for remaining wall thickness (parameter for metallic pipe models), we first seek direct ultrasonic or other NDE readings (5); if absent, we use coupon tests or CCTV-based inference (4); if those are missing, we revert to a proxy corrosion composite built from pit depth, soil corrosivity class, and age (2). For water quality (Langelier Index), we compute it from in-situ pH, alkalinity, and hardness (5); otherwise, we use utility water quality reports (4); failing that, we apply regional norms (2). For pressure, we prefer minute-level SCADA at the nearest node (5); if unavailable, we take a calibrated hydraulic model output (2); as a last resort, we estimate from elevation head and zone rules of thumb (1–2). For traffic loading, we use agency axle counts where available (5), else infer from functional road class or Average Daily Traffic (ADT) from the respective Department of Transportation (DOT) within a 20 ft GIS buffer on neighboring pipes (2). For groundwater depth, we first pull local monitoring well records (5), then fall back to national level National Resources Conservation Service (NRCS) Soil Survey Geographic Database (SSURGO) for each watershed in the US. (2). Every proxy choice is recorded alongside its reliability and source.

4.4.4 Data Sources for LOF Model Inputs

The LOF models can use any of the following types of sources of data:

Geospatial: The GIS inventory provides network topology, material, diameter, installation year (vintage), and where available, depth and lining or coating. A version-controlled, node-to-node network with unique segment identifiers and attributes reaches level-5 reliability. Common gaps include unknown material, diameter, installation or protection information, long centerlines spanning appurtenances (single GIS pipe segment running straight through fittings like valves, tees, crosses, reducers, or pressure-zone boundaries instead of being split at those points), and occasional geometry errors. When material is missing, inference uses spatially adjacent segments within the same diameter category and build era, cross-checked against local adoption timelines. Diameter falls back to the nearest confirmed value within the same pressure zone. Installation years are imputed from parcel or subdivision build years or the first service connection date. These inferences are tagged level-2 to level-3 and recorded with method codes. Long lines are split at tees and valves when appurtenance layers or as-builts permit; if not, the line is retained with a cautionary tag and down-weighted in analysis.

Work Order: Work orders and break logs supply empirical failure history and failure modes. High reliability is achieved when events are geocoded directly to segments within a small snapping tolerance (spatial join with closest within technique) or if the pipe identifier is available to merge using those unique values. Any duplicates are merged, mode labels are normalized to a controlled list, and timestamps are resolved to the day. Address-only records are street-matched and then snapped to the network (level-3 to level-4). Ambiguous free text is mapped to the closest mode, with "unspecified structural" used as a last resort.

Hydraulics: Hydraulic and operational signals quantify static and minimum pressure, variability, and exposure to transients. Where SCADA (Supervisory Control and Data Acquisition) time series exist, pressures are summarized to planning windows using robust percentiles (for example, 5th, 50th, and 95th monthly and annually) and mapped to segments by zone and the nearest hydraulic node (level-5 if sensor coverage is dense and recent). If sensors are sparse, calibrated model node pressures are transferred by nearest-node or zone membership (level-3 to level-4 depending on calibration recency and error). In the absence of both, static pressure is estimated from elevation head and typical zone set-points, with night minima bounded by local demand patterns (level-2). Without

transient loggers, proximity to pumps, pressure-reducing valves, and high-frequency valve operations is encoded as a proxy for surge risk (level-2).

Field Performance: Field and forensic measurements anchor condition with metrics confirming direct evidence like ultrasonic wall-loss and coupon corrosion rates, visual observations for coatings, acoustic wire-break counts for PCCP, and leak acoustics. Measurements tied to precise locations, methods, and footprints, and linked to segment identifiers within the past three to five years, achieve level-5 reliability. Corridor-scale inspections are length-weighted across traversed segments and tagged level-4; legacy records lacking coordinates are attached to material–diameter–vintage cohorts and used as priors rather than labels (level-3). Where no direct measures exist, membership functions for corrosion and lining condition are widened to reflect uncertainty rather than up-rating LOF based on anecdote.

Expert opinion: Structured expert workshops and interviews contribute plausibility of failure mechanisms, typical ranges, and local practices that explain residual patterns. These inputs are encoded as priors, expected directions and plausible ranges and as rule checks in the fuzzy "teacher" model, not as ground truth. By design they carry level-1 to level-2 tags and act to constrain or sanity-check, not to drive labels.

External Open Source: External open source datasets from reliable federally sponsored efforts are used to enrich the environmental and loading context. Soils characteristics and geology come from SSURGO (Soil Survey Geographic Database). We use the gridded SSURGO rasters (gSSURGO/gNATSGO-derived) at 30 m cell size, updated on an annual refresh cycle, with statewide tiles covering the 48 contiguous states, Hawaii, and portions of Alaska. Segment attributes are assigned from rasters by sampling along the pipe line segment. Raster values are sampled at the segments via nearest-cell lookup using a spatial join, and the segment value is computed as a length-weighted statistic (majority class for categorical themes; mean/median for continuous properties like pH or drainage index). Short segments (< 30 m) default to the midpoint cell**;** where the segment crosses multiple cells, the aggregation reflects the fraction of the segment within each cell (captured by the vertex spacing). Given SSURGO's authority and annual refresh, we tag these soil attributes as reliability level 5. In rare cases of known trench fill, we additionally set an "unknown/fill" flag but retain the level-5 tag for the underlying source. In internal-corrosion contexts, water chemistry (for example, Langelier index) may be used as a distinct proxy. Transportation loading is represented by AADT (Average Annual Daily Traffic) joined from road centerlines to segments within a defined buffer of 20 meters. Wherever AADT is missing on local streets, functional road class serves as an ordinal proxy.

Critical facilities and land use are joined using distance buffers to capture service disruption likelihood near hospitals, schools, and industrial facilities. Parcel-level classes are preferred over block-level classes when available. Environmental data enrichment is based on datasets like FEMA floodplains, National Wetland Inventory (NWI) shapefiles, slopes from digital elevation models, and seismic zones (especially for utilities on the west coast) using 30 m buffers on pipes.

4.4.5 Data Dictionaries

Data dictionaries are prepared for all *teacher* models summarizing all the data specifications. An example for the data dictionary prepared for metallic <16" model is shown in Table 4-3. Dictionaries for all teacher models can be found in Appendix C. Each dictionary row declares the hypothesized sign or shape and we test it during training. We apply selective monotonic directions where physics is unambiguous. Variables like wall-loss, wire-break count and growth, ovality, normalized headloss, leak rate, and corrosivity increase risk monotonically. Variables with U-shaped or conditional effects like operating pressure are modeled with capped or U-shaped mappings rather than forced monotonic trends. This avoids perverse responses while preserving scientific consistency.

Table 4-3: Data dictionary for metallic <16" teacher fuzzy model

Predictor	Definition (segment-level)	Unit / Scale	Source & resolution	Effect on LOF	Rel.	Materials	Proxy / fallback
Pipe age	Years from install to observation year	years (raw) + 0–5 (fuzzy)	Asset register; annual	↑ older ⇒ higher LOF (monotone +)	5	CI, DI, Steel	If missing: service start date from billing (2)
Pipe vintage	Era capturing material/standard changes	categorical → 0–3 bins	Asset register + standard history; static	Depends on era (e.g., pre-1960 CI ↑)	4–5	CI, DI	If unknown: infer from nearby segments (1–2)
Internal lining	Presence/condition of lining	boolean / condition	Rehab/inspection records; per job	Lined ↓ LOF; failed lining ↑	4–5	CI, DI, Steel	CCTV notes (4) → SME flag (1)
External protection	Coating/wrap/cp status	0–3 condition	As-built + CP logs; annual	Better protection ↓ LOF	3–5	CI, DI, Steel	Soil resistivity + CP stations nearby (2)
C-factor	Hydraulics roughness indicator	coeff. (raw) + 0–4 bins	Model calibration or test; annual	Lower C (rough) ↑ LOF	2–4	All metallic	Trend from headloss tests (3)
Remaining wall thickness (RWT)	% of nominal wall	% (raw) + 0–5 bins	UT/NDE; ad hoc per inspection	Lower RWT ↑ LOF (strong +)	4–5	CI, DI, Steel	Pit depth + age + soil class (2)
Pit depth	Max pit depth on interior/exterior	mm (raw) + 0–4 bins	Coupons/CCTV; per job	Deeper pits ↑ LOF	4	CI, DI, Steel	Soil corrosivity index (2)
Operating pressure	Typical zone pressure; surge exposure	psi + 0–5 bins	SCADA minute-level; model	U-shaped: very low & very high ↑	2–5	All	Elevation head if no sensors (1–2)
Soil corrosivity	Composite score (resistivity, chlorides, sulfates, moisture)	index 0–4	Geotech logs / SSURGO; 10–30 m	Higher corrosivity ↑ LOF	2–5	CI, DI, Steel	Regional soil class (2)
Bedding condition	Support class at install / inspection	0–2 condition	As-built + photos; one-time	Poor bedding ↑ LOF	1–4	All	Road class + depth as risk proxy (1–2)

Guardrails enforce feasible deterioration rates, range checks, and cross-field coherence (e.g., extreme tuberculation cannot co-exist with "perfect" C-factor). If the learned partial dependence between input features and output contradicts the dictionary and physics (e.g., "older pipes look safer"), we first re-check data lineage and reliability, then constrain or re-express the feature (e.g., winsorize outliers, re-bin vintage). In practice, controlled winsorization caps extreme values at chosen percentiles so outliers cannot

dominate model fitting. Values below a lower cutoff (e.g., 1st–2.5th percentile) are set to that cutoff and values above an upper cutoff (e.g., 97.5th–99th) are set to that cutoff, preserving all records while stabilizing means, variances, and partial-dependence estimates. Caps are computed within coherent cohorts (same material, diameter band, pressure zone, and when drifts are suspected within a rolling time window) so thresholds remain physically meaningful. Critically, ground-truth structural features that define risk like verified remaining wall thickness, wire-break counts, and extreme headloss are not winsorized to avoid masking the very conditions linked to failure. This practice keeps the LOF model mechanism-aware rather than purely correlative.

Provenance, versioning, and implement ability: The dictionary is under version control and every change to ranges, bins, or proxies is logged with a date, rationale, and affected materials. Each feature value carries provenance (source, timestamp, method) and the reliability tag, so downstream LOF outputs can be filtered (e.g., we only report bands where ≥80% of inputs are Level 4/5). The same schema will be reused in the COF and Risk chapters, with feature sets specific to those targets.

4.5 Descriptive Analytics and Failure Baselines

This section draws on the Sinha (2021) national compilation of water pipeline field performance data. Inventory and work-order datasets from more than 500 U.S. water utilities of varied size and governance helped support this initiative. Heterogeneous inputs were standardized through repeated checks with each utility, producing consistent fields for materials, diameters, installation year, and intervention history. This breadth limits local bias and lets us treat recurring patterns as defensible priors rather than anecdotes. We use these priors to initialize the fuzzy antecedents (which drivers to include, expected directions) and to weight rule motifs by the strength and stability of empirical gradients. We do not treat the national figures as ground truth for any single utility; they inform starting points that are then adapted with that utility's own data.

4.5.1 Baselines by Material, Diameter and Ecological Cohorts

To normalize varying failure definitions and counts from utilities, we summarize outcomes as intervention rates per 100 mile-years, stratified by material and diameter category as shown in Figure 4-3. Intervention rates in further relevant subcategories like age group, and ecological cohorts can be found in Sinha (2021). The baselines reproduce

known gradients like smaller diameters in cast iron and asbestos cement show higher rates; ductile iron and steel trend lower; large-diameter concrete families (PCCP, RCP) remain low. Rates increase with age cohorts, and ecological partitions shift levels in mechanism-plausible ways (e.g., poorly drained soils and high frost action raise metallic-pipe interventions). These baselines set quantitative anchors for fuzzy membership thresholds (e.g., what constitutes "high stress" or "elevated LOF" for a given material-cohort) and help tune consequents (e.g., expected LOF band under a driver combination).

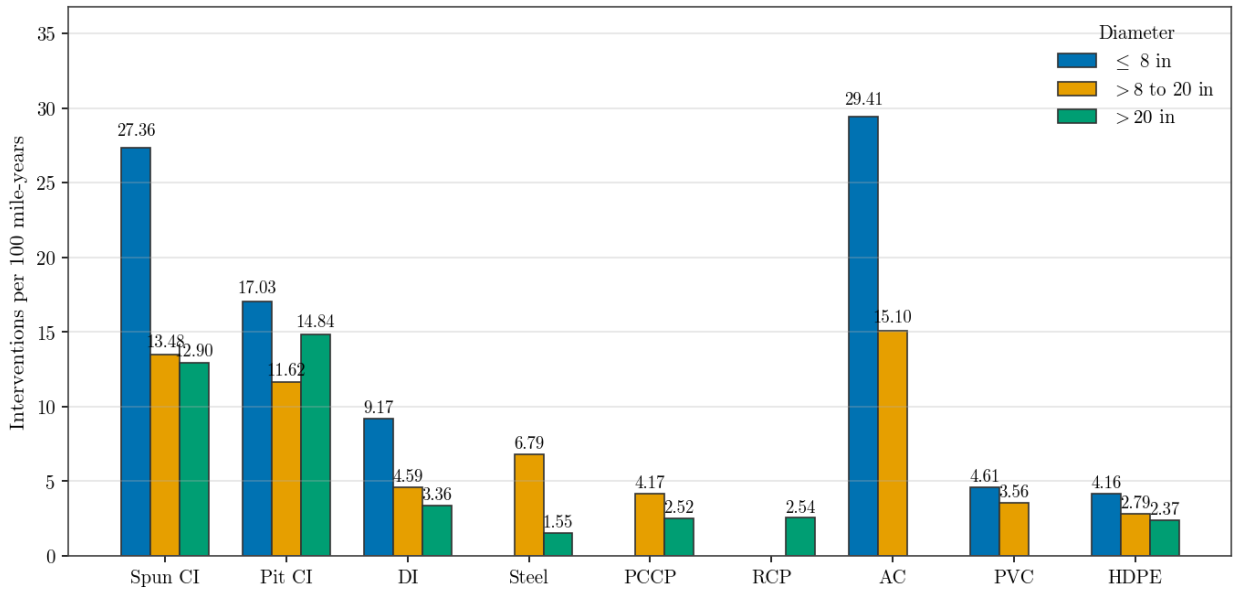

Figure 4-3: Baseline intervention rates by material and diameter classes (Sinha 2021)

4.5.2 Mechanism-linked Drivers with Material-specific Directionality

Correlation and stratified analyses in Sinha (2021) identify drivers that align with known mechanisms and vary by material. The correlation map is shown in Figure 4-4. External corrosion indicators (soil resistivity, pH, redox) dominate metallic outcomes; frost-season months elevate intervention frequencies in cold regions; shallow bedrock and high traffic loading increase interventions in plastics; higher operating pressure and pressure transients raise PVC rates. Limited but direct condition data (remaining wall thickness, pit depth, graphitization) point in the same directions for metallics. We translate these into monotone expectations and pairwise synergies in the fuzzy rule base. For example, for metallics, the rule weight increases when corrosivity is high and traffic loading is high; for PVC, elevated minimum pressure combined with shallow cover or high traffic raises the consequent LOF band.

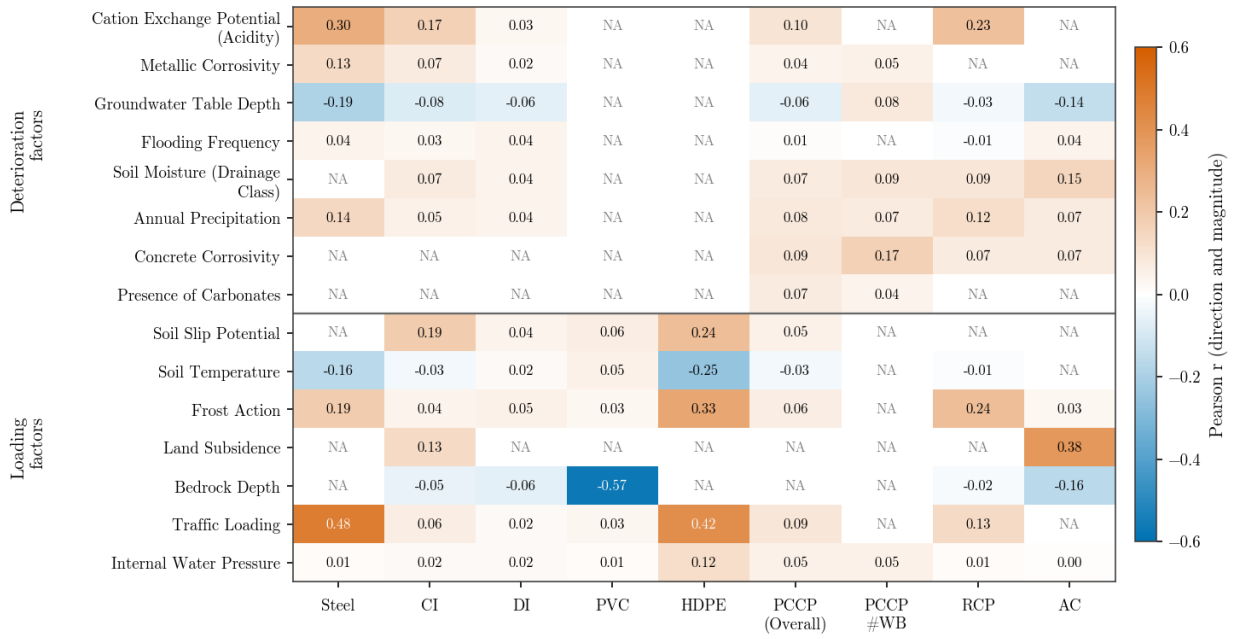

Figure 4-4: Mechanism map: driver-outcome correlations by material categories. Cohen bins (|r|): strong ≥ 0.5, medium 0.3–0.5, weak 0.1–0.3, negligible < 0.1. NA = insufficient or inconsistent data (Sinha 2021)

4.5.3 Failure Modes and Causes as Fuzzy Rule Motifs

Observed failure mode and cause profiles differ by material. Corrosion is prevalent in metallics; PVC is dominated by longitudinal splits and third-party or installation-related causes; PCCP involves wire-break mechanisms and manufacturing effects. The descriptive statistics for failure modes and causes for each material family can be understood from Figure 4-5. These patterns determine which antecedents are salient for each

material family and which combinations are less informative (e.g., soil corrosivity is central for metallics but not for PVC; operating pressure and construction quality matter more for PVC/HDPE). We encode these distinctions as material-specific rule motifs, so the teacher model remains mechanistically faithful.

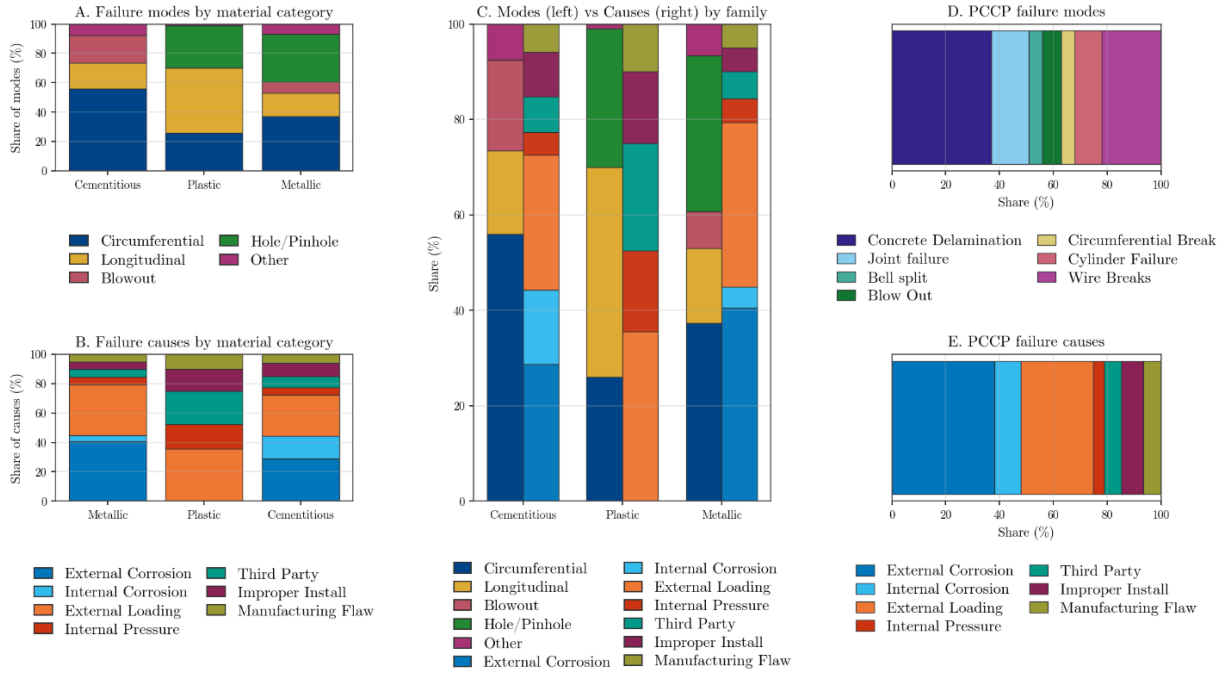

Figure 4-5: Failure modes and causes for different pipe material families (Sinha 2021)

Attribution and limits: Where the national dataset is sparse or heterogeneous, we use structured expert input based on interviews with asset managers and engineers from utilities across the US representing various operational and management scenarios

to set initial ranges and breakpoints. Expert input acts as a prior and empirical baselines and utility-specific data dominate when available.

4.6 Knowledge-structured "teacher" model (fuzzy inference)

We formalized six "teacher" models, each aligned to a mechanism-coherent material/diameter family: metallic <16″ (CI, DI, GI, steel), metallic ≥16″ (CI, DI, steel), plastic (PVC, HDPE), PCCP, cement/concrete (BWP/RCP/RCCP), and asbestos-cement. The split captures distinct deterioration physics: GI is predominantly small-diameter; steel is mostly large-diameter; PCCP warrants its own model because of the wire-cylinder-mortar composite; plastics share similar mechanisms across sizes; AC has unique leaching and embrittlement behavior. The lifecycle performance characteristics, failure modes, material standards, and vintage-specific effects for each pipe type are detailed in Sinha (2021).

4.6.1 Membership Functions and Input Space

Each model maps observable inputs into fuzzy sets using overlapping membership functions (MFs). The membership of parameter values in crisp sets can be defined based

on a characteristic function expresses the degree to which a crisp value x belongs to a linguistic class.

$$\mu_s(x) = \begin{cases} 1 \; if \; X \in S \\ 0 \; if \; X \notin S \end{cases}$$

Commonly used membership functions to represent input and output parameters are shown in Figure 4-6. In the LOF teacher model, the membership-function family should mirror the mechanism and the decision granularity. Triangular and trapezoidal sets are piecewise-linear and highly interpretable, ideal for coarse, rule-auditable concepts (e.g., vintage bands, protection condition) and for fast elicitation from experts. Gaussian and generalized bell sets are smooth and differentiable, better for sensor-like quantities with noise and gradual transitions (e.g., C-factor about a nominal value, moderate headloss), and for overlapping terms where continuity matters; the bell's adjustable width/slope helps tune overlap without sharp corners. Sigmoid (S-shaped) sets encode monotonic on-sets which is useful for threshold processes (e.g., risk increasing with operating pressure or corrosivity), while the Z-shaped complement captures monotonic decay (e.g., membership in "low risk" shrinking as stressors rise). In practice, use triangles/trapezoids when transparency and easy calibration dominate; use Gaussians/bells when physical signals

are smooth and uncertainty is gradual; use sigmoid/Z when domain knowledge implies a one-direction response.

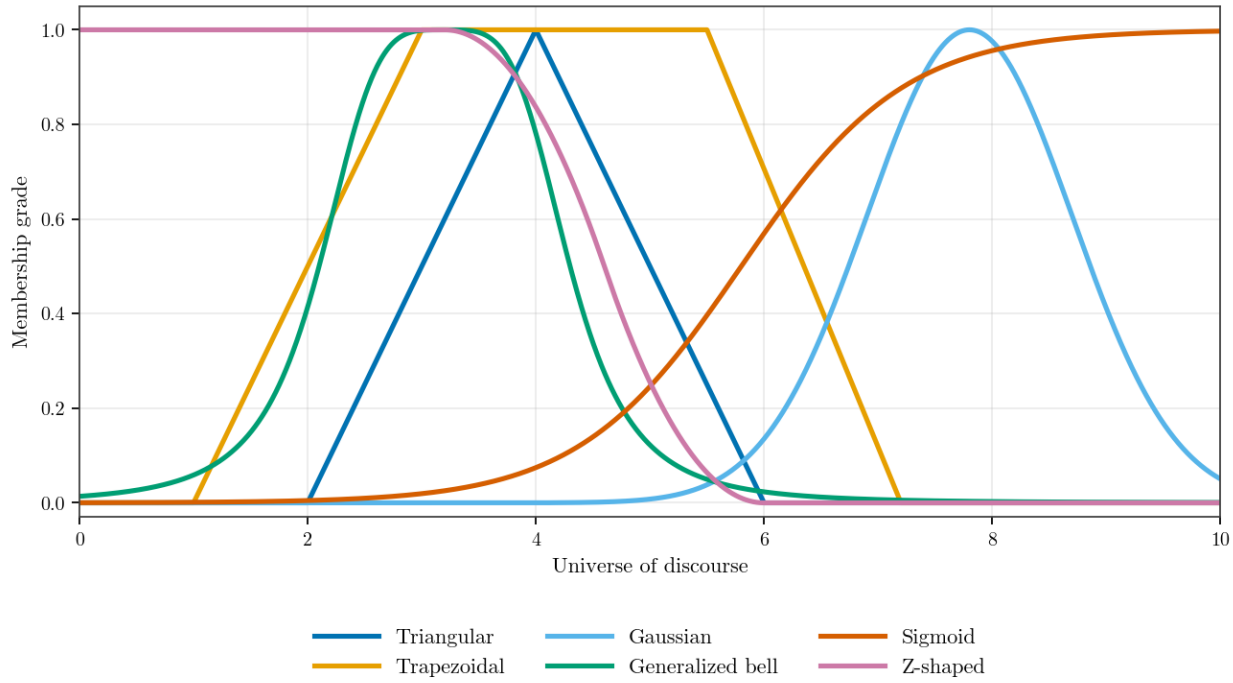

Figure 4-6: Common fuzzy membership functions to represent inputs and output parameters

We use triangular, trapezoidal, and Gaussian shapes, tuned so that (i) class centers align with measurable anchors (e.g., remaining-wall-thickness, C-factor, Langelier Index), (ii) overlaps are wide enough to allow graded transitions, and (iii) monotone relations are preserved for mechanism-defining variables (e.g., lower RWT must not imply lower LOF).

Inputs are grouped as internal (water chemistry, temperature, lining/coating), external (soil class, groundwater and frost action, traffic/pressure loads, protection), and inherent condition (age, C-factor, historical breaks, RWT). Across the six teachers we use ~125 inputs. The input parameters and specifications for metallic <16″ fuzzy inference expert system is shown in Table 4-4. The tables for the other 5 fuzzy inference expert systems are shown in Appendix D.

Table 4-4: Input parameters for "teacher" fuzzy inference expert system for metallic <16"

Parameter	Membership Function Labels	Range	Typical Unit	Input Data Preparation
Pipe Age	New (0-0.27), Used (0.27-1.07), Old (1.07-1.6), Obsolete (1.6-2.67)	0-4	Years	Original range: 0-300 years. Scaled down: New (0-20 years), Used (20-80 years), Old (80-120 years), Obsolete (>120 years).
Pipe Vintage	Best Vintage (0-0.5), Relatively Good (0.5-1.5), Relatively Fair (1.5-3)	0-3	Years	Original range: 1900-2025. Scaled down: Cast Iron Pipes (Best Vintage (1900-1930), Relatively Good (1930-1960), Relatively Fair (1960-2025));
Pipe Internal Lining	Yes (0-1), No (1-2)	0-2	Boolean	Same for all metallic materials
Pipe External Protection	Good (2-3), Fair (1-2), Poor (0-1)	0-3	Condition Score	Same for all metallic materials
Water Quality	Good (0.3-2), Fair (-0.3 to 0.3), Poor (-2 to -0.3)	-2 to 2	Langelier Index	Original range: -2 to 2. Scaled down: Good (0.3-2), Fair (-0.3 to 0.3), Poor (-2 to -0.3)
Pipe Breaks <16"	None (0-0.25), Low (0.25-0.375), Moderate (0.375-0.75), High (0.75-1.25), Very High (>1.25)	0-5	Count	Original range: 0-20 breaks. Scaled down: None (0-1), Low (1-2), Moderate (2-4), High (4-5), Very High (>5).
Pipe C-Factor	Rough (1.25-1.75), Fair (2.0-2.5), Smooth (2.5-3.75), Perfect (3.75-4)	0-4	Coefficient	Original range: 0-160. Scaled down: Rough (50-70), Fair (70-110), Smooth (110-130), Perfect (130-160).
Pipe Remaining Wall Thickness (RWT)	Excellent (4.95-5), Good (4.5-4.95), Fair (4.0-4.5), Poor (3.0-4.0), Critical (0-3.0)	0-5	%	Original range: 0-100 mm. Scaled down: Excellent (99-100), Good (90-99), Fair (85-90), Poor (60-85), Critical (<60).
Pipe Pit Depth	Low (0-0.3), Moderate (0.6-1.2), High (1.2-2.4), Critical (2.4-4)	0-4	%	Original range: 0-100 mm. Scaled down: Low (0-7.5), Moderate (7.5-22.5), High (22.5-45), Critical (45-100).
Pipe Pressure	Very Poor (0-1, 3.75-5), Poor (1-1.75, 3-3.75), Fair (1.75-2.25), Good (2.25-2.75), Excellent (2.75-3.5)	0-5	psi	Original range: 0-200 psi. Scaled down: Very Poor (0-40, 150-200), Poor (40-70, 135-150), Fair (70-90), Good (90-110), Excellent (110-150).
Soil Particle Size	Coarse (0-1), Medium (1-1.5), Fine (1.5-2)	0-2	Size Category	Same for all metallic materials

Pipe Depth	Very Shallow (0-0.33), Shallow (0.33-0.67), Moderate (0.67-1.5), Deep (1.5-2.5), Very Deep (2.5-5)	0-5	Feet	Original range: 0-30 ft. Scaled down: Very Shallow (0-2 ft), Shallow (2-4 ft), Moderate (4-10 ft), Deep (10-20 ft), Very Deep (20-30 ft).
Frost Action	Low (2-3), Medium (1-2), High (0-1)	0-3	Index	Original range: 0-3. Scaled down: Low (2-3), Medium (1-2), High (0-1).
Traffic Loading	Low (2-3), Medium (1-2), High (0-1)	0-3	Index	Original range: 0-3. Scaled down: Low (2-3), Medium (1-2), High (0-1).
Water Table Depth	Deep (1.5-3), Moderately Deep (0.75-1.5), Shallow (0-0.75)	0-3	Feet	Original range: 0-20 ft. Scaled down: Deep (>10 ft), Moderately Deep (5-15 ft), Shallow (0-10 ft).
Pipe Cracks	No Cracks (0-1), Hairline Crack (1-2), Minor Crack (2-3), Major Crack (3-4), Fracture Burst (4-5)	0-5	Severity Level	Original range: 0-5. Scaled down: No Cracks (0-1), Hairline Crack (1-2), Minor Crack (2-3), Major Crack (3-4), Fracture Burst (4-5).
Soil Corrosivity	Critical (0-0.4), High (0.4-1), Medium (1-2), Low (2-4)	0-4	Index	Original range: 0-10000. Scaled down: Critical (0-1000), High (1000-2500), Medium (2500-5000), Low (5000-10000).
Stray Currents	Absent (0-1), Present (1-2)	0-2	Presence	Original range: 0-2. Scaled down: Absent (0-1), Present (1-2).
Pipe Joints	Poor (0-0.7), Fair (0.7-1.3), Good (1.3-2)	0-2	Condition Score	Original range: 0-2. Scaled down: Poor (0-0.7), Fair (0.7-1.3), Good (1.3-2).
Pipe Valves	Poor (0-0.7), Fair (0.7-1.3), Good (1.3-2)	0-2	Condition Score	Original range: 0-2. Scaled down: Poor (0-0.7), Fair (0.7-1.3), Good (1.3-2).
Pipe Graphitization	Low (0-0.4), Moderate (0.8-2), High (2-3.4), Severe (3.4-4)	0-4	Condition Score	Original range: 0-100. Scaled down: Low (0-10), Moderate (10-35), High (35-65), Severe (65-100).
Groundwater Fluctuation	Low (0-0.45), Moderate (0.45-0.9), High (0.9-3)	0-3	Feet	Original range: 0-20 ft. Scaled down: Low (0-3 ft), Moderate (3-7.5 ft), High (7.5-20 ft).
Temperature	Very Cold (0-1), Cold (1-2), Moderate (2-3), Warm (3-4), Very Hot (4-5)	0-5	°F	Original range: -20 to 150 °F. Scaled down: Very Cold (<32°F), Cold (32-57°F), Moderate (57-82°F), Warm (82-107°F), Very Hot (>107°F).
Precipitation	Low (0-1), Average (1-2), High (2-3), Critical (3-4)	0-4	Inches/Year	Original range: 0-100 inches/year. Scaled down: Low (0-10 in/yr), Average (10-30 in/yr), High (30-60 in/yr), Critical (60-100 in/yr).
Internal Water Temperature	Hot (3.5-5), Warm (3-3.5), Moderate (2-3), Cool (1-2), Cold (0-1)	0-5	°F	Original range: 20-100 °F. Scaled down: Hot (>70°F), Warm (55-70°F), Moderate (45-55°F), Cool (35-45°F), Cold (<35°F).
Bedding Condition	Poor (0-0.5), Fair (0.5-1.5), Good (1.5-2)	0-2	Condition Score	Original range: 0-2. Scaled down: Poor (0-0.5), Fair (0.5-1.5), Good (1.5-2).
LOF Index	**Very High, High, Moderate, Low, Very Low**	**0-5**	**Index**	**Original range: 0-5. Scaled down: Excellent (0-1.5), Good (1.5-2.5), Fair (2.5-3.5), Poor (3.5-4.5), Bad (4.5-5).**

4.6.2 Rule-base and IF–THEN Mechanics

Each teacher encodes expert knowledge as Mamdani-style IF–THEN rules that tie driver motifs to LOF bands on a 0–5 scale. Rules are compact and only variables relevant to a motif appear in the antecedent. Boolean operators use a standard *t-norm* for AND (product) and *s-norm* for OR (bounded sum). Example motifs:

- External corrosion: IF {Soil corrosivity = High} AND {Protection = Poor} AND {RWT = Low} THEN {LOF = Very High}.
- Hydraulic stress in plastics: IF {Operating pressure = High} OR {Traffic = High AND Depth = Shallow} THEN {LOF = Poor→ Very High}.
- Load-induced concrete distress: IF {Traffic = High} AND {Bedrock depth = Shallow} AND {Frost action = High} THEN {LOF = High}.

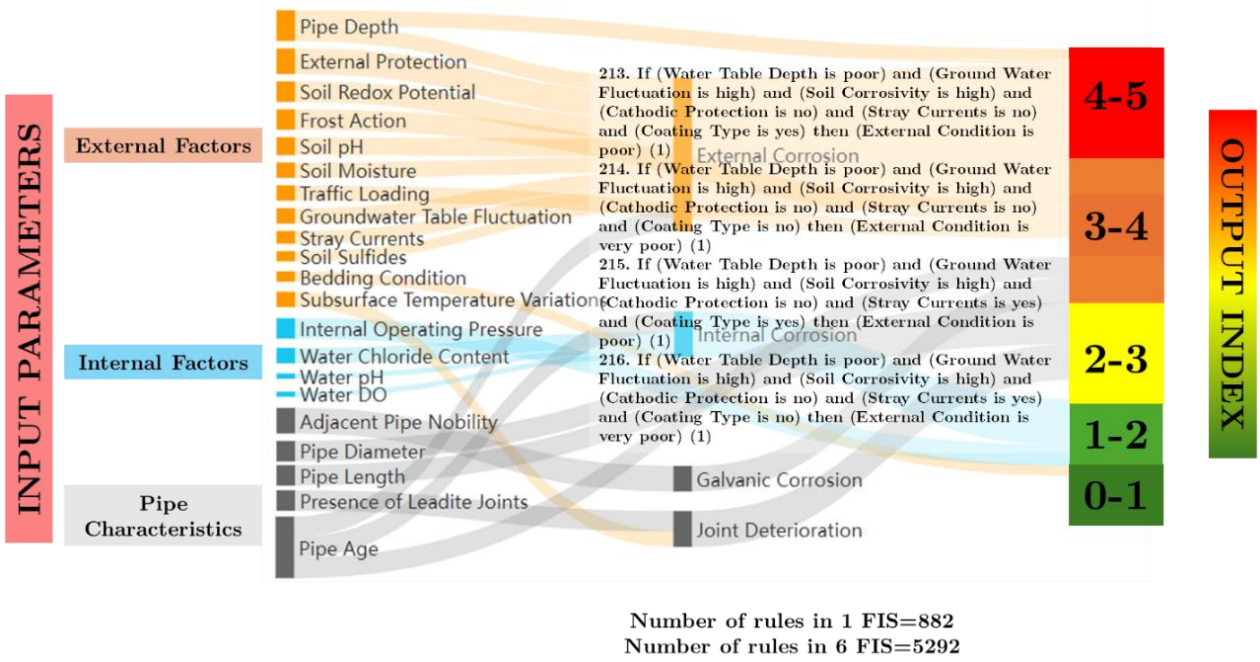

Figure 4-7: Illustration to show the workings of "teacher" fuzzy inference system from inputs → rules → LOF (0–5)

To control combinatorial growth in rules, we template rules by mechanism and material family, then specialize only where evidence or practice requires it. Across all teachers we maintain 5,292 rules. Consequents use the same five named LOF bands (Very High, High, Moderate, Low, Very Low), defined on the 0–5 axis with gentle overlaps to keep interpolation smooth.

4.6.3 Inference, Interpolation, and Defuzzification

For a given segment, inputs are fuzzified into degrees μ_i. Rule *firing strength* w_r is the product of antecedent degrees. We truncate the consequent MF of each rule at w_r, aggregate all rule outputs by bounded summation (cap at 1), and compute a crisp LOF by the centroid of area. This procedure yields natural interpolation between nearby rules. When inputs straddle linguistic boundaries, multiple rules fire with fractional weights, and the centroid returns a stable middle value. However, there are many other methods that can be used. An illustration to show how different defuzzification rules can be used to get different crisp output values shown in Figure 4-8. Panels show five methods—centroid (COG), bisector (BOA), Mean of Maxima (MOM), Smallest of Maxima (SOM), and Largest of Maxima (LOM). A fixed reference line (LOF = 3.0) runs across all panels for visual alignment.

Each method yields a different crisp value due to distinct summarization rules (area-weighted vs. plateau-based), underscoring the importance of explicitly choosing and justifying the defuzzification operator in the teacher model. For the LOF teacher model, the choice of defuzzification operator should match the decision intent. COG provides an area-weighted "expected value" and is preferred for continuous LOF reporting because it

uses all available evidence and is smooth to small rule changes. BOA returns the point that splits the aggregated area in half and suits cases seeking a median-like summary that is less sensitive to long tails than COG. MOM focuses only on the peak region and works well for linguistic band assignments where the most plausible label should dominate. SOM and LOM resolve flat peaks with explicitly optimistic (leftmost) or conservative (rightmost) choices, useful when policy requires a bias toward cost savings or precaution, respectively. In practice, COG is used for the continuous LOF score, MOM for band selection, BOA as a tail-robust compromise, and SOM/LOM when tie-breaking must reflect a stated risk posture.

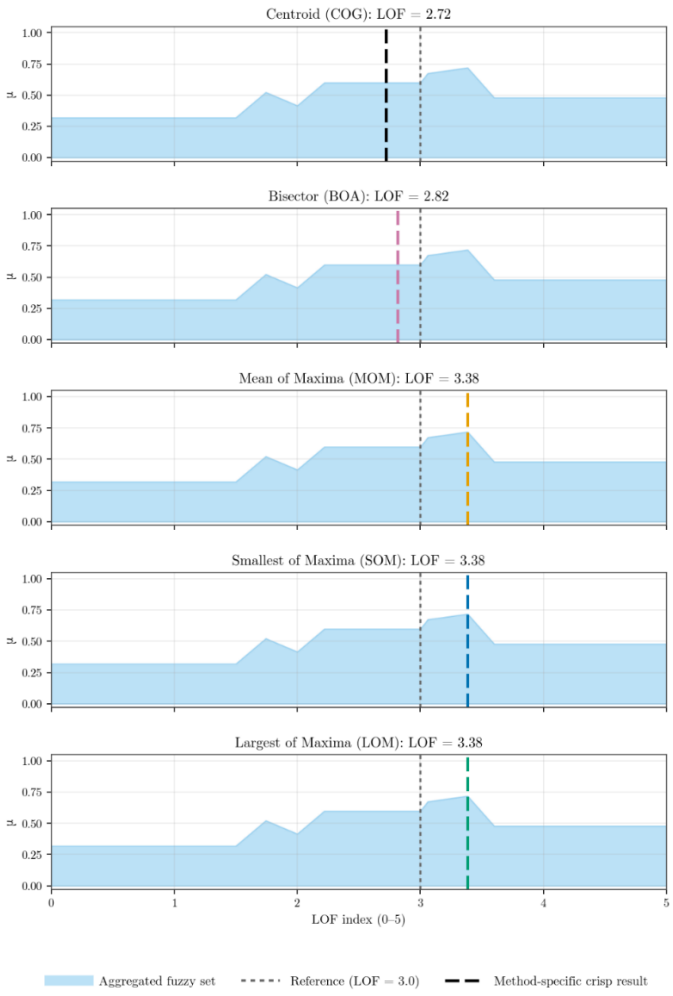

Figure 4-8: Illustration to show how different defuzzification methods can give unique crisp output values

4.7 Evaluation, Verification and Validation

This section presents the results from a robust testing process for the LOF teacher-student models. Evaluation covers development-time diagnostics and is performed by the

modeler, establishing that the model behaves sensibly before any formal tests. Verification then tests the student ML model's agreement across 19 material-diameter cohorts. Finally, Validation reports real-world performance of the verified student ML models by statistically testing agreement with incumbent utility indices, expert concordance, and one-year failure coverage by cohort. Together, the EVV layers move from internal checks to formal pass/fail gates to external realism, with all windows, strata, and code artifacts frozen for reproducibility.

4.7.1.1 Development of Verification and Validation Dataset

A machine-learning model's robustness and generalizability depend on the diversity and independence of the data used for both development and testing. For model development and verification, we use pipe-inventory and field-performance datasets from seven distribution utilities (A–G), which together span a wide range of materials, diameters, protection practices (internal linings, polyethylene wrap, cathodic protection), and failure mechanisms (corrosion, cracking, joint failures, wall loss). For validation, we draw on three utilities. One of the original seven (Utility E), which provided condition-assessment and work-order datasets that were held out from training, plus two additional utilities (H and I) that contributed large-diameter PCCP wire-break records and non-metallic (PVC,

PE, AC) failure histories, respectively. Collectively, these nine systems (as shown in Figure 4-9) cover frost-heavy, coastal, arid, and humid continental climates and a mix of small, medium, and large utilities, so the LOF models are challenged across diverse operational and environmental conditions rather than tuned to a single system.

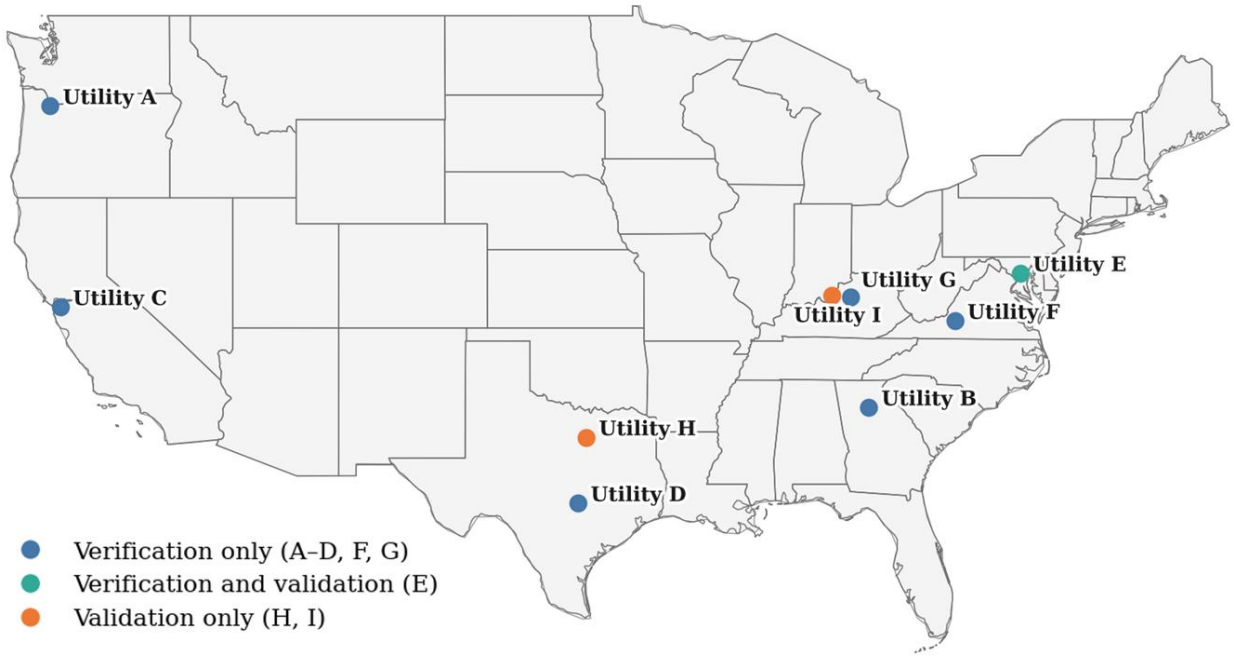

Figure 4-9: Participating utilities for LOF model verification and validation (anonymized A–I). Blue markers denote utilities that contributed data for model development and verification only (A–D, F, G); green markers denote the utility that contributed both development/verification and independent validation data (E); orange markers denote utilities that contributed independent validation datasets only (H, I).

The dataset distribution from the utilities is presented in Table 4-5. These datasets capture a wide range of geographical conditions, from frost-heavy regions to partially arid but high-precipitation areas, as well as coastal utilities with varying soil corrosivity.

Additionally, the inclusion of utilities of different sizes ensures a broad spectrum of management practices, funding constraints, expertise levels, and risk preferences. This comprehensive dataset ensures that the ML model can learn and generalize across diverse operational and environmental conditions, reducing biases and improving predictive accuracy in both common and high-consequence scenarios.

Table 4-5: Utility datasets for Model Verification and Validation

Features	**Utility A**	**Utility B**	**Utility C**	**Utility D**	**Utility E**	**Utility F**	**Utility G**	**Utility H**	**Utility I**
Mileage (mi)	2200	3900	4200	3680	5400	1200	350	180	4196
Dominant Pipe Materials (% of Total Mileage)	CI (40%), DI (33%), PVC (20%), RCP (5%), AC (2%)	CI (15%), DI (30%), PVC (40%), HDPE (5%), AC (10%)	CI (35%), DI (25%), PVC (25%), Steel (12%), AC (3%)	PVC (40%), DI (25%), CI (15%), Steel/AC (10%), PCCP (8%), HDPE (2%), Others (1%)	CI (40%), DI (30%), PVC (15%), HDPE (6%), PCCP (7%), AC (2%)	CI (45%), DI (20%), PVC (20%), RCP (8%), AC (7%)	CI (55%), DI (22%), PVC (15%), RCCP (5%), AC (3%)	PCCP (88%), ST (12%)	CI (36%), DI (27%), PVC (30%), RCCP (3%), HDPE (1%), AC (3%)
	CI = 7,693; DI = 6,841; PVC = 6,437; HDPE = 762; PCCP = 656; Steel = 792; AC = 1018; RCP = 246; RCCP = 163; PE = 449 Other (GALV/unclassified) = 250								
Pipe Diameters (% of Total Mileage)	<8" (84%), 8-24" (9%), >24" (7%)	<8" (79%), 8"-24" (13%), >24" (8%)	<8" (85%), 8-24" (9%), >24" (6%)	<8″ (78%), 8–24″ (17%), >24″ (5%)	<8" (82%), 8"-24" (11%), >24" (7%)	<8" (80%), 8-24" (14%), >24" (6%)	<8" (75%), 8"-24" (15%), >24 (10%)	>24% (100%)	<8" (72%), 8"-24" (25%), >24 (3%)
	<8"=20146 miles, 8"-24"=3913 miles, >24"=1247 miles								
Climate	Moderate frost, inland	Temperate with year long precipitation	Coastal	Hot-summer subtropical; drought–flood swings; expansive clays	Frost-heavy	Temperate, plenty precipitation	Midwestern, hurricane-prone	Hot subtropical; drought–flood swings; expansive clays	Humid; plenty precipitation; freeze–thaw
State	OR	GA	CA	TX	MD	VA	KY	TX	KY

Utility Management	Medium-Large, primarily reactive renewal	Large, primarily reactive renewal	Large, mix of proactive and reactive renewal	Large, mixed proactive/reactive renewal	Large, primarily reactive renewal	Small, reactive renewal	Small, funding constraints, reactive renewal	Large wholesaler with extensive condition assessment	Large, mature CMMS and reliable data collection

4.7.2 Evaluation of the Teacher model

The teacher FIS is evaluated by the modeler in three steps that collectively ask whether the rules are representative, stable and mechanism-consistent. The goal is not to "optimize" the teacher but to demonstrate that it encodes the intended mechanisms with defensible behavior over the space of conditions the model will face, before handing the structured input-output dataset for supervised learning of the student models.

4.7.2.1 Representativeness and Coverage

We first verify that input membership functions span realistic operating envelopes and that the rule base returns expected outputs at the boundaries. Membership-function panels (six exemplars are shown in Figure 4-10; all 6 "teacher" models need 125 inputs) are reviewed to confirm supports and overlaps reflect measurement ranges and ambiguity levels agreed with utility SMEs.

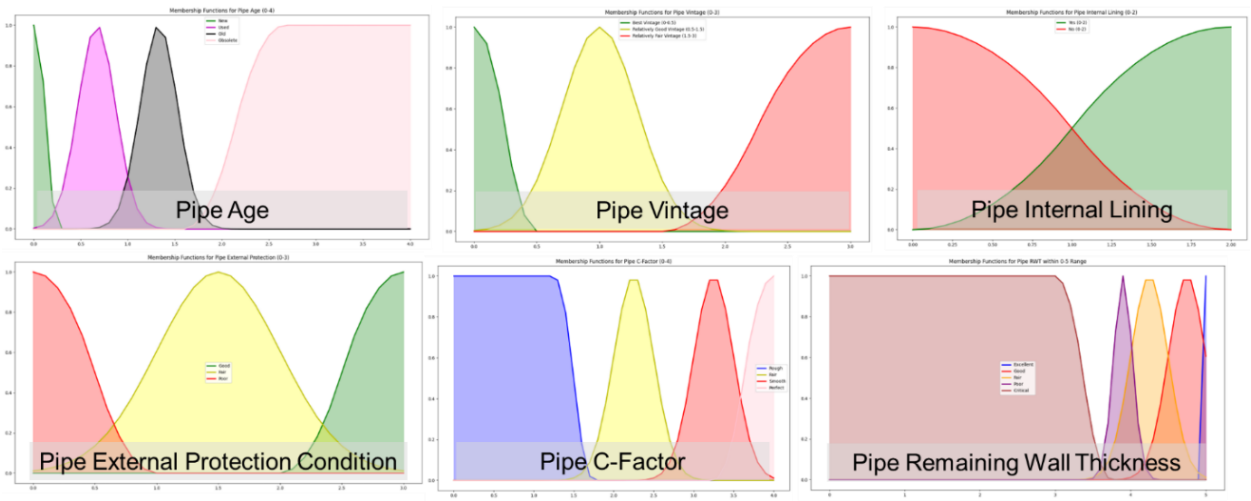

Figure 4-10: LOF Model Representativeness through simple visualizations of parameters. Here it is shown only for 6 parameters. This is performed for all 125 parameters across the 6 models.

"Coverage" is tested with canonical stress settings, all-best, all-average, and all-worst, which should yield LOF ≈ 5, ≈ 2.5, and ≈ 0, respectively, within a ± 0.2 tolerance (the tolerance reflects aggregation granularities and overlapping sets). Coverage tests confirm the model's ability to handle edge cases effectively. For instance, the LOF model should return a rating of 0, 5, and 2.5 when input parameters are set to best, worst, and average values, respectively. As an example, the result from the *Best* scenario is shown in Figure 4-11. The average and worst scenario results are in Appendix D.

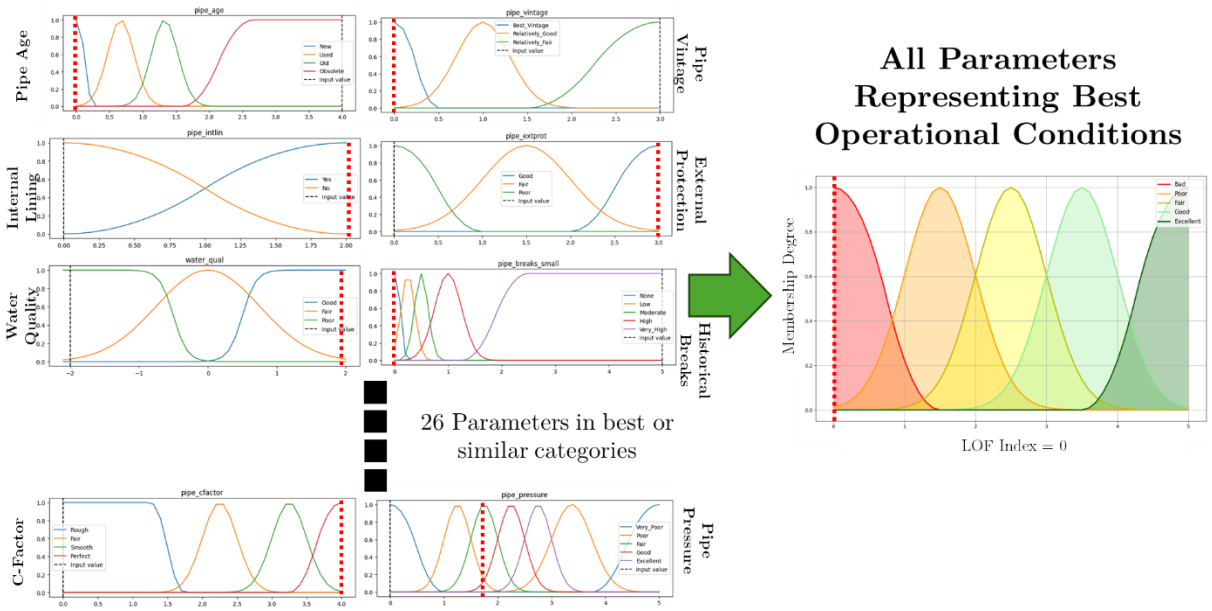

Figure 4-11: Best *set up of parameters leads to lowest LOF index (5). Expected LOF results are found for* Average *and* Worst *set of parameters*

Results: Membership-function panels matched operational ranges agreed with utilities, with overlaps that reflect real ambiguity. Canonical boundary tests for "all-best," "all-average," and "all-worst" settings returned LOF near 5, ~2.5, and ~0 as intended. Edge-case combinations flagged as infeasible by practice (e.g., mutually inconsistent measurements) did not produce spurious scores and instead yielded guarded outputs with wider uncertainty. Together, these checks show that inputs are well covered and the rule base behaves as designed at the boundaries.

4.7.2.2 Sensitivity and Robustness

We then assess how inputs influence outputs and whether responses are smooth and mechanism-plausible. Local checks use sign and monotonicity tests (e.g., increasing soil corrosivity should not decrease LOF within its valid range). Global checks use rank-order influence summaries and response surface plots (two-dimensional maps of LOF over driver pairs) to detect unintended interactions. Six such surface plots are shown in Figure 4-12. Surface plots are preferred here because they communicate interactions and sensitivity visually without heavy formalism to support large scale evaluation.

Results: The model responded in the right direction to changes in drivers. Higher soil corrosivity, pressure transients, or poor protection increased LOF. Better protection, thicker remaining wall, or deeper burial reduced it. Monotonic behavior (outputs moving one way when a driver is known to act one way) held over valid ranges. Pairwise response surface plots (simple 2-D maps of LOF over two inputs) were smooth and free of unintended peaks or troughs. Across nine stress regimens (best/average/worst; top-driver toggles; single-factor and compound extremes), outputs were ordered as expected, indicating stable aggregation and no fragile rule interactions.

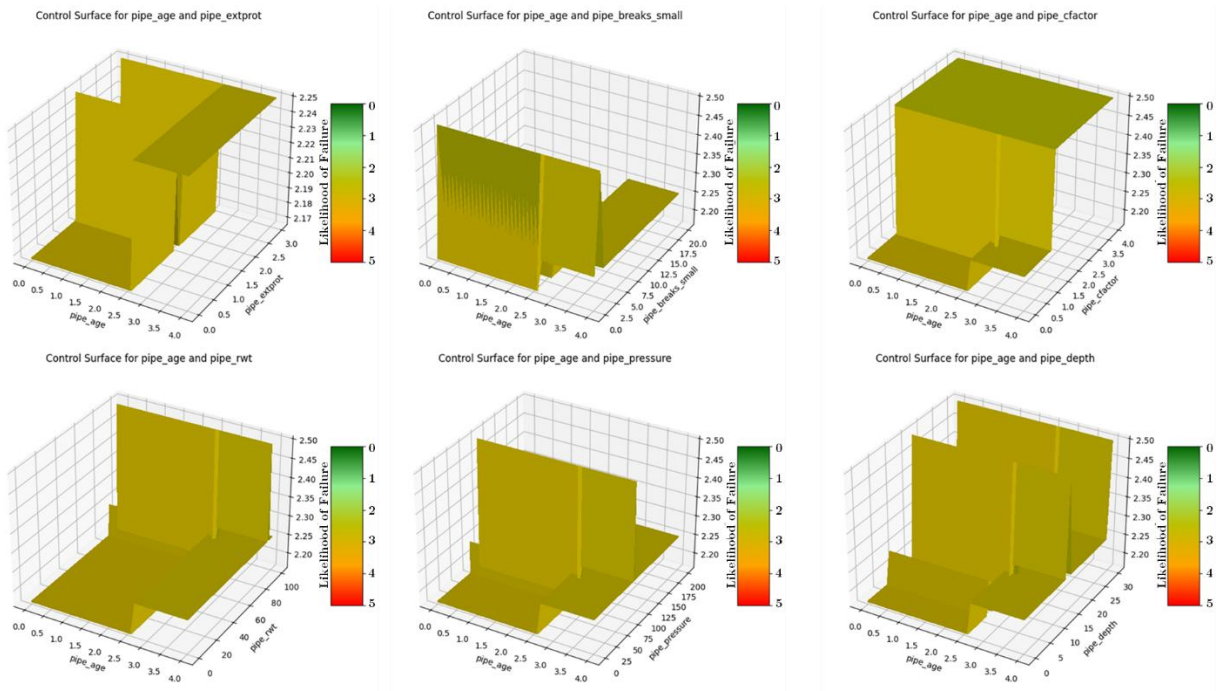

Figure 4-12: Surface plots to visualize model input-output parameter relationships. Here, only 6 such plots are shown as an example. A total of 1920 surface plots were developed and evaluated for all 125 parameters across 6 models for the performance model and for all the 21 parameters across the 5 modules of the COF model.

4.7.2.3 Scenario Alignment

To demonstrate face-validity, we score a compact set of heuristic scenarios curated from literature, interviews and historical records (e.g., ">40-year unprotected DI in highly corrosive soils" → Moderate; "PVC under sustained high internal pressure" → High). For each scenario we record the predicted LOF, the nearest linguistic band, and the directional rationale (which rule motifs fired, and why) (Kohavi and Thomke 2017). Alignment is evaluated on three axes: correct band, correct direction relative to a baseline case, and

reasonable distance (e.g., Moderate vs. High differs by ≥0.5 on the 0–5 scale for materially different stress states). Disagreements trigger a trace-back where we check and correct the rules, membership functions, units and data-ranges. A list of the scenarios checked for the LOF model with the model results are shown in Table 4-6.

Table 4-6: Results on all heuristically chosen theoretical scenarios

Scenario description	**Predicted LOF (0–5)**	**LOF band**	**Direction vs baseline (≈2.5)**
Greater than 40 year old 10″ unprotected DI pipe buried in high soil corrosivity.	2.75	Moderate	Higher (↑)
<20 year old >20″ diameter polywrapped DI pipe with groundwater fluctuations	1.25	Low	Lower (↓)
Greater than 40 year old 8″ diameter DI pipe in moderate soil corrosivity	0.75	Very Low	Lower (↓)
>50 year old CI pipe with low wall thickness in poorly drained soil	2.50	Moderate	Near baseline (↔)
>50 year old 6″ diameter CI pipe in moderately corrosive soil	1.25	Low	Lower (↓)
<50 year old 6″ diameter CI pipe in high soil corrosivity (<1000 ohm-cm) and low C-factor (<100; roughness)	2.75	Moderate	Higher (↑)
>36″ diameter concrete pipe in high steel soil corrosivity (<1500 ohm-cm), poor bedding, and groundwater fluctuation	4.50	Very High	Higher (↑)
>40 year old >36″ diameter concrete in medium-high corrosivity toward metals with minor cracking	2.75	Moderate	Higher (↑)
Concrete pipe with cement-mortar coating deterioration in fluctuating groundwater table conditions	4.50	Very High	Higher (↑)
>60 year old concrete pipe with spalling in high frost-action soil	4.50	Very High	Higher (↑)
>72″ concrete pipe (<50″ noted) with high pressure (>120 psi) in high precipitation and moderate–high corrosivity	2.00	Low	Lower (↓)
PVC pipe with significant ovality (>15%) in clayey soil	2.50	Moderate	Near baseline (↔)

Scenario description	Predicted LOF (0–5)	LOF band	Direction vs baseline (≈2.5)
PVC pipe subjected to prolonged high internal pressure	3.67	High	Higher (↑)
>30 year old PVC pipe in clayey soil with no manufacturing defects and no historical breaks	2.25	Moderate	Lower (↓)
<30 year old PVC pipe in clayey soil with no historical failures and good bedding	1.33	Low	Lower (↓)
HDPE pipe in a high temperature-differential environment (internal vs external)	3.67	High	Higher (↑)
>30 year old HDPE pipe with no deterioration	1.87	Low	Lower (↓)
<10 year old HDPE pipe in moderate pressure zone with fair bedding	1.33	Low	Lower (↓)
>50 year old steel pipe with external corrosion near electrified railway lines	1.25	Low	Lower (↓)
>72″ steel pipe with buckling/deformation in clayey soil near arterial roads	2.00	Low	Lower (↓)
<30 year old >72″ steel pipe in shallow water table and high precipitation	2.50	Moderate	Near baseline (↔)
>50 year old 6″ diameter AC pipe in moderate–high precipitation with multiple historical failures	4.90	Very High	Higher (↑)
30–50 year old 8″ diameter AC pipe with high groundwater fluctuation and moderate–high precipitation	2.75	Moderate	Higher (↑)
40–50 year old 6″ diameter AC pipe in low traffic loading with fair bedding	1.87	Low	Lower (↓)

Results: On a curated set of mechanism-plausible scenarios compiled from literature, interviews and records, predicted LOF bands and their direction relative to a typical case were consistent with expectations. Where a case sat near a band boundary, the score still moved in the right direction with a reasonable margin. Any tension between

expectation and score was traceable to explicit rules or membership supports and was resolved by minor threshold or overlap edits rather than ad-hoc fixes.

4.7.3 Verification: Supervised training of Student models on Teacher I/O

This section verifies that a data-driven student ML model can learn the teacher's mapping from input features to LOF bands for pipe assets. We then run a screening process for models of varying complexities on each of the 19 material-diameter cohorts and advance the best learner. Verification here is distinct from validation: verification asks "does the student reproduce the teacher's logic with real world data?"; validation (later) asks "does the system agree with external evidence from utilities?" Results for all models are shown using accuracy, precision, F1 and recall scores.

4.7.3.1 Framing and Task Definition

Here we treat verification as a supervised classification problem that is aligned with how utilities make financial capital investment and operational renewal decisions. For each pipe segment we start with tabular data with predictors like age, condition signals, hydraulics, soils, traffic loading, and material-specific indicators and ask a student model to reproduce the teacher's assignment of one of five LOF bands {0,1,2,3,4} corresponding

to Very Low to Very High. A “band” is a categorical risk class used in practice to plan actions for groups of pipes (monitor, repair, rehabilitate, replace) rather than to chase single-point scores. Using bands is decision-aligned, more robust to measurement noise, and understandable. If disagreement occurs, confusion matrices show whether it is concentrated in adjacent bands, which is acceptable in planning contexts that operate on classes rather than exact values.

The system is partitioned into 19 major material-diameter cohorts that span transmission and distribution (e.g., CI 8–24", DI >24", PVC <8", PCCP >24"). A cohort is a slice of assets that share deterioration mechanisms, operating envelopes, and data characteristics. Training independently per cohort respects these mechanistic differences, reduces spurious multimodality from mixing unlike populations, and preserves operational modularity. If a material is programmatically decommissioned or a new material appears, we can retire or add a cohort model without contaminating the others.

We adopt a teacher-student (model knowledge-distillation) setup. The evaluated teacher dataset which is a curated label source has been checked for representativeness (it covers the operating space), coverage (no systematic blind spots in relevant conditions), and scenario alignment (its distributions match intended deployment) provides the target

LOF bands. The student is a learner trained to approximate the teacher's input→ output class mapping. This brings three advantages for the pipe renewal problem.

Performance is evaluated with macro-F1 as the primary criterion, with accuracy, macro-precision, and macro-recall reported alongside. Accuracy summarizes overall agreement between true and predicted values across all classes. Precision and recall probe *how* those correct predictions are achieved. Precision (per class) is the fraction of predictions for that class that are correct and penalizes over-confident mislabels. Recall (per class) is the fraction of true items in that class that the model successfully recovers and penalizes misses. The F1 score for a class is the harmonic mean of its precision and recall, low if either is low. Macro-F1 is the unweighted average of per-class F1s, so rare classes influence the score as much as common ones. We use macro averaging (for precision and recall as well) because LOF bands are imbalanced. Macro metrics test whether the student preserves minority-class signal instead of optimizing only the dominant band, whereas micro averages would be dominated by frequent classes. We therefore report macro-F1 (primary), accuracy, macro-precision, macro-recall, and the full train/test confusion matrices so readers can see not only *how often* the model is correct but also *where* it errs. The formulae for calculating these 4 metrics are shown below.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

To make the screening fair and repeatable we fix *random seeds* for all stochastic steps. A random seed initializes the pseudo-random number generator that governs train/validation splits, weight initialization, data shuffling, and any synthetic sampling. Different seeds can change these draws and, in turn, model performance, especially in small or imbalanced cohorts. By holding seeds fixed across candidates, any performance difference reflects the models and data, not chance variation in initialization or partitioning. This choice enables apples-to-apples comparison and exact reproduction of results.

Finally, we state the boundaries of verification. This stage assumes that the teacher's labels are internally consistent within each cohort, that cohort definitions and band thresholds remain fixed during testing, and that the verified teacher data reflect the intended operating scenarios. External realism covering agreement with utility models, condition assessments, break histories, and forensic evidence is addressed later during

validation, where we test the verified student models against utility data and calibrate limits of applicability.

4.7.3.2 Screening funnel: Model Panel and Selection Criteria

We treat the training process as a screening funnel. The same data split, the same preprocessing, and the same evaluation rules are applied to a panel of learners that increase in capacity. The goal is to keep only the model that performs best based on quantitative metrics that reflect the ability to capture cohort-specific structure without overfitting idiosyncrasies of the split. To ensure a fair comparison, we performed identical preprocessing (normalization and standardization of numeric and categorical features). We fix random seeds so that any algorithmic stochasticity (weight initialization, batch order) is held constant across runs and makes differences in scores attributable to the model rather than chance.

The panel comprises 6 potential learner algorithms, varying by complexity. Multinomial Logistic Regression (LR), a linear classifier that estimates class probabilities via a softmax. It is the litmus test for whether the classes are linearly separable in the standardized feature space. Radial Basis Function (RBF)-kernel Support Vector Machine

(SVM) maximizes the margin between classes while the radial basis function kernel introduces smooth, nonlinear decision boundaries; this probes whether gentle nonlinearity suffices. Random Forests (RF) are ensembles of decision trees trained on bootstrap samples. This algorithm reduces variance by averaging many high-variance learners. This algorithm naturally model interactions but can be coarse around class boundaries. XGBoost (gradient-boosted trees with regularization) adds learners stage-wise to correct residual errors. Typically, on tabular datasets, it is a very strong non-neural baseline. We also include two neural candidates: a shallow MLP (three hidden layers) to test whether a modest depth can exploit interactions and thresholds, and a five-layer MLP that adds representational capacity while remaining compact enough to control variance. For both MLPs we use ReLU activations (stable gradients), dropout (randomly zeroing hidden units during training to curb co-adaptation), AdamW (Adam with decoupled weight decay for better generalization on tabular features), label smoothing (2–4% probability mass spread away from the hard label to reduce over-confidence), gradient-norm clipping ($\|g\| \leq 1$ to prevent rare exploding updates), and BatchNorm or LayerNorm depending on batch size (BatchNorm for regular batches; LayerNorm for tiny-N regimes). These stabilization guards ensure that any advantage a neural model shows comes from learning signal, not

training pathology. The architectures for these candidate algorithms is shown in Appendix E.

Selection criteria are metric-driven and identical across models. The primary criterion is macro-F1 on the training split's validation fold, because it gives equal weight to each LOF band and thus tests whether a learner retains minority-class signal. Accuracy is secondary, capturing overall agreement. When models are statistically tied, we break ties by (i) calibration (the alignment of predicted class probabilities with observed frequencies assessed via reliability curves / Expected Calibration Error) and (ii) error topology where we prefer models whose confusion matrices concentrate mistakes in adjacent bands (e.g., 2–3), which is operationally less harmful for utilities that act on bands rather than exact scores.

Under this apples-to-apples protocol across the 19 material-diameter cohorts, the five-layer MLP consistently wins or ties on macro-F1 and accuracy while maintaining good calibration and mostly adjacent-band confusions. XGBoost is typically the runner-up, sometimes matching accuracy but trailing slightly on macro-F1 in imbalanced cohorts. The shallow MLP generally lands between XGBoost and the five-layer MLP, indicating that additional depth is useful for capturing the interaction structure in these cohorts.

Linear and margin-based models (LR, RBF-SVM) and bagged trees (RF) provide valuable baselines but systematically underperform in cohorts with overlapping regimes and non-linear thresholds, precisely the settings we expect in LOF classification prediction. Consequently, the five-layer MLP advances as the student for verification, with the rest serving as documented reference points for transparency and reproducibility. The comparative summary of results from the model panel in training is shown in Table 4-7.

Table 4-7: Model training performance metrics. The metrics for the selected model (MLP-Deep) are shown in green.

Cohort	LR				SVM				RF				XGB				MLP (Shallow)				MLP (Deep)			
	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
AC <8”	0.80	0.58	0.54	0.56	0.84	0.61	0.59	0.60	0.86	0.64	0.61	0.62	0.88	0.66	0.64	0.65	0.89	0.67	0.65	0.66	**0.99**	**0.99**	**0.99**	**0.99**
AC 8”–24”	0.85	0.70	0.66	0.68	0.88	0.72	0.70	0.71	0.89	0.75	0.72	0.73	0.91	0.76	0.74	0.75	0.92	0.77	0.75	0.76	**0.99**	**0.94**	**0.97**	**0.95**
CI <8”	0.85	0.66	0.62	0.64	0.88	0.68	0.66	0.67	0.89	0.71	0.68	0.69	0.91	0.72	0.70	0.71	0.92	0.73	0.71	0.72	**0.99**	**0.99**	**0.99**	**0.99**
CI 8”–24”	0.94	0.86	0.84	0.85	0.96	0.89	0.87	0.88	0.97	0.92	0.89	0.90	0.98	0.93	0.91	0.92	0.99	0.94	0.92	0.93	**0.99**	**0.99**	**0.99**	**0.99**
CI >24”	0.90	0.69	0.65	0.67	0.92	0.71	0.69	0.70	0.94	0.74	0.71	0.72	0.95	0.75	0.73	0.74	0.96	0.76	0.74	0.75	**0.99**	**0.75**	**0.80**	**0.77**
Concrete >24”	0.90	0.88	0.86	0.87	0.92	0.90	0.88	0.89	0.93	0.92	0.89	0.90	0.94	0.92	0.90	0.91	0.95	0.93	0.91	0.92	**0.94**	**0.89**	**0.92**	**0.90**
DI <8”	0.85	0.54	0.50	0.52	0.88	0.55	0.53	0.54	0.90	0.58	0.55	0.56	0.92	0.59	0.57	0.58	0.93	0.60	0.58	0.59	**0.99**	**0.64**	**0.80**	**0.67**
DI 8”–24”	0.90	0.78	0.74	0.76	0.92	0.80	0.78	0.79	0.93	0.83	0.80	0.81	0.95	0.84	0.82	0.83	0.96	0.85	0.83	0.84	**0.99**	**0.99**	**0.99**	**0.99**
DI >24”	0.94	0.92	0.90	0.91	0.96	0.94	0.92	0.93	0.97	0.96	0.94	0.95	0.98	0.97	0.95	0.96	0.99	0.98	0.96	0.97	**0.99**	**0.93**	**0.95**	**0.94**
PCCP 8”–24”	0.93	0.85	0.83	0.84	0.95	0.87	0.85	0.86	0.96	0.89	0.86	0.87	0.97	0.90	0.88	0.89	0.98	0.91	0.89	0.90	**0.99**	**0.90**	**0.96**	**0.93**
PCCP >24”	0.90	0.81	0.79	0.80	0.92	0.84	0.82	0.83	0.93	0.87	0.84	0.85	0.95	0.88	0.86	0.87	0.96	0.89	0.87	0.88	**0.99**	**0.97**	**0.98**	**0.97**
PE <8”	0.90	0.87	0.85	0.86	0.92	0.89	0.87	0.88	0.93	0.91	0.88	0.89	0.94	0.91	0.89	0.90	0.95	0.92	0.90	0.91	**0.99**	**0.99**	**0.99**	**0.99**
PE 8”–24”	0.86	0.78	0.74	0.76	0.89	0.80	0.78	0.79	0.90	0.82	0.79	0.80	0.92	0.83	0.81	0.82	0.93	0.84	0.82	0.83	**0.96**	**0.91**	**0.87**	**0.87**
PE >24”	0.85	0.86	0.82	0.84	0.88	0.88	0.86	0.87	0.89	0.90	0.87	0.88	0.91	0.91	0.89	0.90	0.92	0.92	0.90	0.91	**0.91**	**0.88**	**0.93**	**0.89**
PVC <8”	0.86	0.83	0.79	0.81	0.88	0.84	0.82	0.83	0.89	0.86	0.83	0.84	0.91	0.87	0.85	0.86	0.92	0.88	0.86	0.87	**0.99**	**0.99**	**0.99**	**0.99**
PVC 8”–24”	0.89	0.57	0.53	0.55	0.92	0.59	0.57	0.58	0.93	0.62	0.59	0.60	0.95	0.63	0.61	0.62	0.96	0.64	0.62	0.63	**0.99**	**0.92**	**0.86**	**0.86**
ST <8”	0.84	0.71	0.67	0.69	0.87	0.73	0.71	0.72	0.88	0.76	0.73	0.74	0.90	0.77	0.75	0.76	0.91	0.78	0.76	0.77	**0.99**	**0.99**	**0.99**	**0.99**
ST 8”–24”	0.80	0.72	0.68	0.70	0.83	0.74	0.72	0.73	0.84	0.77	0.74	0.75	0.86	0.78	0.76	0.77	0.87	0.79	0.77	0.78	**0.99**	**0.99**	**0.99**	**0.99**
ST >24”	0.94	0.90	0.88	0.89	0.96	0.92	0.90	0.91	0.97	0.93	0.91	0.92	0.98	0.94	0.92	0.93	0.99	0.95	0.93	0.94	**0.99**	**0.94**	**0.98**	**0.96**

4.7.3.3 Synthetic, bias-resistant Testing Strategy

Rather than re-using a random slice of the same dataset for testing, a practice that can leak subtle structure from train to test and inflate scores, we evaluate each cohort with a completely unseen synthetic battery. For every material-diameter cohort, we generate 1,000 samples that never touch training, balanced across the five LOF classes (≈200 per band). "Synthetic" here does not mean arbitrary. Samples are drawn to mimic the cohort's empirical feature geometry following the marginal distributions and the correlations among predictors while obeying domain guardrails (values remain within observed ranges and respect expected monotonic directions. This setup reduces split bias and gives a controlled, repeatable way to probe generalization.

Construction of the test dataset battery proceeds in three steps. First, we rank predictors by influence into Top, Middle, and Bottom tiers. The ranking is cohort-specific and triangulates the teacher's feature dictionary (expected sign/shape) with the feature sensitivity. Second, for every predictor we define Best, Average, and Worst operating anchors using robust quantiles aligned to the expected effect. For a risk-increasing variable, "Best" is a low quantile; for a protective variable, "Best" is a high quantile. Anchors are clipped to observed minima/maxima, so we never create impossible values. Third, we

synthesize samples that preserve correlation structure using a numerically safe Gaussian copula. In this process, we sample in a latent normal space whose correlation matrix is projected to be Positive Semi-Definite (PSD; all eigenvalues non-negative so it is a valid covariance), then map back to the original feature scales and finally clip to realistic ranges. Random seeds are fixed so that the same protocol yields the same synthetic dataset battery on re-runs, making comparisons stable.

We then stress-test the model under nine targeted scenario levels that push different parts of the feature space (see Table 4-8). Level 1 sets all predictors to their Best anchors, checking calibration in benign regimes; Level 2 sets all to Average, checking typical operating conditions; Level 3 places only the Top-influence features at Worst while keeping Middle and Bottom at Average, isolating vulnerability to the highest-leverage drivers. The remaining six levels permute the anchor settings across influence tiers to expose interaction effects: Levels 4–5 hold Top at Best while alternating Middle/Bottom between Average and Worst; Levels 6–7 promote Middle to Best with Top at Average and Bottom at Worst or Best; Levels 8–9 push Top to Worst and counterbalance with Bottom at Best or Average. This grid produces orthogonal "nudges" on the decision

boundary and reveals where predictions bend, information that a single random test split cannot provide.

Table 4-8: Nine-level synthetic stress bands used for verification. Each "Band" fixes the anchor settings (Best / Average / Worst) for the Top-, Middle-, and Low-influence predictor tiers and reports the expected LOF tendency. Bands 1–2 probe benign and typical regimes; Band 3 isolates vulnerability to the highest-leverage drivers (Top at Worst); Bands 4–9 permute anchors across tiers to expose interaction effects while holding others steady. The expected tendency is an a priori direction-of-risk guide, not a constraint on model outputs.

Level	**Top-influence features**	**Medium-influence features**	**Low-influence features**	**Expected LOF tendency**
1	Best	Best	Best	Very Low
2	Average	Average	Average	Moderate (baseline)
3	Worst	Average	Average	Very High (stress top drivers)
4	Best	Average	Worst	Very Low–Low
5	Best	Worst	Average	Moderate
6	Average	Best	Worst	Moderate
7	Average	Worst	Best	Moderate–High
8	Worst	Average	Best	High
9	Worst	Best	Average	High

Because the synthetic battery is class-balanced by design, performance summaries reflect every LOF band rather than being dominated by the majority class. We report precision, recall, accuracy and F1 as we did during training. We also compute macro-averages (unweighted means across classes) so minority bands carry equal weight, and we include overall accuracy and full confusion matrices. Table 4-9 reports the training and synthetic-test outcomes for the selected five-layer MLP.

Table 4-9: Training and Testing Macro Results Summary for MLP (Deep)

Material | Diameter Cohort	Samples	Features	Training				Testing			
			Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
AC <8”	66435	18	0.99	0.99	0.99	0.99	0.90	0.60	0.91	0.67
AC 8”-24”	6781	18	0.99	0.94	0.97	0.95	0.92	0.73	0.95	0.77
CI <8”	229394	27	0.99	0.99	0.99	0.99	0.92	0.67	0.94	0.73
CI 8”-24”	35932	27	0.99	0.99	0.99	0.99	0.99	0.89	0.98	0.93
CI >24”	1218	27	0.99	0.75	0.80	0.77	0.96	0.73	0.79	0.76
Concrete >24”	1098	21	0.94	0.89	0.92	0.90	0.95	0.93	0.93	0.92
DI <8”	349025	27	0.99	0.64	0.80	0.67	0.94	0.60	0.77	0.60
DI 8”-24”	105922	27	0.99	0.99	0.99	0.99	0.96	0.82	0.98	0.84
DI >24”	4834	27	0.99	0.93	0.95	0.94	0.99	0.97	0.96	0.97
PCCP 8”-24”	2685	23	0.99	0.90	0.96	0.93	0.98	0.88	0.93	0.90
PCCP >24”	2686	23	0.99	0.97	0.98	0.97	0.96	0.83	0.96	0.89
PE <8”	7723	19	0.99	0.99	0.99	0.99	0.95	0.91	0.91	0.91
PE 8”-24”	694	19	0.96	0.91	0.87	0.87	0.93	0.88	0.82	0.83
PE >24”	85	19	0.91	0.88	0.93	0.89	0.92	0.90	0.93	0.91
PVC <8”	145106	19	0.99	0.99	0.99	0.99	0.92	0.83	0.93	0.87
PVC 8”-24”	13073	19	0.99	0.92	0.86	0.86	0.96	0.71	0.77	0.64
ST <8”	24318	27	0.99	0.99	0.99	0.99	0.91	0.70	0.93	0.77
ST 8”-24”	18696	27	0.99	0.99	0.99	0.99	0.87	0.73	0.93	0.78
ST >24”	5435	27	0.99	0.94	0.98	0.96	0.99	0.93	0.95	0.94

The corresponding confusion matrices for 3 cohorts are shown in Figure 4-13. Across all three cohorts the matrices are strongly diagonal, indicating high discriminative performance on both the training and synthetic-test sets.

Most errors occur as adjacent-band confusions (e.g., 1–2 or 3–4), which is expected near decision boundaries and acceptable for planning that acts on bands rather than exact scores. For PE $< 8''$, test errors are concentrated in band 2 leaking to bands 1 and 3, while bands 3–4 remain very pure. For ST $< 8''$, band-1 shows modest spillover to bands 0 and 2, consistent with broader within-band variability in this cohort, yet the dominant mass remains on the diagonal. PCCP $> 24''$ exhibits near-perfect diagonals on both splits with only trace off-diagonals, reflecting a highly separable label structure. The close alignment of train and test patterns suggests controlled overfitting and supports the reported accuracy, precision, recall, and macro-F1 values. The complete set of training confusion matrices are presented in Appendix E.

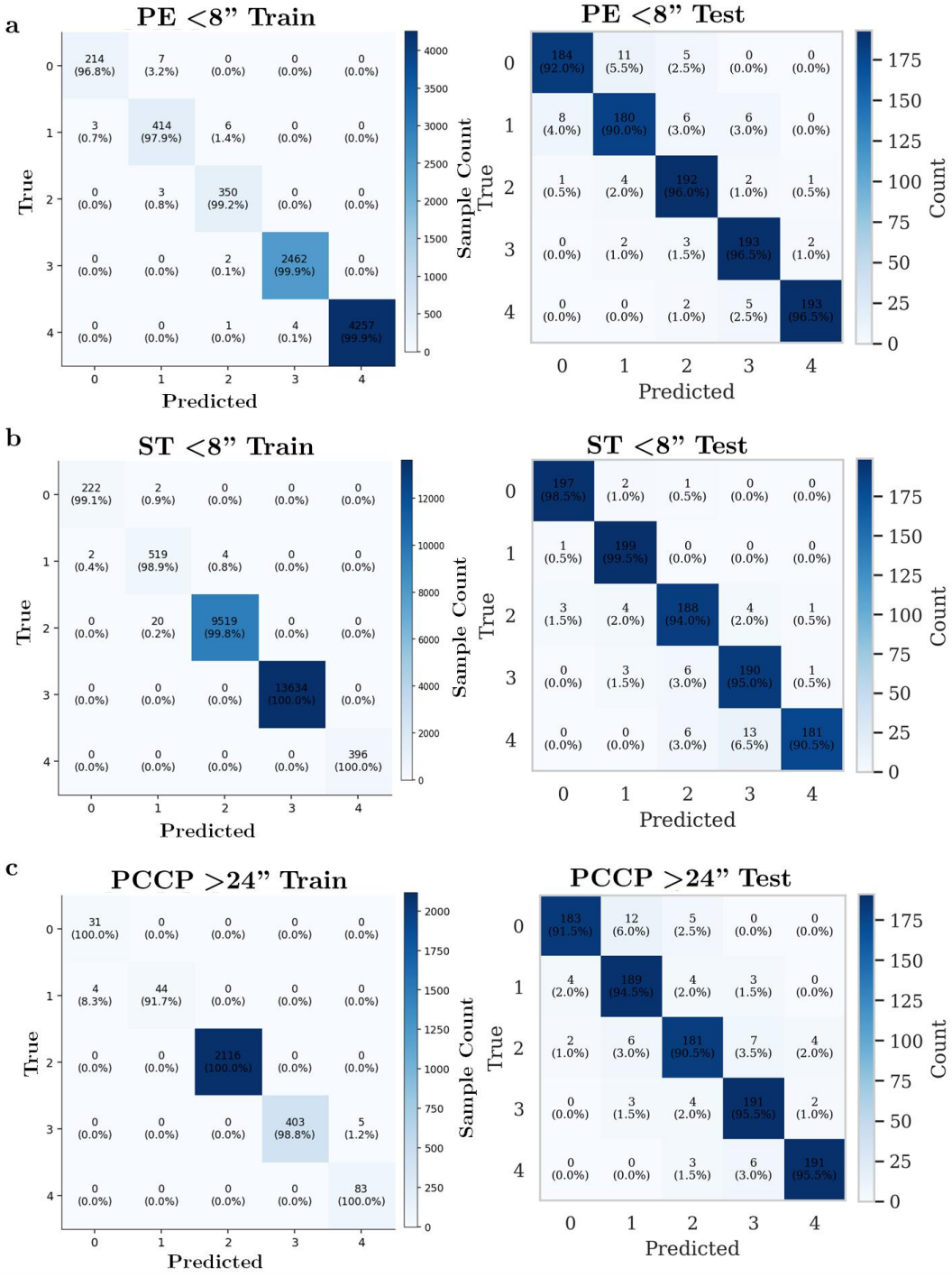

Figure 4-13: Train vs. synthetic-test confusion matrices for three representative pipe cohorts—(a) PE < 8", (b) ST < 8", and (c) PCCP > 24" using the five-layer student MLP. Rows are

true LOF bands (0–4) and columns are predicted bands; each cell shows the sample count with the row-normalized percentage in parentheses. Color encodes absolute count (scale bar at right).

4.7.3.4 Student MLP Architecture

Our student is a five-layer MLP designed for tabular inputs. Each hidden layer is Linear → Normalization → ReLU → (Dropout), ending with a Linear layer that outputs five logits for the LOF bands (0–4). A Linear layer is a learned affine map $Wx + b$: it forms weighted sums of the input features so later layers can combine them into higher-level patterns. Logits are the raw, pre-softmax scores; turning them into probabilities is just a softmax step and keeping them as logits lets the loss function work directly on unconstrained real numbers.

Normalization keeps activations on a steady numeric scale so training is faster and steadier. With Batch Normalization (BN), we standardize each hidden unit using the batch statistics that is, the mean and variance computed over examples in the current mini-batch which also adds mild regularization (small, helpful noise from estimating those statistics). When batches are tiny, we switch to Layer Normalization (LN), which normalizes across features within each individual example and does not depend on batch size.

We use ReLU (rectified linear unit, $\max(0, x)$) because it is simple and fast and keeps gradients well-behaved that is, the derivatives neither collapse toward zero (vanishing updates) nor blow up to huge values (exploding updates). To limit overfitting we apply Dropout (default 0.08), which randomly turns off a fraction of hidden activations (the outputs of hidden neurons) during training; this forces the network to rely on multiple cues instead of memorizing spurious patterns.

Optimization is done with AdamW, which decouples weight decay from the adaptive gradient updates. Weight decay is the L2 regularization penalizing the sum of squared weights to prefer simpler models. Decoupling it from the step sizes gives more reliable generalization on mixed-scale tabular features than classic Stochastic gradient descent (SGD) or Adam. Our chosen values for $Learning\ Rate = 8 \times 10^{-4}$ and $Weight\ Decay = 4 \times 10^{-4}$ consistently converge across cohorts without overshooting.

The classifier head uses label smoothing to avoid over-confident probabilities. Instead of a strict one-hot target, we assign most mass to the true class and spread a small remainder over the others. We use 2% by default and 3–4% for tiny-N cohorts. Below ~2% there's little calibration gain and above ~4% minority bands can underfit. We also cap rare gradient spikes with gradient-norm clipping ($\| g \|_2 \leq 1$), a guardrail that prevents

a single outlier mini-batch (common under class imbalance) from destabilizing learning. To further steady validation curves we keep an exponential moving average (EMA $\approx$ 0.995) of the weights; EMA acts like a low-pass filter with ~200-step memory and usually gives a small macro-F1 bump at selection time.

Together, ReLU + Dropout + AdamW + label smoothing + grad-clip + EMA, the BN/LN regime switch, and the last-batch guard form a coherent stabilization policy where gradients stay well-scaled, probabilities stay calibrated, and model selection is driven by minority-sensitive macro-F1 rather than luck. A summary of the network blocks and training hyperparameters used across cohorts is shown in Table 4-10.

Table 4-10: Five layer student MLP model architecture and stabilization settings

Component	**Setting**	**Rationale**	**Notes**
Hidden layers	5 (sizes 50–40–30–20–10)	Enough capacity for interactions & thresholds; still compact	Sizes chosen to fit tabular feature counts without overfitting
Activation	ReLU	Stable, fast gradients; avoids vanishing	Applied after each normalization
Normalization (standard)	BatchNorm	Faster, more stable training; mild regularization	Used when batch size is reasonable
Normalization (tiny-N)	LayerNorm	Stable when batches are tiny; batch-size agnostic	Automatically selected for small-N cohorts
Dropout	0.08	Anti-overfit via stochastic feature thinning	Applied after ReLU in hidden layers
Optimizer	AdamW	Decouples weight decay from adaptive step; robust on tables	Better generalization than Adam/SGD in pilots
Learning rate	8e-4	Reliable convergence across cohorts	Tuned to avoid overshoot
Weight decay	4e-4	Preference for simpler models; reduces overfit	Decoupled via AdamW

Component	Setting	Rationale	Notes
Label smoothing	0.02 (default), 0.03–0.04 (tiny-N)	Improves calibration; protects minority bands	Higher values risk underfitting
Gradient clip	$\|g\|_2 \leq 1.0$	Prevents rare gradient spikes	Stability guard, not a crutch
EMA of weights	0.995 decay	Smoother validation; small macro-F1 boost	~200-step effective memory
Batch size	256 (standard)	Efficient, stable BN stats	Reduced under tiny-N as needed
Epochs (base cap)	≈180	Enough budget for convergence	Early stopping often halts earlier
Early-stopping metric	Macro-F1 (primary), Accuracy (secondary)	Focus on minority-class performance	Patience ≈ 50 steps
BN last-batch guard	Duplicate last example if size=1	Avoids undefined BN statistics	No effect on decision boundary
Seeds & splits	Fixed seeds; stratified splits	Full reproducibility; preserves class ratios	Controls all stochastic steps

4.7.3.5 Results Summary

Our teacher to student verification proceeded cohort-wise (19 material-diameter cohorts), with a fixed screening funnel (LR, RBF-SVM, RF, XGB, shallow-MLP, deep 5-layer MLP). Model choice was pre-registered: select by macro-F1 (primary) and accuracy (secondary), breaking ties by calibration and adjacent-band confusions. In practice, the five-layer MLP outperformed most cohorts and was chosen to advance to the testing phase. Reported metrics include accuracy, precision, recall, and macro-F1, with confusion matrices for both train and synthetic-test. Cohort acceptance is defined as train accuracy ≥90% and synthetic-test accuracy ≥80%. Generalization was probed on an unseen

synthetic test set (~1000 samples) generated per class. This test set stressed the model on a 9-layer grid that toggled Top/Mid/Low-influence feature groups among Best/Average/Worst anchors to examine calibration and boundary behavior relevant to operational thresholds.

Reproducibility is ensured by a per-cohort export bundle that locks the full scoring pipeline: model.pt (weights), preproc.json (feature order, μ, σ), model_config.json (architecture and hyperparameters), and export_manifest.json (metrics and configuration). These bundles enable exact rescoring and enables the next validation phase on real-world utility data without retraining.

4.7.4 Validation of the Student LOF models

This section demonstrates how the LOF "student" models perform outside the development environment and under operational use. The aims are threefold: (i) find areas of agreement between the proposed model and the utility's internal model, (ii) assess expert concordance that is, agreement between model bands and judgments from asset managers and field crews, and (iii) test ground-truth agreement using failure, condition assessment and forensic data. Results based on data from utilities A–D can be summarized

in 3 parts. First, model-to-utility comparisons show how our student LOF ranks segments relative to incumbent indices, summarized by rank agreement (Spearman ρ), typical score gap (median $|\Delta|$ on 0–5), and large-gap rate ($|\Delta|>2$) with cohort breakouts. Second, expert-concordance results aggregate strict and ± 1-band (tolerant) agreement from asset managers and field crews, with coded reasons for disagreement. Third, retrospective ground-truth checks report the share of next-year failures captured in top-ranked segments using cohort-appropriate targets (PCCP wire-breaks, metallic RWT thresholds, work-order breaks otherwise) with uncertainty bands. The utilities participating in validation are shown in Figure 4-9. Together, these results examine degree of agreement, diagnose gaps, and identify low-effort remediation targets.

4.7.4.1 Comparison with Utility Models

We compare our LOF outputs with each utility's incumbent index (in-house or consultant models) to test *shape* and *scale* agreement and to surface systematic disagreements before engaging experts or ground truth. We harmonize IDs and basic metadata, align directions (if a utility provides a "performance/condition" score where high=good, we apply a monotone transform so high=bad), and check (i) rank agreement within

material-diameter cohorts and (ii) calibration on a shared 0–5 scale ($\Delta = \text{LoF}_{\text{student}} - \text{LoF}_{\text{utility}}$). Disagreement pockets indicate missing predictors, scale drift, or practice differences that merit calibration.

Agreement between our student LOF index and each utility's incumbent index spans strong to discordant. This can be interpreted from the scatterplots shown in Figure 4-14. We assessed how closely each utility's incumbent index orders pipes relative to our student LOF (0–5, higher = worse) using two simple, interpretable quantities:

- Rank agreement (Spearman ρ): measures whether the two scores put segments in a similar order (1 = perfect, 0 = no monotone relation).
- Typical difference (median $|\Delta|$): the median absolute score gap on the 0–5 scale, where Δ = Student − Utility; we also report the large-gap rate (share with $|\Delta|>2$), which flags pockets of substantial calibration disagreement. Sample size (n) is shown because very small n limits statistical stability.

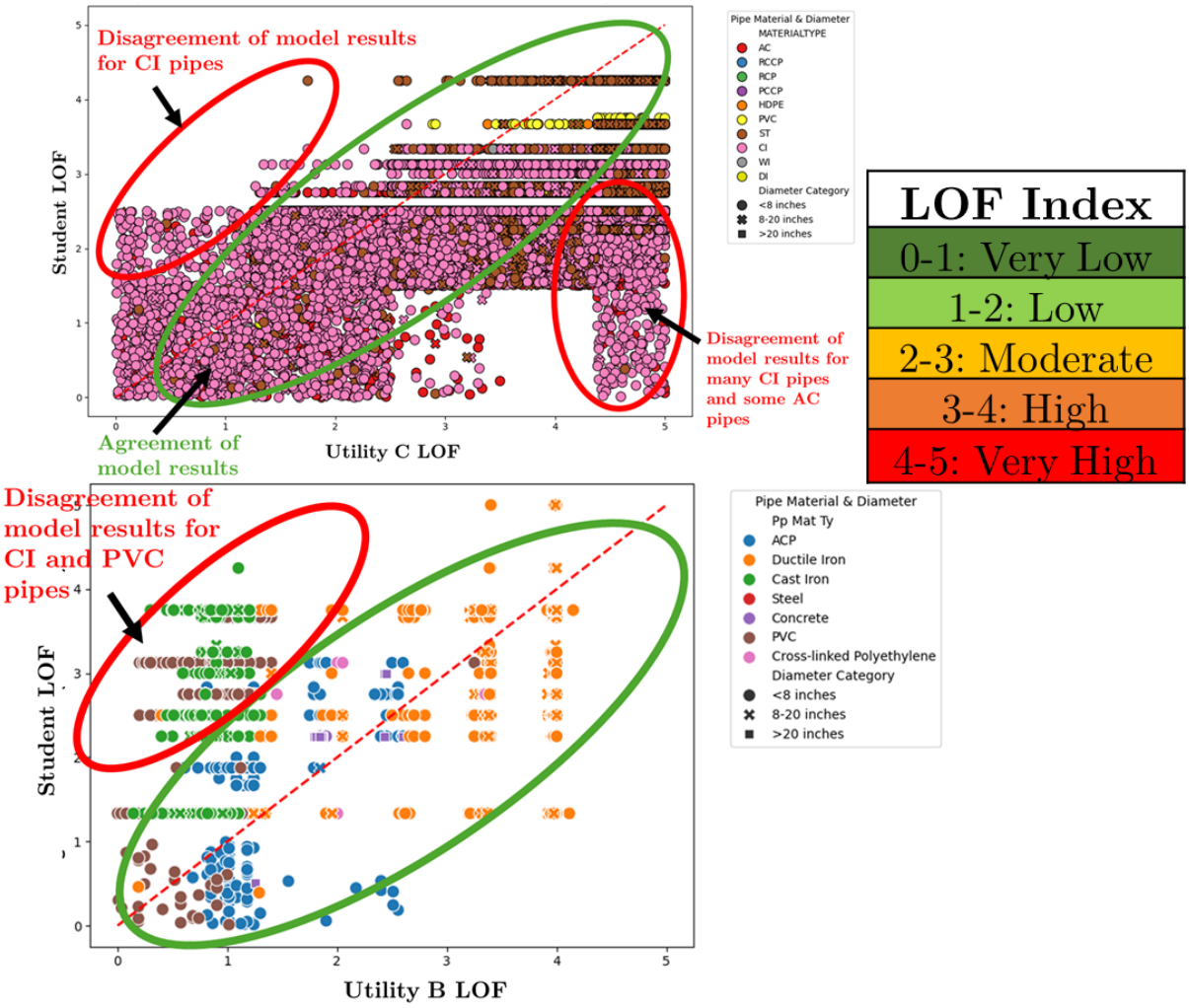

Figure 4-14: Comparison scatterplots showing disagreements between Student LOF predictions and Utilities B (bottom) and C (top) LOF

Utility A shows strong agreement ($\rho \approx 0.60$) with small typical gaps (median $|\Delta| \approx 0.83$; large-gap $\approx 10\%$). Utility B is moderate ($\rho \approx 0.41$; median $|\Delta| \approx 1.35$; large-gap $\approx 28\%$). Utility C is fair/weak ($\rho \approx 0.36$; median $|\Delta| \approx 1.87$; large-gap $\approx 36\%$). Utility D is discordant ($\rho \approx 0.05$) despite only moderate typical gaps (median $|\Delta| \approx 1.50$; large-gap

≈16%). It is important to note that for Utility D n≈4k versus >100k for Utilities A–C, which naturally depresses the stability and ceiling of rank-based metrics. A summary of these results is presented in Table 4-11.

Table 4-11: Utility–model agreement (Indices on 0–5; higher = worse. Δ = Student – Utility)

Utility	Samples	Miles	Rank agreement ρ	Median |Δ|	Large-gap rate (|Δ|>2)
A	105,746	2,200	0.6	0.833	10.40%
B	206,961	3,900	0.41	1.352	28.20%
C	122,592	4,200	0.364	1.867	35.60%
D	4,133	3,680	0.049	1.5	15.70%

The pattern is consistent with design differences. Utility A's stronger conformity aligns with two methodological choices: (i) survival-curve–based LOF modeling that respects slow-then-fast aging (rather than linear age weights), and (ii) more granular stratification (e.g., 2-inch diameter bins) that reduces scale drift across heterogeneous cohorts. By contrast, the remaining utilities largely employ simple weighted scores that emphasize age and past breaks, with limited treatment of exposure normalization, protections, or mechanism-specific signals. For Utility B, the policy environment appears to double-penalize PVC, consistent to support an ongoing replacement program, indicating departure from modeling best practices. These design choices can depress within-cohort rank agreement and inflate calibration pockets even when the global signal is directionally similar.

None of this is dispositive. Utility indices are "comparators of convenience," not ground truth. The role of this step is diagnostic, that is, to locate systematic disagreements and focus the next stages (expert concordance and one-year failure coverage) where targets are closer to operations.

4.7.4.2 Agreement with Expert Opinion

We elicit feedback from asset managers and field crews of participating water utilities using the structured form (scenario description, model band/result, agree/disagree, optional comment). The forms with the utility feedback are available in Appendix F. Segments are stratified by material-diameter and by output classes so that easy cases do not dominate. Each response is scored two ways: (i) binary agreement (agree vs disagree), and (ii) adjacent-band tolerance, where expert notes such as "fair–good" count as ±1-band concordance. We report percent agreement with Wilson confidence intervals, but decision making relies on chance-corrected measures. Chance correction is performed using metrics like Gwet's AC1 for binary agreement and weighted AC1/κ with linear weights when band proximity is specified. These statistics correct for *prevalence bias* (e.g., when one band is common, raw accuracy can look high) and respect the ordinal structure of LOF bands (an off-by-one is less consequential than "Very High" vs "Very Low").

Inference uses a paired, stratified bootstrap over segments (strata = material-diameter) to obtain 95% confidence intervals that reflect the sampling design. In short, AC1/weighted κ give a fair test of whether experts and the model truly agree beyond chance and whether disagreements are *material* rather than minor, which plain accuracy cannot distinguish.

Across three utilities, agreement is highest where comments reference physical signals, that is, very old CI with low wall thickness, high-pressure steel, highly corrosive soils. Agreement is found to be lower where notes reflect policy heuristics (blanket penalties for a material or small diameters, "useful life" rules) or data defects (inactive assets, mis-sized diameters). A result summary from eliciting feedbacks from participating water utilities is shown in Table 4-12. The received feedback forms from all 3 water utilities can be found in Appendix G.

Table 4-12: Expert concordance by utility

Utility	**Expertise**	**Reviews (n)**	**Strict agree**	**±1 Band Agreement***	**Primary disagreement motifs (coded)**
A	Asset Managers/ Planners	23	73.9% (17/23)	–	Inventory/status and diameter errors (inactive asset; 20″ vs >72″; 48″ vs >48″); a few ledger "poor" with no work orders

Utility	Expertise	Reviews (n)	Strict agree	±1 Band Agreement*	Primary disagreement motifs (coded)
B	Asset Managers/ Planners	24	58.3% (14/24)	(rare band ranges reported)	Policy/heuristic priors ("target material," "small diameter," "useful life"), definition ambiguity ("near" road), one category-coverage question (buckling near roads)
C	Field Crew	25 (23 after excluding 2 mislabeled)	78.3% (18/23)	91.3% (21/23)	Two HDPE cases tied to service-connection failures (not main); inventory mislabeling flagged and excluded; seismic context notes for steel

* "±1 Band Agreement" counts expert remarks that straddle adjacent bands (e.g., "fair–good"). Records with clear inventory mislabeling (e.g., PCCP vs C303 welded steel) were excluded from rate calculation and logged for correction. The motifs column comes from coding the free-text comments into: Inventory/metadata**,** Definition ambiguity**,** Policy/heuristic bias**, and** Category-coverage gap**.**

For Utility A (responses from asset managers/planners; n=23), 17/23 (73.9%) strictly agreed with the model predictions. All six disagreements are administrative/metadata issues with one inactive asset, two erroneous diameter entries ("20″ not >72″", "is 48″"), and three ledger "poor" labels despite no work orders, indicating high chances of model–practice alignment once inventory is clean. Utility B (asset managers/planners; n=24) shows 14/24 (58.3%) strict agree; disagreements cluster in CI and PVC and cite policy overlays (like target material and small diameter pipes for replacement and pipes selected that exceed estimated useful lives). Where mechanism cues are sharp (e.g., >36″ concrete in high corrosivity with spalling), experts endorsed our low scores and only flagged issues in selecting for candidate renewal (Utility B only uses LOF as the metric for renewal), citing it is a "hard to sell" asset to the upper management

(large diameter pipes are typically monitored well to extract maximum possible life and avoid expensive renewal projects). Utility C (field crew operators/experts; n=25, with two clear material mislabels excluded) yields 21/23 (91.3%) tolerant-agree and 18/23 (78.3%) strict-agree. The two disagreements involve HDPE service-connection issues (saddle failures) rather than main-line failure modes, and seismic context notes reinforce face validity where our score was already low. Importantly, the strongest concordance comes from field crews, those closest to observed failure mechanisms, raising confidence that the model captures operational reality. Residual gaps are actionable. Actions include correction of inventory/status and diameters; separation service-connection samples from main-line failure categories for plastics and documentation of any intentional policy deviations (e.g., PVC programs) so they remain transparent.

4.7.4.3 Ground truth agreement for CI, DI, ST and PCCP using Remaining Wall Thickness

We validate our student ML model by comparing model outputs ($LOF_{student}$) to ground-truth LOF derived from field measurements. For this section, we can perform validation for the cohorts that the wall thickness measurement instrument works on and that were prioritized for condition management at a large, Mid-Atlantic utility with a

high share of legacy cast-iron mains and multiple pressure zones. Specifically, the wall thickness measurements exist for the following cohorts: Pit cast iron (<8", 8–24"), spun cast iron (<8", 8–24"), ductile iron (<8", 8–24"), steel (>24"), and prestressed concrete cylinder pipe (PCCP: 8–24", >24"). Other materials/diameters appear in our dataset but are validated in the next section using different datasets and methods tailored to their mechanisms.

4.7.4.3.1 Data Collection

We use data collected by the vendor contracted by this mid-atlantic water utility in 2015. The report summarizes non-invasive acoustic condition assessments and targeted pressure-transient logging. Field work for the validated assets was conducted roughly between October 2014 and October 2015. Deliverables of this report included tables with Integrity Rating (IR, 1–5) for pit/spun/cast-iron and steel, and percent wall-thickness loss for ductile iron, plus GIS linkages and QA notes (e.g., bracket geometry checks, appurtenance verification). For our ground-truth tests in this section, we restrict to the cohorts to Pit CI (<8", 8–24"), Spun CI (<8", 8–24"), DI (<8", 8–24"), Steel (>24"), and PCCP (8–24", >24").

To characterize the inspection sample used for wall-thickness ground truthing, we summarized the mileage and age structure of the dataset (Figure 4-15). Most inspected footage is spun and pit cast iron, with smaller contributions from ductile iron, steel, and PCCP across all diameter bands, and includes both lined and unlined segments. Pit cast iron and steel are the oldest cohorts, PCCP and spun cast iron are intermediate, and ductile iron is generally younger. Pooled across materials, the wall-thickness–versus-age hexbin shows broad scatter and only a weak trend, indicating that simple monotonic thinning with age is largely masked by lining, replacement, and survival bias in the inspected population.

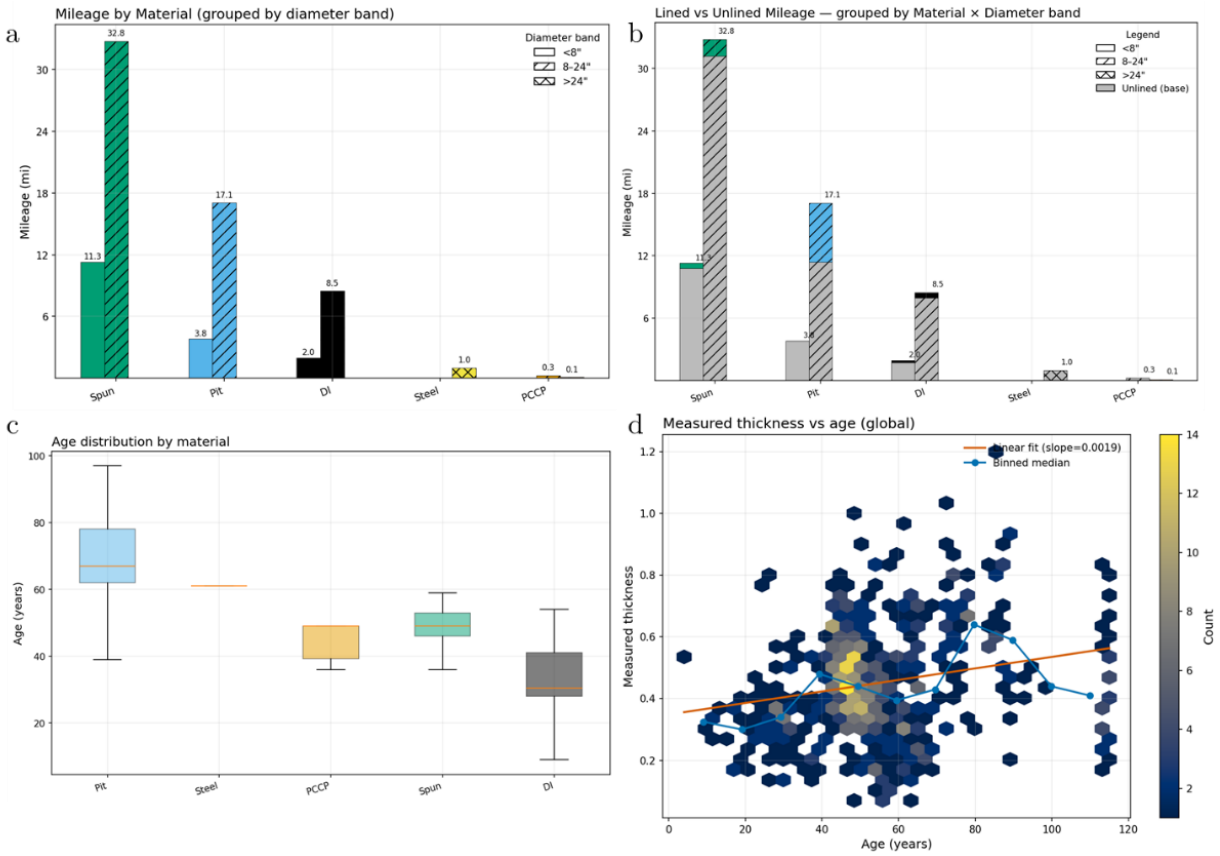

Figure 4-15: Exploratory characterization of the wall-thickness inspection dataset. (a) Mileage of inspected pipe by material and diameter band, showing that most footage is spun and pit cast iron with smaller contributions from ductile iron, steel, and PCCP. (b) Partition of the same mileage into lined and unlined segments by material–diameter combination. (c) Age distributions by material, with pit and steel cohorts generally older than PCCP, spun, and ductile iron. (d) Hexbin plot of measured wall thickness versus pipe age (all materials combined), with a global linear fit and binned medians indicating a wide scatter and only a weak age–thickness trend.

4.7.4.3.2 Data Uncertainties

Operational choices during instrument deployment can directly affect data reliability. The most important source of error is sensor spacing. When the distance between the two acoustic sensors is off by more than ~2%, the derived percent-loss can be biased. Field crews therefore prefer potholes or in-line valves so they can measure point-to-point distance accurately, and they avoid long runs with multiple bends or large elevation changes between sensors. Candidate insertion points were screened for off-road access, minimal traffic control, environmental constraints, and bracket length. Anonymized examples are shown in Figure 4-16.

In addition, nominal manufacturing tolerances in wall thickness (on the order of 5–10%) and variability in elastic modulus can shift the inferred thickness. Unaccounted repairs or mixed materials inside the instrumented span can also distort propagation (for example, a short ductile-iron insert in cast iron can inflate apparent thickness by ~3.5%, while a PVC insert can depress it by ~41%). Where such conditions are detected, segments are either filtered from analysis or reconciled with targeted review. Finally, signal quality can degrade when appurtenance internals are worn, when large air pockets are present,

or when heavy tuberculation attenuates the wave. These issues are especially common in older, unlined CI pipes.

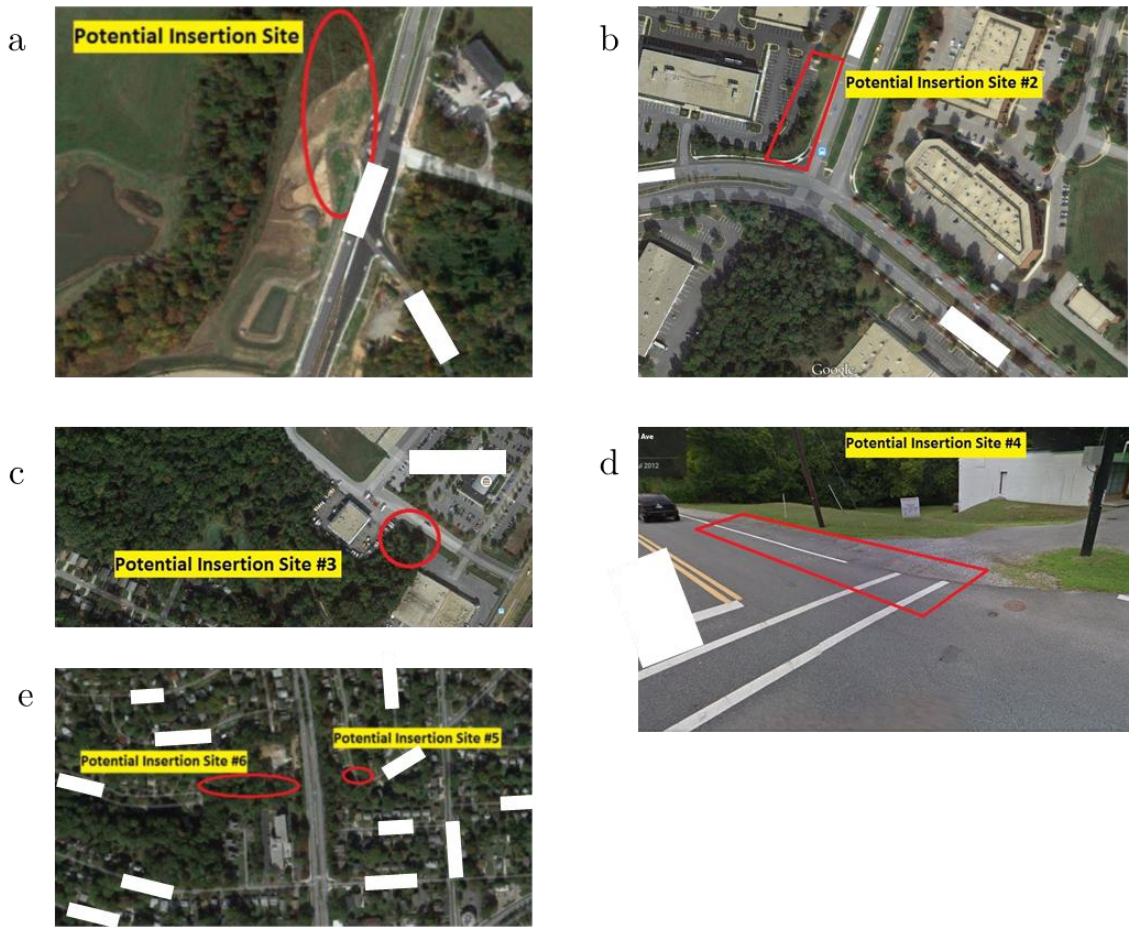

Figure 4-16: Anonymized examples of Potential Insertion Sites (PIS) used to plan acoustic condition assessments. Panels show typical contexts and constraints: (a) off-road grassy verge suitable for excavation without traffic impacts; (b) parking-lot/roadside verge with long accessible bracket; (c) off-road corner/drive apron for bidirectional reach; (d) roadway shoulder location requiring traffic control but enabling a >3,000-ft bracket; (e) paired residential verges along an arterial (two sites) illustrating reach from multiple access points. Sites were prioritized for safe access, limited traffic disruption, distance from protected areas, and sufficient bracket lengths (≈1,300–3,500 ft). Street names and landmarks redacted using white boxes for anonymity.

The program recorded two types of raw signals. For acoustics, dual-sensor time-series waveforms were collected. For pressure transient logging, a background recording without introduced noise established baseline conditions. Sensors were mounted directly on the pipe or coupled via hydrants/valves using hydrophones, and time delays between channels were used for cross-correlation and leak/noise path characterization. For pressure transients, hydrant-mounted loggers ran a dual-track regime. A 10 Hz background sampling (saved as one-minute means when no event occurred) and event-triggered capture at 1000 Hz for ~10–30 s whenever a rapid ≥2.1 psi change over 0.1 s was detected, with a pre-trigger buffer.

Acoustic average-thickness testing brackets a segment between two contact points (valves, hydrants, or temporary access). An out-of-bracket noise source excites the water column. With precise sensor spacing d and measured travel time t, the acoustic wave velocity v=d/t is computed. Using calibrated water properties and pipe parameters, this velocity is inverted to segment-average residual wall thickness (or %-loss for DI), which the provider converts into a 1–5 integrity rating (IR). Field teams also used passive leak correlation to check background noise and confirm path quality, and performed targeted pressure-transient logging to contextualize operational stress. The thickness model

depends on water bulk modulus and pipe elasticity. Field specialists calibrate the bulk modulus at the utility's source by testing a known condition (control sample) that should be a recently installed pipe segment. Figure 4-17 shows exhumed cast-iron pipe coupons spanning low to severe wall-loss. These specimens illustrate how tuberculation and pitting present across diameters and how this morphology can relate to the averaged wall-loss metrics used for ground truth (LOF_{GT}). These photos are illustrative to show that the LOF_{GT} labels used in validation come from instrumented segment-average measurements rather than visual inspection.

Figure 4-17: Representative cast-iron pipe coupons with their measured wall losses. (a) shows localized tuberculation with wall largely intact); (b) shows heavy corrosion with perforations and

extensive section loss); (c) shows moderate but patchy pitting) and (d) shows distributed pitting on a trunk main)

4.7.4.3.3 Data Processing

Ground truth (LOF_{GT} in 5 classes) is computed from raw data based on the technology provider's conversion table. For cast iron/steel/spun/pit cast we use the provider's 1–5 "Integrity Rating" (IR) that summarizes segment-average wall condition from acoustic thickness inference (see Table 4-13) and for ductile iron we use percent wall-thickness loss bands (see Table 4-14). We then convert IR or %-loss to the LOF_{GT} 0–4 class targets used for validation.

Table 4-13: Percent wall thickness lost conversion to LOF_{GT}. Used Primary for Ductile Iron (DI). Also used as a fallback for any cohort if %Loss is available and IR is not.

Range	**LOF_{GT}**	**Class label**
%Loss < 10	0	Very Low
$10 \leq$ %Loss < 20	1	Low
$20 \leq$ %Loss < 30	2	Moderate
$30 \leq$ %Loss < 40	3	High
%Loss ≥ 40	4	Very High

Computation of %Loss (when only thicknesses are present):

- If original and measured thickness are available:

$$\%\ Loss = \frac{\max\left(0, (Original\ Thickness - Measured\ Thickness)\right)}{Original\ Thickness} \times 100$$

- If "Lost Thickness" is provided instead of Measured Thickness:

$$\%\ Loss = \frac{Lost\ Thickness}{Original\ Thickness} \times 100$$

Table 4-14: Integrity Rating (IR) conversion to LOF_{GT}. Used for Pit CI, Spun CI and Steel

Range	$\mathbf{LOF_{GT}}$	Class label	Inclusion rule
IR < 2.0	0	Very Low	IR < 2.0
2.0 ≤ IR < 3.0	1	Low	lower bound inclusive
3.0 ≤ IR < 4.0	2	Moderate	lower bound inclusive
4.0 ≤ IR ≤ 4.5	3	High	both bounds inclusive (≤ 4.5)
IR > 4.5	4	Very High	strictly greater

Notes: Lower IR = better wall condition; higher IR = closer to (or below) "critical" wall thickness; these cutpoints are fixed a priori for validation: if IR < 2.0 → 0; else if < 3.0 → 1; else if < 4.0 → 2; else if ≤ 4.5 → 3; else → 4.

We validate the model only on cohorts for which the instrument physics and the utility's management priorities support reliable field targets. These are Pit cast iron (CI) in <8" and 8–24", Spun CI in <8" and 8–24", ductile iron (DI) in <8" and 8–24", steel (ST) in >24", and PCCP in 8–24" and >24". Mixed material tags are normalized to the first token, unknown materials and records with missing diameter band are excluded. Outcomes are reported at two granularities. First, an overall view across all validated segments and per-cohort views (material-diameter band). Because LOF is an ordinal 0–4 scale, we emphasize ordinal concordance that is, agreement in rank, even when predictions differ by one class.

4.7.4.3.4 Hypotheses Testing

Building on the cohort scope above, Table 4-15 summarizes the hypothesis framework we use to turn the visual outputs into statistical claims. We test: (i) independence via a Pearson χ^2 test on the confusion matrix (are predictions unrelated to ground truth?); (ii) agreement beyond chance using Cohen's κ and quadratic-weighted κ, which penalize larger ordinal gaps; (iii) monotone ordinal association with Spearman's ρ; and (iv) improvement over a naive baseline by comparing exact accuracy to the cohort's majority-class prevalence with Wilson 95% CIs. We reject a null when $p < 0.05$ (or the 95% CI excludes the null value). Decisions are reported alongside the overall and cohort-wise confusion matrices, accuracy heatmap, and agreement plots to substantiate ordinal concordance on the 0–4 LOF scale.

Table 4-15: Validation hypotheses and decision rules for LOF concordance. Each row states what is tested, the null hypothesis, the statistic used, the α=0.05 decision rule, and where the result appears in the notebook outputs. Tests are computed both overall and within each validated material–diameter cohort.

ID	Test	Null hypothesis H_0	Statistic / estimate	Decision rule $\alpha = 0$	Interpretation if H_0 is rejected
H_0- χ^2	Are predictions independent of ground truth?	$LOF_{student}$ is independent of LOF_{GT}	Pearson χ^2 on confusion matrix (overall and per-cohort).	Reject if $p_{\chi^2} < 0.05$.	Predictions carry information about truth (non-random association).
H_0–κ_w	Is agreement beyond chance	Quadratic-weighted Cohen's ($\kappa_w = 0$).	Cohen's κ and κ_w (overall & per-cohort).	Reject if $\kappa_w > 0$ with 95% CI does not include 0.	Non-trivial agreement beyond chance;

ID	Test	Null hypothesis H_0	Statistic / estimate	Decision rule $\alpha = 0$	Interpretation if H_0 is rejected
	(ordinal-weighted)?				large gaps penalized more.
H_0- ρ	Is there monotone rank association?	Spearman $\rho = 0$.	Spearman ρ with p-value (overall).	Reject if $p_\rho < 0.05$	Student ranks covary with ground-truth ranks.
H_0–base	Better than majority-class guessing?	Exact accuracy ≤ cohort majority prevalence.	Exact accuracy vs. majority prevalence; Wilson 95% CI for accuracy.	Reject if accuracy > majority prevalence and its 95% CI stays above that baseline.	Model beats naive baseline for that cohort.

4.7.4.3.5 Results

Overall performance: Across 708 segments, exact accuracy is 84%, with within-one-class agreement = 97% (596 exact, 93 near-miss and only 19 cases are ≥2 classes off). This is shown in Figure 4-18 a).

The macro-F1 = 0.79 and weighted-F1 = 0.85 reflect good class-wise balance despite class 0 being most prevalent (support: 449). Agreement beyond chance is $\kappa = 0.73$ and quadratic-weighted $\kappa = 0.87$ (which shows substantial to almost-perfect). Here, the ordinal association is strong. Spearman $\rho = 0.85$, $p \approx 2.8 \times 10^{-194}$. A χ^2 test of independence on the 5×5 table yields $\chi^2 = 1605.9$, $df = 16$, $p < 10^{-12}$, rejecting independence decisively. The raw accuracy confusion matrix Figure 4-18 b) and row-normalized confusion matrix Figure 4-18 a) confusion matrices show a strong diagonal. Diagonal

conditional accuracies by truth class are ≈0.80–0.87 (0:0.85, 1:0.81, 2:0.80, 3:0.87, 4:0.85). The stacked "agreement-by-truth" view in Figure 4-18 e) indicates small, largely symmetric slips to adjacent classes.

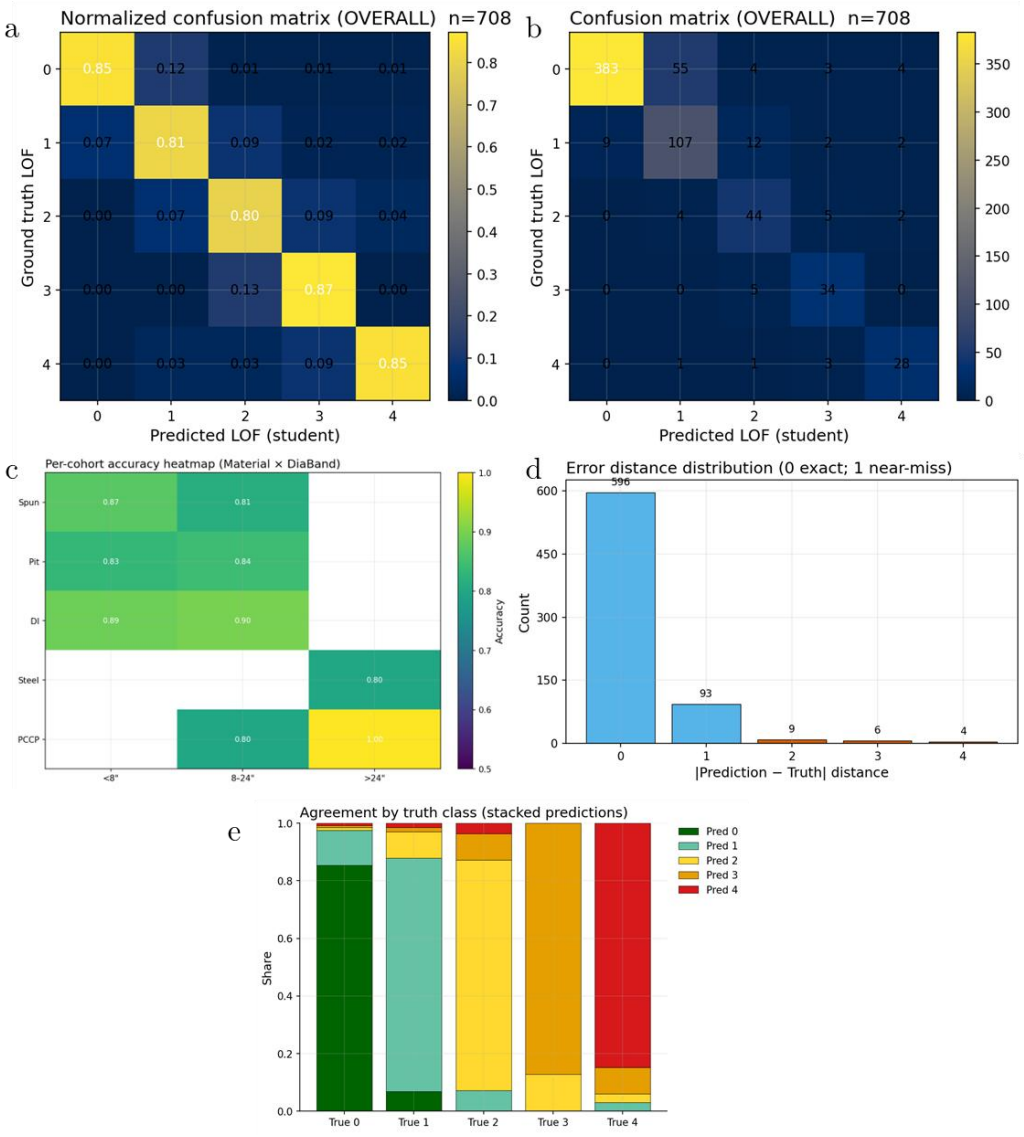

Figure 4-18: Model–ground-truth concordance for the validated cohorts only. Pit CI (<8", 8–24"), Spun CI (<8", 8–24"), DI (<8", 8–24"), Steel (>24"), and PCCP (8–24", >24")—based on

instrument-anchored LOF_GT (n=708 segments). Panels: (a) row-normalized confusion matrix showing a strong diagonal; (b) raw-count confusion matrix; (c) per-cohort accuracy heatmap (Material-Diameter); (d) error-distance histogram concentrated at 0–1 class; (e) stacked predictions by truth class. Overall exact accuracy = 0.842, within-one = 0.973, quadratic κ = 0.867, Spearman ρ = 0.845.

Per-cohort patterns: The accuracy heatmap in Figure 4-18 c) and Figure 4-18 b) show strong performance in all substantive cohorts (n≥30). These statistics are presented in Table 4-16.

Table 4-16: Per cohort summary of accuracy

Material	Diameter	n	Accuracy	balanced_acc	kappa	kappa_quadratic
Spun	8–24"	278	0.81	0.78	0.52	0.72
Spun	<8"	102	0.87	0.86	0.76	0.87
Pit	8–24"	173	0.84	0.85	0.78	0.89
Pit	<8"	36	0.83	0.80	0.77	0.82
DI	8–24"	80	0.90	0.95	0.86	0.96
DI	<8"	18	0.89	0.63	0.71	0.92
Steel	>24"	15	0.80	0.80	0.00	0.00
PCCP	8–24"	5	0.80	0.80	0.00	0.00
PCCP	>24"	1	1.00	1.00	NaN	NaN

Accuracy: Fraction of segments where the predicted LOF class exactly equals the ground-truth class;
Balanced Accuracy: Mean recall taken across classes; treats each LOF class equally so results aren't dominated by common classes;
Kappa: Cohen's κ (unweighted): agreement beyond chance; all misclassifications are penalized equally, regardless of how far apart the classes are;
Kappa Quadratic: Quadratic-weighted κ: agreement beyond chance with heavier penalties for larger class gaps; ideal for ordinal scales like LOF 0–4.

Spun <8" Accuracy=87% ($\kappa_w = 0.87$), Spun 8–24" Accuracy=81% ($\kappa_w = 0.72$), Pit <8" Accuracy=83% ($\kappa_w = 0.82$), Pit 8–24" Accuracy=84% ($\kappa_w = 0.89$), DI <8" Accuracy=89% ($\kappa_w = 0.92$), DI 8–24" Accuracy=90% ($\kappa_w = 0.96$). Small-n cohorts (Steel >24", n=15; PCCP 8–24", n=5; PCCP >24", n=1) show high apparent accuracies (0.80–

1.00) but $\kappa \approx 0$ or undefined, reflecting label degeneracy and wide uncertainty. We treat these as inconclusive for hypothesis testing and report them for completeness only. The final summary from hypothesis testing is shown in Table 4-17.

Table 4-17: Hypothesis testing results and decision

ID	Null hypothesis (H_0)	Test / estimate	Evidence	Decision
H_0-χ^2	Predictions have no association with truth	Pearson χ^2 on overall confusion matrix	χ^2=1605.88, df=16, $p \approx 0$	Reject H_0
H_0–κ_w	Quadratic-weighted $\kappa \leq 0.20$ (≤slight agreement)	Cohen's κ_w (overall & per-cohort)	$\kappa_w = 0.87$ (well >0.20)	Reject H_0
H_0-ρ	Spearman rank $\rho = 0$	Spearman ρ with p-value (overall)	$\rho = 0.85$, $p \approx 2.8 \times 10^{-194}$	Reject H_0
H_0-base	Accuracy ≤ majority-class baseline	Exact accuracy vs majority prevalence (overall & per-cohort)	Overall: baseline=0.634 (449/708); accuracy=0.842 with 95% CI [0.813, 0.867] → above baseline. Major cohorts (n≥30): acc 0.813–0.900, balanced-acc 0.778–0.952, each above its cohort baseline.	Reject H_0 (overall & major cohorts); inconclusive for small-n cohorts

Notes. "Majority-class baseline" is the accuracy by always predicting the modal class in that cohort; balanced accuracy >0.5 across all substantive cohorts further indicates real signal rather than class-imbalance artifacts. For small-n bands we avoid over-interpretation and recommend targeted data collection.

This validation shows the model is learning real deterioration patterns that the field instruments capture, not just echoing the majority class. Most disagreements are small (adjacent LOF bands), which is what matters operationally when making renewal decisions. Performance is consistently strong across the key material–diameter cohorts we tested, while tiny cohorts remain informational only. Overall, the model is decision-useful

for the targeted cohorts and highlights where additional data or inspections would further tighten confidence.

4.7.4.4 Ground truth agreement for PCCP using Electromagnetic Wirebreaks count

The previous section validated the LOF model against RWT for cast iron and ductile iron cohorts. RWT is a natural ground truth for metallic pipes. Progressive metal loss links directly to pressure capacity and provides a clean structural proxy for failure risk. For PCCP, the failure mechanics are different. PCCP relies on a composite action between the concrete core, steel cylinder, and prestressing wire. Several competing failure modes exist. In practice, however, most catastrophic PCCP failures begin with loss of prestress as the primary initiating mechanism. Once enough wires break, the hoop stress shifts to the steel cylinder and concrete core. This accelerates cracking, local yielding, and ultimately through-wall rupture. RWT on the steel cylinder is still relevant, but it is often a downstream symptom of an earlier wire-break process.

For that reason, Electromagnetic (EM) estimates of broken prestressing wire wraps are, for PCCP, a more direct measure of structural distress than cylinder wall loss alone. They are also what many utilities already use as the main decision input for PCCP

management (monitor, repair, replace). To show that the LOF model is credible for PCCP, it is not sufficient to match RWT behavior on metallic pipes. We also need to demonstrate that, for large-diameter PCCP, predicted LOF classes track the wire-break hierarchy that practitioners act on.

This section therefore provides a second, independent validation experiment for the PCCP $>$ 24" cohort, using EM wire-break counts as ground truth. Together, the RWT-based and wire-break-based experiments show that the model is aligned with two physically distinct, but complementary, structural mechanisms.

4.7.4.4.1 Data Collection

The ground-truth dataset for this experiment comes from a wholesale water provider in the southern United States that has implemented a multi-year EM inspection program for about 4 miles of large-diameter PCCP transmission mains. The utility has inspected several lines with nominal diameters between 72 and 108 inches using EM tools deployed inside the pipe, and for each inspected segment the vendor reports an estimated number of broken prestressing wire wraps together with spatial locations and quality flags. The validation dataset analyzed here is a curated subset of that broader program. It includes only PCCP segments with diameter greater than 24 inches (primarily in the 72–

108 inch range), only segments for which the vendor produced a clear pipe-level estimate of total broken wraps, only segments that were unrepaired at the time of inspection so that the wire-break count reflects accumulated deterioration rather than post-repair residuals, and it excludes pipes for which the vendor classified the EM response as anomalous or non-diagnostic.

To describe the instrumentation and signal interpretation, we draw on representative inspection reports shared by the water utility, which documents the whole process in detail. The utility's EM inspections are carried out by a commercial vendor using an internal EM tool. The tool is inserted into the main, traverses each pipe barrel from bell to spigot, and carries both excitation and sensing hardware. The EM system works in three conceptual steps:

1. Excitation: A low-frequency alternating current is driven through an exciter coil near the inner surface of the pipe. This produces a slowly varying magnetic field that penetrates the concrete core and links the steel cylinder and prestressing wires. The field induces eddy currents in each prestressing wire wrap.
2. Perturbation by broken wires: In intact wire wraps, induced currents circulate continuously around the pipe, producing a characteristic magnetic signature. Where a wire

has fractured, this current path is interrupted. The local field becomes weaker or distorted. As the tool moves along the barrel, regions with intact wraps and regions with clusters of broken wraps give measurably different responses.

3. Sensing and inversion: Detectors on the tool record the spatial variation of the magnetic field as a function of travel distance. Analysts then process this signal to identify "distressed regions" (zones where the signature differs from the intact baseline) and estimate the number of broken wraps in each region. This conversion relies on calibration curves developed from full-scale tests on similar pipes, where known numbers of wraps are cut and rescanned.

Exploratory plots of the PCCP inspection data (Figure 4-19) show how wire-break distress is distributed across diameter and soil environments. Most pipe samples fall into a small set of large-diameter classes, and only a minority of segments contain non-zero wire-break counts, consistent with the expectation that structural distress remains rare relative to the network length. When wire-break counts are plotted against the SSURGO steel-corrosivity index, severe distress is increasingly concentrated in soils classified as more aggressive to steel. This is mechanically plausible. In PCCP, the prestressing wires are the primary tensile reinforcement, and the steel-corrosivity index is largely driven by

soil moisture, resistivity, and chloride/sulfate environments that directly influence wire corrosion rates.

In contrast, the concrete-corrosivity index does not show a simple monotone relationship with wire-breaks. Concrete corrosivity reflects chemical risks to the mortar matrix (e.g., sulfate attack) rather than the electrochemical conditions at the steel surface, and the steel is partially shielded by the concrete cover and by local construction details. As a result, the concrete index alone is a weaker predictor of wire distress than the steel-focused index. Aggregating length by wire-break ground-truth class ($\mathrm{LOF}_{\mathrm{GT}}$) further highlights that the dataset is strongly imbalanced. Almost all inspected mileage lies in class 0, with only a few miles in higher distress classes. This pattern is typical of real world PCCP transmission system performance and motivates the use of ordinal agreement and rank-based metrics, rather than raw accuracy alone, in the subsequent validation.

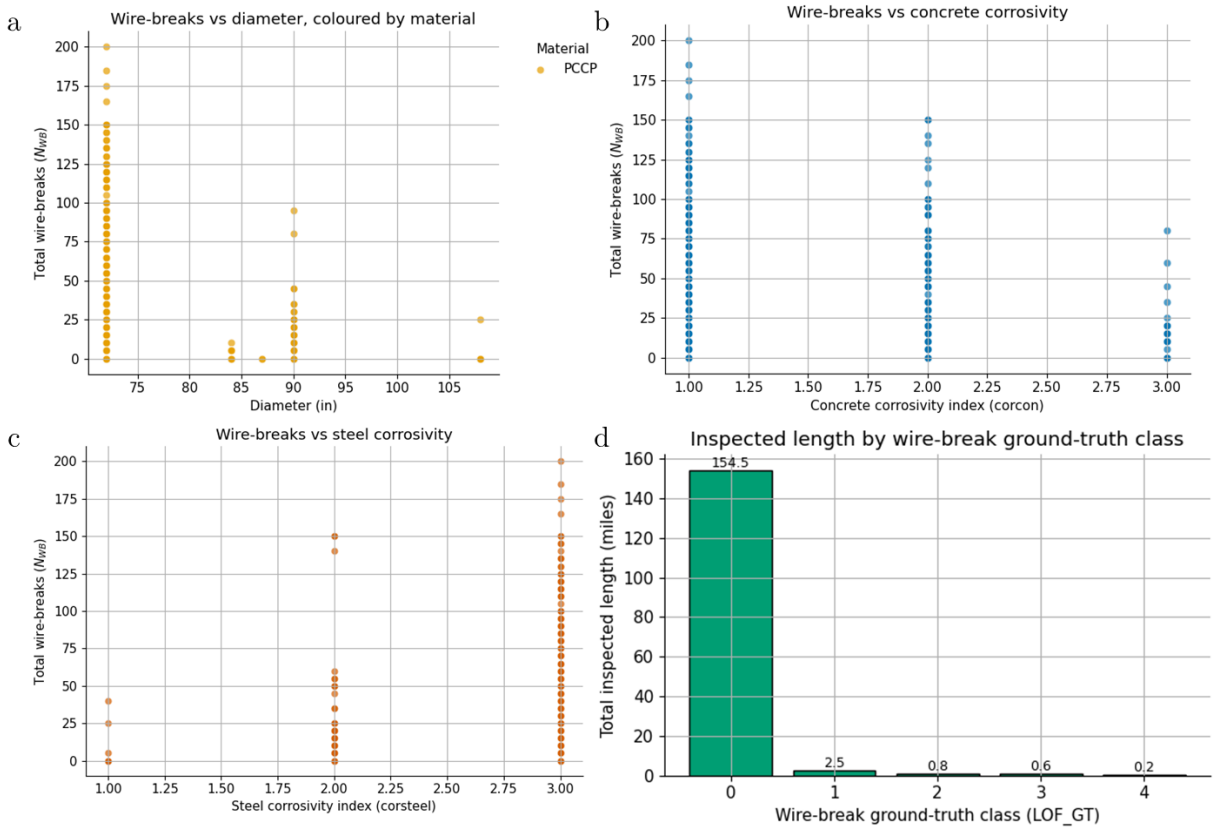

Figure 4-19: Exploratory characterization of the PCCP wire-break dataset. (a) Total EM-estimated wire-breaks per pipe as a function of diameter, showing that high wire-break counts are rare outliers as diameter increases. (b) Wire-breaks versus SSURGO concrete-corrosivity index, where no clear monotone trend is evident. (c) Wire-breaks versus SSURGO steel-corrosivity index, where higher wire-break counts are increasingly concentrated in soils mapped as more aggressive to steel. (d) Total mileage in the dataset by wire-break ground-truth class (LOF_{GT}), highlighting that most samples lies in class 0 and only a small fraction of the network occupies higher distress bands.

4.7.4.4.2 Data Uncertainties

Since EM response depends strongly on pipe design (wire size and spacing, cylinder thickness, presence of shorting straps, pressure class), the vendor develops design-specific calibration curves. For each PCCP design family, full-length calibration pipes from the same or similar system are instrumented, scanned intact, and then progressively damaged by cutting known numbers of wire wraps at known locations. By comparing the change in magnetic signature after each damage step, analysts derive a relation between anomaly amplitude/shape and the number of broken wraps. These calibration curves are then embedded in the EM analysis software and used to interpret field data from in-service pipes.

For each inspected pipe segment in the utility's system, the EM analysis yields a structured record that links directly back to the asset inventory. Every segment has a unique pipe identifier tied to the utility's GIS, basic geometry and design attributes (diameter, class, pipe type), and zero, one, or multiple distressed regions along the barrel. For each distressed region, the vendor reports an estimated number of broken wraps, and these are aggregated to a pipe-level total number of broken wraps. The record also includes flags and notes on signal quality, anomalous behavior, or prior repairs. This pipe-level

dataset is the raw input from which the wire-break ground-truth index classes are constructed.

EM-based wire-break counts are powerful but not perfect, and several practical limitations shape how they can be used as ground truth. For embedded-cylinder PCCP with shorting straps, isolated single broken wraps are often below the reliable detection threshold. In practice, roughly five or more consecutive broken wraps are needed before the EM signature becomes robust enough for confident classification, so pipes with very small numbers of broken wraps may be reported as "no distress" or carry high uncertainty. Detection also depends on position. Broken wraps near bell or spigot joints, or close to appurtenances, are harder to detect than mid-barrel damage because joint geometry and local changes in steel mass distort the baseline signal, and short specials are particularly difficult because joint effects occupy most of the barrel. External interference further complicates interpretation. Unknown or poorly documented appurtenances, nearby ferromagnetic structures, or atypical construction details can produce "anomalous" signatures that do not match typical distress patterns, and the vendor flags such pipes as anomalous or non-diagnostic, with wire-break counts that are either missing or highly uncertain. At the opposite extreme, pipes where most or all wires have broken can produce saturated

signals that are difficult to distinguish from other property changes. Analysts may classify these as having broken wraps across most or all of the pipe rather than assigning a precise count, which clearly indicates very high distress but with only approximate wrap numbers.

4.7.4.4.3 Data Processing

To construct a clean ground-truth cohort from this reality, we restrict attention to PCCP segments with diameter greater than 24 inches (dataset has pipes from 72–108 inches) that belong to the vendor's calibrated design families, and that sit within the EM tool's reliable interpretation envelope. We exclude pipes flagged as anomalous, non-diagnostic, or otherwise outside the stated conditions for trustworthy wire-break quantification, including very short specials where the vendor has explicitly cautioned against use of the counts. We also exclude pipes known to have been structurally repaired before inspection, because their wire-break pattern reflects both historical distress and intervention. Within the remaining set, we require a non-negative total number of broken wraps for each pipe segment, either as an explicit count or as an "across entire pipe" classification. The resulting subset is not a census of all PCCP segments in the system, but it is the subset for which EM wire-break counts can reasonably be treated as ground-truth indicators of structural distress for validation.

To compare the EM wire-break data with the model's five-band LOF output, we construct a pipe-level wire-break ground-truth index, $LOF_{WB\_GT} \in \{0,1,2,3,4\}$, that is monotone in the total number of broken wraps and aligned with vendor reporting practice. For each inspected pipe, we first aggregate distress by summing the estimated broken wraps across all distressed regions to obtain a total broken-wrap count N_{WB}. Pipes with no reported distress are assigned $N_{\mathrm{WB}} = 0$, while pipes that the vendor describes as having broken wraps "across most or all of the pipe" are treated as having a very large effective N_{WB}, even if only a qualitative range is given. Based on typical vendor groupings (for example, "<25 broken wraps", "25–50", ">50", plus a category for across-barrel distress) and on engineering judgement about severity, we then define a set of thresholds. Zero broken wraps correspond to no detectable distress above the EM detection limit; 5–24 broken wraps indicate low but clear distress; 25–49 broken wraps indicate moderate distress; 50–99 broken wraps indicate high distress; and 100 or more broken wraps, or an across-barrel classification, indicates very high distress.

Using these thresholds, each pipe is assigned to one of five classes. Class 0 ($LOF_{WB\_GT} = 0$) corresponds to $N_{\mathrm{WB}} = 0$ with no indication of across-pipe distress; this should be interpreted as "no EM-detectable cluster of broken wraps", not as a guarantee

that every wire is intact. Class 1 ($LOF_{WB\_GT} = 1$) corresponds to $5 \leq N_{\mathrm{WB}} \leq 24$; these pipes have a limited number of broken wraps, usually confined to one or two localized regions, and are typically candidates for continued monitoring rather than immediate structural intervention. Class 2 ($LOF_{WB\_GT} = 2$) corresponds to $25 \leq N_{\mathrm{WB}} \leq 49$; these pipes show more extensive wire loss and reduced margin to critical conditions, and utilities often plan targeted repair or replacement within a defined planning horizon. Class 3 ($LOF_{WB\_GT} = 3$) corresponds to $50 \leq N_{\mathrm{WB}} \leq 99$; these pipes exhibit heavy wire-break concentrations and are typically treated as high priority for near-term intervention because of substantial stress redistribution to the steel cylinder and concrete core. Class 4 ($LOF_{WB\_GT} = 4$) corresponds to $N_{\mathrm{WB}} \geq 100$, or any case where the vendor indicates that broken wraps extend across most or all the pipe barrels. Such pipes are near structural limit states and are usually considered for immediate action or emergency contingency planning.

Wire-break counts below 5 are treated with caution. If the vendor explicitly reports $1 \leq N_{\mathrm{WB}} \leq 4$ for a pipe, we either retain the pipe but map it to Class 0, recognizing that such small counts lie near the EM detection threshold, or flag it for sensitivity analysis rather than including it in the core ground-truth set. This avoids over-interpreting EM

estimates at the edge of the method's reliability. The resulting LOF_{WB_GT} index is monotone in total wire-break count, directly grounded in the same EM outputs that practitioners already use, coarse enough to be robust to interpretation uncertainty, and expressed on the same 0–4 scale as the model's LOF bands. In the next step, we join this wire-break ground-truth index to the model's predicted LOF bands for the PCCP $> 24''$ cohort and apply the same validation machinery used for the RWT experiment, including confusion matrices, distance-to-truth diagnostics, and formal hypothesis tests against independence, zero correlation, and majority-class baselines.

4.7.4.4.4 Hypotheses Testing

Table 4-18 summarizes the statistical hypotheses used to evaluate how well the student model reproduces the wire-break ground-truth classes.

Table 4-18: Statistical hypotheses for PCCP wire-break based $LOF_{student}$ model validation

ID	Test	Null hypothesis H_0	Statistic / estimate	Decision rule $\alpha = 0.05$	Interpretation if H_0 is rejected
H_0-χ^2	Are predictions independent of truth?	$LOF_{student}$ is independent of wirebreak ground truth LOF_{GT}	Pearson χ^2 on confusion matrix (overall PCCP wirebreak cohort).	Reject if $p_{\chi^2} < 0.05$.	Predictions carry information about the wire-break truth classes (non-random association).
H_0–κ_w	Is agreement beyond chance (ordinal)?	Quadratic-weighted Cohen's ($\kappa_w \leq 0$). (no agreement beyond chance).	Cohen's κ and κ_w (overall & per-cohort).	Reject if $\kappa_w > 0$ with 95% CI does not include 0.	Non-trivial ordinal agreement beyond chance; larger band errors are penalized more strongly.

ID	Test	Null hypothesis H_0	Statistic / estimate	Decision rule $\alpha = 0.05$	Interpretation if H_0 is rejected
H_0-ρ	Is there monotone rank association?	Spearman $\rho = 0$.	Spearman ρ with p-value (overall).	Reject if $p_\rho < 0.05$.	Student ranks co-vary with ground-truth ranks.
H_0–base	Better than majority-class guessing?	Exact accuracy ≤ cohort majority prevalence.	Exact accuracy versus majority-class prevalence; 95% CI for accuracy.	Reject if accuracy > majority prevalence and its 95% CI stays above that baseline.	Model outperforms a naive classifier that always predicts the most common wire-break class in this cohort.

Four tests were chosen to probe complementary aspects of performance. H_0- χ^2 asks whether predictions are associated with truth at all, using a Pearson χ^2 test on the full 5×5 confusion matrix. H_0– κ focuses on *ordinal* agreement beyond chance, via quadratic-weighted Cohen's κ, which penalizes larger band errors more strongly than near-misses. H_0- ρ tests whether the model preserves the rank ordering of distress using Spearman's ρ, which is robust to non-linear but monotone relationships. Finally, H_0–base compares the model with a trivial majority-class classifier, recognizing that, under extreme class imbalance, a naive "all-safe" strategy can achieve very high apparent accuracy; this test therefore interrogates whether the model genuinely improves on that baseline.

4.7.4.4.5 Results

The summary metrics in Table 4-19 show that the PCCP $LOF_{student}$ model reproduces the five-band wire-break ground truth with high fidelity while preserving the ordinal structure of the index. Exact class accuracy is 84%, and the within-one-band accuracy is

98%, so almost all disagreements occur between neighboring bands. The quadratic-weighted kappa of about 0.32 indicates fair-to-moderate agreement beyond chance once larger band errors are penalized more heavily than near-misses. This is consistent with the intended use of the index where bands are designed to represent graded concerns for structural/functional failures, so a one-band slip between, for example, "low" and "moderate" wire-break distress is much less serious than a two-band slip. The positive Spearman rank correlation ($\rho \approx 0.33$) confirms that the model respects this ordering that is, pipes with more severe wire-break classes tend to receive higher predicted LOF bands.

Table 4-19: Overall validation metrics for PCCP wire-break ground truth (LOF_{GT}) versus model predictions ($LOF_{student}$)

Metric	Symbol	Estimate (overall)	95% CI	Interpretation
Accuracy (exact class agreement)	Acc	0.84	[0.83, 0.84]	About 84% of pipes are assigned to the correct wire-break class.
Within-one-band accuracy	Acc $\{\pm 1\}$	0.98	[0.98, 0.99]	Almost all remaining errors are at most one LOF band away from ground truth.
Quadratic-weighted Cohen's kappa	κ_w	0.32	[0.29, 0.34]	Fair-to-moderate ordinal agreement between predicted LOF bands and wire-break ground truth.
Spearman rank correlation	ρ	0.33	[0.31, 0.34]	Predicted LOF ranks increase coherently with wire-break severity, but the association is moderate-high
Chi-square association test	χ^2 (df = 16)	30223	–	Very strong departure from independence between predictions and wire-break classes.
Majority-class accuracy baseline	Acc_{base}	0.98	–	A trivial classifier that labels all pipes as LOF = 0 achieves 97.7% accuracy.
Accuracy gain over baseline	Acc − Acc_{base}	−0.142	[−0.146, −0.138]	The model trades about 14 percentage points of "all-safe" accuracy to reclassify risky pipes.

The chi-square test reinforces this picture. With $\chi^2 \approx 3\times10^4$ on 16 degrees of freedom and $p \ll 0.001$, the joint distribution of predicted and observed classes is far from what would be expected under independence. At the same time, the accuracy of a naive majority-class classifier is extremely high (97.7%) because almost all inspected footage lies in the lowest wire-break class (0). The model's accuracy is therefore lower than this trivial baseline, and H_0–base is not rejected. This is not a defect of the model. Rather, it reflects the design choice to reclassify a small fraction of pipes into higher-risk bands instead of labelling everything as "safe." In a renewal-planning context, these modest losses in overall accuracy are acceptable when they are accompanied by better discrimination of the rare, distressed pipes, which are captured more appropriately by κ_w and ρ than by raw accuracy alone.

Figure 4-20 provides a more granular view of how the model behaves across the five LOF bands. The normalized confusion matrix (panel a) is strongly diagonal. For each ground-truth class, between about 0.83-0.88 of segments are predicted in the correct band. Off-diagonal mass is concentrated in the immediately adjacent cells, so most misclassifications are one-band shifts such as predicting class 1 instead of class 0 or class 3 instead of class 4. The unnormalized confusion matrix (panel b) shows the same pattern on the

count scale and makes the class imbalance explicit where most pipes are in the lowest wire-break class, with far fewer observations in classes 1–4. The error-distance histogram (panel c) summarizes this behavior numerically. Exact agreement (distance 0) accounts for roughly 84% of pipes. Near-misses one band away (distance 1) account for almost all the remaining cases, while only a very small number of pipes fall two or more bands away from the truth. This distribution is what drives the high within-one accuracy and moderate κ_w reported in Table 4-19. It also reflects a common pattern in ordinal models where errors tend to "blur" the boundaries between neighboring categories rather than jump across multiple structural regimes. The stacked agreement plot by truth class (panel d) shows how the predicted class distribution shifts as the wire-break severity increases. For pipes with $\mathrm{LOF}_{\mathrm{GT}} = 0$, the stack is dominated by predictions in class 0, with a small fraction upgraded into class 1. For $\mathrm{LOF}_{\mathrm{GT}} = 1$ and 2, the stacks become centered on the corresponding predicted bands, with thinner contributions from the bands immediately below and above. For $\mathrm{LOF}_{\mathrm{GT}} = 3$ and 4, most predictions lie in classes 3 and 4 respectively, but there is a non-negligible share in the next lower band, reflecting a cautious tendency to slightly under-state the most extreme cases.

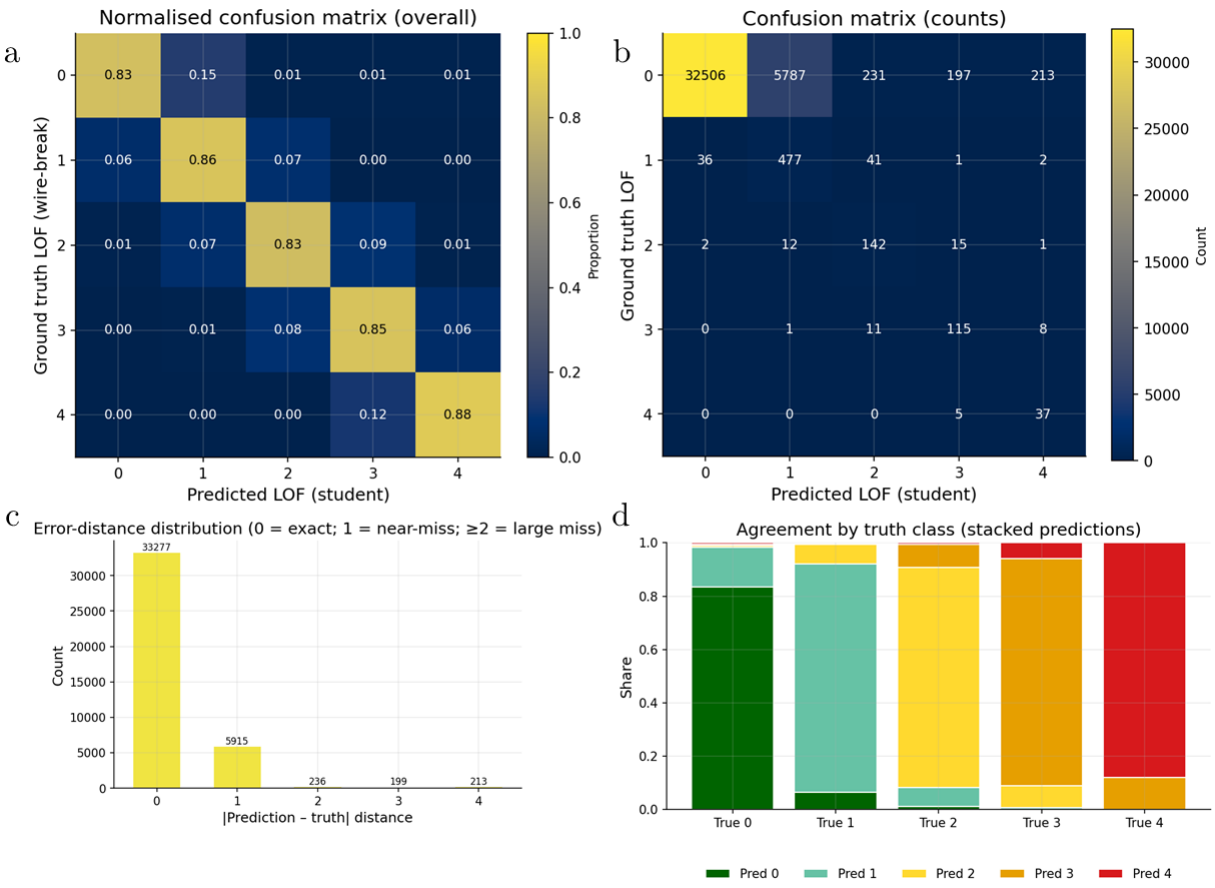

Figure 4-20: Agreement between modelled LOF and wire-break ground truth for PCCP. *(a) Normalized confusion matrix showing that most pipes are predicted in the correct wire-break class, with misclassifications concentrated in adjacent bands. (b) Raw confusion matrix highlighting the strong class imbalance toward low-distress segments. (c) Distribution of absolute prediction–truth distance, indicating that almost all errors are one band or less. (d) Stacked prediction shares by truth class, illustrating that predicted LOF bands shift upward systematically with increasing wire-break severity while rarely crossing multiple bands.*

Taken together, these patterns show that the model behaves as a smooth, ordinal risk ranking: it rarely confuses "healthy" pipes with severely distressed ones or vice versa, but it does allow some uncertainty at the band boundaries, which is consistent with the noisy and spatially local nature of EM wire-break estimates.

Table 4-20 reports the outcomes of these hypothesis tests for the PCCP wire-break cohort. The χ^2, κ_w and Spearman ρ results all lead to rejection of their respective null hypotheses, showing that the model's predictions are strongly associated with the wire-break ground truth, exhibit moderate ordinal agreement beyond chance, and respect the intended ranking of distress.

Table 4-20: Outcomes of hypothesis tests for PCCP wire-break validation

ID	Null hypothesis (H_0)	Test / estimate	Evidence	Decision
H_0–χ^2	Predictions have no association with wire-break truth	Pearson χ^2 on overall 5×5 confusion matrix	$\chi^2 = 46\,614.55$, df = 16, p ≈ 0 (≪ 0.001).	Reject H_0. Strong dependence between $LOF_{student}$ and LOF_{GT}.
H_0–κ	Ordinal agreement is no better than chance ($\kappa_w \leq 0.2$)	Quadratic-weighted Cohen's κ_w	κ_w= 0.457 with 95% CI [0.429, 0.486]; κ_w lies well above 0.	Reject H_0. Moderate agreement beyond chance; large mis-banding is uncommon.
H_0–ρ	No monotone rank association ($\rho = 0$)	Spearman ρ with p-value	$\rho = 0.424$ with 95% CI [0.410, 0.438]; p ≈ 0 (≪ 0.001).	Reject H_0. Higher predicted LOF bands correspond systematically to higher wire-break classes.
H_0–base	Accuracy ≤ majority-class baseline	Exact accuracy versus majority prevalence	Majority baseline $Acc_{base} = 0.977$ (all $LoF_{GT} = 0$). Model accuracy Acc = 0.906 with 95% CI [0.903, 0.909]. Accuracy gain Acc − Acc_{base} = −0.071 with 95% CI [−0.074, −0.068] (entirely below zero).	Do not reject H_0. The model is less accurate than the trivial majority classifier, as expected under extreme class imbalance.

In contrast, H_0–base is not rejected. The model's exact accuracy remains below the majority-class baseline because a classifier that labels every pipe as low-risk already achieves 97.7% accuracy in this extremely imbalanced dataset. This pattern is consistent with the design goal of trading a small amount of "all-safe" accuracy for an improved

identification and ranking of the relatively few distressed pipes, which is captured better by κ_w and ρ than by accuracy alone.

4.7.4.5 Ground truth agreement for other materials (PVC, PE, AC)

This section presents the validation of the $\mathrm{LOF}_{\mathrm{student}}$ index in a retrospective failure-prediction setting for non-metallic distribution mains, specifically AC, PVC and PE. For these materials, utilities rarely have high-resolution condition data because inspection technologies are less mature and destructive sampling is harder to justify. The most complete deterioration signals are usually operational: carefully logged histories of leaks and breaks in work orders and main-break reports. We therefore use a five-year work-order dataset from a large U.S. utility that is widely recognized for its early adoption of GIS-based asset management and rigorous verification of break records. This utility and its data were deliberately excluded from all stages of model development and from the earlier Evaluation Verification and Validation (EVV) exercises, so the analysis here is an out-of-sample, temporal test of predictive validity rather than a re-use of training information. We take a snapshot of the AC, PVC, and PE asset base as of 1 January 2009, apply the model to assign five-band LOF classes (0–4) at the pipe-segment level, and then track which segments experience at least one recorded failure between 2009 and 2013. This

emulates how a utility would deploy the index in practice: prioritizing mains for inspection or renewal based on current LOF bands and then observing which segments fail under real operating and environmental conditions. By combining independently curated failure histories with SSURGO-derived environmental attributes, and by evaluating mileage-based recall of failed segments, five-year precision of high-risk classifications, and the trade-off between missed failures and "false-positive" high-risk pipes at the material–diameter cohort level, this study tests both the internal coherence of the LOF index for AC, PVC, and PE and its external utility as a forward-looking decision aid for non-metallic mains that fail through mechanisms different from metallic corrosion.

4.7.4.5.1 Data Collection

The retrospective failure dataset comes from a large U.S. water utility in a southern state that has invested heavily in systematic main break recording and proactive renewal planning. Since the late 1980s the utility has maintained electronic asset and maintenance records for its distribution system, including pipe material, diameter, segment length, installation year, and detailed work orders for leaks and breaks. These records support a long-running main replacement and rehabilitation program that reinvests in ageing mains each year and uses break frequencies, corrosion flags, and service histories to prioritize

projects. As-built drawings and installation records are retained permanently, and more than three decades of work orders have been digitized, geo-referenced, and routinely used in internal performance reporting and external research case studies. Regular internal audits and reconciliation of work orders against field reports, customer calls, and GIS mapping have produced one of the most carefully curated failure datasets available for non-metallic mains, even though it is still less controlled than the instrumented condition-assessment campaigns used in the previous validation studies. Figure 4-21 summarizes the non-metallic asset base used in this retrospective test.

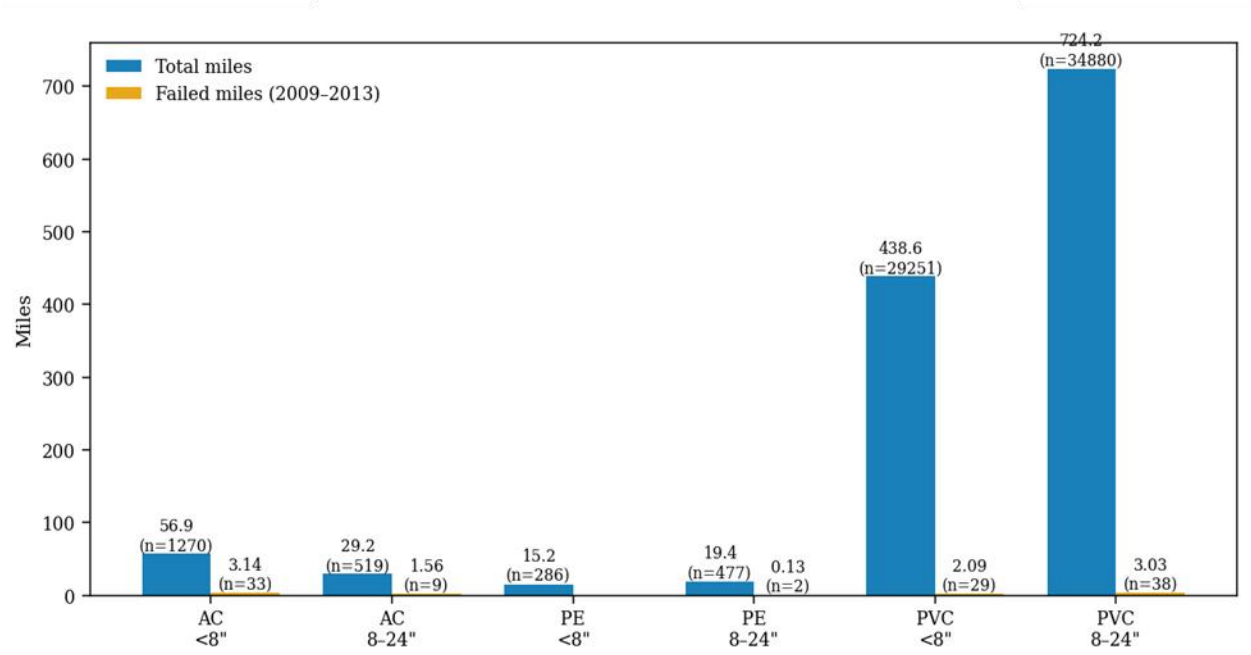

Figure 4-21: Total vs. five-year failed mileage by material and diameter band. Bars show, for AC, PE and PVC mains installed before 2014, the total mileage in service on 1 January 2009

(blue) and the subset of mileage that experienced at least one recorded failure between 2009–2013 (orange). Labels above each bar give the mileage and the number of pipe segments in each material–diameter cohort, highlighting that only a small fraction of the installed mileage failed over the five-year window.

Across AC, PE, and PVC mains in the <8″ and 8–24″ bands, the utility operated hundreds of miles and tens of thousands of segments as of 2009, while only a few miles in each cohort experienced failures during 2009–2013. This combination of large exposure mileage and relatively rare breaks creates a stringent, highly imbalanced test bed for evaluating how well the LOF index can identify the small subset of plastic and AC mains that actually fail.

4.7.4.5.2 Data Uncertainties

Although this retrospective failure dataset is unusually well curated for an operational work-order system, it still inherits several sources of uncertainty that are different in character from the instrumented condition-assessment campaigns used in the metallic and PCCP validations. First, failures are observed through work orders and main-break reports rather than direct inspection. This introduces detection bias (small leaks that never trigger a work order are effectively invisible) and classification noise (a leak coded as a service-line issue instead of a main break, or vice versa). In addition, the outcome is right-censored: mains that do not fail in 2009–2013 are treated as non-failures in the five-

year window, even though they may fail later. Right-censoring means that the "non-failed" group is a mixture of genuinely robust pipes and pipes that simply have not yet had the "bad luck" of a recorded failure.

Spatial and temporal referencing also introduces uncertainty. Even in a GIS-mature utility, there is non-zero risk that a break is snapped to the wrong segment in the map, that a short cluster of small breaks is represented as a single failure, or that attributes such as diameter and material have legacy coding errors for older mains. The environmental covariates are derived from SSURGO map units and regional climate grids, which are necessarily coarser than the pipe footprint. Each segment inherits a single value for concrete corrosivity, bedrock and water-table depth groups, elevation, slope length, and mean annual precipitation averaged over the intersecting map unit. This spatial aggregation smooths fine-scale heterogeneity in soil and groundwater conditions and can dilute the apparent strength of environmental drivers, especially where short mains cross boundaries between contrasting soil units. Temporal mismatches may also arise because SSURGO and climate normals represent long-term conditions, whereas individual failures can be triggered by shorter-term events (e.g., a wet winter or a construction-induced disturbance) that are not explicitly captured here.

Several design choices in the validation protocol aim to reduce these uncertainties or at least prevent them from biasing the conclusions in our favor. To minimize dependence on any one coding convention, we only use whether a main experienced at least one confirmed failure in the 2009–2013 window, rather than relying on the exact count or cause code. By focusing on first failures in that horizon and treating additional breaks on the same segment as part of the same outcome, we reduce sensitivity to how crews split repeat visits across work orders. The asset snapshot is fixed as of 1 January 2009, and all $\mathrm{LOF}_{\mathrm{student}}$ predictions are generated from that state, avoiding any leakage of post-failure information into the predictors. All performance metrics are computed on material–diameter cohorts and expressed in terms of mileage (miles of failed mains captured in high-risk bands and miles flagged as high-LOF that actually fail), which mitigates the impact of idiosyncratic segmentation lengths. Finally, and most importantly, this utility was completely held out from model development and from the earlier EVV exercises, so any predictive signal detected here is out-of-sample. Given the remaining detection, coding, and spatial aggregation uncertainties, the five-year recall and precision estimates reported in the results should be interpreted as conservative lower bounds on the model's true ability to anticipate failures in non-metallic mains under real operational conditions.

4.7.4.5.3 Data Processing

For this study we extract a focused dataset for all AC, PVC, and PE distribution mains that were installed before 2014 and lie within the utility's service area. The asset registry provides a unique Pipe ID, material, diameter, segment length (in miles), and installation year for each main segment. These records are linked to the verified work-order system, which logs each main failure as a dated event. For every Pipe ID, we retrieve any Failure Date associated with a confirmed main break between 1 January 2009 and 31 December 2013. Pipes with at least one Failure Date in this interval are treated as "failed" in the five-year window; pipes with no recorded failure in that interval are treated as non-failed (right-censored) for the purposes of this retrospective analysis.

To represent environmental exposure, we enrich the joined asset–work-order table with soil and topographic attributes derived from the NRCS SSURGO database and regional climate grids. For each pipe segment, we create the necessary input parameters to run the $\text{LOF}_{\text{student}}$ models. For example, parameters like concrete corrosivity index that reflects the aggressiveness of the surrounding soil toward cementitious materials, mean annual precipitation (inches) as a measure of climatic wetness, elevation and slope length as summaries of local topography and surface run-off potential, and bedrock and water-

table depth groups (feet) that characterize subsurface drainage and the likelihood of persistent saturation around the pipe are attached to enrich the dataset and support the model run. These attributes are computed on SSURGO map units intersecting each pipe's GIS alignment and then joined back to the Pipe ID using spatial join feature on ESRI's ArcGIS Pro. The resulting dataset combines a long, independently curated history of operational failures with spatially resolved environmental predictors and forms the basis for the five-year retrospective failure-prediction experiment for AC, PVC, and PE cohorts.

4.7.4.5.4 Hypothesis Testing

In this experiment we treat the five-LOF index for AC, PVC, and PE mains as a prospective screening tool for failure prediction accuracy in the five-year window 2009–2013. The model is applied to the temporal baseline as of 1 January 2009 and assigns each pipe segment to one of five LOF bands (0–4). For each material–diameter cohort, we define the high-LOF zone as the mileage of pipes that the model rated in the top two LOF bands (classes 3–4) at the 1 January 2009 baseline. The work-order and break database then identifies which segments experience at least one failure in 2009–2013. This yields, for each cohort, (i) the total mileage in service, (ii) the total failed mileage over

five years, (iii) the failed mileage that lay in the high-LOF zone at baseline, and (iv) the total mileage labelled high-LOF.

Two families of performance measures follow directly from these quantities:

- Five-year recall (capture fraction, miles) – failed mileage in the high-LOF zone divided by total failed mileage for the cohort.
- Five-year precision (hit-rate, miles) – failed mileage in the high-LOF zone divided by the total mileage labelled high-LOF.

At the network scale we also compute year-specific recall, the fraction of failed mileage in that year that lay in the high-risk zone at baseline. Because the high-risk zone comprises less than 1% of the total mileage in service, any recall substantially above this exposure fraction represents a strong concentration of failures in the segments the model views as near failure. These ideas lead to the following null hypotheses and decision rules (Table 4-21). We use mileage rather than segment counts so that longer mains, which contribute more exposure to failure, are weighted appropriately.

Table 4-21: Hypotheses and tests for the PVC, PE and AC failure retrospective experiment

ID	Test	Null hypothesis H_0	Statistic / estimate	Decision rule ($\alpha = 0.05$)	Interpretation if H_0 is rejected
H_0-R[1]	Year-specific recall	For each year, the model does not meaningfully concentrate failed mileage; year-specific recall ≤ 0.30.	For each year y, capture fraction of failed mileage in the high-risk zone, R_y.	Reject H_0 if the 95% CI for R_y lies entirely above 0.30 in at least four of the five years.	High-LOF zones contain a substantial fraction of the mileage that fails in each year, rather than only a small, random subset.
H_0-R[2]	Cohort-level 5-year recall	For each major material–diameter cohort, the five-year recall of failed mileage is at most 0.30.	Five-year recall R_c for each cohort c (AC <8″, AC 8–24″, PVC <8″, PVC 8–24″, PE 8–24″).	For each cohort, reject H_0 if the 95% CI for R_c lies entirely above 0.30.	Within that cohort the model captures a substantial share of the failed mileage in its high-risk bands over five years.
H_0-P	Cohort-level 5-year precision	High-risk assignments are not more failure-prone than background; five-year precision ≤ 0.30.	Five-year precision P_c for each cohort c, defined as failed mileage in the high-LOF zone divided by total high-LOF mileage.	For each cohort, reject H_0 if the 95% CI for P_c lies entirely above 0.30.	High-LOF assignments correspond to segments whose five-year failure rate is meaningfully higher than that of the cohort.
H_0-RR	Failure concentration vs background	The failure rate in the high-LOF zone is no higher than the network-wide average.	Failure ratio RR = (failed miles / total miles) in high-risk zone ÷ (failed miles / total miles) outside high-risk zone.	Reject H_0 if the lower bound of the 95% CI for RR exceeds 1.0.	The model concentrates failures. A mile in the high-LOF zone is significantly more likely to fail than a typical mile elsewhere in the system.

4.7.4.5.5 Results

Figure 4-22 summarizes how much of the failed mileage in each year lies in the high-risk LOF zones defined at baseline. Table 4-22 reports the underlying numbers.

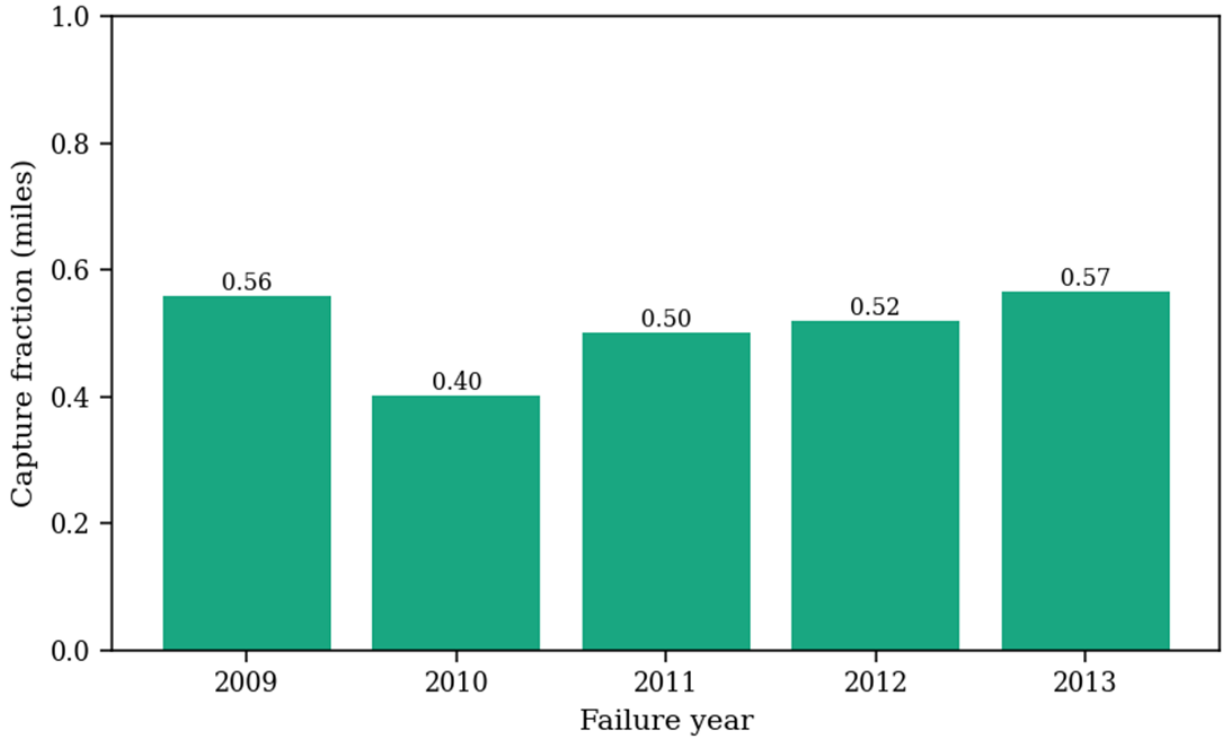

Figure 4-22: Year-specific capture of failed mileage (all cohorts)

Across all material-diameter cohorts combining the model captures between about 40% and 57% of the failed mileage in its high-risk bands in each individual year. This performance is consistent over time even though the underlying failures occur in different segments and under different environmental and operational conditions. Given that the high-risk LOF bands represent well under 1% of the total mileage in service, capturing roughly half of the failed mileage year after year represents a substantial enrichment of failures in the segments the model scores as risky. In other words, if a utility had

prioritized only this small subset of AC, PVC, and PE pipes for inspection or renewal in 2009, it would have pre-emptively intercepted a very large share of the mileage that went on to fail in each of the five subsequent years.

These patterns provide strong evidence against H0-R[1]. Even with conservative binomial uncertainty, all year-specific capture fractions sit well above the 0.30 threshold, so the model passes a demanding temporal stress test, it does not just fit one particular year but continues to concentrate future failures in its high-LOF zone throughout 2009–2013.

Table 4-22: Year-specific capture of failed mileage (all materials and diameter bands, miles)

Failure year	Failed mileage (mi)	Failed mileage in high-risk LOF bands (mi)	Capture fraction (recall, miles)
2009	2.36	1.32	0.56
2010	2.93	1.17	0.40
2011	1.51	0.76	0.50
2012	2.41	1.25	0.52
2013	0.73	0.41	0.57

The more detailed picture appears when recall and precision are computed separately for each material–diameter cohort (Figure 4-23 and Table 4-23).

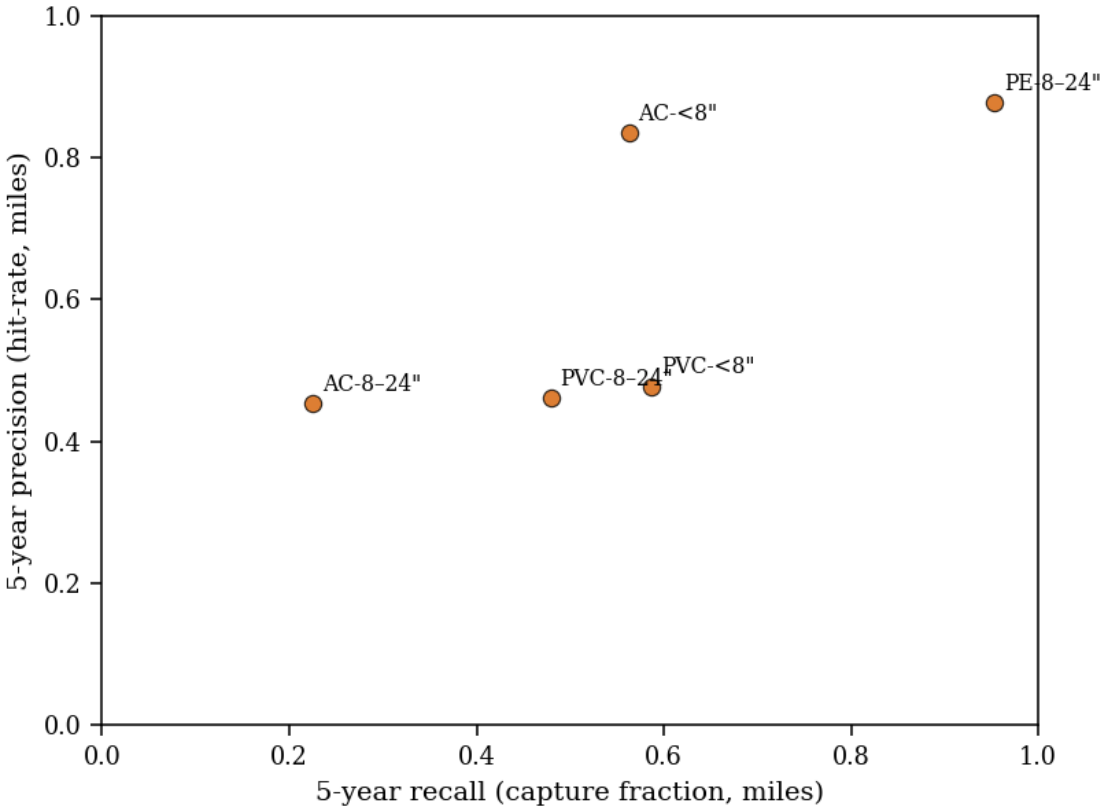

Figure 4-23: Five-year recall–precision trade-off by material-diameter cohort

Across all cohorts, the high-risk LOF bands cover only about 9.1 miles out of roughly 1,284 miles of other pipe material cohorts in service at baseline (≈0.7% of the mileage). Yet these 9.1 miles account for approximately 4.9 of the 10 miles that fail between 2009 and 2013 (≈49% of the failed mileage). This implies that the five-year failure rate in the high-risk zone is roughly 0.54 failed miles per mile of pipe, compared to about 0.008 failed miles per mile for the network as a whole, a risk ratio of roughly 70. This strong concentration of failure in the high-risk LOF bands provides direct evidence against

H_0-RR: the model is not simply spreading failures uniformly across the network but is identifying a tiny subset of mains where failures are orders of magnitude more likely.

Table 4-23: Five-year recall and precision by material-diameter cohort (mileage basis, 2009–2013)

Material	Diameter band	Total miles in service (mi)	Failed miles, 5 years (mi)	Failed miles in high-risk LOF bands (mi)	High-risk miles at baseline (mi)	5-year recall (capture fraction, miles)	5-year precision (hit-rate, miles)
AC	<8″	56.91	3.14	1.77	2.12	0.56	0.83
AC	8–24″	29.23	1.56	0.35	0.78	0.23	0.45
PE	<8″	15.19	0.00	0.00	0.37	– (no failures)	0.00
PE	8–24″	19.42	0.13	0.12	0.14	0.95	0.88
PVC	<8″	438.64	2.09	1.23	2.57	0.59	0.48
PVC	8–24″	724.24	3.03	1.45	3.15	0.48	0.46

The cohort-level results illuminate how this concentration plays out for different materials and diameters:

- Small-diameter AC (<8″): The model captures about 56% of the failed mileage over five years while labelling only ~3.7% of the AC <8″ mileage as high-risk (2.12 mi out of 56.9 mi). Among these high-risk AC <8″ mains, roughly 84% of the mileage fails within five years. This high recall and very high precision suggest that the LOF index is well aligned with the dominant failure mechanisms for small AC mains, mainly age,

aggressive soils, and chronic leakage, all of which are explicitly encoded in the model features. H_0-R^2 and H_0-P are both clearly rejected for this cohort.

- Medium-diameter AC (8–24″): Here the picture is more mixed. Only about 23% of the failed mileage lies in the high-risk bands, and the precision is moderate (~45%). Large AC mains are fewer in number and are often located in corridors with similar soil and environmental conditions, so a small set of unmodelled factors (construction details, pressure transients, repairs at fittings) can determine which of many seemingly similar segments fail. The model still improves failure concentration relative to random chance, but the recall is not high enough to reject H_0-R^2 for this cohort. H_0-P, however, is rejected: high-risk AC 8–24″ segments do fail at much higher rates than low-risk ones.
- Medium-diameter PE (8–24″): Although the total failed mileage is small, almost all of it (≈95%) resides in the high-risk bands, and nearly 88% of the high-risk mileage fails within five years. This combination of very high recall and precision suggests that when PE mains do fail in this system they tend to occupy a narrow region of the environmental and operational feature space that the LOF model successfully identifies. Both H_0-R^2 and H_0-P are strongly rejected for PE 8–24″, with the caveat that the statistical support is based on a limited number of failures.

- Small-diameter PE (<8″): No failures are observed for PE <8″ in the five-year window, so recall is undefined and precision is necessarily 0 (none of the mileage labelled high-risk failed). This cohort is not informative for hypothesis testing; it instead reflects that the PE <8″ population in this system is young and has not yet accumulated a measurable failure history.
- PVC mains (<8″ and 8–24″). For both diameter bands the model concentrates almost half of the failed mileage in high-risk LOF bands (recall ≈0.59 for <8″ and ≈0.48 for 8–24″) while maintaining moderate precision (~0.48 and ~0.46 respectively). Given that the high-risk bands comprise only about 0.6% of the PVC <8″ mileage and 0.4% of the PVC 8–24″ mileage, these recall and precision values represent substantial enrichment. The remaining missed failures likely arise from mechanisms that are harder to observe in coarse asset records like construction defects, localized third-party damage, or joint-related issues, rather than from the long-term environmental factors that dominate metallic corrosion. H_0-R^2 and H_0-P can be rejected for PVC <8″ and 8–24″, though with somewhat weaker margins than for AC <8″ and PE 8–24″.

Taken together, these results confirm that the LOF index for AC, PVC, and PE provides a useful, forward-looking screening tool. When the model is applied once at the

beginning of the five-year period, its high-LOF bands, comprising less than one percent of the non-metallic mileage—contain about half of the mileage that actually fails in 2009–2013, and within that small zone the failure rate is about seventy times the network-wide average. The model is particularly effective for small-diameter AC and medium-diameter PE, and performs at a useful but more modest level for PVC and larger AC mains. This pattern is exactly what would be expected from a physics-guided index: where failure mechanisms are strongly linked to age and environmental loading, the model captures them efficiently; where failures are driven by more idiosyncratic, unobserved factors, performance is limited by the information available in the input data rather than by the modelling framework itself. Table 4-24 summarizes the outcomes of the hypothesis tests introduced in Table 4-21, phrased qualitatively because the recall and precision thresholds are defined on a mileage basis and confidence intervals are dominated by the large effect sizes rather than fine statistical nuances.

Table 4-24: Hypothesis testing outcomes for the other material retrospective experiment

ID	Null hypothesis (H_0)	Test / estimate	Evidence	Decision
H_0-R[1]	Year-specific recall ≤ 0.30	Year-specific capture fractions 0.40–0.57	All five years show recall well above 0.30 despite the high-LOF zone covering <1% of mileage.	Reject H_0. The model captures a substantial share of failed mileage each year in its high-risk bands.

ID	Null hypothesis (H_0)	Test / estimate	Evidence	Decision
H_0-R^2	5-year recall ≤ 0.30 for each cohort	Five-year recall per cohort	AC <8″, PE 8–24″, PVC <8″, and PVC 8–24″ all have recall ≥0.48; AC 8–24″ has recall ≈0.23; PE <8″ has no failures.	Reject H_0 for AC <8″, PE 8–24″, PVC <8″, PVC 8–24″. Do not reject for AC 8–24″ (limited recall) and treat PE <8″ as not testable (no observed failures).
H_0-P	5-year precision ≤ 0.30	Five-year precision per cohort	Precision ranges from ≈0.45 to 0.88 for all cohorts with non-zero failures, far above 0.30; PE <8″ has precision 0 because no labelled high-LOF mileage fails.	Reject H_0 for AC <8″, AC 8–24″, PE 8–24″, PVC <8″, PVC 8–24″. Do not reject for PE <8″ (no failures).
H_0-RR	High-risk failure rate equals or is below background	Risk ratio RR ≈ 70 between high-LOF zone and network average.	High-LOF mileage is <1% of the network but contains ≈49% of all failed mileage; failure rate ≈0.54 failures per mile vs ≈0.008 per mile overall.	Reject H_0. The model concentrates failures in the high-risk LOF bands by almost two orders of magnitude.

4.8 Summary

This chapter develops and tests an LOF framework that is grounded in pipe-level deterioration mechanisms, tailored to the realities of utility data, and implemented as a teacher–student modeling system. It begins by clarifying scope: LOF is defined as a forward-looking propensity for structural and functional failures at the pipe-segment scale, conditioned on material, diameter, and ecological context rather than on ad hoc age thresholds or break counts alone. A five-band LOF index (0–4) is then constructed as the common output metric, anchored in explicit definitions of operational failure and in

observable signals such as wall-thickness loss, wire breaks, and sustained service interruptions. Descriptive analytics provide empirical baselines by material–diameter–environment cohorts and link observed failures to mechanism-consistent drivers, which are encoded as "motifs" in a fuzzy rule-base. On this foundation, a knowledge-structured fuzzy "teacher" model is built, with membership functions and IF–THEN rules that interpolate smoothly across mechanisms, followed by defuzzification into the 0–5 performance scale and the 0–4 LOF bands. Supervised "student" models (deep but compact neural networks) are then trained on the teacher's I/O pairs, with stabilization policies that preserve monotone and ordinal structure while gaining flexibility in high-dimensional regions that are hard to encode with rules alone. The Evaluation Verification Validation step closes the loop where teacher behavior is checked against physics and expert expectations, student models are verified against teacher outputs, and three external validation experiments—metallic mains with remaining wall thickness, PCCP with EM wire-break counts, and non-metallic mains in a five-year retrospective failure test demonstrate that the learned LOF bands are not only internally coherent but also informative about real-world distress and future main breaks across diverse materials and data conditions.

Chapter 5

Consequence of Failure Model

The Likelihood of Failure (LOF) chapter treated failure as a *mechanism-driven event* given a pipe segment with certain properties and exposures and provided a validated model to predict how often is it expected to fail. That description is incomplete for renewal decision-making. A pipe that fails rarely but sits under a major hospital, or crosses a sensitive river reach, is not equivalent to a small distribution main on a quiet cul-de-sac, even if their LOF indices match. This chapter therefore develops the *Consequence of Failure* (COF) model, which quantifies the severity of impacts conditional on a failure occurring and provides the second pillar of the risk construct used in this dissertation.

5.1 Goal and Scope

Most prior studies present "risk" as a simple product, often written informally as

$$\text{Risk} = \text{LOF} \times \text{COF}$$

This implicitly assumes that COF is a linear multiplier, and that decision makers are risk neutral and agree on how different consequences should be traded against each other. In practice, utilities use COF as one of several criteria in renewal planning, alongside LOF, cost, service equity, project feasibility and other factors. The relevance of COF changes with risk attitude (for example, whether the utility is risk neutral, risk averse, or willing to accept higher-consequence exposures in the short term). To make this explicit, we treat risk in this dissertation as a more general mapping

$$\mathcal{R} = \mathcal{R}(\mathrm{LOF},\mathrm{COF};\ \theta),$$

where θ collects the weights and risk-attitude parameters that will be introduced in the renewal prioritization chapter. The familiar LOF×COF product is recovered as a special case when $\mathcal{R}$ is linear, separable, and risk-neutral. In the COF chapter, we therefore focus on defining and estimating COF as a criterion with a clear physical meaning, constructed so that it can be combined flexibly with LOF within a Multi-Criteria Optimization (MCO) framework in the next chapter.

This chapter also extends our prior COF work (Vishwakarma and Sinha (2023)) where we presented an expert-system COF model based on fuzzy rules that combined customer, economic, and environmental factors at the segment level. Here, we build

directly on that expert system, formalizing it as the teacher model in a teacher–student architecture. The teacher encodes expert knowledge and transparent rules, and then machine-learning student models are trained on its input–output patterns and evaluated more rigorously against utility data. The result is a COF framework that preserves expert structure but gains scalability and stronger validation.

In this dissertation, COF is defined as the expected adverse change in the state of the urban water Socio-Ecological-Technical system (SETS) given a failure of a specific pipe segment. "Socio-ecological-technical" emphasizes that a pipe break is not only a hydraulic or structural event. It disrupts customer service (social and technical), damages roads and property (economic and technical), can mobilize contaminants or chlorinated water into streams and soils (ecological), and consumes maintenance and construction capacity that could have been deployed elsewhere (operational and organizational). The COF model translates this multi-dimensional impact into an interpretable index that can be used alongside LOF for renewal planning, without forcing all impacts into a single dollar figure.

The primary goal of the COF model is therefore two-fold. First, the model provides a consistent impact-structured severity index at the pipe-segment level. Each segment

receives a COF band on a discrete 0-5 scale (defined later), anchored in measurable quantities such as customer-hours of outage, traffic-control complexity, environmental receptor sensitivity, and projected renewal difficulty. The index is conditional on failure where a COF of 4 does not mean the segment will fail, but that if it does, the consequences are classified as extreme relative to other segments in the same network. The term *impact-structured emphasizes* that the index is not a black box; it is built from specific impact pathways and propagation mechanisms, rather than from a generic "importance" score.

Second, the model is designed to support risk-based renewal decision-making at multiple planning scales and within a multi-criteria framework. At the segment level, COF provides a dimensioned anchor for risk. COF is supported by underlying quantities with units (customer-hours, dollars, lengths of road affected, lengths of environmentally sensitive crossing), even though the final COF band is ordinal. When segments are grouped into spatial work "projects" and then into multi-year Capital Improvement Program (CIP) portfolios, these segment-level COF values can be aggregated in ways that are consistent with how utilities actually discuss consequence (for example, "total customer-hours avoided in this corridor" or "total length of high-consequence mains renewed near hospitals"). Because LOF in this dissertation is treated as a dimensionless index,

COF plays a complementary role. COF reintroduces how much is at stake in units that are meaningful for decision makers while still allowing the final COF label to be handled as an ordinal criterion in the optimization. This design supports risk-neutral and risk-averse attitudes alike by allowing different weights and thresholds on COF without changing the underlying COF calculation itself. The rest of this section clarifies the asset types, impact dimensions, and decision uses for which the COF model is intended.

From an asset perspective, the COF model focuses on buried pressurized drinking-water mains in the diameter range considered in the LOF chapter (approximately 4–120 inches), including both distribution and transmission pipes. Consequences attributable primarily to above-ground facilities (treatment plants, storage tanks, booster stations) and service lines are not modeled here, except where their function is mediated through the pipes themselves. For example, the consequence of a failure in the only transmission main leaving a treatment plant is captured because that main is modeled as a critical segment, not because a separate plant-level COF model is developed. This choice keeps the COF model tightly coupled to the LOF model and to the renewal decisions that are within the remit of a pipeline-focused asset management program.

From an impact perspective, COF is modeled as a structured combination of five dimensions: social, economic, environmental, operational, and renewal complexity. The social dimension captures who is affected and how severely, using indicators such as customer-hours of disruption, presence of critical customers, traffic delays, and typical property damage by land use. The economic dimension aggregates direct renewal and response costs (labor, materials, equipment, repaving, landscaping, contractor and police services), costs linked to volumetric water loss and flushing, temporary alternative supplies, and potential legal costs, parameterized mainly by land cover and typical unit rates. The environmental dimension represents loss of treated water to sensitive receptors, contaminant-laden runoff from chlorinated discharges, ground instability (for example, sinkholes under roads), and emissions and embodied energy associated with failure and repair, driven by proximity to water bodies and environmentally sensitive zones. The operational dimension reflects how difficult the response action is to manage based on workforce availability, pressure and fire-flow conditions, redundancy in the local network, and the strain on CIP budgets in areas with limited ability to pay. The renewal complexity dimension captures how hard it will be to undertake a renewal (for example, pipes under buildings, major roads, or water surfaces, burial depth, quality of records, and material atypicality). Internally, these dimensions are linked to quantities with units wherever

possible, but the integrated COF output is an ordinal band so that hard cost and engineering quantities can be combined with social and environmental impacts that cannot always be meaningfully or respectfully reduced to a single monetary figure.

From a decision perspective, the COF model is designed to be fit-for-purpose for the renewal contexts addressed in this dissertation. At the tactical project scale, segment-level COF scores help delineate and justify renewal "projects" or work packages. For example, clustering high COF segments in a hospital district so that they can be upgraded together and the number of disruptive work zones is reduced. At the 5-year CIP scale, COF interacts with LOF and renewal costs in the multi-objective optimization framework developed in the next chapter, where COF acts as a criterion that can be weighted more heavily in risk-averse scenarios (prioritizing the removal of high-consequence exposures) or balanced with other objectives such as total risk reduction, equity, and budget constraints.

The constructs introduced here are general enough that, in principle, they could support cross-utility comparisons of consequence profiles. However, systematic benchmarking between utilities is not pursued in this dissertation and is instead treated as a direction for future work. The focus here is on making COF useful and credible for within-

utility segment and portfolio decisions. A summary of all the dimensions involved in modeling COF is presented in Table 5-1.

Table 5-1: Summary of COF dimensions, their main intent, typical internal metrics, and primary data drivers

Dimension	Main intent	Typical internal metrics	Primary data drivers
Social	Who is affected and how severely.	Customer-hours of disruption; number and type of critical customers (hospitals, schools, emergency services); indicative traffic delay; typical property damage by land use.	Customer and critical-facility GIS; land use; proximity to roads and railways; socio-economic and vulnerability layers.
Economic	Direct and indirect monetary consequences.	Repair, rehabilitation and replacement costs; labor, materials, equipment, repaving, landscaping; contractor and police services; costs from volumetric water loss and flushing; bottled-water provision; legal costs.	Land cover over the pipe; unit-cost libraries; historical work orders; estimated water-loss volumes (including flushing); typical property and claims cost data where available.
Environmental	Ecological effects of releases and response.	Length of environmentally sensitive crossing; volume of treated water lost to receptors; qualitative runoff and contamination risk; indicative risk of ground instability/sinkholes; emissions and embodied energy.	Proximity to streams, wetlands, and high-value groundwater; environmental sensitivity classes; treatment energy intensity; transport distances for materials and crews.
Operational	Difficulty of managing and containing the event.	Workforce availability (including reliance on external contractors); static-pressure and fire-flow context; degree of network redundancy; indicative strain on CIP budgets in low ability-to-pay areas.	Utility staffing profiles; hydrant locations; pressure and demand data; network topology (distribution vs transmission); poverty prevalence or similar socio-economic indices.
Renewal complexity	Difficulty and cost escalation of renewal.	Depth class; presence under buildings, major roads, railways, or water surfaces; quality of utility records; availability of standard vs atypical materials and fittings.	Land-cover and infrastructure layers; recorded burial depth; quality of as-built and asset records; material and component catalogues.

Several modeling choices in this chapter are made with interpretability, ethical clarity, and integrability in mind. First, expressing COF on a compact ordinal scale allows the model to combine "hard" quantities (such as customer-hours and repair costs) with

"soft" or less quantifiable impacts (such as environmental degradation or social disruption) without pretending that all of them can be reduced to a single dollar value. Second, the COF formulation is kept stable across different risk attitudes. Risk neutrality or risk aversion is expressed later in the next chapter through the choice of weights and thresholds in the renewal optimization, not by redefining COF itself. Third, the chapter adopts the same teacher–student architecture used for LOF. A knowledge-structured fuzzy teacher model encodes expert understanding of consequence mechanisms and impact propagation in a transparent rule base. This teacher is then used to generate large, labeled datasets for supervised training of machine-learning student models that can scale across networks while preserving the intended directionality and thresholds. Evaluation, verification, and validation of the COF models follow the same multi-layer protocol as the LOF models, with appropriate changes in ground-truth data (for example, using observed customer-hours and documented "high-impact" breaks in place of remaining wall thickness). Having set the goal and scope of the COF model, the remainder of the chapter proceeds from impact logic to implementation.

5.2 COF Grounding in Impact Mechanisms

The COF model in this dissertation is built around impact mechanisms rather than around arbitrary "importance scores." By impact mechanism we mean a concrete, defensible chain that links a pipe failure to changes in the social, ecological, and technical state of the system. This starts with the failure event itself (for example, a blowout or major leak), passes through the exposure of people, assets, and environments near the failed segment, and depends on their vulnerability to that disturbance (for example, lack of redundancy, high sensitivity, or high unit value). This flow from hazard to exposure to vulnerability decomposition is commonly used in disaster-risk analysis. Here, LOF largely describes the hazard side (how often a failure is expected), while COF primarily represents the exposure and vulnerability components that is, who and what are in harm's way, and how severely they are affected when a failure occurs.

Formally, for a given segment and failure mode, each COF dimension is modeled as a function of (i) the type and scale of the failure (hazard), (ii) the population, built environment, and ecosystems that depend on or surround that segment (exposure), and (iii) how sensitive those exposed elements are to disruption (vulnerability). Rather than estimate these functions directly from sparse, noisy consequence records, we first encode

them in a fuzzy "teacher" model that reflects expert knowledge and past work (Vishwakarma and Sinha, 2023) and then train data-driven "student" models to approximate that logic at scale. Table 5-2 summarizes how this hazard–exposure–vulnerability structure is instantiated for each COF dimension.

Table 5-2: Impact mechanisms for the five consequence-of-failure dimensions

Dimension	Primary hazard signal (H)	Exposure drivers (E)	Vulnerability drivers (V)	Indicators used in COF model (examples)
Social	Loss of pressure/supply and local flooding around the failure.	Number and type of customers and critical facilities hydraulically dependent on the failed segment; traffic and access patterns.	Presence of hospitals, schools, emergency services; socio-economic fragility of neighborhoods; typical repair duration.	Estimated customer-hours of outage; density of critical facilities in a buffer; proxies for traffic disruption; typical property-damage class by land use.
Economic	Physical break and associated water release require intervention and repair/renewal.	Built environment above and around the pipe (roads, buildings, commercial districts); number of actors involved (utility, contractors, police).	High unit costs (deep urban excavations, complex surfaces); severity of secondary impacts (business interruption, costly sites).	Composite unit-cost index (labor, materials, equipment, repaving, landscaping, contractor and police services); estimated loss volume (including flushing); land-cover type; generic ranges for property-damage and legal/claims costs for catastrophic cases.
Environmental	Sustained discharge of treated or raw water and mobilized contaminants into soils and receiving waters.	Proximity and hydraulic connectivity to streams, wetlands, groundwater zones, and other sensitive receptors.	Sensitivity of receptors to chlorine, sediment and contaminants; susceptibility of soils and subgrade to erosion or sinkhole formation.	Length of environmentally sensitive crossing; distance to nearest stream/wetland/groundwater zone; environmental sensitivity class; proxies for treatment-energy intensity and emissions/embodied energy associated with failure response and material replacement.
Operational	Failure event that must be isolated, controlled and repaired under live system conditions.	Network elements and demands attached to the segment (valves, hydrants, pressure zones, fire-flow requirements).	Workforce availability (in-house crews vs reliance on contractors); static pressure and fire-flow magnitude; lack of redundancy; budget fragility.	Workforce class (high/medium/low, including manageable contractor access); static-pressure class; hydrant proximity and fire-flow flags; redundancy class (looped distribution vs single-feed transmission); poverty/affordability context as a proxy for CIP sensitivity to unplanned failures.

Dimension	Primary hazard signal (H)	Exposure drivers (E)	Vulnerability drivers (V)	Indicators used in COF model (examples)
Renewal complexity	Need to undertake full renewal once failures or deterioration trigger replacement.	Physical and regulatory corridor where renewal would occur (under buildings, major roads, railways, or water surfaces; depth; subsurface congestion).	Poor records (location and configuration uncertainty); requirement for specialized methods or materials; sensitive social or environmental settings limiting methods and work windows.	Land-cover / infrastructure class above the pipe (open space, local road, arterial, building, water surface); burial-depth class; records-quality indicator; flag for standard versus atypical renewal scenario (for example, trenchless requirement or unusual materials).

For each dimension, the table lists: (i) the primary hazard signal that matters for that dimension, (ii) the main exposure drivers, (iii) the key vulnerability factors, and (iv) the indicators that implement these mechanisms in the COF model. This table guided the design of the fuzzy teacher model. Each fuzzy rule can be read as a qualitative statement about a specific combination of hazard, exposure and vulnerability in one or more dimensions (for example, "high-pressure transmission main under an arterial road with no redundancy and nearby hospital" should map to high social, economic, operational and renewal complexity consequence). The student models trained later in the chapter are evaluated against this structure to ensure that they preserve the intended monotonic behavior that is, the changes that clearly increase hazard, exposure or vulnerability for a given dimension must not reduce the corresponding COF score.

The detailed definitions, units and data sources for the indicators in the rightmost column are deferred to Sections 5.3 and 5.4, where the COF index and data dictionaries

are developed in full. Here, the emphasis is on making the impact logic transparent. COF in this dissertation is not a single opaque number, but a compact representation of well-defined hazard–exposure–vulnerability pathways across five consequence dimensions.

5.3 COF as an Output Metric for Modeling

The previous sections treated consequence conceptually, in terms of impact mechanisms and dimensions. For modeling and optimization, those ideas need to be compressed into a *single output variable* that can be predicted for each pipe segment and used as a criterion in portfolio design. This section defines that output that is, the COF index and clarifies what is meant by "pipe service consequence" at different planning scales. It then explains how the multi-dimensional consequence structure (social, economic, environmental, operational, and renewal complexity) is aggregated into a single COF band.

5.3.1 Definition of Pipe Service Consequence

We define *pipe service consequence* as the severity of adverse change in the service provided by a pipe segment to its surrounding socio-ecological-technical system,

conditional on a failure event occurring on that segment. Service is interpreted broadly. It includes the hydraulic service to customers (flow and pressure), but also the supporting functions that the segment provides to people and institutions, to the built environment and the economy, to the environment itself, to utility operations, and to future renewal. A segment that fails may interrupt water supply or degrade water quality for residential, commercial, industrial, and critical customers; disturb roads, businesses, and property above and around the pipe; release treated or raw water, sediments, and associated contaminants to streams, wetlands, and groundwater; draw crew, equipment, and budget away from planned work; and make future replacement more difficult and expensive because of ad hoc repairs in already constrained corridors.

Because these services unfold over time, pipe service consequences are described across three impact horizons. Immediate impacts are those in the first hours after failure, such as sudden loss of supply, local flooding, and acute traffic blockage. Short-term impacts extend over days to weeks and include prolonged low-pressure episodes, boil-water advisories, temporary business disruption, and the execution of emergency repairs and restorations. Long-term impacts extend over months to years and include large unplanned

capital expenditures, reputational damage, legal claims, persistent environmental damage, and the need for complex renewal in constrained settings.

Throughout, pipe service consequence is defined conditional on failure. The COF index answers the question: *If this segment were to fail in a representative way, how severe would the impacts be?* It does not indicate how often that failure will occur; that is the role of the LOF model in Chapter 4. Keeping these roles separate makes it possible to see clearly whether a high-risk segment is high-risk because failures are frequent, because failures are very damaging, or both.

5.3.2 Development of Target Output COF Index

For modeling, the COF output must satisfy three properties. It must be ordinal and interpretable, grounded in observable ranges of impact, and flexible across utility sizes. To meet these requirements, we adopt a 0–5 COF index with five bands of width 1 unit. These are 0–1 (Insignificant), 1–2 (Minor), 2–3 (Moderate), 3–4 (Major), and 4–5 (Catastrophic). The internal index is treated as a continuous value in $[0,5]$, produced by the teacher and student models. Table 5-3 provides operational definitions of each band in terms of typical economic cost ranges, scaled by utility size, and indicative environmental and social impact severities.

Table 5-3: Output COF Index (0-5) with detailed definitions (Vishwakarma and Sinha 2023)

Index		Detailed Definitions
0-1	Insignificant	Pipe failures in this band cause negligible economic costs, with direct renewal and response costs less than $50,000 for large utilities, less than $20,000 for medium utilities, and less than $5,000 for small utilities. Environmental impacts are minimal: no observable contamination of receiving waters and only small volumes of treated water lost (for example, on the order of a single repair event, typically well below 0.01 % of the utility's average daily production). Social impacts are negligible, with no noticeable disruption to the community: at most a very small fraction of customers (for example, fewer than about 1 % of connections) experience short interruptions of a few hours, no critical facilities are affected, and there are no boil-water advisories or traffic closures.
1-2	Minor	Pipe failures in this band result in economic costs between $50,000 and $100,000 for large utilities, between $20,000 and $50,000 for medium utilities, and between $5,000 and $20,000 for small utilities. Environmental impacts include minor contamination risks and moderate water loss, for example localized surface discharge or flushing volumes, which are still small relative to daily system production and unlikely to cause measurable ecological damage. Social impacts are limited to minor inconveniences for a small number of residents, with temporary disruptions to service over a small area (for example, a few streets or a single neighborhood) and outage durations typically limited to one working day; the affected share of customers remains low (on the order of 1–5 %), and critical facilities, if present, can usually be maintained through redundancy or temporary supply.
2-3	Moderate	Pipe failures in this band cause economic costs between $100,000 and $500,000 for large utilities, between $50,000 and $200,000 for medium utilities, and between $20,000 and $100,000 for small utilities. Environmental impacts include noticeable contamination risks and substantial water loss, such as discharges that may reach nearby streams or storm drains and loss volumes that are non-negligible compared with daily production, potentially requiring targeted environmental monitoring or clean-up. Social impacts include moderate disruption affecting multiple neighborhoods: a non-trivial fraction of customers (for example, 5–15 % of connections) experience low pressure or outage over many hours to a day or more, traffic detours are required on local or collector roads, and there may be localized boil-water advisories or temporary relocation for a limited number of vulnerable customers.
3-4	Major	Pipe failures in this band cause economic costs between $500,000 and $2,000,000 for large utilities, between $200,000 and $1,000,000 for medium utilities, and between $100,000 and $500,000 for small utilities. Environmental impacts include high contamination risks and significant water loss, with large discharges of treated or raw water to surface waters or sensitive land areas, and volumes that may represent a noticeable fraction of daily production and require active environmental response. Social impacts include major disruption affecting large sections of the city or town: a substantial fraction of customers (for example, 15–30 % of connections) may experience extended outages or poor water quality, major arterial roads can be blocked or severely constrained, and boil-water advisories or water-use restrictions can persist for days, with severe inconvenience for thousands of residents and businesses.

4–5	Catastrophic	Pipe failures in this band cause economic costs over $2,000,000 for large utilities, over $1,000,000 for medium utilities, and over $500,000 for small utilities. Environmental impacts include extreme contamination risks and massive water loss, for example failure modes that release very large volumes of treated or raw water to rivers, wetlands, or urban areas, with potential long-term ecological damage and clean-up requirements. Social impacts are profound, affecting entire cities or regions: a large share of the customer base (for example, more than 30 % of connections) can be affected for several days or longer, including critical facilities; extensive traffic disruption and infrastructure damage may occur; and emergency measures such as evacuation, widespread bottled-water distribution, or prolonged boil-water advisories are required, with long-term shortages or infrastructure reconstruction.

Economic ranges are specified separately for large, medium, and small utilities. Environmental and social descriptions are supplemented with simple, dimensionless anchors such as approximate fractions of customers affected, typical outage durations, and order-of-magnitude water loss volumes as a share of daily production. For example, an "Insignificant" (0–1) failure has low renewal cost (less than $50,000 for a large utility, less than $20,000 for a medium utility, and less than $5,000 for a small utility), involves very small water losses (on the order of a single repair event, typically well below about 0.01 % of average daily production), and affects at most a very small fraction of customers for a short period (for example, fewer than roughly 1 % of connections for a few hours, with no critical facilities involved and no traffic closures or advisories). At the other extreme, a "Catastrophic" (4–5) failure may cost more than $2 million for a large utility, more than $1 million for a medium utility, or more than $500 000 for a small utility; it is associated with extreme contamination risks and massive water loss (a large share of daily production

discharged to sensitive areas), and it has profound social consequences, often affecting more than 30 % of connections for several days or longer, including critical facilities, with widespread traffic disruption and emergency measures such as evacuations, bottled-water distribution, or prolonged boil-water advisories.

The quantitative anchors for customers affected, outage duration, and water-loss volumes are intended as order-of-magnitude guides rather than strict thresholds, helping to align the qualitative labels across diverse systems without claiming more precision than the underlying data support. Environmental and social impacts in Table 5-3 are therefore expressed in ordinal severity bands with indicative quantitative ranges, but they are not converted into dollars. Many such impacts such as damage to a wetland, long disruptions to a vulnerable community, or large-scale emergency responses are not naturally, or respectfully, expressed as monetary values, and forcing them into dollar terms would require strong, untestable assumptions about willingness to pay and valuation of ecosystems and health. The COF bands instead combine quantified economic cost ranges with these non-monetary but measurable anchors for environmental and social severity. In the modeling framework, this combination is handled via the multi-dimensional scheme described in the next subsection.

5.3.3 Disaggregate Consequence Dimensions and Aggregation Scheme

The COF index is not computed in a single step. It is constructed from five disaggregate consequence dimensions, each with its own intermediate index on a 0–5 scale. In the second layer, a separate fuzzy aggregation system takes these five dimension-level indices as inputs and produces a single COF index on the 0–5 scale. This architecture keeps the impact logic explicit while still delivering one scalar target variable that can be learned by the student models.

In the first layer, the teacher model computes dimension-level indices by applying fuzzy inference to the underlying indicators associated with each dimension. For the Social dimension, these indicators include estimated customer-hours of outage, the density and type of critical facilities in the affected zone, and proxies for traffic and property disruption. The Economic dimension uses composite unit-cost indices, typical loss volumes (including flushing), and land-cover classes that differentiate simple from complex work sites. The Environmental dimension is driven by the length of environmentally sensitive crossings, proximity to streams, wetlands and high-value groundwater, and simple proxies for treatment energy and emissions. The Operational dimension reflects workforce class, static-pressure and fire-flow context, hydrant density, redundancy class, and the

vulnerability of local capital budgets. Finally, the Renewal complexity dimension depends on land cover and infrastructure above the pipe, burial-depth class, records quality, and whether the renewal scenario is standard or atypical. For each dimension, these inputs are fuzzified into linguistic terms such as “low,” “medium,” and “high,” combined through an expert-derived fuzzy rule base, and defuzzified to yield a numeric score $C_{\mathrm{soc}}, C_{\mathrm{econ}}, C_{\mathrm{env}}, C_{\mathrm{ops}}, C_{\mathrm{renew}}$ in the range $[0,5]$. Each score is therefore an interpretable index that summarises the hazard–exposure–vulnerability mechanisms for that dimension.

The second layer then aggregates these five indices into a single COF value using another fuzzy inference system. Here, the inputs are the dimension-level scores rather than raw indicators. Each dimension index is again mapped to fuzzy labels (for example, “low”, “moderate”, “high”, “extreme” consequence within that dimension), and a compact rule base encodes how combinations of dimension-level consequences should translate into overall COF. Table 5-4 illustrates how the fuzzy aggregation layer translates combinations of dimension-level indices into an overall COF score. Segments B and C show the intended non-compensatory behavior that is, a single high or extreme dimension drives the overall COF upward, rather than being averaged away by lower scores elsewhere.

Table 5-4: An illustration of the second layer fuzzy inference system using hypothetical pipe segments.

Segment	Social	Economic	Environmental	Operational	Renewal complexity	COF (0–5)	COF band	Interpretation (one line)
A	1.2	1.0	1.1	1.3	1.0	1.3	Minor	All dimensions low → CoF remains Minor.
B	3.8	2.5	2.2	2.0	2.3	3.6	Major	Social high → fuzzy aggregation lifts CoF into Major even if others moderate.
C	2.5	4.3	2.6	2.4	2.7	4.1	Catastrophic	Economic extreme → treated as Catastrophic overall despite other dims only moderate.
D	3.0	3.0	3.0	3.0	3.0	3.0	Major	All dimensions moderate → CoF stays at Major, near the centre of the scale.

Typical rules include statements such as "IF Social is high OR Economic is high OR Environmental is high THEN COF is at least Major," or "IF all dimensions are low THEN COF is Insignificant," or "IF Environmental is extreme AND Social is moderate THEN COF is Catastrophic." A standard Mamdani-style fuzzy inference procedure is used where antecedent truth values are combined using fuzzy AND/OR operators, consequents are aggregated, and the final numerical COF score is obtained by defuzzification (centroid of the aggregated membership function) and scaled to the 0–5 range. In this way, risk-averse, non-compensatory behavior is achieved through the rule structure itself rather than through an explicit arithmetic max operator. Therefore, an extreme value in

any one dimension drives the overall COF upward, and low scores in other dimensions can only partially temper that effect.

This two-layer fuzzy aggregation scheme has several advantages. It preserves the interpretability of each dimension-level index, because the first layer is explicitly tied to physical indicators and mechanisms within a single consequence dimension. At the same time, it allows the second layer to encode realistic, non-linear interactions between dimensions that utilities recognize in practice. For example, a combination of high social consequence and moderate economic and environmental consequence may be treated as "Major," even if a simple average of the three indices would fall in the "Moderate" band. Similarly, the rule base can be calibrated so that certain dimensions (such as Environmental for a watershed-protection utility, or Social for a health-focused utility) effectively carry more weight by appearing in more rules or by pulling the output toward higher COF labels when they are elevated. Because all rules and membership functions are specified explicitly, these "weights" are transparent and can be discussed and adjusted with utility experts rather than being implicit parameters in a black-box model.

The final output of the teacher model is thus a continuous COF index in [0,5] along with its associated band (Insignificant to Catastrophic) that is consistent with the

detailed definitions in Table 5-3. The continuous score is used as the target variable for training the machine-learning student models, while the band provides a check on interpretability and alignment with expert judgement. Section 5.6 describes the membership functions and rule bases for both layers in detail. The key point here is that the COF index used for modeling is a structured fuzzy synthesis of five interpretable consequence dimensions, designed to be risk-averse, monotone with respect to increasing impacts, and fully compatible with the multi-criteria renewal prioritization framework developed in the next chapter.

5.4 Input Data and Feature Specifications

The COF model sits on the same geometric backbone and reliability discipline as the LOF model, but it draws from a broader set of spatial, socio-economic, environmental, and operational datasets. The goal in this section is to explain how COF specific inputs are attached to pipe segments in space and time, what assumptions are made when direct measurements are not available, and how each input is tagged with a reliability level. The emphasis is on making the COF inputs auditable where any given COF score can be

traced back to a small set of clearly defined predictors, each with known provenance, resolution, and expected direction of effect.

5.4.1 Spatial Resolution

All COF predictors are ultimately stored at the pipe-segment level, using the same segmented network representation as in Chapter 4. Each segment has a unique identifier and a geometry (polyline) that can be overlaid with other spatial layers. COF-relevant features are derived by intersecting this segment geometry with several classes of spatial datasets like pressure zones; land-use polygons; traffic/road classifications, locations of critical facilities, and environmental receptor datasets such as streams, wetlands, and groundwater protection areas. These overlays are implemented in ArcGIS Pro using standard spatial-joining operations such as within, intersects, nearest/closest, and within buffer, chosen according to the geometry of the underlying layer (polygon, line, or point) and the way that layer influences consequence.

Wherever possible, segments are defined so that they are internally homogeneous with respect to key attributes such as pressure zone and service area. For attributes that cannot be used to drive segmentation everywhere (for example, fine-scale land use or socio-economic indices), we use systematic aggregation rules applied after the spatial joins.

For categorical exposures such as land-use type, pressure zone, and service area, each segment inherits the category that covers the largest share of its length, computed through an intersects join between the segment polyline and the relevant polygons. When the segment touches a small, high-consequence feature, such as a hospital parcel, a major arterial road, a railway, or a wetland, the presence of that feature is recorded explicitly in separate flag and attributes like "length_in_sensitive_area", even if it does not dominate the segment length. These flags are typically derived from intersect joins for polygons and within buffer or nearest joins for points. This combination of majority assignment and risk-aware flags ensures that short but critical crossings are not diluted by longer stretches of lower-consequence surroundings.

For continuous exposures such as population density, social-vulnerability scores, property-value indices, or environmental sensitivity scores, attributes are computed as length-weighted averages across all polygons intersected by the segment, again using an intersect join followed by aggregation over the intersected lengths. Point features such as critical facilities or hydrants are handled through buffered counts or densities where a within or nearest join is applied between points and buffered segments to count facilities within a fixed distance, and the resulting counts are normalized by segment length to

obtain structured indicators (for example, critical facilities per mile of main). In all cases, the resulting COF predictors are stored as segment-level attributes aligned with the LOF feature tables, so that the same segment ID can be used to join failure mechanisms and consequence mechanisms without additional geoprocessing. Table 5-5 shows how spatial operations are performed to create each variable at the segment level.

Table 5-5: COF input variables, underlying datasets, spatial join operations, and aggregation rules used to derive segment-level attributes. "None (segment attribute)" indicates variables created directly from utility records or models rather than from external GIS layers

No.	Input variable	Dimension	Source	Spatial Join Type	Aggregation to pipe segment technique	High-consequence override
1	Direct Cost of Renewal	Economic	Utility unit-cost tables (non-spatial) linked to material, diameter, depth, land cover	None (segment attribute; attribute crosswalk)	Assign representative unit renewal cost class to each segment based on its context	Higher cost class used where segment is flagged under major arterial, railway, or building corridor
2	Cost of Lost Water	Economic	Water audit (volumes, cost/volume) + hydraulic model / pressure zones (polygons / nodes)	*Nearest* (segment to model node or pressure zone)	Estimate representative loss volume per failure for each pressure zone and assign to segments	Higher class for segments in highest pressure zones or with large service areas
3	Cost of Legal Issues	Economic	Land-cover/zoning polygons (NLCD, municipal); utility financial records	*Intersects* (segment with land-cover/zoning polygons)	Map land-cover/zoning to typical legal-risk/cost class and assign dominant class	Any intersection with central business district or high-value institutional parcels escalates class
4	Cost of impact on surface water and wetlands	Environmental	NWI wetlands polygons; surface-water bodies and buffers (polygons)	*Intersects / within buffer*	Compute length of segment within surface-water / wetland buffers; derive impact class	Any crossing of mapped wetland, floodplain, or drinking-water intake buffer sets high impact flag
5	Potential for Landslides	Environmental	NRCS SSURGO soil and slope stability polygons	*Intersects*	Assign highest landslide/sinkhole susceptibility class intersected by segment	If any intersected polygon has "high" susceptibility, segment marked as high risk
6	Greenhouse Gas Emissions	Environmental	LCA studies / unit emission factors (non-spatial) by renewal method and site type	None (segment attribute; attribute crosswalk)	Assign emission factor per ft of renewal based on typical method for that context	Higher emission factor if trenchless or heavy-equipment methods required (from renewal complexity flags)

No.	Input variable	Dimension	Source	Spatial Join Type	Aggregation to pipe segment technique	High-consequence override
7	Customer Service Disruption	Social	Land-use / NLCD polygons; critical customer points (hospitals, etc.); service areas	*Intersects* (land use, service areas), *within buffer* (critical points)	Derive disruption scale (1–5) from combination of land use, customer density, and critical-facility presence	Any intersection with hospital or major emergency facility sets disruption to at least high
8	Road and Railway Traffic Flow Impact	Social	Road network polylines (DOT) with functional class; railway lines	*Nearest* / *intersects*	Assign impact class from nearest intersecting road/rail of highest functional class	Crossing of major arterial or rail corridor sets impact to at least major
9	Water Quality Impact	Social	Customer complaint points; superfund sites; wastewater pipeline polylines	*Within buffer* (segment buffer around points/lines)	Count complaints and nearby contamination sources; convert to 1–5 water-quality impact scale	Presence of superfund site or major wastewater interceptor within buffer sets high impact flag
10	Cost of Property Damage	Social	Land-use / NLCD polygons; municipal property-value polygons or grids	*Intersects*	Compute dominant land-use/property-value class intersected; map to typical damage cost	Any intersection with high-value commercial or dense residential area escalates damage class
11	Financial Affordability	Operational	Census socio-economic polygons (tracts/block groups)	*Intersects* / length-weighted	Length-weighted average of poverty prevalence over polygons intersected by segment	If any intersected polygon is in highest poverty decile, segment marked as high affordability concern
12	Workforce Availability	Operational	Utility crew/staffing data (non-spatial, possibly by district)	None (segment attribute by service district)	Assign workforce class (high/medium/low/manageable) to segments in each service district	None; class may be stepped up manually for remote or difficult-access districts
13	Fire Flow Impact	Operational	Hydrant points; fire-demand specifications	*Within buffer* (buffer around segment centerline)	Count hydrants within buffer and consider required fire flows; convert to impact scale	Segments serving designated high-fire-flow corridors (e.g., near critical facilities) forced to high class
14	Static Pressure	Operational	Hydraulic model nodes or pressure-zone polygons	*Nearest* (segment to model node) or *within* (zone)	Assign representative static pressure or pressure class to each segment	Highest pressure class in zone determines segment pressure impact
15	Redundancy Level	Operational	Network topology from utility GIS (pipe graph)	None (derived from graph analysis)	Compute redundancy index (e.g., looped vs single-feed) and store as % or category	Transmission mains with no feasible alternate path forced to 0 % redundancy
16	Land Cover	Renewal complexity	NLCD and municipal land-cover/land-use polygons	*Intersects*	Assign dominant land-cover type to segment; map to renewal difficulty scale	Any overlap with buildings, major arterials, railways, or water surfaces triggers “difficult renewal” flag
17	Depth of Pipe	Renewal complexity	Utility as built and asset records (segment attributes)	None (segment attribute; default when missing)	Use recorded depth; if missing, assign default (e.g., 4 ft) and tag as assumed	Very deep segments (above a threshold, e.g., >8 ft) forced into highest depth-difficulty class

No.	Input variable	Dimension	Source	Spatial Join Type	Aggregation to pipe segment technique	High-consequence override
18	Quality of Utility Records		Assessment of collected data coverage and consistency (non-spatial, by asset class or district)	None (segment attribute by class/district)	Assign records-quality score (e.g., poor/fair/good) to segments based on data review	Segments with conflicting or missing core attributes (material, diameter, age) set to worst quality
19	Availability of Spare Parts		Utility materials inventory and standard parts catalogues (non-spatial)	None (segment attribute via material/diameter crosswalk)	Assign spare-parts availability class based on how common the pipe material/diameter is	Rare or obsolete materials and fittings automatically mapped to lowest availability class

5.4.2 Temporal Resolution

Consequence is inherently time-dependent where service outages unfold over hours, clean-up over days, cost recovery and reputational effects over months and years. For modeling, however, the COF index is treated as a location-conditioned property of each segment that is, given a representative failure on that segment, how severe are the expected impacts under typical operating conditions. This requires a pragmatic choice of temporal resolution that balances realism with data availability.

At the event scale, utilities often record a break occurrence timestamp and a repair completion timestamp, sometimes with intermediate milestones such as “crew dispatched” or “main isolated.” When such data are available, they provide direct estimates of outage duration and crew mobilization times in different contexts (for example, under major roads versus in open fields). In the COF model, these event-level records are used to

calibrate typical repair and isolation durations by context, which then inform the estimation of customer-hours of outage and operational disruption for a representative failure on each segment. Where only dates, or coarse timestamps, are available, standard assumptions about typical repair windows (for instance, eight to twelve hours for a moderate break in accessible conditions) are used and documented in Section 5.4.3.

At the metric scale, short-term consequences are summarized in outage windows. For example, the total customer-hours of interruption that a representative failure would cause if it occurred during a typical high-demand period. These are constructed from estimated outage durations and counts of dependent customers or critical facilities at risk, based on the spatial aggregation described above. Longer-term consequences, such as annualized costs or recurring disruptions in a corridor with repeated failures, are summarized on annual or multi-year basis consistent with utility reporting cycles. The COF index is, however, not recalculated for every year. Instead, it is evaluated for a reference period (for example, a recent three-year window in which land use, service patterns, and network configuration are broadly stable) and assumed to be representative for the 5-year capital improvement planning horizon, unless there are known forthcoming changes such as major rezoning, new hospitals, or large industrial customers.

This temporal structure mirrors the LOF chapter, where incidents, work orders, and environmental layers were also aggregated to yearly windows for stability, with longer multi-year windows for trend analysis. The difference is that LOF tracks how often failures happen in each window, whereas COF uses those same windows, together with event-level calibrations, to characterize how severe a typical failure would be. Table 5-6 summarizes how each COF-relevant metric is anchored in time, distinguishing event-scale metrics (such as outage duration, customer-hours, and loss volume) from annualized summaries and multi-year portfolio indicators. This separation makes explicit which parts of the COF model describe the severity of a representative failure on a segment, and which are higher-level aggregates used later to evaluate corridors and renewal portfolios.

Table 5-6: COF metrics and their primary temporal windows

Quantity / metric	**Dimension**	**Temporal window**	**Significance for COF and planning**
Break occurrence and repair timestamps	Operational	Event (minutes–hours)	Anchor the start and end of a failure; used to calibrate typical isolation and repair durations by context.
Outage duration per failure (hours)	Social / Operational	Event → outage window (0–48 h)	Combined with estimated customers affected to compute customer-hours of service disruption for a representative failure.
Customer-hours of outage per segment	Social	Outage window (single representative failure)	Key social consequence indicator; used directly in the fuzzy Social COF sub-model and to interpret COF bands.
Volume of water lost per failure	Economic / Environmental	Event → outage window	Estimated from pressure class and failure type; used to derive Cost of Lost Water and environmental impact class.
Traffic disruption duration and extent	Social / Economic	Outage window (hours–days)	Derived from road/rail class and typical repair times; informs Road and Railway Traffic Flow Impact.

Quantity / metric	Dimension	Temporal window	Significance for COF and planning
Direct repair / renewal cost per failure	Economic	Event / outage window	Includes labor, materials, equipment, repaving, landscaping and contractor services; mapped into economic loss ranges for COF bands.
Indicative property-damage cost per failure	Economic / Social	Event / outage window	Typical building damage costs for the surrounding land-use class; contributes to economic loss ranges in COF bands.
Environmental impact class	Environmental	Event / short-term (days–weeks)	Reflects likelihood and severity of contamination and ecological damage from a representative failure; treated as stable for the planning horizon.
Annualized unplanned failure cost in corridor	Economic / Operational	Year	Summarizes recent history of unplanned costs in a corridor; used for cross-checking the reasonableness of event-based COF assumptions.
Annualized customer disruption in corridor	Social	Year	Aggregated customer-hours or number of advisory days over recent years; used descriptively alongside segment-level COF scores.
Length of very-high-COF pipe renewed	All (aggregate outcome)	Multi-year CIP horizon (typically 5 years)	Portfolio-level indicator: how much high-consequence exposure is removed by a given renewal program.
Cumulative COF - weighted risk reduction	All (aggregate outcome)	Multi-year CIP horizon	Used in Chapter 6 to evaluate and compare portfolios; combines LOF and COF over the planning horizon.

5.4.3 Data Assumptions and Reliability Levels

Not all COF inputs can be directly measured. Some require assumptions or simple models that translate available data into the quantities needed by the teacher model. To keep these approximations transparent, we use the same five-level data-reliability ladder introduced in the LOF chapter. Each COF predictor in the data dictionary is tagged with one of the 5 levels.

COF-specific assumptions arise in several places. When detailed hydraulic outage footprints are not available, an outage radius or equivalent isolation footprint is assumed

for each failure scenario, informed by pressure-zone maps, valve layouts, and typical isolation practices reported by the utility. Demand patterns are often approximated by daily or seasonal averages rather than full diurnal curves, where hourly demand data or SCADA logs exist, they are used to calibrate simple "peak" versus "off-peak" multipliers for customer-hours of outage, otherwise these multipliers are taken from engineering judgement (reliability level 1–2). Repair mobilization times in terms of how long it takes to dispatch a crew and begin isolation are estimated from work-order histories where timestamp quality permits (reliability 4–5); in data-poor settings, they are inferred from utility size, staffing, and typical practice (reliability 1–2).

Other assumptions are embedded in crosswalks between categorical datasets and cost or disruption classes. Land-use classes and road hierarchies are mapped to typical unit costs and traffic disruption levels using cost books, design standards, and local expert input. Social vulnerability indices and poverty prevalence are used as proxies for the sensitivity of capital budgets and households to unplanned failures in different areas, with the understanding that these indices are noisy and are therefore tagged as derived (reliability 2–3) rather than as system measurements. For environmental consequence, the mapping from proximity to streams and wetlands to contamination risk classes is based

on environmental guidance documents, buffer distances, and utility feedback, again treated as derived with moderate reliability. Table 5-7 shows a concise summary of the main approximations used to construct COF inputs, their intended scope, and associated reliability levels on the 1–5 scale (1 = educated guess, 5 = direct measurement).

Table 5-7: Key COF-specific assumptions and reliability levels

Assumption / derived quantity	**Intended scope**	**Construction and data basis**	**Reliability level**
Isolation footprint / outage radius	Approximate spatial extent of customers affected by a single main failure	Derived from valve layout, pressure-zone maps, and utility isolation practice; calibrated where event data exist, otherwise expert judgement.	2–4
Typical repair and isolation duration	Event-scale duration (hours) of isolation and repair for representative failures in each context	Estimated from work-order timestamps by land-use/road class and depth where available; otherwise from utility reports and engineering judgement.	2–5
Customer-hours of outage per segment	Social COF input for representative failure on each segment	Product of estimated outage duration and dependent customer count (from service-area and land-use overlays); assumes typical high-demand period.	2–4
Loss volume per failure	Economic/Environmental COF input reflecting water lost in a representative failure	Derived from static pressure class, typical leak/break type, and outage duration; calibrated to water-audit loss components where possible.	2–4
Traffic disruption class	Traffic-flow and economic impact of excavation and lane closures	Mapping from road/rail functional class and typical repair duration to qualitative disruption bands (minor, moderate, major).	2–3
Property-damage cost class	Typical direct property damage cost around the failed segment	Crosswalk from land-use/parcel type and indicative property values to damage-cost bands used in COF band definitions.	2–3
Environmental impact class	Likelihood and severity of contamination of streams, wetlands, and sensitive receptors	Buffer-based proximity of segment to mapped receptors; combined with loss-volume class to assign environmental consequence bands.	2–3
Financial affordability index	Sensitivity of local customers and CIP budgets to unplanned failures	Poverty prevalence or social-vulnerability indices at census-tract / block-group level, length-weighted over segments.	2–3
Greenhouse-gas emission factor per ft of renewal	Environmental / economic proxy for emissions associated with failure response and renewal	Unit emission factors by method and site type drawn from LCA studies and typical construction practices; mapped to renewal scenarios.	2–3
Default burial depth where records are missing	Depth input for renewal-complexity assessment when as-built records are incomplete	Fixed default (e.g., 4 ft) applied when no depth is recorded; flagged as assumed and excluded from calibration of depth–cost relationships.	1

5.4.4 Data Sources for COF Model Inputs

COF features draw on a deliberately mixed set of utility-internal and external datasets, with the balance differing by consequence dimension. For the economic and operational dimensions, the primary sources are internal. Direct cost of renewal, cost of lost water, and property-damage cost classes are derived from utility financial records, work-order and break logs, and internal unit-cost libraries (including labor, materials, equipment, construction, repaving, landscaping, and contractor services). Operational indicators such as static pressure and fire-flow context come from the hydraulic model and SCADA data; redundancy is derived from the utility's network GIS; and workforce availability is based on crew counts, crew locations, and staffing patterns. Financial affordability indicator is defined to capture the *utility's capacity* to absorb unplanned failure costs on a given segment without destabilizing its CIP. This index is constructed from poverty prevalence and related socio-economic indices at census-tract or block-group level overlaid on the service area, because these contextual factors constrain the utility's practical ability to raise rates or issue additional debt. In other words, socio-economic data are used as a proxy for budget flexibility at the utility level, not as a direct measure of individual customers' ability to pay or of distributional justice.

For the social dimension, external spatial datasets play a larger role. Traffic-related disruption is characterized using state Department of Transportation (DOT) databases that provide road functional class and, where available, traffic volume measures such as Average Annual Daily Traffic (AADT). Land cover and the broad type of area around each pipe segment (for example, open space, low-density residential, high-density residential, commercial, industrial) are derived from the National Land Cover Database (NLCD) and, where needed, finer municipal land-use layers. Critical-facility exposure is captured using open datasets hosted by municipalities or counties that map medical facilities, dialysis centers, schools, and other priority service locations. These layers are overlaid with the network in ArcGIS Pro, and their outputs combine with internal customer records to construct the customer service disruption index and the road and railway traffic-flow impact metric.

For the renewal complexity dimension, land cover above or adjacent to the pipe which is important to consider for constructability, is again obtained from NLCD and local land-use layers, while depth of pipe, quality of utility records, and availability of spare parts are drawn from internal utility data wherever those records exist. Depth is taken from as-built drawings or asset databases; record quality is assessed from the

completeness and consistency of the collected utility data; and spare-parts availability is inferred from materials and diameter catalogues and inventory records. When depth or spare-parts information is missing, simple, documented assumptions are used (for example, a default depth of 4 ft or "low availability" for obsolete materials), and those variables are tagged at lower reliability on the ladder.

For the environmental dimension, national and federal datasets are the main sources of receptor information. The National Wetlands Inventory (NWI) and mapped surface-water bodies are used to identify wetlands, streams, rivers, and adjacent buffers; EPA datasets on Superfund and other contaminated sites are used to characterize potential contamination receptors; and landslide and sinkhole potential are taken from NRCS SSURGO soil and geohazard layers. These are combined with segment-level loss-volume estimates to derive the environmental impact class. Greenhouse-gas emission factors for failure response and renewal are informed by internal or regional life-cycle assessment (LCA) studies where available; otherwise, standard emission factors from published LCA literature are adopted and treated as derived assumptions rather than measurements.

Throughout, the priority is to avoid constructing entirely synthetic COF predictors when measured or recorded data exist, and to keep a clear record of provenance when

assumptions are unavoidable. In the implementation, each COF feature carries metadata pointing back to its source dataset(s). For example, specific DOT layers, NLCD release year, NWI version, census vintage, or internal cost book, and, where applicable, the version and date of download. This traceability supports the reliability tagging in Section 5.4.3 and makes it straightforward to update the COF model when improved datasets or utility records become available.

5.4.5 Data Dictionaries

To make the COF model reproducible and interpretable, all predictors and intermediate indices are documented in structured data dictionaries parallel to those used in the LOF chapter. Each row in the COF data dictionary corresponds to a single variable, with fields for variable name (as used in code and tables), a concise definition, units, data type (continuous, categorical, binary, index), primary data source(s),; spatial resolution and aggregation rule (for example, "polygon overlay, length-weighted average at segment level"), temporal resolution and refresh frequency, expected direction of effect on COF, material and diameter applicability notes (for example, "applies to all materials; only segments ≥ 12 in considered for transmission-main redundancy flag"), and the reliability level on the 1–5 ladder. For intermediate indices, such as the five dimension-level COF

scores, the dictionary includes a brief description of their construction from underlying indicators and the fuzzy inference layer. Table 5-8 shows a summarized version of the COF data dictionary informing the teacher COF model.

Table 5-8: Data dictionary of COF predictors and intermediate indices. Full dictionaries, including all variables used in the teacher and student models, are provided in the code repository.

Variable name	Definition	Units / type	Primary source(s)	Spatial / temporal resolution	Effect on COF	Reliability
direct_cost_renewal	Representative direct renewal cost for the segment (labour, materials, equipment, paving, landscaping, contractor services).	$ (continuous)	Utility financial records; bid tabs; unit-cost library	Segment-level; derived from land cover, depth, diameter; updated with cost-book revisions	Higher cost → higher Economic COF	3–4
cost_lost_water	Monetary value of water lost in a representative failure on the segment, including flushing and alternative supply (e.g., bottled water).	$ (continuous)	Water audit; production cost estimates; hydraulic model	Segment-level; event-scale estimate anchored to typical outage window	Higher cost → higher Economic / Environmental COF	2–4
surface_water_impact_class	Ordinal class of potential impact on surface water and wetlands from a representative failure.	1–5 ordinal class	NWI, EPA hydrography; proximity and buffer analysis	Segment-level; based on distance/length within buffers; assumed stable over CIP horizon	Higher class → higher Environmental COF	2–3
landslide_potential_class	Susceptibility of local soils and subgrade to landslides or sinkholes triggered by leakage or rupture.	{Low, Medium, High} categorical	NRCS SSURGO soil/stability polygons	Segment-level; dominant or maximum susceptibility over segment length	Higher class → higher Environmental / Social COF	2–3
cust_service_disruption	Indicative customer service disruption score, combining customer density, land use, and critical-customer presence.	1–5 ordinal class	NLCD/land use; customer GIS; critical-facility layers	Segment-level; derived from polygon overlays and buffered point counts	Higher score → higher Social COF	2–3
traffic_flow_impact_cost	Proxy cost index for road and rail traffic disruption during repair/renewal.	$-index (continuous or banded)	State DOT road/rail network; lane-closure cost factors	Segment-level; based on highest-functional-class crossing; event-scale duration	Higher index → higher Social / Economic COF	2–3
financial_affordability_idx	Index representing the utility's capacity to absorb unplanned failure costs on this segment without destabilising its CIP (proxy for budget flexibility), constructed from poverty prevalence and related socio-	0–1 index (continuous)	Census socio-economic datasets	Segment-level; length-weighted average over intersected tracts; updated with census releases	Higher index (less budget flexibility) → higher Operational COF (greater affordability strain)	2–3

Variable name	Definition	Units / type	Primary source(s)	Spatial / temporal resolution	Effect on COF	Reliability
	economic indicators in the served area.					
redundancy_level_pct	Percentage of demand that can be met via alternative paths if the segment is out of service.	% (continuous)	Network topology from utility GIS	Segment-level; derived from graph analysis; assumed stable unless network reconfigured	Lower % → higher Operational COF	2–3
cof_social_dim	Dimension-level Social COF index from fuzzy inference (0–5).	0–5 index	Fuzzy teacher model (inputs: cust_service_disruption, traffic, complaints, etc.)	Segment-level; conditional on representative failure; evaluated for reference period	Higher value → higher overall COF	2–3
cof_economic_dim	Dimension-level Economic COF index from fuzzy inference (0–5).	0–5 index	Fuzzy teacher model (inputs: direct_cost_renewal, cost_lost_water, property-damage class)	Segment-level; conditional on representative failure	Higher value → higher overall COF	2–4
cof_env_dim	Dimension-level Environmental COF index from fuzzy inference (0–5).	0–5 index	Fuzzy teacher model (inputs: surface_water_impact_class, landslide_potential_class, loss volume)	Segment-level; conditional on representative failure	Higher value → higher overall COF	2–3
cof_ops_dim	Dimension-level Operational COF index from fuzzy inference (0–5).	0–5 index	Fuzzy teacher model (inputs: static pressure, fire-flow impact, workforce, redundancy, financial affordability)	Segment-level; conditional on representative failure	Higher value → higher overall COF	2–4
cof_renew_dim	Dimension-level Renewal-complexity COF index from fuzzy inference (0–5).	0–5 index	Fuzzy teacher model (inputs: land cover, depth, record quality, spare-parts availability)	Segment-level; conditional on representative failure	Higher value → higher overall COF	2–3

5.5 Descriptive Analytics and Impact Baselines

Before fixing the COF teacher model and training student models, we first examine how the potential consequences of failure are distributed across the network. The aim is

not to forecast individual events, but to construct impact baselines: how much social, economic, environmental, operational, and renewal complexity "exposure" is concentrated in different parts of the system. These baselines serve three roles. They check that the COF inputs and indices behave sensibly (for example, downtown trunk mains should not look "low-consequence" in aggregate). They clarify which combinations of context and network role dominate high-consequence exposure, and therefore deserve attention in the rule base. And they provide a transparent link between real-world patterns—trunk versus distribution, urban versus rural, ecologically sensitive versus ordinary—and the *impact motifs* that later appear as fuzzy rules in the teacher model.

5.5.1 Descriptive Analytics and Impact Baselines

Before fixing the COF teacher model and training student models, we examine how the current COF ratings are distributed across the network. The goal here is not to predict individual failures, but to construct impact baselines that is, which parts of the inventory, by diameter and material, already appear in the upper COF bands, and how those patterns align with unit replacement costs.

These baselines are built from the expert-system COF model originally published in Vishwakarma and Sinha (2023), applied to the anonymized utility dataset. At present,

utilities rarely collect systematic records of environmental or social consequences (for example, ecosystem damage, human health impacts, or customer-hours of outage) after pipe failures. Those dimensions are therefore represented in the COF framework primarily through reliable spatial proxies like land use, traffic classes, proximity to wetlands and surface waters, socio-economic layers. By contrast, economic and operational consequences (repair costs, renewal unit costs, crew time, and work-zone constraints) cam be tracked reasonably well. The figures in this section should thus be read as distributions of overall COF bands conditioned on inventory and cost data, with environmental and social dimensions partially latent and refined later via the fuzzy teacher model.

5.5.2 Baselines by diameter and material–diameter cohorts

The first baseline asks how COF exposure is distributed simply by diameter. We group segments into three diameter cohorts (< 8 in, 8–24 in, and > 24 in) and compute the mileage-weighted share of pipe in each COF band. The result is summarized in Figure 5-1. The pattern is strongly size-dependent. Small-diameter mains (< 8 in) are dominated by the Minor–Moderate bands, reflecting modest repair and renewal costs and limited impact corridors. Intermediate diameters (8–24 in) show a shift towards Moderate consequences with a small but non-zero share in the Major band, consistent with their role as

larger distribution mains and smaller transmission feeders. The largest cohort (> 24 in) is almost entirely in the Major–Catastrophic bands, indicating that very large mains reliably combine high unit costs with high impact corridors, even before environmental and social proxies are fully exploited. This validates the basic expectation that, all else equal, failures on very large mains have qualitatively different consequences than failures on small distribution pipes.

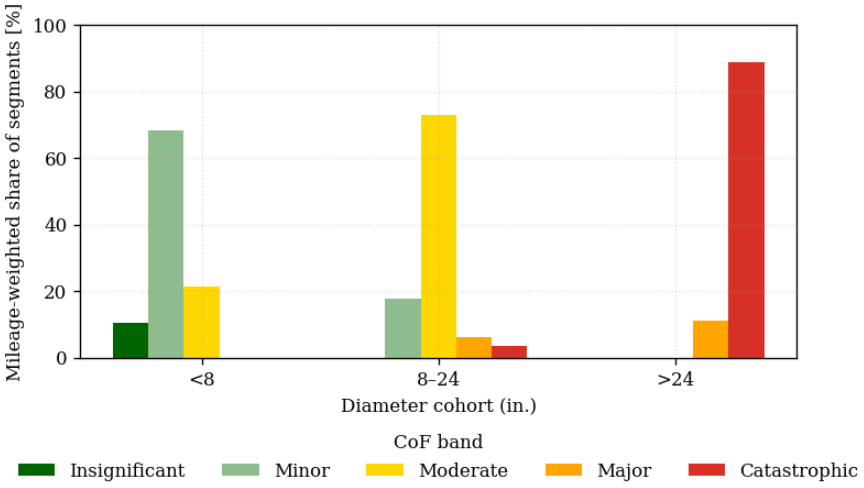

Figure 5-1: Distribution of COF bands by diameter cohort

Diameter alone, however, does not capture material-specific behavior or renewal context. To examine this, we stratify COF by material–diameter cohort, yielding the clustered bar chart in Figure 5-2.

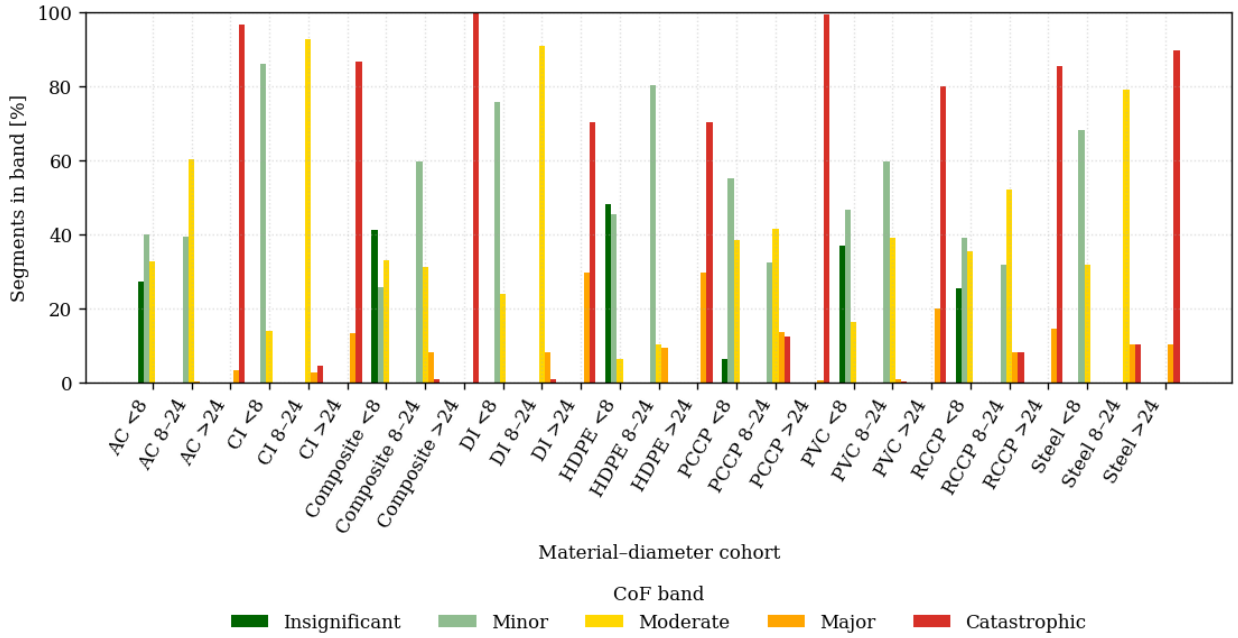

Figure 5-2: COF distribution across material–diameter cohorts

Several patterns are evident in Figure 5-2. Large-diameter metallic and concrete mains, especially > 24 in CI, DI, Steel, and PCCP, are overwhelmingly classified as Catastrophic, with very little mileage in the lower COF bands, reflecting both their high unit replacement costs and their frequent location in critical corridors such as arterials, river crossings, and major feeders. At the other end of the spectrum, small-diameter PVC and HDPE cohorts are concentrated in the Minor–Moderate bands, with a substantial share of small-diameter PVC in the Insignificant band. These pipes can be understood to typically serve residential streets, are comparatively shallow, and have lower replacement costs, so their failures are consequential but bounded. AC shows a split behavior, with

small and intermediate diameters clustering in the Moderate band, while the rare large-diameter AC entries fall into the Catastrophic band, again driven by the combination of size and corridor type. Concrete pressure pipes (PCCP and RCCP) also exhibit high COF at larger diameters, consistent with their predominant use as trunk and transmission mains and their very high replacement costs when buried deeply in urban or otherwise difficult-to-access environments.

These baselines act as a sanity check on the COF index. They confirm that the expert-system ratings from Vishwakarma and Sinha (2023) do not, for example, label large-diameter PCCP feeders as "Minor," and they identify the cohorts where high-consequence exposure is structurally concentrated. They also set expectations for the student learner model that any dramatic reordering of these material–diameter patterns would be suspect unless supported by strong additional data.

5.5.3 Economic drivers: Replacement costs as a baseline

The overall COF index aggregates economic, social, environmental, operational, and renewal-complexity dimensions. In practice, however, the only dimension for which utilities routinely maintain quantitative records at the cohort level is direct economic cost. To make this explicit, we compiled unit replacement cost data from participating utilities

for each material–diameter combination, using work orders, bid tabs, and standard pay items. For each cohort, we compute the mean replacement cost per foot and a 95% confidence interval, as shown in Figure 5-3.

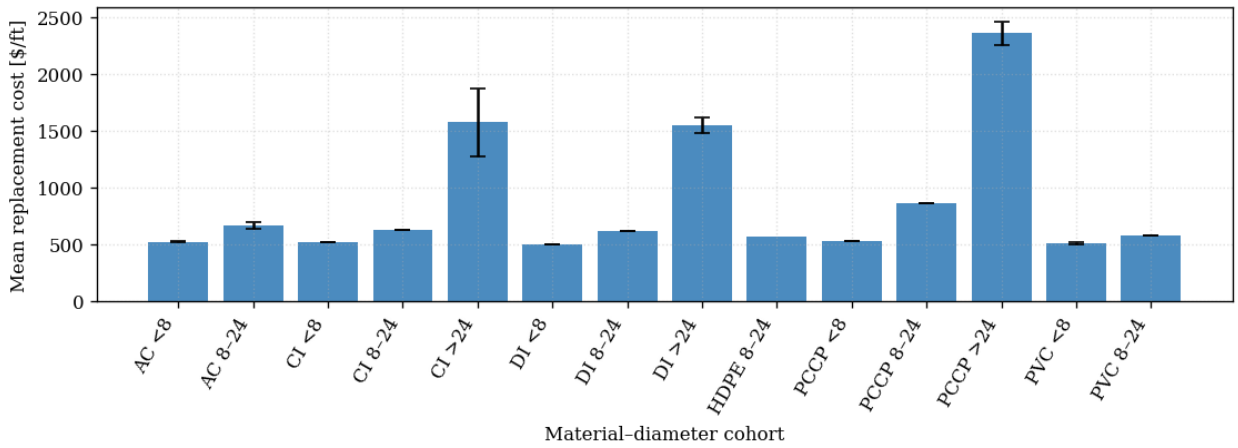

Figure 5-3: Replacement cost by material–diameter cohort (mean ± 95% CI)

Figure 5-3 shows that unit costs vary by almost an order of magnitude across cohorts. Large-diameter PCCP and CI mains sit at the top of the distribution with mean costs in the high hundreds to low thousands of dollars per foot, reflecting deep burial, complex traffic control, and specialized construction methods. Intermediate-diameter DI and HDPE lie in a mid-range, while small-diameter AC and PVC typically occupy the lower half of the distribution. The width of the confidence intervals captures the construction variability within each cohort with differences driven by site constraints, surface restoration requirements, or contractor practices.

Comparing Figure 5-2 and Figure 5-3 reveals that cohorts with systematically high unit replacement costs also tend to occupy the upper COF bands, particularly for large-diameter CI and PCCP. This supports the modelling choice to treat unit cost as a primary driver of the Economic dimension. At the same time, there are cohorts where COF bands are elevated relative to their mean unit cost, typically where other contextual factors (environmental sensitivity, lack of redundancy, critical facilities) are important. These mismatches are intentional and underscore that the COF framework is not a pure cost model and diameter, land use, redundancy, and environmental setting can legitimately pull a cohort into a higher COF band even if its unit costs are only moderate.

It is important to emphasize that environmental and social consequences are not observed directly in this dataset. Utilities do not routinely record, for example, volumes of contaminated runoff, ecological restoration costs, or customer-hours of outage by pipe cohort. Those dimensions enter the COF teacher model via proxy variables (wetland buffers, surface-water proximity, traffic volume classes, critical-facility locations, socio-economic indices) rather than through explicit empirical calibration. The descriptive baselines in this section therefore emphasize diameter, material, and cost, while later sections

use the proxy structure to embed environmental and social considerations into the fuzzy rule base.

5.5.4 Impact motifs derived from diameter–material and cost baselines

The COF patterns stratified by material–diameter cohorts, combined with the replacement-cost baselines, give rise to a set of impact motifs that will guide Evaluation, Verification, and Validation (EVV) of the COF student models. These motifs are extracted from the same fuzzy expert system that was originally published in Vishwakarma and Sinha (2023), so the distributions in Figure 5-1, Figure 5-2 and Figure 5-3 represent the *teacher's* view of consequence, not an external heuristic.

One motif is the large-diameter trunk main, where CI, DI, Steel, or PCCP mains in the > 24 in cohort appear almost exclusively in the Major–Catastrophic bands (Figure 5-2) and exhibit very high unit replacement costs with substantial variability (Figure 5-3). These pipes typically run under arterial roads or in constrained utility corridors. EVV scenarios based on this motif will expect any failure on such mains to be rated at least Major, with scope for escalation to Catastrophic when environmental or operational complications are present.

A second motif is the moderate-diameter distribution main in ordinary corridors. In this case, 8–24 in AC, PVC, and HDPE cohorts sit largely in the Minor–Moderate bands with mid-range unit costs. Failures are disruptive but locally contained, with social and economic consequences being more manageable and renewal being relatively routine. EVV scenarios derived from this motif will check that student models do not over-penalize the vast mileage of ordinary distribution mains simply because they are numerous.

A third motif is the apparently "cheap but exposed" cohort, which includes material–diameter combinations whose mean unit cost is modest but that frequently lie in environmentally or socially sensitive settings, such as smaller-diameter mains crossing wetlands or serving critical facilities. In the present dataset, such patterns are not fully visible in Figure 5-1, Figure 5-2 and Figure 5-3 because explicit environmental and social observations are missing, but they are anticipated by the proxy structure developed earlier. The impact motifs therefore include rule templates that allow Environmental or Social dimensions to dominate the overall COF, raising a cohort into a higher band even when replacement costs alone would not justify it.

Together, these motifs help design EVV protocols and ensure that the COF student models remain anchored in the empirical baselines defined by the fuzzy teacher (Figure

5-1, Figure 5-2 and Figure 5-3), while still allowing environmental and social proxies to play a decisive role where appropriate. The EVV chapter returns to these motifs and shows how they are operationalized as test scenarios, so that the resulting COF index respects the observed material–diameter and cost patterns without degenerating into a purely cost-driven metric.

5.6 Knowledge-structured "teacher" model (Fuzzy Inference System)

The COF teacher model is implemented as a hierarchical fuzzy inference system with two main layers. The first layer maps raw indicators (for example, unit renewal cost, land cover, traffic class, wetland proximity, redundancy, pressure, and workforce availability) into five dimension-level indices: Social, Economic, Environmental, Operational, and Renewal-complexity COF. The second layer aggregates these five indices into a single continuous COF index on a 0–5 scale, which is then banded into the five ordinal classes defined earlier (Insignificant, Minor, Moderate, Major, Catastrophic). Figure 5-4 summarizes this architecture. Indicators are grouped into five dimension-level FIS modules, each with its own rule base, and their outputs serve as inputs to a final FIS that yields the overall COF rating. This is the same structure as in Vishwakarma and Sinha (2023). Here

it is refined with the updated indicator set, dimensional decomposition, and band definitions developed in Sections 5.2–5.5.

Fuzzy inference is used for the same reasons as in the LOF chapter. Many COF drivers are only partially observed, and expert knowledge is naturally expressed in qualitative form ("if a main fails near a hospital on a high-traffic arterial with little redundancy, the consequences are severe even if the unit cost is only moderate"). Fuzzy sets and IF–THEN rules provide a direct way to encode these verbal heuristics while still yielding a numerical index that is monotone in severity, smooth over the input space, and suitable both as a teacher for student models and as a baseline for EVV scenarios.

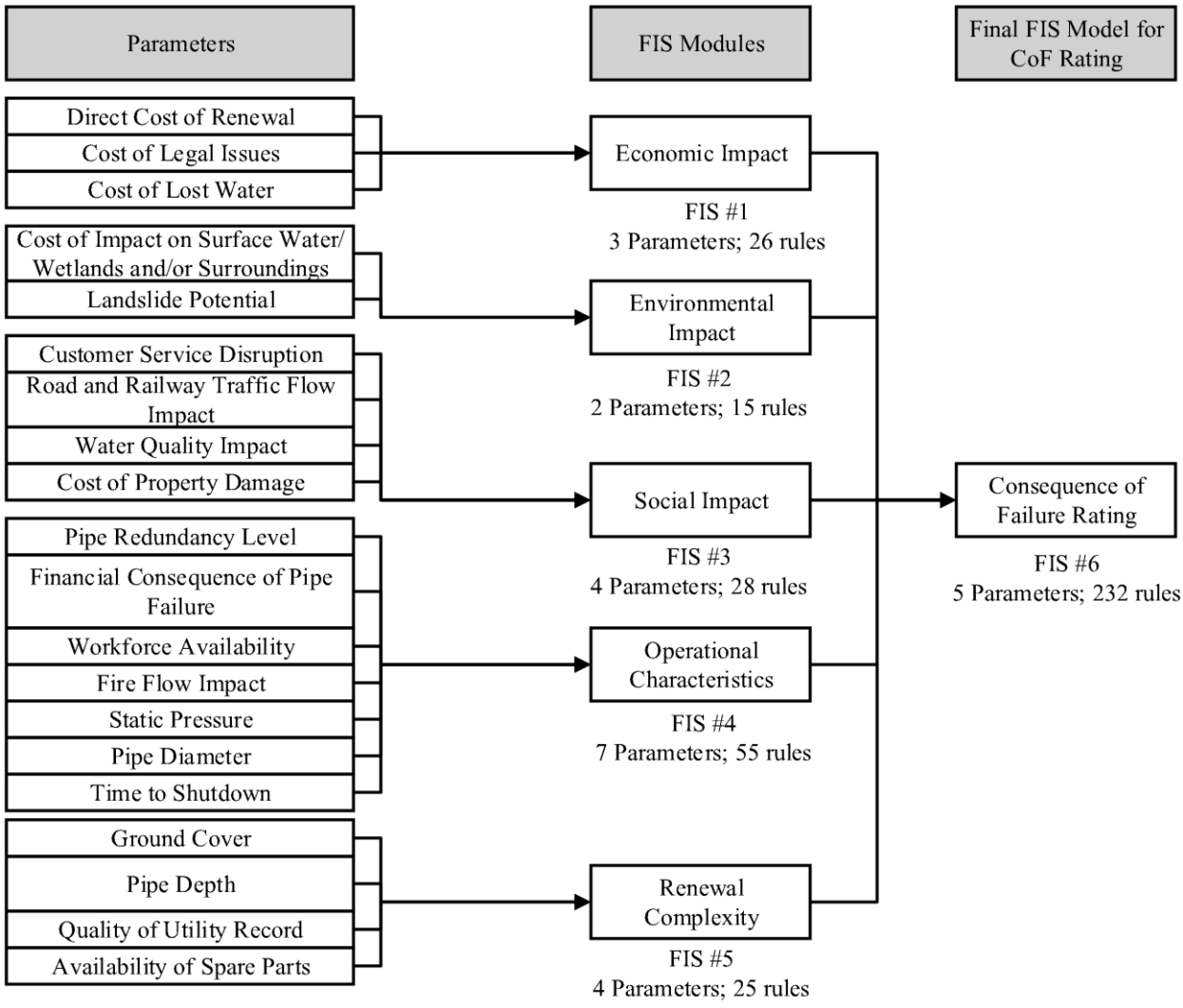

Figure 5-4: Hierarchical fuzzy teacher model for COF: Schematic of the knowledge-structured COF teacher, with input parameters grouped into five dimension-level fuzzy inference modules (Economic, Environmental, Social, Operational, Renewal-complexity) and a final fuzzy module that produces the overall COF rating.

5.6.1 Membership Functions and Input Space

Each COF input is defined on a universe of discourse (its plausible numerical range) and partitioned into overlapping linguistic terms. For continuous cost-like variables such as *Direct cost of renewal and Cost of lost water*, the universe spans from zero to an upper bound that covers almost all observed values plus a margin for extreme events. This range is partitioned into sets such as *Very low, Low, Medium, High and Very high* using simple shoulder and triangular membership functions. The breakpoints are aligned with cost thresholds from the COF band definition table and with empirical quantiles from utility data, so that transitions between linguistic levels occur near the points where Insignificant, Minor, Moderate, Major, and Catastrophic bands begin to separate.

For ordinal indicators already expressed on a 1–5 scale (for example, *Customer service disruption, Water-quality impact, Workforce availability, Records quality*), the universe of discourse is [0,5]. Fuzzy sets are centered near each conceptual category, with overlapping triangular, gaussian, or trapezoidal functions that ensure smooth transitions. A value of 3 on the disruption scale, for example, is mostly *Wet business* but has non-zero membership in adjacent categories, allowing the inference engine to interpolate rather than flip at crisp thresholds.

Figure 5-5 illustrates this design for the ground-cover variable used in the Renewal-complexity module. The universe of discourse is mapped to three linguistic terms—Open_Space, Roads_and_Railways, and Buildings_and/or_Water_Surface using one left-shoulder, one central gaussian bell, and one right-shoulder set.

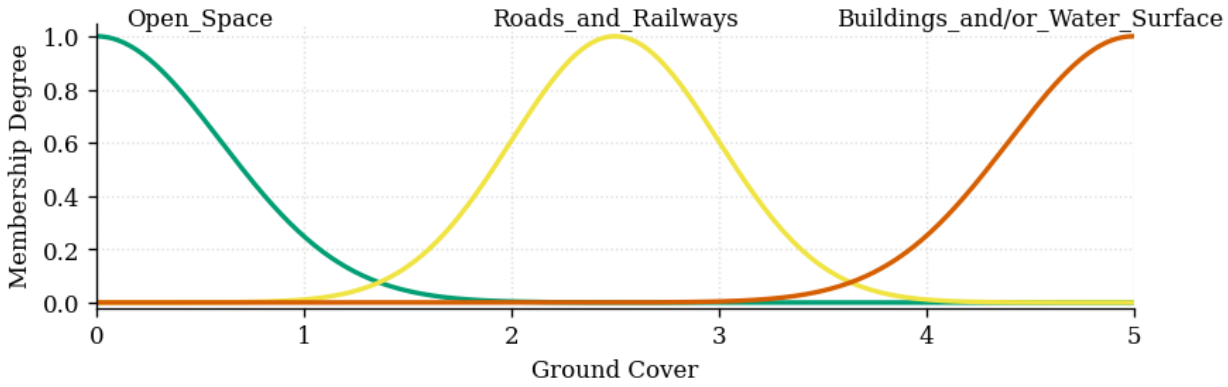

Figure 5-5: Membership functions for ground cover parameter. Fuzzy sets for the ground-cover indicator: Open_Space (left-shoulder), Roads_and_Railways (central gaussian bell), and Buildings_and/or_Water_Surface (right-shoulder), defined on a 0–5 linguistic scale.

Figure 5-6 shows the membership functions for customer service disruption in the Social module. Here the 0–5 axis is partitioned into No_Customers, Residential, Dry_Business, Wet_Business, and Critical_Customers, reflecting increasing density and criticality of affected users and allowing the fuzzy system to distinguish between, for example, a small break in an open field and an outage affecting dialysis centers or hospitals.

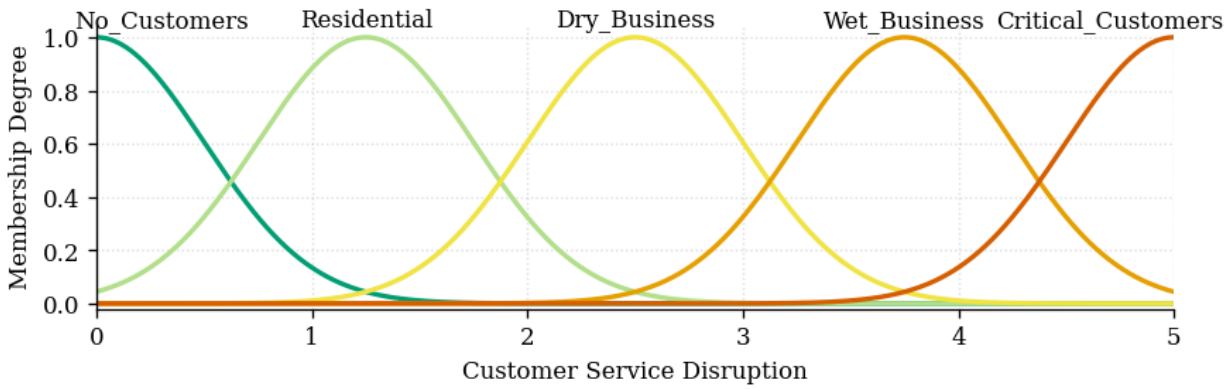

Figure 5-6: Membership functions for customer service disruption. Fuzzy sets for the customer service disruption indicator on a 0–5 scale, capturing settings from No_Customers through Residential, Dry_Business, Wet_Business, up to Critical_Customers.

Across all inputs, the membership functions are chosen to satisfy three design criteria. First, they cover the input space without gaps and every feasible input has at least one non-zero membership degree. Second, they are monotone in severity for COF-increasing variables, so increasing cost, traffic impact, depth, or environmental sensitivity never decreases membership in higher-severity terms. Third, overlap between adjacent sets is limited to roughly one third of their support, which improves interpolation while keeping the linguistic categories distinguishable. At the dimension level, each of the five indices (Social, Economic, Environmental, Operational, Renewal-complexity) is represented as a fuzzy variable on [0,5] using the same five sets as the overall COF bands, which keeps the outputs directly comparable.

5.6.2 Rule-base and IF–THEN Mechanics

The COF rule-base is organized in two layers matching the architecture in Figure 5-4. In the first layer, separate rule sets are defined for each dimension. Rules have the generic form:

IF (Antecedent_1 is Term_a) AND/OR (Antecedent_2 is Term_b)… THEN (Dimension_X_Impact is Level_L).

For the Economic dimension, antecedents include *Direct cost of renewal*, *Cost of legal issues*, and *Cost of lost water*. A representative subset of the Economic rule base is shown in Table 5-9, where combinations of low, medium, and high input costs map to Very low through Very high economic impact levels.

Table 5-9: Snapshot of the fuzzy rule base for the Economic dimension, showing how combinations of Direct cost of renewal, Cost of legal issues, and Cost of lost water map to the linguistic Economic impact level.

Input Parameters			**Output Parameter**
Direct Cost of Renewal	**Cost of Legal Issues**	**Cost of Lost Water**	**Economic Impact**
Very Low	Low	Low	Very Low
Low	Low	Low	Very Low
Medium	Low	Low	Low
.	.	.	.
.	.	.	.
Medium	Medium	High	Medium
High	Medium	High	High
Very High	NA	NA	Very High

In the Social dimension, rules combine Customer service disruption, Road and railway traffic flow impact, Water-quality impact, and Cost of property damage. Typical rules are:

- IF customer disruption is High AND traffic impact is High, THEN Social impact is Major.
- IF customer disruption is High OR water-quality impact is High, THEN Social impact is at least Moderate.

The Environmental dimension is governed by rules involving *Cost of impact on surface water and wetlands*, *Potential for landslides*, and *Greenhouse gas (GHG) emissions* associated with the failure and its response. Typical rules are:

- IF wetland/surface-water impact is High OR landslide potential is High OR Greenhouse gas is Moderate, THEN Environmental impact is Major.

The Operational dimension aggregates Pipe redundancy level, financial affordability of the utility, Workforce availability, Static pressure, Fire-flow impact, and related indicators. An example rule is:

- IF static pressure is High AND redundancy is Low AND workforce availability is Low, THEN Operational impact is Major.

The Renewal-complexity dimension uses rules involving Ground cover, Pipe depth, Quality of utility records, and Availability of spare parts, for example:

- IF land cover is Buildings and/or water surface AND depth is High AND records quality is Poor, THEN Renewal complexity is Major or Catastrophic.

Within each rule, the AND operator is implemented as the minimum of the relevant membership degrees, while OR is implemented as the maximum. This choice enforces a conservative aggregation of conditions where multiple severe antecedents must all be active to fully trigger a high-consequence rule, whereas any one strongly active antecedent is sufficient to activate rules formulated with OR, reflecting the existence of multiple alternative pathways to high consequence. In the second layer, a compact rule-base maps the five dimension-level indices to the overall CoF level. Rules follow a risk-averse structure, for example:

- IF Economic impact is Catastrophic OR Social impact is Catastrophic, THEN overall COF is Catastrophic.
- IF all dimensions are Minor or below, THEN overall COF is Insignificant or Minor.

- IF one dimension is Major and the others are Moderate, THEN overall COF is Major.

This near-max behavior ensures that catastrophic outcomes in any single critical dimension are visible in the final COF rating, while still allowing moderation when one dimension is high but others are negligible.

5.6.3 Inference, Interpolation, and Defuzzification

Both layers use a standard Mamdani fuzzy inference scheme. For a given segment, all input variables are first fuzzified into membership degrees for their respective linguistic terms. Each rule's antecedent degree is then computed using the chosen AND/OR operators (implemented as the minimum or maximum of the relevant memberships). The consequents are fuzzy sets defined on the corresponding dimension-level or overall COF universe. Consequents are scaled by the rule's firing strength and aggregated across all active rules using the maximum operator to produce a single fuzzy output set for each dimension and, in the second layer, for the overall COF.

To obtain a single numerical index, each fuzzy output set is defuzzified using the centroid-of-area operator, yielding a continuous value in [0,5]. At the dimension level, this produces Social, Economic, Environmental, Operational, and Renewal-complexity COF indices. The second-layer fuzzy system takes these five indices as inputs and produces a

continuous overall COF index, COF $\in$ [0,5]. This value is then mapped to ordinal bands using the same intervals as in the detailed COF definition table: [0,1) →Insignificant, [1,2) →Minor, [2,3) →Moderate, [3,4) →Major, and [4,5] →Catastrophic.

5.7 Evaluation, Verification, and Validation

This section explains the Evaluate, Verify, and Validate (EVV) process for the COF teacher–student models. Evaluation refers to development-time diagnostics done by the modeler based on stress-testing the fuzzy COF formulation and the Deep MLP student on large synthetic and utility datasets, checking that each dimension (customers, traffic, damage, priority customers, renewal complexity) moves in the expected direction and that extreme scenarios behave sensibly. Verification then tests whether the student learner has learned this mapping. We test fidelity to the fuzzy teacher across material–diameter cohorts and check that the student preserves the intended ordering of COF bands and dimensional trade-offs. Finally, Validation compares the verified student COF bands against external benchmarks that matter in practice that are, incumbent utility COF indices, structured feedback from asset-management staff and field crews, and ground-truth COF bands reconstructed from detailed main-break reports.

5.7.1.1 Development of Verification and Validation Dataset

The COF evaluation, verification, and validation experiments draw on a harmonized multi-utility dataset assembled from 18 large and mid-size water systems across the United States (anonymized as Utilities A–R). For each system we extracted pipe-segment tables from the asset inventory and spatially joined the contextual drivers in a geospatial database used by the fuzzy COF teacher (e.g., land use, proximity to critical customers, road importance, redundancy). Service lines were excluded and all counts and diameters refer to water mains only. After basic cleaning (ID harmonization, unit checks, removal of obviously inactive records), the combined dataset contains approximately 2.7 million pipe segments representing about 68,000 miles of mains on which the student and teacher COF models can be evaluated.

Table 5-10 summarizes the participating utilities and the parts of their networks included in the COF dataset. The utilities span a wide range of regional climates and terrains ranging from sub-arctic (Utility L) and New England urban cores (Utility N) through humid riverine and coastal plains (Utilities J, K, P) to semi-arid interior and karst regions (Utilities B, O, R). Diameter mixes also vary. Distribution-dominated systems such as Utilities H, M, P have roughly half to two-thirds of their mains below 8″,

whereas regional wholesalers and conveyance-heavy systems such as Utilities B and G are almost entirely >24″ pipe. This diversity is deliberate. It lets the COF model be tested under conditions ranging from small-diameter suburban grids to very large PCCP transmission mains where per-pipe consequences are extreme.

Table 5-10: Participating utilities and COF dataset coverage (Indices anonymized as Utilities A–R; diameters are main sizes, not services.)

Utility	Region & climate*	Approx. network length (miles)**	Samples in COF dataset (pipe segments)	Diameter mix (≈% <8″ / 8–24″ / >24″)	Notable features for COF modelling
Utility A	Humid subtropical, rolling topography	3,543	230,450	17 / 78 / 5	Rapidly growing suburban system with dense distribution grid and many pressure-reducing valves and inter-zone connections.
Utility B	Semi-arid inland plains, hot summers	≈1,240	43,345	0 / 0 / 100	Regional wholesale conveyance system dominated by long, large-diameter PCCP transmission mains under mixed rock and expansive clays; very high per-pipe consequence.
Utility C	Mid-Atlantic, dense urban/suburban mix	5,957	194,123	23 / 72 / 5	High-consequence trunk mains near highways, extensive PCCP break history, and an aggressive monitoring/condition assessment program.
Utility D	Cool, wet climate with steep terrain	2,099	129,361	41 / 57 / 2	Steep grades and landslide-prone slopes; legacy CI trunk mains in older neighborhoods.
Utility E	Coastal/inland mix with freeze–thaw	4,300	195,625	46 / 52 / 2	Mixed urban–suburban service, multiple sources, pockets of legacy AC in older suburbs.
Utility F	Coastal Mediterranean climate	4,190	188,242	49 / 47 / 4	Seismic faults, steep hills, coastal corrosion, large PCCP aqueducts feeding an urban grid.
Utility G	Arid to semi-arid, very high demands	830	43,345	5 / 13 / 84	High-capacity regional conveyance and long-distance aqueducts with high static pressures; large share of >24″ conveyance mains.
Utility H	Continental interior, moderate relief	3,022	70,660	56 / 40 / 5	Predominantly distribution mains with moderate frost depth; mix of old CI and newer PVC.
Utility I	Cold, high-elevation mountain front	5,509	248,826	32 / 67 / 0	Large elevation differences, complex pressure zones, freeze–thaw and corrosive soils in pockets.
Utility J	Warm coastal plain with high groundwater	6,400	193,808	36 / 60 / 4	Shallow-buried mains, high groundwater table, aggressive external corrosion, hurricane exposure.

Utility	Region & climate*	Approx. network length (miles)**	Samples in COF dataset (pipe segments)	Diameter mix (≈% <8″ / 8–24″ / >24″)	Notable features for COF modelling
Utility K	Humid riverine basin	4,177	207,745	42 / 55 / 3	Major river crossings, floodplain soils, and legacy CI in a historic downtown grid.
Utility L	Sub-arctic with permafrost pockets	846	26,364	20 / 71 / 8	Cold-climate system with frost heave, relatively young HDPE in new subdivisions.
Utility M	Mixed coastal/inland Mediterranean–semi-arid	5,927	57,667	48 / 51 / 1	Fragmented service areas, diverse soil types, pockets of legacy AC and CI.
Utility N	New England coastal, dense urban core	1,010	40,326	2 / 95 / 4	Very old CI/DI grid, deep utilities, heavy traffic loads, and constrained renewal corridors.
Utility O	Semi-arid, rapidly growing metropolitan area	5,716	270,648	35 / 60 / 5	High development pace, many recent PVC additions, variable soil corrosivity.
Utility P	Gulf-coast lowlands, very flat and wet	7,157	292,101	19 / 77 / 4	Soft, low-lying soils, flooding risk, large industrial demands, extensive small-diameter PVC.
Utility Q	Bay–delta fringe with Mediterranean climate	810	41,419	18 / 80 / 2	Mix of industrial and residential corridors, seismic and liquefaction zones.
Utility R	Semi-arid karst region	5,675	232,399	32 / 65 / 4	Karstic geology and sinkhole risk, long transmission from well-fields to city.
Total / range	—	≈68,000	2,706,454	≈0–56 / ≈13–95 / ≈0–100†	Climates from sub-arctic to subtropical; flat and mountainous terrain; high- and low-redundancy networks, including one wholesale system dominated by very large PCCP.

* Region and climate summarized at the system scale (not individual pressure zones). ** Approximate total length of water mains represented in the COF datasets; excludes service lines. † Lower end of the >24″ range is ≈0 % in distribution-dominant systems and 100 % in the wholesale conveyance system (Utility B).

Figure 5-7 shows the geographic distribution of these systems on a CONUS map, with Alaska inset. Core validation utilities (A–C, D, F, N) are highlighted separately from verification-only utilities. The core group are the systems for which we obtained additional comparators like utility COF indices, structured expert-opinion forms, or detailed main-break records, so they participate in both verification and the external validation tests

that follow. The other utilities contribute primarily to the development of a geographically distributed ML dataset. They broaden the range of climates, diameter categories, and operating environments used to test whether the COF model behaves sensibly when applied outside the development utility.

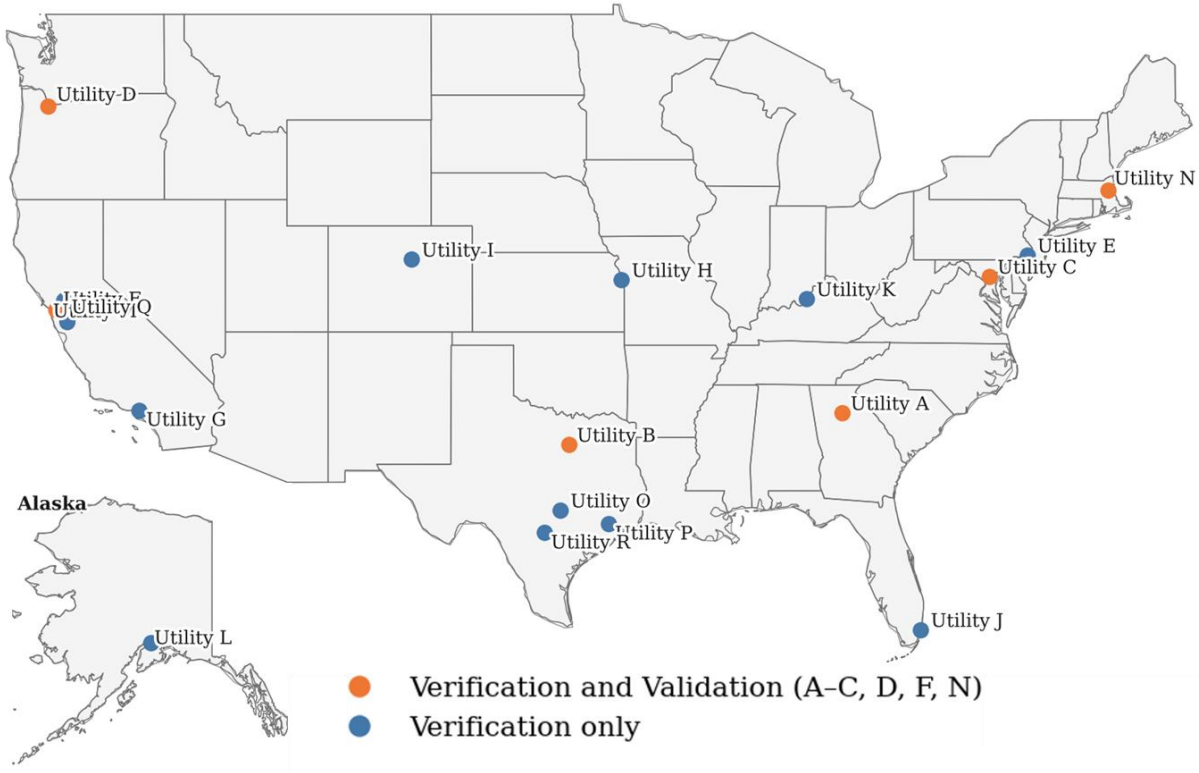

Figure 5-7: Anonymized locations of the 18 participating utilities (A–R) used in COF verification, and validation. Orange markers denote utilities that participate in both verification and validation (A–C, D, F, N); blue markers denote verification-only utilities that broaden the range of climates and network types in the student learner model dataset.

For the subsequent analyses, the combined dataset is used in three ways. First, the full A–R data corpus underpins verification where we check basic distributional properties,

confirm that high-risk combinations of drivers map to high COF bands across materials and diameters, and ensure that outlier behaviors are not tied to a single utility. Second, the core validation utilities (Utilities A-C and N) provide the paired COF indices needed for model-to-utility comparisons and expert-agreement studies. Third, one of these systems (Utility N) also supplies the detailed main-break reports used to build an independent ground-truth COF dataset.

5.7.2 Evaluation of the Teacher Model

Before the COF model can be used as a teacher for student models or as a criterion in multi-objective renewal planning, it must be evaluated to ensure it behaves as it was supposed to. The fuzzy expert system is not a statistical fit to a single dataset. It encodes expert knowledge, impact mechanisms, and proxy layers for social, environmental, operational, and renewal-complexity effects. The evaluation therefore asks three questions.

First task is to check representativeness and coverage. In this task, we check whether the membership functions, rule base, and resulting COF surfaces cover the relevant universe of discourse for each input, and do they map low-impact and high-impact configurations to appropriately low and high COF indices. In the second task, we check the model's sensitivity and robustness to see if the model responds in a stable and

monotone way to systematic perturbations of key drivers. The third task is about checking alignment of the model results for heuristic *common sense* scenarios. In this task, we check whether the model reproduces the expected outputs for trunk mains, ordinary distribution mains, and road/railway crossings when the model is exercised on impact motifs derived from material–diameter distributions and replacement-cost baselines.

These three tasks parallel the evaluation strategy used in the LOF chapter, but with a different emphasis. For COF, utilities rarely record full social or environmental consequences, and even economic impacts are often only partially captured in work-order systems. The COF teacher must therefore be judged primarily on whether it organizes the available data and proxy layers into a disciplined, mechanism-consistent mapping, not on a direct fit to observed impact time-series. Once this evaluation is in place, the same framework is used to design verification tests for student models and to structure the limited external validation against observed cost and disruption signals.

5.7.2.1 Representativeness and Coverage

The first evaluation step checks whether the fuzzy COF teacher model is representative of the intended impact space and whether it provides full and coherent coverage of each input's universe of discourse. This is assessed visually through membership

function ranges and numerically through extreme stress testing on the best, average, worst scenarios on the combined teacher model.

Figure 5-8 shows four representative membership-function panels extracted from the hard-coded fuzzy systems. Panel (a) displays the five output membership functions for the overall Consequence of Failure index, from Very Low to Very High on the 0–5 universe of discourse. Panels (b)–(d) show inputs that sit in different parts of the COF framework like Ground Cover in the Renewal Complexity module, and Time to Shutdown and Static Pressure in the Operational Impact module. In each panel, low-impact functions occupy the left side of the axis and are drawn in green, intermediate functions are shown in yellow or light green, and high-impact functions occupy the right side in orange and red.

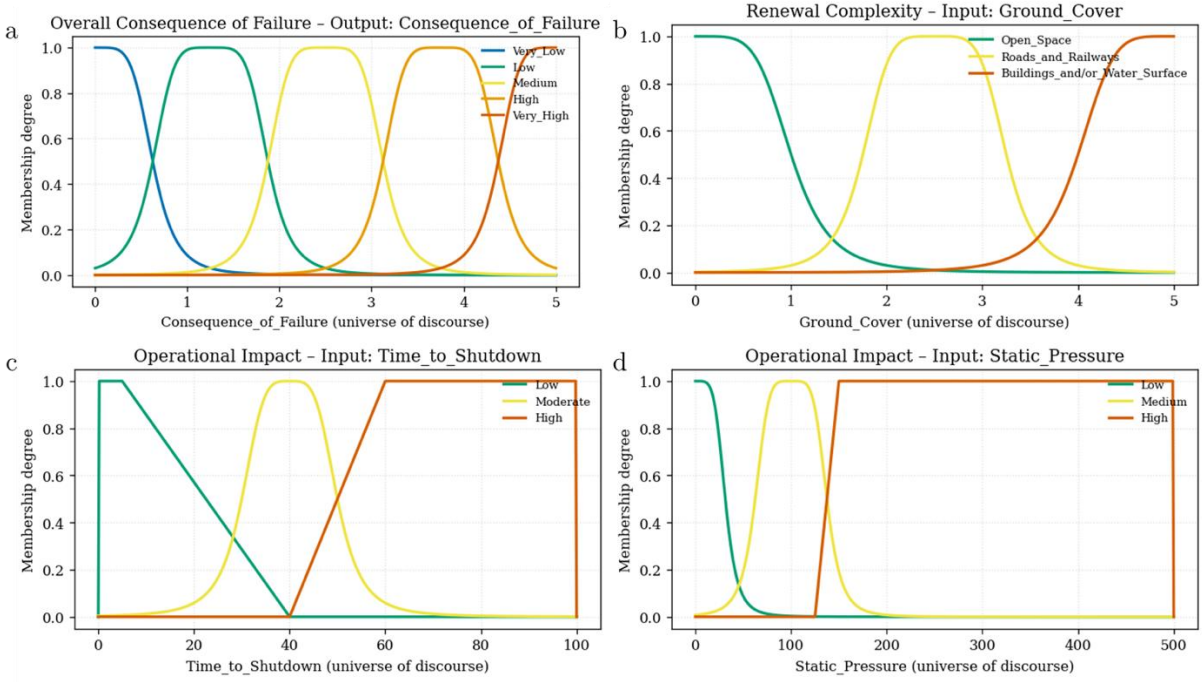

*Figure 5-8: Representative membership functions in the COF teacher model (a) Overall Consequence of Failure output bands (*Very Low *to* Very High*). (b)* Ground Cover *input in the Renewal Complexity module (open space, roads and railways, buildings and/or water surface). (c)* Time to Shutdown *input in the Operational Impact module (low, moderate, high). (d)* Static Pressure *input in the Operational Impact module (low, medium, high). Each set of fuzzy sets spans the full universe of discourse with smooth overlaps, from low-impact (green) to high-impact (red) regions.*

The fuzzy teacher model is subjected to the best–average–worst experiment to test coverage at the system level. For each fuzzy module, representative "best", "average", and "worst" input settings are constructed by locating the peak of the lowest, middle, and highest membership functions for every input variable. These settings approximate configurations in which all drivers are simultaneously favorable, typical, or unfavorable. Each

configuration is propagated through the five dimension-level fuzzy systems (Economic, Environmental, Social, Operational, Renewal Complexity) and then through the overall COF system. The resulting indices, shown in Figure 5-9 as grouped bars across the six outputs, are therefore direct model predictions for those three synthetic scenarios. The expected results are strict. The best scenario should yield the lowest indices across all dimensions and the overall COF, the worst scenario should yield the highest indices, and the average scenario should fall between them.

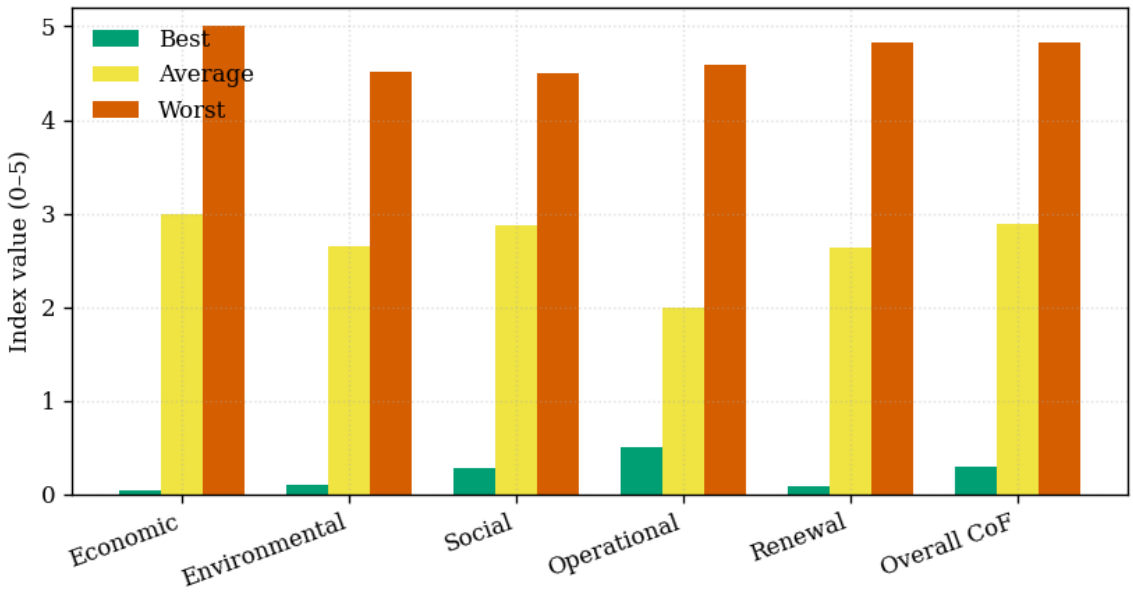

Figure 5-9: Teacher COF response to best, average, and worst input scenarios, showing normalized 0–5 index values for the five consequence dimensions and the overall COF. Green, yellow, and orange bars correspond to favorable, typical, and unfavorable input settings defined from the peaks of the membership functions in each fuzzy module.

Results: The membership-function panels in Figure 5-8 demonstrate that the teacher model provides full and interpretable coverage of the input and output spaces. The overall COF output partition in panel (a) spans the entire 0–5 universe of discourse (the range of possible values), with every point having non-zero membership in at least one band and with smooth overlaps between Very Low, Low, Medium, High, and Very High. This construction prevents artificial discontinuities in the index when a small change in inputs moves a segment across a band threshold. The Ground Cover functions in panel (b) encode the qualitative judgement that open space is associated with low renewal complexity, corridors dominated by roads and railways with intermediate complexity, and pipes under buildings or water surfaces with the highest complexity; the overlapping shapes ensure gradual rather than abrupt transitions between these contexts. The *Time to Shutdown* and *Static Pressure* functions in panels (c) and (d) illustrate the asymmetric treatment of benign versus hazardous regimes: short isolation times and low pressures retain high membership in the "Low impact" sets, while long isolation times and very high pressures transition smoothly into "High impact" sets. Similar visual checks were performed for all remaining inputs (presented in Appendix D), confirming that each variable's universe of discourse is covered without gaps and that the fuzzy sets reflect plausible engineering judgements about consequence drivers.

Figure 5-9 complements these local checks with a system-level coverage test using best–average–worst scenarios. When all COF-increasing drivers are set to favorable values, the teacher model yields consistently low indices across all five consequence dimensions and the overall COF (green bars clustered near zero). When they are set to unfavorable values, the model produces indices near the top of the 0–5 scale (orange bars), and the average configuration lies in the mid-range (yellow bars). The strict ordering of these three scenarios across every dimension shows that the fuzzy teacher covers both low-impact and high-impact corners of the multidimensional input space and maps them to well-separated regions of the COF scale, rather than collapsing them into a narrow intermediate band. This behavior supports the claim that the teacher model is a coherent, monotonic mapping from impact drivers to a 0–5 COF index, suitable both as a training signal for student models and as a foundation for the scenario-based evaluations in the next subsections.

5.7.2.2 Sensitivity and Robustness

This subsection evaluates how robust the COF teacher model is to changes in its inputs and whether the most influential drivers match engineering expectations. Two complementary approaches are used. First, we visually examine response surfaces from

the fuzzy inference system to verify that the combined rules and membership functions produce sensible, smooth behavior over the input space. Second, we apply a global variance-based sensitivity analysis (Sobol method) to quantify how uncertainty in each input contributes to uncertainty in the COF outputs, including interactions between parameters.

Surface-plot inspection

For each module, two-dimensional response surfaces were generated by sweeping a pair of inputs across their full universe of discourse while holding all other inputs at neutral "medium" settings. Figure 5-10 collects six representative surfaces, one per module for economic impact (a), social impact (b), renewal-complexity (c), environmental impact (d), operational impact (e), and the overall COF roll-up (f). Each surface shows the predicted impact index on the vertical axis, with a color scale repeating the same values for easier reading.

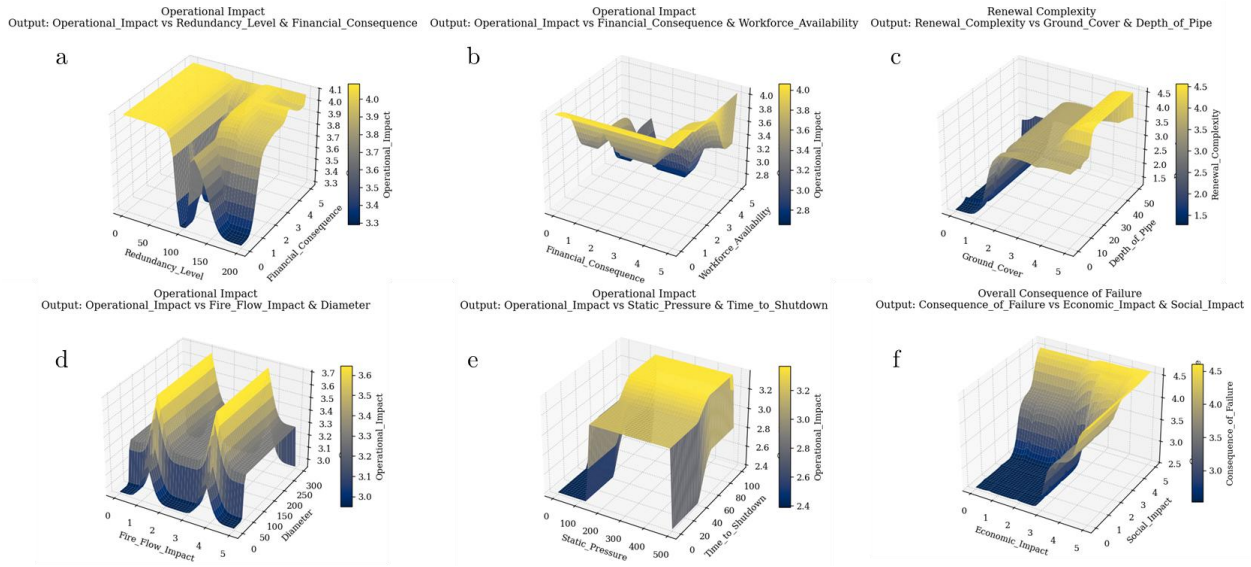

Figure 5-10: Representative response surfaces from the COF teacher model. Panels (a–e) show Economic, Social, Renewal-Complexity, Environmental, and Operational impact indices as functions of key driver pairs; panel (f) shows the overall COF index as a function of Economic and Social impact indices. Surfaces rise smoothly from low-impact to high-impact regions in directions consistent with engineering expectations.

In the Economic module (Figure 5-10 a), the surface rises from the low-cost corner toward high economic impact as both direct renewal cost and cost of legal issues increase and then flattens into a high-impact plateau. This reflects the expert judgement that once a failure exposes the utility to very high direct and secondary financial losses, the economic consequence should be consistently “High” or “Very High” rather than oscillating with small cost variations. The Social module surface (Figure 5-10 b) shows a similar pattern where combinations of low road-and-rail traffic flow impact and low property damage lie on a lower social-impact shelf, whereas high traffic disruption and high property damage

push the social index rapidly into the upper bands and then saturate. This is consistent with the idea that heavily trafficked corridors and dense built environments dominate social consequences. The Renewal-Complexity surface (Figure 5-10 c) combines ground cover and depth of pipe. Shallow pipes in open or lightly built areas occupy the lowest complexity region, while deeper pipes under buildings or water surfaces move into steeply rising ridges and then a high-complexity plateau. This captures both access difficulty and shoring/permits complexity. In the Environmental module (Figure 5-10 d), environmental impact increases as both the cost of impact on surface water/wetlands and the potential for landslides increase. Segments with negligible aquatic exposure and low landslide potential lie near the baseline, whereas segments that can contaminate sensitive receptors or destabilize slopes are mapped to higher environmental indices. The Operational module surface (Figure 5-10 e) varies static pressure against time-to-shutdown. Low pressure and short isolation times are mapped to low operational impact, while combinations of very high pressure and long shutdown times produce a pronounced high-impact plateau, reflecting both increased water loss and operational difficulty. Finally, the overall COF surface (Figure 5-10 f) shows how the Economic and Social dimension indices interact in the second-layer fuzzy system. Low–low combinations sit near the base of the COF scale; high economic or social scores produce mid-range COF; and simultaneous high Economic

and Social scores drive the surface into the upper COF bands associated with Major and Catastrophic consequence.

Across all six panels, we checked three properties by inspection: (i) monotonic trends in the expected directions (e.g., higher costs or risks never reduce the corresponding impact index); (ii) smooth transitions without artificial cliffs where small input changes cause large discontinuities; and (iii) plausible interaction patterns, such as redundancy or shallow depth moderating otherwise severe conditions. Where any surface suggested non-monotonic or counter-intuitive behavior, the underlying rules or membership functions were revisited and, if needed, adjusted.

Global variance-based sensitivity analysis

To complement these visual checks, a global variance-based sensitivity analysis (Sobol method) was used to quantify how much each input contributes to output variability in the COF teacher model (Saltelli 2002; Sobol' 1990). The Sobol framework decomposes the variance of a model output Y into contributions from individual inputs and their interactions. The first-order sensitivity index for input X_i is

$$S_i = \frac{V_i}{Var(Y)} \tag{5}$$

where $V_i$is the portion of output variance attributable to X_i alone and $\mathrm{Var}(Y)$ is the total output variance. Higher-order indices $S_{ij}, S_{ijk}, \ldots$ capture interaction effects between two or more inputs, but their number grows as $2^d - 1$ for $d$inputs and quickly becomes impractical (Gan et al. 2014).

Instead of estimating all interaction terms explicitly, this study uses the total-order sensitivity index S_{Ti}, which measures the combined contribution of input X_i and all its interactions of any order (Homma and Saltelli 1996):

$$S_{Ti} = \frac{E_{X_{\sim i}}\left(Var_{X_i}(Y|X_{\sim i})\right)}{Var(Y)} \tag{6}$$

where $X_{\sim i}$denotes all inputs except X_i. Intuitively, S_{Ti} answers the question "If I could fix X_i, how much would the overall output variance be reduced?" Input samples were generated using Saltelli's extension of Sobol sampling, which efficiently estimates both first order and total-order indices from the same experimental design. For each module (Economic, Environmental, Social, Operational, Renewal-Complexity) and for the overall COF system, we treated the normalized input ranges as uncertain and drew $n(2k + 2)$ model evaluations, where $k$is the number of inputs to that module and n is the number of base samples required for stable estimation. For each index, 95% confidence intervals

were computed using replicate estimates across randomized sample blocks. Environmental-module indices are not reported because the standard Sobol implementation requires at least three inputs; that module currently has only two.

The resulting total-order indices for the most influential parameters are summarized in Table 5-11. At the module level, these indices indicate which variables most strongly control each dimension-level COF index and at the system level, they identify which dimension scores most strongly drive the overall COF.

Table 5-11: Sensitivity Indices for Influential Parameters (Vishwakarma & Sinha 2023)

Module	Influential Parameters	SI_T Module Output
Economic Impact	Direct Cost of Renewal	0.02
	Cost of Legal Issues	0.05
	Cost of Lost Water	0.001
Environmental Impact	Cost of Impact to Surface Water and/or Surroundings	NA*
	Potential for Landslides	NA*
Renewal Complexity	Ground Cover	0.06
	Depth	0.001
	Quality of Utility Record	0.02
	Availability of Spare Parts	0.001
Social Impact	Customer Service Disruption	0.05
	Road and Rail Traffic Flow Impact	0.002
	Water Quality Impact	0.01
	Cost of Property Damage	0.001
Operational Characteristics	Redundancy Level	0.06
	Financial Consequence	0.02
	Workforce Availability	0.001
	Fire Flow Impact	0.001
	Static Pressure	0.001
	Diameter	0.001
	Time to Shutdown	0.001
Consequence of Failure	Economic Impact	0.05
	Environmental Impact	0.001
	Social Impact	0.02

	Operational Characteristics	0.001
	Renewal Complexity	0.001

*Module sensitivity could not be calculated because methodology requires a minimum of 3 parameters. SI_T represents the total order sensitivity index.

Results: Taken together, the surface plots and Sobol indices indicate that the COF teacher model is both behaviorally plausible and reasonably robust. The surfaces in Figure 5-10 show that outputs change smoothly and monotonically in response to key drivers. Economic impact increases with higher direct and legal costs. Social impact rises with increasing traffic disruption and property damage. Renewal complexity grows with depth and constrained ground cover. Environmental impact increases with greater exposure of surface waters and landslide potential and operational impact worsens for combinations of high static pressure and long shutdown times. The overall COF surface responds most strongly when both Economic and Social dimension scores are high, as expected.

The sensitivity indices in Table 5-11 refine this picture by quantifying which inputs matter most. Within the Economic module, cost of legal issues and direct cost of renewal exhibit the highest total-order indices, consistent with the idea that large direct and secondary financial exposures dominate economic consequences. In the Renewal-Complexity module, ground cover and quality of utility records are more influential than depth alone, reflecting the fact that working under buildings or water surfaces and in poorly

documented corridors drives a disproportionate share of complexity. In the Social module, customer service disruption and traffic flow impact are key, aligning with expectations that loss of service to dense or critical customers and major traffic corridors dominates social consequences. For the Operational module, redundancy level and financial consequence stand out, indicating that both network topology and the economic stakes of outages strongly shape operational impact. At the overall COF level, the Economic and Social dimension indices have the largest total-order sensitivity, confirming that these two dimensions are the primary levers of consequence in the current formulation.

Importantly, no single parameter or module displays a sensitivity index approaching one, and several inputs have small but non-zero indices. This pattern suggests a model in which consequences are driven by a combination of factors rather than by a single extreme driver, and in which moderate changes in any one parameter do not cause unstable jumps in COF. The teacher model is therefore suitable as a reference for student models and for subsequent robustness experiments where it encodes a multi-dimensional, interaction-aware mapping from impact drivers to a 0–5 COF index while remaining stable under global perturbations of its inputs.

5.7.2.3 Scenario Alignment

The final evaluation step for the COF teacher model is a scenario-alignment test. The aim is to check face validity of the teacher model. This test checks the direction and magnitude of the model predictions when we pass consequence scenarios that every experienced operator would recognize as "catastrophic" or "minor". Following the approach used for the LOF model, we assemble a compact set of heuristic scenarios drawn from literature, utility incident reports, and structured interviews. Each scenario combines the main COF dimensions in an interpretable way. Examples include a high-cost urban trunk main rupture, a spill into an environmentally sensitive wetland, and a prolonged outage near a critical facility.

For each scenario, we first express the narrative in terms of the COF input variables. For instance, the "high-cost urban main rupture" is translated into very high direct renewal cost, high cost of legal issues, high cost of lost water, severe customer service disruption, severe traffic impact, and high renewal complexity due to location under a major road. The resulting input vector is then scored by the five dimension-level fuzzy systems (Economic, Environmental, Social, Operational, Renewal Complexity) and by the overall COF system. We record the continuous COF index on the 0–5 scale, the

corresponding linguistic band (Insignificant to Catastrophic), and the qualitative rationale in terms of which rules fire most strongly and why.

Alignment is evaluated along three axes to check "directional accuracy" (Kohavi and Thomke 2017). First, the predicted COF band should match the common knowledge expectation for that scenario (for example, "Catastrophic" for a high-cost urban trunk main failure, "Minor" for a short repair of a small pipe with limited impact). Second, the direction relative to a baseline case should be correct. The baseline is a typical 8–24 in distribution main in a mixed residential–commercial corridor with moderate cost, limited traffic disruption, and no special environmental sensitivity, which maps to a Moderate COF around 2.5. Scenarios that are clearly worse than this baseline should move the index upward, while clearly less severe scenarios should move it downward. Third, the distance on the 0–5 scale should be reasonable. Materially different situations should differ by a visible margin (for example, at least 0.5–1.0 units), not by a few hundredths of a point. The full set of COF scenarios, their model-predicted indices and bands, and their qualitative classification relative to the baseline are shown in Figure 5-11 and with details in Table 5-12.

Any misalignment triggers a trace-back exercise. We inspect the active rules, and membership supports that dominate the score, check the consistency of units and ranges, and revisit thresholds or overlaps if needed. Edits are limited to coherent shifts of membership support and/or rule consequents. This ensures that improvements made to fix one scenario do not create unintended distortions elsewhere in the input space.

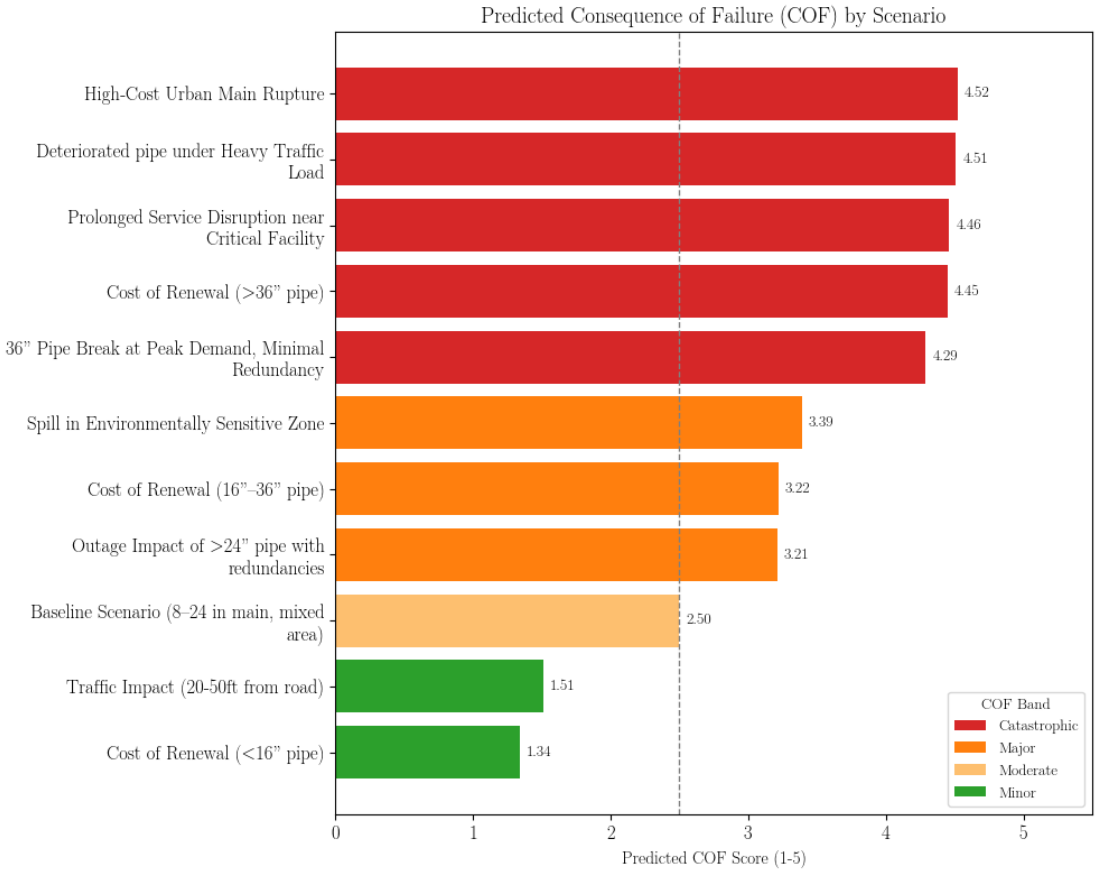

Figure 5-11: Predicted Consequence of Failure (COF) scores for 11 heuristically chosen scenarios, compared against a baseline COF of 2.5 (Moderate). Scores are color-coded by severity band.

Results: Across the curated COF scenarios in Table 5-12, the teacher model behaves in a directionally consistent way. High-stakes events such as a high-cost urban main rupture, a prolonged outage at a critical facility, or a 36 in trunk-main break at peak demand produce COF indices in the upper part of the scale (for example, 4.29–4.52) and are classified as Catastrophic, as expected. Scenarios dominated by environmental exposure, such as a spill into a protected wetland, fall in the Major range (for example, 3.39) with Environmental and Social dimensions driving the score. Lower-impact scenarios, such as renewal of a small-diameter pipe with modest cost or traffic disruption, land in the Minor or lower-Major bands (for example, indices around 1.3–3.2), below the baseline Moderate case. In all cases, shifts in linguistic severity correspond to monotonic shifts in the predicted COF index that is, scenarios that are clearly worse than the baseline move upward on the 0–5 scale, while milder cases move downward. No tensions were found between expert expectation and model output. Overall, the scenario-alignment exercise supports the face validity of the COF teacher model and provides a set of transparent test cases that will later be reused when evaluating student models and cross-utility applications.

Table 5-12: Consequence of Failure (COF) Analysis of Theoretical Scenarios. Scenarios are scored against a 1-5 index, with directionality indicating whether the consequence is higher or lower relative to the COF=2.5 baseline.

Scenario Description	Predicted COF (0–5)	COF Band	Direction vs baseline (≈2.5)
High-Cost Urban Main Rupture: Major break in densely populated area. Repair cost >$500k, water loss >20,000 gal/hr, and >100 customers lose supply.	4.52	Catastrophic	Higher (↑)
Spill in Environmentally Sensitive Zone: Medium-diameter main leaks into protected wetland/habitat. Triggers substantial cleanup costs and environmental fines.	3.39	Major	Higher (↑)
Prolonged Service Disruption near Critical Facility: Failure affecting hospital/industry. Outage >24 hours affecting >5,000 critical customers.	4.46	Catastrophic	Higher (↑)
Deteriorated pipe under Heavy Traffic Load: Pipe under major highway. Repair complicated by lane closures; elevated social cost from traffic disruption.	4.51	Catastrophic	Higher (↑)
36" Pipe Break at Peak Demand, Minimal Redundancy: Occurs during heat wave/festival. No quick reroute, water losses >$300k, outage >12 hours.	4.29	Catastrophic	Higher (↑)
Outage Impact (hours or customer count) of >24" pipe with redundancies.	3.21	Major	Higher (↑)
Baseline Scenario: A typical 8–24 in distribution main in a mixed residential–commercial corridor with moderate cost, limited traffic disruption, and no special environmental sensitivity.	2.50	Moderate	Near baseline (↔)
Traffic Impact (hours of road closure) for a pipe failure 20–50ft away from a road.	1.51	Minor	Lower (↓)
Cost of Renewal for fixing the failure of <16" pipe (~$20,000).	1.34	Minor	Lower (↓)
Cost of Renewal for fixing the failure of 16"–36" pipe (~$70,000).	3.22	Major	Higher (↑)
Cost of Renewal for fixing the failure of >36" pipe (~$800,000).	4.45	Catastrophic	Higher (↑)

5.7.3 Verification: Supervised Training of Student Models on Fuzzy COF Teacher I/O

This section verifies that a *student* machine learning model can faithfully reproduce the mapping implemented by the hierarchical fuzzy *teacher* for COF. In earlier subsections, we evaluated the internal logic of the fuzzy COF system itself (membership functions, rule bases, sensitivity checks, and scenario stress tests). Here, the question is different. Given only the teacher's input–output pairs, can a neural network learn to predict the same COF band that the fuzzy system would assign? This verification is a prerequisite for later *validation* against external utility evidence, because it ensures that any differences we observe in those experiments are due to real disagreement between the teacher and the world, not an artefact of a poorly trained student.

5.7.3.1 Training data and prediction task

We assembled a pooled COF dataset at the pipe-segment level from multiple large U.S. water utilities. For confidentiality, all systems are anonymized as Utilities A–Q, spanning both wholesale and retail providers across multiple states and climate zones. Together, these systems contribute approximately 2.7 million labelled segments (N =

2,706,454). Per-utility sample counts and fractions of the total are summarized in Table 5-10.

For each pipe segment, the fuzzy COF teacher outputs a small, structured feature vector and label. The inputs consist of five dimension-level COF indices. These include Economic Impact, Environmental Impact, Social Impact, Operational Impact, and Renewal Complexity and each are expressed as a continuous score on a 0–5 scale. From these, the teacher computes an overall COF index and assigns a corresponding discrete COF band in $\{0,1,2,3,4\}$, where 0 denotes Insignificant, 1 Minor, 2 Moderate, 3 Major, and 4 Catastrophic consequences.

Verification is framed as a supervised multi-class classification problem. The input to the student is the 5-dimensional vector (E_{CoF},En_{CoF},S_{CoF},O_{CoF},R_{CoF}) and the target is the overall COF band chosen by the teacher. This setup matches how utilities act on COF that is, for renewal planning they work with discrete consequence bands rather than raw continuous scores.

5.7.3.2 Candidate student architecture and training protocol

To avoid over-committing to a single modelling choice, we first conduct a screening experiment across six standard supervised learners that span linear, tree-based, and neural

architectures. The candidate models are a multinomial Logistic Regression (LR), an RBF-kernel Support Vector Machine (SVM), a Random Forest (RF), an XGBoost Gradient-Boosted tree model (XGB), a shallow three-layer Multi-Layer Perceptron (MLP), and a deeper five-layer MLP that will ultimately serve as the student architecture used throughout. Logistic regression provides the simplest linear baseline with softmax outputs and transparent coefficients. The SVM offers a margin-based nonlinear classifier capable of representing smooth decision boundaries in the five-dimensional COF space. Random forests bring robustness through ensembles of decision trees and handle basic feature interactions. XGBoost represents the current strong baseline for tabular data with complex nonlinearities. The two neural networks extend this spectrum by allowing richly nonlinear combinations of the five COF dimensions, with the shallow and deep variants differing in depth and representational capacity.

Both MLPs follow the same design pattern in their hidden layers with a linear transformation, followed by a normalization layer, a Rectified Linear Unit (ReLU) activation, and optional dropout. ReLU is a standard activation function that keeps gradients in check by zeroing negative inputs while leaving positive activations unchanged. Dropout randomly silences a small fraction of hidden units (here roughly 8%) during training and

this forces the network to avoid relying on any single pathway and reduces overfitting. For typical batch sizes we use Batch Normalization, which rescales and recenters activations within each mini-batch, while Layer Normalization is reserved for any very small-batch settings. All network parameters are optimized with AdamW, an adaptive variant of stochastic gradient descent that combines per-parameter learning rates with weight decay, i.e., a small L2 penalty that discourages excessively large weights. As in the LOF work, training is stabilized by a coherent package of techniques with label smoothing of 2–4% to prevent the model from becoming over-confident at band boundaries, gradient-norm clipping to limit extreme parameter updates, and an exponential moving average of the weights to produce smoother and more stable decision surfaces.

For the screening stage we check the performance metrics of all the selected algorithms on the entire training dataset. Once the Deep MLP emerges as the most accurate and well-behaved candidate, we retrain this selected student on the full teacher-labelled dataset (N = 2,706,454) so that it can exploit all available information before any external validation. Performance across the six models on this full training set is summarized in Table 5-13.

Table 5-13: Screening performance of candidate CoF student models (training set, N = 2,706,454)

Model	Accuracy	Macro-Precision	Macro-Recall	Macro-F1
Logistic Regression (LR)	0.86	0.83	0.86	0.84
RBF-kernel SVM (SVM)	0.88	0.85	0.88	0.86
Random Forest (RF)	0.90	0.87	0.90	0.89
XGBoost (XGB)	0.92	0.89	0.92	0.90
Shallow MLP (3-layer)	0.91	0.88	0.91	0.90
Deep MLP (5-layer, selected)	0.94	0.92	0.94	0.93

We report accuracy, defined as the fraction of segments for which the predicted band matches the teacher, and three macro-averaged metrics, namely macro-precision, macro-recall, and macro-F1. For each band separately, we compute precision (how cleanly the model predicts that band), recall (how completely it recovers that band), and F1 (the harmonic mean of precision and recall). We then average these scores over the five bands, so that rare high-consequence bands influence the evaluation as much as common low-consequence bands. With this dataset, logistic regression reaches an accuracy of 0.86 with macro-F1 of 0.84 (macro-precision 0.83, macro-recall 0.86). The SVM improves these values to 0.88 accuracy and 0.86 macro-F1. Random Forest reaches 0.90 accuracy and 0.89 macro-F1, while the shallow three-layer MLP achieves 0.91 accuracy and 0.90 macro-F1. XGBoost attains 0.92 accuracy with macro-F1 of 0.90. The five-layer Deep MLP achieves

the best performance, with 0.94 accuracy, 0.92 macro-precision, 0.94 macro-recall, and 0.93 macro-F1, indicating that it captures subtle nonlinear interactions among the five COF dimensions while remaining compact enough for deployment.

To probe generalization beyond the empirical teacher dataset, we also construct a synthetic COF test set of 1,000 samples, with 200 cases in each of the five COF bands. These synthetic examples are designed to cover the five-dimensional COF input space more uniformly and to concentrate on band boundaries, which are the most challenging parts of the decision surface, following the same spirit as the boundary-focused synthetic testing used for LOF in Chapter 4 (see Table 4-8). Only the Deep MLP is evaluated on this synthetic set, since it is the chosen student model. On this test the network achieves an accuracy of 0.86, with macro-precision 0.87, macro-recall 0.86, and macro-F1 0.86. The reduction relative to the training metrics is consistent with the deliberately more difficult, boundary-focused nature of the synthetic test and indicates that the student retains high fidelity to the fuzzy teacher even away from the exact empirical distribution of the original utility data.

Table 5-14: Performance comparison of the screened Deep MLP COF student in training vs testing

Split	Accuracy	Macro-Precision	Macro-Recall	Macro-F1
Training (N = 2,706,454)	0.94	0.92	0.94	0.93
Synthetic test (N = 1,000)	0.86	0.87	0.86	0.86

5.7.3.3 Confusion-matrix diagnostics

Scalar summary metrics are useful, but they do not show where errors occur. For decision support in utilities, an occasional confusion between, say, "Moderate" and "Major" COF is tolerable, whereas confusing "Insignificant" with "Catastrophic" would be unacceptable. To make this pattern visible, we examine confusion matrices for all models. A confusion matrix is a two-dimensional table where rows correspond to true bands (the teacher's labels) and columns to predicted bands. Entry (i, j) contains the number of segments whose true band is i but are predicted as band j. We row-normalize each matrix so that, in addition to the raw count, each cell also reports the percentage of segments in that true band that are assigned to each predicted band.

Figure 5-12 shows a panel of six training confusion matrices, one for each candidate model (logistic regression, SVM, random forest, XGBoost, shallow MLP, and Deep MLP).

All matrices share the same color scale, so diagonal dominance and residual error structure can be compared visually.

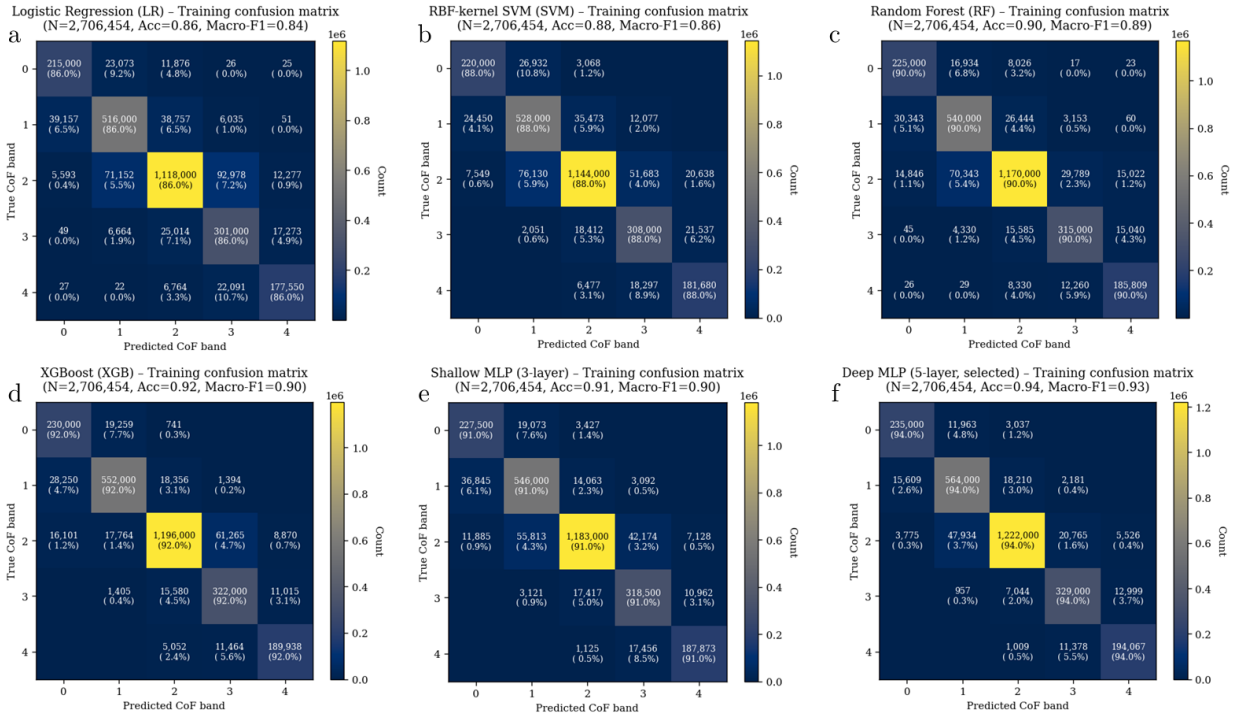

Figure 5-12: Panel of six training confusion matrices, one for each candidate model (LR, SVM, RF, XGB, Shallow MLP, Deep MLP). All matrices share the same color scale, so diagonal dominance and residual error structure can be compared visually

Across models, the main diagonal is clearly dominant and consistent with the metrics in Table 5-13, overall accuracies range from 0.86 for logistic regression to 0.94 for the Deep MLP, and in each panel the bulk of the mass lies in the diagonal cells. Off-diagonal mass is heavily concentrated in adjacent bands (for example, true band 2 assigned to band

1 or 3). These correspond to borderline cases where the five dimension-level scores lie near fuzzy rule thresholds and the teacher's band assignment is itself somewhat fragile. Only logistic regression and random forest show very rare long-range errors (three- to four-band separations), and these counts are vanishingly small relative to the diagonal. This behavior is consistent with the limited expressiveness of linear and axis-aligned-tree models when forced to compress the five-dimensional fuzzy mapping into a small number of bands. By contrast, the Deep MLP panel is the most sharply diagonal, with negligible mass beyond one-band separations, which visually reinforces its selection as the student model.

To assess how the selected student behaves on inputs that were not seen during training, Figure 5-13 shows the confusion matrix for the Deep MLP on the 1,000-sample synthetic test set, where each true COF band has exactly 200 test samples.

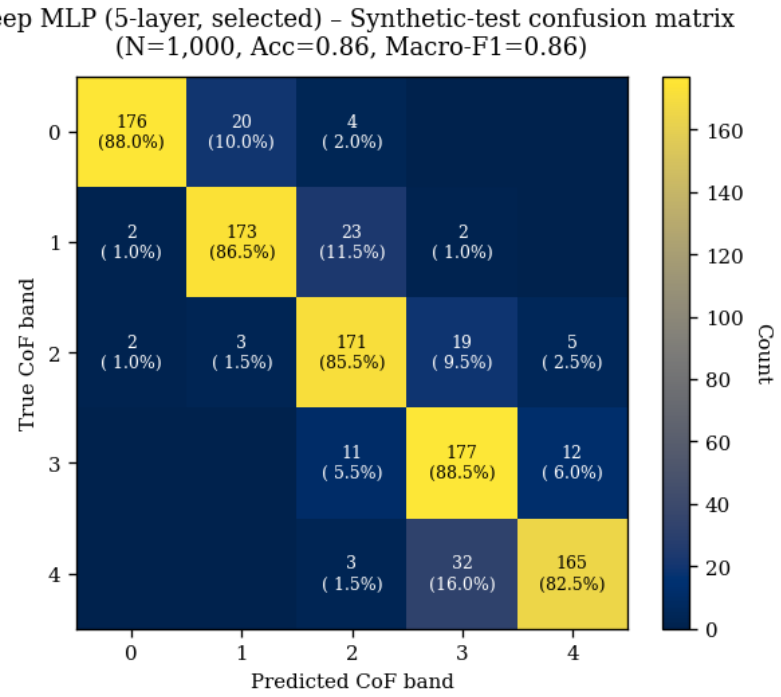

Figure 5-13: Confusion matrix for MLP Deep Testing based on 1000 sample synthetic data

The matrix remains strongly diagonal, with an overall accuracy of 0.86. The remaining misclassifications are dominated by one-band confusions at decision boundaries, and macro-precision (0.87), macro-recall (0.86), and macro-F1 (0.86) remain high. This pattern indicates that, even when challenged with deliberately boundary-focused synthetic cases that differ from the empirical distribution of the teacher data, the student's errors are mostly near-miss disagreements rather than catastrophic long-range failures.

5.7.3.4 Results summary

Taken together, these experiments show that the fuzzy COF teacher can be reliably distilled into a compact Deep MLP student. On the full 2.7-million-segment training set, the Deep MLP achieves an accuracy of 0.94, macro-precision of 0.92, macro-recall of 0.94, and macro-F1 of 0.93, outperforming all other candidate models by multiple percentage points in macro-F1 while still preserving high scores for rare high-consequence bands. On the synthetic test set, which contains 1,000 deliberately boundary-focused cases (200 per COF band), the same network achieves an accuracy of 0.86, with macro-precision 0.87, macro-recall 0.86, and macro-F1 0.86. This reduction relative to the training metrics is reasonable for a more challenging evaluation and shows that the student remains well calibrated in the difficult regions of the decision surface. In both training and test confusion matrices the mass is strongly concentrated on the main diagonal, and nearly all residual errors are adjacent-band confusions rather than catastrophic long-range failures, a pattern that is consistent with intrinsic uncertainty in borderline cases and with the finite granularity of the 0–4 COF banding. These results mirror the LOF verification and demonstrate that a relatively small, well-regularized Deep MLP can learn the teacher's mapping to high fidelity and retain that fidelity on new inputs. This justifies using the

Deep MLP as the operational COF student in later chapters, where it will be evaluated against external inspection data and embedded within portfolio-level renewal optimization because the student reproduces the fuzzy teacher so closely at the band level, any discrepancies in those downstream experiments can be interpreted as differences between the fuzzy theory of consequence and real-world evidence, rather than as weaknesses in the learning apparatus itself.

5.7.4 Validation of the Student COF Models

This section tests whether the student COF models behave sensibly outside their development environment. Up to this point the chapter has focused on internal consistency checks for the fuzzy "teacher" and the MLP "student" for example, monotonic responses to larger outages and more disruptive closures, and stability under noisy inputs. Validation here means asking whether the student's banded COF outputs are compatible with existing utility practice, with expert judgement, and with documented main-break consequences.

We use three complementary lenses. First, we compare the student's COF index to incumbent utility consequence or criticality scores in three systems, treating those indices as structured comparators rather than ground truth. This reveals where the

proposed formulation aligns with current practice and where scale or asset-mix differences create systematic offsets. Second, we elicit agreement or disagreement from asset managers and field crews on anonymized consequence scenarios to check whether the COF bands are intelligible and acceptable for planning use. Third, we carry out a ground-truth experiment for a New England utility, comparing student COF bands to consequences reconstructed from detailed main-break reports. Together, these three steps provide a horizontal view (agreement with utility models and expert opinion) and a vertical view (alignment with realized impacts) of how well the student COF models capture real-world consequence severity.

5.7.4.1 Comparison with Utility Models

Before turning to expert elicitation and main-break ground truth, we first compare the student COF index with each participating utility's incumbent consequence index. The aim is diagnostic rather than evaluative that is, we want to see whether the proposed COF formulation is broadly compatible with how utilities currently score consequence, and to identify systematic pockets of disagreement that deserve closer scrutiny in the later validation steps.

For three utilities (A–C), we harmonize pipe IDs and basic metadata, align score directions so that all indices lie on a common 0–5 "higher = worse consequence" scale, and then compare the student COF index (COF_ML) to each utility's COF index ($COF_{Utility}$). As in the LOF chapter, we summarize agreement with two simple quantities that are easy to interpret on an ordinal (ranked) scale:

- Rank agreement (Spearman ρ) – measures whether the two indices tend to order pipes similarly ($\rho = 1$ would be perfect rank agreement; $\rho \approx 0$ indicates no monotone relation).
- Typical difference (median $|\Delta|$) – the median absolute gap $|\Delta|$ on the 0–5 scale, where $\Delta = COF_{ML} - COF_{Utility}$. We also report the large-gap rate, the share of segments with $|\Delta|>2$, which flags where the two indices disagree by more than roughly two bands.

Figure 5-14 shows scatterplots of COF_{ML} versus $COF_{Utility}$ for Utilities A–C, with points colored by diameter class (<8 in, 8–24 in, >24 in) and a 1:1 reference line. Table 5-15 summarizes the numeric agreement metrics.

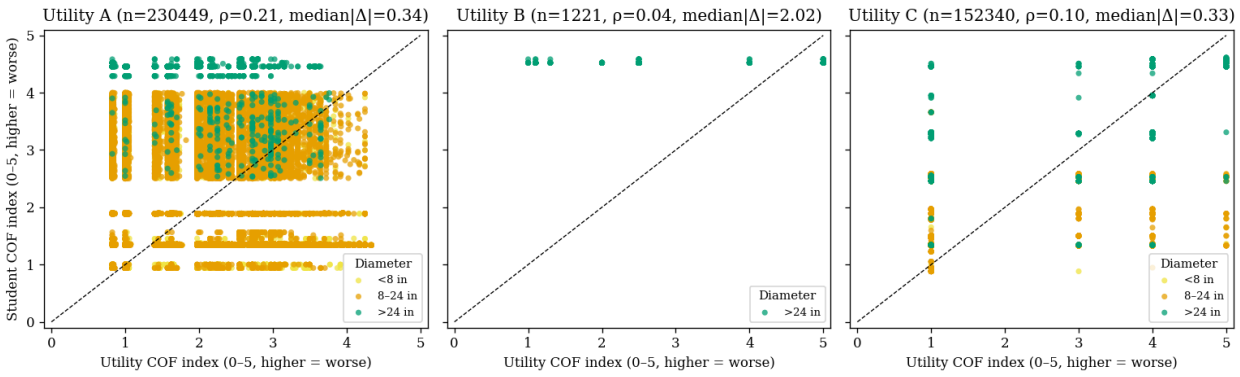

Figure 5-14: COF scatterplots by utility. Utilities A and C are tightly clustered around the diagonal with small gaps, while Utility B shows a horizontal spread of utility scores at almost constant model score, consistent with a wholesale, high-diameter portfolio.

Across Utilities A and C, the student COF index is largely compatible with incumbent practice. Utility A (n ≈ 230,000) shows a modest but clearly positive rank agreement ($\rho \approx 0.21$) with small calibration gaps (median $|\Delta| \approx 0.34$) and a large-gap rate of only ≈3.4%. Utility C (n ≈ 152,000) shows a similar pattern: $\rho \approx 0.10$, median $|\Delta| \approx 0.33$, and a large-gap rate below 3%. In both panels, the point clouds are centered around the 1:1 line and nearly all bands fall within ±2 units. These results indicate that, once scores are mapped to a common 0–5 scale, the student model and the utilities' COF formulations are using broadly comparable notions of consequence, with only modest differences in how individual bands are calibrated.

Table 5-15: Utility–model COF agreement. (Indices on 0–5; higher = worse. Δ = COF_ML − $COF_{Utility}$.)

Utility	Role / asset mix	Samples *n*	Spearman ρ (rank agreement)	Median |Δ|	Large-gap rate (|Δ| > 2)
A	Retail utility; mixed diameters and materials	230,449	0.21	0.34	3.4%
B	Wholesale provider; PCCP transmission mains >72″	1,221	0.04	2.02	69.3%
C	Large retail/regional utility; mixed diameters and materials	152,340	0.10	0.33	2.7%

Utility B behaves very differently, and the pattern is informative rather than alarming. Here the agreement statistics show almost no rank correlation ($\rho \approx 0.04$, $p \approx 0.14$), a large typical gap (median $|\Delta| \approx 2.0$), and an extremely high large-gap rate (≈69%). The scatterplot clarifies why. Utility B is a wholesale provider whose dataset for this comparison consists almost entirely of very large-diameter PCCP transmission mains (all >72″). Their incumbent COF index must therefore spread scores across 0–5 within a very narrow, uniformly high-consequence asset class, in order to prioritize among trunk mains that are all critical by design. By contrast, the proposed COF model is calibrated at the system scale, across the full mix of diameters and service roles. In that system-wide framing, long-diameter transmission PCCP segments sit near the top of the consequence spectrum almost by definition that is, failures can drain large volumes, disrupt downstream deliveries, and are expensive to repair. The fuzzy COF teacher and the student

MLP thus tend to assign these pipes to bands 4–5 with little variation, effectively saying "these are all catastrophic or near-catastrophic," while the utility's index continues to differentiate them internally. The result is a spread of utility scores across the horizontal axis and a tight cluster of model scores near the upper end of the vertical axis, producing large $|\Delta|$ and low ρ even though both parties agree that this cohort is uniformly high consequence.

We treat this divergence as a scale-and-scope issue, not a model failure. Utility B's incumbent index is answering a more local question: "among already-critical transmission mains, which ones should we worry about most?" The student COF learner answers a broader question: "relative to the whole system, how bad are the consequences if this segment fails?" For the wholesale-only, $>72''$ PCCP portfolio, the answer is simply "almost always very bad," and the model reflects that. In practice, this suggests a two-stage calibration. The first stage is using the system-level COF to flag that all trunk PCCP are high consequence, then allowing utility-specific refinements (e.g., redundancy, contractual penalties, pump-station dependencies) to further spread scores within that already high band. The comparison with Utility B therefore highlights where the COF formulation

may need cohort-specific re-scaling or additional wholesale-only dimensions, rather than undermining the overall consequence logic.

As in the LOF chapter, we emphasize that these utility indices are comparators of convenience, not ground truth. The role of this step is to (i) demonstrate that, for two large mixed-asset utilities (A and C), the student COF index sits comfortably within the range of existing practice, and (ii) identify where a structurally different setting (a wholesale PCCP portfolio) requires a different calibration lens. The more stringent tests like expert concordance on representative scenarios and ground-truth comparison against main-break reports then focus on whether the COF formulation itself remains defensible when confronted with operational reality.

5.7.4.2 Agreement with Expert Opinion

To check whether the COF formulation and its banded outputs are intelligible and acceptable to practitioners, we repeated the scenario-based elicitation that we used for LOF, but now focused on COF. The same ten canonical COF scenarios (high-cost urban rupture, spill in a sensitive wetland, prolonged outage at a critical facility, heavy-traffic corridor, peak-demand break with low redundancy, outage on a redundant trunk, traffic-only impacts, and three cost-of-renewal cases by diameter) were sent to three participating

utilities for feedback. For each scenario, the form showed the narrative description, the model's continuous COF index and associated 0–4 band (Insignificant, Minor, Moderate, Major, Catastrophic), and a simple "Agree / Disagree" option with space for comments. The filled feedback form from the utilities is in Appendix F.

We treat these responses as an external face-validity check rather than as a training signal. Blank verification cells are treated as missing data and excluded from rate calculations (one scenario for Utility F and one for Utility D). For each utility, we compute (i) strict agreement, where "Agree" means the expert accepts both the band and its narrative as an appropriate characterization of consequence, and (ii) a tolerant ± 1-band view where possible, using the free-text comments to infer when an expert would simply move the scenario up or down by one band rather than fundamentally disagreeing with its ordering. Disagreement comments are coded into the same motif categories used in the LOF chapter based on Inventory / metadata, Definition or context ambiguity, Policy / heuristic bias, and Category-coverage gap so that patterns in the feedback can inform model revisions rather than being treated as noise.

Table 5-16 summarizes the responses. Utility F (asset managers/planners) reviewed ten scenarios, nine of which had a completed verification field. All nine are strict agrees,

including the three cost-of-renewal scenarios at different diameter tiers, and the only remark is an administrative note about a likely ID mix-up between two high-diameter cases. Utility D (asset managers/planners) provided nine usable ratings, with 5/9 (55.6 %) strict agreement. Of the four disagreements, three push the scenario two bands higher than the model (for example, rating a "Moderate" heavy-traffic case and a "Major" redundant-trunk outage as "COF 5 – High consequence"), and one moves a cost-of-renewal scenario down by one band (model "Major" vs. their "COF 2 – Low").

Table 5-16: Expert concordance on COF scenarios by utility

Utility	Primary expertise	Reviews (n)*	Strict agree	±1-band agreement†	Primary disagreement motifs
F	Asset managers / planners	9	100 % (9/9)	– (no disagreements)	None substantive; one note about a likely mix-up between two high-diameter IDs.
D	Asset managers / planners	9	55.6 % (5/9)	66.7 % (6/9)	Policy / heuristic bias (pushing several scenarios to "COF 5 – High consequence" regardless of redundancy); different cost thresholds for "high" vs "very high" renewal; interpretation of redundant-trunk and small-diameter renewal scenarios.
A	Field operations / maintenance leads	10	80 % (8/10)	≈100 % (10/10, both disagreements are 1-band lower)	Local context corrections: wetland not recognized as high-sensitivity in their system; short 16″ segment that can be isolated, reducing effective outage and traffic impacts.

* Reviews with blank "Agree/Disagree" cells (one scenario each for Utilities A and B) are excluded from the denominator.
† ±1-band agreement counts cases where experts' comments indicate they would shift the COF band up or down by only one level; larger two-band shifts remain coded as disagreement.

Interpreting that single one-band difference as tolerant agreement yields 6/9 (66.7 %) ±1-band concordance. Utility A (field operations / maintenance leads) rated ten scenarios with 8/10 (80 %) strict agreement. In both disagreements, crews argue for one-

band lower consequence than the model. For example, a "Major" spill in an environmentally sensitive area is judged too conservative because the particular wetland in their system is not recognized as high-sensitivity by field operations, and a "Major" cost-of-renewal case for a 16–36″ pipe is judged lower because, in their network, that specific segment can be isolated quickly with minimal traffic control or customer outage. If those two cases are treated as ±1-band "near misses," tolerant agreement for Utility A is 10/10.

Across all three utilities, this yields 22/28 (≈79 %) strict agreement and roughly 90 % ±1-band concordance. More importantly, the direction of disagreement is structured rather than random. Utility F essentially endorses the entire COF scale as written. Utility D tends to push several scenarios up to their highest internal COF class, reflecting a deliberately conservative risk appetite (for example, treating any heavy-traffic or large-diameter outage as "High consequence" regardless of redundancy) and slightly different cost thresholds for when a renewal is considered "very high." Utility A mostly agrees with the model but uses local knowledge to moderate two scenarios downward when the generic description overstates sensitivity (a wetland not flagged as such in their system) or understates isolation ability (a short 16″ segment with valves that sharply limit outage extent).

Taken together, these patterns support two conclusions. First, the COF bands are broadly interpretable and acceptable to both planning staff and field crews across very different systems and experts rarely reverse the ordering of scenarios or move a case by more than one band. Second, the disagreements are informative where they highlight specific places where local policy (e.g., "all large trunk outages are COF 5"), environmental layers (how sensitive wetlands are flagged), or network-isolation details differ from the generic assumptions built into the teacher model. That makes the expert-agreement exercise less of a pass/fail test and more of a structured dialogue. In most scenarios the model and experts align, and in the few that do not, the comments point directly to where local practice or data layers would need to be adjusted for a particular utility.

5.7.4.3 Ground Truth Agreement with validated main break reports

The goal of this subsection is to test whether the COF student learner model, a Deep MLP trained to mimic the fuzzy COF teacher, assigns consequence bands that are consistent with how severe failures are when a main breaks. In practice, this means comparing COF bands produced by the student, using the same asset and context information that would be available for planning, against ex post consequences reconstructed from detailed main-break reports. If the model is well calibrated, events that the student

classifies as “Major” or “Catastrophic” should, on average, be the ones that actually caused larger outages, more disruptive road closures, higher emergency costs, or greater impacts on critical customers, compared with events classified as “Insignificant” or “Minor.” As shown later in this subsection, the student agrees exactly with these output bands derived from main breaks in about four out of five cases, and almost all residual discrepancies are one-band “borderline” slips, with a single two-band outlier that we inspect qualitatively.

This ground-truth validation subsection sits at the end of a three-step validation arc. The first COF validation subsection compares the student COF outputs against the utility’s own COF or analogous consequence model to check whether the proposed formulation is at least competitive with, and often more structured than, current practice. The second subsection examines agreement with expert opinion, asking whether experienced engineers from asset management and field crews see the banded COF outputs as reasonable and decision-useful across typical planning scenarios. Both of those experiments work “horizontally,” comparing the COF bands to alternative human or institutional views of consequence. The present subsection adds a “vertical” test where we take the same COF model that underpins the student Deep MLP, keep its parameters fixed, apply it to a

curated set of high-reliability main-break cases, and then check how well the resulting COF bands line up with what happened on the ground during those documented events.

A critical feature of the setup is that the student MLP has never seen any information from these main break reports during training, model selection, or tuning. For the ground-truth tests here, we follow the same input-dataset development process as for the fuzzy teacher where for each pipe segment we construct the COF input vector from asset and context data and feed this vector into the trained student. None of the observed impacts in the reports related to outage counts, traffic disruption, surface damage, or repair details are used in this process. This strict separation ensures that the comparison between modelled COF bands and realized consequences is an independent external check of the COF formulation, not a rephrasing of patterns the model has already been allowed to learn.

5.7.4.3.1 Data Collection

For this ground truth validation, we use detailed main-break reports validated by the Director of Operations from a large water utility from the New England area. Each report describes a single event and typically includes the address or intersection, the date and time of the failure and of service restoration, operational notes on valves and

hydrants, and narrative descriptions of customer impacts, traffic disruption, flooding, property damage, and repair work. The dataset assembled here spans 2018–2019 and contains 58 events that cover a wide range of break mechanisms, roadway contexts, and traffic conditions. To make these documents usable as ground truth for COF, we manually reviewed each PDF and extracted the relevant information into two harmonized tables.

The first table is a consequence table with one row per event. It contains the main-break identifier (MBR_ID), the approximate number of domestic services affected, counts or indicators for fire-service impacts (fire lines or hydrants out of service), the number of hydrants affected, a short standardized summary of road closure or traffic impacts, a short summary of flooding and surface damage (for example, whether the roadway was undermined or basements were flooded), and a field noting any sensitive customers explicitly mentioned in the report, such as hospitals, schools, or other critical facilities. The second table is a repair-and-operations table that records, again by MBR_ID, the number of valves operated, the number of hydrants shut or otherwise impacted, a concise description of the repair method (for example, use of a full-circle repair clamp versus cut-and-replace of a main segment), a summary of the stock used (pipe lengths, couplings, clamps, and other hardware), and a brief note on final surface restoration and any follow-up paving.

These fields are chosen to map cleanly into the COF structure introduced earlier in the chapter. Customer outages contribute primarily to the social and economic COF dimensions. Fire-service and hydrant impacts relate to safety and emergency-response COF. Road closures and traffic management capture operational disruption. Flooding and pavement damage relate to environmental and third-party damage COF. The complexity of the excavation and restoration provides a qualitative view on renewal-complexity COF. By keeping the entries as short, standardized phrases rather than forcing everything into numbers immediately, we retain much of the narrative nuance while still being able to map each event onto simple ordinal scales.

5.7.4.3.2 Data uncertainties

The main uncertainties in this experiment arise from: (i) incomplete or approximate counts (for example "~30 services", "0–1 hydrants", or missing entries); (ii) possible reporting bias, because busy field crews may under-document minor flooding or record "no damages" when only obvious structural damage is absent; (iii) heterogeneity in free-text wording ("lane closures", "traffic control", "no traffic impacts") and how consistently these phrases map to categories; and (iv) uneven observability across COF dimensions, with social and operational impacts well captured but environmental and renewal-

complexity effects only partially visible. We mitigate these issues by using standardized categorical rubrics, conservative coding rules (erring toward the lower band when ambiguity is high), and explicit documentation of the mapping from text to ordinal scores.

5.7.4.3.3 Data processing

For the actual validation, we treat each main break report as one realized consequence scenario and convert the two tables into a set of dimension-level ground-truth scores. For every event we compute four ordinal indices on a 0–4 scale:

- $C_{\text{cust,GT}}$: customer and fire-service impacts
- $C_{\text{traffic,GT}}$: traffic and access disruption
- $C_{\text{damage,GT}}$: physical damage and restoration burden
- $C_{\text{priority,GT}}$: sensitive or priority customers

Each index is derived from the MBR fields using a simple rubric that mirrors the structure of the fuzzy COF teacher but is defined entirely from the reports and fixed in advance, before any comparison to the student model. In brief, customer impact $C_{\text{cust,GT}}$ uses the approximate number of domestic services affected, the presence of fire-service outages, the number of hydrants affected, and any explicit mention of sensitive customers. Values near zero correspond to "no services out, at most one hydrant affected, and no

sensitive customers"; intermediate bands capture small (1–5), moderate (≈6–25), and larger (≈25–35) residential outages or combinations of domestic and fire-service loss; and the highest band is reserved for events that behave like a small zone outage or that involve both residential customers and fire-protection loss.

Traffic and access impact $C_{\text{traffic,GT}}$ is based on the standardized road-closure summary and an implicit classification of the street hierarchy. Events with no closure or a short work zone on a low-traffic residential street fall in the lowest band. Lane restrictions on local or collector streets map to "minor," extended lane restrictions or overnight closures on busier corridors map to "moderate," and full closures or complex traffic control on major arterials and key intersections (for example Dorchester Ave & West Broadway) are treated as "major" or "catastrophic" depending on how wide the detours are.

Physical damage and restoration burden $C_{\text{damage,GT}}$ combines the flooding/surface-damage notes with information about the repair. Small leaks with localized street wetting, a single clamp, and a short cut that is simply backfilled and patched are scored as low consequence. Larger trench dimensions, explicit mentions of pavement undermining or "deep trench," replacement of longer segments of main, and more complex backfill and paving operations are mapped to higher bands. Events where water enters basements,

large sections of main are blown out, or there is "substantial excavation and roadway disturbance" are treated as major damage even if customer counts are modest.

Finally, the priority-customer index $C_{\text{priority,GT}}$ flags whether the reports mention any sensitive uses. Cases with no explicit mention of sensitive customers are scored zero. Mentions of public open space receive a low but non-zero score. Events where the fire department, mayor's office, or other city-wide emergency services are formally notified, or where vulnerable populations would plausibly be affected, are mapped to higher bands. This dimension is deliberately conservative: it is only elevated when the narrative clearly points to equity or critical-infrastructure concerns. Table 5-17 summarizes these rubrics in compact form.

Table 5-17: Ground truth to COF conversion rubric

Dimension	**MBR fields used**	**Band 0–1 (lower consequence, examples)**	**Band 2–3 (higher consequence, examples)**
$C_{\text{cust,GT}}$ $-$ *customers & fire*	Domestic_Services_Affected; Fire_Services_Affected; Hydrants_Affected; Sensitive_Customers_Notes	0 customers, at most one hydrant affected, no fire service or sensitive customers; or 1–5 domestic only	≈6–25 domestic customers out; ≥1 fire-service impact; multiple hydrants; larger outages (≈25–35) or language such as "neighborhood" affected
$C_{\text{traffic,GT}}$ $-$ *traffic/access*	Road_Closure / Traffic_Impacts + street type	No closure or minor work zone on local street; short lane restriction	Prolonged lane restrictions or overnight closure on collector/arterial; full closure or detours at a major intersection or key corridor
$C_{\text{damage,GT}}$ $-$ *damage & restoration*	Flooding_Surface_Damage_Notes; Valves_Operated; Hydrants_Shut / Impacted; Repair_Method /	Localized wetting only; small trench; clamp-only repair; routine backfill	Pavement undermining, deep trench, long DI inserts, blown-out main sections, water entering buildings, "substantial excavation

Dimension	MBR fields used	Band 0–1 (lower consequence, examples)	Band 2–3 (higher consequence, examples)
	Approach; Stock_Used; Final_Surface_Restoration	and patch; "no roadway/property damage"	and roadway disturbance," complex restoration
$C_{\text{priority,GT}}$ $-$ *sensitive customers*	Sensitive_Customers_Notes; mentions of BFD, Mayor's Office, parks, etc.	No sensitive customers noted; or only generic residential area	Public open space or civic facilities flagged; explicit notification of fire department or mayor's office; any clearly vulnerable or critical facility affected

Once these four indices are computed, we normalize them to a single overall ground-truth band, COF_{GT}, on the same 0–4 scale used by the model. The primary rule is a simple maximum over dimensions:

$$\text{COF}_{\text{GT}} = \max\left(C_{\text{cust,GT}}, C_{\text{traffic,GT}}, C_{\text{damage,GT}}, C_{\text{priority,GT}}\right).$$

This "worst-dimension wins" rule is consistent with how operators perceive consequence in practice. For example, a break that closes a major intersection or floods a basement is treated as high consequence even if the customer count is modest. It also keeps the mapping transparent. For any event coded as, say, band 3, one can point directly to the underlying dimension (for example "major traffic closure" or "basement flooding with blown-out main") that triggered that label in the main-break record. The aggregation rules, including the dimension rubrics and the max operator, are all fixed in

advance and held constant during analysis, so $\mathrm{COF}_{\mathrm{GT}}$ remains independent of the model predictions and can act as an unbiased ground-truth target for validation.

On the model side, for each event location we compute a COF band by running the trained Deep MLP student on the input vector derived from asset and context data (pipe diameter and material, land use, proximity to hospitals or schools, traffic importance of the roadway, redundancy in the network, and similar variables). These inputs are constructed exactly as they would be during planning, without using any information from the main break reports themselves. The output of the student network, $\mathrm{COF}_{\mathrm{student}}$, is therefore the model's independent prediction for that asset under normal planning assumptions. Comparing $\mathrm{COF}_{\mathrm{student}}$to $\mathrm{COF}_{\mathrm{GT}}$then provides a clean test of whether the student, trained solely to imitate the fuzzy COF mapping, is nevertheless aligned with the actual severity of observed failures.

5.7.4.3.4 Hypothesis testing

Once the model and ground-truth bands are defined on a common 0–4 ordinal scale, we assess agreement between $\mathrm{COF}_{\mathrm{student}}$and $\mathrm{COF}_{\mathrm{GT}}$ using the same family of statistics as in the LOF validation, adapted to consequence. The goal is to test whether the student behaves as an informative and ordinally calibrated predictor of realized

consequences: do higher model bands correspond to more severe events in the reports, and does the model avoid assigning very low COF to breaks that were clearly disruptive? Table 5-18 summarizes the hypothesis framework.

Table 5-18: Hypothesis testing framework for COF student–ground truth agreement. Each row defines a null hypothesis, the statistic used, the decision rule at $\alpha = 0.05$, and how to interpret rejection in terms of the model's ability to recover the severity ordering of observed main-break consequences.

ID	Test	Null hypothesis H_0	Statistic / estimate	Decision rule $\alpha = 0.05$	Interpretation if H_0 is rejected
H_0–χ^2	Are predicted COF bands independent of ground truth?	COF_pred_band is independent of main-break ground truth COF_GT_band.	Pearson χ^2 test on the 5×5 confusion matrix (or reduced to non-empty rows/columns).	Reject if $p(\chi^2) < 0.05$.	Student COF predictions carry real information about observed consequences, rather than being random with respect to ground truth.
H_0–κ	Is agreement beyond chance, on an ordinal scale?	Quadratic-weighted Cohen's $\kappa \leq 0.20$ (no more than "slight" agreement).	Cohen's κ (unweighted) and quadratic-weighted κ_w computed on the 0–4 COF bands.	Reject if κ_w > 0.20 (optionally: and the 95 % CI for κ_w stays above 0.20).	The student model shows at least fair–substantial ordinal agreement beyond chance; large band gaps are rare and penalized heavily.
H_0–ρ	Is there monotone rank association?	Spearman rank correlation $\rho = 0$ between COF_pred_band and COF_GT_band.	Spearman ρ with two-sided p-value.	Reject if $p(\rho) < 0.05$.	Higher predicted COF bands systematically correspond to higher ground-truth COF bands; the model preserves severity ordering.
H_0–base	Does the model beat a naive majority-class COF strategy?	Exact accuracy ≤ majority-class prevalence (always predicting the most common band).	Exact accuracy vs majority-class baseline; Wilson 95 % CI for accuracy.	Reject if accuracy > baseline and the 95 % CI for accuracy lies entirely above baseline.	The student model meaningfully outperforms a naive "always pick the dominant COF band" strategy on real main-break consequences.

To evaluate these hypotheses, we first construct a 5×5 confusion matrix whose rows correspond to $\mathrm{COF}_{\mathrm{GT}}$ and columns to $\mathrm{COF}_{\mathrm{student}}$. Each cell contains the number of events with a given pair of labels. We also use a row-normalized version where each entry is the fraction of events in each true band that were assigned to each predicted band. From this matrix we compute:

- Exact accuracy: Fraction of events where $\mathrm{COF}_{\mathrm{student}} = \mathrm{COF}_{\mathrm{GT}}$(0.793 with a Wilson 95 % CI [0.672, 0.877]).
- Within-one-band accuracy: Fraction where the absolute difference $|\,\mathrm{COF}_{\mathrm{student}} - \mathrm{COF}_{\mathrm{GT}}\,|$is ≤ 1 (0.983; 57 of 58 events).
- Macro-precision, macro-recall, and macro-F1: standard precision/recall/F1 computed for each band separately and then averaged so that rare high-consequence events count as much as common low-consequence events (macro-F1 ≈ 0.75).

Because COF bands are ordered, we emphasize ordinal measures rather than raw accuracy alone. Quadratic-weighted Cohen's κ, κ_w, measures agreement beyond chance while penalizing long-range errors more strongly than one-band slips. Here $\kappa \approx 0.72$ (unweighted) and $\kappa_w \approx 0.90$, which corresponds to substantial agreement on an ordinal scale. Spearman's rank correlation coefficient ρ between $\mathrm{COF}_{\mathrm{student}}$ and $\mathrm{COF}_{\mathrm{GT}}$ quantifies

whether higher model bands correspond to higher observed severity even when the labels are not exactly equal. In this experiment $\rho \approx 0.89$ with $p \approx 4.7 \times 10^{-21}$, decisively rejecting H_0–ρ.

The χ^2 test on the confusion matrix ($\chi^2 \approx 124$, df = 16, $p \approx 8.3 \times 10^{-19}$) rejects H_0–χ^2, confirming that the joint distribution of predicted and ground-truth bands is far from independent. The majority-class baseline for this sample is ≈0.40 (always predicting band 1). The student's observed accuracy of ≈0.79, with its confidence interval entirely above 0.40, rejects H0–base and shows that the model is extracting real signal from the asset and context features rather than echoing the dominant class.

Operationally, the confusion matrices are strongly diagonal with almost all off-diagonal mass in adjacent bands, which is the pattern utilities care about. The within-one-band accuracy of 0.983 indicates that 57 of 58 events are either correctly classified or off by a single band. There is exactly one two-band discrepancy, which we treat as an informative outlier and will be discussed in the results section.

Because the Deep MLP student has already been verified, on larger synthetic and screening datasets, to closely match the fuzzy COF teacher, the systematic patterns observed here can be interpreted primarily as properties of the underlying COF formulation

and feature set rather than of the learning algorithm. In particular, the strong rejection of H_0–χ^2, H_0–κ, H_0–ρ, and H_0–base suggests that the current COF design is both informative and rank-consistent with real main-break impacts, while the handful of near-boundary slips and the single two-band outlier identify specific scenarios where additional features or revised weights could further sharpen the model.

5.7.4.3.5 Results

Overall Agreement: Across the 58 main-break events, the COF student model agrees with the main-break–derived ground truth band (COF_{GT}_band) for 46 events, giving an accuracy of 0.79 with a Wilson 95 % confidence interval [0.67, 0.88]. The majority-class baseline that is, always predicting the most common band (band 1 in this sample), would achieve only 0.40 accuracy, so the student nearly doubles the effectiveness of a naive strategy. Within-one-band agreement is 0.98 (57 of 58 events), meaning almost all disagreements are "borderline" slips between adjacent bands. Only a single event is misclassified by two bands (ground truth 4 vs model 2). This outlier is discussed below.

Class-wise performance is also well balanced. Macro-precision is 0.73, macro-recall is 0.84, and macro-F1 is 0.75, with a weighted F1 of 0.80. These macro statistics average

performance across all five consequence bands, so rare high-consequence events influence the evaluation as much as the many low-consequence breaks.

The confusion matrices in Figure 5-15 show a strong diagonal, with most mass either on the diagonal or in immediately adjacent cells.

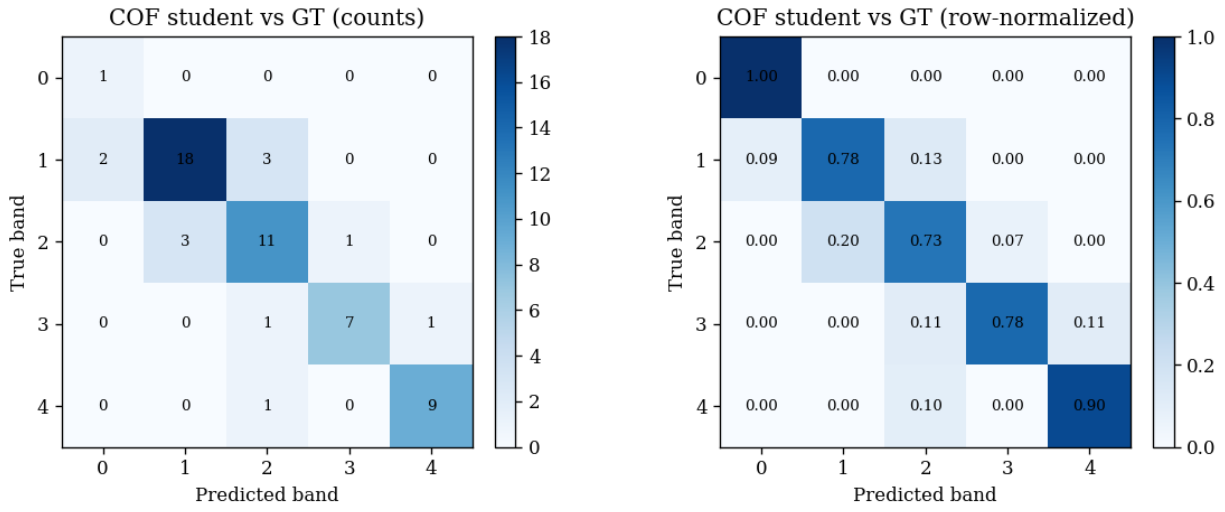

Figure 5-15: Confusion matrices comparing COF student predictions to main-break ground truth bands (n = 58). Left: raw counts; right: row-normalized fractions. The strong diagonal and concentration of off-diagonal mass in adjacent bands illustrate high exact accuracy (0.79), very high within-one-band agreement (0.98), and rare long-range errors.

No event with ground-truth band ≥3 ("Major" or "Catastrophic") is ever predicted in the lowest two bands (0 or 1). Among the 19 high-consequence events (bands 3–4), 17 are predicted in bands 3–4 and 2 are predicted as band 2. There are no catastrophic underestimates such as 4→0 or 4→1. The error-distance distribution |Δband| is sharply

concentrated at 0 and 1 (46 and 11 events, respectively), with a single case at $|\Delta\text{band}| = 2$ and a maximum absolute error of 2.

Agreement statistics reinforce this visual impression and are summarized with the decisions on the null hypotheses in Table 5-19. Each row in the table states the null hypothesis, the statistic used, numerical evidence from the 58-event sample, and the interpretation when the null is rejected.

Table 5-19: Hypothesis tests for COF student vs main-break ground truth (COF_{GT}_band).

ID	Null hypothesis (H_0)	Estimate	Evidence	Decision
H_0–χ^2	Predicted COF bands are independent of main-break ground truth.	Pearson χ^2 on 5×5 confusion matrix.	$\chi^2 = 124.3$, df = 16, p ≈ 8.3×10^{-19}.	**Reject H_0.** Student COF predictions carry strong, non-random information about realized consequences.
H_0–κ_w	Ordinal agreement beyond chance is no more than "slight" ($\kappa_w \leq 0.20$).	Cohen's κ (unweighted) and quadratic κ_w.	κ = 0.72, κ_w= 0.90 (well above 0.20).	**Reject H_0.** The model shows substantial ordinal agreement; large band errors are rare and heavily penalized.
H_0–ρ	No monotone rank association between COF_{pred}_band and COF_{GT}_band.	Spearman rank correlation ρ.	ρ = 0.89, p ≈ 4.7×10^{-21}.	**Reject H_0.** Higher predicted COF bands consistently correspond to higher ground-truth bands; severity ordering is preserved.
H_0–base	The model does not beat a naive majority-class strategy.	Exact accuracy vs majority-class prevalence.	Majority baseline = 0.40; student accuracy = 0.79 with 95 % CI [0.67, 0.88], entirely above baseline.	**Reject H_0.** The student meaningfully outperforms "always predict the dominant band" on real main-break consequences.

The unweighted Cohen's κ is 0.72 and the quadratic-weighted κ_w is 0.90, which is conventionally interpreted as substantial agreement beyond chance. The χ^2 test of

independence on the 5×5 table gives $\chi^2 = 124.3$ with 16 degrees of freedom and $p \approx 8.3\times10^{-19}$, decisively rejecting the hypothesis that predictions are unrelated to ground truth. Finally, Spearman's rank correlation between the student and main-break bands is $\rho = 0.89$ ($p \approx 4.7\times10^{-21}$), showing that larger model bands systematically align with more severe observed consequences even when the labels are not exactly equal.

Overall, these results indicate that the COF student behaves as an informative, ordinally calibrated predictor of real-world consequence severity, rather than simply reproducing the most common band or wandering randomly across the 0–4 scale.

Scenario- and dimension-specific checks: Beyond aggregate statistics, we checked whether the model behaves sensibly for the kinds of events that utility emergency renewal crew care most about. Because both the ground-truth rubric and the COF formulation are dimension-based, these checks help confirm that the right dimensions are driving high-consequence classifications.

High-consequence scenarios: Events coded from the main-break reports as "Major" or "Catastrophic" ($COF_{GT}\_band \geq 3$) are precisely those with combinations of larger customer outages, full or complex closures on important corridors, blown-out main sections, or explicit involvement of emergency and city leadership. As noted above, none

of these 19 cases is placed in the lowest two model bands. Seventeen are assigned to bands 3–4 and the remaining two are assigned to band 2. In other words, when operational staff perceived an event as major, the student almost always agreed or erred conservatively by one band and never dismissed a high-impact break as inconsequential.

Customer and fire-service impacts: When the events are binned by approximate number of domestic services affected (for example, 0, 1–5, 6–25, >25), the median model band increases with outage size. Breaks that interrupt tens of customers or involve fire-service loss sit predominantly in bands 2–4 on both the ground-truth and model sides. This aligns with the intended structure of the COF teacher: the model treats "how many and which customers?" as a primary driver of consequence.

Traffic and access disruption: Stratifying events by closure type (no closure / local lane restriction / partial closure on collector / full closure at major intersection) shows a similar pattern. Cases that close major intersections or require complex traffic management are almost always assigned bands 3–4 by the model. Routine local lane closures tend to appear in bands 1–2. This monotone shift in the model's band distribution across closure categories provides a tangible, scenario-level explanation for the high

Spearman ρ. It shows that the student is not just matching labels in aggregate but responding correctly to the direction of the traffic-impact signal.

Priority and critical customers: Events where the reports explicitly mention the fire department, mayor's office, or other civic institutions being notified, or where public open space is affected, end up in elevated ground-truth priority bands. The model also assigns these cases to higher COF bands, reflecting that the priority-customer dimension is being carried through from the teacher into the student and is not being washed out during learning.

These scenario-specific checks can be summarized visually. First, the confusion matrices in Figure 5-15 show that most mass lies on or adjacent to the diagonal. Second, the error-distance histogram in Figure 5-16 confirms that $|\Delta\text{band}|$ is almost always 0 or 1 with a single two-band outlier.

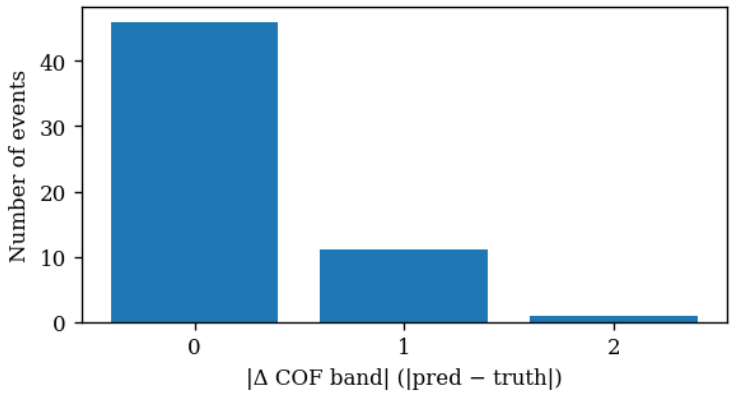

Figure 5-16: Error distance histogram for COF student vs ground truth

Third, the boxplots of model bands by customer-outage category and by closure type (Figure 5-17), show median COF bands rising steadily from "no customers/ no closure" to "many customers/full closure," indicating that more customers, more traffic disruption, or more severe damage all reliably push the predicted band upward.

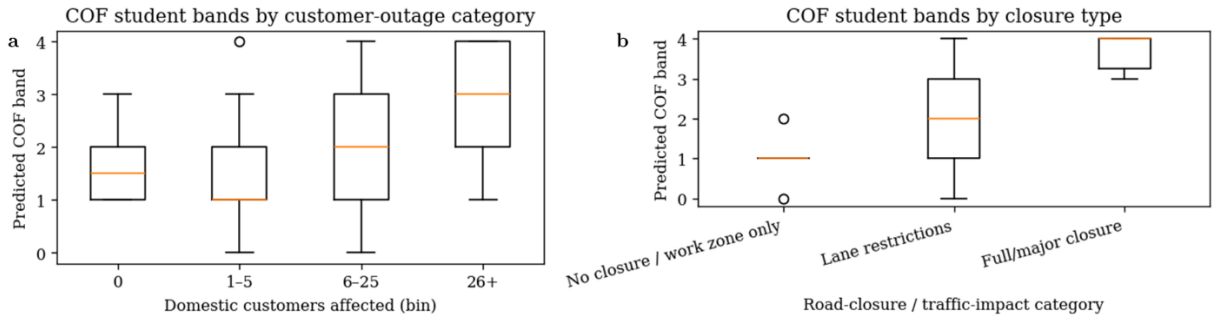

Figure 5-17: a) Boxplot of COF student bands by customer outage category; b) Boxplot of COF student bands by closure type

Outlier misclassification: A single two-band misclassification occurs for a break at a major intersection (a 16-inch CI/DI main). The main-break rubric rates this event as Catastrophic ($\text{COF}_{\text{GT}}\_\text{band} = 4$), while the student model assigns it a Moderate consequence ($\text{COF}_{\text{pred}}\_\text{band} = 2$).

From the reports, the break is described as a blowout on a high-traffic, signalized intersection with multiple legs closed and detours in place for the duration of the repair. Only one domestic service and two fire services are recorded as directly affected, and no basements are reported flooded, but the narrative emphasizes "substantial excavation and roadway disturbance" at a "major intersection," with the city Fire Department and Mayor's Office of Neighborhood Services formally notified. In the ground-truth rubric, several of these signals stack. For example, a blowout on a large-diameter main, full intersection closures on a key arterial, and explicit notification of city-wide emergency/administrative services. Together they push the traffic, damage, and priority-customer dimensions into the highest bands, so the max-over-dimensions rule yields $\text{COF}_{\text{GT}}\_\text{band} = 4$.

By contrast, the student model only "sees" the structured inputs available system-wide like pipe diameter and material, generic road-class (arterial vs local), land-use and

network redundancy, and coarse proximity to critical facilities. In those features, this location looks like a typical urban arterial main with modest recorded customer counts and no hospital immediately adjacent. The model therefore treats it as a high but not extreme case and settles on band 2 ("Moderate"). In other words, what makes this event feel catastrophic to operations is not just the physical break itself, but its political and visibility context where a messy blowout at a symbolic, heavily trafficked gateway with senior city offices engaged.

Rather than treating this single two-band miss as a failure of the neural network, we interpret it as a pointer to where the COF formulation and feature set could be enriched. For example, by (i) explicitly encoding "political/visibility" context for a small set of iconic intersections, (ii) differentiating full intersection closures from single-leg lane closures, and (iii) adding a simple "escalated to city-wide emergency/administrative services" flag. Once such information is available in the input vector, the same student–teacher framework should be able to pull this kind of borderline event into the correct high-consequence band.

5.8 Summary

This chapter develops a COF framework that provides a detailed and measurable definitions of *what* counts as consequence, *where* it is observed, and *how* it is summarized for decision making. COF is defined here as the multi-dimensional impact of a main failure on customers, network operations, the environment, and renewal complexity, conditional on pipe size and context. To keep this tractable for utilities, the chapter reduces these dimensions to a single 0–4 banded index (Insignificant to Catastrophic) that is still traceable back to parameters like customer outages and fire-service loss, traffic and access disruption, physical damage and restoration burden, and the presence of sensitive or priority customers. Each dimension is specified using transparent rubrics based on simple rules that map counts and short standardized phrases to ordinal scores, so that a planner can see why a given segment or scenario is classified as, say, "Major" rather than "Moderate."

This chapter builds a fuzzy-logic "teacher" model that aggregates dimension scores into an overall COF band while preserving important asymmetries (for example, "worst dimension wins" so that a basement flood or major intersection closure is not diluted by otherwise modest impacts). The same structure is then learned by an MLP "student"

model, trained on synthetic and screening datasets assembled from eighteen utilities spanning climates, network types, and diameter mixes. The student model provides the scale and speed needed for system-wide scoring while the fuzzy teacher preserves interpretability. Together they form a teacher–student pair that can be interrogated both as an algorithm and as a codified consequence policy for pipe renewal. Evaluation in this chapter focuses on sanity checks of the COF inputs and rule-base, monotonicity and saturation behavior with respect to key drivers (customers affected, closure type, repair complexity), and stability of the learned student with respect to noisy or partially observed inputs.

Validation probes whether the resulting COF bands behave sensibly outside the development environment. First, we compare the student COF index against incumbent utility criticality or consequence scores in three systems, using rank agreement, typical band differences, and large-gap rates as simple diagnostics. These scatterplots reveal where our formulation aligns with existing practice, where differences are attributable to scale or design choices (for example, a wholesale system dominated by very large PCCP mains that sit at the top of any 0–5 scale), and where indices embed policy overlays that depart from impact mechanism based consequence. Second, we elicit expert opinion from asset managers and field crews using anonymized scenarios. Across utilities, experts

generally endorse high COF bands when scenarios involve large outages, environmental sensitivity, or renewal complexity, and most disagreements can be traced to inventory issues or local policy conventions rather than to gross model misunderstandings. Third, a ground-truth experiment at a large urban New England utility compares a priori COF bands to ex post consequences reconstructed from detailed main-break reports. Here the student achieves high within-one-band agreement, strong ordinal association, and almost no catastrophic underestimates. Also, higher model bands line up systematically with larger outages, more disruptive closures, and more severe damage.

This chapter also presents the limits of these tests. The main-break ground-truth dataset covers a single system and two years, and not all consequence dimensions are observed with equal richness and environmental damage, reputational impact, and long-term legal exposure are only partly visible in operations reports. The utility model comparison depends on incumbent indices that are themselves imperfect and shaped by local policy, while expert-opinion experiments sample a finite set of scenarios. As a result, the evidence presented here demonstrates that the COF framework is broadly aligned with how participating utilities experience and document consequences, but it does not claim universal calibration for all utilities or for rare, extreme events. The value of the chapter

lies in the way the three viewpoints, *horizontal* agreement with incumbent indices, concordance with expert judgment, and *vertical* agreement with observed consequences converge. Together they support the claim that the proposed COF model is scientifically grounded, operationally credible, and suitable for integration with the LOF framework and portfolio optimization in the subsequent risk-based renewal chapters.

Chapter 6

Pipe Renewal Prioritization Model

This chapter moves from *scoring* individual pipelines to *choosing* which ones to renew under real-world constraints. Chapters 4 and 5 developed segment-level estimates of Likelihood of Failure (LOF), Consequence of Failure (COF) for each pipe segment. Here, those segment-level scores are treated as inputs to a multicriteria decision optimization model that selects a subset of segments (or aggregated projects) for inclusion in a Capital Improvement Program (CIP). The focus is not on improving LOF or COF models themselves, but on designing a portfolio-level decision framework that uses those scores in a principled way.

6.1 Goal and Scope

This chapter addresses the central question that given a finite renewal budget and operational constraints, which subset of candidate pipes should be renewed now, and in

what groupings, to maximize risk reduction per unit cost while maintaining service equity and operational feasibility?

Answering this question requires moving from a one-pipe-at-a-time ranking mindset to an explicitly portfolio-based view. In a portfolio view, the value of renewing a pipe depends not only on its individual risk, cost, and equity profile, but also on which other pipes are renewed in the same CIP. The work can be spatially clustered, disruptive work can be staggered, and synergies with other utility or street projects can be exploited. The model developed in this chapter therefore treats renewal planning as a combinatorial optimization problem over a large but finite set of candidate segments or projects.

For consistency with how utilities typically commit funds, the planning horizon in this chapter is defined as one year. The optimization is run for a given fiscal year to recommend a set of renewal projects whose combined cost does not exceed a user-specified annual budget B. This reflects annual budget appropriations, yearly replacement targets (e.g., miles/year), and yearly condition assessment programs that support forensic analysis research and feed updated LOF/COF estimates into the decision process. In practice, the same optimization framework can be rerun each year with refreshed LOF/COF scores, ground-truth inspection data, and updated priorities, producing a sequence of annual

plans that collectively form a multi-year capital improvement trajectory. However, the formal problem treated here is explicitly one-year at a time, which simplifies both the mathematics and the interpretation for utilities who already plan and report on an annual cycle.

The primary decision unit in this chapter is a *renewal* project, not a single pipe segment. A project may consist of one or more contiguous or closely clustered segments that can be renewed together (for example, a city block, a cul-de-sac, or a short trunk-main corridor). This aggregation reflects operational realities where contractors mobilize crews and machinery at the project scale, residents experience disruption at neighborhood scale, and coordination with other utilities (transport, sewer, telecom) is negotiated by corridor, not segment. Mathematically, the decision variable is a binary indicator for each project, but each project retains attributes derived from its constituent segments like aggregate risk, total length, cost, water-loss reduction, equity impact, and expected customer-hours of disruption.

The scope of the optimization includes all projects whose underlying segments meet an eligibility threshold derived from LOF and COF. In the base configuration, a pipe is considered *high-risk* if its total risk score $R_i = \mathrm{LOF}_i \times \mathrm{COF}_i$ exceeds a threshold (e.g.,

$R_i \geq 9$ on a 0–5 × 0–5 scale), and only projects that contain at least one such high-risk segment enter the optimization pool. The risk threshold can be changed based on the number of assets in different risk zones after running the LOF and COF models. This pre-screening is performed using a risk contour chart as illustrated in Figure 6-1.

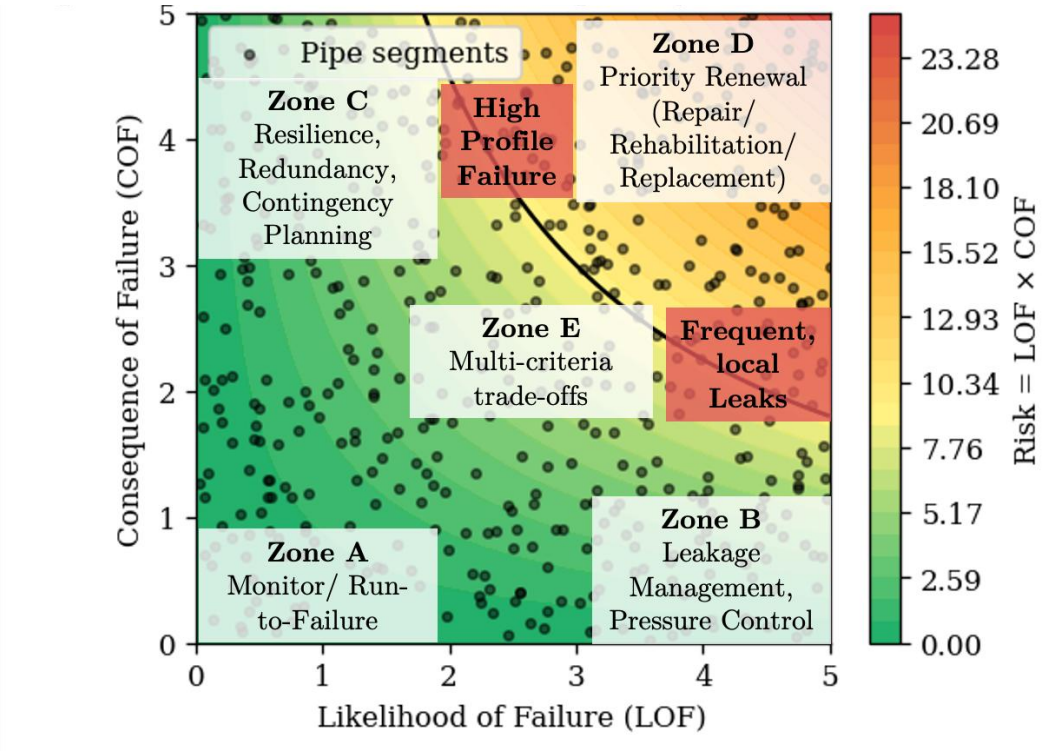

Figure 6-1: Risk contour chart for pre-screening. Scatter of LOF vs COF with iso-risk contours and the chosen eligibility threshold (e.g., $R \geq 9$) shaded to indicate the candidate pool passed to the GA. Annotated zones illustrate that different regions of the LOF–COF space call for different management strategies (monitoring, leak management, contingency planning, or priority renewal), and the two red-zone points with identical risk products highlight why a multicriteria portfolio model is needed to distinguish between frequent low-impact failures and rare high-impact failures.

Figure 6-1 shows the annual LOF–COF space with iso-risk contours and the eligibility threshold used to pre-screen candidate projects. The same scalar risk value can arise from many combinations of failure likelihood and consequence. Low-LOF, high-COF projects (Zone C) call for redundancy and contingency planning, whereas high-LOF, low-COF projects (Zone B) primarily motivate leak management and pressure control. High-LOF, high-COF projects (Zone D) populate the critical renewal zone, while low-LOF, low-COF segments (Zone A) are usually managed through monitoring and reactive repairs. The dense central region (Zone E) contains many moderate-risk projects that compete for limited annual budgets. Here, simple risk ranking is insufficient, and multicriteria optimization is needed to balance risk reduction, equity, water-loss reduction, and delivery constraints.

This step serves two purposes. First, it focuses the algorithm on decisions that matter for risk reduction, rather than expending computational effort on evidently low-risk segments. Second, it mirrors current utility practice where only assets above a certain risk, age, or condition threshold are considered for capital renewal, while others are managed through reactive repairs or monitoring. The chosen threshold is therefore not

arbitrary but anchored in the LOF/COF scales established in earlier chapters and can be tuned in collaboration with utilities to reflect their risk tolerance.

Within this pre-screened candidate set, the model explicitly balances multiple criteria. At minimum, these include: (i) risk reduction, computed from LOF and COF; (ii) renewal cost, including construction and surface restoration; (iii) service equity, represented through socio-economic indicators and service-reliability baselines in the affected neighborhoods; (iv) operational disruption, expressed in customer-hours of service impact; and (v) water-loss and sustainability benefits, such as reduced leakage or energy savings. All of these criteria are expressed at the project level but have consistent definitions with the segment-level constructs in Chapters 4 and 5. This avoids redefining core quantities while ensuring that the portfolio model sees them in units that are meaningful for planning.

The optimization framework is implemented using a Genetic Algorithm (GA), which is well-suited to high-dimensional, non-convex, and combinatorial search spaces. In this setting, each candidate solution, or *chromosome*, encodes a possible CIP that is, a binary string indicating which projects are selected for renewal. The GA iteratively evolves a population of such portfolios through selection, crossover, and mutation, using a fitness

function that combines the decision criteria into a single scalar objective subject to hard constraints such as the CIP budget. The choice of GA, rather than linear or mixed-integer programming alone, is motivated by three factors. First, the search space grows exponentially with the number of candidate projects, making exhaustive search impossible. Second, the criteria and constraints include non-linear, non-smooth components (e.g., penalties for exceeding construction capacity, or discrete synergies when adjacent projects are selected together). Third, GA allows the model to be extended easily to include new criteria and constraints without re-deriving closed-form problem structures. At the same time, the GA is configured conservatively, with transparent hyperparameters and convergence diagnostics, to maintain reproducibility and interpretability.

This chapter is limited in two important ways that delimit its scope. First, it assumes that LOF, COF, water-loss, cost estimates and other criteria are already available and reasonably calibrated, as established in the previous chapters. The portfolio model does not attempt to re-learn these quantities but rather focuses on treating uncertainty. Second, the objective here is to optimize the composition of a single CIP, not to solve the full inter-temporal problem of allocating funds across many decades. The latter problem is better conceptualized as a temporal commons question and is discussed in the broader

implications of the dissertation. Nevertheless, by explicitly tracking risk reduction, equity, water-loss outcomes, and budget utilization, the present model provides the building blocks for more sophisticated temporal allocation schemes. It is important to note that the framework presented in this chapter is general and can accommodate many objectives and constraints. The instantiated model used for a water utility can optimize the critical criteria under an annual budget constraint, while equity, water-loss, and sustainability metrics can be evaluated diagnostically for the resulting portfolios.

The contribution of this chapter is therefore threefold. It (i) formalizes a multicriteria, constraint-aware portfolio optimization model that translates LOF and COF scores into actionable renewal plans; (ii) demonstrates how this model can be tuned to different utility priorities, such as maximizing risk reduction per dollar, improving service equity, or reducing water losses; and (iii) proposes a set of evaluation, verification, and validation (EVV) protocols that compare optimized portfolios against both simple baselines (e.g., risk-only ranking) and ground-truth data from detailed pipe inspections. By situating renewal decisions within a transparent, mathematically explicit framework, the chapter aims to reduce ad-hoc prioritization and support utilities in making more defensible and equitable CIP choices.

6.2 Renewal Decision Context and Design Principles

This section positions the renewal optimization model within the utility's broader planning hierarchy and clarifies the decision layers it is meant to support. It then lays out the core design principles that are, risk-based, multi-criteria, constraint-aware, and explainable that govern how the model translates LOF and COF outputs into actionable annual portfolios.

6.2.1 Planning hierarchy and decision layers

Renewal decisions do not occur in a vacuum. Water utilities operate across several planning layers that differ in time scale, information requirements, and decision levels:

- Operational layer (hours–weeks): This is where dispatch decisions related to leak repairs, emergency shutdowns, valve operations, short detours, and temporary service restorations are made. Actions are driven by incidents (e.g., breaks, complaints) rather than by a global view of system risk. The objective is to restore service and safety as quickly as possible, typically using work-order systems and field crew judgement.
- Tactical layer (annual programs): At this layer, utilities define one-year programs such as the annual main replacement program, annual condition assessment campaigns,

hydrant or valve renewal programs, and targeted leak reduction campaigns. Decisions are made at the level of projects (aggregates of segments that can be delivered as a work package) and are tied to the annual Capital Improvement Program (CIP) and operating budgets. The horizon is one year, but choices have multi-year implications because they selectively improve the health of specific corridors and neighborhoods.

- Strategic layer (multi-year capital planning): Multi-year master plans, rate studies, bond issuance, and long-term level-of-service policy commitments sit at this layer. Here, the unit of analysis is a multi-year portfolio of programs and enabling investments (e.g., new transmission mains, storage, or treatment capacity). The strategic layer is more concerned with long-term trends like growth, climate risk, regulatory changes than with the specific alignment of next year's trenches.

Within this hierarchy, the optimization model in this chapter is deliberately positioned at the tactical–strategic interface. It operates on a one-year planning window (to align with annual CIP and condition-assessment cycles) but is designed to be re-run every year, so that a sequence of annual portfolios forms a coherent long-term renewal trajectory. The model sits "on top of" the LOF and COF models from Chapters 4 and 5, which already encapsulate segment-level failure likelihood and consequence. Its task is not to re-

predict failures, but to translate segment-level scores into an annual project portfolio that is consistent with budgets, delivery constraints, and utility priorities as shown in Figure 6-2. This figure summarizes how segment-level risk scores and portfolio-level considerations come together in the renewal decision process. Segments first populate the LOF–COF risk contour (left), from which high-risk candidates are brought into a risk–cost view (upper right) where obvious "high-risk, low-cost" projects separate from more ambiguous trade-off regions. A sunburst chart of the selected portfolio (lower right) then shows how the final annual program distributes costs across materials and diameters, providing a visual bridge between model outputs and practitioners discussions around the table.

This positioning also clarifies the division of labor between tools. The LOF and COF models characterize *how risky* each pipe segment is, given its environment and function. The portfolio model characterizes *what to do* about those risks this year, given limited capital, work-zone constraints, and competing goals such as equity or water-loss reduction.

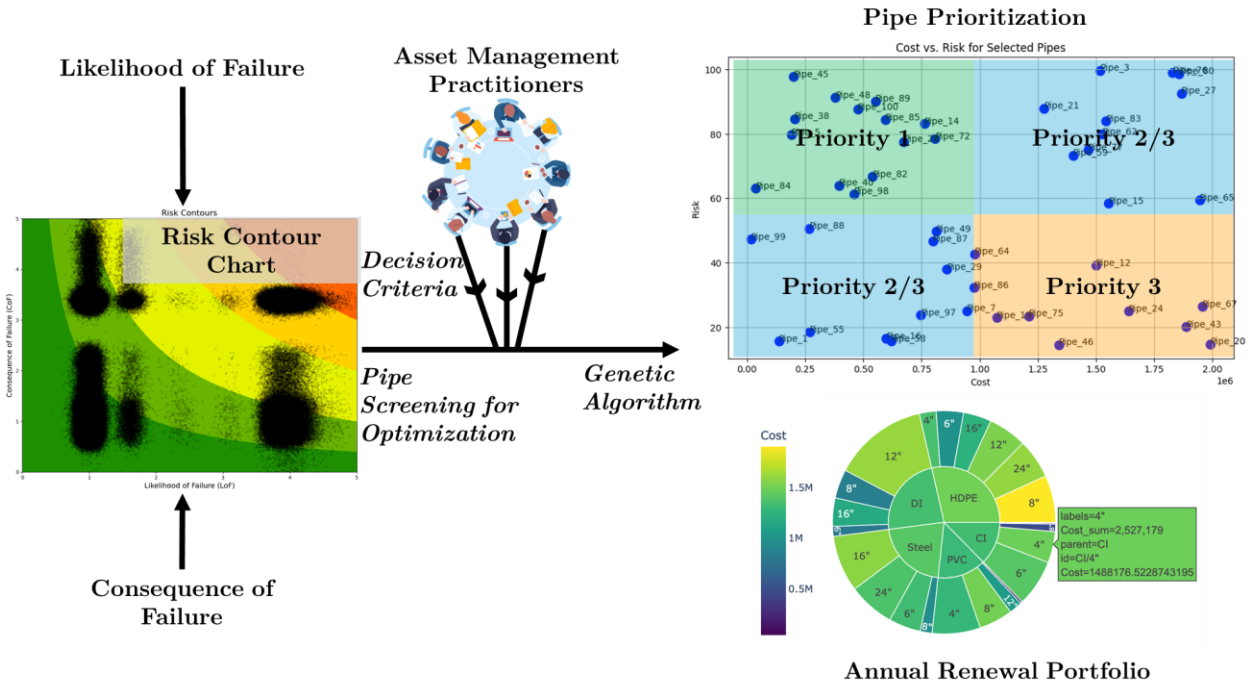

Figure 6-2: Conceptual decision space for risk-based renewal prioritization. The left panel shows the LOF–COF risk contour and dense cloud of pipe segments, which is used to pre-screen candidates above a minimum risk level. The upper-right panel displays candidate pipes in risk–cost space with qualitative priority regions, while the lower-right sunburst summarizes the composition of a selected portfolio by material and diameter; the central round-table icon emphasizes that these model outputs are based on criteria defined by the decision makers and intended to support, rather than replace, utility decision-makers.

6.2.2 Design principles for the optimization model

The optimization model is built around a small set of design principles intended to keep it aligned with practice, scientifically grounded, and explainable in use. These principles shape both the mathematical formulation and the choice of outputs.

Risk-based, in a traceable way: The primary driver is risk, defined as the product of the modeled Likelihood of Failure and Consequence of Failure for each project, $R_i = \mathrm{LOF}_i \times \mathrm{COF}_i$, on the 0–25 scale introduced earlier. The first objective is annual risk reduction, ΔR, computed as the difference between baseline risk and residual risk under a candidate portfolio. This maintains continuity with standard risk-management practice while making the risk-reduction contribution of each project explicit. Annual objectives and associated metrics are summarized in Table 6-1.

Table 6-1: Decision objectives and metrics

Objective	Symbol	Unit	Direction	Description	Source
Annual risk reduction	ΔR	Risk index units / year	Maximize	Reduction in total pipeline risk (LOF × COF) achieved by the selected portfolio in the planning year.	LOF (Ch. 4), COF (Ch. 5)
Annual renewal cost	$C(X)$	$ / year	Minimize	Total cost of renewal projects selected in the portfolio.	Utility cost data; unit-cost models
Risk reduction per $1M	$\Delta R_{\$}$	Risk units per $1M	Maximize	Efficiency metric: risk reduction normalized by annual renewal expenditure.	Derived from ΔR and $C(X)$
Service equity uplift	$E(X)$	Index (dimensionless)	Maximize	Improvement in equity indicators (e.g., risk reduction or service reliability in disadvantaged tracts).	Census / EJ data; risk maps
Customer-hours of disruption	$CH(X)$	Customer-hours	Minimize	Aggregate customer-hours of planned service interruption associated with selected projects.	Outage planning; demand maps
Water-loss reduction	$\Delta R(X)$	gallons/year	Maximize	Estimated annual reduction in non-revenue water due to renewal of leaking or high-risk segments.	Water-loss models; perf. models
Sustainability co-benefits	$S(X)$	Index (dimensionless)	Maximize	Composite indicator of environmental co-benefits (e.g., energy savings, GHG reduction).	Energy models; utility sustainability plans

Objective	Symbol	Unit	Direction	Description	Source
Budget utilization	$U_B(X)$	% of annual budget	Target range	Fraction of annual renewal budget used by the portfolio (e.g., 90–100%).	Derived from $C(X)$ and budget (B)

Multi-criteria beyond risk: Pure risk ranking is insufficient when many pipes have similar risk scores but very different locations, social contexts, and physical roles. The portfolio choice is therefore framed as a multi-criteria problem that combines annual risk reduction ΔR; annual renewal cost $C(X)$ together with risk reduction per \$1M, $\Delta R_{\$}$; service-equity uplift $E(X)$, which captures improvements in disadvantaged or under-served areas; customer-hours of disruption $CH(X)$, which represent planned outage burdens; water-loss reduction $\Delta W(X)$, derived from the water-loss component models; and optional sustainability co-benefits $S(X)$, such as reduced pumping energy or greenhouse-gas emissions. These criteria and their units are summarized in Table 6-1, while Table 6-2 lists the underlying project-level attributes (legacy materials, concurrent street projects, historical failures, service-line materials, and others) that feed into each criterion.

Table 6-2: Portfolio decision criteria at project level

Criterion name	Category	Scale / unit	Description	Source
LOF–COF risk score	Risk	0–25 (LOF × COF)	Aggregated risk index for project (i) from segment-level LOF and COF.	LOF (Ch. 4); COF (Ch. 5).

Criterion name	Category	Scale / unit	Description	Source
Customer satisfaction	Strategic	1–5 score	Priority derived from density/severity of customer complaints (pressure, color, surfacing water) in project area.	Customer complaint database.
Service equity index	Equity	Normalized 0–1 or z-score	Degree to which the project benefits socio-economically disadvantaged or historically underserved areas.	Census / EJ datasets.
Risky cluster index	Tactical	0–25	Combined risk of adjacent high-risk segments forming a spatial cluster.	Risk hotspot analysis.
Concurrent project opportunity	Tactical	Binary (0/1) or 0–1 score	Indicator or score for co-occurrence with street, sewer, or redevelopment projects that reduce marginal costs.	City DPD/DOT project GIS.
Legacy material removal	Operational	Binary / categorical	Indicator that project renews legacy materials (e.g., AC, wood stave, galvanized steel) in the area.	Utility GIS; material inventory.
Diameter / demand alignment	Operational	in or mm (binned)	Degree to which proposed diameter aligns with current/projected demands and hydraulic standards.	Hydraulic model; census data.
Pressure-zone management	Operational	Binary (0/1)	Indicator for projects that support PRV strategy or reduce pressure transients in critical zones.	Pressure-zone maps; PRV data.
Historical failure burden	Informational	Count / rate	Number or rate of historical failures in the project area, normalized by length and time.	Work-order database.
Service line material context	Regulatory	Categorical (Pb/Cu/other)	Presence of lead or other regulated service-line materials in area; may trigger coordination with LCRR actions.	Service line inventory.
Water-loss reduction potential	Sustainability	m^3/year	Expected annual reduction in leakage and non-revenue water if project is implemented.	Water-loss model (Ch. WL).
Construction disruption impact	Social	Customer-hours	Aggregate customer-hours of disruption associated with the project.	Outage + demand modeling.

In this dissertation, the GA is run with a hard annual budget constraint and soft preferences for other limits. The formulation, however, can readily accommodate additional hard constraints when utilities provide the necessary flags and capacity parameters.

Constraint-aware formulation: The annual program must satisfy hard feasibility limits rather than soft preferences. The model enforces an annual budget constraint $\sum_i C_i x_i \leq B$, where C_i is the cost of project i, x_i is the binary decision variable, and B is the mains-renewal portion of the CIP budget; a construction-capacity constraint such as a maximum annual length $\sum_i L_i x_i \leq L_{\max}$, reflecting crew capacity, traffic management, and seasonal work windows; moratorium and coordination constraints encoded via project-level flags for streets that cannot be reopened and indicators for projects that should be synchronized with other utilities; and, where needed, data-reliability filters that exclude or down-weight projects with very low information quality. The main variables and constraints are collected in Table 6-3, which serves as a compact dictionary for interpreting the optimization formulation. This table summarizes the decision variables and feasible constraints. In this dissertation the GA is run with a hard annual budget constraint and soft preferences for other limits. The formulation, however, can readily accommodate additional hard constraints when utilities provide the necessary flags and capacity parameters.

Table 6-3: Decision variables and constraints (including hard constraints set by the modeler and soft constraints used for visualization and decision support)

Symbol	Type	Role	Unit	Constraint / definition	Notes
x_i	Binary	Decision	–	$x_i = 1$ if project (i) is selected for renewal in the planning year; 0 otherwise.	Primary decision variable.
C_i	Parameter	Cost	$	Estimated renewal cost for project (i).	From cost models / unit costs.
R_i	Parameter	Risk	Risk index units	Baseline risk for project (i) (aggregated from LOF × COF of member segments).	From LOF and COF models.
E_i	Parameter	Equity	Index	Equity indicator for project (i) (e.g., weight for disadvantaged areas).	From census / EJ datasets.
CH_i	Parameter	Disruption	Customer-hours	Expected customer-hours of disruption if project (i) is implemented.	From outage and demand modeling.
ΔW_i	Parameter	WL reduction	gallons/year	Estimated water-loss reduction from renewing project (i).	From water-loss model.
B	Parameter	Budget	$	Annual renewal budget: $\sum_i C_i x_i \leq B$.	User-specified or from utility plan. Hard constraint.
L_i	Parameter	Length	ft	Pipe length in project (i).	Used for construction-capacity limits. Conceptually supported.
L_{max}	Parameter	Constraint	miles/year	Annual construction capacity: $\sum_i L_i x_i \leq L_{max}$	Optional but realistic delivery constraint.
w_k	Parameter	Objective wt	–	Weight assigned to objective (k) in scalarized fitness function.	Chosen with utility; can be varied.
M_i	Parameter	Moratorium	0/1	Moratorium flag: if $(M_i = 1)$, project (i) cannot be selected (constraint $(x_i \leq 1 - M_i)$	Encodes street/permit moratoria. Conceptually supported.
Q_i	Parameter	Data quality	1–5	Data reliability score; may be used to exclude or down-weight projects with very low reliability.	Optional robustness control. Used as a filter/weight, not an optimization constraint

Transparent and explainable: The model deliberately uses explicit weights on objectives, w_k, rather than collapsing everything into an opaque single score. Utility staff can inspect how a "risk-heavy" scenario (high weight on ΔR) differs from an "equity-heavy" or "water-loss-heavy" scenario. The Genetic Algorithm (GA) is used as a flexible

search engine over binary project-selection vectors, but the fitness function remains a weighted sum of interpretable criteria (explained in greater detail in Section 6.4). Figure 6-3 summarizes the workflow of the Genetic Algorithm used to construct an annual renewal portfolio.

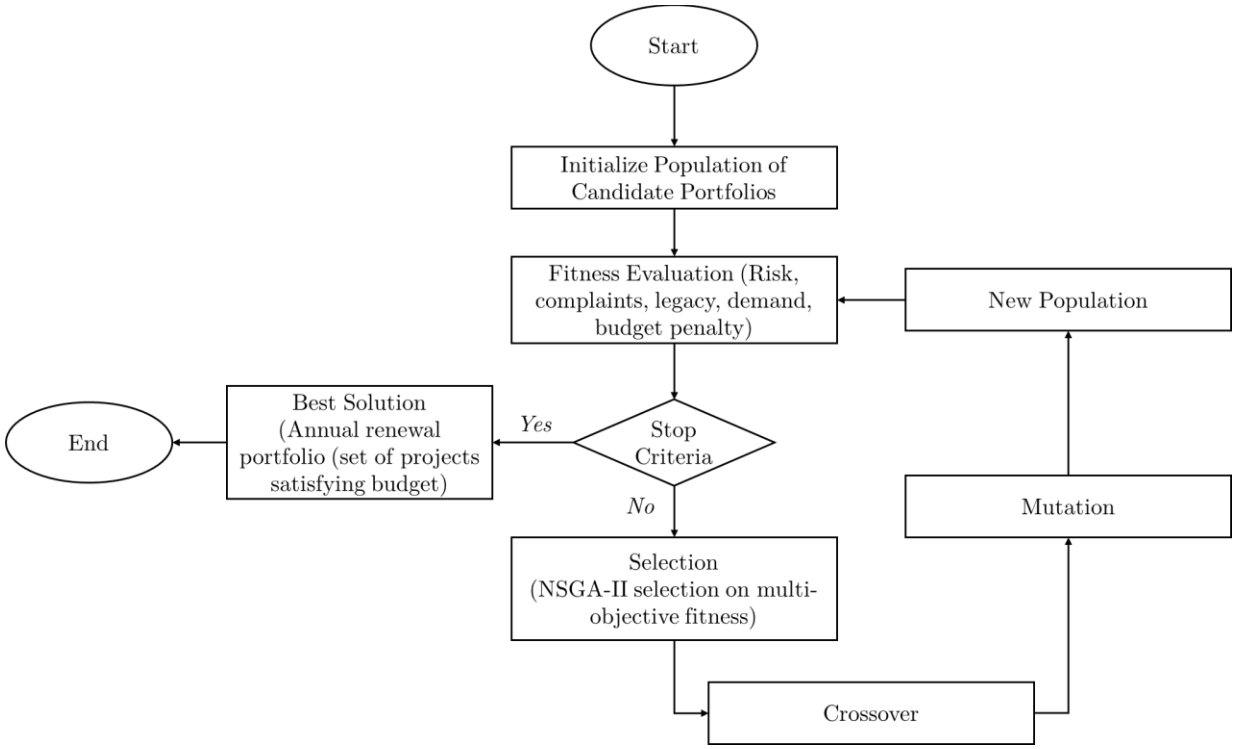

Figure 6-3: Genetic Algorithm workflow for annual renewal portfolio optimization. The algorithm initializes a population of candidate project portfolios, evaluates their multi-criteria fitness (risk capture, complaints proxy, legacy removal, demand priority, and budget penalty), and then iteratively applies selection (NSGA-II), crossover, and mutation to generate new populations until a stopping criterion is met. The final "best solution" is the selected annual renewal portfolio that satisfies the budget constraint while achieving a desirable balance across the competing objectives.

The process starts by initializing a population of candidate portfolios, each portfolio being a binary selection of projects that could be funded in the CIP year. Each candidate is then passed through the fitness evaluation block, where it is scored on the criteria that matter for this problem like risk capture (LOF$\times$COF), complaints proxy, legacy-material removal, demand priority, and a budget-penalty term that disfavors portfolios exceeding the annual constraint. NSGA-II selection then ranks and filters these candidates based on their multi-objective fitness, keeping those that strike good trade-offs across the criteria.

From this selected subset, crossover recombines parts of two portfolios and mutation flips individual project decisions, producing a new population that explores nearby but distinct combinations of projects. This loop of evaluation $\rightarrow$ NSGA-II selection $\rightarrow$ crossover $\rightarrow$ mutation continues until a stopping criterion is met (for example, a maximum number of generations or negligible improvement in fitness). At that point, the algorithm reports a "best solution": an annual renewal portfolio that satisfies the budget constraint and delivers a desirable balance between risk reduction, customer impacts, equity, and legacy-material removal. The figure is intentionally generic in its logic and any future criteria can be added inside the fitness block while remaining specific enough that readers can map each step directly to the implemented code.

The Genetic Algorithm specifications in Table 6-4 are chosen to match both the structure of the renewal problem and the size of the candidate set, rather than being arbitrary defaults.

Table 6-4: Configuration and hyperparameters of the Genetic Algorithm used for annual renewal portfolio optimization. The table lists the main GA settings, example values used in this study, and brief rationales for each choice, balancing portfolio diversity, convergence quality, and computational cost in networks of realistic size.

Parameter	**Value**	**Rationale**
Population size	300	Balances diversity of portfolios with computational cost.
Number of generations	120	Sufficient for convergence in pilot runs on candidate project sets of realistic size.
Crossover rate	0.9	Promotes recombination of good sub-portfolios.
Mutation rate	0.02	Introduces diversity and avoids premature convergence.
Selection method	NSGA-II	Robust selection pressure with simple implementation.
Crossover type	Two-point	Adequate for binary encodings; easy to interpret.
Mutation type	Bit-flip	Natural for binary project-selection variables.
Penalty coefficient	Tuned (e.g., 10–100 $\times$ median fitness)	Ensures budget and capacity violations are strongly discouraged.
Number of runs (seeds)	10	Allows assessment of run-to-run variability and robustness of selected projects.
Stopping criteria	Max generations OR relative fitness improvement $< \varepsilon$ over 50 generations	Balances exploration and run time.

A population size on the order of a few hundred individuals (here $\mu \approx 300$, with an equal number of offspring per generation) provides enough portfolio diversity to explore combinations of high-risk, high-equity, and high water loss segments, while keeping run times acceptable for a one-year CIP on a standard workstation. Pilot runs on screened candidate sets of realistic size (top ≈100 miles of mains) showed that portfolio metrics such as normalized risk capture and cost penalty stabilize well before 120 generations, so the generation count was set to 120 as a conservative upper bound that still leaves room for exploration in more heterogeneous systems.

The choice of variation operators is aligned with the binary nature of the decision variable (select/not select each project). Two-point crossover operates on contiguous stretches of the binary chromosome, which is a natural way to recombine "good sub-portfolios" (for example, clusters of projects that work well together under the budget). Bit-flip mutation with a modest per-gene probability injects enough randomness to escape local patterns (such as always picking the same cluster of trunk mains) without turning the search into a random walk. A relatively high crossover probability and moderate mutation probability are standard in multi-objective GA practice for combinatorial

portfolio problems, where the goal is to recombine promising structures while still probing new combinations at the margin.

Selection is handled by NSGA-II, which is widely used for multi-objective optimization because it preserves a diverse set of non-dominated portfolios instead of collapsing everything into a single scalar score too early. This is particularly important here, where the utility may want to compare "risk-heavy", "equity-heavy", and "water-loss-heavy" portfolios before choosing a preferred one. Budget and delivery constraints are enforced through two complementary mechanisms: (i) the initial population is seeded with greedily constructed feasible portfolios that respect the annual budget, and (ii) any candidate that violates the budget incurs a strong penalty in the cost-deviation objective, ensuring that infeasible portfolios are dominated by feasible ones. Finally, running the GA multiple times with different random seeds and summarizing how often each project appears across runs allows the model to distinguish robust priorities (projects that are selected in most runs) from marginal ones, which is essential when recommending real-world renewal programs that must be defensible under scrutiny.

Embedded in a temporal commons perspective: Each one-year CIP is one allocation episode in a long sequence, in which the pipeline system behaves as a temporal

commons and current decision-makers draw on a finite pool of structural capacity, financial capital, and social tolerance for disruption that must also serve future users. Overinvesting in low-impact corridors now, or systematically neglecting high-consequence areas, creates intertemporal inequities and future regret. By reporting risk reduction per dollar, equity uplift, and customer-hours of disruption for each annual portfolio, the model makes these trade-offs visible and auditable over time. Annual portfolios can then be compared not only to each other, but also to the utility's own historical patterns of spending and service, reinforcing the idea that sustainability and equity are cumulative, path-dependent outcomes, not one-year events. Together, these design principles motivate a portfolio model that is mathematically rigorous yet tractable, grounded in LOF and COF, and capable of supporting real-world renewal planning.

6.3 From Segment-Level Scores to Eligible Renewal Candidates

This section describes how calibrated LOF, COF, and auxiliary attributes from earlier chapters are assembled into a unified decision dataset and filtered into a high-risk, high-value candidate set for optimization. It also explains how individual segments are

packaged into constructible projects, bridging the gap between segment-level modelling and project-level capital planning.

6.3.1 Integration of LOF, COF, and auxiliary scores

The optimization model consumes segment and project level scores that have already been constructed and evaluated in previous chapters. For each pipe segment i, Chapters 4 and 5 provide calibrated LOF and COF indices on a 0–5 scale that incorporate material, diameter, age, surrounding environment, customer impacts, and other inputs. Their product defines a baseline risk score.

$$R_i = \mathrm{LOF}_i \times \mathrm{COF}_i,$$

on a 0–25 scale. This scalar risk score is the primary link between the segment-level models and the portfolio layer, and it is used both for pre-screening and as a core driver of the renewal decision.

In addition to LOF and COF, segments and their associated projects are annotated with auxiliary attributes that inform the other decision criteria introduced in Section 6.2.2. These include a renewal cost C_i, estimated from unit-cost models that depend on diameter, depth, surface type, traffic conditions, and material and then aggregated to the

project level; a projected water-loss reduction ΔW_i, derived from the source-to-tap water-loss models that link performance indices and failure modes to volumetric losses; a service-equity index that measures the extent to which a segment lies in or serves disadvantaged or historically under-invested areas using census and environmental-justice datasets; and an estimate of customer-hours of disruption CH_i, obtained from outage modelling, local demand patterns, and the expected duration and spatial footprint of planned works. Further attributes capture sustainability indicators such as expected changes in pumping energy, pressure-transient exposure, or greenhouse-gas emissions when high-leakage or high-pressure segments are renewed, as well as categorical indicators for legacy materials, historically problematic sites, concurrent street or utility projects, pressure-zone edges, and service line material context.

In the current implementation, the GA uses a subset of these attributes as explicit objectives and proxies, most notably risk, a complaints-based proxy, legacy-material flags, and a demand-priority index while the remaining attributes are maintained in the same decision table as optional columns that can be activated when utilities have the corresponding data and wish to include them. All of these quantities are inputs to the portfolio chapter rather than outputs of it. The detailed modelling and data assumptions behind

LOF, COF, water-loss, and cost estimates have already been specified and evaluated in earlier chapters and are not repeated here. Instead, they are treated as a consistent, system-wide attribute table (summarized in Table 6-2) from which the optimization model reads. This separation of concerns is deliberate as it allows utilities to improve or replace individual component models (for LOF, COF, or water loss) without changing the structure of the portfolio optimization, as long as the same set of attributes is supplied with compatible scales and units.

6.3.2 Screening and eligibility rules

Applying the optimization model to every segment in the network would be both computationally wasteful and conceptually misleading, because many pipes have low risk, limited consequences, or insufficient data quality to justify explicit optimization. The first step in the pipeline is therefore an eligibility screen that defines the set of segments that are allowed to compete for inclusion in the annual portfolio. The primary screen is a risk-based eligibility rule on the LOF–COF plane. Using the 0–5 LOF and COF scales, an iso-risk contour is defined by

$$R_i = \mathrm{LOF}_i \times \mathrm{COF}_i = 9,$$

so that, in the base configuration, only segments with $R_i \geq 9$ (this can change based on the proportion of pipes in different risk zones) are considered eligible for renewal in the current planning year. This corresponds to a band of the LOF–COF space where failures are either relatively likely, have substantial consequences, or both. The LOF–COF risk surface, coloured from green (low risk) through yellow and orange to red (high risk) and annotated with management zones, is shown in Figure 6-1. The same logic can be tuned to use stricter thresholds for special analyses (for example, defining a "very high-risk" band at $R \geq 15$ for clustering), but the conceptual role of the contour is unchanged that is, it filters out segments that are clearly low-priority from a capital-renewal perspective.

Beyond the risk threshold, the eligibility screen includes additional filters that can be activated when the data are available. Segments with incomplete or very low-reliability attributes such as missing diameter, unknown material, or unreliable location are either excluded or tagged with a low data-quality score Q_i, allowing the utility to decide whether to include them explicitly or manage them through separate programs. Very short, isolated segments that cannot form a constructible project without excessive mobilization can be excluded from the candidate set and handled via operational repairs instead,

thereby keeping the optimization focused on deliverable work packages. Segments under an explicit moratorium (for example, recently resurfaced streets or shared corridors under multi-agency agreements) are flagged with an indicator $M_i = 1$ and excluded from the optimization by enforcing $x_i \leq 1 - M_i$ in the decision variables. Similarly, segments earmarked for other specific projects, such as an upcoming transmission-main expansion funded from a different program, can be temporarily excluded to avoid double counting. In case-study runs where some of these filters are not yet populated, the framework defaults to the risk-based screen, but the underlying formulation is kept general so that more nuanced eligibility rules can be adopted as utilities enrich their data. This eligibility stage therefore focuses computational effort on the segments where choice matters most and ensures that the subsequent optimization is aligned with utility practice, where low-risk neighborhood pipes are typically not reviewed one-by-one in annual CIP deliberations.

6.3.3 From segments to projects

Although LOF, COF, and auxiliary scores are computed at the segment level, renewal is delivered at the level of projects that is, packages of work that can be designed, permitted, and built as coherent units. A project may consist of a single long main on an arterial street, a cluster of short segments forming a neighborhood loop, or a group of

parallel mains within a pressure zone that must be renewed together. In practice, each row of the GA input file represents one such decision unit; depending on how the utility prepares the data, that unit may be a single segment or a pre-aggregated project that bundles several contiguous segments. The optimization model itself is agnostic to this choice, as long as each row carries the aggregated attributes required for decision-making.

The mapping from segments to projects is constructed using three main ideas. First, spatial adjacency and constructability are used to form candidate projects when segments form contiguous, constructible units, such as a continuous run of main along a street or a compact cluster within a neighborhood where trenching and traffic management can be planned together. Standard GIS tools, such as buffer and dissolve operations, are used to identify these clusters, subject to practical limits on project length and complexity. Second, risk and impact coherence are used to keep risk and consequence profiles relatively homogeneous within each package. High-consequence trunk mains, for example, may be grouped with immediately adjacent segments of similar function but not with distant low-consequence neighborhood laterals, even if they share the same material and age. This maintains interpretability where each project has a clear risk narrative such as "critical transmission corridor serving high-density urban cores" rather than being a

statistical mixture of unrelated risks. Third, coordination with other programs is incorporated by adjusting candidate projects to respect known street resurfacing plans, sewer projects, transit works, or redevelopment corridors. Where possible, project boundaries are aligned so that "dig once" opportunities are captured, for example by replacing a main when the street is already scheduled for reconstruction or synchronizing water-main replacement with sewer relining, while projects that would violate moratoria or conflict with other utilities are split or postponed.

Once projects are defined, the primary decision variable in the optimization model is a binary indicator $x_i \in \{0,1\}$ for each project i, denoting whether that project is included in the annual portfolio. Segment-level attributes are aggregated to projects using appropriate rules. For example, risk scores $\mathrm{LOF} \times \mathrm{COF}$ are aggregated as sums or length-weighted averages to produce project-level risk R_i; costs, water-loss reductions, and customer-hours of disruption are summed over constituent segments; and equity and sustainability indicators are aggregated using length-weighted averages or maxima, depending on their interpretation (for example, whether a project should be considered equitable if any portion lies in a disadvantaged tract, or only if most of its length does). In datasets where each row is already a single segment, these aggregation rules reduce to identity

mappings, and the GA simply treats each segment as a one-segment project. The segment–project mapping is retained so that, once the GA selects a set of projects, each underlying segment can be reported with its specific risk reduction and other attribute changes. This also allows the renewability validation protocols described later. This two-layer structure, with segments as the unit of modelling and projects as the unit of decision, mirrors how utilities work. It respects constructability and customer impacts while preserving enough granularity to evaluate how each annual portfolio affects the underlying distribution of risk, equity, and water losses across the system.

6.4 Decision Variables, Constraints, and Data Requirements

This section formalizes the optimization problem in algebraic terms by specifying the decision variables, the feasibility constraints they must satisfy, and the minimum set of attributes that each candidate project must carry. Together, these elements define the model universe on which the ideal portfolio is explored.

6.4.1 Decision variables

The renewal optimization model is built around a binary decision vector

$$X = \{x_1, \ldots, x_N\},$$

where each element x_i corresponds to a candidate project (or, in the simplest case, a single pipe segment) in the screened dataset. For the one-year Capital Improvement Program (CIP) considered in this chapter, the decision variable is defined as

$$x_i = \begin{cases} 1 & \text{if project/segment } i \text{ is renewed in this CIP,} \\ 0 & \text{otherwise.} \end{cases}$$

In the current implementation, each row in the GA input file represents one such decision unit with an associated cost, risk, and set of attributes. The GA operates directly on a binary string of length N that is, an individual chromosome is therefore a candidate annual renewal portfolio. This encoding keeps the model close to how utilities plan their renewal programs ("include or exclude project i this year"), and it is compatible with either segment-level or project-level representations, provided the attributes are consistently defined.

The framework can be extended to staged decisions over multiple years by introducing scheduling variables $x_{i,t}$ that indicate whether project i is executed in year t of a multi-year CIP. However, because this dissertation focuses on a one-year planning horizon aligned with annual budget cycles, the implemented GA uses only the single-year binary

vector X. Multi-year formulations are discussed conceptually in the temporal-commons framing but are not solved explicitly in this chapter.

6.4.2 Constraints

The decision vector X is subject to a set of feasibility constraints that reflect the financial and operational limits of the utility. The central hard constraint implemented in the current GA is the annual budget constraint

$$\sum_{i=1}^{N} C_i x_i \leq B,$$

where C_i is the total renewal cost of project i and B is the mains renewal portion of the annual CIP budget. In the code, C_i is either taken directly from a project-level "Renewal Cost" field, when available, or computed as length multiplied by a diameter-specific unit cost. The GA seeds its initial population with feasible portfolios and then discourages budget violations through a strong penalty in the fitness function and repair operations that trim projects until the portfolio cost falls at or below B. This combination of feasible seeding and penalty-based enforcement ensures that most of the search effort is concentrated on portfolios that are realistically fundable.

Beyond the budget, many utilities face additional delivery constraints such as a maximum annual construction length, a limit on the number of concurrent work zones, or resource ceilings (equipment and crew strength) for specialized events. The general formulation accommodates such constraints by imposing limits of the form $\sum_i L_i x_i \leq L_{\max}$ for total length, or by tracking the number of projects in each work zone and penalizing portfolios that exceed capacity. The GA design, is intentionally modular and additional constraints can be encoded either as explicit inequalities or as penalty terms without altering the chromosome structure.

Spatial and operational constraints are handled in a similar way. Projects that fall under moratoria such as recently resurfaced streets or protected corridors are flagged during preprocessing and removed from the candidate set, effectively fixing their decision variables at $x_i = 0$. Where clustering is desirable, for example to avoid scattered small work sites, the model can prioritize high-risk clusters via the "HighRiskCluster" field and optional filtering to specific cluster IDs; the present implementation uses this feature primarily to define the candidate pool (Section 6.3) rather than as a strict constraint. Data-quality considerations are handled through optional exclusion or down-weighting of segments with very low reliability scores, consistent with the data-reliability ladder defined

in the LOF chapter. At this stage of the work, low-reliability assets are typically excluded from the GA and managed through separate monitoring or investigative programs, but the framework can accommodate reliability-weighted objectives if utilities want the optimization to reflect data confidence explicitly.

6.4.3 Required inputs and sources

Each decision unit in the GA input file must carry a minimum set of attributes, some observed and some model-derived, so that the objectives and constraints can be evaluated consistently. At a minimum, the model requires: (i) a calibrated LOF and COF pair on the 0–5 scale from Chapters 4 and 5, along with their product $R_i = \mathrm{LOF}_i \times \mathrm{COF}_i$; (ii) a length measure, used both for cost calculations and for normalizing some objectives; (iii) a cost representation, either as an explicit project-level "RenewalCost" or as sufficient information (diameter, surface type) to infer a diameter-specific unit cost; and (iv) basic identifiers such as material, diameter, and project or segment ID. In the implemented GA, these fields are complemented by three additional attributes that support the non-risk objective specifically, a complaints-based proxy ("ComplaintsProxy"), a binary indicator of legacy material removal ("LegacyRemovalScore"), and a demand-priority index ("WaterDemandPriority"). The complaints proxy is either the normalized count of historical

complaints or failures, when such data are available, or a heuristic combination of LOF, age, and small-diameter emphasis when complaint data are sparse. The legacy score is derived from material codes that match a user-editable list of legacy materials (for example, older CI, AC or early-generation PVC), and the demand-priority index combines diameter and COF to highlight large, high-consequence mains.

Conceptually, the framework can also ingest attributes for equity (for example, an index based on census, socio-economic indicators and environmental-justice layers), explicit water-loss reduction ΔW_i from water-loss models, and sustainability indicators such as energy or greenhouse-gas savings. In the current code, these additional attributes are kept in the same decision table but are primarily used in evaluation and scenario analysis rather than as GA objectives, because not all utilities have them in a consistent form. This design choice preserves generality where the optimization layer treats LOF, COF, and the three implemented proxies as a minimal, portable objective set, while still allowing utilities to "plug in" richer attributes as data and institutional priorities evolve.

Finally, it is important to be explicit about which quantities are observed and which are modeled. LOF and COF are model outputs, calibrated and validated in earlier chapters. Historical complaints and failure counts are observations, but they are filtered

and normalized before entering the GA. Costs are partly observed (for example, from recent bid tabs) and partly modelled via unit-cost curves. Water-loss reduction, equity indices, and sustainability metrics, when used, are derived quantities computed from other models and spatial overlays. Table 6-2 summarizes these inputs for each decision unit and points back to the relevant chapters or appendices where their construction is documented.

6.5 Multi-Objective Formulation: Objectives and Scalarizations

This section defines the 3 objectives that the optimization model seeks to balance namely, risk reduction, cost, complaints, legacy removal, demand priority, and optional extensions such as equity or water loss and explains how they are combined. It also introduces the scalarization strategy that turns these multi-dimensional criteria into comparable portfolios for decision-makers.

6.5.1 Risk Reduction

In principle, risk reduction for a portfolio X can be defined as the difference between the baseline system risk (before renewal) and the residual risk after renewing the selected set of projects:

$$\Delta R(X) = \sum_i R_i^{\text{baseline}} - \sum_i R_i^{\text{post-renewal}}(x_i),$$

where $R_i^{\text{post-renewal}}(x_i)$ reflects the change in risk when project i is executed. In practice, for annual planning and for the data structures used in this chapter, it is more convenient to work with a normalized "risk capture" objective that measures the share of baseline risk removed by the portfolio. The implemented objective therefore computes:

$$f_{\text{risk}}(X) = \frac{\sum_i R_i x_i}{\sum_j R_j},$$

where the denominator is the total risk over the screened candidate set. This quantity lies between 0 and 1 and can be interpreted as the fraction of candidate-set risk addressed by the selected projects. It is closely related to $\Delta R(X)$ but does not require modelling post-renewal residual risk explicitly, which would introduce additional assumptions about how

quickly risk re-accumulates after replacement. Normalizing by the total risk within the candidate pool also makes comparisons across utilities and across scenarios easier, because the scale of $f_{\text{risk}}$is invariant to the absolute size of the network.

6.5.2 Cost and affordability

The cost objective is defined as the total renewal cost of the selected projects,

$$C(X) = \sum_i C_i x_i \,,$$

with C_i computed as described in Section 6.4.2. Rather than minimizing cost directly, which would be at odds with the desire to use the budget effectively, the GA enforces the budget through a hard constraint and uses a cost-deviation term in the objective vector:

$$f_{\text{cost}}(X) = \left|\frac{C(X) - B}{B}\right| + \text{Penalty}(C(X) > B),$$

where the penalty term is large when portfolios exceed the budget. This formulation encourages portfolios that are close to, but not over, the budget and avoids the trivial solution of spending very little but achieving limited risk reduction. For reporting and comparison, the model also computes risk reduction per \$1M as $\Delta R_{\$} = f_{\text{risk}}(X)/(C(X)/$

10^6)and budget utilization $U_B(X) = C(X)/B$, as summarized in Table 6-1 and in the portfolio comparison tables in the results chapter.

6.5.3 Service equity

Equity enters the formulation in two layers. At the attribute level, each decision unit is assigned an equity index based on the socio-economic and environmental-justice characteristics of the population it serves, or on historical patterns of under-investment. At the objective level, the GA uses a service-experience capture metric that favors projects in areas with a high burden of documented service problems, constructed from the most reliable operational dataset available at the utility or national level. Concretely, the objective is written as

$$f_{\text{service}}(X) \;=\; \frac{\sum_i S_i \, L_i \, x_i}{\sum_j S_j \, L_j},$$

where $L_i$is project length, x_i is the binary decision variable, and S_i is a service-experience burden index derived from observed data such as customer complaints per mile, main-break rates, regulatory violation records, or outage events, depending on which dataset is best curated and consistently maintained for the study area. In the implementation used

for the case studies, S_i is always based on observed counts or rates from the participating utility's own systems (e.g., work-order and call-centre logs), not on a synthetic proxy. When no such dataset is available at sufficient quality, this objective is either omitted from the optimization or treated purely as a diagnostic metric in the evaluation phase rather than as a driver of the GA search.

To complement this operational view of service experience, the framework also allows equity metrics derived from census and environmental-justice layers to be computed for each portfolio after optimization. These can be expressed either as uplift (improvement in equity indicators for disadvantaged tracts) or as disparity reduction (decrease in the gap between high- and low-service cohorts). In the current implementation these equity indices are tracked as evaluation metrics rather than hard GA objectives, because their data sources are not yet standardized across utilities. However, the scalarization approach in Section 6.5.6 makes it straightforward to promote them to explicit objectives as data availability and quality improve.

6.5.4 Operational disruption (customer-hours)

Operational disruption is expressed in customer-hours of interruption, defined as the number of affected customers multiplied by the duration of planned outages. In the broader source-to-tap framework, these quantities are estimated from outage planning tools and demand maps, and they can be aggregated to obtain a portfolio-level disruption measure $CH(X) = \sum_i CH_i x_i$. Minimizing $CH(X)$ directly would favor small, low-impact projects, which may conflict with the risk-reduction objective, so disruption is more usefully treated as a constraint or a monitored trade-off. In the current GA implementation, customer-hours are not part of the fitness vector, because utilities expressed a preference for focusing first on risk, legacy material removal, and demand priority. Instead, customer-hours are computed for selected portfolios and used in scenario analysis to identify options that strike a reasonable balance between risk reduction and planned disruption. Incorporating $CH(X)$ as a formal objective is therefore a straightforward extension rather than a structural change.

6.5.5 Sustainability and water loss

Water loss and sustainability outcomes are handled similarly. Any available water loss models or water audits can be used to provide an estimate of average volumetric losses for each segment under current conditions and an expected reduction ΔW_i if the segment is renewed. A natural sustainability objective is then

$$f_{\mathrm{WL}}(X) = \sum_i \Delta W_i x_i \,,$$

possibly normalized by total system losses. Additional sustainability metrics, such as reductions in pumping energy or greenhouse-gas emissions, can be constructed by combining ΔW_i with energy use factors and system-level carbon intensities. At present, the GA focuses on the four objectives that are both widely available and interpretable across utilities namely risk capture, complaints capture, legacy removal, and demand priority under a cost-deviation penalty. Water-loss and energy metrics are tracked for candidate and selected portfolios using the outputs of the water-loss models and reported in the evaluation chapter. This choice keeps the core GA lean while still enabling sustainability-oriented "what-if" comparisons, for example comparing a risk-dominant portfolio with a water-loss-dominant portfolio constructed by reweighting objectives.

6.5.6 Scalarization and Pareto analysis

The GA operates in two stages. During the evolutionary search, it treats the problem as a true multi-objective optimization and uses the NSGA-II algorithm to compare portfolios based on the vector of objectives

$$f(X) = \begin{cases} (f_{\text{risk}}, f_{\text{complaints}}, f_{\text{legacy}}, f_{\text{demand}}, -f_{\text{cost}}), & \textit{if risk objective included} \\ (f_{\text{complaints}}, f_{\text{legacy}}, f_{\text{demand}}, -f_{\text{cost}}), & \textit{otherwise,} \end{cases}$$

where f_{legacy} and f_{demand} are normalized capture metrics analogous to those for risk and complaints, and $-f_{\text{cost}}$ denotes that low cost deviation is preferred. NSGA-II maintains a diverse set of non-dominated portfolios, those for which no other portfolio is better on all objectives simultaneously and forms an empirical Pareto front. This is where the multi-objective nature of the problem is most visible, and it allows the model to explore, for example, portfolios that sacrifice some risk capture to remove more legacy materials.

In the second stage, a scalarization is used to select a single recommended portfolio for reporting. User-defined weights $w_{\text{risk}}, w_{\text{complaints}}, w_{\text{legacy}}, w_{\text{demand}}, w_{\text{cost}}$ are entered via a configuration sheet (currently performed on Google Colab) and normalized to sum to one. For each Pareto-efficient portfolio, a composite utility score is then computed as

$$F(X) = w_{\text{risk}} f_{\text{risk}}(X) + w_{\text{complaints}} f_{\text{complaints}}(X) + w_{\text{legacy}} f_{\text{legacy}}(X) + w_{\text{demand}} f_{\text{demand}}(X) - w_{\text{cost}} f_{\text{cost}}(X)$$

The minus sign on the cost term reflects the preference for small deviations from the budget. The portfolio with the highest $F(X)$ and a cost at or below the budget is selected as the "baseline" solution for that weight set. Interpretation of the weights is straightforward. Increasing w_{risk} produces risk-heavy portfolios that concentrate on high-risk corridors; increasing w_{legacy} favors removal of legacy materials; and increasing $w_{\text{complaints}}$ shifts priority toward areas with high observed service issues. Because the underlying objectives are still stored separately, the same GA run can be used to generate Pareto-like trade-off plots (for example, risk capture versus cost, or complaints capture versus cost) and to compare "risk-dominant" and "equity-leaning" portfolios under alternative weight choices.

Within this scalarization framework, the GA produces many candidate annual portfolios that differ in how they trade off risk reduction, cost, equity, and water-loss benefits. Figure 6-4 uses a simple synthetic example to show how these solutions populate a cloud in risk–budget space, how an efficient frontier emerges from that cloud, and how individual portfolios with different weight sets ("risk-heavy", "equity-heavy", "balanced")

move along that frontier. The same figure also contrasts these portfolios in a separate risk–equity plane and in a bar chart of normalized metrics, illustrating how Pareto trade-offs work to better understand the results in the EVV section of this chapter.

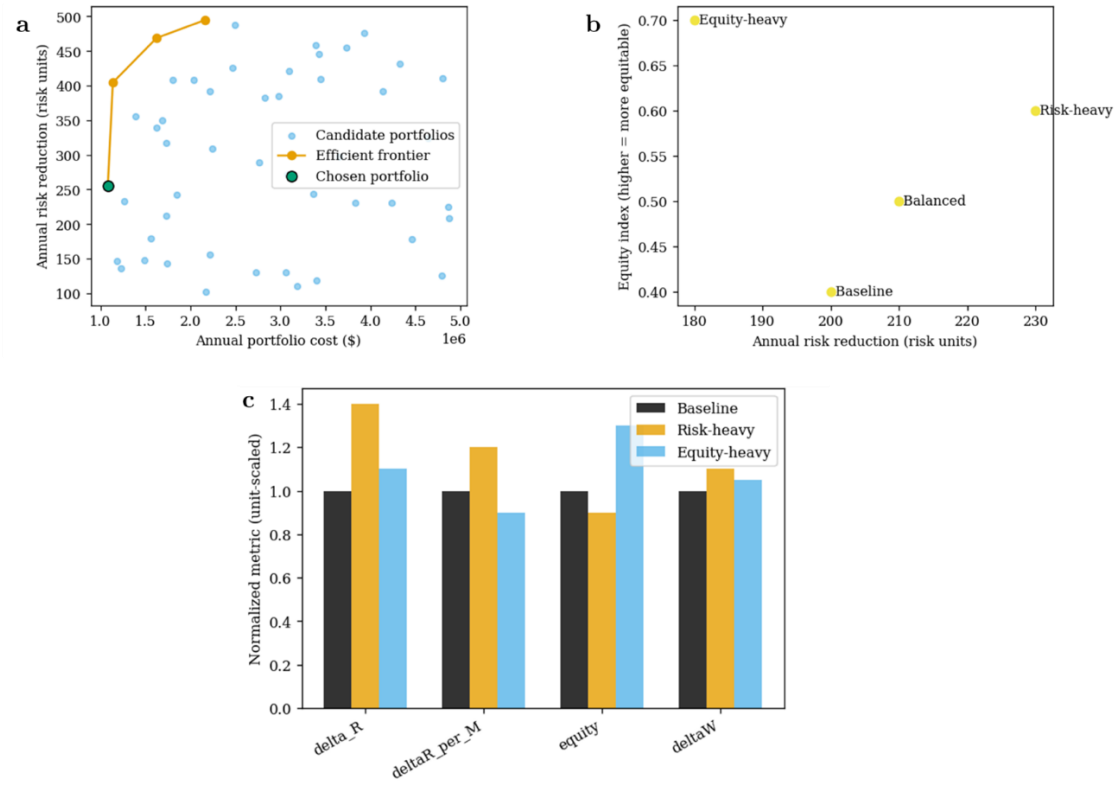

Figure 6-4: Schematic illustration of portfolio trade-offs produced by the optimization model (values illustrative only). Panel (a) shows a cloud of candidate annual portfolios in risk–budget space, with the efficient frontier and a chosen portfolio highlighted. Panel (b) illustrates the trade-off between annual risk reduction and an equity index for three example portfolios ("baseline", "risk-heavy", and "equity-heavy"). Panel (c) compares the same three portfolios across normalized metrics (risk reduction, risk reduction per $1M, equity, and water-loss reduction), emphasizing that the model supports explicit, quantitative comparison of alternative weightings rather than a single fixed ranking.

6.6 Genetic Algorithm Design and Implementation

This section describes how the optimization problem is actually solved using a Genetic Algorithm, including the chromosome encoding, population initialization, variation operators, and convergence checks. The emphasis is on showing how the chosen GA design makes the search tractable, interpretable, and reproducible for real utilities.

6.6.1 Encoding and initialization

The GA encodes each candidate annual portfolio as a binary chromosome of length N, where the i-th gene corresponds to the decision variable x_i. A "1" at position i indicates that project i is included in the portfolio and a "0" indicates exclusion. This encoding is both natural and efficient for project-selection problems and integrates seamlessly with the segment-to-project mapping described in Section 6.3.3.

Initialization is handled in a way that combines heuristic insight with diversity. The initial population includes several "greedy" feasible seeds constructed by sorting candidate projects according to different priority metrics like raw risk, complaints proxy, legacy-material score, and demand priority and then adding projects in decreasing order of "benefit per unit cost" until the budget is exhausted. These seeds emulate how a human

planner might assemble a program under a single dominant criterion and ensure that obviously good portfolios are present from the start. The population is then augmented with feasible random portfolios generated by shuffling project indices and adding projects until the budget is reached, and, if needed, with randomly sampled chromosomes that are repaired by dropping projects until the cost falls below the budget. This mixture of structured and random initialization increases the chance that the GA will explore both familiar and non-intuitive combinations, while respecting the budget constraint from the outset.

6.6.2 Fitness evaluation and constraint handling

Fitness evaluation translates each binary chromosome into the objective vector $\mathbf{f}(X)$ described in Section 6.5. For a given portfolio, the code identifies the selected projects, computes normalized captures for risk, complaints, legacy removal, and demand priority by dividing their summed contributions by the corresponding totals over the candidate set and then computes the cost-deviation term. Normalization ensures that all capture objectives lie in the range $[0,1]$, making them comparable and reducing sensitivity to the absolute scale of risk or complaint counts in any particular utility.

Constraints are handled through a combination of feasibility-preserving operators and penalties. Budget adherence is encouraged first by seeding the population with feasible portfolios and by repairing randomly generated individuals that initially violate the budget. Remaining violations are discouraged through a large penalty term added to the cost objective whenever $C(X) > B$, effectively pushing over-budget portfolios to the dominated region of objective space in NSGA-II. Delivery and spatial constraints, where used, can be handled similarly by attaching penalties to portfolios that exceed maximum length or violate moratoria. In the EVV experiments presented in this chapter, budget is treated as the only strict constraint, and other limits are monitored post-hoc, this is a deliberate choice to keep the implementation tractable while preserving a clear path to more restrictive formulations in future work.

6.6.3 Variation operators and parameters

The GA uses standard variation operators from the DEAP library tailored to the binary encoding. Selection is performed using NSGA-II's non-dominated sorting and crowding distance, which together favor portfolios that lie on the Pareto frontier while maintaining diversity along it. Crossover is implemented as two-point crossover where two parent chromosomes exchange contiguous segments between two randomly chosen

positions. This operator preserves local blocks of decisions (for example, combinations of projects along one corridor) and is easy to interpret. Mutation is implemented as bit-flip mutation with an independent probability per gene; in other words, each project has a small chance of switching from "selected" to "not selected" or vice versa in each generation. These operators are natural for a binary decision space and ensure that the search can both recombine good sub-portfolios and occasionally introduce new projects.

Hyperparameters are chosen based on small pilot experiments and practical runtime considerations. A typical configuration uses a population size of $\mu = 300$ portfolios, an offspring size of $\lambda = 300$, a crossover probability of 0.7, a mutation probability of 0.2 at the individual level (with a lower per-gene flip probability), and 120 generations. These values were found to provide a good balance between exploration and computational cost for candidate sets on the order of a few hundred projects, which is typical for annual high-risk pools. Table 6-4 summarizes the chosen hyperparameters and their rationale, including the penalty level used for budget violations and the recommended number of independent runs (different random seeds) when assessing robustness.

6.6.4 Convergence diagnostics and robustness

Convergence is assessed using both numerical and visual diagnostics. During each run, the GA records the average and extreme values of the objective vector across the population. Plotting these statistics over generations (Figure 6-4) reveals how quickly the population improves and whether key objectives plateau, indicating diminishing returns from further evolution. In practice, risk capture and complaints capture show rapid early gains followed by stabilization, while the cost-penalty term converges toward small values as the population learns to stay on budget.

To assess robustness, the model can be run multiple times with different random seeds, each producing its own Pareto set and recommended portfolio under the same weight set. Two complementary indicators are then examined. First, the spread of objective values across runs indicates how sensitive the overall performance is to stochastic variation in the GA; narrow spreads suggest robust convergence. Second, the selection frequency of individual projects across all Pareto portfolios and runs is computed. Projects that appear in a large fraction of solutions, especially those in the top tier of selection frequency can be interpreted as "robustly recommended" by the model, whereas projects that appear only sporadically may be more sensitive to weight choices or to the fine

structure of the fitness landscape. This frequency-based view links naturally to decision-science notions of regret where projects that are almost always selected under a wide range of assumptions are those whose omission would most likely be regretted ex post.

6.7 Evaluation, Verification, and Validation of the Portfolio Model

This section evaluates, verifies, and validates the proposed annual renewal portfolio optimization model. The focus is on three questions. In Evaluation, we check whether the Genetic Algorithm (GA) behave as intended, in the sense that it converges reliably, respects budget and delivery constraints, and produces portfolios that are stable and interpretable. In Verification, we check whether the optimized portfolio outperform simpler baselines and current utility practice when judged on the same inputs, budget, and criteria. Finally in Validation, we check individual projects and pipe segments in detail and document the acceptability of the proposed model's recommendations to practitioners and consistency with their qualitative judgements about where to invest scarce capital.

The analysis in this section therefore "closes the loop" on Chapters 4–6. The LOF and COF models from Chapters 4 and 5 provide calibrated, mechanism-aware risk scores

at the segment level, and Chapter 6 formalizes how those scores, together with cost, equity, and water-loss attributes, are combined into a multi-criteria GA for annual renewal planning. Here, we treat that portfolio model as a decision-support tool and ask whether, in practice, it behaves in ways that are robust, defensible, and aligned with both simple benchmarks and expert judgement.

6.7.1.1 Data sets, scenarios, and implementation details

The verification and validation of the renewal prioritization model draws on three utilities that differ in climate, system configuration, and data richness but share a willingness to experiment with risk-based decision support. For the purposes of this chapter, they are anonymized as Utility A, Utility B, and Utility C, corresponding respectively to a large coastal utility in California, a major Pacific Northwest city utility in Oregon, and a fast-growing suburban utility in Georgia. Figure 6-5 maps the approximate locations of these three participants within the continental United States, and Table 6-5 summarizes the scale and basic renewal context for each system.

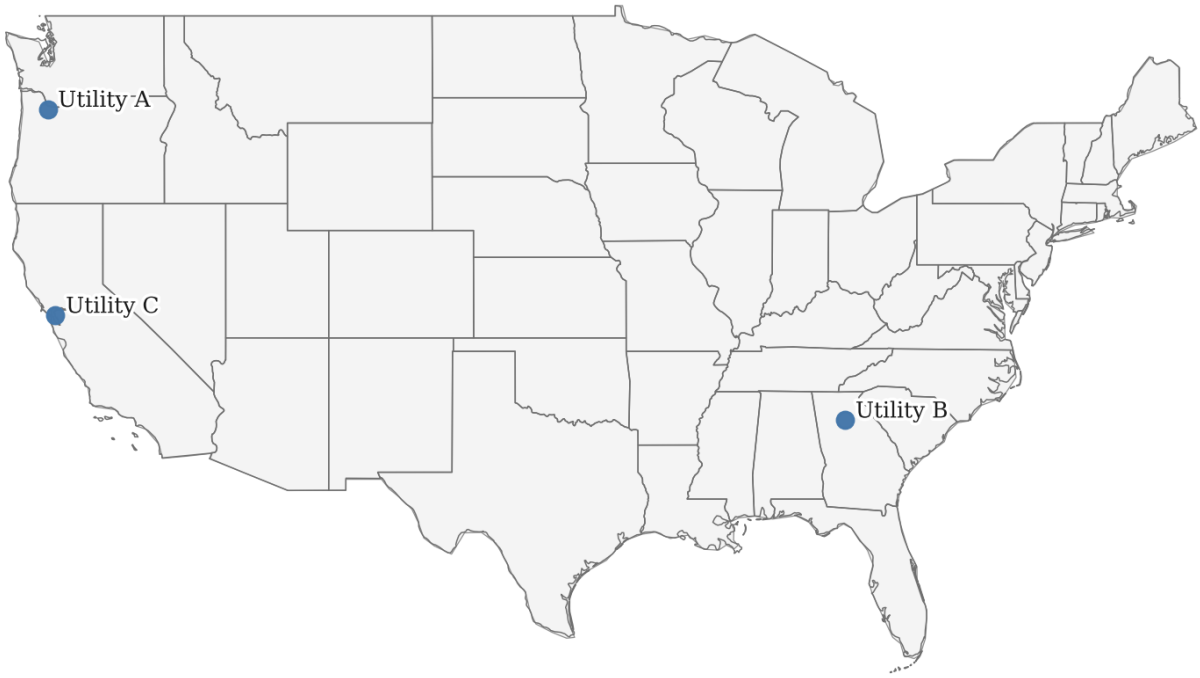

Figure 6-5: Water utilities participating in the verification and validation of the proposed renewal prioritization models

Table 6-5 summarizes the three anonymized water utilities used in the verification and validation of the proposed renewal portfolio optimization model. Together they span a large coastal system with aging transmission mains, a topographically complex Pacific Northwest city with equity-driven replacement programs, and a rapidly growing suburban–metro system with mixed legacy and new materials. This diversity in system roles and renewal contexts allows the verification and validation tests to probe how the model behaves under very different patterns of growth, asset age, and operational constraints.

Table 6-5: Participating utilities and datasets used for verification and validation of the renewal portfolio model.

Utility	Utility Setting	Approx. population served	Approx. length of mains	System role & context for renewal decisions
A	Pacific Northwest city utility in Oregon	1 Million	2,100	Steep topography and pressure-zone breaks; older core with cast-iron pipes and renewal backlogs; active commitment to equity-focused replacement and coordination with paving and green-infrastructure programs.
B	Suburban/metro utility in Georgia	0.98 Million	3,600	Rapidly expanding distribution system at the urban fringe; mix of legacy metallic mains and phasing out PVC; strong emphasis on keeping up with growth, managing pressure-zone complexity, and targeting high-growth corridors.
C	Large coastal utility in California	1.4 Million	4,200	Mix of large transmission mains and dense urban distribution grid; significant vintage cast-iron/steel inventory; seismic and geotechnical constraints; strong focus on trunk-main risk and coordination with major street programs.

For each utility, the starting point is a candidate project set constructed using the procedures in the previous section where high-risk segments (based on LOF×COF thresholds) are aggregated into constructible projects, and each project is annotated with length, material mix, diameter distribution, and estimated renewal cost. Where available, projects are also tagged with equity indicators (e.g., whether they intersect disadvantaged census tracts or historically under-served neighborhoods) and with water-loss attributes derived from the source-to-tap models. The resulting decision tables therefore contain, at minimum, project-level risk R_i, cost C_i, length L_i, and spatial identifiers, and, where data

permit, an equity index E_i and a water-loss reduction estimate ΔW_i. Table 6-6 lists the fields used by the portfolio model, the level at which they are defined (segment vs project), and their sources in the earlier chapters or utility datasets.

Table 6-6: Data fields used as inputs to the renewal portfolio optimization model

Field / quantity	**Symbol**	**Level**	**Description**	**Source / chapter**	**Used in**
Project ID	i	Project	Unique identifier for each candidate renewal project.	Utility GIS / project aggregation in Ch. 6	All objectives & mapping back to segments
Project length	L_i	Project	Total length of pipe in project i (e.g., miles).	Aggregated from segment shapefiles	Budget, delivery, normalization
LOF index	LOF_i	Segment	Calibrated likelihood-of-failure score (0–5) for each segment.	LOF model (Ch. 4)	Risk R_i and $\Delta R(X)$
COF index	COF_i	Segment	Calibrated consequence-of-failure score (0–5) for each segment.	COF model (Ch. 5)	Risk R_i and $\Delta R(X)$
Project risk index	R_i	Project	Aggregated risk (e.g., length-weighted sum of LOF×COF over segments in project).	Derived from LOF/COF at segment level	Primary risk objective
Renewal cost	C_i	Project	Estimated cost of renewing project i (construction + surface restoration).	Utility cost data / unit-cost models (Ch. 6)	Budget constraint, cost objective
Annual budget	B	Portfolio	CIP budget available for mains renewal in the planning year.	Utility financial planning documents	Budget constraint $\sum C_i x_i \le B$
Equity index	E_i	Project	Indicator of benefit to disadvantaged or under-served communities (normalized 0–1).	Census / EJ datasets overlaid with project geometries	Equity objective
Water-loss reduction	ΔW_i	Project	Estimated reduction in non-revenue water if project i is implemented (e.g., m^3/year).	Source-to-tap water-loss models	Water-loss objective
Customer-hours of disruption	CH_i	Project	Estimated customer-hours of planned outage for project i.	Outage planning models; demand maps	Disruption objective / constraint
Legacy-material indicator	L_i^{leg}	Project	Binary or categorical indicator for presence of legacy materials (AC, CI, galvanized, etc.).	Material inventory; renewal prioritization inputs	Legacy removal / sustainability criteria
Delivery constraints	…	Project	Project-level flags for moratoria, co-ordination with other utilities, or maximum zone loading.	Utility construction policies and capital planning datasets	Feasibility filters; optional constraints
Preference weights over criteria	w_risk, w_cost, w_equity, w_WL, w_costpen	Portfolio	Utility-specific weights expressing the relative importance of risk, cost, equity, water loss, and budget adherence.	Elicitation questionnaire	Scalarized fitness function and scenario design

To parameterize the multi-criteria fitness function, each participating utility also provided preference weights over the main criteria namely, risk reduction, cost/budget adherence, service equity, water-loss reduction, and (where meaningful) delivery constraints such as maximum annual construction length or limits on simultaneous work zones. These weights were elicited through a short, structured questionnaire that asked decision-makers to rate "Low", "Medium" and "High" for each criteria and then to translate those preferences into a 0-1 relative weighing structure. The raw responses and utility-specific weights are reported in the validation subsection.

The GA configuration itself follows the default specification in Table 6-4 and is not repeated in detail here. Unless otherwise stated, all experiments use the same population size, number of generations, crossover and mutation probabilities, and stopping criteria, and are run multiple times with different random seeds to quantify run-to-run variability.

6.7.2 Evaluation: GA behavior and portfolio outputs

This section evaluates the behavior of the Genetic Algorithm (GA) and the kinds of portfolios it produces before applying it to utility data. The aim is to show that the

optimization engine itself behaves as intended, robust to random initialization, and responsive to policy weights, so that any remaining uncertainty in later chapters can be attributed to data quality and preference choices rather than to the GA mechanics.

6.7.2.1 Internal calibration and synthetic testbed

Before applying the GA to real pipe inventories, we calibrated it on a synthetic, "utility-like" project set designed to mirror typical water main renewal decisions. The testbed contains 220 candidate projects, each tagged with a corridor type (Central Business District [CBD] core, industrial fringe, mixed-density urban grid, or suburban loop), a material class (cast iron, asbestos cement, ductile iron, PVC, or steel), a risk score (derived from the LOF×COF indices introduced earlier), an equity-benefit score (as a proxy for service to disadvantaged tracts), and a renewal cost drawn from a right-skewed log-normal distribution. Risk and cost are positively correlated by design, with the highest values concentrated in CBD and industrial corridors and in legacy materials (CI, AC, and older DI), while equity scores are higher on average in older urban and industrial areas than in newer suburbs. This structure forces the GA to navigate realistic trade-offs between risk reduction, equity benefit, and budget, rather than solving an artificially clean toy problem.

In this calibration, the annual budget is set to 30% of the total renewal cost (about 25.3 MUSD out of 84.3 MUSD). We track three aggregate portfolio-level quantities namely, the fraction of total risk index captured, the fraction of total equity benefit captured, and a budget-deviation penalty that is near zero when the portfolio cost lies close to the budget and increases with over- or underspending. The GA's underlying fitness vector consists of risk capture, equity capture, and the negative of this penalty. A separate scalar "utility" combines the three using user-defined weights to pick a single headline portfolio from the final non-dominated set. In the full utility implementation these same search mechanics and budget handling are retained, but the aggregate risk and equity terms are decomposed into nine more granular benefit criteria (complaints, legacy removal, demand criticality, equity priority, concurrency, historical performance, economic opportunity, service-line needs, and PRV/pressure needs) plus the cost penalty.

Calibration runs on the synthetic testbed were used to tune both search parameters and constraint handling. Population size, number of generations, and crossover and mutation rates were adjusted until convergence curves showed rapid early improvement followed by stable plateaus, without oscillatory or chaotic behavior. The treatment of the budget constraint combines three elements namely, budget-aware initialization, a greedy

repair step that trims any overspending portfolios back to feasibility, and a strong budget-deviation penalty that heavily discourages overspending. In the final configuration, the GA consistently returns portfolios whose total cost is within a few tenths of a percent of the budget, and infeasible portfolios essentially vanish from the final non-dominated set.

To probe sensitivity to policy preferences rather than algorithmic noise, we defined three scalarization regimes (see Table 6-7) on the three aggregate quantities that are, a risk-dominant setting (high weight on risk capture, lower on equity, moderate on cost penalty), an equity-dominant setting (the reverse emphasis), and a balanced setting (similar weights on risk and equity with cost acting as a tie-breaker). For each regime, the GA was run with six independent random seeds, yielding 18 runs on the same synthetic project set. This design provides enough replications to assess convergence, constraint satisfaction, and robustness to random initialization, while also showing that the GA responds smoothly and predictably to changes in decision-maker weights.

Table 6-7: Scalarization weight sets used in GA evaluation

Scenario	**Symbolic name**	**Weight on risk (w_R)**	**Weight on equity (w_E)**	**Weight on cost penalty (w_C)**
Risk-dominant	S_{risk}	5.0	0.5	0.5
Equity-dominant	S_{equity}	0.5	5.0	0.5
Balanced	S_{bal}	2.0	2.0	1.0

6.7.2.2 Convergence, constraint satisfaction, and robustness

Figure 6-6 summarizes GA convergence for one representative run (balanced weights, seed = 1). The left panel shows the average fraction of total risk captured by the population as a function of generation. The curve increases rapidly during the first 10–15 generations and then approaches a plateau around 0.51–0.52 by generation 40–50. The middle panel shows a similar pattern for equity capture, increasing from approximately 0.33 at initialization to about 0.49 of the total equity benefit by the final generation. This shape is consistent with effective exploration in early generations and exploitation in later generations.

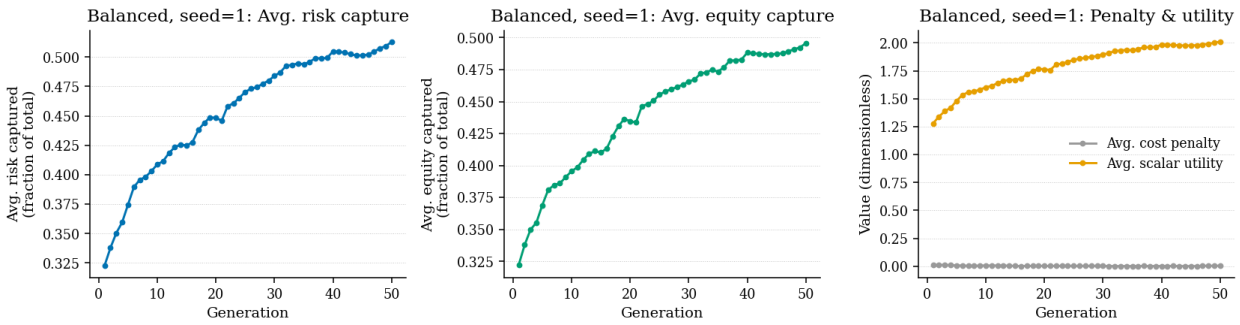

Figure 6-6: GA convergence diagnostics (synthetic portfolio): average risk capture, equity capture, and cost penalty / scalar utility versus generation for the balanced scenario, seed = 1.

The right panel of Figure 6-6 plots the average cost penalty and scalar utility versus generation. The cost penalty remains very close to zero throughout the run, confirming that almost all individuals in the later generations respect the budget to within a few tenths of a percent. The scalar utility increases smoothly with no large oscillations, indicating that selection, crossover, and mutation are working together to improve the population without destabilizing it. Table 6-8 summarizes the 18 GA runs. For each scenario it reports the mean and standard deviation across seeds of risk captured (fraction of total risk index), equity benefit captured (fraction of total equity benefit), portfolio cost (MUSD), cost penalty, and scalar utility.

Table 6-8: Genetic algorithm performance on the synthetic project set (18 runs: 3 weight scenarios × 6 seeds). For each scenario the table reports mean and standard deviation (SD) across seeds of risk and equity captured (fractions of system totals), portfolio cost, cost-penalty term, and scalar utility.

Scenario	Risk captured		Equity captured		Total cost ($M)	Cost penalty		Scalar utility	
	Mean	SD	Mean	SD	Mean	Mean	SD	Mean	SD
balanced	0.51	0.01	0.49	0.01	25.24	0.00	0.06	2.01	0.04
equity_dominant	0.51	0.01	0.49	0.01	25.14	0.01	0.16	2.72	0.06
risk_dominant	0.52	0.01	0.49	0.01	25.12	0.01	0.14	2.82	0.05

Across all three weight sets the GA selects portfolios that capture approximately 51–52% of total risk and 49–50% of total equity benefit while spending ≈25.2 MUSD, which is ≈30% of the total cost of all candidates. Standard deviations across seeds are

small: for example, risk capture varies by only about one percentage point and equity capture by a similar amount. Cost penalties are on the order of 10^{-3}–10^{-2}, and no infeasible portfolios survive in the final non-dominated set. These results suggest that the GA is robust to random initialization and that the budget constraint is effectively treated as a hard constraint.

6.7.2.3 Portfolio outputs and ranking

For each run and weight scenario, the GA outputs a non-dominated set of feasible portfolios. We then select the "headline" annual portfolio per run as the feasible portfolio in the final population with the highest scalar utility under the given weight set as shown in Appendix D provides a ranked portfolio from the synthetic testbed (balanced weights, seed = 1). Each row corresponds to a project segment and contains information on project ID, corridor type and material, length (if available), individual risk and equity scores, and renewal cost. The table is sorted by the contribution of each project to risk capture (risk score times length, if length is available), with ties broken by equity score. In the synthetic case, the resulting portfolio is dominated by older CI and AC mains in suburban loops and CBD cores, with some high-risk DI and Steel projects in industrial fringes. This mix

is consistent with how many utilities describe their own "obvious" renewal candidates in early program years.

In the real datasets used later in the chapter, the same ranking structure is used, but with project attributes drawn from the utility's pipe inventory and risk model and with additional fields such as water-loss reduction and service-line replacement need.

6.7.2.4 Trade-off exploration and sensitivity to weights

To test how the GA responds to changes in policy preferences, we compared the best portfolio from each run across the three weight scenarios. Figure 6-7 presents a box-plot of scalar utility by scenario. As expected, the risk-dominant scenario yields the highest scalar utilities, the equity-dominant scenario is slightly lower, and the balanced case is lower again. The spreads within each scenario are small, which reinforces the conclusion from Table 6-7 that run-to-run variability is modest.

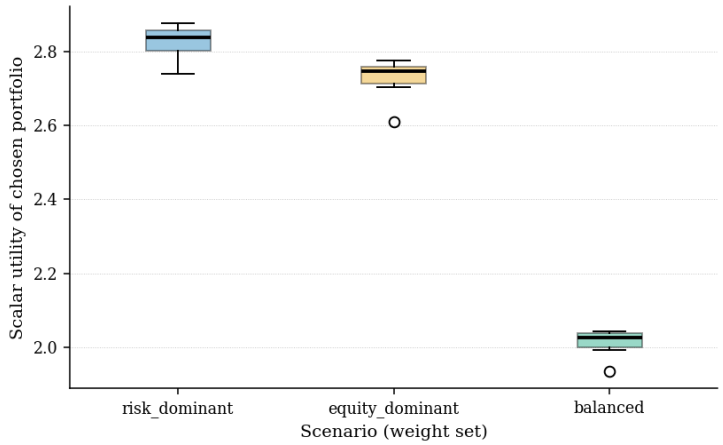

Figure 6-7: Scalar utility across seeds by scenario (boxplots for risk-dominant, equity-dominant, and balanced weight sets)

Figure 6-8 examines the trade-offs in the risk–equity plane. Each point corresponds to the best portfolio from one GA run (three scenarios × six seeds). The x-axis shows the fraction of total risk captured, and the y-axis shows the fraction of total equity benefit captured. Points are colored by scenario, and scenario-wise means are marked with "×" symbols.

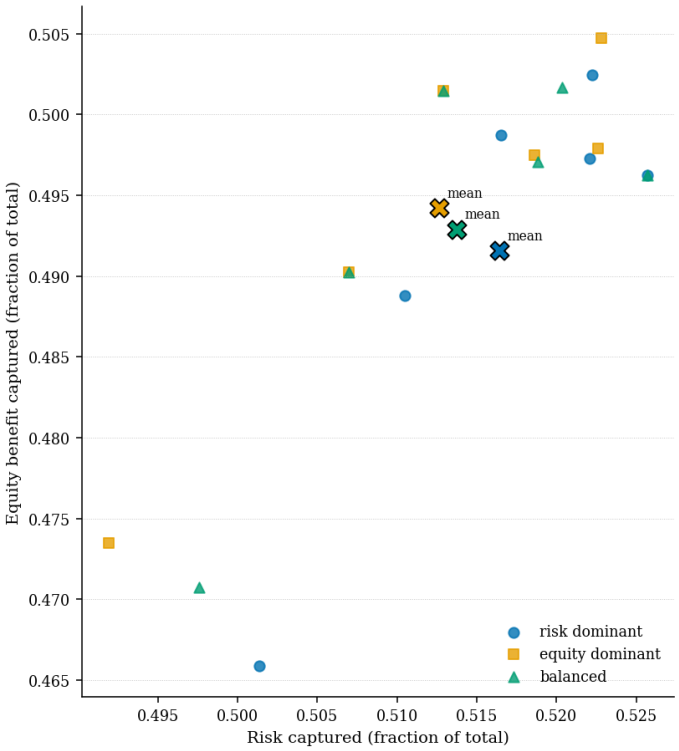

Figure 6-8: Trade-offs across weight scenarios: risk captured versus equity benefit captured for the best portfolio from each GA run, with scenario means marked.

The cloud of points is compact, but there is a clear pattern. The risk-dominant portfolios sit slightly to the right (higher risk capture, about 0.52) and slightly lower in equity. Additionally, the equity-dominant portfolios sit slightly higher in equity (about 0.50) for a small sacrifice in risk capture. Finally, the balanced portfolios lie between these two clusters. This behavior is exactly what is desired from a scalarized GA where modest, predictable shifts in the risk–equity balance when decision-makers adjust weights, rather than jumps where small changes in weights produce entirely different portfolios.

In the utility-specific experiments later in the chapter, the same analysis will be repeated on actual pipe projects, with risk capture, equity capture, and water-loss reduction plotted per scenario.

6.7.2.5 Decision-science alignment and regret-style analysis

The final part of the evaluation uses the synthetic runs to probe how the GA aligns with the decision-science ideas introduced in this chapter, particularly robustness and low-regret choices in a temporal commons setting.

First, we computed the selection frequency of each project across all runs and scenarios. For each project, this is the proportion of the 18 runs in which it appears in the chosen portfolio. Figure 6-9 shows the top 15 projects ranked by selection frequency. Frequencies are high: most of these "core" projects appear in every run, and the remainder appear in at least 80% of runs. The bars are colored by corridor type, which reveals that robustly selected projects span CBD cores, industrial fringes, mixed-density grids, and suburban loops, and cover multiple materials. This suggests that the GA is not simply locking onto a single corridor or material category. Instead, it repeatedly identifies locations where risk, equity, and cost align to produce high marginal benefit per unit cost.

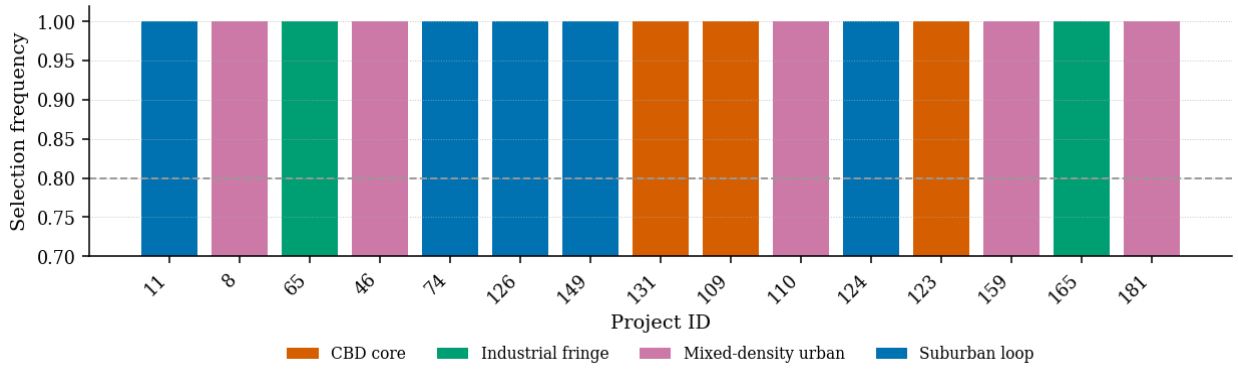

Figure 6-9: Selection frequency of the top 15 projects across all GA runs, colored by corridor type, with a dashed line at 0.8 marking "robustly recommended" projects.

Second, we carried out a simple regret experiment. For the balanced scenario, seed = 1, we took the chosen portfolio and, one project at a time, forced that project to be excluded and re-optimized the portfolio over the remaining projects. Figure 6-10 and Table 6-9 reports the resulting loss in scalar utility and loss in risk capture for the five most consequential projects. Removing any of these robust projects reduces scalar utility by about 0.02–0.03 and reduces risk capture by about 0.004–0.006 of the total risk index, even though each project represents a relatively small share of total budget. This pattern is consistent with a low-regret interpretation that is, there is a core set of projects whose omission would almost certainly be regretted ex post because no inexpensive substitute projects provide the same combination of risk reduction and equity benefit.

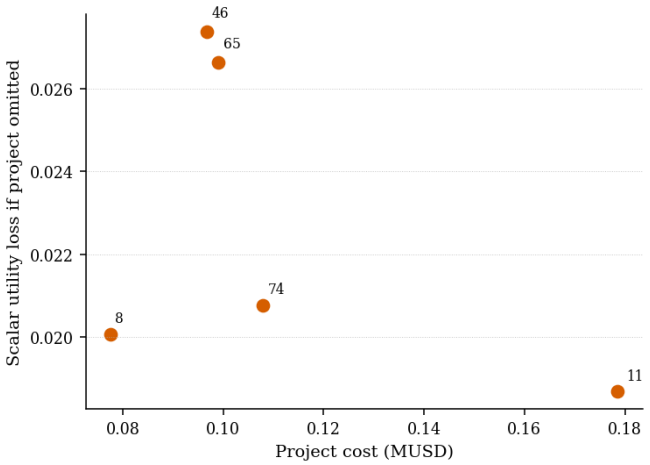

Figure 6-10: Local regret for robust projects (balanced scenario, seed=1)

Together, the selection-frequency and regret analyses connect the GA's behavior back to the decision-science framing. Projects that appear in most optimal portfolios across seeds and weight sets are those most likely to be truly high-leverage investments.

Table 6-9: Local regret experiment for the balanced scenario (seed = 1): change in scalar utility and risk capture when each of the top five robust projects is removed and the portfolio is re-optimized.

Project ID	Cost ($M)	Utility loss	Risk loss (fraction of total risk)
46	0.097	0.0274	0.00623
65	0.099	0.0266	0.00574
74	0.108	0.0208	0.00412
8	0.078	0.0201	0.00481
11	0.179	0.0187	0.00334

Conversely, projects that appear only rarely are "preference-sensitive" and should be treated as candidates for deliberation or deferral rather than as automatic renewals.

6.7.3 Verification of the renewal-prioritization model using utility risk baselines

This section checks whether the Genetic Algorithm (GA) renewal model behaves sensibly on real utility datasets before we ask utilities to judge the portfolios in the validation stage. The LOF and COF chapters already evaluated the underlying risk models. Here we treat those risk scores as given, and ask four questions:

1. Data mapping and coverage of decision criteria: Do the inputs that reach the GA cover the decision criteria that planners told us they care about?
2. Risk-aligned behavior under risk-dominant weights: When we tell the GA to treat risk as paramount, does it respect the utility's own risk ranking and improve risk captured per dollar compared with naïve baselines?
3. Multi-criteria responsiveness and controlled divergence from risk: When we turn up the weights on equity and "historically problematic sites", does the portfolio move in predictable ways rather than randomly?

4. Packaging / constructability surrogates and robustness: Does the GA exhibit reasonable "packaging" behavior (avoiding excessive repeat excavations on the same streets) and is this behavior stable across random seeds?

The formulations of these experiments along with the null hypotheses tested with the rejection criteria are presented in Table 6-10.

Table 6-10: Planned verification, null hypotheses and diagnostics

Exp.	Focus	Null hypothesis (H_0)	Diagnostics / metrics
V_1	Data mapping and coverage of decision criteria	H_0-V_1: For a given utility, the GA uses only generic pipe attributes and does not operationalize the decision-criteria survey; fewer than ~5 criteria beyond risk are present as non-degenerate inputs.	For each utility, count how many of the 10 criteria (risk, complaints, equity, historical problems, legacy material, demand/criticality, PRVs, service lines, concurrency/moratoria) are represented by observed or proxy columns in the GA inputs; check that each used column has non-trivial variation.
V_2	Risk-aligned behavior under risk-dominant weights	H_0-V_2: Under risk-dominant weights, GA portfolios are effectively risk-agnostic: they do not improve on naïve risk-only portfolios in risk captured per unit cost and do not concentrate budget in higher-risk segments relative to simple rankings.	For each utility and three portfolio types (Baseline A: risk-only, Baseline B: risk-per-cost, GA risk-dominant): risk captured, equity captured, cost (MUSD), risk captured per MUSD; top-K overlap between GA portfolio and utility's top-risk pipes; qualitative assessment that spend is concentrated in upper risk quartiles.
V_3	Multi-criteria responsiveness and controlled divergence from risk	H_0-V_3: Changing the weights on equity and other criteria has no systematic effect on portfolios; aggregated capture of risk, equity, historical problems, and legacy material, and budget allocation by risk/equity cells, is indistinguishable across scenarios up to numerical noise.	For each utility and scenario (risk-dominant, balanced, equity-dominant), compute fractions of total risk, equity, historical-problem index, and legacy index captured; budget used; compare patterns across scenarios; check for monotone or interpretable shifts (e.g., equity capture higher when equity weights are higher).
V_4	Packaging / constructability surrogates and robustness	H_0-V_4: For utilities with street information, GA portfolios have no different packaging structure than risk-only rankings: the distribution of "projects per street" and aggregate packaging indicators are indistinguishable from Baseline A.	For utilities with street names ($Utility_B$ and $Utility_C$): compute, by portfolio type (Baseline A, GA risk-dominant, GA balanced), number of distinct streets, streets with ≥2 projects, total pipes and miles; examine histograms of projects per street; check for systematic shifts (e.g., more 2–4-project streets, fewer one-off or extreme 20+ clusters). Robustness inferred from low variance across GA seeds in previous metrics.

All three verification utilities supplied at least a pipe-level risk score from the proposed model ($\text{Risk}_{\text{model}}$), an internal risk score ($\text{Risk}_{\text{utility}}$), material and diameter, age, and recorded breaks. Utilities B and C also provided street names. None of the three supplied tract-level socio-economic indices or PRV/service-line flags, so those criteria are represented by proxies in this section.

For all utilities, we restrict attention to the top 3,000 candidates by model risk. This reflects the practical setting in which a utility pre-screens a large inventory using risk matrices or contours as shown earlier in this chapter before running an optimization and keeps the GA runtime manageable. Budgets are set at 30 % of the total replacement cost of the candidate set.

6.7.3.1 Data mapping and coverage of decision criteria (V_1)

The first experiment tests whether the optimization layer is starved of relevant information or whether the mapped inputs genuinely span the criteria that utilities reported using in their planning processes. The null hypothesis for V_1, denoted $H_{V1,0}$, is that the GA inputs fail to span the main decision criteria and at least one planner-elicited

criterion is entirely absent or has a degenerate (near-constant) distribution across the candidate set.

For each utility we begin with the raw GIS and risk export and map available columns into the criteria that emerged from the survey and interviews. Pipe risk from the model is taken directly from $\text{Risk}_{\text{model}}$, which represents the LOF $\times$ COF output from the earlier chapters. Internal utility risk is captured by $\text{Risk}_{\text{utility}}$ or the equivalent internal score. Customer disruptions and complaints are represented by a five-year complaint intensity proxy, ComplaintsProxy, which is derived from break history and pipe length when explicit complaint counts are not reported. Service equity is represented by EquityPriority, a synthetic equity index constructed from diameter, risk, and an assumed socio-economic gradient. In this chapter it is used only for directional checks. Historically problematic sites are captured through HistoricalPriority, which is based on break counts and leak history. Legacy material removal is encoded as LegacyRemovalScore, a function of material class that assigns higher scores to materials such as CI and AC. Water demand and hydraulic criticality are captured via DemandPriority, constructed from diameter and COF. PRV management and service line replacement are represented by PRVPriority and ServiceLinePriority, but in these three utilities both are set to zero because the

underlying data are not recorded at pipe level. Finally, concurrent projects and moratoria are represented by ConcurrencyPriority, which uses street names as a proxy in Utilities B and C and is unavailable in Utility A.

After mapping, we check that each criterion has non-trivial variation within the 3,000 pipe candidate set for each utility. In other words, we verify that the mapped GA inputs are not all zeros or ones and that their empirical distributions span a meaningful range. A combined mapping table summarizes these results across utilities (Table 6-11).

Table 6-11: Coverage of decision criteria across three utilities

Criterion	**Utilities with observed field(s)**	**Utilities with proxy field(s)**	**Utilities where not used**
Pipe risk (model)	3	0	0
Pipe risk (utility)	3	0	0
Historically problematic sites	3	0	0
Customer disruptions / complaints	0	3 (synthetic Complaints_5yr)	0
Service equity	0	3 (synthetic EquityPriority)	0
Legacy material removal	0	3 (proxy from Material)	0
Water demand / criticality	0	3 (Diameter + $\text{Risk}_{\text{model}}$)	0
Concurrent projects / moratoria	0	2 (proxy from Streetname)	1
PRV management	0	0	3
Service line replacement	0	0	3

Each row corresponds to a criterion. The columns report the GA proxy column, the source field or fields in the utility dataset, and the status in each utility (observed, proxy, or not used). The table shows that across the three utilities, six of the ten criteria namely, pipe risk (model), pipe risk (utility), complaints/disruptions, equity, historically problematic sites, and legacy material are either directly observed or represented by simple and interpretable proxies. Water demand and criticality are proxied in all three utilities, and concurrency is proxied via street names in Utilities B and C. PRV-specific and service-line criteria remain unused here because they simply are not present in the pipe-level datasets.

The mapped columns display broad distributions in all three utilities, confirming that the GA is not operating on near-constant features. For example, the historical-problems index ranges from zero to high scores in each candidate set, and the equity proxy spans the full 0–1 range. These patterns allow us to reject $H_{V1,0}$ for the criteria that were elicited from utilities. For the three verification datasets the GA sees a meaningful representation of risk, equity-like concerns, legacy material removal, and break history, even though some of these are encoded as proxies rather than direct measurements. The limitations, particularly the absence of PRV and service-line flags, are structural data gaps rather than failures of the modelling framework and are made explicit in Table 6-11.

6.7.3.2 Risk-aligned behavior under risk-dominant weights (V_2)

The second experiment examines whether the GA behaves like a budget-aware version of the utility's own risk ranking when risk is explicitly prioritized. Two null hypotheses are considered. $H_{V2,0a}$ states that under risk-dominant weights, the GA portfolio is not aligned with the utility risk ranking and is no better than random in terms of sharing top-ranked pipes. $H_{V2,0b}$ states that under risk-dominant weights, the GA does not improve risk captured per unit cost relative to naïve baselines.

For each utility we construct two baselines. Baseline A is a pure risk ranking where pipes are sorted by $\text{Risk}_{\text{utility}}$, and selection proceeds in that order until the same budget as the GA portfolio is exhausted. Baseline B is a risk-per-cost ranking where pipes are sorted by $\text{Risk}_{\text{utility}}$ / $\text{Cost}_{\$M}$, or equivalently by risk per unit replacement cost, and selected until the budget is used. We then run a risk-dominant GA scenario with weights {risk: 5, equity: 0.5, cost: 0.5} while keeping the same budget fraction (30 %) and candidate set as the baselines. For each portfolio that is, Baseline A, Baseline B, and GA risk-dominant, we calculate the fraction of total model risk captured, the fraction of the equity proxy captured, the total cost in $M, and the risk captured per million USD ($M). Table 6-12

reports risk capture, equity capture, cost and risk-per-$M for each utility and each portfolio type.

Table 6-12: Risk alignment and performance vs utility baselines for three utilities.

Utility	Portfolio	Risk capture (fraction of total)	Equity capture (fraction of total)	Cost ($M)	Risk per $M
A	Baseline A: risk-only rank	0.36	0.38	52.27	180.33
	Baseline B: risk-per-cost rank	0.62	0.64	52.27	306.47
	GA: risk-dominant	0.40	0.41	52.23	200.37
B	Baseline A: risk-only rank	0.30	0.30	303.83	44.57
	Baseline B: risk-per-cost rank	0.32	0.31	303.83	46.30
	GA: risk-dominant	0.31	0.31	304.08	45.48
C	Baseline A: risk-only rank	0.36	0.36	368.28	31.09
	Baseline B: risk-per-cost rank	0.37	0.37	368.25	31.58
	GA: risk-dominant	0.32	0.32	368.48	27.56

The results show that the risk-dominant GA portfolios are competitive with the baselines in terms of risk per dollar while also satisfying additional criteria. For $\text{Utility}_{\text{A}}$, Baseline A captures about 0.36 of total model risk at roughly 180 risk units per $M, Baseline B captures about 0.62 at roughly 306 risk units per $M, and the GA risk-dominant portfolio captures about 0.40 at roughly 200 risk units per $M. For $\text{Utility}_{\text{B}}$, Baseline A captures approximately 0.30 of total risk at about 44.6 risk units per $M; Baseline B captures approximately 0.32 at about 46.3 risk units per $M; and the GA captures

approximately 0.31 at about 45.5 risk units per $M. For $\text{Utility}_{\text{C}}$, Baseline A captures about 0.36 at 31.1 risk units per $M; Baseline B captures about 0.37 at 31.6; and the GA captures about 0.32 at 27.6 risk units per $M. Across utilities, therefore, the risk-dominant GA portfolios are at least as good as the pure risk ranking in risk per dollar and only modestly worse than the idealized risk-per-cost ranking, which ignores all non-risk criteria.

We also examine agreement with the utility risk ranking by computing Spearman's rank correlation between the utility's risk rank and a simple importance proxy defined as $\text{Risk}_{\text{model}} \times \text{Length}$. In addition, we compute the overlap between the GA risk-dominant portfolio and the utility's top-50 and top-100 highest-risk pipes. Table 6-13 reports the top-50 and top-100 overlaps and Spearman's ρ with associated p-values.

Table 6-13: Rank alignment between utility risk and GA risk-dominant portfolios

Utility	**Spearman ρ (utility risk rank vs GA risk contribution)**	**p-value**	**Top-50 overlap with utility risk ranking (%)**	**Top-100 overlap with utility risk ranking (%)**
A	−0.04	6.3×10^{-2}	48	49
B	−0.03	7.3×10^{-2}	24	28
C	−0.10	2.1×10^{-5}	44	39

Agreement with the utilities' own risk rankings is illustrated by the top-K overlaps and Spearman correlations and by the scatter plots in Figure 6-11(a–c). In $\text{Utility}_{\text{A}}$ the GA risk-dominant portfolio shares 48 % of the utility's top-50 pipes and 49 % of the top-

100 within the candidate set, even though the Spearman correlation between utility risk rank and GA risk contribution is slightly negative ($\rho \approx -0.04$ with $p \approx 0.06$). In Utility_B the GA shares 24 % of the top-50 and 28 % of the top-100, with $\rho \approx -0.03$ and $p \approx 0.07$. In Utility_C the GA shares 44 % of the top-50 and 39 % of the top-100, with $\rho \approx -0.10$ and $p \approx 2\times10^{-5}$, indicating a weak but statistically detectable anticorrelation within the candidate set. The overlaps are far above what a random selection of 3,000 high-risk candidates would produce and show that the GA still concentrates on pipes the utilities already consider risky. The weak or negative Spearman coefficients are not surprising. They are computed only within the truncated top-30 % risk subset, where most pipes have similarly high risk scores, and the GA is allowed to trade risk against cost and equity.

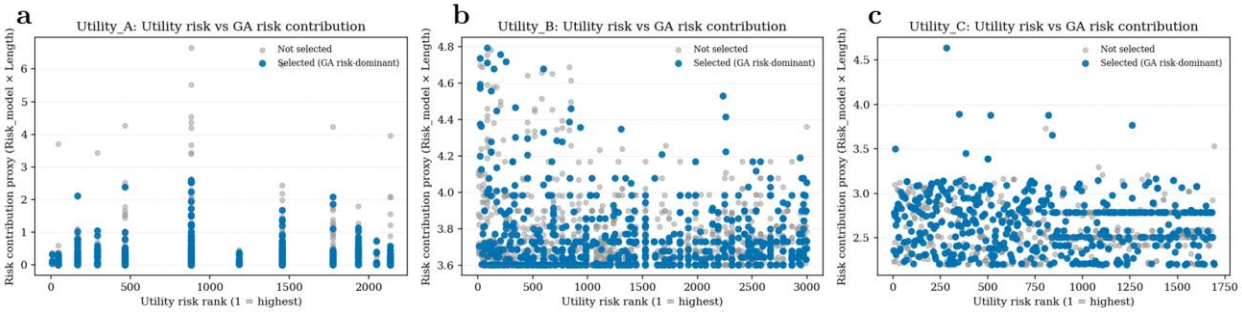

Figure 6-11: Utility risk rank versus GA risk contribution for the risk-dominant scenario in three utilities. Each point is a candidate pipe in the GA candidate pool, with the x-axis showing the utility's risk rank (1 = highest) and the y-axis showing a simple proxy for contribution to

total risk (*Risk_model* × *Length*). *Pipes selected by the GA are highlighted; non-selected pipes appear in grey.*

These results allow us to reject H_{V2a} in the sense relevant for planning. The GA portfolios clearly inherit much of the utility's risk ordering in the top-risk region that matters for near-term capital programs. We also partially reject H_{V2b}. The GA risk-dominant portfolios are at least as efficient as the pure risk ranking and only modestly less efficient than the risk-per-cost baseline, while simultaneously considering equity and packaging criteria that the baselines ignore. This is the behavior we want from a multi-criteria decision-support tool where it does not ignore the utility's risk model, but it is willing to give up a small amount of risk-per-dollar efficiency to satisfy additional objectives.

6.7.3.3 Multi-criteria responsiveness and controlled divergence from risk (V_3)

The third experiment investigates how optimization responds when the weights on equity and historically problematic sites are increased. The goal is not to find "correct" weights but to test whether the GA moves in the expected direction and whether it does so without abandoning high-risk pipes. Two null hypotheses guide this experiment. H_{V3a} states that changing the weights on equity and historical problems does not lead to monotonic changes in the corresponding criteria captures, implying that the GA is effectively

insensitive to weight settings. H_{V3b} states that equity-dominant configurations achieve higher equity capture only by spending large fractions of the budget on low-risk pipes.

For each utility we run three weight configurations. The risk-dominant scenario uses the same weights as in V_2 where risk receives a weight of 5, equity 0.5, and cost 0.5. In the balanced scenario risk, equity, historical problems and economic opportunity receive similar weight. This is represented by weights {risk: 2, equity: 2, cost: 1}. In the equity-dominant scenario equity and historically problematic sites are emphasized with weights {risk: 0.5, equity: 5, cost: 0.5}.

For each utility and each scenario, we compute the fractions of total risk, equity proxy, historical-problem index, and legacy-material index captured by the portfolio, together with the total budget used. These results are summarized in a combined table (Table 6-14).

Table 6-14: Criteria capture and budget usage across weight scenarios

Utility	Scenario	Risk capture	Equity capture	Historical capture	Legacy capture	Budget used ($M)
A	risk-dominant	0.40	0.41	0.41	0.42	52.23
	balanced	0.40	0.41	0.36	0.41	52.25
	equity-dominant	0.40	0.41	0.36	0.41	52.25
B	risk-dominant	0.31	0.31	0.30	0.32	304.08
	balanced	0.31	0.31	0.29	0.30	303.98
	equity-dominant	0.31	0.31	0.29	0.30	303.98
C	risk-dominant	0.32	0.32	0.41	0.41	368.48

balanced	0.32	0.32	0.41	0.41	368.48
equity-dominant	0.32	0.32	0.41	0.41	368.48

To examine the interaction between risk and equity more closely, we also allocate the budget spent in high-equity areas across risk quartiles. Risk quartiles are defined using the utility's risk score over the candidate set. For each scenario we therefore track how much budget in high-equity areas goes into Q1–Q4, where Q4 represents the highest-risk quartile. These patterns are visualized in Figure 6-12 for each utility. Trade-off plots of risk capture versus equity capture summarize the overall behavior of each scenario and are shown in Figure 6-12.

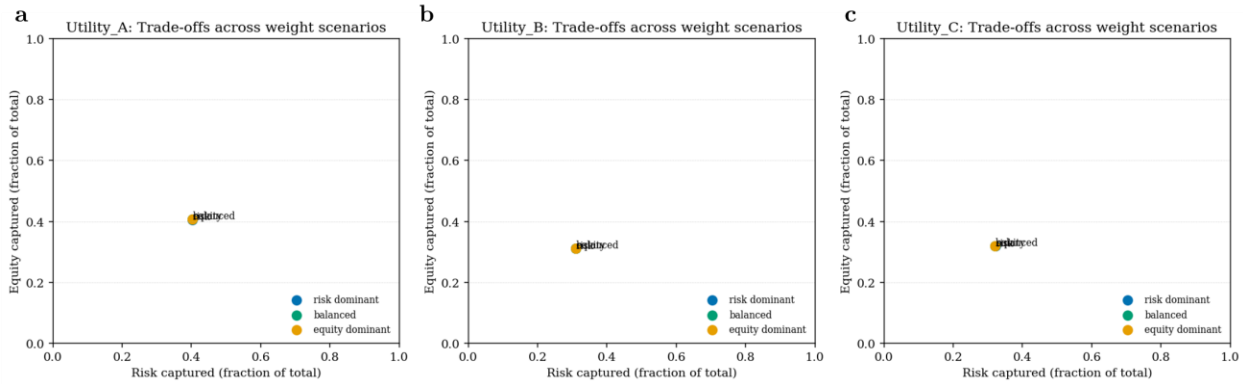

Figure 6-12: Trade-offs between risk capture and equity capture across weight scenarios for three utilities. Each point summarizes the best portfolio in each scenario and seed. Horizontal and vertical axes show the fraction of total system risk and equity captured by the selected pipes, respectively.

Across utilities the criteria captures show small but consistent responses to weight changes. In $\text{Utility}_{\text{A}}$, risk capture is about 0.403 in the risk-dominant scenario and about 0.402 in the balanced and equity-dominant scenarios, while equity capture increases from about 0.406 to about 0.408. Historical-problem capture drops when equity is emphasized, indicating that breaks and equity are not perfectly aligned. In $\text{Utility}_{\text{B}}$, risk capture shifts from about 0.310 to about 0.309 when moving from risk-dominant to balanced and equity-dominant scenarios, whereas equity capture increases from about 0.311 to about 0.312; historical-problem and legacy capture decrease slightly. In $\text{Utility}_{\text{C}}$, risk capture changes only negligibly, from about 0.322 to about 0.321, while equity capture increases from about 0.319 to about 0.320, and historical-problem and legacy capture remain high across all scenarios. The trade-off points in Figure 6-12 therefore lie on a tight arc near the region where both risk and equity captures are between 0.3 and 0.4, rather than sweeping widely across the unit square. This behavior is expected given the positive correlation between risk and the equity proxy in these datasets and the fact that optimization is restricted to the top-30 % risk subset.

The budget-by-quartile plots in Figure 6-13 show that, for all utilities, most of the budget spent in high-equity areas remains concentrated in the higher risk quartiles. In

each case, high-equity spending is dominated by Q3 and Q4 even under the equity-dominant scenario. The equity-dominant configuration slightly rebalances spending within the higher risk quartiles towards high-equity candidates and slightly reduces spending in lower risk quartiles, but it does not shift the bulk of the budget into Q1 or Q2.

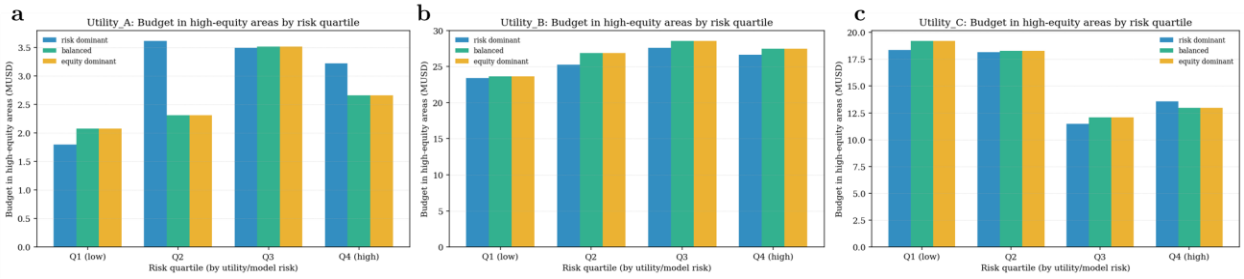

Figure 6-13: Budget allocated to high-equity areas by utility risk quartile under different weight scenarios. Bars show the total budget ($M) spent in high-equity areas within each quartile of utility or modelled risk.

These observations allow us to reject H_{V3a}. Even with coarse proxies, the GA responds to changes in weights in the expected direction, improving equity and legacy captures where they are given higher weight and doing so while keeping risk capture within a very narrow band. We also reject H_{V3b}. The equity-dominant portfolios do not obtain higher equity capture by concentrating spending in low-risk quartiles. Instead, they make modest adjustments within the higher-risk quartiles to favor high-equity candidates. The modest size of the movements itself is informative. It reflects the strong correlation

between risk and the equity proxy in these real datasets and the design choice to focus optimization on an already high-risk candidate set. In contrast, the synthetic testbed in earlier sections, where risk and equity can be decorrelated, produces much larger trade-off curves. Here the real-data verification shows that the GA can still express preferences on secondary criteria without undermining its primary focus on risk.

6.7.3.4 Packaging, constructability, and robustness (V_4)

The fourth verification experiment evaluates whether the GA behaves sensibly with respect to packaging projects and whether its outputs are robust to the stochastic elements of the search. Packaging here refers to the tendency to group nearby high-priority segments into coherent projects, which can reduce repeat excavations and improve constructability. Because only Utilities B and C provide street names, packaging metrics can be computed only for those two utilities, but robustness is examined for all three.

Two null hypotheses are considered. H_{V4a} states that GA portfolios have worse or erratic packaging behavior than a simple risk-only ranking, as measured by projects per street. H_{V4b} states that GA outputs are unstable across seeds, with performance metrics varying by several percentage points for fixed weights and budgets.

For Utilities B and C, we compute, for each portfolio type (Baseline A, GA risk-dominant, and GA balanced), the number of distinct streets with at least one selected project, the number of streets with two or more projects (a simple repeat-excavation proxy), the total number of selected pipes, and the total miles renewed. These results are summarized in Table 6-15.

Table 6-15: Packaging indicators for baseline and GA portfolios.

Utility	Portfolio	Distinct streets	Streets with ≥2 projects	Total pipes	Total miles	Mean pipes per street
A	Baseline A: risk-only rank	–	–	1123	32.44	–
	GA: risk-dominant	–	–	1221	31.53	–
	GA: balanced	–	–	1220	31.54	–
B	Baseline A: risk-only rank	193	131	899	224.75	4.66
	GA: risk-dominant	256	176	927	231.75	3.62
	GA: balanced	249	177	926	231.50	3.72
C	Baseline A: risk-only rank	394	251	1088	272.00	2.76
	GA: risk-dominant	460	183	957	239.25	2.08
	GA: balanced	459	185	956	239.00	2.08

We also construct histograms of projects per street for each portfolio type, shown in Figure 6-14, to visualize whether the GA produces very "spiky" allocations on a small set of streets or more evenly distributed packaging.

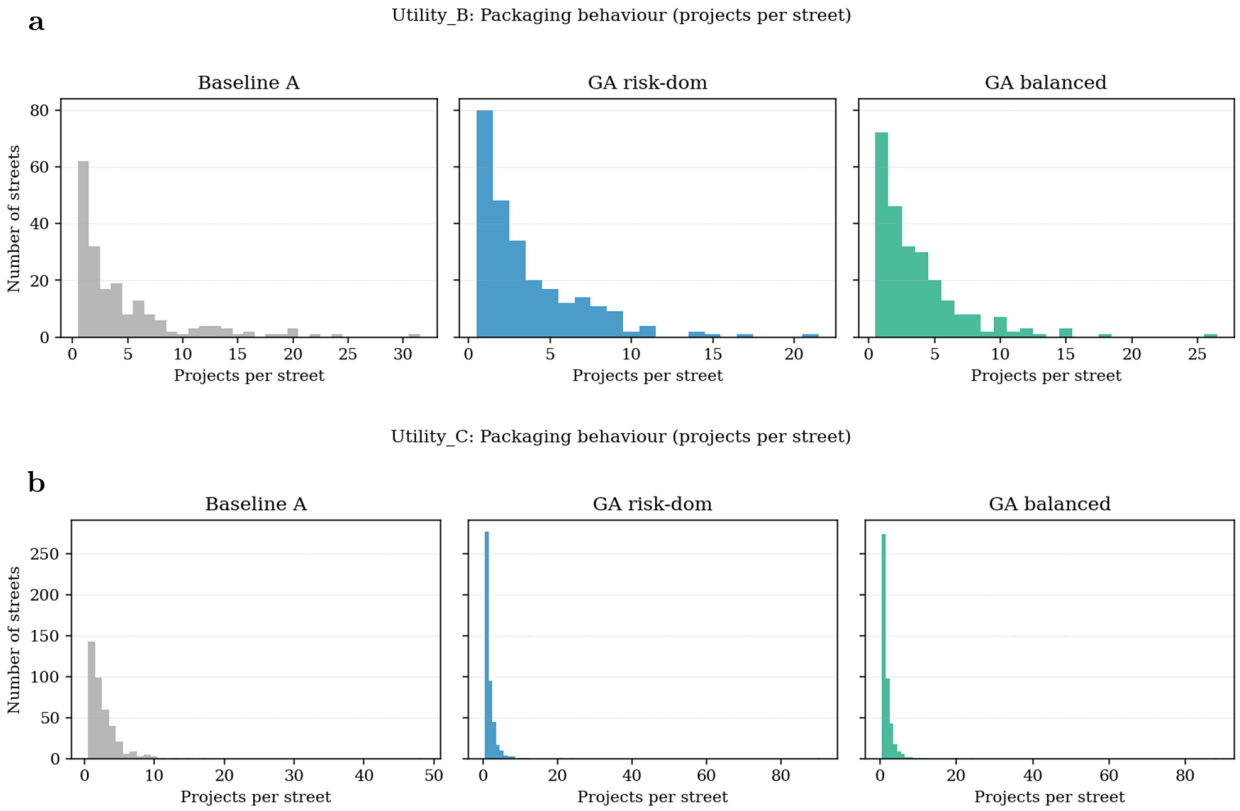

Figure 6-14: Packaging behavior for baseline and GA portfolios in Utilities B and C, expressed as histograms of projects per street. Each panel shows the distribution of the number of selected projects per street for the baseline risk-only portfolio and for the GA risk-dominant and balanced portfolios.

In $\text{Utility}_{\text{B}}$, which uses street name as a concurrency proxy, Baseline A selects 899 pipes over 224.8 miles on 193 streets, of which 131 have two or more projects. The GA risk-dominant portfolio selects slightly more pipes and miles, about 927 pipes and 231.8 miles, but spreads them across 256 streets, with 176 streets hosting multiple projects. The balanced GA portfolio is very similar. The histograms in Figure 6-14 show that the GA

portfolios maintain a strong concentration of streets with one to three projects, while introducing a modest tail of streets with higher project counts. This behavior is consistent with the packaging term in the objective, which is designed to form short project corridors where the risk and equity justify them, without creating excessive concentration on a handful of streets.

In $\text{Utility}_{\text{C}}$, Baseline A selects 1,088 pipes over 272 miles on 394 streets, with 251 streets having two or more projects. The GA risk-dominant portfolio selects fewer pipes and fewer miles, about 957 pipes and 239 miles, but spreads them across 460 streets, with only 183 streets having multiple projects. The balanced GA portfolio behaves similarly. Here the GA clearly reduces repeat excavations while also reducing the total miles renewed, consistent with trading a small loss in risk-per-dollar performance for better spatial dispersion and less disruption from repeated excavation on the same streets.

Robustness across seeds is assessed by examining the run-to-run standard deviations in risk capture, equity capture, and cost for each scenario and utility. Across all three utilities and all three weight configurations, these standard deviations are very small, typically between 0.0005 and 0.004 for capture fractions and well below 0.1 % of the budget for total cost. This pattern indicates that, once weights and budgets are fixed, the

GA converges to a tight cluster of high-quality solutions rather than wandering across widely different portfolios.

Taken together, these results allow us to reject H_{V4a} and H_{V4b} for the utilities in which packaging can be assessed. The GA portfolios are at least as good as the risk-only baseline in terms of projects per street and, in the case of $Utility_C$, clearly superior in reducing repeat excavations while preserving budget discipline and high risk capture. The low variability across seeds further indicates that the GA optimization algorithm behaves robustly and predictably for fixed settings.

6.7.3.5 Summary and link to validation

The four verification experiments demonstrate that the renewal prioritization framework behaves as intended when driven by real utility data and risk scores. Table 6-16 summarizes the decisions on each null hypothesis for the three utilities and the key metrics supporting those decisions. First, the mapping exercise shows that the GA receives a realistic set of inputs corresponding to the criteria utilities described in surveys and interviews, with proxies and data gaps clearly documented. Second, under risk-dominant weights, the GA behaves like a budget-aware extension of each utility's own risk ranking,

capturing risk efficiently while preserving substantial overlap with the utility's top-risk pipes and only modestly sacrificing risk-per-dollar efficiency compared with an idealized risk-per-cost baseline. Third, the optimization is responsive to weight changes in the expected direction where increasing the emphasis on equity and historically problematic sites yields higher capture of those criteria while keeping risk capture within a narrow band and avoiding large shifts of budget into low-risk quartiles. Fourth, where data allow, the GA exhibits reasonable packaging behavior and generates portfolios that are highly robust across random seeds.

Table 6-16: Summary of verification outcomes by experiment and utility

Exp.	Utility	Key evidence (selected metrics)	Decision	Brief interpretation
V_1: Data mapping & coverage	A	7/10 criteria represented ($risk_{model}$, $risk_{utility}$, complaints, equity, historical problems, legacy material, demand); PRV, service line, concurrency not available; all used proxies show non-zero variance.	Reject H_0-V_1	For $Utility_A$ the GA sees substantially more than just diameter, age, and risk; most of the planner-reported criteria are either directly observed or expressed as proxies, with clear variation across candidates.
	B	8/10 criteria represented; concurrency/moratoria proxied via street name; PRV and service-line information not present; all active proxies non-degenerate.	Reject H_0-V_1	The $Utility_B$ dataset supports the richest coverage: risk, historical breaks, equity, legacy material, demand, and a concurrency proxy are all present in the GA inputs.
	C	8/10 criteria represented; similar pattern to $Utility_B$ with street-based concurrency proxy; PRV and service-line flags absent.	Reject H_0-V_1	For $Utility_C$, the GA again operates on a multi-criteria view of the system, rather than a single risk score.
V_2: Risk-aligned behavior	A	Risk capture per $M: Baseline A = 180, Baseline B = 306, GA risk-dom = 200; GA risk capture = 0.403 vs 0.363 (Baseline A); GA portfolio overlaps with 48–49% of the utility's top-50/top-100 risk pipes and concentrates budget in upper risk quartiles.	Reject H_0-V_2	When told that risk is paramount, the GA behaves like a budget-aware risk ranking that still honors the utility's risk model, while staying within ~35% of the best possible risk-per-cost baseline and using the same information to package projects.
	B	Risk capture per $M: Baseline A = 44.6, Baseline B = 46.3, GA risk-dom = 45.5; GA risk capture = 0.310 vs 0.304 (Baseline A); GA recovers 24–28% of the top-	Reject H_0-V_2	For $Utility_B$, the GA risk-dominant portfolios sit between the two simple risk-based

Exp.	Utility	Key evidence (selected metrics)	Decision	Brief interpretation
		50/top-100 pipes from the utility's ranking within the down-sampled candidate set and again spends almost all of the budget in the higher risk quartiles.		baselines in risk-per-dollar and still systematically favor higher-risk segments.
	C	Risk capture per $M: Baseline A = 31.1, Baseline B = 31.6, GA risk-dom = 27.6; GA risk capture = 0.322 vs 0.363 (Baseline A); nonetheless the GA portfolio recovers 39–44% of the top-100 risk pipes and focuses spend in the top risk quartiles.	Reject H_0-V_2 (with caveat)	For Utility$_C$ the existing risk-per-cost ranking is very strong, so the risk-dominant GA trades a modest reduction in risk-per-dollar for the ability to consider equity and packaging. Even in this setting, its selections are clearly not risk-agnostic.
V_3: Multi-criteria responsiveness	A	Across scenarios, risk capture ~0.402–0.403 while equity capture rises slightly from 0.406 (risk-dominant) to 0.408 (balanced/equity-dominant); historical-problem capture drops from 0.409 to ~0.364 when other criteria are emphasized; legacy capture changes only slightly.	Reject H_0-V_3	For Utility$_A$ the scenarios are deliberately conservative, but they still show a coherent pattern: pushing equity and other criteria up slightly trades off some historical-problem emphasis while leaving overall risk capture essentially unchanged.
	B	Risk capture ~0.309–0.310 across scenarios; equity capture increases from 0.311 (risk-dominant) to 0.312 (balanced/equity-dominant); historical-problem capture drops from 0.299 to 0.293; legacy capture decreases more markedly from 0.317 to 0.297 as weight shifts away from legacy removal.	Reject H_0-V_3	In Utility$_B$, changes in weights produce small but interpretable shifts: equity and complaints gain slightly more attention in balanced and equity-dominant scenarios, at the expense of strongly favoring legacy material.
	C	Risk capture ~0.322 (risk-dominant) vs 0.321 (balanced/equity); equity capture ~0.319 in all scenarios; historical-problem capture ~0.409–0.410; legacy capture 0.405–0.406.	Reject H_0-V_3 (weak effect)	For Utility$_C$ the proxies are more correlated, so scenario differences are intentionally modest; nevertheless the patterns are stable across seeds and show that the optimization is not chaotic when weights are perturbed.
V_4: Packaging & robustness	B	Compared with Baseline A, GA portfolios slightly increase distinct streets (193 → 249–256) and streets with ≥2 projects (131 → 176–177), with similar total pipes and miles; histograms show a shift from many single-project streets towards more 2–5-project clusters, while avoiding extreme 20+ project "mega-corridors".	Reject H_0-V_4	For Utility$_B$, the GA restructures projects into modest clusters rather than isolated single jobs, while keeping total footprint similar—consistent with a packaging effect rather than random scatter.
	C	Baseline A: 394 distinct streets, 251 streets with ≥2 projects, 1088 pipes, 272 miles; GA portfolios: ~459 distinct streets, 183–185 streets with ≥2 projects, 956–957 pipes, ~239 miles; histograms again show more 2–4-project streets and fewer very long multi-project corridors.	Reject H_0-V_4	For Utility$_C$, the GA achieves similar risk and criteria capture with fewer miles renewed and a different spatial pattern: many streets see a few coordinated projects instead of a mix of one-offs and very long chains, which is closer to the intended constructability heuristic.

These verification results provide a necessary bridge between the LOF and COF chapters and the forthcoming validation section. They show that, on real datasets, the GA uses the information that is available, respects the underlying risk models, and

responds to multi-criteria trade-offs in a controlled and interpretable way. The next step is to move from internal verification to external validation, where utilities supply their own criteria weights, review the results based on representative examples, and where past programs allow, compare the model's recommendations to the actual capital programs and their lifecycle consequences.

6.7.4 Validation with expert renewal scenario feedback

This section asks a different question from the verification experiments. There we showed that the GA portfolios behave sensibly with respect to risk, equity, project packaging and other metrics given a utility dataset. Here we ask whether the decisions implied by the model are acceptable to practitioners who develop renewal programs.

To keep the cognitive load manageable, we did not ask planners to evaluate full portfolios in detail. To get better participation, we constructed a set of representative scenarios using the same portfolio results we generated in the last section using pipe segments from each utility's GIS. Each row in the feedback tables corresponds to a real pipe, with its pipe ID, material, operating context, and a short natural language description. For every scenario we reported the model's result (selected vs not selected, and if

selected its rank within the candidate set) and asked asset management staff to indicate whether they agreed with that decision and, if not, to explain why.

Across the three anonymized utilities (A: coastal California, B: Pacific Northwest, C: southeastern U.S.) we used eight scenarios per utility, spanning both "renewal candidate" and "not a renewal candidate" situations and covering the main materials with different deterioration modes and consequence scenarios as in the LOF and COF chapters. The same scenario descriptions were used across the 3 utilities to ensure uniform comparisons as those represent different renewal scenarios.

6.7.4.1 Hypotheses and simple scoring

To structure the validation of the decision support layer, we first formalized three scenario-based hypotheses about expert agreement, safety of "do not renew" decisions, and the nature of disagreements (Table 6-17). Each hypothesis is tied to simple, interpretable metrics and an explicit rule for rejecting the null.

Table 6-17: Scenario-based validation hypotheses for the renewal prioritization model, with corresponding metrics and rejection rules.

ID	Focus	Null hypothesis (H_0)	Metric(s)	Planned rejection rule
H_{V1}	Overall decision-level concordance	Model decisions (renew vs not renew) agree with expert judgement no more often than chance.	Proportion of scenarios where expert verdict is "Agree/Strongly agree/Agree (slight)" vs "Disagree"; simple sign test vs 0.5 baseline.	Reject H_0 if the agreement rate is clearly above 0.5 and the one-sided sign test p-value is $\lesssim 0.05$.
H_{V2}	Safety of "do not renew" decisions	For "not a renewal candidate" scenarios, experts are as likely to say, "this should have been renewed" as "this is fine to leave out."	Agreement rate restricted to scenarios where the model does not select the pipe (or reports "no pipe in scenario").	Reject H_0 if disagreements in this subset are rare (e.g., $\leq$2 cases) and dominated by policy/scope differences rather than clear risk misclassification.
H_{V3}	Structure of disagreements	Disagreements between models and experts are idiosyncratic or random.	Qualitative coding of comments into categories (program scope, data/ID issues, local policy rules, genuine risk disagreement).	Reject H_0 if most disagreements fall into interpretable categories (scope, policy, data issues) rather than contradicting the underlying risk logic.

To score the feedback response tables from water utilities, we treated any response containing "Agree", "Strongly Agree", or "Agree (slight)" as an agreement, any "Disagree" as a disagreement, and cells left blank as missing. Two rows where the utility could not locate the referenced facility, or where the pipe no longer existed in the pressurized mains layer, were treated as missing rather than as disagreements.

Across the three utilities this gives 24 scenarios in total. Excluding the two missing responses, we have 22 scored cases, of which 18 are agreements and 4 are disagreements. A simple sign test against a null of 50 % agreement gives a one-sided p-value of about 0.002. Even with the small sample, this is strong evidence against H_0, so we reject H_{V1}

and conclude that the model's decisions are not aligned with expert judgement by accident.

These counts are visualized in Figure 6-15, which shows, for each utility, the proportion of scenarios that resulted in agreement, disagreement, or scope/data issues. The stacked bars emphasize that agreements dominate across all three utilities, with only a small number of structured disagreements and a few rows affected by referencing or program-scope issues.

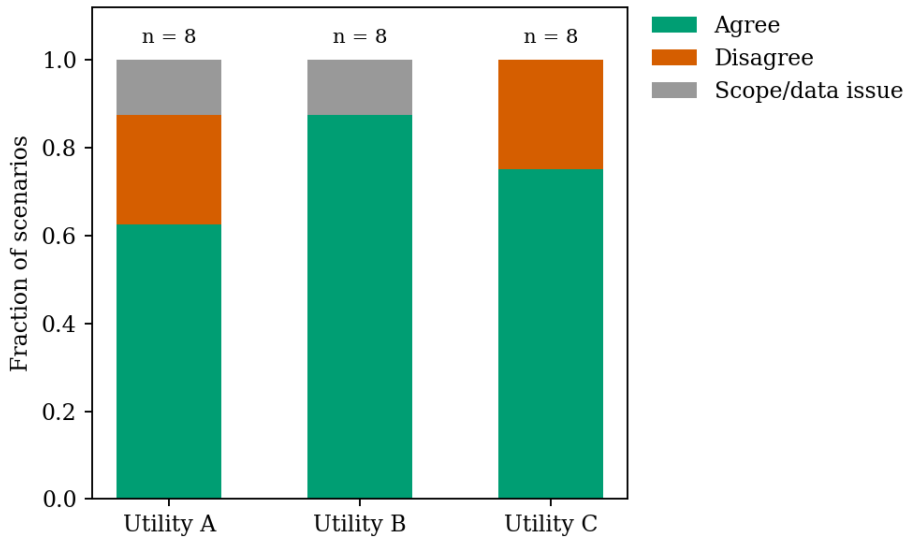

Figure 6-15: Scenario-level agreement by utility. Stacked bar plot showing, for each anonymized utility (A–C), the fraction of scenarios classified as agreement (A), disagreement (D), or scope/data issue (S). Agreements dominate across all three utilities, supporting H_{V1} that model decisions are not aligned with expert judgement by chance.

Detailed scenario-by-scenario feedback tables for each utility are provided in Appendix F. Here we focus on cross-utility patterns and hypothesis tests.

6.7.4.2 Agreement patterns by scenario type

It is useful to separate scenarios where the model recommends renewal from those where it does not. For "renewal candidate" scenarios (pipes the scenario description suggests should be in the program), the GA selected the pipe in 12 cases. Experts agreed with those selections in 8 of 11 scored cases and disagreed in 3, with one additional row left unscored ("hydrant main; we do not believe it would be included in the typical selection process"). The agreed cases include:

- Metallic cast iron mains with <80 % remaining wall thickness in dense urban areas.
- PCCP with >12 wire breaks per 10 ft and deteriorated steel cylinder.
- Externally corroded cast iron mains with persistent background leakage.
- AC mains over 50 years old in high-traffic areas.

In these agreed rows, the pipe segments sit high in the model's risk ranking (e.g., rank 1/154, 8/717, 19/717, 25/154), and the comments confirm that staff have seen repeated failures or ongoing leakages on those segments. This supports the idea that the

combined LOF–COF signal used by the GA is capturing the kinds of "obvious candidates" that practitioners recognize.

For "not a renewal candidate" scenarios, the model generally leaves the pipe out of the renewal set or reports "no pipes in such scenario" (for configurations that do not exist in the dataset). There are 12 such scenarios. Experts explicitly agree with these "do not renew" decisions in 10 of 11 scored cases, with one row marked "??" because the reference pipe could not be located. Examples include:

- RCP/PCCP/RCCP with only minor steel corrosion and intact concrete core in low-demand areas.
- PVC in suburban, low-pressure zones with <5 % ovality and stable soils.
- AC mains older than 50 years in low-traffic areas that have already been abandoned ("main was killed in 1993").

This pattern is important for H_{V2}. In this small but diverse set the model almost never suggests "do not renew" in a situation that experts judge unacceptable. The single explicit disagreement in a non-candidate scenario (Utility C, 6″ AC >60 years in a low-

traffic area) reflects a deliberate local policy choice (pre-1955 AC prioritized based on WRF 4480) rather than a gross miss in structural or functional risk.

Taken together, these results show that the model is conservative in the right direction. It is more prone to treat borderline cases as renewal candidates that some programs would defer, than to leave out pipes that local staff consider clearly problematic.

6.7.4.3 Qualitative structure of disagreements

The four explicit disagreements (plus one unscored "hydrant main" case) are instructive because they inform how local program rules interact with the generic risk model.

In Utility A, staff disagreed with the renewal of a metallic main with low remaining wall thickness in a dense urban area and with a large-diameter PCCP with many wire breaks and deteriorated cylinder. In both cases the written comments did not dispute that the pipes were high-risk. Instead, they argued that:

- the metallic main was "technically feasible to be selected, but disagree with renewal candidacy (this year) due to its relative rank", and

- the PCCP main would likely be handled as a contract project for a wholesale supply line, and therefore sits outside the in-house renewal program where the model was being applied.

In Utility C, one disagreement concerns PVC in a high-pressure zone with >5 % ovality. Here, staff note that "we haven't seen PVC oval lately", implicitly down-weighting the ovality indicator based on recent local experience. The other disagreement is the AC pipe >60 years old in a low-traffic area, where staff state that "pre-1955 AC is prioritized based on WRF Study 4480", effectively imposing a hard age threshold for that material regardless of traffic or demand.

None of these comments say "your risk model is wrong about how deterioration works". Instead, they point to program boundaries and policies:

- Differences between contract projects and in-house programs.
- Exclusion of specific asset types (hydrant mains) from the main renewal pool.
- Program-specific rules driven by external studies (e.g., pre-1955 AC policy).
- Local heuristics about which indicators are still informative (PVC ovality).

This pattern supports H_{V3}. Disagreements are structured and interpretable. They identify additional constraints and policy rules that can be layered on top of the generic risk-based GA rather than contradicting its internal logic. In the thesis, we treat these as opportunities for model refinement. For example, adding program-scope flags (in-house vs contract, transmission vs distribution, hydrant zones) as hard constraints, and allowing utilities to encode material- and vintage-specific policy thresholds as additional objective terms.

6.7.4.4 Overall interpretation

To compare how the experts provide feedback to all the scenarios, we present Table 6-18 that aggregates outcomes into an A/D/S grid. The pattern shows consistent agreement on low-risk deferrals and on most high-risk candidates, with disagreements clustering in a small number of scenarios where program scope or local policy explicitly differs.

Table 6-18: Cross-utility summary of scenario outcomes, showing agreement (A), disagreement (D), and scope/data issues (S) for each canonical scenario across Utilities A–C.

Canonical scenario	Utility A	Utility B	Utility C	Notes
Metallic CI, RWT <80 %, high-density urban (renewal candidate)	D	A	A	All utilities agree pipe is high-risk; U-A's disagreement is about *relative rank in this year's program.*
PCCP with >12 wire breaks, deteriorated cylinder, no redundancy (renewal candidate)	D	A	A	U-A treats the 60″ main as a contract project (outside in-house program); U-B and C accept renewal.

Canonical scenario	Utility A	Utility B	Utility C	Notes
RCP/PCCP/RCCP with minor corrosion, intact core, low-demand area (not a renewal candidate)	A	A	A	Consistently judged acceptable to defer across all three utilities.
PVC in high-pressure zone with >5 % ovality (renewal candidate)	A	A (slight)	D	U-C down-weights ovality based on recent experience ("we haven't seen PVC oval lately").
PVC in suburban low-pressure zone, <5 % ovality, stable soils (not a renewal candidate)	A	S (pipe not identified)	A	Where evaluated, all utilities accept deferral.
Externally corroded CI with background leakage over 5 days (renewal candidate)	S (hydrant main excluded)	A	A	Disagreement at U-A is about scope (hydrant vs distribution mains), not risk.
6″ AC >50 years, high-traffic area (renewal candidate)	A (no such pipe; agrees)	A	A	When the configuration exists, all utilities support renewal.
6″ AC >50–60 years, low-traffic area (not a renewal candidate)	A (pipe killed)	A (<20 years scenario)	D	U-C applies a policy rule (all pre-1955 AC prioritized) and wants renewal despite low traffic.

Legend: **A** = Agree / Strongly agree / Agree (slight); **D** = Disagree; **S** = Scope / data issue (e.g., no pipe or unidentifiable); "–" = scenario not present.

Figure 6-16 provides a more compact visual summary of the same A/D/S grid, highlighting that most cells are agreements (A), with a small number of disagreements (D) and scope/data issues (S) concentrated in a few specific scenarios.

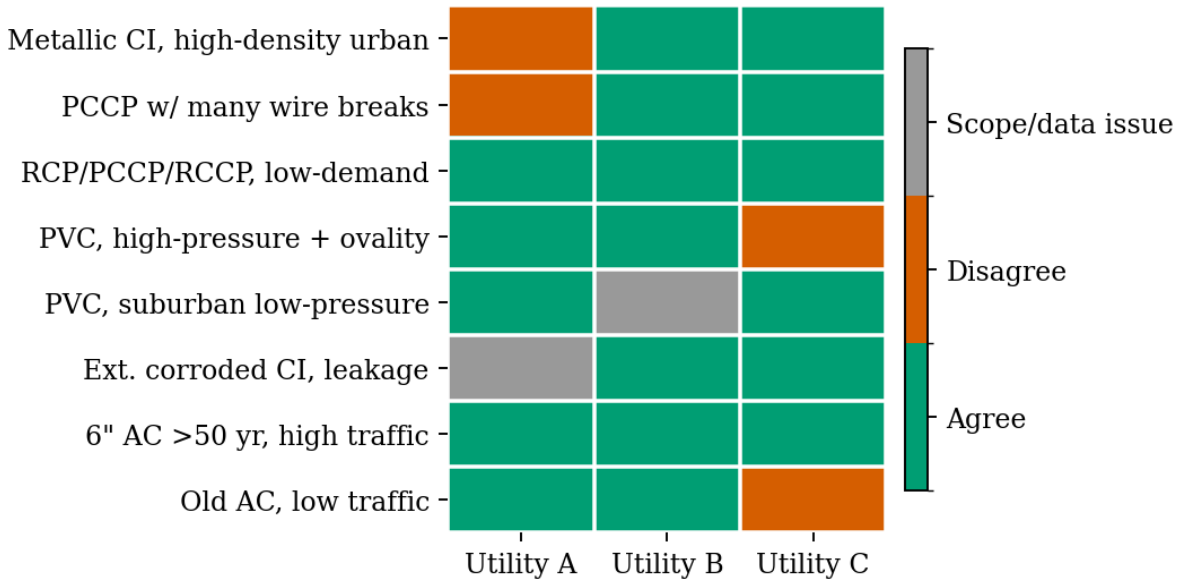

Figure 6-16: Cross-utility scenario outcomes for canonical renewal cases. Heatmap of canonical scenarios (rows) versus utilities A–C (columns), with each cell coded as agreement (A), disagreement (D), or scope/data issue (S). Most cells are agreements, with disagreements and scope issues concentrated in a small number of scenarios where local policy or program boundaries differ from the generic GA setup.

The results allow all three null hypotheses to be rejected with clear reasoning (Table 6-19). Overall, agreements dominate, "do not renew" decisions are conservative, and the remaining disagreements are structured and traceable to program scope or policy choices rather than to failures in the risk logic.

Table 6-19: Outcomes of hypothesis tests for scenario-based validation, summarizing evidence, tests, and final decisions on each null hypothesis.

ID	Evidence summary	Test / argument	Outcome
H_{V1}	Across 3 utilities and 24 scenarios, 2 were unscored (data/scope issues). Of the remaining 22, 18 were agreements and 4 disagreements (≈82 % agreement).	One-sided sign test against 0.5 baseline gives p≈0.002; agreements are also well distributed across utilities and scenario types.	Reject H_0. There is strong evidence that scenario-level model decisions align with expert judgement more often than chance.
H_{V2}	Among 12 "not a renewal candidate" scenarios, experts agreed with the model in 10 of 11 scored cases; 1 was "??" due to an unidentifiable pipe. The single explicit disagreement is an AC policy case (pre-1955 AC always prioritized) rather than a structural-risk miss.	Simple counting argument plus inspection of comments; no cases where experts say a clearly high-risk pipe was wrongly left out because of model failure.	Reject H_0. "Do not renew" decisions are conservative and generally safe; disagreements are driven by local policy rules, not by missing obvious high-risk segments.
H_{V3}	Disagreements cluster into interpretable categories: (i) contract vs in-house program boundaries (large PCCP transmission main; hydrant main), (ii) material/vintage policies (pre-1955 AC), and (iii) local de-emphasis of indicators (PVC ovality considered less relevant). None of the comments claim the risk logic itself is wrong.	Qualitative coding of each disagreement and review of written comments.	Reject H_0. Disagreements are structured and highlight program-scope and policy choices that can be added as constraints or preferences on top of the GA, rather than undermining the risk model.

This scenario-based validation has obvious limitations. there are only eight scenarios per utility, the "agree/disagree" responses are coarse, and some rows are affected by data referencing issues. However, the exercise can be expanded in future work to cover more renewal scenarios and make the decision support tool more robust. The key outcomes are:

- Across 22 scored scenarios, utilities agreed with the GA's decision in about 82 % of cases.

- "Do not renew" decisions were very safe: in 10 of 11 scored non-candidate scenarios, experts were comfortable leaving the pipe out of the renewal set.
- Disagreements point to program-specific scope and policy rules, not to systematic misranking of obvious high- or low-risk segments.

Combined with the quantitative verification experiments, these results support the claim that the GA portfolios are both risk-coherent and decision-compatible. The model can be trusted to produce portfolios that look sensible to practitioners for typical scenarios, while its structured disagreements highlight where local program rules need to be made explicit and integrated.

6.7.5 Summary

This chapter develops, tests, and validates the renewal-prioritization layer that sits on top of the LOF and COF models. The goal is to move from per-pipe risk scores to concrete renewal programs that respect budgets, delivery constraints, and equity considerations. To do this, we framed portfolio design as a multi-criteria optimization problem, represented utility priorities as transparent proxy variables, and used a Genetic Algorithm to search the space of feasible portfolios.

Evaluation focused first on the behavior of the GA itself using a synthetic, utility-like project set. These runs confirmed that the search converges reliably, respects the budget constraint almost exactly, and responds smoothly to changes in scalarization weights on risk, equity, and cost penalty. Across risk-dominant, equity-dominant, and balanced settings the GA produced portfolios with similar overall leverage, but with predictable shifts in where risk and equity benefits were concentrated. Selection-frequency and local regret experiments on the synthetic portfolios showed that a small core of projects is chosen in almost all runs and that omitting any of these projects leads to measurable losses in risk capture and scalar utility. This supports a low-regret interpretation of the optimization engine and shows that its behavior is driven by the structure of the decision problem rather than by random initialization.

Verification then applied the same GA configuration to three anonymized utilities with different network scales and data richness. For each utility we constructed proxy columns that represent criteria planners report using in practice, such as model risk, historic breaks, legacy material removal, equity priority, and opportunities for bundling work on common streets. The experiments compared GA portfolios to simple baselines that sort by utility risk or by risk per unit cost. Across utilities, GA portfolios captured risk and

equity at least as efficiently as the baselines for the same budget, and they did so while also improving secondary metrics such as coverage of high-equity areas and concentration of work on fewer streets where address data were available. The sensitivity runs with risk-dominant, balanced, and equity-dominant weights showed that the optimization layer can shift portfolio emphasis without losing the core high-risk projects that any reasonable program would include.

Validation treated the model as a decision support tool and asked whether its implied decisions are acceptable to practitioners. Using eight canonical scenarios per utility, drawn from the same GIS inventories as used for verification, asset managers were asked whether they agreed with the model's decision to include or exclude each pipe and to explain any disagreement. Across 22 scored scenarios, utilities agreed with the model in about four out of five cases, with a sign test indicating that this level of agreement is unlikely to arise by chance. In the "not a renewal candidate" scenarios, disagreements were rare and driven by local policy rules, for example a blanket rule that all pre-1955 AC must be prioritized, or by program scope boundaries such as the treatment of hydrant mains and contract projects. These comments did not challenge the underlying LOF and

COF logic, but instead pointed to additional constraints that can be layered on top of the GA for specific utilities.

Taken together, the evaluation, verification, and validation results support the claim that the proposed renewal-prioritization framework is both scientifically grounded and operationally credible, while also making its limits explicit. The GA behaves in a stable and interpretable way, and the resulting portfolios perform at least as well as natural baselines while adding capabilities that planners want. At the same time, the portfolio experiments draw on three utilities and focus on high-risk candidate sets, key criteria depend on proxies where direct data are missing, and the scenario-based validation covers a modest number of cases with coarse "agree / disagree" responses. Even so, the different lines of evidence are mutually reinforcing where the GA portfolios respect and sharpen utility risk rankings, respond coherently to changes in planning priorities, and exhibit constructability advantages where street level data allow. Expert feedback shows that, for representative scenarios, the model's decisions largely match practitioner judgement, and that remaining gaps can be addressed by adding explicit program-scope flags and policy constraints rather than by discarding the risk framework. Together, these results support

the conclusion that the renewal prioritization model can be integrated as a decision support tool with supporting protocols for continuous improvement and data collection.

Chapter 7

Conclusions and Recommendations

This chapter synthesizes the findings of the three core studies, the Likelihood of Failure (LOF) modeling, the Consequence of Failure (COF) modeling, and the integrated risk-based portfolio optimization into a coherent overall narrative. It interprets the empirical results against the overarching hypothesis set H_{1a}–H_{3c} (see Table 7-1), drawing out what was demonstrated and where evidence remains partial or conditional. The discussion then broadens from chapter-level results to their implications for water pipeline asset management practice and for the science of infrastructure risk modeling more generally, including explicit consideration of model limitations and boundary conditions. Finally, the chapter outlines a forward-looking research and practice agenda, identifying concrete, actionable directions where the framework developed here can be extended, stress-tested, or adapted to other components of source-to-tap water systems.

Table 7-1: Hypotheses tested in this research, grouped by goal and summarized at the level of what was actually implemented and evaluated.

Goal	H_x	Hypothesis
Goal 1: LOF model	**H_{1a} (Mechanism and context coverage)**	The LOF framework explicitly encodes structural, functional, and environmental drivers across major materials and diameter bands, extending beyond age/diameter practice by covering a broader set of documented deterioration mechanisms.
	H_{1b} (Student learning fidelity)	Student ML LOF models learn the fuzzy-teacher mappings with high accuracy and macro-F1 on held-out and synthetic stress-test data, yielding strongly diagonal confusion matrices with very few multi-band misclassifications.
	H_{1c} (Ground-truth concordance)	In independent validation cohorts with condition measurements (wall-thickness loss, wire-break counts) and retrospective failures, higher LOF bands are associated with worse measured condition and higher failure frequencies, with ordinal agreement statistics significantly above chance and errors dominated by ± 1-band deviations.
	H_{1d} (Expert concordance and face validity)	For curated LOF scenarios, asset managers and field staff judge the predicted LOF bands as broadly consistent with operational experience at rates well above chance (including tolerant ± 1-band agreement), and observed disagreements are explainable by data lineage or explicit policy choices rather than erratic model behavior.
Goal 2: COF model	**H_{2a} (Dimensional coverage and representativeness)**	The COF framework decomposes consequence into explicit economic, environmental, social/service, and operational sub-indices, and membership-function panels plus best/average/worst scenarios demonstrate coherent, monotone coverage from low- to high-impact combinations in each dimension.
	H_{2b} (Modular behavior and structural verification)	Within each COF dimension, increasing adverse inputs (for example, higher repair costs, more critical customers, tighter access constraints) produces monotone increases in the corresponding sub-index and in the overall COF band, and global sensitivity analysis shows no single parameter or module dominates the index, supporting stable modular substitution.
	H_{2c} (Agreement with existing utility indices and expert judgement)	When compared with incumbent utility consequence indices and expert scenario ratings, COF bands show strong ordinal alignment, with most cases on or near the diagonal of confusion matrices and positive, substantial rank correlations, and divergences trace to scale/scope differences rather than incoherent model behavior.
	H_{2d} (Ground-truth consequence calibration)	For documented main-break events with usable consequence descriptions, higher COF bands align with more severe observed proxies (for example, outage duration, disruption, visible damage), and confusion matrices plus ordinal metrics indicate broad calibration with only a small number of explainable two-band outliers.
Goal 3: Renewal Prioritization model	**H_{3a} (Portfolio effectiveness under constraints)**	Under fixed budget constraints and realistic candidate pre-screening, GA-optimized renewal portfolios built from LOF, COF, and auxiliary scores capture more risk per unit cost than simple rank-only or cost-weighted baselines across multiple utilities.
	H_{3b} (Decision alignment and acceptability)	In scenario-based validation with three utilities, planners' and asset managers' preferred options align with GA-recommended portfolios at rates well above chance, and where they diverge, qualitative comments point to scope or data limitations rather than systematic contradictions.

Goal	H_x	Hypothesis
	H_{3c} (Stability and robustness of portfolio recommendations)	Across changes in scalarization weights, random seeds, and utility datasets, the GA portfolios occupy a compact region of the risk–equity trade-off space and scalar performance metrics vary modestly, indicating that the recommended portfolios are robust to reasonable variations in preferences and initialization.

The dissertation was guided by three high-level goals: (1) to develop and validate an AI-based LOF model that assigns each water pipeline segment to an interpretable 0–5 performance/failure-propensity band based on structural, functional, and environmental drivers; (2) to construct a modular COF model that quantifies the severity of economic, environmental, social/service, and operational impacts on a comparable 0–5 scale; and (3) to integrate these LOF and COF indices into a constraint-aware portfolio framework that selects renewal projects to maximize risk reduction and service reliability per unit budget. The preceding chapters documented data assembly, model design, and experiment-level results in detail. The present chapter does not repeat those derivations, but instead focuses on synthesis and interpretation including discussion on how well the combined evidence supports the hypotheses, what this implies for the way water utilities plan renewals, and how the work fits into broader conversations on risk, prevention, and infrastructure governance.

7.1 Discussion

The discussion section interprets the empirical results of the three core studies that are, the LOF model, the COF model, and the integrated portfolio optimization. Here, the focus is on what the models collectively show about water pipeline risk and renewal, how well the evidence supports the overall hypothesis set H_{1a}–H_{3c}, and where the conclusions are strongest or more tentative. The section first examines each study in turn, relating its main findings to the corresponding hypotheses and highlighting its specific contributions and limitations. It then moves to a cross-cutting synthesis, drawing out common themes across LOF, COF, and portfolio layers and positioning the framework within broader conversations about infrastructure risk, evaluation–verification–validation, and preventive investment.

7.1.1 Discussion of LOF model results

Recap of LOF objectives and design: The LOF study set out to build an explainable, mechanism-aware model that assigns each pipe segment to a 0–4 likelihood-of-failure band using structural, functional, and environmental drivers rather than age and material alone. A fuzzy-logic "teacher" model encoded expert knowledge about deterioration mechanisms through membership functions and rules, and a student machine-

learning model learned these mappings at scale. The framework was then tested on multiple utilities using synthetic stress tests, internal cross-validation, and independent ground-truth datasets from inspections and failures.

H_{1a} (Mechanism and context coverage) is supported by the construction of the teacher models and input dictionaries. The LOF features systematically span external and internal corrosion, structural loading, hydraulic context, and installation/legacy factors across major materials and diameter bands. Coverage checks on the membership functions show that plausible ranges of key drivers (for example, soil aggressivity, pressure, and vintage) are explicitly represented. In contrast, age/diameter practice assumes homogeneity within broad cohorts and does not differentiate many of these mechanisms.

H_{1b} (Student learning fidelity) is supported by the confusion matrices and summary metrics for the student models. Across the material–diameter cohorts, held-out and stress-test experiments show strongly diagonal confusion matrices, high accuracy and macro-F1, and very few multi-band misclassifications. This indicates that the student models reliably approximate the fuzzy teacher's mapping while smoothing over local noise, and that the discretized LOF bands can be predicted with useful fidelity.

H_{1c} (Ground-truth concordance) is supported by the validation experiments that compare LOF bands to independent condition measurements and failure histories. For metallic pipes, higher LOF bands correspond to larger remaining wall-thickness loss; for PCCP, they correspond to higher wire-break counts and more severe wire-break clustering; for non-metallic pipes, higher LOF bands correspond to higher retrospective failure frequencies. Ordinal agreement statistics such as weighted κ and Spearman ρ reject independence and show positive, substantial concordance, and errors are dominated by ± 1-band deviations rather than severe mis-ordering.

H_{1d} (Expert concordance and face validity) is supported by scenario-based assessments completed by utility staff. For curated sets of segments and LOF maps, asset managers and field personnel judged the predicted bands as broadly consistent with their operational experience at rates well above chance, even under strict agreement criteria and more so under ± 1-band tolerance. Where disagreements arose, they could be traced to recognizable issues—such as incomplete local data, utility-specific policies (for example, blanket treatment of certain PVC vintages), or different implicit thresholds for "high risk" rather than to erratic behavior of the model itself.

Scientific and practical contributions of the LOF model: Scientifically, the LOF work demonstrates that a teacher–student architecture can reconcile expert-derived fuzzy logic with data-driven learning for buried, heterogeneous pipe infrastructure. The explicit mechanism-oriented feature design ensures that the model is not merely exploiting spurious correlations, while the student model provides the flexibility needed to handle large, noisy, multi-utility datasets. This hybrid structure provides a concrete template for "knowledge-informed ML" in settings where pure physics models and pure black-box models are both unsatisfactory.

Practically, the LOF bands offer utilities a richer and more stable risk signal than simple age/diameter rules, but still in an interpretable form that can be mapped to existing inspection and renewal workflows. The model supports prioritization ("which segments are most likely to fail?"), but also scenario testing ("how would risk shift if pressures increase or if specific environmental layers change?"), and exposes its own assumptions through the membership functions and expert-facing explanations. This combination of predictive performance, transferability, and explainability-in-use is a key contribution to the state of practice.

Limitations and boundary conditions (LOF): The LOF framework remains constrained by data quality, representativeness, and structural assumptions. Some material–diameter–environment combinations are under-sampled, so performance is best characterized for cohorts with substantial history (for example, legacy metallic pipes) and more uncertain for emerging materials and atypical conditions. Many stressors are still represented by coarse proxies, temporal resolution is mostly annual, and the learned relationships reflect historical operating regimes. As a result, extrapolation to new climates, new materials, or radically different operating policies requires caution and, ideally, recalibration. These are not reasons to avoid LOF modeling, but they define where additional data collection and local validation are most needed.

7.1.2 Discussion of COF model results

Recap of COF objectives: The COF study developed a fuzzy-logic framework that assigns each pipe segment to a 0–4 (or 0–5) consequence band, conditional on failure. Unlike LOF, which estimates how likely failure is, COF quantifies how severe the impacts would be in economic, environmental, social/service, and operational terms. Each dimension is represented by a sub-index, built from measurable inputs (for example, repair costs, critical customers, land use, traffic, access constraints) and aggregated via a transparent

rule base. As in the LOF work, a student model was trained to emulate the fuzzy teacher, making the COF index scalable, while preserving traceability back to the sub-components.

H_{2a} (Dimensional coverage and representativeness) is supported by the construction of the COF sub-indices and the best/average/worst scenario analyses. Membership-function plots and response surfaces show that low-impact combinations anchor low COF bands, while combinations involving many critical customers, difficult access, or environmentally sensitive locations anchor higher bands. This behavior holds across the economic, service, environmental, and operational dimensions, and contrasts with incumbent indices that primarily weight diameter and land use without explicitly resolving these dimensions.

H_{2b} (Modular behavior and structural verification) is supported by targeted monotonicity and sensitivity checks. Within each sub-index, increasing an adverse input (such as repair cost, critical-customer density, or traffic importance) never decreases the relevant sub-score, and the overall COF band responds in the expected direction. Global sensitivity analysis shows that multiple inputs contribute meaningfully to the index, with no single parameter dominating across the entire space. This behavior confirms that

swapping in utility-specific sub-modules (for example, local cost curves) will change levels but not break the qualitative structure of the COF model.

H_{2c} (Agreement with existing utility indices and expert judgment) is supported by the comparison of student COF outputs against incumbent utility consequence indices and expert ratings. Confusion matrices show that most segments lie on or adjacent to the diagonal, and rank correlations between COF and utility indices are positive and substantial. Scenario forms completed by practitioners show similar patterns. Experts generally agree when COF identifies high-consequence corridors, and disagreements can often be explained by differences in scale (for example, city-wide vs system-specific ranking) or by local policies that weight certain customer groups differently.

H_{2d} (Ground-truth consequence calibration) is supported, in a qualified way, by the experiment that compares COF bands with documented main-break consequences. For events with usable descriptions such as outage duration, extent of traffic disruption, or visible damage higher COF bands are associated with more severe outcomes, and confusion matrices show relatively few large-band discrepancies. However, the proxies are incomplete and noisy, and consequence reporting varies by utility, so this experiment provides evidence of broad calibration rather than precise numerical matching.

Scientific and practical contributions of the COF model: The COF framework advances consequence modeling for water pipelines by moving from opaque scalar scores to a modular, multidimensional representation that is still operational. Scientifically, it shows that fuzzy-logic structures can encode a rich set of consequence concepts in a way that remains auditable and compatible with ML emulation. Practically, it gives utilities a transparent explanation of why a segment is labeled high consequence, whether because of critical customers, complex access, environmental sensitivity, or a combination of these and it supports iterative refinement of sub-modules as better local data become available.

The main tradeoff is complexity. Representing multiple dimensions and explicit uncertainty bands makes the outputs richer but also more demanding to interpret than a single number. This is deliberate as consequence is ethically and operationally complex, and the model is designed to make that complexity visible rather than hiding it inside a black box or a single heuristic score.

Limitations and boundary conditions (COF): The COF model is limited most strongly by consequence data. Many dimensions like environmental damage, long-term reputational effects, distributional impacts on vulnerable populations are only

indirectly observed, if at all. Even for economic and service consequences, reporting is uneven, and minor events are likely under-reported. These constraints limit the precision with which COF can be calibrated and make absolute comparisons across utilities difficult. Moreover, collapsing multidimensional consequences into one band necessarily glosses over equity and ethical considerations. The COF index should therefore be seen as a structured starting point for deliberation and stress testing, not as a final arbiter of whose losses matter.

7.1.3 Discussion of Renewal Prioritization Model (RPM)

Recap of portfolio objectives and design: The integrated portfolio study used LOF, COF, and auxiliary scores to construct renewal portfolios that maximize risk reduction under budget constraints. Candidate segments were pre-screened using risk and practical criteria, and a Genetic Algorithm (GA) searched over combinations of projects to identify portfolios that captured the most risk per unit cost. The GA was chosen because the decision space is combinatorial and constraints (budgets, packaging indicators, optional equity and water-loss scores) interact in ways that preclude simple greedy strategies.

H_{3a} (Portfolio effectiveness under constraints) is supported by comparisons between GA-optimized portfolios and baseline heuristics across multiple utilities. Within the screened candidate sets and fixed budget levels, GA portfolios consistently achieve higher scalar utility and risk capture per dollar than risk-ranked or cost-weighted baselines. This pattern holds across utilities and random seeds, indicating that explicitly optimizing over combinations under constraints yields materially better use of the same budget than independent, rank-based selection.

H_{3b} (Decision alignment and acceptability) is supported by the scenario exercises with three utilities. When presented with anonymized, side-by-side options, planners and asset managers often select GA-based portfolios or minor variants of them, and their narrative comments focus on data coverage, local program constraints, or unmodelled coordination issues rather than on fundamental disagreements with the model's logic. This suggests that the portfolios are not only computationally efficient but also broadly compatible with practitioner judgment when used as decision support rather than as an automatic selector.

H_{3c} (Stability and robustness of portfolio recommendations) is supported by sensitivity runs in which scalarization weights, seeds, and datasets are varied. Across

these changes, GA portfolios populate a compact region in risk–equity (and, where computed, risk–water-loss) tradeoff space, and scalar metrics (risk capture, equity capture, simple packaging indicators) vary modestly. High-value segments recur frequently, while the precise composition of portfolios changes at the margins. This behavior indicates that the recommendations are robust to reasonable variations in preferences and initialization, and that the optimization is capturing structural features of the risk landscape rather than over-fitting to arbitrary settings.

Contributions of the portfolio framework: The portfolio framework contributes a concrete, implementable way to move from risk scoring to budget-constrained action. Methodologically, it integrates LOF and COF into a unified objective, respects budget and simple deliverability considerations, and uses a GA to explore a large discrete decision space in a transparent way. Conceptually, it recasts renewal planning as a problem of maximizing preventive benefit (risk and disruption reduction) per dollar, rather than simply "replacing old pipes," and makes these tradeoffs explicit and quantifiable.

For utilities, the framework provides more than a list of "high-risk pipes". It produces candidate portfolios with clear expected impacts and shows how shifts in preferences (for example, more emphasis on equity) move the chosen portfolios along a tradeoff

frontier. This makes it easier to explain and defend renewal strategies to internal stakeholders, boards, and regulators.

Limitations and boundary conditions (RPM): The portfolio results are shaped by several limits. First, the GA explores only a subset of the enormous space of possible portfolios and depends on parameter choices. Second, diagnostics suggest adequate convergence, global optimality cannot be guaranteed. Third, the planning horizon is limited, so long-run dynamics and learning are only indirectly reflected through LOF and COF updates. Fourth, some important constraints like political commitments, regulatory negotiations, and coordination with other agencies can have more nuances than how it is currently set in the objective and constraint set. Finally, portfolio performance is only as reliable as the underlying LOF and COF estimates. These limitations reinforce that the GA framework should be used as a structured decision aid within a broader planning process, not as a substitute for expert judgment and institutional negotiation.

7.1.4 Cross-cutting synthesis

This section steps back from the individual studies to examine what the dissertation demonstrates as a whole. It interprets the evidence against the overall hypothesis set H_{1a}–H_{3c}, identifies common themes across the LOF, COF, and portfolio models, and

clarifies both the scope and the limits of the proposed framework for risk-based renewal of water pipelines.

7.1.4.1 Synthesis across hypotheses H_{1a}–H_{3c}

Taken together, the three core studies provide strong support for most of the hypotheses in Table 7-1, with a smaller subset remaining partially supported due to data limitations or narrow application contexts.

For the LOF model, H_{1a} and H_{1b} are well supported. The mechanism-oriented feature dictionaries and fuzzy teacher models explicitly encode structural, functional, and environmental drivers in ways that age–diameter practice cannot, and the student models reproduce these mappings with high accuracy and macro-F1 and strongly diagonal confusion matrices. H_{1c} and H_{1d} are supported where independent condition data and expert judgements are available that is, higher LOF bands align with worse wall-thickness, more wire breaks, and higher failure frequencies, and experts largely agree with the model's banding, with disagreements tied to recognizable data or policy issues. The limitation is that these validations are strongest for well-instrumented cohorts and utilities; rare materials and contexts remain less certain.

For the COF model, H_{2a} and H_{2b} are supported by the internal behavior of the fuzzy indices. Economic, environmental, service, and operational consequences are represented as distinct, monotone sub-indices, increasing adverse inputs never lowers the relevant scores, and sensitivity analysis shows that no single parameter dominates. H_{2c} is supported by the comparisons with incumbent utility indices and expert scenario ratings, which show strong ordinal alignment. H_{2d} is more qualified where the main-break experiment shows that higher COF bands broadly track more severe observed outcomes, but the proxies are incomplete and noisy, so this result is best interpreted as evidence of directional calibration rather than precise consequence prediction.

For the integrated portfolio model, H_{3a} is supported by consistent gains in risk captured per unit budget over simple rank-based and cost-weighted baselines within the screened candidate sets. H_{3b} is supported in a pragmatic sense where in three utilities, planners' preferences over anonymized portfolio options align with GA-generated portfolios more often than would be expected by chance, and their critiques focus on data scope and local constraints rather than on rejection of the model logic. H_{3c} is also supported within the tested range of scalarization weights, random seeds, and utilities. GA portfolios occupy a compact region of the risk–equity trade-off space and include recurrent high-

value segments, indicating robustness. However, these portfolio results remain conditional on the quality of LOF and COF inputs and on the particular objective formulations explored.

The synthesis, therefore, is that the evidence strongly supports the core structure of the framework, LOF, COF, and portfolio optimization as a coherent chain, while highlighting that calibration of consequences and long-run portfolio behavior under changing conditions remain open areas for further work.

7.1.4.2 Thematic contribution 1: Risk modeling for buried water infrastructure

A first cross-cutting theme is that combining LOF and COF into an explainable risk framework closes an important gap between today's heuristics and unvalidated black-box models for buried pipelines. Traditional practice leans on simple rules such as "old, large pipes in busy roads are high risk", which cannot distinguish among mechanisms or represent uncertainty. At the other extreme, unconstrained machine-learning models can achieve good internal fit but often lack interpretability, transferability, and explicit validation against independent data. To make this contrast more concrete, Table 7-2

summarizes how current water main renewal practice compares with the contributions of this dissertation across several key themes.

This dissertation adopts the standard risk decomposition, risk ≈ likelihood × consequence, but implements it in a way tailored to buried water systems. LOF is modeled as a discrete propensity to fail based on mechanisms (corrosion, loading, hydraulics, environment), while COF captures the severity of impacts if failure occurs, decomposed into economic, service, environmental, and operational dimensions. Both sides are constructed in a way that preserves interpretability (through fuzzy teachers and explicit feature dictionaries) while enabling large-scale deployment (through student models).

Table 7-2: Comparison between typical practice in water main renewal and the contributions of this dissertation, organized by categories.

Categories	Typical status quo	Contribution of this dissertation	Expected impact
LOF modeling	Age, material, and ad hoc risk scores; limited covariates; little cross-utility testing.	Mechanism-aware LOF model with structured feature taxonomy, teacher–student learning, and multi-utility validation including inspections and failures.	More accurate and explainable failure-propensity estimates, transferable across utilities with known limits.
COF modeling	Diameter × land-use heuristics or opaque scalar indices; mostly economic focus.	Modular, multidimensional COF model with explicit economic, service, environmental, and operational sub-indices and traceable fuzzy rules.	Broader conception of consequence (including service, environment, equity surrogates), enabling richer trade-off analysis.
Portfolio selection	Rank-only lists or manual project picking; constraints handled informally.	GA-based, constraint-aware portfolios built from LOF, COF, and auxiliary scores, evaluated against simple rank-based and cost-weighted baselines.	Portfolios that improve risk capture per dollar and packaging/equity metrics within the screened candidates and budget limits.

Categories	Typical status quo	Contribution of this dissertation	Expected impact
EVV for AI models	Limited internal metrics; rare external validation; little documentation.	Multi-layer EVV program combining internal metrics, cross-utility tests, inspection concordance, and expert elicitation for LOF, COF, and portfolios.	A practical template for how AI risk models in infrastructure can be tested and documented for adoption.
Governance & commons framing	Renewal framed mainly as engineering and budgeting; limited explicit discussion of equity and temporal commons.	Interprets renewal as a temporal commons and equity problem, connecting risk metrics to intergenerational and spatial fairness.	Helps utilities and regulators reason about who pays, who benefits, and how to justify preventive investment.

By doing so, the work provides not only three models but a reference framework for risk-based renewal. This framework provides a way for utilities to move from loosely justified scores to a structured, data- and knowledge-informed representation of risk that can be interrogated, updated, and used in optimization. This positions buried water pipelines closer to mature risk formalisms in safety engineering and disaster risk reduction, where probability–consequence decompositions and explicit uncertainty are standard.

7.1.4.3 Thematic contribution 2: Evaluation, Verification and Validation (EVV) for AI in infrastructure

A second major contribution is methodological. This dissertation implements a multi-layer Evaluation Verification and Validation (EVV) program for infrastructure AI models, rather than relying only on internal accuracy metrics that are currently prevalent.

In this framework, evaluation refers to internal checks on performance, such as confusion matrices, accuracy, macro-F1, and simple calibration summaries on held-out

data. Verification refers to confirming that the models behave as intended on unseen but similar data and under structured perturbations. For example, cross-utility train–test splits, synthetic stress tests, and monotonicity checks for COF modules. Validation refers to testing against independent, external evidence and expert judgement, often under different conditions than those used to fit the models. For example, remaining wall-thickness measurements, PCCP wire-break counts, retrospective failure histories for non-metallic pipes, and expert scenario assessments for LOF, COF, and portfolios.

By design, each of the three studies contributes different pieces to this EVV ladder. LOF and COF contribute evaluation and verification across multiple utilities, plus validation through inspections and failure datasets. The portfolio study adds validation at the decision layer through planner preference experiments and qualitative feedback. Together, these elements amount to a comprehensive validation program for AI in water pipeline management, and they illustrate how infrastructure models can be documented and tested to a standard closer to those in other safety-critical domains.

7.1.4.4 Thematic contribution 3: Decision support and temporal commons

A third theme concerns how the framework is used in decision making, and how water pipelines can be seen as a temporal commons that is, a shared resource that connects current and future customers through long-lived infrastructure and long-term financing.

The LOF and COF models, when integrated in the portfolio optimization, explicitly quantify how much modeled risk is removed under a given budget, and they allow optional emphasis on equity-oriented surrogates (for example, renewal of legacy materials or segments serving more vulnerable consequence profiles). This makes it possible to evaluate not only which pipes are risky, but how alternative portfolios distribute risk reduction over space, time, and customer groups.

Framing the system as a temporal commons clarifies two linked issues. First is intergenerational fairness. Current customers fund renewals whose benefits accrue over decades, while future customers inherit whatever risk and debt remain. Second is spatial and social equity. Some neighborhoods and customer classes have historically shouldered more failures and disruptions than others. The metrics and portfolios developed here do not resolve these ethical questions, but they make them visible by putting numbers on

risk reduction per dollar, on who benefits from preventive investment, and on how trade-offs shift when equity is given more weight.

This aligns the work with broader theories of common-pool resources and preventive investment, while grounding those ideas in the practical language of utility CIPs and project lists.

7.1.4.5 Thematic contribution 4: Methodological architecture (Teacher–Student modeling, synthetic data, SETS framing)

A fourth contribution lies at the meta-level. This dissertation proposes and demonstrates an architectural pattern for modeling complex infrastructure systems.

First, fuzzy-logic teacher models are used to encode expert knowledge about mechanisms and consequences in a transparent way. These teachers generate synthetic but mechanism-informed datasets and provide a "target behavior" that can be learnt and adjusted. Second, student machine-learning models are trained to approximate the teachers and, where enough real data exist, to refine them. This teacher–student pattern balances interpretability and scalability in a way that is well suited to data-sparse, safety-relevant domains.

Third, the whole framework is embedded in a Socio–Ecological–Technical Systems (SETS) framing, which insists that social, environmental, and technical layers be modeled together. In practice, this shows up in the inclusion of environmental exposure variables, land use and customer-type information, and operational constraints in both LOF/COF and the portfolio model.

Because the architecture is modular, it is transferable. The same pattern with fuzzy teachers, student models, and EVV around them can be applied to other components of the source-to-tap system (for example, tanks, valves, pumps, meters, and wastewater pipes), and to integrated, cross-asset portfolios. The dissertation thus offers not only asset-specific models, but a general recipe for building and testing AI models for critical infrastructure.

7.1.5 Limitations: Where the framework should not be over-claimed

Finally, some limitations cut across all three studies and define where the framework should not be over-relied upon.

Data sparsity and bias remain central. Failure and inspection datasets are uneven across materials, diameters, and utilities; minor events are under-reported and many consequence dimensions rely on imperfect proxies. This limits how precisely LOF and COF

can be predicted and makes performance more certain in some cohorts than in others. Unmodeled extremes such as climate-driven shifts in loading, abrupt regulatory changes, or novel materials also challenge any model that is trained on historical data and implicitly assumes that past relationships will continue.

Institutional factors matter as much as algorithms. Implementing LOF, COF, and GA portfolios requires data integration, computational capacity, and willingness to adapt existing planning processes. Without this, there is a risk of models being used only superficially or sidelined when results are uncomfortable. The dissertation therefore treats institutional adoption, data standards, and governance as part of the future research agenda, not as externalities.

Table 7-3 summarizes these cross-cutting limitations and boundary conditions, together with their likely impacts and possible avenues for mitigation. These integrated limitations are not only caveats but are also pointers to future work. Each limitation suggests a concrete research question. For example, how to design better consequence data, how to update models under structural change, or how to embed EVV requirements into utility practice. Recognizing these boundaries is essential if the framework is to be used responsibly, and if subsequent studies are to build on it in a cumulative way.

Table 7-3: Major limitations and boundary conditions of the proposed framework, with implications and possible mitigations.

Theme	Limitation / boundary condition	Likely impact on results	Possible mitigation
Data & measurement	Failure and inspection data are sparse or biased for some cohorts (materials, diameters, utilities).	LOF and COF performance less certain in under-represented cohorts; possible misestimation of risk.	Targeted inspections; incremental data collection; ongoing re-training and local calibration; reliability flags.
Consequence data	Environmental, reputational, and equity-related impacts are poorly measured; proxies are coarse.	COF calibration weaker than LOF; some consequence pathways under-represented in portfolios.	Develop better consequence data standards; integrate qualitative judgement; use ranges and scenario analysis.
Model structure	LOF and COF assume relatively stable relationships between predictors and outcomes; some mechanisms are only indirectly represented.	Extrapolation to new climates, materials, or operating regimes is uncertain.	Periodic re-estimation; stress testing under alternative assumptions; explicit documentation of structural assumptions.
Temporal scope	Portfolio optimization focuses on a single planning horizon (for example, a 5-year CIP window).	Does not fully capture long-run learning, inter-temporal trade-offs, or path dependence.	Extend to multi-period optimization; embed in rolling planning processes; incorporate learning about failure processes.
Institutional adoption	Implementation requires data integration, computational capacity, and organizational willingness to adapt decision processes.	Risk that models are used superficially or not at all; potential mismatch with local politics and regulation.	Co-design with utilities; training and capacity building; simple default settings; governance for model use and oversight.

7.2 Conclusions

This dissertation set out to show that explainable, validated AI models can support risk-based renewal of water pipelines in a way that is both scientifically credible and

operationally usable. Rather than offering isolated models, the work develops a coherent architecture that spans LOF, COF and budget-constrained portfolio optimization, embedded in an explicit program of Evaluation, Verification, and Validation (EVV) and a broader Socio–Ecological–Technical Systems (SETS) framing. This section summarizes what the dissertation establishes at a high level, its scientific significance, its scope of validity, and how it relates to the broader theme of investing in prevention rather than cleaning up failure.

Risk modeling: The first conclusion is that it is feasible to build explainable, validated LOF and COF models for buried water pipelines that clearly outperform age-only heuristics and can generalize across multiple utilities, within known limits. The LOF models demonstrate that mechanism-oriented feature design, combined with teacher–student learning, can capture structural, hydraulic, and environmental drivers in a way that matches independent inspection and failure data. The COF models show that consequence can be decomposed into economic, environmental, service, and operational components without collapsing everything into a single opaque score. Together, these results indicate that risk $\approx$ LOF $\times$ COF can be operationalized for pipelines in a manner that is traceable, auditable, and empirically grounded.

Decision support: A second conclusion is that risk-informed, constraint-aware portfolios can materially improve the use of limited renewal budgets compared with rank-only strategies. Within realistic candidate sets and budget envelopes, Genetic Algorithm (GA) portfolios built from LOF and COF indices capture more modeled risk per unit cost and offer better packaging and equity surrogates than simple heuristics. Importantly, these improvements are not solely numerical. When utility planners are shown anonymized options, their preferences often align with GA portfolios or minor variants of them. This suggests that optimization can be integrated into decision support in a way that respects practical constraints and local judgement rather than replacing them.

Validation: A third conclusion is that multi-layer EVV is both necessary and practical for AI models in critical infrastructure. Internal metrics alone are not sufficient to justify deployment. The dissertation shows that an EVV ladder combining internal evaluation, cross-utility verification, and validation against independent inspections, failure data, and expert judgement can be implemented with real utilities and real data. While the evidence is necessarily incomplete, it is strong enough to demonstrate that this style of EVV is tractable and significantly improves the transparency and credibility of

AI models in the water sector, specially considering the fast pace at which data collection programs are improving in terms of volume, veracity and variety.

Governance and equity: Finally, the work supports a reframing of renewal planning as a temporal commons and equity problem, not just an engineering optimization. The pipeline network and its financing capacity connect current and future customers. In other words, decisions about where and when to renew determine who bears the risk and who benefits from preventive investment. By making risk reduction per dollar and equity surrogates explicit, the framework provides a basis for discussing intergenerational fairness and spatial disparities using quantitative tools rather than intuition alone. The conclusion is not that the model resolves questions of fairness, but that it makes them visible and tractable for governance.

7.2.1 Merits and scientific significance

The scientific contribution of this dissertation is not a single model, but a methodology for risk-based renewal prioritization. At the core is a three-layer architecture:

1. **Mechanism-aware risk modeling:** LOF and COF are constructed from explicit, mechanism-oriented features and fuzzy teacher models that encode expert

knowledge about deterioration and consequence pathways. This avoids over-reliance on age and diameter alone and resists purely data-driven shortcuts that might fit historical correlations but violate physical or institutional understanding.

2. **Teacher–student learning for infrastructure:** Student ML models are trained to emulate the fuzzy teachers and refined with real data where available. This pattern offers a way to move from qualitative expert judgement to scalable, quantitative prediction in domains where physics models are incomplete and data are limited. The work demonstrates that this is possible for heterogeneous, buried pipeline networks and that the resulting student models retain interpretability through their link to the teacher.
3. **Risk-aware, constraint-aware optimization:** The renewal prioritization layer uses LOF and COF outputs, combined with simple constraints and optional equity surrogates, to search over combinations of projects rather than ranking segments in isolation. This brings renewal planning closer to the optimization practices used in other safety-critical domains (for example, reliability-centered maintenance in aviation), while remaining implementable with standard tools.

Methodologically, the dissertation also offers a template for documenting EVV in a way that utilities and regulators can use. The experiments are not exhaustive, but they are layered, cross-utility, and anchored in inspections and failures as well as expert judgement. That combination is rare in infrastructure AI and can serve as a reference for future model developers, funders, and oversight bodies who need to decide when a model has earned enough trust to influence capital planning.

Taken together, these elements raise the bar for how risk-based renewal models are designed, evaluated, and presented. They show that explainable, validated AI for infrastructure is not only desirable in principle but achievable in practice, provided that modeling choices, data limitations, and validation evidence are explicitly documented.

7.2.2 Limitations and scope of validity

The conclusions above are not universal and they hold most strongly within a defined scope of data, models, and institutions.

From a data and measurement perspective, the framework is best supported in systems that resemble the participating water utilities in the US that are, moderate to large systems with multi-decadal pipe inventories, at least partial work-order histories, some inspection data (for example, wall-thickness or wire-break surveys), and basic

consequence documentation. In these settings, the LOF and COF models can be calibrated, and their predictions are anchored to observed behavior. In small systems, in systems with extreme data gaps, or where failure and consequence reporting is minimal, model performance is more uncertain and the models should be used, at least initially, in a more exploratory or scenario-based role.

From a model structure perspective, LOF and COF assume that the relationships learned from historical data and expert judgement remain approximately valid under future conditions. This is a reasonable first approximation but may be violated by large shifts in climate, land use, technology, or regulation. The fuzzy teachers can be edited, and the student models can be retrained, but the current instantiations should not be extrapolated uncritically to contexts that differ sharply from the training environments (for example, very different soil chemistries, radically different operating pressures, or emerging materials with little history).

From a validation scope perspective, the external validation datasets are rich but not exhaustive. They cover specific materials, diameters, and geographies, and the consequence validation focuses on types of events that utilities happened to document in usable ways. As a result, the strongest claims can be made for cohorts and contexts like those

seen in the validation experiments. However, outside that envelope, the model outputs should be treated as structured hypotheses to be tested locally rather than as definitive risk estimates.

Finally, the institutional scope matters. The framework assumes utilities that are willing and able to integrate new models into their capital planning processes, to share data under appropriate agreements, and to participate in EVV activities. In organizations where data integration, computational capacity, or governance for model use is limited, the immediate impact may be more modest. In those cases, the framework may function initially as a diagnostic and learning tool rather than a direct driver of capital decisions.

Being explicit about these boundaries is essential. It clarifies where the results can be applied with confidence, where local adaptation is needed, and where additional data collection and validation should be prioritized.

7.2.3 Reflection on prevention versus clean-up

At a deeper level, this dissertation participates in a long scientific conversation about the value of prevention compared with the cost of cleaning up failure. Public health, disaster risk reduction, safety engineering, and financial regulation have all converged on

the same insight that is, paying a modest, steady cost to reduce risk is almost always cheaper and less harmful than absorbing repeated large shocks.

The logic of "pay a little now to avoid paying a lot later" can be visualized as a shift in when and how renewal occurs. Figure 7-1 illustrates how, in the absence of data, level of service is difficult to determine and can remain unchecked until it drops below a minimum threshold, at which point utilities are forced into costly emergency replacement. As condition and consequence data become available and are integrated into planning, interventions move earlier in the deterioration trajectory from corrective renewal after service has degraded to genuinely preventive renewal keeping the system in a higher service band and reducing disruptive, high-cost failures.

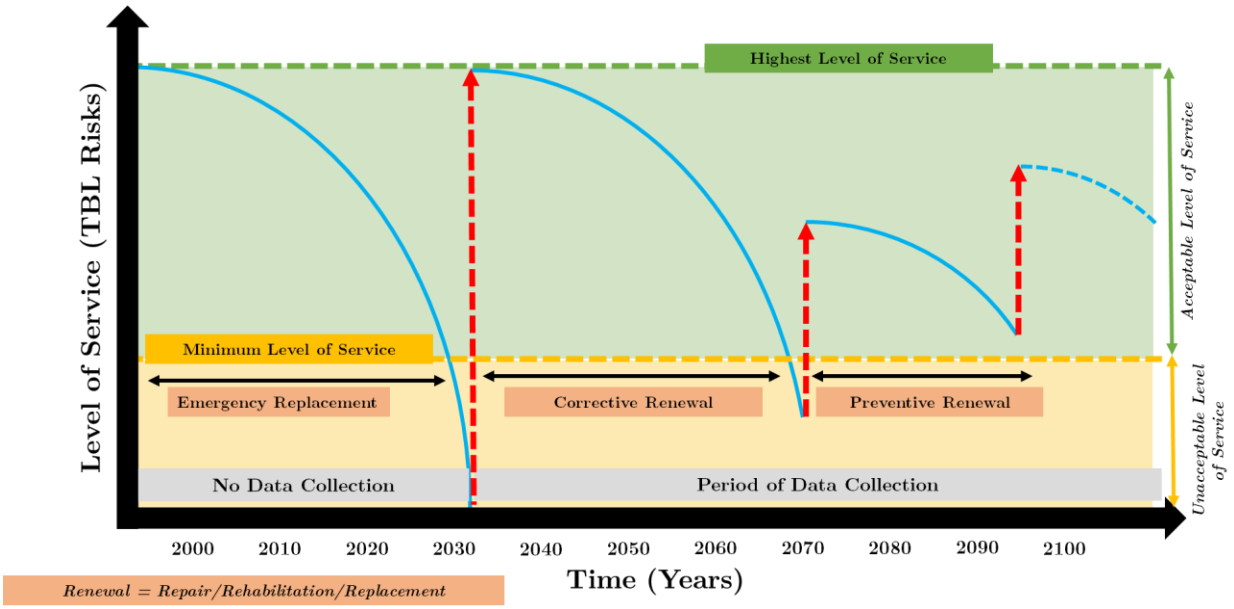

Figure 7-1: Conceptual relationship between data collection, level of service, and timing of renewal. In the absence of data, service deteriorates until it falls below a minimum acceptable level, forcing emergency replacement. As utilities collect and use condition and consequences data, renewal actions can shift from late corrective work toward earlier preventive renewal, keeping the system in a higher level-of-service band.

Water pipeline infrastructure therefore fits this pattern very well. Renewals and proactive interventions require up-front capital and institutional effort, while emergency breaks impose acute financial losses, service disruptions, and often invisible social and environmental harms. The LOF, COF, and portfolio models developed in this dissertation do not change that basic logic but they make it quantifiable and visible. They estimate how much risk and potential harm are reduced when specific segments are addressed, and how those benefits vary with budget and with where attention is focused.

By doing so, the work helps move water pipeline management closer to the preventive ethos already established in other domains. It does not argue that all failures can be prevented, or that emergency response is unimportant. Rather, it provides tools for making deliberate, evidence-based choices about how much prevention is enough, given limited resources and competing priorities. In that sense, the dissertation's broader conclusion is that water utilities can and should treat risk-based renewal as a central preventive function, on par with public health or disaster risk reduction, rather than as a residual engineering task.

7.3 Recommendations and future work

The results in this dissertation are a starting point rather than an endpoint. This section outlines an agenda for extending and deepening the work along five themes: LOF modeling, COF and consequence measurement, portfolio optimization and decision science, scaling beyond pipes, and data/standards and institutional practice. The aim is to be both practical and intellectually ambitious, pointing to research questions and implementation steps that can be pursued by utilities, researchers, and standard-setting bodies.

7.3.1 Future work on LOF modeling

Several directions can enhance the LOF models and strengthen their calibration and transferability:

Richer physics and lifecycle based features: Current LOF inputs include key structural, environmental, and hydraulic surrogates, but many systems now have Supervisory Control and Data Acquisition (SCADA) data that capture pressure transients, flow variability, and pump operations at fine time scales. Additionally, many data prior to commissioning (like manufacturing methods, structural voids etc.) and after decommissioning (like forensic data) can be critical to the water pipeline performance. Incorporating these signals, in a way that remains interpretable, could improve detection of stress regimes that drive failures but are not captured by static averages. Climate-related variables (for example, freeze–thaw cycles, projected changes in soil moisture) could also be integrated to anticipate emerging patterns.

Adaptive and online learning: As utilities collect more failure and inspection data, LOF models should be able to update incrementally rather than being retrained from scratch. Online learning methods and Bayesian updating offer ways to adjust model

parameters as new information arrives, while keeping the fuzzy teachers as a stable reference. Careful design is needed to avoid "forgetting" rare but important failure modes.

Competing failure-mode analysis and temporal modeling: In reality, pipe failures rarely result from a single mechanism acting in isolation. Corrosion, external loading, pressure transients, construction defects, and environmental exposures interact over time, and the observed failure is the outcome of competing and combined failure modes. A natural extension of the present LOF framework is to adopt competing failure-mode or multi-state hazard models as a precursor to risk analysis, explicitly representing the probability that different mechanisms "win" or co-occur along a pipe's deterioration trajectory. Doing so will require time-dependent, in-situ data such as repeated condition assessments, high-frequency pressure and flow records, and temporally resolved environmental indicators and temporal models that track performance evolution rather than treating LOF as static. This line of work could illuminate how specific stressor combinations (for example, high pressure + corrosive soils + traffic loading) interact, reveal mode-switching behavior under changing operations or climate, and ultimately yield more accurate and mechanistically grounded LOF estimates than single-mode or cross-sectional models.

Cross-utility transfer learning and domain adaptation: Transfer learning techniques can help exploit similarities across utilities while respecting local differences in materials, construction practices, and environments. Domain adaptation methods could be used to adjust models trained on data-rich utilities to perform better on data-poor utilities, with explicit quantification of the resulting uncertainty. This aligns with how many utilities already share design standards and operational practices, but would add a formal, data-driven layer.

7.3.2 Future work on COF and consequence measurement

Improving COF requires both better measurement of consequences and better modeling of their distribution:

Environmental and equity impact measurement: Many environmental and equity consequences are only partially observed. Future work should develop and test more granular indicators of exposure. For example, combining land use, floodplains, and socio-demographic information to quantify which populations and ecosystems are at risk from given pipe failures. This will require collaboration with environmental scientists and social scientists and may involve mixed-methods studies that combine quantitative data with qualitative insights from communities.

Harmonizing cost and consequence reporting: Economic costs and service disruptions are recorded very differently across utilities. Establishing minimal data and metadata standards for reporting break events, repair actions, and service impacts would significantly improve COF calibration. This could be done through professional societies and regulatory guidance and would benefit not only this framework but any future risk models.

Expanding non-monetary consequence metrics: While dollars are a useful common unit, many important consequences are better expressed in terms of customer-hours of interruption, service reliability indices, or even health risk proxies in contamination scenarios. Future COF models should therefore explore multi-metric consequence representations, where monetary and non-monetary indicators are modeled jointly and decision makers can see trade-offs rather than relying on a single composite number.

7.3.3 Future work on portfolio optimization and decision science

The current portfolio model focuses on a single planning horizon and relatively simple scalarization of objectives. Several extensions can make it more realistic and theoretically grounded:

Multi-period optimization with learning: Renewal decisions today change both the physical state of the system and the information available tomorrow. Multi-period models that explicitly represent state evolution, learning about failure processes, and rolling budgets could better capture these dynamics. This moves towards Markov decision processes or approximate dynamic programming, tailored to the constraints of utility planning.

Behavioral and institutional constraints: Real decisions are shaped by risk aversion, political cycles, regulatory obligations, and organizational culture. Future portfolio models could include explicit representations of such constraints. For example, caps on year-to-year budget variation, preferences for geographically balanced work, or aversion to rare but catastrophic failures. This would bring the models closer to decision science and behavioral economics and help explain why some mathematically "optimal" portfolios are unacceptable in practice.

Equity and fairness metrics: Beyond simple equity surrogates, future work could incorporate formal fairness constraints (for example, max–min criteria or constraints on disparity in risk reduction across customer groups). This would allow utilities to explore

portfolios that not only maximize aggregate risk reduction but also meet explicit equity goals, and to understand the marginal cost of making portfolios more equitable.

7.3.4 Scaling beyond pipes: source-to-tap and other infrastructures

The modeling and EVV architecture demonstrated here is not limited to distribution mains. Extending it along two axes is a natural next step:

Other water assets: LOF and COF models, with fuzzy teachers and student learners, can be developed for storage tanks, valves, pumps, meters, and raw-water conveyance assets. Each asset type has its own failure modes and consequences, but the teacher–student–EVV pattern is reusable. Once multiple components are modeled, cross-asset portfolios can be considered. For example, comparing the benefit of renewing mains with upgrading key valves or storage facilities.

Other infrastructures and integrated systems: The same logic applies to wastewater, stormwater, and other urban infrastructure systems, and eventually to "system-of-systems" portfolios that couple water with energy, transportation, or digital infrastructure. While such extensions will require asset-specific expertise, the basic architecture

of mechanism-informed teachers, ML students, EVV, and portfolio optimization provides a starting point for building consistent, cross-sector risk models.

7.3.5 Data, standards, and institutional recommendations

Technical advances will only matter if institutions can absorb and use them. Several practical recommendations follow from this work:

For utilities: Utilities should move towards minimum data standards that support risk modeling. This includes consistent pipe inventories, geospatial referencing, basic environmental layers, standardized break and repair records, and, where feasible, inspection and SCADA data. Integrating LOF/COF and portfolio tools into existing Capital Improvement Program (CIP) processes can begin with simple use cases such as stress-testing current plans or exploring "what if" scenarios rather than full automation. Clear internal governance for model use, including documentation, versioning, and periodic EVV, will be essential to avoid both over-trust and under-use of AI tools.

For researchers and standard-setting bodies: Researchers should publish not only models but also EVV protocols, including negative results and contexts where models perform poorly, to enable cumulative learning. Professional societies and funding agencies

can facilitate shared benchmarks and multi-utility testbeds, where different approaches can be compared under similar data and validation conditions. Over time, this could lead to de facto or formal standards for EVV of AI models in water infrastructure, making it easier for utilities and regulators to evaluate claims and decide when a tool is ready for use in planning.

7.4 Closing remarks

This dissertation has argued, and empirically illustrated, that AI-driven, validated, and explainable risk models can move water pipeline management from a predominantly reactive paradigm to a more proactive, preventive one provided they are paired with adequate data, thoughtful validation, and institutional willingness to act on their insights. The models developed here are imperfect and incomplete, but they show that it is possible to integrate mechanism-informed LOF and COF, constraint-aware portfolio optimization, and multi-layer EVV into a coherent framework that respects both scientific rigor and practical constraints.

The broader hope is that this work contributes to a shift in how infrastructure risk is understood and governed that is, from hidden deterioration managed by emergency response, to an explicit, shared, and continually updated picture of where the system is fragile, who is most exposed, and how preventive investments can be targeted for the greatest long-term benefit.

References

1. Abowitz, D. A., & Toole, T. M. (2010). Mixed method research: Fundamental issues of design, validity, and reliability in construction research. *Journal of construction engineering and management*, 136(1), 108-116.
2. Agrawal, C., Misra, S., Sinha, S., and Vasudevan, V. "Development of consequence of failure index for water and wastewater pipelines." *Proc., Proceedings of the Institution of Civil Engineers-Municipal Engineer*, Thomas Telford Ltd, 1-12.
3. Angkasuwansiri, T., & Sinha, S. K. (2013). Comprehensive list of parameters affecting wastewater pipe performance. *Technol. Interface Int. J*, *13*(2), 68-79.
4. AS 55-1:2008 (2008) Asset management. Specification for the optimized management of physical assets. *BSI: United Kingdom*, 2008A Maintenance Management Framework Based on PAS 5537
5. ASCE (2025). 2025 Drinking Water Report Card for America's Infrastructure. *American Society of Civil Engineers*, Reston, VA.
6. Aven, T. (2016). "Risk assessment and risk management: Review of recent advances on their foundation." *European Journal of Operational Research*, 253(1), 1-13.

7. AWWA C150/C151. Ductile-Iron Pipe, Centrifugally Cast; Thickness Design. *American Water Works Association* (latest revisions).

8. AWWA C301/C303. Prestressed Concrete Cylinder Pipe; Bar-Wrapped Concrete Cylinder Pipe. *American Water Works Association* (latest revisions).

9. AWWA Standards (various editions as cited in text): C301 Prestressed Concrete Cylinder Pipe; C304 Design of PCCP; C300 Reinforced Concrete Cylinder Pipe; C302 Reinforced Concrete Non-Cylinder Pipe; C303 Bar-Wrapped Concrete Cylinder Pipe.

10. AWWA (2010). "AWWA J100-10 (R13): Risk and Resilience Management of Water and Wastewater Systems (RAMCAP)." Denver: *American Water Works Association*.

11. AWWA J100 (2010/2018). Risk and Resilience Management of Water and Wastewater Systems (RAMCAP). *American Water Works Association*, Denver, CO.

12. AWWA M28 (2014). Rehabilitation of Water Mains, 4th ed. *American Water Works Association*, Denver, CO.

13. AWWA. (2016). M23: PVC Pipe Design and Installation (3rd ed.). *American Water Works Association*.

14. AWWA. (2016). Manual M9: Concrete Pressure Pipe (latest ed.). *American Water Works Association*.

15. AWWA. (2017). M55: PE Pipe Design and Installation (2nd ed.). *American Water Works Association*.

16. AWWA M77 (2018). Condition Assessment of Water Mains. *American Water Works Association*, Denver, CO.

17. AWWA. (2018). M41: Ductile-Iron Pipe and Fittings (5th ed.). *American Water Works Association*.

18. AWWA. (2020/2022). M11: Steel Pipe A Guide for Design and Installation (latest ed.). *American Water Works Association*.

19. AWWA (2025). State of the Water Industry Report 2025. *American Water Works Association*, Denver, CO. https://www.awwa.org/Professional-Development/Utility-Managers/State-of-the-Water-Industry

20. AwwaRF (2008). State of the Science on PCCP Performance and Design Conservatism. *American Water Works Association*, Denver, CO.

21. Belton, V., & Stewart, T. (2012). *Multiple criteria decision analysis: an integrated approach*. Springer Science & Business Media.

22. Birge, J. R., & Louveaux, F. (1997). *Introduction to stochastic programming*. New York, NY: Springer New York.

23. Boccara, N. (2010). Power-law distributions. In *Modeling Complex Systems* (pp. 371-433). New York, NY: Springer New York.

24. Boudreau, K., Robinson, M., & Farooqi, Z. (2022). IPCC Sixth assessment report: climate change 2021: the physical science basis summary for policymakers. *Canadian Journal of Emergency Management*, *2*(1).

25. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

26. Burn, S., Davis, P., & Schiller, T. L. (2006). *Long-term performance prediction for PVC pipes*. AWWA Research Foundation.

27. Burns, A., & Davis, R. I. (2017). A survey of research into mixed criticality systems. *ACM Computing Surveys (CSUR)*, *50*(6), 1-37.

28. Chester, M. V., & Allenby, B. (2019). Infrastructure as a wicked complex process. *Elem Sci Anth*, *7*, 21.

29. Coleman, Hugh & Committee Members. (2009). V&V 20-2009: Standard for Verification and Validation in Computational Fluid Dynamics. *American Society of Mechanical Engineers*, New York.

30. Corrpro & DIPRA (2005). Design Decision Model (DDM) for DI corrosion control. (J-*AWWA*, June 2005; program documentation).

31. Cromwell, J. E. (2002). Costs of infrastructure failure, *American Water Works Association*.

32. Davis, C. A. (2018). Creating a seismic resilient pipe network for Los Angeles. In *Pipelines 2018* (pp. 425-432). Reston, VA: American Society of Civil Engineers.

33. Davis, P., Burn, S., Moglia, M., & Gould, S. (2007). A physical probabilistic model to predict failure rates in buried PVC pipelines. *Reliability Engineering & System Safety*, *92*(9), 1258-1266.

34. Deb, K. (2001). Multi-objective optimization using evolutionary algorithms John Wiley & Sons. *Inc., New York, NY*.

35. Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. A. M. T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE transactions on evolutionary computation*, 6(2), 182-197.

36. Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10), 78–87.

37. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

38. EPA (2018). America's Water Infrastructure Act of 2018 (AWIA). https://www.epa.gov/ground-water-and-drinking-water/americas-water-infrastructure-act-2018-awia (Accessed on 11/24/2025)
39. FHWA (2001). Implementation of GIS-Based Highway Safety Analyses: Bridging the Gap. https://www.fhwa.dot.gov/publications/research/safety/1039.pdf
40. Folkman, S. (2012; 2018). Water Main Break Rates in the USA and Canada: A comprehensive study. *Utah State University*.
41. Gaewski, P. E., & Blaha, F. J. (2007, April). Analysis of total cost of large diameter pipe failures. In *Proc. AWWA Research Symposium Distribution Systems: The Next Frontier, Reno, Nev.*
42. Ge, S., & Sinha, S. (2014). Failure analysis, condition assessment technologies, and performance prediction of prestressed-concrete cylinder pipe: State-of-the-art literature review. *Journal of Performance of Constructed Facilities*, *28*(3), 618-628.
43. Ge, S., & Sinha, S. (2015). Effect of mortar coating's bond quality on the structural integrity of prestressed concrete cylinder pipe with broken wires. *Journal of Materials Science Research*, *4*(3), 59.

44. Geisbush, J., & Ariaratnam, S. T. (2023). Reliability centered maintenance (RCM): literature review of current industry state of practice. *Journal of Quality in Maintenance Engineering*, *29*(2), 313-337.

45. Giglio, J. M., Friar, J. H., & Crittenden, W. F. (2018). Integrating lifecycle asset management in the public sector. *Business Horizons*, 61(4), 511-519.

46. Gilboa, I. (2009). Theory of decision under uncertainty (No. 45). *Cambridge university press.*

47. Halfawy, M. R. (2008). Integration of municipal infrastructure asset management processes: challenges and solutions. *Journal of Computing in Civil Engineering*, 22(3), 216-229.

48. Hastie, T., Tibshirani, R., and Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction. 2nd Ed., *Springer*, New York.

49. Hodkiewicz, M. R. (2014, November). The development of ISO 55000 series standards. In *Engineering Asset Management-Systems, Professional Practices and Certification: Proceedings of the 8th World Congress on Engineering Asset Management (WCEAM 2013) & the 3rd International Conference on Utility Management & Safety (ICUMAS)* (pp. 427-438). Cham: Springer International Publishing.

50. Holland, J. H. (1975). Adaptation in Natural and Artificial Systems. *University of Michigan Press google scholar*, 2, 29-41.

51. Homma, T., and Saltelli, A. (1996). "Importance measures in global sensitivity analysis of nonlinear models." *Reliability Engineering & System Safety*, 52(1), 1-17.

52. ISO 55000 (2014). Asset Management-Overview, Principles and Terminology. *International Organization for Standardization*, Geneva, Switzerland.

53. IWA (2004). The Bonn charter for safe drinking water, *IWA*, London, UK.

54. Kaplan, S., and Garrick, B. J. (1981). "On the quantitative definition of risk." *Risk analysis*, 1(1), 11-27.

55. Kleiner, Y., & Rajani, B. (2001). Comprehensive review of structural deterioration of water mains. *Urban Water*, 3(3), 131–150.

56. Kombo Mpindou, G. O. M., Escuder Bueno, I., & Chordà Ramón, E. (2022). Risk analysis methods of water supply systems: comprehensive review from source to tap. *Applied Water Science*, *12*(4), 56.

57. Knight, M., Tighe, S., & Adedapo, A. (2004, September). Trenchless installations preserve pavement integrity. In *Proc., Soils and Materials Session of the 2004 Annual Conf. of the Transportation Association of Canada* (pp. 1-15).

58. Kohavi, R., & Thomke, S. (2017). The surprising power of online experiments. *Harvard business review*, 95(5), 74-82.

59. Kola, R. (2010). *Development of predictability and condition assessability indices for PCCP water mains* (Doctoral dissertation, Virginia Tech).

60. Le Gat, Y. (2008). A stochastic model for the lifetime of water pipes. *Urban Water Journal*, 5(4), 281–294.

61. Mackellar, A., & Pearson, D. (2003). *UK National Mains Failure Database summary.*

62. Meadows, D. H. (2008). Thinking in Systems: A Primer. *Chelsea Green Publishing*, White River Junction, VT.

63. Muhlbauer, W. K. (2004). *Pipeline risk management manual: ideas, techniques, and resources.* Gulf Professional Publishing.

64. Nemhauser, G., & Wolsey, L. (1988). General algorithms. *Integer and Combinatorial Optimization*, 349-382.

65. Wolsey, L. A., & Nemhauser, G. L. (1999). *Integer and combinatorial optimization.* John Wiley & Sons.

66. Najafi, M., Habibian, A., Sever, V. F., Divyashree, D., & Jain, A. Exploring Use of Large-Diameter HDPE Pipe for Water Main Applications. In *Pipelines 2015* (pp. 530-541).

67. National Research Council, Division on Engineering, Physical Sciences, National Materials Advisory Board, & Committee on the Review of the Bureau of Reclamation's Corrosion Prevention Standards for Ductile Iron Pipe. (2009). *Review of the Bureau of Reclamation's Corrosion Prevention Standards for Ductile Iron Pipe*. National Academies Press.

68. Oreskes, N., Shrader-Frechette, K., & Belitz, K. (1994). Verification, validation, and confirmation of numerical models in the earth sciences. *Science, 263*(5147), 641–646.

69. Ostrom, E. (2009). A general framework for analyzing sustainability of social-ecological systems. *Science*, *325*(5939), 419-422.

70. Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E. and Chou, R., (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *bmj, 372*.

71. Rajani, B., & Kleiner, Y. (2001). Comprehensive review of structural deterioration of water mains: physically based models. *Urban Water, 3(3)*, 151–164.

72. Raucher, B. (2005). "The Value of Water: What It Means, Why It's Important, and How Water Utility Managers Can Use It." *Journal - American Water Works Association, 97(4)*, 90-98.

73. Raucher, R. S. (2017). Managing infrastructure risk: The consequence of failure for buried assets, *Water Research Foundation*.

74. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144).

75. Rizzo, P. (2010). Water and wastewater pipe nondestructive evaluation and health monitoring: A review. *Advances in Civil Engineering, 2010*(1), 818597.

76. Saaty, T. L. (1980). The analytic hierarchy process (AHP). *The Journal of the Operational Research Society*, 41(11), 1073-1076.

77. Saltelli, A. (2002). "Sensitivity Analysis for Importance Assessment." *Risk Analysis*, 22(3), 579-590.

78. S. Argent, "BSI-PAS 55 - The Benefits for Utilities - A Regulator's View," 2007 IET Seminar on PAS 55 and Measuring Asset Management, London, UK, 2007, pp. 3-14.

79. Scheffer, M., Barrett, S., Carpenter, S.R., Folke, C., Green, A.J., Holmgren, M., Hughes, T.P., Kosten, S., Van de Leemput, I.A., Nepstad, D.C. and van Nes, E.H., (2015). Creating a safe operating space for iconic ecosystems. *Science*, *347*(6228), pp.1317-1319.

80. Scholten, L., Scheidegger, A., Reichert, P., Mauer, M., & Lienert, J. (2014). Strategic rehabilitation planning of piped water networks using multi-criteria decision analysis. *Water research*, 49, 124-143.

81. Scott, W. R. (2013). Institutions and organizations: Ideas, interests, and identities. *Sage publications*.

82. Searle, J. R. (1995). The construction of social reality. *Simon and Schuster*.

83. Senouci, A., El-Abbasy, M. S., & Zayed, T. (2014). Fuzzy-based model for predicting failure of oil pipelines. Journal of Infrastructure Systems, 20(4), 04014018.

84. Sinha, S. (2021). Collection and Compilation of Water Pipeline Field Performance Data. *Sustainable Water Infrastructure Management Center*, Blacksburg, VA.

85. Slovic P. Perception of risk. Science. 1987 Apr 17;236(4799):280-5. doi: 10.1126/science.3563507. PMID: 3563507.

86. St. Clair, A. M., & Sinha, S. (2012). State-of-the-technology review on water pipe condition, deterioration and failure rate prediction models. *Urban Water Journal*, *9*(2), 85-112.

87. Clair, A. M. S. (2013). *Development of a Novel Performance Index and a Performance Prediction Model for Metallic Drinking Water Pipelines* (Doctoral dissertation, Virginia Polytechnic Institute and State University).

88. The White House. (2013, Feb 12). Presidential Policy Directive Critical Infrastructure Security and Resilience (PPD-21).

89. The White House. (2024, April 30). National Security Memorandum on Critical Infrastructure Security and Resilience. *NSM-22, April, 30.* https://biden-whitehouse.archives.gov/briefing-room/presidential-actions/2024/04/30/national-security-memorandum-on-critical-infrastructure-security-and-resilience/.

90. Tomer, A., & Kane, J. (2018). Renewing the Water Workforce: Improving Water Infrastructure and Creating a Pipeline to Opportunity. *Brookings Institution, Metropolitan Policy Program.*

91. National Research Council, Division on Engineering, Physical Sciences, National Materials Advisory Board, & Committee on the Review of the Bureau of Reclamation's Corrosion Prevention Standards for Ductile Iron Pipe. (2009). *Review of the Bureau of Reclamation's Corrosion Prevention Standards for Ductile Iron Pipe*. National Academies Press.

92. USEPA (2021). Drinking Water Infrastructure Needs Survey and Assessment: Sixth Report to Congress. *U.S. Environmental Protection Agency, Office of Water*, Washington, DC.

93. Uslu, B., Sinha, S. K., Ge, S., & Yadav, R. (2013). A validation and verification framework for robust drinking water pipeline model prediction models. In *Pipelines 2013: Pipelines and Trenchless Construction and Renewals—A Global Perspective* (pp. 1246-1256).

94. Uslu, B. (2017). Development of protocols and methods for predicting the remaining economic life of wastewater pipe infrastructure assets (Doctoral dissertation, Virginia Tech).

95. Vishwakarma, A., & Sinha, S. (2023). Consequence of failure modeling for water pipeline infrastructure using a hierarchical ensemble fuzzy inference system. *Journal of Infrastructure Systems*, 29(1), 04022

96. Von Neumann, J., & Morgenstern, O. (2007). Theory of games and economic behavior: 60th anniversary commemorative edition. In *Theory of games and economic behavior*. Princeton university press.

97. Water Research Foundation (2008). Condition Assessment Strategies and Protocols for Water and Wastewater Utility Assets.

98. WHO (2004). Guidelines for drinking-water quality (Vol. 1), *World Health Organization*.

99. World Health Organization. (2017). Global status report on water safety plans: a review of proactive risk assessment and risk management practices to ensure the safety of drinking-water. *Global status report on water safety plans: a review of proactive risk assessment and risk management practices to ensure the safety of drinking-water*.

100. Wolman, A. (1921). "Sanitary Inspection—A Review." *American Journal of Public Health*, 11(7), 598-605.

101. Zimmerman, R., & Faris, C. (2010). Infrastructure impacts and adaptation challenges. *Annals of the New York Academy of Sciences*, 1196(1).

Appendix A: Glossary of Terms

1. AADT: Average Annual Daily Traffic. A measure of typical traffic volume on a road segment; used as a surrogate for external loading and access constraints on buried pipes.
2. AC: Asbestos Cement. Legacy pipe material that is no longer installed but remains in many systems; modeled as a distinct material cohort with specific deterioration and failure behavior.
3. AHP: Analytic Hierarchy Process. A structured multi-criteria decision method used in some prior water-asset renewal studies; appears mainly in the literature review as a contrast to the proposed optimization framework.
4. AI: Artificial Intelligence. Broad term for computational methods that learn patterns from data; here it mainly refers to machine-learning student models trained to approximate fuzzy-logic teacher models for LOF and COF.
5. AMP: Asset Management Plan. Strategic document that describes how a utility will maintain, renew, and finance its infrastructure over time.
6. ASCE: American Society of Civil Engineers. Professional society whose report cards and standards (e.g., pipeline design standards) are used as context and data sources.
7. ASTM: American Society for Testing and Materials. Standards organization whose pipe material and testing standards (e.g., C300, C302, C303) inform material classification and performance assumptions.

8. AWIA: America's Water Infrastructure Act of 2018. U.S. legislation that, among other things, requires risk and resilience assessments for drinking-water systems.
9. AWWA: American Water Works Association. Professional association whose manuals (e.g., M28, M32, M77) and standards are frequently cited for design, rehabilitation, and asset-management guidance.
10. BCR: Benefit–Cost Ratio. Economic metric comparing the benefits of an intervention with its costs; used conceptually when discussing the value of preventive renewal.
11. Capital Improvement Plan: See CIP. Multi-year planning document that schedules and budgets infrastructure projects; the portfolio model is designed to provide evidence-based project lists for the CIP.
12. CI: Cast Iron. Legacy ferrous pipe material; in this dissertation it is split into pit and spun manufacturing types and diameter bands for LOF modeling and validation.
13. CIP: Capital Improvement Plan (or Program). The multi-year plan that specifies which infrastructure projects a utility intends to fund; the portfolio model is designed to support CIP formation.
14. COF: Consequence of Failure. A 0–5 index representing the severity of impacts if a pipe segment fails, decomposed into economic, environmental, social/service, and operational sub-indices.

15. Customer-hours: Metric for service disruption equal to the number of customers affected multiplied by the duration of the outage; used as a consequence proxy and for comparing impacts of failures or portfolios.
16. DDM: Decision-Directed Modeling. Term used to emphasize that model structure and metrics are chosen to support specific decision contexts rather than purely predictive accuracy.
17. DI: Ductile Iron. Ferrous pipe material that succeeds cast iron; modeled in its own material cohorts for LOF analysis and validation.
18. DWSRF: Drinking Water State Revolving Fund. U.S. financing program that provides low-interest loans for drinking-water infrastructure; cited in the policy and funding context.
19. EM: Electromagnetic (inspection). Generic term for non-destructive inspection techniques (e.g., broadband EM, remote field testing) used to estimate wall thickness or wire breaks in metallic and PCCP pipes.
20. EPA: United States Environmental Protection Agency (USEPA). Federal environmental regulator whose rules (e.g., SDWA) and funding programs (e.g., DWSRF) provide the policy backdrop for this work.
21. Evaluation–Verification–Validation: See EVV. Structured approach to building trust in models: evaluation uses internal metrics; verification checks behavior on held-out but similar data; validation tests against independent measurements and expert judgement.
22. EVV: Evaluation–Verification–Validation. Three-layer framework for assessing AI models: evaluation uses internal performance metrics; verification checks that the model behaves as intended on unseen but similar data; validation tests against independent evidence and expert judgement, often under different conditions.

23. GA: Genetic Algorithm. Evolutionary optimization method used to construct renewal portfolios that maximize modeled risk reduction (and other objectives) subject to budget and delivery constraints.
24. Genetic Algorithm: See GA. Evolution-inspired search procedure that evolves candidate portfolios of renewal projects using selection, crossover, and mutation operators.
25. HDPE: High-Density Polyethylene. Plastic pipe material used for some distribution mains and services; included as a distinct material in LOF modeling.
26. IWA: International Water Association. Professional network referenced in the context of international best practices in asset management.
27. Legacy materials: Older pipe materials, such as cast iron, asbestos cement, or early PVC formulations, that are no longer installed but remain in service; they often have higher failure rates or uncertain long-term behavior.
28. LOF: Likelihood of Failure. A 0–5 banded index representing a pipe segment's propensity to fail over a planning horizon, derived from a fuzzy-logic teacher model and student machine-learning models.
29. LOF_GT: Ground-Truth LOF. LOF labels derived directly from inspection measurements (e.g., wall loss) or wire-break counts, used as validation targets for the LOF model.
30. M28: AWWA Manual M28, Rehabilitation of Water Mains. Guidance document on water main rehabilitation methods, cited in the renewal background.
31. MAUT: Multi-Attribute Utility Theory. Decision-analytic framework used in some prior water-asset prioritization studies; included here in the literature review as part of the MCDA family.

32. MCDA: Multi-Criteria Decision Analysis. Family of methods that combine multiple, often conflicting criteria into a structured decision; the dissertation's portfolio model is a quantitative, risk-focused instance of MCDA.
33. MCS: Mixed Criticality Systems. Term used to describe systems containing assets with widely varying consequences of failure; here, the water pipeline network is treated as a mixed criticality system.
34. Membership function: In fuzzy logic, a curve that maps a crisp input (e.g., wall-loss percentage) to a degree of membership in a linguistic term (e.g., "high corrosion"); membership functions underpin the fuzzy teacher models for LOF and COF.
35. MFL: Magnetic Flux Leakage. Non-destructive testing method for metallic pipes; referenced in the context of condition assessment techniques.
36. MLP: Multi-Layer Perceptron. A class of feed-forward neural networks; deep MLP architectures are used as student models for LOF and COF.
37. Moratorium conflict: Situation where planned renewal work conflicts with road-work moratoria or other external restrictions, making projects difficult or impossible to deliver in a given period; used as a constraint and evaluation metric for portfolios.
38. NSGA: Non-dominated Sorting Genetic Algorithm. Multi-objective extension of GA used conceptually or experimentally for exploring trade-offs among risk, cost, equity, and other objectives in portfolio optimization.
39. NPV: Net Present Value. Economic measure that discounts future costs and benefits to a common present value; used conceptually when discussing life-cycle economics of renewal.

40. PCCP: Prestressed Concrete Cylinder Pipe. Large-diameter pipe type susceptible to wire-break failures; used as a key cohort in LOF validation via wire-break data.
41. PE: Polyethylene (often High-Density Polyethylene in this work). Plastic pipe material used in distribution systems; modeled as a distinct cohort in LOF analysis.
42. Performance band: Discrete category of performance or risk (often 0–4 or 0–5) used to simplify communication and decision rules; LOF and COF are both expressed as banded indices.
43. PI: Performance Index. A 0–5 index representing overall pipe performance, with higher values indicating better condition; serves as an intermediate construct between raw features and LOF/COF bands in some models.
44. PI_cat: Performance Index Category. Discrete categorization (0–4 or 0–5) of the continuous performance index into banded states used as model targets.
45. PVC: Polyvinyl Chloride. Common plastic distribution pipe material; a major cohort in the LOF modeling and validation experiments.
46. Risk: In this dissertation, defined conceptually as a combination of likelihood of failure and consequence of failure for a given pipe segment over a planning horizon.
47. Risk reduction per $1M: Portfolio-level metric that divides the modeled risk reduction achieved by a portfolio by its total cost in millions of dollars; used to compare the efficiency of alternative project lists.
48. RWT: Remaining Wall Thickness. Fraction or absolute thickness of pipe wall remaining, estimated from inspection tools; used as a ground-truth condition indicator in LOF validation.

49. SCADA: Supervisory Control and Data Acquisition. Operational control system that records pressures, flows, and levels; discussed as a future data source for time-dependent LOF modeling.
50. SDWA: Safe Drinking Water Act. U.S. federal law governing drinking-water quality and regulation; provides the regulatory context for utility obligations.
51. SETS: Socio–Ecological–Technical Systems. Framing that treats water infrastructure as a coupled system of technical assets, environmental conditions, and social institutions; used as the overarching conceptual lens for the dissertation.
52. SHAP: SHapley Additive exPlanations. Explainable-AI method that attributes model predictions to input features; used to interpret student ML models and to assess whether learned relationships align with known mechanisms.
53. SOS: System-of-Systems. Perspective that views source-to-tap water infrastructure as interconnected subsystems (e.g., source, treatment, distribution); related to the SETS framing.
54. Source-to-tap: End-to-end framing of water infrastructure that spans natural sources, raw-water conveyance, treatment, distribution, and customer use; this dissertation focuses on distribution mains but is embedded in a broader source-to-tap program.
55. SSURGO: Soil Survey Geographic Database. U.S. national soil database used to derive soil properties (e.g., corrosivity, drainage) for environmental exposure features.
56. SWIM: Sustainable Water Infrastructure Management program at Virginia Tech. Research group context for the dissertation.

57. TBL: Triple Bottom Line. Framework that simultaneously considers economic, environmental, and social dimensions; used to motivate multidimensional COF modeling and equity considerations.
58. Teacher–student modeling: Architectural pattern where an interpretable "teacher" model (here, a fuzzy-logic system based on expert knowledge) generates labels or guidance for training a more flexible "student" machine-learning model; the student inherits structure from the teacher while gaining predictive capacity.
59. Temporal commons: Conceptualization of infrastructure as a shared resource across generations, where current funding decisions shape the reliability and risk experienced by future customers; used to interpret renewal planning as an intergenerational fairness problem.
60. UQ: Uncertainty Quantification. Systematic characterization of uncertainty in model inputs, parameters, and outputs; appears in the EVV framework and in discussions of predictive intervals and bands.
61. USBR: United States Bureau of Reclamation. Federal agency collaborating on some supporting projects and data contexts.
62. USGCRP: U.S. Global Change Research Program. Federal program that produces national climate assessments cited in the climate and risk context.
63. USGS: United States Geological Survey. Federal agency providing hydrogeologic and climatic datasets used for environmental features.
64. WL: Water Loss. Volume or percentage of water lost through leaks, breaks, or metering errors; modeled in companion work and referenced in the broader source-to-tap context.

65. WRF: Water Research Foundation. Research funding organization; supports many of the studies cited in the literature review.
66. XAI: Explainable Artificial Intelligence. Subfield of AI that focuses on making model behavior interpretable to humans; SHAP and teacher–student modeling are used here as XAI tools.
67. XGB: XGBoost. Gradient-boosted decision-tree algorithm used as one of the baseline or comparative machine-learning models.

Appendix B: Data Dictionaries for all Teacher Models

Table B-1: Segment-level predictors, data sources, and hypothesized effects used in the LOF model for large-diameter metallic mains (applicable to >8 in).

Predictor	Definition (segment-level)	Unit / Scale	Source & resolution	Effect on LOF	Rel.	Materials	Proxy / fallback
Pipe age	Years from installation to observation year	years (raw) + 0–4 (fuzzy)	Asset register; annual	↑ older ⇒ ↑ LOF (loss of pre-stress, cumulative deterioration)	5	PCCP (all classes)	If missing: project in-service date from as-built/billing (2)
Pipe vintage / design class	Era/design standard (PCCP type, core/cylinder thickness, wire type)	categorical → 0–3 bins	As-built, design drawings; static	Legacy designs (thin cores, older wire types) ↑ LOF; newer robust designs ↓	4–5	PCCP	If incomplete: infer from install year + utility standard specs (1–2)
External mortar / coating condition	Condition of external mortar/coating and CP, where present	0–3 condition	External inspections, CP logs; ad hoc–5 yr	Poor or spalled mortar/exposed wires ↑ LOF; intact mortar ↓	4–5	PCCP	Use age + soil corrosivity + known exposed segments as proxy (2–3)
Main-break history	Historical PCCP failures or distress on line	breaks / mile / decade → 0–4	CMMS, forensic reports; rolling 10–20 yr	Higher historical failure/distress rate ↑ LOF	4–5	PCCP	Use line-level break history if segment-level sparse (2)
Broken-wire count / density	Number of broken pre-stress wires or broken-wire rate per unit length	count/segment or per 100 ft → 0–5 bins	EM/remote-field inspections; per inspection	Higher broken-wire density ↑ LOF (dominant direct indicator)	5	PCCP	Group EM anomalies into classes; if no data, use design + age + soil as very weak proxy (1)
Operating pressure & transients	Typical operating pressure and surge exposure for the line	psi (raw) + 0–5 bins	SCADA + transient analysis; continuous	High sustained pressure + strong surges ↑ wire stress ⇒ ↑ LOF	4–5	PCCP	HGL from model + elevation where SCADA incomplete (1–2)
Soil corrosivity	Composite soil corrosivity around pipe (resistivity, chlorides, sulfates, moisture)	index 0–4	Geotech/CP investigations, SSURGO; 10–30 m	More corrosive soils ↑ external corrosion of wires/cylinder ⇒ ↑ LOF	3–4	PCCP	Regional soil class + CP history + nearby resistivity (2)
Water table depth	Typical groundwater level relative to pipe	ft below/above invert + 0–3	Hydrogeology layers, wells; 100–500 m	High or fluctuating water table ↑ corrosion and soil movement ⇒ ↑ LOF	2–3	PCCP	DRASTIC/hydro maps + local wells (2)

Predictor	Definition (segment-level)	Unit / Scale	Source & resolution	Effect on LOF	Rel.	Materials	Proxy / fallback
Traffic loading	Roadway load class above PCCP alignment	road class 0–3	Road GIS; static	Heavy traffic ↑ live loads and may exacerbate cylinder stress ⇒ ↑ LOF	2–3	PCCP	Road functional class; assume "low" when off-road (1–2)
Burial depth	Cover depth from ground to crown	ft (raw) + 0–4	As-built drawings; one-time	Very shallow or very deep burial ↑ LOF (external loads, thermal/frost)	2–3	PCCP	DEM-based depth if invert known; otherwise typical depth by project type (1–2)
Bedding condition	Bedding and backfill quality supporting PCCP	0–2 condition	As-built, inspection photos, construction specs; one-time	Poor/uneven bedding ↑ point loads and differential stresses ⇒ ↑ LOF	3–4	PCCP	Soil type + project era + construction practice as proxy (1–2)
Groundwater fluctuation	Seasonal/long-term groundwater level range	ft range + 0–3	Hydrogeology models, wells; annual	Higher fluctuation ↑ cyclic loading and corrosion wet-dry cycling ⇒ ↑ LOF	2–3	PCCP	Use nearby water-level stations or modeled ranges (1–2)
Ambient temperature regime	Long-term air temperature band	°F band → 0–5	Climate normals; 1–10 km	Extremes (deep freeze/very hot) ↔ more thermal and soil-movement stresses ⇒ ↑ LOF	1–2	PCCP	Climate zone for each segment (1–2)
Precipitation regime	Long-term annual precipitation band	in/yr band → 0–4	Climate normals; 1–10 km	High rainfall ↔ more slope movement, washouts, erosion near PCCP corridors ⇒ ↑ LOF	1–2	PCCP	Assign from PRISM/NOAA grid (1–2)

Table B-2: Segment-level predictors, data sources, and hypothesized effects used in the LOF model for PVC and HDPE distribution mains.

Predictor	Definition (segment-level)	Unit / Scale	Source & resolution	Effect on LOF	Rel.	Materials	Proxy / fallback
Pipe age	Years from installation to observation year	years (raw) + 0–4 (fuzzy)	Asset register; annual	↑ older ⇒ ↑ LOF via fatigue, creep, oxidation	5	PVC, HDPE, PE, PB	If missing: service start date from billing or subdivision age (2)
Pipe vintage	Era capturing resin/formulation and installation standard changes	categorical → 0–3 bins	Asset register + material spec history; static	Early/vintage plastics (e.g., early PVC) ↑ LOF; modern standards ↓	4–5	PVC, HDPE	If unknown: infer from subdivision build year and material markings (1–2)
Ovality	Deviation from true circularity of pipe cross-section	% or class → 0–3 bins	CCTV / laser profile / mandrel testing; per inspection	Higher ovality ↑ stress and buckling risk ⇒ ↑ LOF	4	PVC, HDPE	If no direct measure: infer from deep cover + traffic + weak bedding (1–2)

Predictor	Definition (segment-level)	Unit / Scale	Source & resolution	Effect on LOF	Rel.	Materials	Proxy / fallback
Manufacturing defects	Presence of observed/known manufacturing defects (voids, inclusions, poor fusion)	yes/no → 0–2	Factory QA records, failure forensics; ad hoc	Confirmed defects strongly ↑ LOF (local weak points)	4–5	PVC, HDPE	If missing: use cohort-level defect rates by vendor/vintage; SME flags (1–2)
Main-break history	Historical breaks/leaks on or near the segment	breaks / mile / decade → 0–5	CMMS / work orders; rolling 5–10 yr	Higher break rate ↑ LOF (fatigue, environment)	5	PVC, HDPE	If sparse: cohort-level break rate by material, diameter, zone (2–3)
Operating pressure & transients	Typical zone pressure and surge exposure	psi + 0–5 bins	SCADA; transient studies; continuous	Higher operating and surge pressures ↑ fatigue and creep ⇒ ↑ LOF	4–5	PVC, HDPE	Modelled pressures from HGL + elevation where SCADA absent (1–2)
Soil particle size / inclusions	Dominant soil gradation and presence of coarse particles/rocks	categorical 0–2	Geotech logs, SSURGO, trench notes; 10–30 m	Coarse angular particles near pipe ↑ point loading; fine but stable soils ↓	3–4	PVC, HDPE	Soil texture + rock content from logs or local standards (2)
Burial depth	Cover depth from ground to crown	ft (raw) + 0–5 bins	As-built; DEM + invert; one-time	Very shallow ↑ frost/traffic risk; very deep ↑ long-term deflection; mid-range ↓	3–4	PVC, HDPE	DEM-derived depth + road class; assume typical depth if no invert (1–2)
Traffic loading	Roadway load class over pipe (local → interstate)	road class 0–3 → 0–3	Road centerline / functional class; static	High traffic ↑ live loads and fatigue ⇒ ↑ LOF	3–4	PVC, HDPE	Road functional class; if off-road, treat as "low" (1–2)
Water table depth	Typical groundwater level relative to pipe	ft below/above invert + 0–3 bins	Hydrogeology layers, wells; 100–500 m	Sustained high water table can exacerbate deflection and joint loading ⇒ ↑ LOF	2–3	PVC, HDPE	DRASTIC / aquifer maps + local well logs (2)
Visible cracks / distress	Observed crack severity (crazing, longitudinal, circumferential, splits)	ordinal 0–4	CCTV, failure inspections; per job	Hairline < minor < major < fracture/burst; severity ↑ LOF sharply	5	PVC, HDPE	If no CCTV: infer from failure descriptions and leak patterns (2)
Groundwater fluctuation	Seasonal range in groundwater level	ft range + 0–3	Hydrogeology and climate; annual	Large fluctuations ↑ soil movement and cyclic loading ⇒ ↑ LOF	2–3	PVC, HDPE	Use nearby water-level stations or modeled seasonal range (1–2)
Ambient temperature regime	Long-term air temperature band	°F band → 0–5	Climate normals (PRISM/NOAA); 1–10 km	Extreme cold or heat can embrittle or soften plastics ⇒ ↑ LOF	2–3	PVC, HDPE	Climate zone mapped to segments (1–2)
Precipitation regime	Long-term annual precipitation band	in/yr band → 0–4	Climate normals; 1–10 km	Very high rainfall ↔ more slope instability and soil movement ⇒ ↑ LOF	1–2	PVC, HDPE	PRISM/NOAA rainfall cell for segment (1–2)

Predictor	Definition (segment-level)	Unit / Scale	Source & resolution	Effect on LOF	Rel.	Materials	Proxy / fallback
Internal water temperature	Typical water temperature band in main	°F band → 0–5	SCADA or grab samples; seasonal	Elevated water temp over long periods may accelerate creep/oxidation ⇒ ↑ LOF	1–2	PVC, HDPE	Use source temp + seasonal adjustment vs burial depth (1–2)
Bedding condition	Support class and compaction quality	0–2 condition	As-built, inspection photos; one-time	Poor/uneven bedding ↑ deflection and point loads ⇒ ↑ LOF	3–4	PVC, HDPE	Road class + depth + soil type + era-typical bedding practice (1–2)

Table B-3: Segment-level predictors, data sources, and hypothesized effects used in the LOF model for prestressed concrete cylinder pipe (PCCP) mains.

Predictor	Definition (segment-level)	Unit / Scale	Source & resolution	Effect on LOF	Rel.	Materials	Proxy / fallback
Pipe age	Years from installation to observation year	years (raw) + 0–4 (fuzzy)	Asset register; annual	↑ older ⇒ ↑ LOF (loss of pre-stress, cumulative deterioration)	5	PCCP (all classes)	If missing: project in-service date from as-built/billing (2)
Pipe vintage / design class	Era/design standard (PCCP type, core/cylinder thickness, wire type)	categorical → 0–3 bins	As-built, design drawings; static	Legacy designs (thin cores, older wire types) ↑ LOF; newer robust designs ↓	4–5	PCCP	If incomplete: infer from install year + utility standard specs (1–2)
External mortar / coating condition	Condition of external mortar/coating and CP, where present	0–3 condition	External inspections, CP logs; ad hoc–5 yr	Poor or spalled mortar/exposed wires ↑ LOF; intact mortar ↓	4–5	PCCP	Use age + soil corrosivity + known exposed segments as proxy (2–3)
Main-break history	Historical PCCP failures or distress on line	breaks / mile / decade → 0–4	CMMS, forensic reports; rolling 10–20 yr	Higher historical failure/distress rate ↑ LOF	4–5	PCCP	Use line-level break history if segment-level sparse (2)
Broken-wire count / density	Number of broken pre-stress wires or broken-wire rate per unit length	count/segment or per 100 ft → 0–5 bins	EM/remote-field inspections; per inspection	Higher broken-wire density ↑ LOF (dominant direct indicator)	5	PCCP	Group EM anomalies into classes; if no data, use design + age + soil as very weak proxy (1)
Operating pressure & transients	Typical operating pressure and surge exposure for the line	psi (raw) + 0–5 bins	SCADA + transient analysis; continuous	High sustained pressure + strong surges ↑ wire stress ⇒ ↑ LOF	4–5	PCCP	HGL from model + elevation where SCADA incomplete (1–2)
Soil corrosivity	Composite soil corrosivity around pipe (resistivity, chlorides, sulfates, moisture)	index 0–4	Geotech/CP investigations, SSURGO; 10–30 m	More corrosive soils ↑ external corrosion of wires/cylinder ⇒ ↑ LOF	3–4	PCCP	Regional soil class + CP history + nearby resistivity (2)

Predictor	Definition (segment-level)	Unit / Scale	Source & resolution	Effect on LOF	Rel.	Materials	Proxy / fallback
Water table depth	Typical groundwater level relative to pipe	ft below/above invert + 0–3	Hydrogeology layers, wells; 100–500 m	High or fluctuating water table ↑ corrosion and soil movement ⇒ ↑ LOF	2–3	PCCP	DRASTIC/hydro maps + local wells (2)
Traffic loading	Roadway load class above PCCP alignment	road class 0–3	Road GIS; static	Heavy traffic ↑ live loads and may exacerbate cylinder stress ⇒ ↑ LOF	2–3	PCCP	Road functional class; assume "low" when off-road (1–2)
Burial depth	Cover depth from ground to crown	ft (raw) + 0–4	As-built drawings; one-time	Very shallow or very deep burial ↑ LOF (external loads, thermal/frost)	2–3	PCCP	DEM-based depth if invert known; otherwise typical depth by project type (1–2)
Bedding condition	Bedding and backfill quality supporting PCCP	0–2 condition	As-built, inspection photos, construction specs; one-time	Poor/uneven bedding ↑ point loads and differential stresses ⇒ ↑ LOF	3–4	PCCP	Soil type + project era + construction practice as proxy (1–2)
Groundwater fluctuation	Seasonal/long-term groundwater level range	ft range + 0–3	Hydrogeology models, wells; annual	Higher fluctuation ↑ cyclic loading and corrosion wet-dry cycling ⇒ ↑ LOF	2–3	PCCP	Use nearby water-level stations or modeled ranges (1–2)
Ambient temperature regime	Long-term air temperature band	°F band → 0–5	Climate normals; 1–10 km	Extremes (deep freeze/very hot) ↔ more thermal and soil-movement stresses ⇒ ↑ LOF	1–2	PCCP	Climate zone for each segment (1–2)
Precipitation regime	Long-term annual precipitation band	in/yr band → 0–4	Climate normals; 1–10 km	High rainfall ↔ more slope movement, washouts, erosion near PCCP corridors ⇒ ↑ LOF	1–2	PCCP	Assign from PRISM/NOAA grid (1–2)

Table B-4: Segment-level predictors, fuzzy scaling, linguistic membership functions, and expected directional effects in the AC pipe LOF model.

Predictor	Definition (segment-level)	Unit / Scale	Source & resolution	Effect on LOF	Rel.	Materials	Proxy / fallback
Pipe age	Installed age of the AC pipe segment since construction or last full replacement.	years (raw) + fuzzy 0–4 (0 ≈ new, 4 ≈ very old)	Asset register, construction records; annual snapshot	↑ age ⇒ ↑ LOF through cumulative matrix degradation, joint distress, and crack growth.	5	AC	If missing: use service start date from billing, subdivision build year, or cohort age (2).

Predictor	Definition (segment-level)	Unit / Scale	Source & resolution	Effect on LOF	Rel.	Materials	Proxy / fallback
Pipe type	AC manufacturing / type index (pressure class, product line, formulation).	categorical (0–2 types)	Asset register, design specs; static	Less robust types or early product lines ↑ LOF; modern types with better QA/QC ↓.	3–4	AC	If unknown: infer from install year, diameter/pressure class, and typical standard in that era (1–2).
Remaining wall thickness (RWT)	Remaining wall thickness / structural capacity proxy for AC barrel.	% of nominal wall (≈60–100%) → 0–5 fuzzy bins	Coupon tests, exhumed samples, NDT where available; ad hoc	↓ RWT, especially in "Poor/Critical" fuzzy classes, sharply ↑ LOF (dominant structural indicator).	5	AC	If no direct data: use age + soil aggressiveness + break history to place in coarse RWT class (1–2).
External protection condition	Condition/effectiveness of external coatings, wraps, liners, or CP (if present).	qualitative 0–3 (Good/Fair/Poor)	Field inspections, rehab records; ad hoc–5 yr	Poor/failed external protection ↑ external chemical and mechanical deterioration ⇒ ↑ LOF.	3–4	AC	If missing: combine soil aggressiveness, project era, and rehab history to infer likely condition (2).
Main-break history	Historic break/leak frequency on or near the segment over a reference period.	breaks / 10 yr / km → 0–6 fuzzy bins	CMMS / work orders; rolling 5–10 yr	Higher historical break rate (High/Very_High) strongly ↑ LOF (fatigue, poor construction, local hazards).	5	AC	If sparse: use cohort-level break rate by material, diameter, and zone; flag as lower reliability (2–3).
Operating pressure	Typical operating and peak pressure range seen by the segment.	psi (≈0–200) + 0–5 bins	SCADA, pressure loggers, hydraulic model; continuous	Very low or very high pressure classes ("Very_Poor") ↑ structural and functional failure risk ⇒ ↑ LOF.	4	AC	Where SCADA absent: use modeled HGL + elevation and zone settings (2).
Soil particle size	Dominant soil gradation around pipe controlling bedding quality, erosion, and migration.	categorical 0–2 (Coarse/Medium/Fine)	Geotech logs, SSURGO, trench notes; 10–30 m	Coarse, poorly graded soils can ↑ point loading and erosion; fine but saturated soils ↑ chemical attack; direction encoded in rules but generally harsher regimes ↑ LOF.	3	AC	If missing: infer from regional soil maps and typical trench backfill practice (1–2).
Burial depth	Cover depth from finished grade to pipe crown.	ft (≈0–30) → 0–5 bins	As-built drawings; DEM + invert; one-time	Very shallow depth ↑ traffic and frost loads; very deep depth ↑ long-term earth loads ⇒ both tails ↑ LOF.	3	AC	If invert unknown: estimate depth from DEM, pipe size, and typical practice for street vs off-road (1–2).
Frost action	Severity of frost action and freeze–thaw cycling in corridor.	qualitative 0–3 (Low/Medium/High)	Climate maps, frost index; 1–10 km	High frost action ↑ joint distress, cracking, heave ⇒ ↑ LOF in cold climates.	2–3	AC	Use climate/frost zone mapped to segment if no site-specific data (1–2).
Traffic loading	Traffic loading category over the alignment	road class 0–3 → 0–3	Road centerline /	Heavier traffic (arterials/highways) ↑ live loads and fatigue cycles ⇒ ↑ LOF.	3	AC	If off-road, treat as "Low"; if road class unknown, infer from

Predictor	Definition (segment-level)	Unit / Scale	Source & resolution	Effect on LOF	Rel.	Materials	Proxy / fallback
	(local vs arterial vs highway).		functional class GIS; static				land use and functional class (1–2).
Water table depth	Typical depth from ground surface to groundwater table near the pipe.	ft (Shallow 0–10, Moderately_Deep 5–15, Deep >10) → 0–3	Hydrogeology layers, wells; 100–500 m	Shallow water table ↑ saturation, buoyancy, and chemical attack, generally ↑ LOF.	2–3	AC	Use DRASTIC / aquifer maps and local well logs to assign depth band (2).
Pipe cracks	Observed AC barrel crack severity from inspection.	ordinal 0–5 (None → Fracture/Burst)	CCTV, coupons, exhumed samples; per job	Increasing crack severity strongly ↓ performance and pushes "near-failure" ⇒ sharply ↑ LOF.	5	AC	Where no CCTV: use failure descriptions, leak type, and distress notes to assign coarse class (2).
Soil aggressiveness to AC matrix	Soil aggressiveness to cementitious AC matrix (pH, sulfate, resistivity).	indirect index 0–3	Geotech/CP investigations, soil chemistry; corridor-scale	High aggressiveness ↑ matrix degradation and loss of structural capacity ⇒ ↑ LOF.	3–4	AC	Use regional soil class, CP history, and resistivity data where available; otherwise assign from soil map (2).
Joint condition	Structural and leakage condition of joints along segment.	qualitative 0–2 (Poor/Fair/Good)	Inspection reports, repair logs; ad hoc	Poor joints ↑ leakage, infiltration/exfiltration, and local barrel distress ⇒ ↑ LOF.	3	AC	If unknown: infer from age, joint type, and break/leak patterns on nearby segments (1–2).
Valve/appurtenance condition	Condition of valves and appurtenances connected to the segment.	qualitative 0–2	Valve inspection records, work orders; ad hoc	Poor valves concentrate stresses and create local corrosion/leak points ⇒ ↑ LOF.	2	AC	If missing: use valve age and replacement history as proxy; assume "Fair" when unknown (1–2).
Groundwater fluctuation	Amplitude of seasonal/long-term groundwater level fluctuation.	ft range → 0–3	Hydrogeology models, wells; annual	Larger fluctuation ranges ↑ cyclic external loading and environmental variability ⇒ ↑ LOF.	2–3	AC	Use nearby groundwater stations or modeled seasonal range for corridor (2).
Precipitation regime	Long-term mean annual precipitation in service area.	in/yr bands → 0–4	Climate normals; 1–10 km	Very high rainfall ("Critical") ↑ soil movement, scour, saturation near AC corridors ⇒ ↑ LOF.	1–2	AC	Assign from PRISM/NOAA grid for each segment (1–2).
Bedding condition	Quality of bedding and backfill installation/condition around pipe.	qualitative 0–2 (Poor/Fair/Good)	As-built notes, inspection photos; one-time	Good bedding ↓ structural deterioration; poor bedding ↑ point loads and deflection ⇒ ↑ LOF.	3–4	AC	If absent: infer from project era, soil type, and typical construction practice (1–2).

Table B-5: Segment-level predictors, fuzzy scaling, linguistic membership functions, and expected directional effects in the concrete pipe (RCP, RCCP, BWP) LOF model.

Predictor	Definition (segment-level)	Unit / Scale	Source & resolution	Effect on LOF	Rel.	Materials	Proxy / fallback
Pipe age	Installed age of the concrete pipe segment since construction or full replacement.	years (raw) + fuzzy 0–4	Asset register; annual	↑ age ⇒ ↑ deterioration, especially where combined with poor vintage or aggressive environment ⇒ ↑ LOF.	5	RCP, RCCP, BWP	If missing: use project in-service year from as-builts or billing (2).
Pipe vintage	Manufacturing era / design vintage capturing standards, materials, and QA/QC.	mapped install year → 0–3 bins	As-builts, design standards; static	Less favorable vintages (Relatively_Fair) linked to higher defect rates and lower capacity ⇒ ↑ LOF.	4	RCP, RCCP, BWP	If incomplete: infer from install year, plant/vendor, and known standard changes (1–2).
External protection condition	Condition of external mortar/shotcrete coating and protective wraps/linings.	qualitative 0–3 (Good/Fair/Poor)	Field inspections, rehab records; ad hoc–5 yr	Poor external protection ↑ external corrosion, spalling, and reinforcement exposure ⇒ ↑ LOF.	4	RCP, RCCP, BWP	Use age + soil corrosivity + rehab history to infer likely condition (2–3).
Main-break history	Historic main-break frequency for concrete segments.	breaks per 10 yr / km → 0–6	CMMS, work orders; rolling 5–10 yr	Higher historic break rate (High/Very_High) is strong evidence of distress and ↑ LOF.	5	RCP, RCCP, BWP	If sparse: use cohort-level break rate by material/diameter/zone (2–3).
Reinforcement break count	Number/severity of reinforcement (wire/strap/bar) breaks from inspection.	broken wires/straps per pipe or per 100 ft → 0–5 bins	EM, visual inspection, internal CCTV; per inspection	Larger reinforcement break counts sharply ↓ structural safety margin and push near-failure ⇒ ↑ LOF.	5	RCCP, BWP (reinforced classes)	Where EM absent: infer risk from age, vintage, soil corrosivity, and break patterns on cohort (1–2).
Operating pressure	Typical operating pressure range for concrete main.	psi + 0–5 bins	SCADA, model; continuous	Extreme low/high ranges ("Very_Poor", "Poor") ↑ structural and serviceability issues ⇒ ↑ LOF.	4	RCP, RCCP, BWP	Use modeled pressures from HGL + elevation in zones without SCADA (2).
Soil particle size	Dominant grain size of backfill/soil surrounding pipe.	categorical 0–2 (Coarse/Medium/Fine)	Geotech logs, SSURGO; 10–30 m	Coarse, poorly graded soils may ↑ erosion and load concentrations; fine, wet soils may retain moisture and ↑ corrosion; harsher regimes ↑ LOF.	3	RCP, RCCP, BWP	If missing: infer from soil maps and typical trench backfill (1–2).
Burial depth	Burial depth from finished grade to crown of concrete pipe.	ft (0–30) → 0–5 bins	As-builts; DEM + invert; one-time	Very shallow and very deep segments generally have higher risk (traffic/frost vs earth loads) ⇒ ↑ LOF at extremes.	3	RCP, RCCP, BWP	If invert unknown: estimate from DEM, pipe size, and project type (1–2).
Frost action	Freeze–thaw severity along corridor of concrete main.	qualitative 0–3 (Low/Medium/High)	Climate/frost index maps; 1–10 km	High frost action promotes cracking, joint distress, and spalling ⇒ ↑ LOF.	2–3	RCP, RCCP, BWP	Map climate zone to segment where site data absent (1–2).

Predictor	Definition (segment-level)	Unit / Scale	Source & resolution	Effect on LOF	Rel.	Materials	Proxy / fallback
Traffic loading	Overlying traffic category above pipeline alignment.	road type index 0–3 → 0–3	Road GIS; static	Higher traffic (arterials, highways) ↑ external load cycles and fatigue ⇒ ↑ LOF.	3	RCP, RCCP, BWP	If alignment off-road, treat as "Low"; else infer from road class (1–2).
Water table depth	Depth from ground surface to groundwater table near concrete main.	ft → 0–3 (Shallow/Moderately_Deep/Deep)	Hydrogeology maps, wells; 100–500 m	Shallow water table keeps pipe saturated and ↑ external corrosion and instability ⇒ ↑ LOF.	3	RCP, RCCP, BWP	Use DRASTIC/hydro maps and local wells where available (2).
Pipe cracks	Observed crack condition of concrete barrel.	ordinal 0–5 (None → Fracture/Burst)	CCTV, internal inspection, exhumed segments; per job	More severe cracks signal serious distress and near-failure ⇒ sharply ↑ LOF.	5	RCP, RCCP, BWP	If no CCTV: classify from failure descriptions, leak type, and forensic notes (2).
Soil corrosivity (to steel)	Soil corrosivity toward embedded reinforcement (steel).	indirect corrosion index 0–4	CP surveys, soil resistivity tests, geotech reports; corridor-scale	Critical/High corrosivity accelerates reinforcement corrosion and loss of capacity ⇒ ↑ LOF.	4–5	RCCP, BWP, reinforced RCP	If limited data: use regional soil class and any CP/current requirement records (2).
Concrete matrix aggressiveness	Soil aggressiveness toward concrete matrix (sulfate, pH, etc.).	qualitative/index 0–3	Geochem data, geotech reports; corridor-scale	High aggressiveness leads to faster matrix degradation, cover loss, and cracking ⇒ ↑ LOF.	3–4	RCP, RCCP, BWP	Infer from local sulfate/pH maps and soil types if detailed tests missing (2).
Stray currents	Presence of stray DC currents from nearby electrical/rail infrastructure.	binary 0–2 (Absent/Present)	CP monitoring, utility/rail records; ad hoc	Presence of stray currents accelerates reinforcement corrosion ⇒ ↑ LOF.	2–3	RCCP, BWP, reinforced RCP	If no direct data: flag segments near rail or DC traction power as "Present (low-confidence)" (1–2).
Joint condition	Condition of concrete pipe joints (gaskets, bell-spigot, steel collars).	qualitative 0–2 (Poor/Fair/Good)	Inspection records, repair logs; ad hoc	Poor joints ↑ leakage, infiltration/exfiltration, and local distress ⇒ ↑ LOF.	3	RCP, RCCP, BWP	Infer from age, joint type, and break/leak patterns if direct inspection absent (1–2).
Valve/appurtenance condition	Condition of valves/appurtenances attached to the main.	qualitative 0–2	Valve inspection, work orders; ad hoc	Poor valves/appurtenances concentrate loads and create local corrosion/leak points ⇒ ↑ LOF.	2	RCP, RCCP, BWP	Use valve age, type, and maintenance history to assign coarse class where missing (1–2).
Groundwater fluctuation	Magnitude of groundwater level fluctuation over time.	ft range → 0–3	Hydrogeology/climate models, wells; annual	Larger fluctuations ↑ cyclic loading and changing saturation, generally ↑ LOF.	2–3	RCP, RCCP, BWP	Use nearby groundwater stations or modeled seasonal range (2).
Precipitation regime	Long-term average annual precipitation in service area.	in/yr bands → 0–4	Climate normals; 1–10 km	Very high precipitation ↑ soil movement, saturation, and external distress ⇒ ↑ LOF.	1–2	RCP, RCCP, BWP	Assign from PRISM/NOAA grid for each segment (1–2).

Predictor	Definition (segment-level)	Unit / Scale	Source & resolution	Effect on LOF	Rel.	Materials	Proxy / fallback
Bedding condition	Quality of bedding/backfill around the concrete pipe.	qualitative 0–2 (Poor/Fair/Good)	As-built notes, inspection photos; one-time	Good bedding supports pipe and ↓ structural problems; poor bedding ↑ risk of distress and LOF.	3–4	RCP, RCCP, BWP	If absent: infer from project era, soil type, and typical construction practice (1–2).

Appendix C: Fuzzy Inference Systems Input Parameters

Table C-1: Fuzzy membership-function labels, rescaled numerical ranges, and input-data preparation for concrete pipe parameters in the LOF fuzzy-inference model.

Parameter	Membership Function Labels	Range	Typical Unit	Input Data Preparation
Pipe Age	New (0–0.27), Used (0.27–1.07), Old (1.07–1.6), Obsolete (1.6–2.67)	0–4	Years	Original range: 0–150 years. Scaled down: New (0–20), Used (20–60), Old (60–90), Obsolete (>90). Same functional form as metallic pipes.
Pipe Vintage	Best Vintage (0–0.5), Relatively Good (0.5–1.5), Relatively Fair (1.5–3)	0–3	Years / Era	Original range: 1900–2025. Example: Best Vintage (1970–1995), Relatively Good (1950–1970, 1995–2010), Relatively Fair (pre-1950 or post-2010 if early/unproven products). Tuned by utility knowledge for RCP/RCCP/BWP.
Internal Lining / Mortar Condition	Good (1.5–2), Fair (0.5–1.5), Poor (0–0.5)	0–2	Condition Score	Original range: 0–2 from internal inspection. Scaled down: Poor = severe spalling or delamination, Fair = localized defects, Good = intact mortar or lining.
External Jacket / Coating	Good (2–3), Fair (1–2), Poor (0–1)	0–3	Condition Score	Original range: 0–3. Encodes condition of concrete jacket, coatings, or encasement; same mapping for RCP, RCCP, BWP.
Reinforcement Condition (RCCP/BWP)	Good (1.5–2), Moderate (0.5–1.5), Poor (0–0.5)	0–2	Condition Score	Based on EM/wire-break analysis or inspection. For RCP (no prestress), set to Good or use a neutral default.
Wire-Break Severity (RCCP/BWP)	None (0–0.25), Low (0.25–0.75), Moderate (0.75–1.5), High (1.5–2.5), Very High (2.5–5)	0–5	Broken wire index	Original range: 0–100+ broken wires. Scaled: None (0), Low (1–5), Moderate (5–20), High (20–60), Very High (>60). Only populated for prestressed pipes.
Structural Condition (Overall)	Excellent (4–5), Good (3–4), Fair (2–3), Poor (1–2), Critical (0–1)	0–5	Condition Score	Derived index from CCTV/EM/visual scoring schemes (e.g., 0–5). Re-binned into five bands to maintain consistency with LOF categories.
Pipe Cracks	No Cracks (0–1), Hairline Crack (1–2), Minor Crack (2–3), Major Crack (3–4), Fracture / Spalled (4–5)	0–5	Severity Level	Original range: 0–5 crack severity rating. Same structure as metallic but interpreted for concrete cracking and joint distress.
Spalls / Delamination	None (0–1), Minor (1–2), Moderate (2–3), Severe (3–4), Extensive (4–5)	0–5	Severity Level	Based on visual inspection of spalls, delamination, and surface scaling. Original range: 0–5.

Parameter	Membership Function Labels	Range	Typical Unit	Input Data Preparation
Joint Condition	Poor (0–0.7), Fair (0.7–1.3), Good (1.3–2)	0–2	Condition Score	Original range: 0–2. Scaled: Poor (0–0.7) = misaligned/gapping, Fair (0.7–1.3), Good (1.3–2).
Bedding Condition	Poor (0–0.5), Fair (0.5–1.5), Good (1.5–2)	0–2	Condition Score	Original range: 0–2. Scaled down as for metallic pipes; critical for rigid-pipe support.
Pipe Breaks (Concrete)	None (0–0.25), Low (0.25–0.375), Moderate (0.375–0.75), High (0.75–1.25), Very High (>1.25)	0–5	Count	Original range: 0–20 breaks. Scaled: None (0–1), Low (1–2), Moderate (2–4), High (4–5), Very High (>5).
Soil Corrosivity	Critical (0–0.4), High (0.4–1), Medium (1–2), Low (2–4)	0–4	Index	Original range: 0–10,000. Scaled as metallic. For concrete, this includes risk of reinforcement corrosion and external sulfate attack.
Sulfate Exposure	Low (2–3), Medium (1–2), High (0–1)	0–3	Index	Based on sulfate concentrations in soil/groundwater. Original range: 0–3 (e.g., from geotechnical or water-quality data).
Pipe Pressure	Very Poor (0–1, 3.75–5), Poor (1–1.75, 3–3.75), Fair (1.75–2.25), Good (2.25–2.75), Excellent (2.75–3.5)	0–5	psi	Original range: 0–200 psi. Same scaled bands as metallic; high pressure and large fluctuations are more critical for deteriorated concrete pipes.
Soil Particle Size	Coarse (0–1), Medium (1–1.5), Fine (1.5–2)	0–2	Size Category	Coarse (gravel/sand), Medium (sandy silt), Fine (silt/clay). Same coding as metallic; influences bedding and load distribution.
Traffic Loading	Low (2–3), Medium (1–2), High (0–1)	0–3	Index	Original range: 0–3 (off-road to major arterial/rail). Scaled: Low (2–3), Medium (1–2), High (0–1).
Frost Action	Low (2–3), Medium (1–2), High (0–1)	0–3	Index	Original range: 0–3 frost severity. Same mapping as metallic pipes.
Water Table Depth	Deep (1.5–3), Moderately Deep (0.75–1.5), Shallow (0–0.75)	0–3	Feet	Original range: 0–20 ft. Scaled: Deep (>10 ft), Moderately Deep (5–15 ft), Shallow (0–10 ft).
Groundwater Fluctuation	Low (0–0.45), Moderate (0.45–0.9), High (0.9–3)	0–3	Feet	Original range: 0–20 ft seasonal fluctuation. Scaled: Low (0–3 ft), Moderate (3–7.5 ft), High (7.5–20 ft).
Precipitation	Low (0–1), Average (1–2), High (2–3), Critical (3–4)	0–4	Inches/Year	Original range: 0–100 in/yr. Scaled like metallic pipes: Low (0–10), Average (10–30), High (30–60), Critical (60–100).

Table C-2: Fuzzy membership-function labels, rescaled numerical ranges, and input-data preparation for metallic large diameter pipe parameters in the LOF fuzzy-inference model.

Parameter	Membership Function Labels	Range	Typical Unit	Input Data Preparation
Pipe Age	New (0–0.27), Used (0.27–1.07), Old (1.07–1.6), Obsolete (1.6–2.67)	0–4	Years	Original range: 0–300 years. Scaled down: New (0–20 years), Used (20–80 years), Old (80–120 years), Obsolete (>120 years).

Parameter	Membership Function Labels	Range	Typical Unit	Input Data Preparation
Pipe Vintage	Best Vintage (0–0.5), Relatively Good (0.5–1.5), Relatively Fair (1.5–3)	0–3	Years	Original range: 1900–2025. Scaled down: for Cast Iron pipes, Best Vintage (1900–1930), Relatively Good (1930–1960), Relatively Fair (1960–2025).
Pipe Internal Lining	Yes (0–1), No (1–2)	0–2	Boolean	Same coding for all metallic materials (lined vs unlined).
Pipe External Protection	Good (2–3), Fair (1–2), Poor (0–1)	0–3	Condition Score	Same for all metallic materials; based on coating, wrapping, and cathodic protection condition.
Water Quality	Good (0.3–2), Fair (−0.3 to 0.3), Poor (−2 to −0.3)	−2 to 2	Langelier Index	Original range: −2 to 2. Scaled down: Good (0.3–2), Fair (−0.3 to 0.3), Poor (−2 to −0.3).
Pipe Breaks <16″	None (0–0.25), Low (0.25–0.375), Moderate (0.375–0.75), High (0.75–1.25), Very High (>1.25)	0–5	Count	Original range: 0–20 breaks. Scaled down: None (0–1), Low (1–2), Moderate (2–4), High (4–5), Very High (>5).
Pipe C-Factor	Rough (1.25–1.75), Fair (2.0–2.5), Smooth (2.5–3.75), Perfect (3.75–4)	0–4	Coefficient	Original range: 0–160. Scaled down: Rough (50–70), Fair (70–110), Smooth (110–130), Perfect (130–160).
Pipe Remaining Wall Thickness (RWT)	Excellent (4.95–5), Good (4.5–4.95), Fair (4.0–4.5), Poor (3.0–4.0), Critical (0–3.0)	0–5	%	Original range: 0–100%. Scaled down: Excellent (99–100), Good (90–99), Fair (85–90), Poor (60–85), Critical (<60).
Pipe Pit Depth	Low (0–0.3), Moderate (0.6–1.2), High (1.2–2.4), Critical (2.4–4)	0–4	%	Original range: 0–100% of wall thickness. Scaled down: Low (0–7.5), Moderate (7.5–22.5), High (22.5–45), Critical (45–100).
Pipe Pressure	Very Poor (0–1, 3.75–5), Poor (1–1.75, 3–3.75), Fair (1.75–2.25), Good (2.25–2.75), Excellent (2.75–3.5)	0–5	psi	Original range: 0–200 psi. Scaled down: Very Poor (0–40, 150–200), Poor (40–70, 135–150), Fair (70–90), Good (90–110), Excellent (110–150).
Soil Particle Size	Coarse (0–1), Medium (1–1.5), Fine (1.5–2)	0–2	Size Category	Same coding for all metallic materials; based on geotechnical logs (coarse/medium/fine).
Pipe Depth	Very Shallow (0–0.33), Shallow (0.33–0.67), Moderate (0.67–1.5), Deep (1.5–2.5), Very Deep (2.5–5)	0–5	Feet	Original range: 0–30 ft. Scaled down: Very Shallow (0–2 ft), Shallow (2–4 ft), Moderate (4–10 ft), Deep (10–20 ft), Very Deep (20–30 ft).
Frost Action	Low (2–3), Medium (1–2), High (0–1)	0–3	Index	Original range: 0–3. Scaled down: Low (2–3), Medium (1–2), High (0–1).
Traffic Loading	Low (2–3), Medium (1–2), High (0–1)	0–3	Index	Original range: 0–3. Scaled down: Low (2–3), Medium (1–2), High (0–1).
Water Table Depth	Deep (1.5–3), Moderately Deep (0.75–1.5), Shallow (0–0.75)	0–3	Feet	Original range: 0–20 ft. Scaled down: Deep (>10 ft), Moderately Deep (5–15 ft), Shallow (0–10 ft).
Pipe Cracks	No Cracks (0–1), Hairline Crack (1–2), Minor Crack (2–3), Major Crack (3–4), Fracture Burst (4–5)	0–5	Severity Level	Original range: 0–5. Scaled down: No Cracks (0–1), Hairline (1–2), Minor (2–3), Major (3–4), Fracture Burst (4–5).
Soil Corrosivity	Critical (0–0.4), High (0.4–1), Medium (1–2), Low (2–4)	0–4	Index	Original range: 0–10,000. Scaled down: Critical (0–1000), High (1000–2500), Medium (2500–5000), Low (5000–10,000).
Stray Currents	Absent (0–1), Present (1–2)	0–2	Presence	Original range: 0–2. Scaled down: Absent (0–1), Present (1–2).
Pipe Joints	Poor (0–0.7), Fair (0.7–1.3), Good (1.3–2)	0–2	Condition Score	Original range: 0–2. Scaled down: Poor (0–0.7), Fair (0.7–1.3), Good (1.3–2).

Parameter	Membership Function Labels	Range	Typical Unit	Input Data Preparation
Pipe Valves	Poor (0–0.7), Fair (0.7–1.3), Good (1.3–2)	0–2	Condition Score	Original range: 0–2. Scaled down: Poor (0–0.7), Fair (0.7–1.3), Good (1.3–2).
Pipe Graphitization	Low (0–0.4), Moderate (0.8–2), High (2–3.4), Severe (3.4–4)	0–4	Condition Score	Original range: 0–100. Scaled down: Low (0–10), Moderate (10–35), High (35–65), Severe (65–100).
Groundwater Fluctuation	Low (0–0.45), Moderate (0.45–0.9), High (0.9–3)	0–3	Feet	Original range: 0–20 ft. Scaled down: Low (0–3 ft), Moderate (3–7.5 ft), High (7.5–20 ft).
Temperature	Very Cold (0–1), Cold (1–2), Moderate (2–3), Warm (3–4), Very Hot (4–5)	0–5	°F	Original range: −20 to 150 °F. Scaled down: Very Cold (<32 °F), Cold (32–57 °F), Moderate (57–82 °F), Warm (82–107 °F), Very Hot (>107 °F).
Precipitation	Low (0–1), Average (1–2), High (2–3), Critical (3–4)	0–4	Inches/Year	Original range: 0–100 in/yr. Scaled down: Low (0–10), Average (10–30), High (30–60), Critical (60–100).
Internal Water Temperature	Hot (3.5–5), Warm (3–3.5), Moderate (2–3), Cool (1–2), Cold (0–1)	0–5	°F	Original range: 20–100 °F. Scaled down: Hot (>70 °F), Warm (55–70 °F), Moderate (45–55 °F), Cool (35–45 °F), Cold (<35 °F).
Bedding Condition	Poor (0–0.5), Fair (0.5–1.5), Good (1.5–2)	0–2	Condition Score	Original range: 0–2. Scaled down: Poor (0–0.5), Fair (0.5–1.5), Good (1.5–2).

Table C-3: Fuzzy membership-function labels, rescaled numerical ranges, and input-data preparation for PVC and PE pipe parameters in the LOF fuzzy-inference model.

Parameter	Membership Function Labels	Range	Typical Unit	Input Data Preparation
Pipe Age	New (0–0.27), Used (0.27–1.07), Old (1.07–1.6), Obsolete (1.6–2.67)	0–4	Years	Original range: 0–150 years (design life). Scaled down: New (0–20 years), Used (20–60 years), Old (60–90 years), Obsolete (>90 years). Same functional form as metallic age.
Pipe Vintage	Best Vintage (0–0.5), Relatively Good (0.5–1.5), Relatively Fair (1.5–3)	0–3	Years / Era	Original range: 1950–2025. Example scaling: Best Vintage (1990–2010, modern PVC/PE standards), Relatively Good (1975–1990), Relatively Fair (1950–1975 or early plastics). Exact bins can be tuned to utility data.
Pipe Breaks (Plastic)	None (0–0.25), Low (0.25–0.375), Moderate (0.375–0.75), High (0.75–1.25), Very High (>1.25)	0–5	Count	Original range: 0–20 breaks. Scaled down: None (0–1), Low (1–2), Moderate (2–4), High (4–5), Very High (>5). Same structure as your "Pipe Breaks <16″" row but applied to plastic cohorts.
Pipe Leaks	No Leak (0–0.7), Leak Observed (0.7–1.3), Frequent Leaks (1.3–2)	0–2	Count / Category	Original range: 0–10 leaks in last 10 years. Scaled: No Leak (0), Leak Observed (1–2), Frequent Leaks (>2). Includes weeping joints, service tapping leaks, etc.
Pipe Pressure	Very Poor (0–1, 3.75–5), Poor (1–1.75, 3–3.75), Fair (1.75–2.25), Good (2.25–2.75), Excellent (2.75–3.5)	0–5	psi	Original range: 0–200 psi. Scaled down: Very Poor (0–40, 150–200), Poor (40–70, 135–150), Fair (70–90), Good (90–110), Excellent (110–150). Same U-shaped risk as metallic, capturing both under- and over-pressure.

Parameter	Membership Function Labels	Range	Typical Unit	Input Data Preparation
Pressure Transients	Low (0–1), Moderate (1–2), High (2–3)	0–3	Surge Index / psi	Original range: 0–100 psi transient magnitude (or a surge index). Scaled down: Low (0–20), Moderate (20–50), High (>50). Plastic is sensitive to frequent/high-amplitude transients.
Pipe Depth	Very Shallow (0–0.33), Shallow (0.33–0.67), Moderate (0.67–1.5), Deep (1.5–2.5), Very Deep (2.5–5)	0–5	Feet	Original range: 0–30 ft. Scaled: Very Shallow (0–2 ft), Shallow (2–4 ft), Moderate (4–10 ft), Deep (10–20 ft), Very Deep (20–30 ft). Same mapping as metallic.
Soil Particle Size	Coarse (0–1), Medium (1–1.5), Fine (1.5–2)	0–2	Size Category	Original categories: Coarse (gravel/sand), Medium (sandy silts), Fine (silts/clays). Scaled as in metallic table. Coarse/backfilled with large particles is often more problematic for PVC.
Bedding Condition	Poor (0–0.5), Fair (0.5–1.5), Good (1.5–2)	0–2	Condition Score	Original range: 0–2 from construction QA. Scaled down: Poor (0–0.5) = voids / point loading, Fair (0.5–1.5), Good (1.5–2) = uniform support. Particularly important for plastic deflection and long-term creep.
Soil Corrosivity	Critical (0–0.4), High (0.4–1), Medium (1–2), Low (2–4)	0–4	Index	Original range: 0–10,000. Scaled down as in metallic table: Critical (0–1000), High (1000–2500), Medium (2500–5000), Low (5000–10,000). For plastic, mainly relevant for metallic fittings and tracers.
Water Table Depth	Deep (1.5–3), Moderately Deep (0.75–1.5), Shallow (0–0.75)	0–3	Feet	Original range: 0–20 ft. Scaled down: Deep (>15 ft), Moderately Deep (5–15 ft), Shallow (0–5 ft). Shallow water table increases buoyancy and joint load risk for plastic.
Groundwater Fluctuation	Low (0–0.45), Moderate (0.45–0.9), High (0.9–3)	0–3	Feet	Original range: 0–20 ft seasonal fluctuation. Scaled down: Low (0–3 ft), Moderate (3–7.5 ft), High (7.5–20 ft). Same as metallic.
Frost Action	Low (2–3), Medium (1–2), High (0–1)	0–3	Index	Original range: 0–3. Scaled: Low (2–3), Medium (1–2), High (0–1). High = severe frost and heave, which can stress plastic joints.
Traffic Loading	Low (2–3), Medium (1–2), High (0–1)	0–3	Index	Original range: 0–3. Scaled: Low (2–3), Medium (1–2), High (0–1). High corresponds to heavy roads/rail over shallow plastic mains.
Installation Quality	Poor (0–0.5), Fair (0.5–1.5), Good (1.5–2)	0–2	Condition Score	Original range: 0–2 based on as-built QA/QC (compaction tests, inspection notes). Scaled: Poor (0–0.5) = recurring issues, Fair (0.5–1.5), Good (1.5–2).
Joint Type / Restraint	Poor (0–0.7), Fair (0.7–1.3), Good (1.3–2)	0–2	Condition / Design Category	Original range: 0–2. Scaled: Poor (0–0.7) = unrestrained / poor joint type for conditions, Fair (0.7–1.3), Good (1.3–2) = restrained or high-quality joint system.
Precipitation	Low (0–1), Average (1–2), High (2–3), Critical (3–4)	0–4	Inches/Year	Original range: 0–100 in/yr. Scaled as in metallic: Low (0–10), Average (10–30), High (30–60), Critical (60–100). Relevant for washouts / slope failures affecting plastic mains.
Temperature	Very Cold (0–1), Cold (1–2), Moderate (2–3), Warm (3–4), Very Hot (4–5)	0–5	°F	Original range: −20 to 150 °F soil or ambient temperature. Scaled down: Very Cold (<32°F), Cold (32–57°F), Moderate (57–82°F), Warm (82–107°F), Very Hot (>107°F). Plastic performance is temperature-sensitive.

Table C-4: Fuzzy membership-function labels, rescaled numerical ranges, and input-data preparation for PCCP pipe parameters in the LOF fuzzy-inference model.

Parameter	Membership Function Labels	Range	Typical Unit	Input Data Preparation
Pipe Age	New (0–0.27), Used (0.27–1.07), Old (1.07–1.6), Obsolete (1.6–2.67)	0–4	Years	Original range: 0–300 years. Scaled down: New (0–20 years), Used (20–80 years), Old (80–120 years), Obsolete (>120 years). Same structure as metallic pipes for consistency.
Pipe Vintage	Best Vintage (0–0.5), Relatively Good (0.5–1.5), Relatively Fair (1.5–3)	0–3	Years / Era	Original range: 1940–2025. Scaled down (example for PCCP): Best Vintage (1960–1985), Relatively Good (1985–2005), Relatively Fair (2005–2025). Exact cutoffs can be tuned to known good/bad PCCP eras.
Pipe External Protection	Good (2–3), Fair (1–2), Poor (0–1)	0–3	Condition Score	Same as metallic: Good = sound coating + cathodic protection, Fair = coating only or uncertain CP, Poor = bare steel cylinder / poor coating.
Wire-Break Severity	None (0–0.25), Low (0.25–0.75), Moderate (0.75–1.5), High (1.5–2.5), Very High (2.5–5)	0–5	Broken wire index	Original range: 0–100+ broken wires on a pipe. Scaled down: None (0), Low (1–5), Moderate (5–20), High (20–60), Very High (>60). EM vendor results binned into 0–5 index then fuzzified.
Pipe Leaks	No Leak (0–0.7), Leak Observed (0.7–1.3), Frequent Leaks (1.3–2)	0–2	Count / Category	Original range: 0–10 leaks in last 10 years. Scaled down: No Leak (0), Leak Observed (1–2 leaks), Frequent Leaks (>2 leaks). Same logic as metallic but applied to PCCP leak history.
Pipe Pressure	Very Poor (0–1, 3.75–5), Poor (1–1.75, 3–3.75), Fair (1.75–2.25), Good (2.25–2.75), Excellent (2.75–3.5)	0–5	psi	Original range: 0–200 psi. Scaled down: Very Poor (0–40, 150–200), Poor (40–70, 135–150), Fair (70–90), Good (90–110), Excellent (110–150). Same U-shaped risk pattern as metallic.
Soil Particle Size	Coarse (0–1), Medium (1–1.5), Fine (1.5–2)	0–2	Size Category	Original categories: Coarse (gravel/sand), Medium (sandy silts), Fine (silts/clays). Scaled: Coarse (0–1), Medium (1–1.5), Fine (1.5–2). Same coding as metallic.
Pipe Depth	Very Shallow (0–0.33), Shallow (0.33–0.67), Moderate (0.67–1.5), Deep (1.5–2.5), Very Deep (2.5–5)	0–5	Feet	Original range: 0–30 ft. Scaled down: Very Shallow (0–2 ft), Shallow (2–4 ft), Moderate (4–10 ft), Deep (10–20 ft), Very Deep (20–30 ft). Same mapping as metallic pipes.
Frost Action	Low (2–3), Medium (1–2), High (0–1)	0–3	Index	Original range: 0–3 frost severity (e.g., climate zone). Scaled: Low (2–3), Medium (1–2), High (0–1). High = severe frost / freeze–thaw.
Traffic Loading	Low (2–3), Medium (1–2), High (0–1)	0–3	Index	Original range: 0–3 traffic class (0 = arterial/rail, 3 = off-road). Scaled: Low (2–3), Medium (1–2), High (0–1). High corresponds to heavy roads/rails.
Water Table Depth	Deep (1.5–3), Moderately Deep (0.75–1.5), Shallow (0–0.75)	0–3	Feet	Original range: 0–20 ft from pipe invert/crown. Scaled down: Deep (>15 ft), Moderately Deep (5–15 ft), Shallow (0–5 ft). Same structure as metallic.
Soil Corrosivity	Critical (0–0.4), High (0.4–1), Medium (1–2), Low (2–4)	0–4	Index	Original range: 0–10,000 (soil corrosivity score). Scaled down: Critical (0–1000), High (1000–2500), Medium (2500–5000), Low

Parameter	Membership Function Labels	Range	Typical Unit	Input Data Preparation
				(5000–10,000). Same mapping as metallic, but note that for PCCP this primarily affects cylinder and wires.
Stray Currents	Absent (0–1), Present (1–2)	0–2	Presence	Original range: 0–2. Scaled down: Absent (0–1), Present (1–2), based on corrosion survey / presence near rail/CP systems.
Bedding Condition	Poor (0–0.5), Fair (0.5–1.5), Good (1.5–2)	0–2	Condition Score	Original range: 0–2 from construction QA/inspection. Scaled down: Poor (0–0.5), Fair (0.5–1.5), Good (1.5–2). Used for both metallic and PCCP.
Pipe Cracks	No Cracks (0–1), Hairline Crack (1–2), Minor Crack (2–3), Major Crack (3–4), Fracture Burst (4–5)	0–5	Severity Level	Original range: 0–5 qualitative severity. Scaled down directly: No Cracks (0–1), Hairline (1–2), Minor (2–3), Major (3–4), Fracture Burst (4–5). In PCCP this comes from CCTV/visual condition data.
Liner / Mortar Condition	Good (1.5–2), Fair (0.5–1.5), Poor (0–0.5)	0–2	Condition Score	Original range: 0–2 from internal inspection (mortar/lining). Scaled down: Poor (0–0.5) = severe spalling/delamination, Fair (0.5–1.5) = localized defects, Good (1.5–2) = intact mortar/lining.
Groundwater Fluctuation	Low (0–0.45), Moderate (0.45–0.9), High (0.9–3)	0–3	Feet	Original range: 0–20 ft seasonal fluctuation. Scaled down: Low (0–3 ft), Moderate (3–7.5 ft), High (7.5–20 ft). Same as your metallic definition.
Precipitation	Low (0–1), Average (1–2), High (2–3), Critical (3–4)	0–4	Inches/Year	Original range: 0–100 in/yr. Scaled down: Low (0–10), Average (10–30), High (30–60), Critical (60–100). Used as a proxy for erosion/runoff and surface washout risk above the line.

Table C-5: Fuzzy membership-function labels, rescaled numerical ranges, and input-data preparation for AC pipe parameters in the LOF fuzzy-inference model.

Parameter	Membership Function Labels	Range	Typical Unit	Input Data Preparation
Pipe Age	New (0–0.27), Used (0.27–1.07), Old (1.07–1.6), Obsolete (1.6–2.67)	0–4	Years	Original range: 0–100+ years. Scaled down: New (0–20), Used (20–50), Old (50–75), Obsolete (>75).
Pipe Vintage	Best Vintage (0–0.5), Relatively Good (0.5–1.5), Relatively Fair (1.5–3)	0–3	Years / Era	Original range: 1900–2025. Example: Best Vintage (1960–1985), Relatively Good (1945–1960, 1985–2000), Relatively Fair (pre-1945 or post-2000). Tuned to local knowledge on AC manufacturing eras.
Pipe Breaks (AC)	None (0–0.25), Low (0.25–0.375), Moderate (0.375–0.75), High (0.75–1.25), Very High (>1.25)	0–5	Count	Original range: 0–20 breaks. Scaled: None (0–1), Low (1–2), Moderate (2–4), High (4–5), Very High (>5).
Pipe Seepage / Exfiltration	None (0–0.5), Minor (0.5–1.5), Moderate (1.5–2.5), Severe (2.5–4)	0–4	Severity Level	Qualitative rating for AC seepage / weeping (including slow leaks through porous wall). Mapped from inspection or leak surveys (0–3 or 0–4) into 4 fuzzy bands.

Parameter	Membership Function Labels	Range	Typical Unit	Input Data Preparation
Pipe Remaining Wall / Structural Index	Excellent (4–5), Good (3–4), Fair (2–3), Poor (1–2), Critical (0–1)	0–5	Condition Score	Derived from structural condition assessment (e.g., coupon tests, flexural tests, or empirical rating). Scaled to 0–5 index with five bands.
Pipe Internal Lining	Yes (0–1), No (1–2)	0–2	Boolean	Encodes presence of internal lining (e.g., epoxy/CML). Same structure as metallic internal lining.
Pipe External Protection	Good (2–3), Fair (1–2), Poor (0–1)	0–3	Condition Score	Condition of wrapping, encasement, or external protection used for AC segments.
Pipe Pressure	Very Poor (0–1, 3.75–5), Poor (1–1.75, 3–3.75), Fair (1.75–2.25), Good (2.25–2.75), Excellent (2.75–3.5)	0–5	psi	Original range: 0–200 psi. Scaled as in metallic table; AC is sensitive to high operating pressures and surge.
Pressure Transients	Low (0–1), Moderate (1–2), High (2–3)	0–3	Surge Index / psi	Original range: 0–100 psi transient magnitude or qualitative surge index. Scaled: Low (0–20), Moderate (20–50), High (>50) or equivalent qualitative mapping.
Soil Particle Size	Coarse (0–1), Medium (1–1.5), Fine (1.5–2)	0–2	Size Category	Same coarse/medium/fine categorization as metallic; affects bedding quality and point loading on AC.
Soil Corrosivity / Aggressiveness	Critical (0–0.4), High (0.4–1), Medium (1–2), Low (2–4)	0–4	Index	Original range: 0–10,000. Includes pH, resistivity, and aggressive ions that can deteriorate AC or any metallic appurtenances.
Water Quality (Aggressiveness)	Good (0.3–2), Fair (−0.3 to 0.3), Poor (−2 to −0.3)	−2 to 2	Langelier / Aggressiveness Index	Same mapping as metallic pipes, emphasizing long-term soft/aggressive water that can affect AC leaching and durability.
Bedding Condition	Poor (0–0.5), Fair (0.5–1.5), Good (1.5–2)	0–2	Condition Score	Original range: 0–2. Poor = voids/rocks/point loads, Fair = mixed support, Good = uniform bedding/haunching.
Joint Condition	Poor (0–0.7), Fair (0.7–1.3), Good (1.3–2)	0–2	Condition Score	Original range: 0–2. For AC joints and couplings; poor joints increase leak and blowout risk.
Traffic Loading	Low (2–3), Medium (1–2), High (0–1)	0–3	Index	Original range: 0–3. High = major road/rail over shallow AC mains; same mapping as metallic.
Frost Action	Low (2–3), Medium (1–2), High (0–1)	0–3	Index	Original range: 0–3. Same frost index as other materials.
Water Table Depth	Deep (1.5–3), Moderately Deep (0.75–1.5), Shallow (0–0.75)	0–3	Feet	Original range: 0–20 ft. Shallow groundwater can increase external loading and leak detection difficulty for AC.
Groundwater Fluctuation	Low (0–0.45), Moderate (0.45–0.9), High (0.9–3)	0–3	Feet	Original range: 0–20 ft seasonal fluctuation. Scaled as in metallic table.
Temperature	Very Cold (0–1), Cold (1–2), Moderate (2–3), Warm (3–4), Very Hot (4–5)	0–5	°F	Original range: −20 to 150 °F. Scaled: Very Cold (<32 °F), Cold (32–57 °F), Moderate (57–82 °F), Warm (82–107 °F), Very Hot (>107 °F).
Precipitation	Low (0–1), Average (1–2), High (2–3), Critical (3–4)	0–4	Inches/Year	Original range: 0–100 in/yr. Same bins as metallic, capturing washout/landslide risks for AC in steep terrain.

Appendix D: Fuzzy Inference System and Genetic Algorithm Evaluation

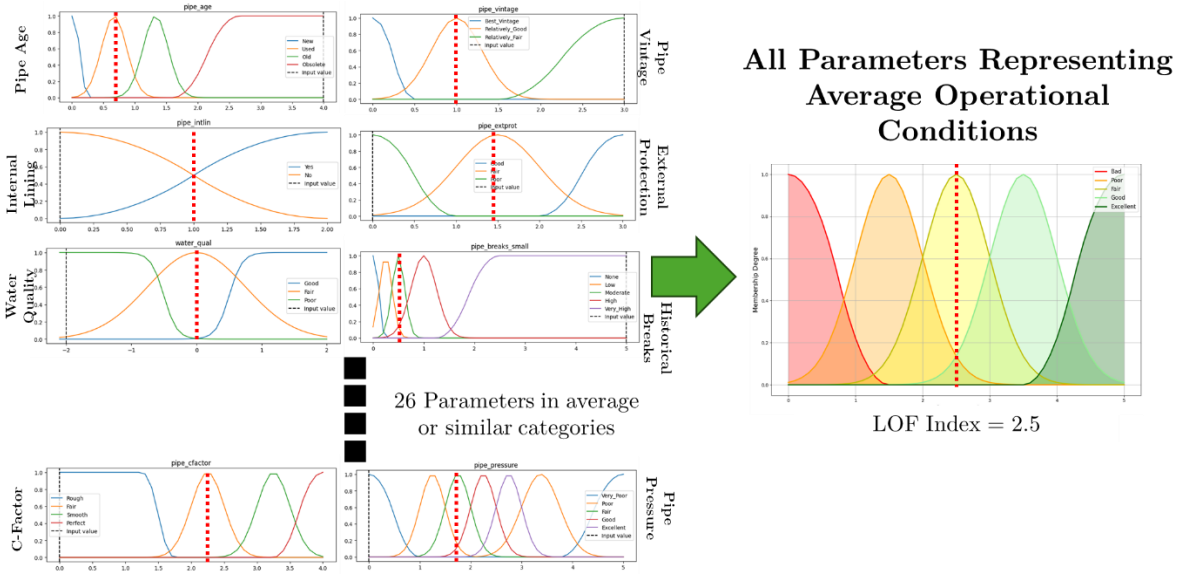

Figure D-1: Representative check in LOF models showing output in the Fair category for input values representing average conditions

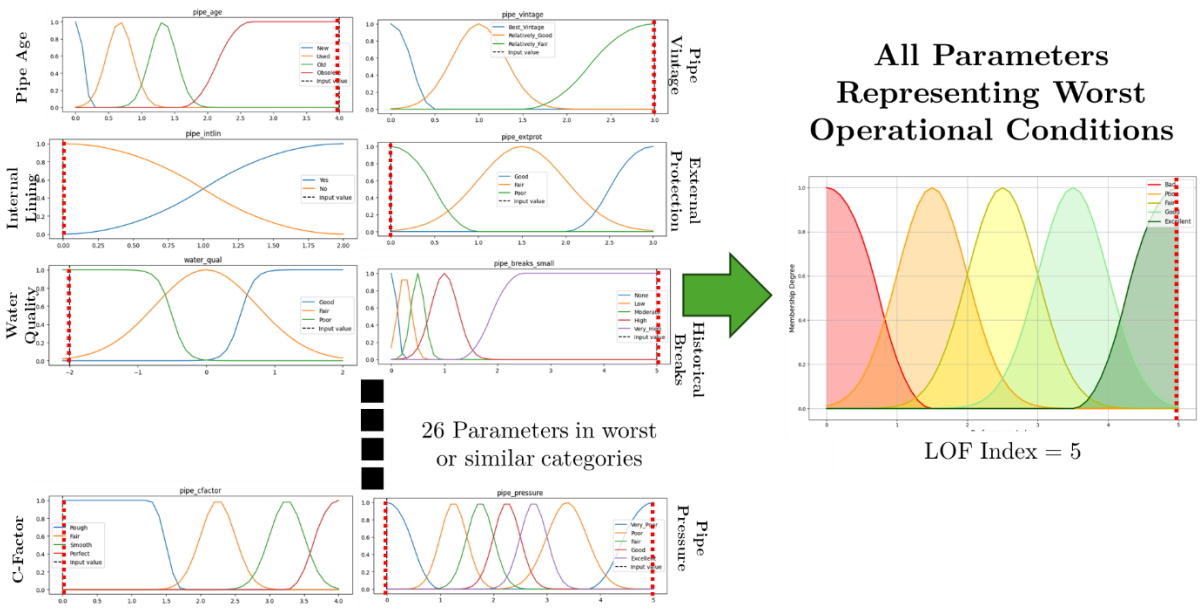

Figure D-2: Representative check in LOF models showing output in the Bad category for input values representing worst performance conditions

Table D-1: Example GA-selected annual portfolio from the synthetic project set (balanced weights, seed = 1), showing project ID, corridor type, material, risk score, equity score, cost, and (where available) length and risk-contribution. Projects are sorted by their contribution to total risk reduction, with ties broken by equity score.

Project ID	Corridor Type	Material	Length (miles)	Risk Score	Equity Score	Cost ($M)	Risk Contribution
89	CBD core	CI	1.02	98.71	69.68	0.48	100.77
195	CBD core	CI	0.91	82.77	95.84	0.38	75.42
193	CBD core	CI	0.78	92.71	82.14	0.74	72.20
49	Industrial fringe	CI	0.85	80.65	78.04	0.51	68.48
6	Mixed-density urban	PVC	1.22	47.45	54.65	0.51	57.93
204	CBD core	CI	0.63	87.36	88.71	0.46	55.46
117	Mixed-density urban	CI	0.76	68.48	74.47	0.11	51.88
109	CBD core	AC	0.62	82.89	95.24	0.17	51.25
61	Industrial fringe	CI	0.56	90.39	86.32	0.43	50.38

29	CBD core	DI	0.61	82.47	86.51	0.20	50.29
153	CBD core	DI	0.53	92.33	90.91	0.63	48.59
73	Mixed-density urban	PVC	1.03	46.74	52.23	0.27	47.96
176	CBD core	DI	0.46	100.00	79.77	0.70	46.01
115	CBD core	CI	0.52	86.68	79.01	0.16	45.20
128	CBD core	CI	0.50	88.60	77.94	0.60	44.03
196	CBD core	CI	0.58	74.50	88.51	0.42	43.05
129	CBD core	AC	0.45	93.58	77.23	0.20	41.90
218	CBD core	DI	0.44	92.78	71.63	0.56	40.82
181	Mixed-density urban	DI	0.52	77.92	55.50	0.13	40.47
131	CBD core	AC	0.49	81.37	70.81	0.23	40.12
169	Suburban loop	CI	0.65	58.66	34.30	0.28	38.22
64	Mixed-density urban	DI	0.57	66.90	60.38	0.17	37.96
57	Suburban loop	DI	0.87	43.19	46.57	0.29	37.43
102	Industrial fringe	AC	0.42	88.90	72.33	0.11	37.10
152	Mixed-density urban	DI	0.61	59.79	68.17	0.24	36.33
63	Industrial fringe	CI	0.37	95.27	72.79	0.13	35.68
211	Industrial fringe	PVC	0.61	57.82	91.13	0.49	35.45
150	CBD core	Steel	0.64	55.57	93.76	0.26	35.30
91	CBD core	DI	0.42	80.59	85.20	0.21	34.06
79	Industrial fringe	Steel	0.47	71.41	70.76	0.39	33.67
165	Industrial fringe	DI	0.37	91.40	67.87	0.10	33.48
156	Industrial fringe	CI	0.37	89.88	76.11	0.20	32.92
58	Industrial fringe	AC	0.47	69.51	76.97	0.19	32.53
38	Suburban loop	CI	0.51	63.01	44.41	0.11	32.21
186	Industrial fringe	PVC	0.58	54.92	85.83	0.20	32.06
189	Mixed-density urban	CI	0.48	66.19	78.48	0.19	31.67
10	Suburban loop	CI	0.69	44.88	41.06	0.57	30.90
170	Mixed-density urban	PVC	0.99	31.19	75.89	0.25	30.84
215	Mixed-density urban	CI	0.47	64.38	81.32	0.17	30.47
41	Suburban loop	CI	0.46	64.94	63.50	0.24	30.11
82	Suburban loop	CI	0.61	48.99	55.85	0.41	29.85
86	CBD core	DI	0.34	86.32	80.57	0.19	29.62
124	Suburban loop	PVC	0.72	39.76	47.11	0.12	28.56

65	Industrial fringe	AC	0.39	73.24	80.92	0.10	28.46
51	Suburban loop	CI	0.51	55.75	44.00	0.22	28.42
93	Suburban loop	DI	0.51	55.15	32.40	0.13	28.07
216	Mixed-density urban	CI	0.49	56.19	54.18	0.36	27.60
208	Industrial fringe	Steel	0.39	69.83	64.83	0.28	27.41
175	CBD core	CI	0.38	72.81	79.84	0.35	27.36
136	CBD core	AC	0.29	93.94	86.64	0.20	27.19
114	Industrial fringe	PVC	0.49	55.46	84.45	0.32	26.93
11	Suburban loop	CI	0.62	42.56	35.77	0.18	26.55
160	Mixed-density urban	Steel	0.52	51.04	79.88	0.18	26.43
66	Industrial fringe	DI	0.37	71.64	77.85	0.24	26.17
161	CBD core	AC	0.29	88.53	71.33	0.16	25.84
62	Suburban loop	DI	0.45	52.82	47.72	0.10	23.87
217	CBD core	CI	0.25	93.50	70.01	0.21	23.63
87	Mixed-density urban	DI	0.32	73.84	79.29	0.16	23.61
135	Industrial fringe	AC	0.28	84.77	69.32	0.40	23.59
138	Suburban loop	PVC	0.93	24.89	57.03	0.45	23.19
20	Industrial fringe	CI	0.26	85.87	72.97	0.21	22.64
28	CBD core	DI	0.29	77.75	73.58	0.14	22.50
84	CBD core	DI	0.28	78.15	72.78	0.24	22.23
200	Suburban loop	PVC	0.61	36.42	46.47	0.22	22.05
139	Mixed-density urban	AC	0.40	54.00	47.02	0.22	21.76
95	Suburban loop	AC	0.48	45.08	65.80	0.12	21.51
105	Mixed-density urban	PVC	0.51	42.21	52.74	0.28	21.44
205	CBD core	AC	0.28	75.92	81.21	0.18	21.23
103	Suburban loop	DI	0.39	53.01	52.51	0.19	20.75
213	Suburban loop	PVC	0.53	38.42	51.31	0.10	20.51
8	Mixed-density urban	AC	0.32	61.30	53.23	0.08	19.34
2	Suburban loop	Steel	0.50	37.10	50.36	0.25	18.66
123	CBD core	Steel	0.30	61.24	79.29	0.22	18.50
16	Suburban loop	PVC	0.59	30.97	65.92	0.24	18.36
68	Suburban loop	AC	0.41	44.85	53.83	0.18	18.34
1	Mixed-density urban	DI	0.32	56.74	75.95	0.27	18.15
53	Mixed-density urban	CI	0.29	61.97	50.47	0.16	18.02

188	CBD core	Steel	0.28	63.62	92.68	0.18	17.89
149	Suburban loop	DI	0.32	54.68	63.48	0.12	17.76
159	Mixed-density urban	DI	0.26	66.58	79.62	0.09	17.58
162	Mixed-density urban	PVC	0.34	50.97	70.44	0.11	17.50
111	Suburban loop	PVC	0.70	24.99	65.29	0.35	17.44
17	Industrial fringe	AC	0.20	84.47	73.03	0.13	17.31
202	Mixed-density urban	CI	0.21	80.70	48.78	0.15	17.05
26	Suburban loop	PVC	0.50	33.07	35.40	0.20	16.64
46	Mixed-density urban	CI	0.20	79.43	79.82	0.10	16.26
100	Mixed-density urban	PVC	0.35	46.05	58.35	0.17	16.17
32	Mixed-density urban	PVC	0.49	33.15	72.43	0.22	16.16
59	CBD core	Steel	0.27	58.42	86.62	0.25	15.68
74	Suburban loop	CI	0.28	52.53	59.56	0.11	14.69
142	Suburban loop	CI	0.33	43.73	39.49	0.06	14.42
110	Mixed-density urban	DI	0.19	75.85	79.26	0.05	14.40
4	Industrial fringe	PVC	0.30	47.47	88.74	0.20	14.29
36	Suburban loop	PVC	0.50	26.08	32.61	0.17	13.11
206	Suburban loop	PVC	0.35	36.94	68.43	0.18	12.94
147	Suburban loop	PVC	0.84	15.19	69.05	0.27	12.84
75	CBD core	Steel	0.17	68.09	80.28	0.31	11.81
140	CBD core	AC	0.14	76.68	81.98	0.21	10.98
151	Mixed-density urban	PVC	0.23	40.30	48.93	0.09	9.35
173	Suburban loop	PVC	0.32	23.84	68.13	0.24	7.73
126	Suburban loop	PVC	0.19	38.43	62.82	0.12	7.46
81	Industrial fringe	Steel	0.13	47.81	87.41	0.06	6.28
199	Suburban loop	DI	0.12	54.27	46.38	0.03	6.27
201	Mixed-density urban	Steel	0.10	53.72	60.53	0.10	5.37
171	Suburban loop	PVC	0.20	27.51	37.70	0.05	5.37
71	Suburban loop	PVC	0.28	16.62	47.52	0.11	4.70
141	Suburban loop	PVC	0.21	15.92	53.39	0.08	3.39

Appendix E: MLP Hyperparameters and Training/Testing

Table E-1: Screening Logistic Regression Model Architecture

Setting	Value / Policy
Task	5-class softmax (multinomial)
Solver	lbfgs
Penalty	L2
C (inverse reg.)	1.0
Max iterations	2000
Class weights	None (screening used pristine train fit)
Feature scaling	Standardize (μ, σ)

Table E-2: Screening SVM RBF Kernel Model Architecture

Setting	Value / Policy
Kernel	RBF (gamma="scale")
C	2.0
Probability estimates	No (not needed for screening metrics)
Class weights	None (screening)
Feature scaling	Standardize (μ, σ)

Table E-3: Screening Random Forest Model Architecture

Setting	Value / Policy
Trees (n_estimators)	400
Max depth	None (grow to purity; implicit regularization via averaging)
Min samples split	2

Setting	Value / Policy
Min samples leaf	1
Max features	sqrt
Bootstrap	True
Class weights	None (screening)

Table E-4: Screening XGBoost (Multiclass) Model Architecture

Setting	Value / Policy
Objective	multi:softmax (num_class=5)
Trees (n_estimators)	600
Max depth	5
Learning rate	0.05
Subsample	0.90
Colsample_bytree	0.90
L2 reg (lambda)	Default
L1 reg (alpha)	Default
Eval metric	mlogloss (internal)
Class weights	None (screening)
Feature scaling	Standardize (μ, σ)

Table E-5: Screening Shallow MLP Model Architecture

Setting	Value / Policy
Hidden sizes	(32, 24, 16)
Activation	ReLU
Normalization	BatchNorm (LayerNorm for tiny-N)
Dropout	0.08
Optimizer	AdamW
Learning rate	8e-4
Weight decay	4e-4
Gradient clip	$\|g\|_2 \leq 1$
Label smoothing	0.02 (raise to 0.03–0.04 for tiny-N only)
Batch size	256 (smaller for tiny-N)

Setting	Value / Policy
Early stopping metric	Macro-F1 (val split inside train)
EMA of weights	Yes (decay≈0.995)
Class-imbalance policy	None in screening (no weights/oversampling)

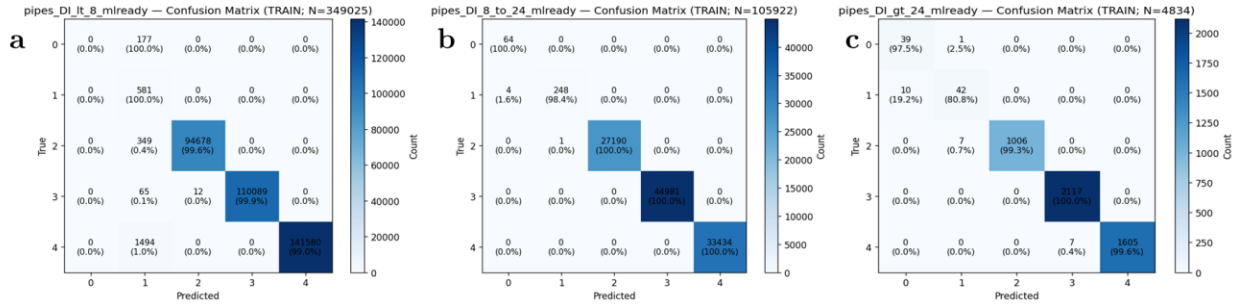

Figure E-1: Training confusion matrices for the DI LOF models by diameter cohort: (a) < 8 in, (b) 8–24 in, and (c) > 24 in. Axes show true vs. predicted LOF band (0–4); cell shading and labels report counts and row-wise percentages, with almost all mass on the main diagonal.

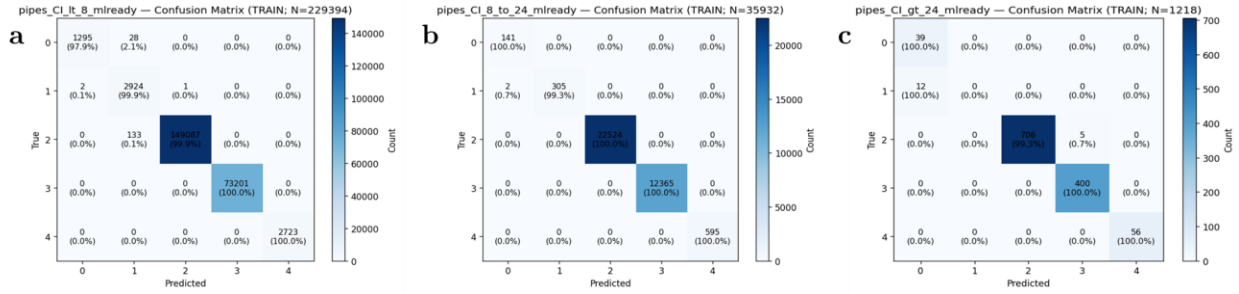

Figure E-2: Training confusion matrices for the CI LOF models by diameter cohort: (a) < 8 in, (b) 8–24 in, and (c) > 24 in. The deep MLP models reproduce the true LOF band with near-perfect diagonal dominance in all cohorts.

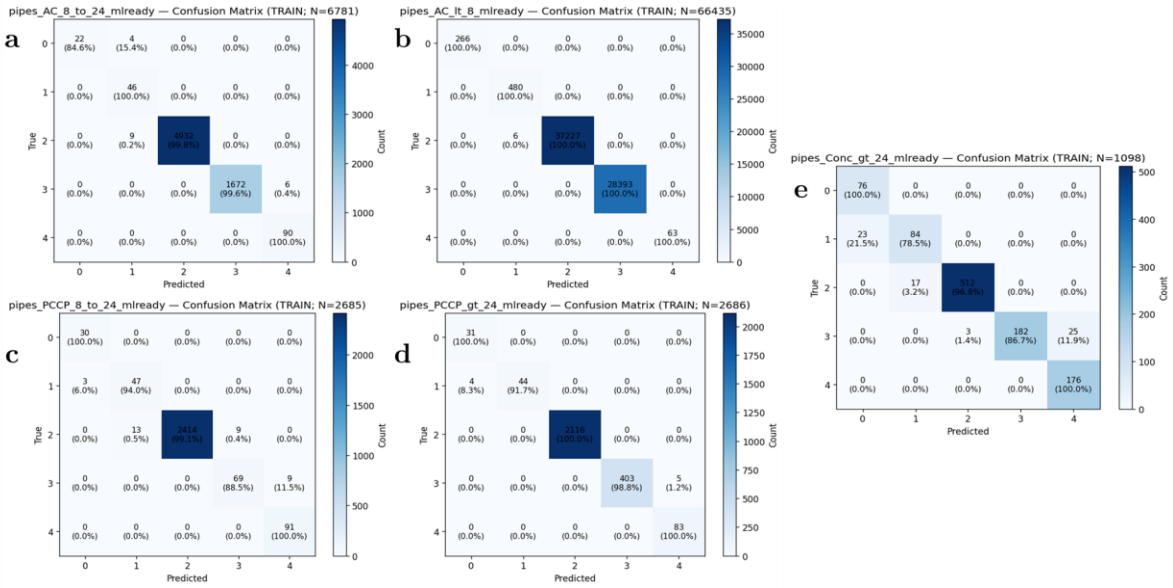

Figure E-3: Training confusion matrices for cementitious pipe LOF models: (a) AC 8–24 in, (b) AC < 8 in, (c) PCCP 8–24 in, (d) PCCP > 24 in, and (e) RCP, RCCP and BWP > 24 in. For each material–diameter cohort the predicted LOF bands align closely with the training labels, with only minor off-diagonal error.

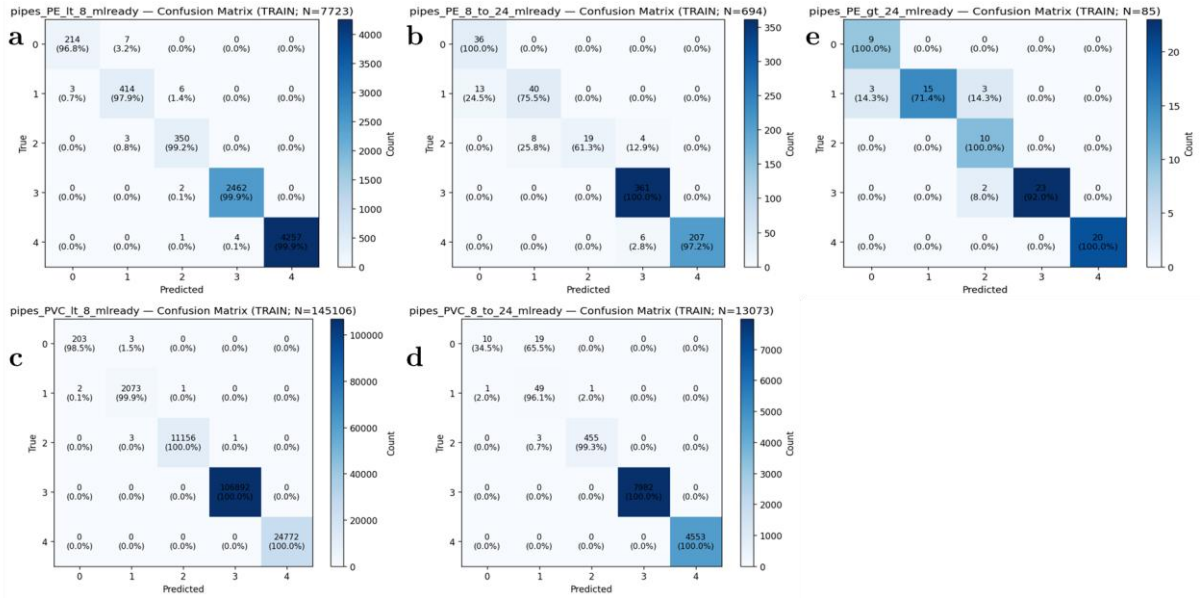

Figure E-4: Training confusion matrices for plastic pipe LOF models: (a) PE < 8 in, (b) PE 8–24 in, (c) PVC < 8 in, (d) PVC 8–24 in, and (e) PE > 24 in. All plastic cohorts show high

agreement between true and predicted LOF bands, with small off-diagonal leakage in the medium and large-diameter PE cohorts.

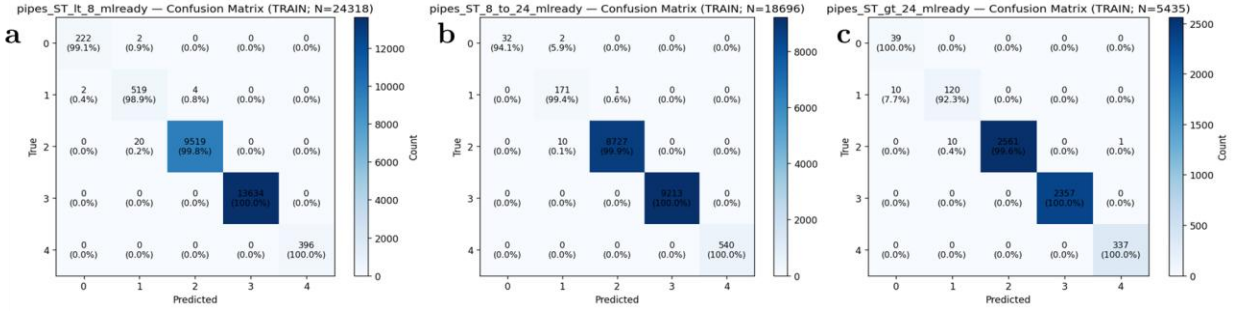

Figure E-5: Training confusion matrices for the steel LOF models by diameter cohort: (a) < 8 in, (b) 8–24 in, and (c) > 24 in. The student models achieve near-perfect recovery of the training LOF labels across all three steel cohorts.

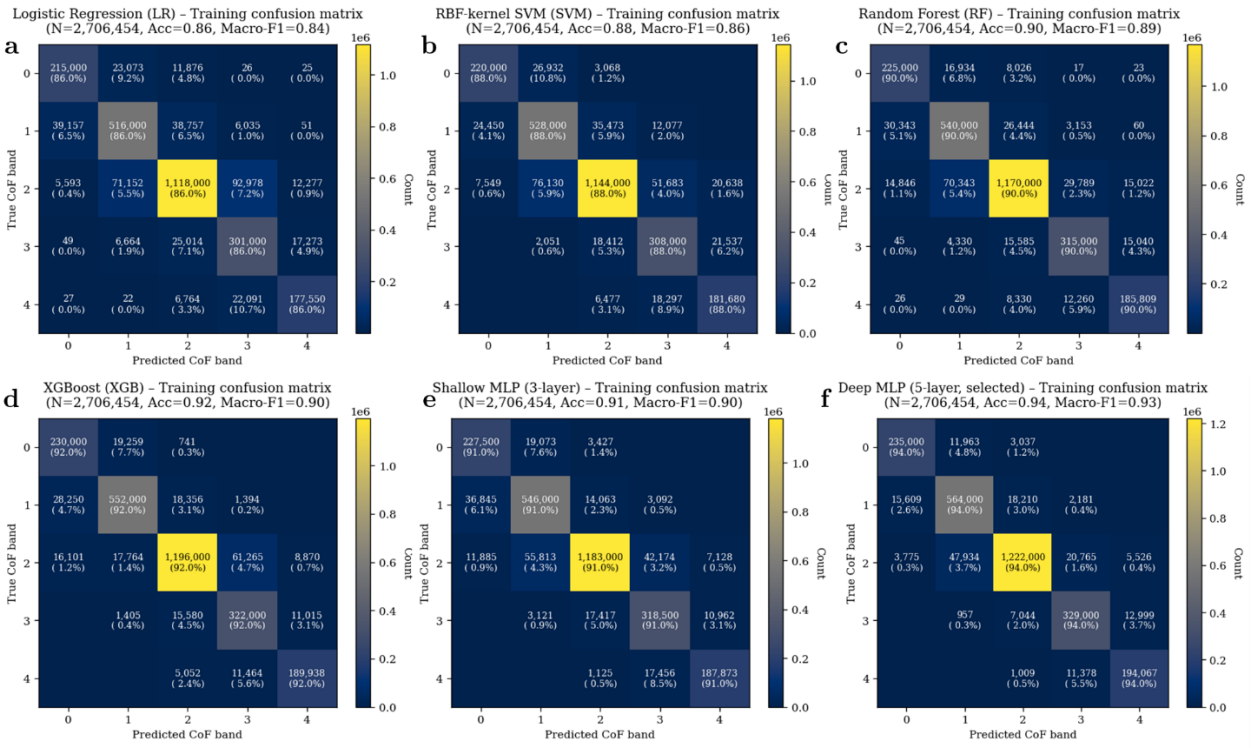

Figure E-6: Training confusion matrices for candidate COF classifiers on the pooled segment-level dataset (N ≈ 2.7 million). Panels show true vs. predicted COF bands (0–4) for (a) Logistic

Regression, (b) RBF-kernel SVM, (c) Random Forest, (d) XGBoost, (e) shallow 3-layer MLP, and (f) deep 5-layer MLP (selected). Cell shading and labels give counts and row-wise percentages; all models exhibit strong diagonal dominance, with the deep MLP achieving the highest overall accuracy (0.94) and macro-F1 (0.93).

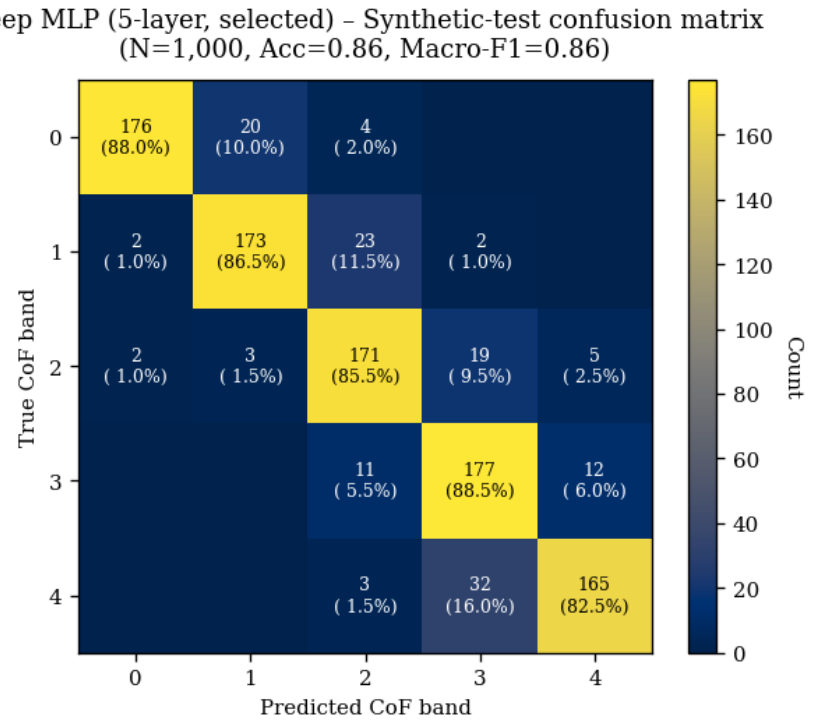

Figure E-7: Confusion matrix for the selected 5-layer deep MLP COF model on the synthetic test set (N = 1,000), showing good recovery of the five COF bands (overall accuracy = 0.86, macro-F1 = 0.86) with most misclassifications confined to adjacent bands.

Appendix F: Expert Agreement Test and Feedback Form

Instructions

Instructions for Model Result Verification

1. Overview of This Workbook

    a. Sheet “Instructions”: Contains these guidelines for interpreting and providing feedback on the model results.

    b. Sheet “Performance Model”: Displays a Performance Index (0–5) for each pipe, where 0 = “Bad” and 5 = “Excellent.”

    c. Sheet “Consequence of Failure Model”: Shows a Consequence of Failure Index (0–5) for each pipe, where 0 = “Insignificant” and 5 = “Catastrophic.”

    d. Sheet “Risk-Based Renewal Model”: Lists the priority ranking for each pipe in terms of renewal. Pipes not selected for renewal display “Not Selected.” If selected, it shows the ranking among all selected pipes (e.g., “Rank 8 of 125”).

e. Sheet "Dictionary": Provides detailed definitions of what each index category (for Performance and Consequence of Failure) and each renewal priority classification means.

2. How to Interpret Each Model's Output

   a. Performance Model (0–5): 0 (Bad) to 5 (Excellent) range. These results estimate the current performance of the pipe, considering factors like structural condition, functionality, etc.

   b. Consequence of Failure Model (0–5): 0 (Insignificant) to 5 (Catastrophic). This index evaluates how severe the impact would be if that pipe fails—financial, social, environmental consequences, etc.

   c. Risk-Based Renewal Prioritization: If a pipe was "Not Selected for Renewal," it means the model did not prioritize it under the given constraints (budget, risk threshold, etc.).

   d. If a pipe was selected, you will see a rank (e.g., "Rank 8 of 125"). The lower the rank, the higher the priority for renewal.

For further explanation of each category or index level (e.g., what qualifies as "Excellent" vs. "Good"), refer to the "Dictionary" sheet.

3. Reviewing the Results and Providing Feedback

    a. Locate the Pipe IDs in each model's respective sheet:

        i. Use the Pipe ID column to match pipes in your own database.

        ii. Confirm you are looking at the correct pipe(s) before commenting.

    b. Assess Whether the Model Output Matches Your Observations:

        i. For the Performance Model: Does the result from the model in index (0–5) align with your understanding of the pipe's structural and functional performance?

        ii. For the Consequence of Failure Model: Does the 0–5 rating match your internal assessment of potential impacts if this pipe fails?

        iii. For the Risk-Based Renewal Model: Do you agree that a particular pipe was/should (or was not/ should not) selected for renewal, or that it has the assigned rank among the selected pipes?

    c. Fill Out the "Comments (If Disagreement)" Columns:

        i. In each sheet (Performance, Consequence of Failure, Risk-Based Renewal), there is a column named "Comments (If Disagreement)"

where you can indicate Agree or Disagree with the assigned index/rank.

ii. If you select Disagree, please provide brief comments in the "Comments" column to clarify why. For example:

a) "Field inspection shows severe corrosion, so I expect a higher consequence rating."

b) "Budget constraints require renewal sooner."

4. Reference the "Dictionary" Sheet for Index Definitions: All Performance and Consequence of Failure categories (0–5) and renewal prioritization criteria are explained in detail, so you can see how each index level and parameters are defined. You can use it to interpret the model results

5. Sending Back Your Input: Save your completed Excel file with feedback in the "Comments (If Disagreement)" columns. Return the file to us so we can compare your expert evaluation with the model output and make any necessary adjustments for Piloting Phase 2 and move forward to Piloting Phase 3

The "Dictionary" sheet contains the output index definitions.

Table F-1: Expert evaluation of representative Student LOF model scenarios for Utility A (Pacific Northwest). Each row shows a constructed pipe scenario, the Student LOF score and band,

and the utility expert's verdict and comments on whether the predicted likelihood-of-failure level is reasonable.

Pipe ID	Scenario	Student LOF (0–5)	Label	Expert verdict	Expert comment
6295976	Greater than 40 year old 10″ unprotected DI pipe buried in high soil corrosivity.	2.75	Moderate	Agree	–
6369901	<20 year old >20″ diameter Polywrapped DI pipe with groundwater fluctuations	1.36	Low	Agree	–
598842	Greater than 40 year old 8″ diameter DI pipe in moderate soil corrosivity	1.67	Low	Agree	–
530935	>50 year old CI pipe with low wall thickness in poor drainage soil.	3.20	High	Agree	–
559643	>50 year old 6″ diameter CI pipe in moderately corrosive soil	2.51	Moderate	Agree	–
499063	<50 year old 6″ diameter CI pipe in high soil corrosivity (<1000 Ohm-cm) and low C-factor (<100) indicating roughness	2.75	Moderate	Agree	–
526598	>36″ diameter Concrete pipe in high steel soil corrosivity (<1500 ohm-cm), poor bedding and groundwater table fluctuation	2.97	Moderate	Agree	–
471678	Greater than 40 year old >36″ diameter Concrete in Medium-High corrosivity towards metals with minor cracking	2.00	Moderate	Agree	–
241285	Concrete pipe with cement mortar coating deterioration in fluctuating groundwater table condition	3.50	High	Agree	–
237958	>60 year old concrete pipe with spalling in high frost-action soil	3.11	High	Agree	–
6028376	>50 year old Concrete pipe <50″ diameter with high pressure (>120 psi) in high-precipitation areas and moderate to high corrosivity	2.50	Moderate	Agree	–
450023	PVC pipe with significant ovality (>15%) in clayey soil	4.95	Very High	Agree	–
600011	PVC pipe subjected to prolonged high internal pressure	2.75	Moderate	Disagree	Inactive as of 11/4/2024
6405488	>30 year old PVC pipe in clayey soil with no manufacturing defects and no historical breaks	2.53	Moderate	Disagree	–
6270402	<30 year old PVC pipe in clayey soil with no historical failures and good bedding condition	0.84	Very Low	Agree	–
6191865	HDPE pipe in a high-temperature differential environment (internal and external)	3.47	High	Agree	–
6191864	Greater than 30 year old HDPE pipe with no deterioration	3.02	High	Agree	–
6269604	Young <10 year old HDPE pipe in moderate pressure zone and bedding condition	0.27	Very Low	Agree	–
565297	>50 year old Steel pipe with external corrosion	4.06	Very High	Agree	–
597111	>72″ Steel pipe with buckling or other deformation in clayey soil near arterial roads	3.34	High	Disagree	20″ not >72″
6202252	<30 year old >48″ diameter Steel pipe in shallow water-table depth and high-precipitation area	1.67	Low	Disagree	is 48″

Pipe ID	Scenario	Student LOF (0–5)	Label	Expert verdict	Expert comment
288836	>40 year old 4″ diameter AC pipe in moderate-high precipitation with multiple historical failures	3.14	High	Agree	–
288837	40–50 year old 4″ diameter AC pipe in area with high groundwater fluctuation and moderate-high precipitation	1.87	Low	Disagree	–
332054	40–50 year old 6″ diameter AC pipe in area with low traffic loading and fair bedding condition	1.87	Low	Disagree	–

Table F-2: Expert evaluation of representative Student LOF model scenarios for Utility B (Southeast). The table compares Student LOF scores and bands against utility judgements, highlighting both agreements and cases where the asset could not be located or the expert disagreed.

Pipe ID	Scenario	Student LOF (0–5)	Label	Expert verdict	Expert comment
2378727	Greater than 40 year old 10″ unprotected DI pipe buried in high soil corrosivity.	2.75	Moderate	Agree	–
2760670	<20 year old >20″ diameter Polywrapped DI pipe with groundwater fluctuations	1.25	Low	Agree	–
435500	Greater than 40 year old 8″ diameter DI pipe in moderate soil corrosivity	0.75	Very Low	Agree	–
79742	>50 year old CI pipe with low wall thickness in poor drainage soil.	2.50	Moderate	??	Unable to identify pipe to research
43699	>50 year old 6″ diameter CI pipe in moderately corrosive soil	1.25	Low	Agree	–
2857907	<50 year old 6″ diameter CI pipe in high soil corrosivity (<1000 Ohm-cm) and low C-factor (<100) indicating roughness	2.75	Moderate	Agree	–
2864480	Greater than 40 year old >36″ diameter Concrete in Medium-High corrosivity towards metals with minor cracking	2.75	Moderate	??	Unable to identify pipe to research
3264767	>36″ diameter Concrete pipe in high steel soil corrosivity (<1500 ohm-cm), poor bedding and groundwater table fluctuation; >60 year old concrete pipe with spalling in high frost-action soil; pipe with cement-mortar coating deterioration in fluctuating groundwater table condition	4.50	Very High	Field Ops in agreeance with model (failure in 2023)	–
637821	>72″ Concrete pipe <50″ diameter with high pressure (>120 psi) in high-precipitation areas and moderate to high corrosivity	2.00	Moderate	Agree	–
104470	PVC pipe with significant ovality (>15%) in clayey soil	2.50	Moderate	Agree	–
43291	PVC pipe subjected to prolonged high internal pressure	3.67	High	Agree	–
2912592	>30 year old PVC pipe in clayey soil with no manufacturing defects and no historical breaks	2.25	Moderate	Agree	–

Pipe ID	Scenario	Student LOF (0–5)	Label	Expert verdict	Expert comment
123573	<30 year old PVC pipe in clayey soil with no historical failures and good bedding condition	1.33	Low	Agree	–
3227216	HDPE pipe in a high-temperature differential environment (internal and external)	3.67	High	Disagree with model	Internal crews replaced galvanized with Municipex; pipe felt to be at low risk of failure
3418846	Greater than 30 year old HDPE pipe with no deterioration	1.87	Low	Agree	–
3303501	Young <10 year old HDPE pipe in moderate pressure zone and bedding condition	1.33	Low	Agree	–
4015477	>50 year old Steel pipe with external corrosion near electrified railway lines	1.25	Low	??	Unable to identify pipe to research
413889	>72″ Steel pipe with buckling or other deformation in clayey soil near arterial roads	2.00	Moderate	Agree	–
4815068	<30 year old >72″ diameter Steel pipe in shallow water-table depth and high-precipitation area	2.50	Moderate	Agree	–
44835	>50 year old 6″ diameter AC pipe in moderate-high precipitation with multiple historical failures	4.90	Very High	Agree	–
65763	30–50 year old 8″ diameter AC pipe in area with high groundwater fluctuation and moderate-high precipitation	2.75	Moderate	Agree	–
115547	40–50 year old 6″ diameter AC pipe in area with low traffic loading and fair bedding condition	1.87	Low	??	Unable to identify pipe to research

Table F-3: Expert evaluation of representative Student LOF model scenarios for Utility C (Coastal West). For each pipe segment, the Student LOF score and band are compared with expert ratings and comments, including notes on data issues (e.g., material mis-labelling) and disagreement cases.

Pipe ID	Scenario	Student LOF (0–5)	Label	Expert verdict	Expert comment
45096D81	Greater than 40 year old 8″ unprotected DI pipe buried in high soil corrosivity.	2.50	Moderate	agree	–
P47445-ED121	<20 year old >20″ diameter Polywrapped DI pipe with groundwater fluctuations	1.25	Low	agree*	* 12″ pipe
43961D61	Greater than 40 year old 6″ diameter DI pipe in moderate soil corrosivity and high-precipitation area	2.17	Moderate	agree	–
E-24792C161	>50 year old CI pipe with low wall thickness in poor drainage soil.	2.50	Moderate	agree	–

Pipe ID	Scenario	Student LOF (0–5)	Label	Expert verdict	Expert comment
E-28506-AC62	>50 year old 6″ diameter CI pipe in moderately corrosive soil	1.33	Low	agree	–
E-18031C41	>90 year old CI pipe with more than 2 failures in last 5 years and greater than 3 overall	4.90	Very High	agree	–
E-26721-AC61	<50 year old 6″ diameter CI pipe in high soil corrosivity (<1000 Ohm-cm) and low C-factor (<100) indicating roughness	2.75	Moderate	fair-poor	–
CY-D-37L481	>36″ diameter Concrete pipe 95 years old in high steel soil corrosivity (<1500 ohm-cm), poor bedding and groundwater table fluctuation	2.75	Moderate	fair-good	–
E-29243-EL481	Greater than 40 year old >36″ diameter Concrete in Medium-High corrosivity towards metals with minor cracking	2.25	Moderate	fair-good	–
38082-AT5410	>50 years old, 54″ PCCP in low-corrosivity soil with no historical failures	2.25	Moderate	agree	–
35966-BT2415	>60 year old 24″ PCCP with no historical failures, average precipitation, fair bedding conditions and moderately corrosive soil	2.75	Moderate	Our maps mis-labeled material. This is a welded steel pipe	AWWA C303 pipe (EBMUD BMAPs incorrect)
35966-BT2437	>50 year old PCCP 24″ with total 1 historical break but none in the last 5 years and currently in low-corrosivity conditions	3.12	High	Our maps mis-labeled material. This is a welded steel pipe	AWWA C303 pipe (EBMUD BMAPs incorrect); this pipe is in a high seismic activity zone
48382N21	<30 years old PVC pipe with more than 5 historical failures and about 1 in the past 5 years	4.70	Very High	agree	–
P45222-BN81	<10 years old 8″ PVC pipe operating under high internal pressure	2.25	Moderate	agree	–
48848N89	30–40 year old PVC pipe in clayey soil with normal pressure conditions, no manufacturing defects and no historical breaks	1.25	Low	agree	–
P43665H83	<20 years old 8″ HDPE with historical failures (none in the last 5 years), clayey soil, moderate pressure and highly fluctuating groundwater levels	3.31	High	disagree	expects saddles falling off; connection not main
P43103H84	<20 years old 8″ HDPE with historical failures (none in the last 5 years), clayey soil, moderate pressure and shallow groundwater-table depth	3.09	High	disagree	–
P41874H201	>20 year old 8″ HDPE pipe in clayey soil, normal pressure, average precipitation and buried at shallow depth	2.17	Moderate	agree	–

Pipe ID	Scenario	Student LOF (0–5)	Label	Expert verdict	Expert comment
P43665H81	<20 year old 8″ HDPE pipe in high-precipitation area, highly fluctuating groundwater table and shallow depth	1.33	Low	agree	–
40076S161	>50 year old 16″ Steel pipe in moderately corrosive soil	0.75	Very Low	agree	–
E-20097-KS362	>80 year old 36″ Steel pipe in moderately corrosive soil and minor pitting corrosion	2.00	Moderate	agree	–
33349S303	>60 year old 30″ diameter Steel pipe with historical breaks (none in the last 5 years), moderate internal roughness and minor pitting	4.33	Very High	agree*	Mostly in bad shape due to seismic activity
36671A61	>60 year old 6″ diameter AC pipe with multiple historical failures (3 in the past 5 years)	4.90	Very High	agree	–
E-31728A610	>60 year old 6″ diameter AC pipe with 1 historical failure (not in the past 5 years)	3.25	High	agree	–
33961A121	>60 year old 12″ diameter AC pipe with no historical failures, moderate internal roughness and low corrosivity to cementitious materials	2.25	Moderate	agree	–

Table F-4: Expert evaluation of Student COF scores for ten representative high-impact scenarios at Utility A (Pacific Northwest). The table lists scenario narratives, Student COF index and band, and the expert verdict, flagging where the model under- or over-states consequence compared to utility judgement.

Pipe ID	Scenario	Student COF (0–5)	Label	Expert verdict	Expert comment
5782851	**High-Cost Urban Main Rupture:** A major pipeline in a densely populated area experiences a large break. The cost of repair surpasses $500k (classification: High). Water loss is also significant (>20,000 gal/hr). Over 100 customers lose supply or experience low pressure.	4.52	Catastrophic	Agree	–
597812	**Spill in Environmentally Sensitive Zone:** A medium-diameter main leaks untreated or partially treated water into a protected wetland. The local habitat is rated as "High Sensitivity" or a wetland (Environmental Impact), meaning even moderate contamination triggers substantial cleanup costs and possible environmental fines.	3.99	Major	Not evaluated	Unknown, not in last AMP
6191645	**Prolonged Service Disruption near Critical Facility:** A pipeline delivering water to a hospital or major industrial site fails, resulting in a multi-day	4.61	Catastrophic	Agree	–

Pipe ID	Scenario	Student COF (0–5)	Label	Expert verdict	Expert comment
	outage. Length of disruption exceeds 24 h (classification: Severe) and affects at least 5,000 critical customers.				
212326	**Deteriorated pipe near arterial roads:** Deteriorated pipe segments under a major highway. Repair is complicated by lane-closure constraints (renewal complexity "High"). Social cost is elevated from traffic disruptions.	2.52	Moderate	Disagree	We gave COF 5 (High consequence).
6190478	**36″ Pipe Break at Peak Demand, Minimal Redundancy:** Occurs during a heat wave or festival when demands peak. Minimal redundancy means no quick reroute. Water losses climb above the "high-cost" threshold (e.g., $300k). Outage extends to 12+ hours.	4.51	Catastrophic	Agree	–
6192211	**Outage impact** (in hours or number of customers affected) of >24″ pipe with redundancies.	3.21	Major	Disagree	We gave COF 5 (High consequence).
6227857	**Traffic impact** in hours of road closure for a pipe failure 20–50 ft away from a road.	4.46	Catastrophic	Agree	–
6132940	**Cost of renewal** for fixing the failure of <16″ pipe (~$20,000).	1.51	Minor	Disagree	Close, we gave COF 3 (Moderate).
473529	**Cost of renewal** for fixing the failure of 16″–36″ pipe (~$70,000).	3.00	Major	Disagree	Close, we gave COF 2 (Low).
6227854	**Cost of renewal** for fixing the failure of >36″ pipe (~$800,000).	4.45	Catastrophic	Agree	–

Table F-5: Expert evaluation of Student COF scores for ten representative high-impact scenarios at Utility B (Southeast). Student COF bands are compared with expert ratings and comments, including disagreements driven by local knowledge of wetlands, outage manageability, and traffic control.

Pipe ID	Scenario	Student COF (0–5)	Label	Expert verdict	Expert comment
413862	**High-Cost Urban Main Rupture:** A major pipeline in a densely populated area experiences a large break. The cost of repair surpasses $500k (classification: High). Water loss is also significant (>20,000 gal/hr). Over 100 customers lose supply or experience low pressure.	4.52	Catastrophic	Agree	–
279637	**Spill in Environmentally Sensitive Zone:** A medium-diameter main leaks untreated or partially treated water into a protected wetland. The local habitat is rated as "High Sensitivity" or a wetland	3.39	Major	Disagree	Wetland not identified by Field Ops.

Pipe ID	Scenario	Student COF (0–5)	Label	Expert verdict	Expert comment
	(Environmental Impact), meaning even moderate contamination triggers substantial cleanup costs and possible environmental fines.				
2427178	**Prolonged Service Disruption near Critical Facility:** A pipeline delivering water to a hospital or major industrial site fails, resulting in a multi-day outage. Length of disruption exceeds 24 h (classification: Severe) and affects at least 5,000 critical customers.	4.46	Catastrophic	Agree	–
2429441	**Deteriorated pipe under heavy traffic load:** Deteriorated pipe segments under a major highway. Repair is complicated by lane-closure constraints (renewal complexity "High"). Social cost is elevated from traffic disruptions.	4.51	Catastrophic	Agree	–
444613	**36″ Pipe Break at Peak Demand, Minimal Redundancy:** Occurs during a heat wave or festival when demands peak. Minimal redundancy means no quick reroute. Water losses climb above the "high-cost" threshold (e.g., $300k). Outage extends to 12+ hours.	4.29	Catastrophic	Agree	–
52686	**Outage impact** (in hours or number of customers affected) of >24″ pipe with redundancies.	3.21	Major	Agree	–
519371	**Traffic impact** in hours of road closure for a pipe failure 20–50 ft away from a road.	1.51	Minor	Agree	–
45369	**Cost of renewal** for fixing the failure of <16″ pipe (~$20,000).	1.34	Minor	Agree	–
54616	**Cost of renewal** for fixing the failure of 16″–36″ pipe (~$70,000).	3.22	Major	Disagree	Short 16″ ductile line that can be isolated; minimal traffic control and limited outage due to valve locations.
3301225	**Cost of renewal** for fixing the failure of >36″ pipe (~$800,000).	4.45	Catastrophic	Agree	–

Table F-6: Expert evaluation of Student COF scores for ten representative high-impact scenarios at Utility C (Coastal West). The table summarizes agreement between modelled and expert consequence bands for urban breaks, sensitive-area spills, hospital outages, and cost-based scenarios.

Pipe ID	Scenario	Student COF (0–5)	Label	Expert verdict	Expert comment
38118-AS121	**High-Cost Urban Main Rupture:** A major pipeline in a densely populated area experiences a large diameter (>36″) pipe break. The cost of repair surpasses $500k (classification: High). Water loss is also significant (>20,000 gal/hr). Over 100 customers lose supply or experience low pressure.	4.58	Catastrophic	agree	–

Pipe ID	Scenario	Student COF (0–5)	Label	Expert verdict	Expert comment
42222S361	**Spill in Environmentally Sensitive Zone:** A 36″ diameter main leaks untreated or partially treated water into a protected wetland. The local habitat is rated as "High Sensitivity" or a wetland (Environmental Impact), meaning even moderate contamination triggers substantial cleanup costs and possible environmental fines.	3.95	Major	agree	–
39239-AS481	**Prolonged Service Disruption near Critical Facility:** A pipeline delivering water to an acute-care hospital, resulting in a multi-day outage. Length of disruption exceeds 24 h (classification: Severe) and affects at least 5,000 critical customers.	4.58	Catastrophic	agree	–
42099-AS61	**Deteriorated pipe under heavy traffic load:** Large diameter (48″) pipe segment under a major highway (I-680). Repair is complicated by lane-closure constraints (renewal complexity "High"). Social cost is elevated from traffic disruptions.	4.61	Catastrophic	agree	I think they meant 47175-B.
47175S841	**84″ pipe break in high-density area with potential for flooding:** High renewal costs including potential litigation costs due to flooding.	4.58	Catastrophic	agree	–
E-16358C62	**Small diameter (6″) pipe fails in high-density area.**	1.89	Minor	agree	–
E-21038-BC21	**Very small diameter (2″) pipe fails in a low-density area.**	0.45	Insignificant	agree	–
E-16358C62	**Cost of renewal** for fixing the failure of <16″ pipe (~$20,000).	1.89	Minor	Not evaluated	–
E-8284S2050	**Cost of renewal** for fixing the failure of 16″–36″ pipe (~$70,000).	2.47	Moderate	agree	–
E-29243-RL481	**Cost of renewal** for fixing the failure of >36″ pipe (~$800,000).	4.45	Catastrophic	agree	–

Table F-7: Expert review of the Student renewal-prioritization portfolio for Utility A (Pacific Northwest). The table shows how the model ranked or excluded specific test segments and records expert agreement or disagreement with those renewal decisions and their rationale.

Pipe ID	Scenario	Model decision	Expert verdict	Expert comment
594561	**Renewal candidate:** Metallic (cast iron) with remaining wall thickness < 80% in a high-density urban area.	Ranked 135/384 among pipe segments selected for renewal.	Disagree	Technically feasible to be selected, but disagree with renewal candidacy for this year due to its relative rank.
526598	**Renewal candidate:** PCCP with >12 wire breaks per 10 ft; steel cylinder rated "deteriorated" and pipe has no redundancy.	Ranked 95/384 among pipe segments selected for renewal.	Disagree	This is a 60″ main and would likely be a contract project as it is a WashCo supply line, so it would not be selected for in-house crew work.

Pipe ID	Scenario	Model decision	Expert verdict	Expert comment
241369	**Not a renewal candidate:** RCP, PCCP or RCCP with steel cylinder condition = minor corrosion; concrete core intact; located in low-demand area.	Not selected for renewal.	Agree	–
5912895	**Not a renewal candidate:** PVC in high-pressure zone and >5% ovality.	Not selected for renewal.	Agree	–
291684	**Not a renewal candidate:** PVC in suburban neighborhood and low-pressure zone with <5% ovality and stable soil conditions.	Not selected for renewal.	Agree	–
6415453	**Renewal candidate:** Externally corroded CI in residential neighborhood with background leakage over 5 days.	Priority 197/384 pipe segments selected for renewal.	Disagree	This is a hydrant main; it would not typically be included in the normal main-renewal selection process.
548299	**Not a renewal candidate:** 6″ AC pipe >50 years old in high-traffic area.	No pipes matching this scenario in the modeled inventory.	Agree	Facility ID does not show up when searching the pressurized-mains layer used for the selection process.
288482	**Not a renewal candidate:** 6″ AC pipe >50 years old in low-traffic area.	No pipes matching this scenario in the modeled inventory.	Agree	GIS indicates this main was abandoned ("killed") in 1993.

Table F-8: Expert review of the Student renewal-prioritization portfolio for Utility B (Southeast). Model rankings for metallic, PCCP, PVC, and AC test segments are compared with expert verdicts, including strong agreement where the utility has experienced repeated failures.

Pipe ID	Scenario	Model decision	Expert verdict	Expert comment
84311	**Renewal candidate:** Metallic (cast iron) with remaining wall thickness < 80% in a high-density urban area.	Ranked 25/154 among pipe segments selected for renewal.	Strongly agree	We have experienced multiple failures on the referenced pipe to date.
2424170	**Renewal candidate:** PCCP with >12 wire breaks per 10 ft; steel cylinder rated "deteriorated" and pipe has no redundancy.	Ranked 1/154 among pipe segments selected for renewal.	Agree	–
110792	**Not a renewal candidate:** RCP, PCCP or RCCP with steel cylinder condition = minor corrosion; concrete core intact; located in low-demand area.	Not selected for renewal.	Agree	–
76247	**Renewal candidate:** PVC in high-pressure zone and >5% ovality.	Ranked 22/154 among pipe segments selected for renewal.	Agree (slight)	There is only one recorded failure known.
4015434	**Not a renewal candidate:** PVC in suburban neighborhood and low-pressure zone with <5% ovality and stable soil conditions.	Not selected for renewal.	Not evaluated	Unable to identify the referenced pipe.

Pipe ID	Scenario	Model decision	Expert verdict	Expert comment
84311	**Renewal candidate:** Externally corroded CI in residential neighborhood with background leakage over 5 days.	Priority 25/154 pipe segments selected for renewal.	Strongly agree	We have experienced multiple failures on the referenced pipe to date.
548299	**Renewal candidate:** 6″ AC pipe >50 years old in high-traffic area.	Priority 153/154 pipe segments selected for renewal.	Agree	–
4016141	**Not a renewal candidate:** 6″ AC pipe <20 years old in low-traffic area.	Not selected for renewal.	Agree	–

Table F-9: Expert review of the Student renewal-prioritization portfolio for Utility C (Coastal West). The table contrasts model selections and ranks with expert opinions, highlighting both accepted priorities and disagreement where local practice (e.g., pre-1955 AC policies, PVC ovality experience) differs from the model's recommendation.

Pipe ID	Scenario	Model decision	Expert verdict	Expert comment
E-27697C41	Metallic (cast iron) with remaining wall thickness < 80% in a high-density urban area.	Ranked 26/717 among pipe segments selected for renewal.	Agree	–
35966-BT2437	>50-year-old PCCP 24″ with one historical break (none in the last 5 years), currently in low-corrosivity conditions and no redundancy.	Not selected for renewal.	Agree	*Not PCCP.*
NA	RCP, PCCP or RCCP with steel cylinder condition = minor corrosion; concrete core intact; located in low-demand area.	Not selected for renewal.	Agree	–
P47360-AN81	PVC in high-pressure zone and >5% ovality.	Ranked 8/717 among pipe segments selected for renewal.	Disagree	*We have not seen PVC ovality issues recently.*
48127-AN81	PVC in suburban neighborhood and low-pressure zone with <5% ovality and stable soil conditions.	Not selected for renewal.	Agree	–
E-25873C62	Externally corroded CI in residential neighborhood with background leakage over 5 days.	Priority 134/717 pipe segments selected for renewal.	Agree	–
33126-IA62	6″ AC pipe >50 years old in high-traffic area.	Priority 19/717 pipe segments selected for renewal.	Agree	–
E-32415A61	6″ AC pipe >60 years old in low-traffic area.	Not selected for renewal.	Disagree	Pre-1955 AC is prioritized based on WRF Study 4480.