Water Resources Research

RESEARCH ARTICLE
10.1029/2024WR039054

Special Collection:
Forcing, response, and impacts of coastal storms in a changing climate

Predicting the Evolution of Extreme Water Levels With Long Short-Term Memory Station-Based Approximated Models and Transfer Learning Techniques

Samuel Daramola[1], David F. Muñoz[1], Paul Muñoz[2], Siddharth Saksena[1], and Jennifer Irish[1]

[1]Department of Civil and Environmental Engineering, Virginia Tech, Blacksburg, VA, USA, [2]Department of Water and Climate, Vrije Universiteit Brussel (VUB), Brussels, VA, USA

Key Points:

- We present a deep learning framework that accurately predicts the evolution of cyclone-induced water levels across multiple domains
- An attention mechanism enhances the framework's recognition of extreme water level patterns within and beyond training locations
- It effectively identifies unseen water level patterns, different from those in training; thus enhancing model's transfer learning capability

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

S. Daramola and D. F. Muñoz,
samueldaramola@vt.edu;
davidmunozpauta@vt.edu

Citation:

Daramola, S., Muñoz, D. F., Muñoz, P., Saksena, S., & Irish, J. (2025). Predicting the evolution of extreme water levels with long short-term memory station-based approximated models and transfer learning techniques. *Water Resources Research*, *61*, e2024WR039054. https://doi.org/10.1029/2024WR039054

Received 27 SEP 2024
Accepted 15 FEB 2025

Author Contributions:

Conceptualization: David F. Muñoz
Formal analysis: Samuel Daramola, David F. Muñoz
Funding acquisition: David F. Muñoz
Investigation: Samuel Daramola, David F. Muñoz

Abstract Extreme water levels (EWLs) resulting from cyclones pose significant flood hazards and risks to coastal communities and interconnected ecosystems. To date, physically based models have enabled accurate prediction of EWLs despite their inherent high computational cost. However, the applicability of these models is limited to data-rich sites with diverse characteristics. The dependence on high quality spatiotemporal data, which is often computationally expensive, hinders the applicability of these models to regions of either limited or data-scarce conditions. To address this challenge, we present a Long Short-Term Memory (LSTM) network framework to predict the evolution of EWLs beyond site-specific training stations. The framework, named LSTM-Station Approximated Models (LSTM-SAM), consists of a collection of bidirectional LSTM models enhanced with a custom attention mechanism layer embedded in the architecture. LSTM-SAM incorporates a transfer learning approach applicable to target (tide-gage) stations along the U.S. Atlantic Coast. Importantly, LSTM-SAM helps analyze: (a) the underlying limitations associated with transfer learning, (b) evaluate EWL predictions beyond training domains, and (c) capture the evolution of EWL caused by tropical and extratropical cyclones. The framework demonstrates satisfactory performance with "transferable" models achieving Kling-Gupta Efficiency (KGE), Nash-Sutcliffe Efficiency (NSE), and Root-Mean Square Error (RMSE) ranging from 0.78 to 0.92, 0.90 to 0.97, and 0.09–0.18 m at the target stations, respectively. We show that LSTM-SAM can accurately predict not only EWLs but also their evolution over time, that is, onset, peak, and dissipation, which could assist in operational flood forecasting in regions with limited resources to set up physically based models.

Plain Language Summary Water levels in rivers, estuaries, and bays rise significantly during hurricanes, leading to severe flood risks and hazards in low-lying areas and interconnected ecosystems. With climate change increasing the frequency of extreme events, it has become crucial to develop models that can accurately simulate extreme water levels in a short time frame and support emergency management for future events. Conventional modeling approaches that help us predict extreme water levels rely on either physically based or data-driven models. Unlike state-of-the-art data-driven models such as deep learning, the former models are site-specific and cannot be applied or transferred to other regions. In this study, we propose a framework that leverages a model trained on extreme water levels from one region to accurately predict those of neighboring regions through a technique known as "transfer learning". We address the limitations associated with this technique, including the inability of transferable models to accurately generalize new input data from those neighboring regions and examine how changes in model parameters influence the development of efficient transferable models. We show that these models can effectively capture the magnitude and timing of extreme water levels, making this framework suitable for early and operational warning systems.

1. Introduction

Tropical and extratropical cyclones are responsible for multiple flood hazards and risks driven by extreme water levels (EWLs) that are exacerbated by climate-related impacts and anthropogenic activities (Hino & Nance, 2021; Khojasteh et al., 2021). Low-lying areas are particularly vulnerable to EWLs as they can lead to severe socio-economic and environmental impacts globally (Rainey et al., 2021; Zscheischler et al., 2020). The United States, accounting for 1.6% of the current global population (129 million people) in low-lying areas (Office for Coastal Management, 2024), has reported more than 391 weather and climate disasters since 1980 (NOAA-NCEI, 2024). In the same period, total reported losses exceeded $2.76 trillion when adjusted for the 2024 Consumer Price Index

Methodology: Samuel Daramola, David F. Muñoz, Paul Muñoz
Project administration: David F. Muñoz
Resources: David F. Muñoz
Software: Samuel Daramola
Supervision: David F. Muñoz, Jennifer Irish
Validation: Samuel Daramola
Visualization: Samuel Daramola, David F. Muñoz
Writing – original draft: Samuel Daramola, David F. Muñoz
Writing – review & editing: Paul Muñoz, Siddharth Saksena, Jennifer Irish

(NOAA-NCEI, 2024). Among these disasters, six of the world's costliest hurricanes resulted in over $50 billion in damages in the United States (Sanders et al., 2023).

1.1. Extreme Water Levels

Extreme water levels in estuarine and coastal systems arise from various flood drivers, including precipitation, river discharge, storm surge, tides, and waves. Yet, these drivers do not necessarily act in isolation but rather synergize resulting in compound flooding (Muis et al., 2019; Parker et al., 2023; Wahl et al., 2017). Extreme water levels become more intense when storm surge co-occur with heavy precipitation during hurricanes (Bevacqua et al., 2019; Wahl et al., 2015), high tide coincide with the peak of a storm surge (Marsooli & Wang, 2020; Thomas et al., 2019), peak river flow and storm surge co-occur along estuarine systems (Moftakhari et al., 2019; Muñoz et al., 2020), and waves and storm surge interact nonlinearly. Due to the increasing frequency and intensity of extratropical (ETCs) and tropical cyclones (TCs) along with the rise of sea levels and ocean temperatures, changes in storminess are expected to play a key role in future EWLs (Anderson et al., 2021; Bloemendaal et al., 2022; Ghanbari et al., 2021). Recognizing the heightened flood risks to coastal communities, it has become imperative for researchers and practitioners to rely on either physically based or data-driven modeling approaches to predict EWLs in terms of peak magnitude and timing.

1.2. Extreme Water Level Prediction

Physically based models are commonly used to predict EWLs and their evolution based on simplified hydrometeorological processes governed by the conservation of mass and momentum (Bates, 2023a; Santiago-Collazo et al., 2019). The accuracy of these models depends on the availability and quality of several spatiotemporal data sets to appropriately characterize input and forcing conditions, topography and bathymetry, land surface roughness, and other key morphologic characteristics (Alipour et al., 2022; Bates, 2023b; Jafarzadegan et al., 2021). Yet, physically based models developed with a fine spatial resolution (e.g., grid-cell size in the order of meters) are often constrained by limited spatial scope and/or high computational demands necessary to solve large-scale flood dynamics (Bilskie et al., 2021; Muñoz et al., 2021). On the other hand, models developed with a coarse spatial resolution can cover broader areas and reduce computational time, but they may lead to less accurate predictions due to a lack of detailed spatiotemporal information around key morphological and hydrodynamic variables in narrow tidal inlets and river channels (Fraehr et al., 2022; Saksena & Merwade, 2015). Importantly, physically based models cannot be transferred or applied to other domains due to site-specific conditions including topographic and bathymetric characteristics (Bates, 2022; Santiago-Collazo et al., 2019). With the increasing need to expedite decision-making processes and circumvent time constraints, data-driven models have become a feasible alternative that can effectively be transferred to other domains under certain conditions.

State-of-the-art data-driven models offer rapid and efficient prediction and forecasting solutions at large scales and have the ability to generalize or identify patterns from the data they are trained on (Hamitouche & Molina, 2022; Lee et al., 2021a). Specifically, neural network (NN) models can learn nonlinear relationships and hidden patterns from sequential time-series data; thus, enhancing the prediction accuracy in hydrological and coastal applications (Li et al., 2021; Zhang et al., 2022). In addition, those models can be updated over time which improves their predictions as more and new information becomes available. Like physically based models, data-driven models can be developed using geographical, morphological, and hydrodynamic features, including forcing drivers from a specific domain. Once trained, these models can generalize learned patterns to neighboring regions using transfer learning (TL) techniques (Shen, 2018).

TL is particularly valuable for addressing the challenge of insufficient training data by leveraging knowledge gained from data-rich training domains (Tan et al., 2018; Zhuang et al., 2021). While TL models are especially advantageous for areas lacking data, it is crucial to first validate their reliability, such as their ability to accurately predict extreme values, in locations with existing data. This validation step allows for refining model architecture, feature selection, and hyperparameter optimization to ensure robust predictions. Once the model's reliability is established, it can be systematically applied to locations without ground truth data by leveraging domain similarities, regional features, and the potential integration of physics-based constraints.

1.3. Transfer Learning

Several studies have implemented TL techniques in NN models to estimate urban flood levels (Seleem et al., 2023; Zhao et al., 2021), predict significant wave height (Obara & Nakamura, 2022), and compound flood hazard characterization of nearby regions to the training domain (Muñoz et al., 2021). Particularly, long short-term memory (LSTM) networks, a variant of recurrent neural network (RNN), have been employed for flood susceptibility assessments, barrage integrity, riverine flood level forecast, and surge prediction (Fang et al., 2021; Kardhana et al., 2022; Liu et al., 2023; Merizalde et al., 2023; Tiggeloven et al., 2021). Other studies have implemented bidirectional LSTM networks (Bi-LSTM) as they demonstrate more satisfactory performances for sequential data. Contrary to conventional LSTM, Bi-LSTM network models process input features in both forward and backward directions (Siami-Namini et al., 2019; Zrira et al., 2024). This in turn enables the models to capture past and future contexts within a sequence, which is advantageous for water level prediction (Bai & Xu, 2021; Fang et al., 2021; Zhang et al., 2022).

However, creating a NN model with effective generalization capabilities beyond its training domain remains a significant hurdle (Bates, 2022; Bentivoglio et al., 2022). Since maintaining consistency in location enhances the accuracy and lead-time of model predictions (Altunkaynak & Kartal, 2021), the geographical areas suitable for effectively applying TL are limited. Also, good generalization relies on the training and test data being similarly distributed and strictly representative of one another to prevent data set shift (Moreno-Torres et al., 2012). In this regard, NN models may not necessarily exhibit the same level of effectiveness in target domains, even if both training and target domains share similar flood thresholds and EWL dynamics. Previous studies have largely attributed the reduced model accuracy to the risk of negative transfer, where information from the training domain may not be entirely applicable to the target areas (Chen et al., 2021; Xu et al., 2023); overfitting during the training phase which prevents the model from capturing generalizable patterns and thereby introducing noise during the transfer process (Peng et al., 2022); as well as, the sensitivity of hyperparameters to the data set, where models with the optimal combinations that yield satisfactory performance in the training domain do not achieve similar result in the target domain (Chen et al., 2021; Dong et al., 2021; Ma et al., 2020).

In this study, we introduce a comprehensive framework that effectively leverages learned data patterns from tide-gauge stations along the U.S. Atlantic Coast and applies them to neighboring tide-gauge stations. Mahmoudi et al. (2024) previously found through cluster analysis that the U.S. Atlantic Coast shares similar hydrodynamic characteristics in terms of flood thresholds and EWL dynamics. Nevertheless, individual tide-gauge stations may have distinct topographic and bathymetric characteristics, which could significantly influence TL applications for storm surge prediction, particularly when predicting EWL in space. Therefore, this study emphasizes the importance of techniques that enhance TL capabilities, focusing on improving the model's ability to recognize EWL patterns beyond the events or conditions observed during training.

Hence, this study addresses the following questions: (a) Is there an efficient transfer learning technique to ensure that the optimal set of hyperparameters for the training domain will also achieve satisfactory model's performance in target domains? (b) Can models be trained to maintain their pattern-capturing abilities beyond the training phase, effectively accounting for "unseen" EWL variability and evolution? (c) How do various combinations of hyperparameters affect the model's pattern recognition capabilities in target domains? Here, we analyze the underlying limitations associated with TL techniques, evaluate EWL predictions beyond training domains, and present accurately captured evolution of EWL dynamics caused by ETCs and TCs. To our knowledge, this study is among the first applications of TL to EWL prediction in coastal domains and beyond hydrological applications.

2. Materials and Data

2.1. Study Area

The proposed LSTM-SAM framework is trained using time-series data from 5 strategically selected tide-gage (training) stations located along the U.S. Atlantic Coast. These stations are Atlantic City, NJ (NOAA ID: 8534720), Sewells Point, VA (8638610), Beaufort, NC (8656483), Fort Pulaski, GA (8670870) and Trident Pier, FL (8721604) (Figure 1). The training stations are selected based on two criteria: (a) exposure to multiple hurricane events and (b) at least 70% consecutive hourly water level (WL) data recorded over 40 years. The latter ensures that the training stations contain EWLs attributed to either TCs (hurricanes) or ETCs (Nor'easter winter storms) to effectively train and validate the LSTM-SAM framework. We then implement a TL approach in the

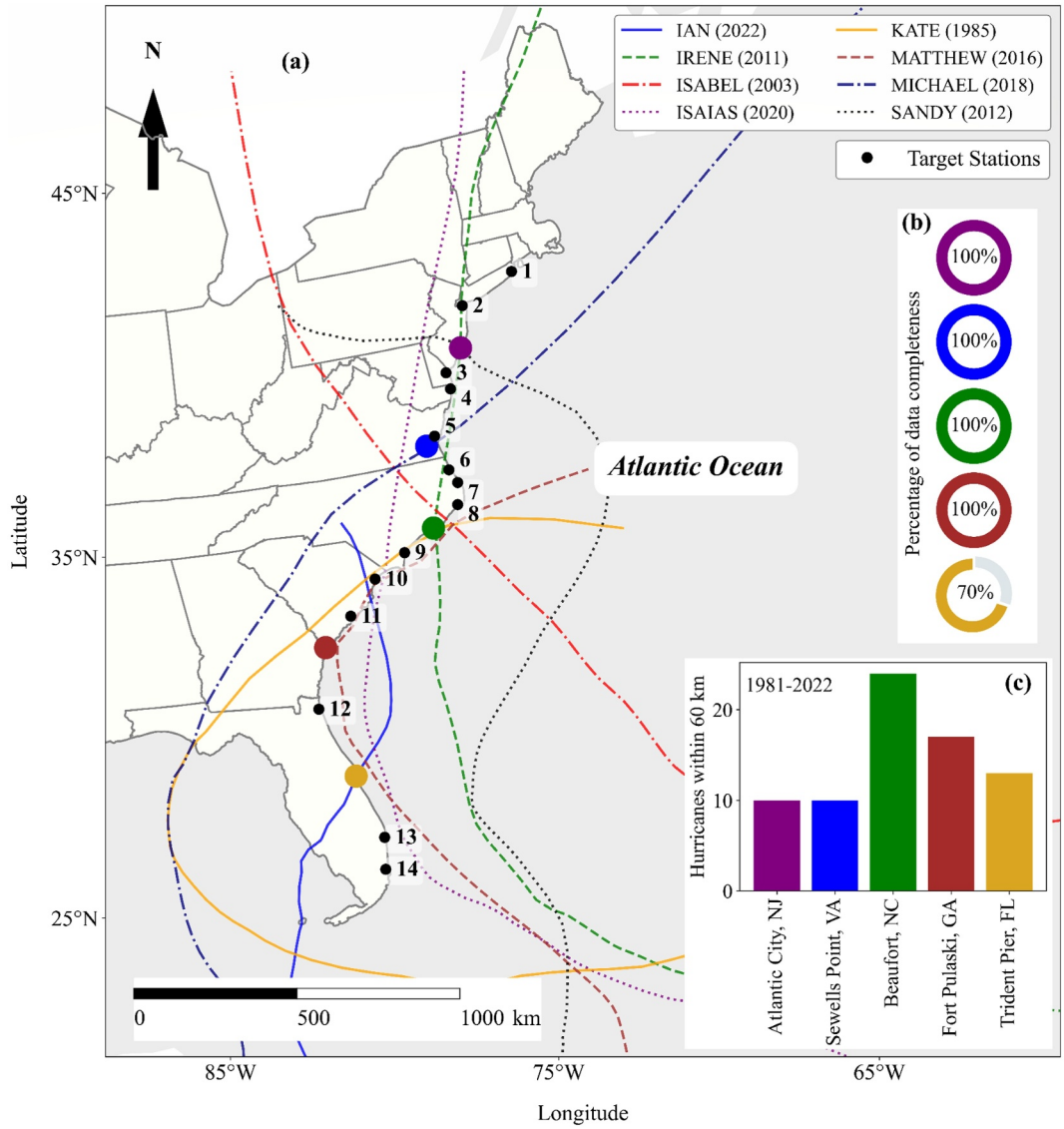

Figure 1. Location of training and target stations along the U.S. Atlantic Coast. (a) Selected training and target stations (numbered from 1 to 14) are shown with colored and black circles, respectively. (b) For each training station, percentage of water level data completeness obtained from the NOAA's Tides and Current portal. (c) Relevant hurricane's best tracks within a 60 km radius of the hurricane's landfall locations.

framework and transfer nonlinear patterns from training to target stations in order to predict the evolution of EWLs. Most of the target stations are directly exposed to the Atlantic Ocean and located in-between the training stations (Figure 1). Those stations include: (a) Montauk, NY (NOAA ID: 8510560), (b) Sandy Hook, NJ (8656483), (c) Lewes, DE (8557380), (d) Ocean City, MD (8570283), (e) Kiptopeke, VA (8632200), (f) Duck, NC (8656483), (g) Oregon Inlet Marina, NC (8652587), (h) USCG Station Hatteras, NC (8654467), (i) Wrightsville Beach, NC (8658163), (j) Springmaid Pier, SC (8661070), (k) Charleston, SC (8665530), (l) Mayport, FL (8720218), (m) Lake Worth Pier, FL (8722670), and (n) Virginia Key, FL (8723214).

2.2. Data Availability

We retrieve WL data from the National Oceanic and Atmospheric Administration (NOAA)'s Tides and Currents portal (https://tidesandcurrents.noaa.gov/map/index.html). Meteorological and wave data are obtained from the European Centre for Medium-Range Weather Forecasts Reanalysis data set (ERA, version 5) produced by the Copernicus Climate Change Service (https://cds.climate.copernicus.eu/). ERA5 data set has a spatial resolution of

31 km that allows for an accurate representation of extreme climate events at large scale, including those driven by either ETCs or TCs (Bian et al., 2021). Specifically, we use hourly wind speed and direction at 10 m elevation, sea level pressure, sea surface temperature, air temperature, precipitation, wave direction, and wave height. These aforementioned data sets have been successfully applied to other NN models that predict hourly non-tidal residuals at tide stations globally with satisfactory results (Bruneau et al., 2020).

2.3. Data Processing

The required data length to effectively train NN models depends on the response time of the system under analysis. For coastal systems, previous studies recommend at least six years of training data consisting of complete consecutive sequences (10 days) in order to achieve consistent proficiency in NN models (Bruneau et al., 2020; Tiggeloven et al., 2021). Following this, we conduct data quality control over the training stations and ensure that the time-series contain complete data sequences to train the Bi-LSTM models. Then, we decompose the time-series data of WL into seasonality, trend, predicted tides, and non-tidal residual (NTR) components using the Seasonal-Trend decomposition using LOESS (STL) and Unified Tidal Analysis and Prediction (UTide) packages in Python (Cleveland et al., 1990; Codiga, 2011). The STL analysis, adept at time-series analysis for its outlier resilience, flexible seasonal adjustment, and trend adaptability, provides comprehensive insights into long-term and seasonal dynamics (Chen et al., 2020). UTide employs a decision tree algorithm, a recognized method for automatically selecting the most relevant constituents from 147 tidal constituents and offers tide prediction correction for records spanning up to one full (18.6-year) nodal cycle (Codiga, 2011; Tedesco et al., 2023; Tiggeloven et al., 2021).

We consider a window size of 40 days and a time step of 3 days for time-series decomposition in order to ensure that at least one full lunar cycle is covered (Figure S1 in Supporting Information S1), including both spring and neap tides and the independence of large storm events by selecting the maximum NTR on a stepped basis (Moftakhari et al., 2024; Rashid et al., 2024; Serafin & Ruggiero, 2014). Shorter moving windows lead to a mixing of information of tidal frequencies whereas longer ones average out flow variability (Jay & Flinchem, 1999; Moftakhari et al., 2013). The time-series decomposition aids to improve deep learning by distinguishing clear, recurring patterns from irregular variations; thereby refining the models' ability to learn from the data and enhancing the accuracy of their predictions (Parker et al., 2023).

In addition to the WL components, we extract meteorological and wave data from the closest grid pixel of ERA-5 data set to tide-gage stations. For this, we calculate the minimum square difference between the latitudes and longitudes of the data points and the specified location, that is, within a radius of 15.5 km. Next, we use the time-series of WL components, meteorological, and wave data as relevant input features to the LSTM-SAM framework in order to predict the target variable (e.g., EWLs and their evolution over time). Both input and target variables are first scaled using the "MinMaxScaler" function from the sklearn library in Python. In this case, MinMaxScaler was selected because it effectively maintains the original data distribution while normalizing it to a scale suitable for neural networks. This function normalizes the range of multisource data and ensures that all features have an identical scale, typically between 0 and 1 (de Amorim et al., 2023) before model training.

3. Methods

The LSTM-SAM framework encompasses a comprehensive approach of data preprocessing, model building, and evaluation for predicting EWL using Bi-LSTM networks (see Equations 1–13 in Supporting Information S1) with a custom attention layer. The pseudocodes describing the steps involved in this approach are provided in Tables S1 and S2 in Supporting Information S1 and discussed in the following subsections.

3.1. Training Approaches and RNN Data Set Construction

The input data (Figure 2a) is processed using two different approaches during the model training: (a) train-test (TT) split, and (b) time-series cross-validation (CV) split, as illustrated in Figure 2b. For the TT split, the data is divided sequentially into an 80% training set (with a fraction at the end serving as the validation set) and a 20% testing set. In contrast, the CV involves iteratively shifting the validation set across the sequential training data (still 80% of the input data), allowing for a more comprehensive evaluation of the model's performance across different time periods. 30% of the training set is the fraction used for validation of the model's learning ability during the model development. Unlike traditional CV strategies, here CV fold does not shuffle the data and

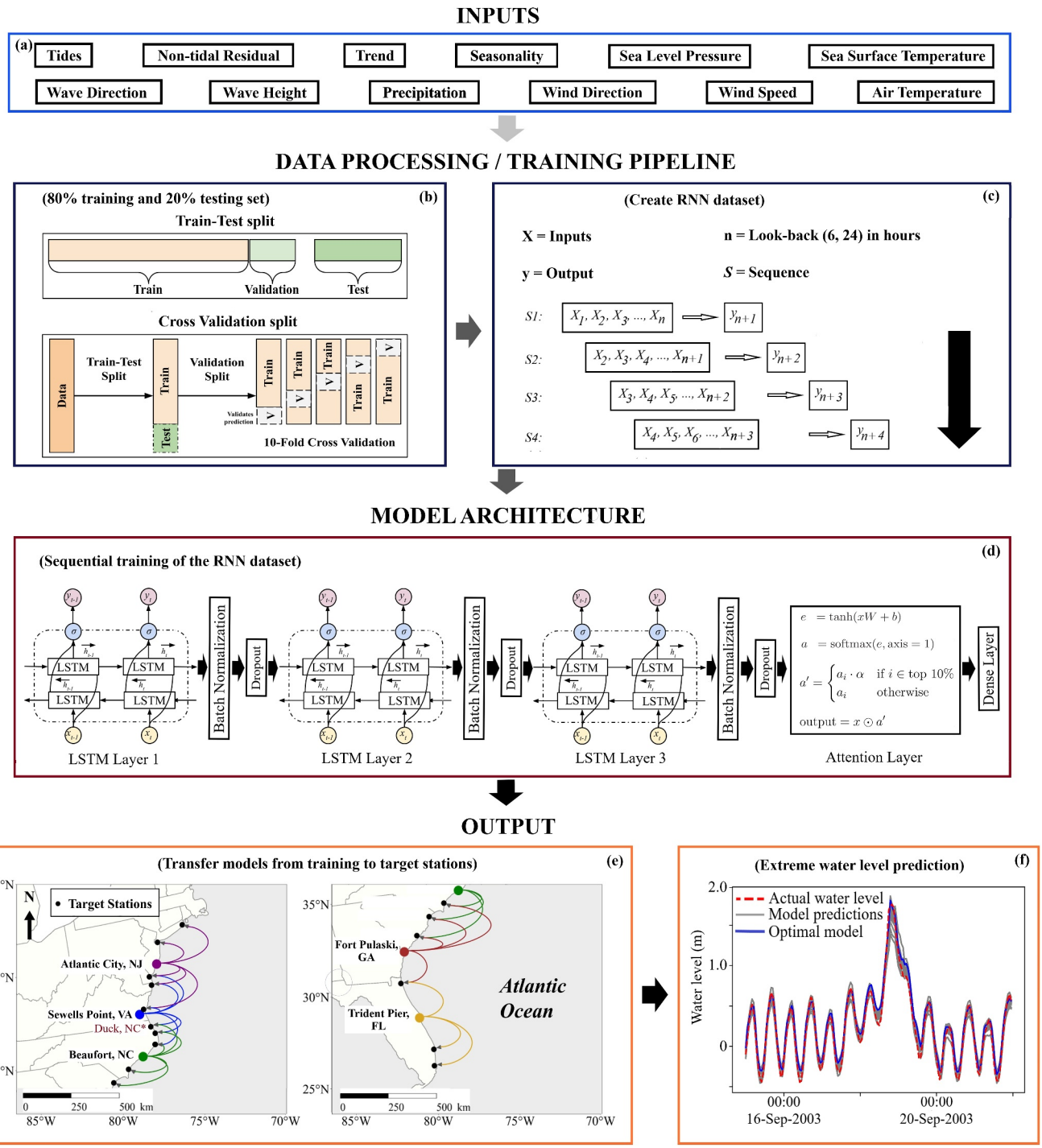

Figure 2. Schematic of the proposed framework, illustrating: (a) model inputs, comprising various hydrometeorological features, (b) data preparation steps, including train-test splitting and time-series cross-validation, (c) the training pipeline for processing inputs and predicting water levels at the training stations, (d) the architecture of the Bidirectional LSTM model with an attention mechanism to enhance pattern recognition, (e) the transfer learning approach for predicting extreme water level evolution at target stations (black circles) using models developed at nearby training stations (colored circles) and (f) extreme water level prediction by transferable models.

therefore keeps the time sequence invariant (Kingphai & Moshfeghi, 2023). We consider 10-fold CV to check the model's performance and potentially improve the prediction accuracy. Moreover, the Bi-LSTM model is trained using the training set for that specific split for each iteration of the loop. As the loop progresses, the size of the

training set increases whereas the validation set consists of data points that come after the training set in time. As a result, the training and validation process involves learning from past data and validating the model's performance on unseen future data, respectively.

A function is designed to preprocess data, creating a RNN data set for effectively training the Bi-LSTM model. It constructs sequences of input features (X) and their corresponding target values (y) based on a specified lookback period (Figure 2b). For each time step, the function extracts a sequence of lookback consecutive time steps from the normalized input features (X_norm) and pairs it with the target value (y_scaled) at the subsequent time step. This results in a data set where each input sequence has a shape of [lookback, number of features], suitable for capturing temporal dependencies in sequence prediction tasks. Moreover, we consider two look-backs of 6 and 24 hr to train the Bi-LSTM, allowing us to evaluate the effects of different temporal resolution on the model's prediction performance.

3.2. Model Architecture

The proposed model architecture consists of three Bi-LSTM layers, with the number of units in each layer set to vary between 32 and 512 in intervals of 32 (Figure 2d). We consider a "L2" regularization method to prevent overfitting by penalizing large weights. We include batch normalization to reduce covariate shift and improve the training by normalizing the inputs of each layer. Also, we include dropout rates in the model architecture to prevent overfitting by randomly disabling a subset of neurons during the training process, thereby allowing the Bi-LSTM network to develop a more generalized understanding of the data and improve its performance on new and/or unseen data. The Bi-LSTM units use the "tanh" and hard "sigmoid" recurrent activation functions. In addition, we add a standard dense (fully connected) layer to output the final prediction. The loss functions consist of both mean absolute error (MAE) and mean squared error (MSE) for different variants of the models with the "Adam" optimizer as suggested in similar WL prediction studies (Huang et al., 2020). An early stopping callback is also employed to monitor the validation loss, stop training if a higher validation loss value is observed after five consecutive epochs, and ultimately prevent overfitting and/or unnecessary computations.

3.3. Attention Mechanism Layer

An attention mechanism layer is added to the model architecture (Figure 2d). It is used to address inherent limitations of conventional RNNs (including Bi-LSTMs), such as losing information from earlier parts of long sequences and difficulties in training models with data of sharp and extreme changes (Rithani et al., 2023). These limitation are due to RNN model development typically based on identifying and exploiting repetitive patterns and correlations within the training data (Bandara et al., 2020; Makridakis et al., 2020). The attention mechanism, therefore, scans through the data, identifies key features, nuanced information and low frequency water levels, increasing their influence in the training process. For more details on attention mechanism, the reader is referred to the study of Niu et al. (2021). Here, we incorporate an attention mechanism layer for a better model generalization during and beyond the training process. Specifically, the layer computes attention scores for each time step using a weight matrix (Glorot) and bias (zeros) initialization (Equations 14 to 15 in Supporting Information S1). In addition, we customize this layer using a factor that amplifies the top 10% of the attention scores (Equation 16 in Supporting Information S1). This factor allows the model to focus more on crucial parts of the sequence, which could be abrupt changes of high or low levels in the data. The choice of the Glorot initializer for the weight matrix in the attention layer is appropriate due to the use of tanh and sigmoid activation functions in the Bi-LSTM units (Glorot & Bengio, 2010). The initializer keeps the scale of the gradients approximately the same in all layers of the Bi-LSTM network. Starting with zero biases ensures that all neurons in a layer initially produce outputs of roughly the same magnitude, which can be a good starting point for symmetric activation functions like tanh.

3.4. Hyperparameter Tuning

We conduct hyperparameter tuning to identify optimal values of Bi-LSTM units, dropout rate, and learning rate within specified ranges to train the models (Table 1). The model architecture relies on a Bayesian optimization technique for hyperparameter tuning that inherently functions in a sequential manner and leverages data from previous evaluations to inform subsequent runs (Wang et al., 2023). Such technique efficiently balances the exploration of new areas in the hyperparameter space with an emphasis on known suitable regions. This is

Table 1
Range of Values Considered for Hyperparameter Tuning

Hyperparameter (3 Bi-LSTM units)	Range of tested values
Bi-LSTM Units	32–512 (step of 32)
Dropout Rate	0.10–0.50 (step of 0.1)
Activation	Tanh, Sigmoid
L2 Regularization	1e−6-1e−3 (log sampling)
Learning Rate (Adam Optimizer)	1e−4-1e−2 (log sampling)
Maximum number of trials	300
Batch Size (b)	32–256 (step of 32)
Loss Function	MAE, MSE
Look-back time (h)	6, 24
Epochs	500 (with early stopping)
Validation Split	30% of training data

particularly useful when each training iteration is computationally intensive since the optimization technique can identify optimal hyperparameters with less time than methods like grid or random search (Marco et al., 2022). Additionally, its capacity to handle high-dimensional hyperparameter spaces and integrate prior knowledge about potential hyperparameters makes it a versatile choice (Bischl et al., 2023). Its proven success in real-world applications and its efficiency in finding robust hyperparameters with limited evaluations position it as a top choice for many practitioners (Wang et al., 2023).

Hyperparameters, identified through a rigorous tuning (or calibration) process on site-specific training data, tend to yield models that perform optimally within their training domains. However, such optimal models may lack the capability to discern patterns at target stations due to differences in hyperparameter values effect on individual data set characteristics (Dong et al., 2021). For example, a smaller batch size might lead to more frequent updates and potentially more nuanced learning, while a larger batch size could provide more learning stability, ignoring finer negligible patterns. Similarly, data sequences of the same resolution within a 6-hr period would provide different temporal patterns and trends compared to a 24-hr period. Each option could be beneficial in one domain and less so in another. To address this limitation, we define fixed values for specific hyperparameters such as batch sizes (32, 64, 128, and 256), look-back times (6 and 24), loss functions (MAE and MSE), and data training strategies (TT and CV) based on other studies (Bandara et al., 2020; Hewamalage et al., 2021), while allowing other hyperparameters, such as the optimal dropout and learning rates, to be determined through tuning. We then ensure that the combinations of options from the hyperparameters occur precisely once, which in turn facilitates the creation of 32 distinctive models with a unique set of hyperparameters at each of the five training stations. However, excessive tuning trials poses a risk of overfitting, where models become overly tailored to the training data and lose their predictive ability on new data sets. Therefore, we limit the tuner search to a maximum of 300 trials, after which the best hyperparameters are used to train the models, generating a spectrum of suitable models for each training domain based on the discussed combination strategy (Table 1).

3.5. Model Prediction and Transfer Learning

The evolution of EWLs in the test data sets is analyzed by focusing on historic hurricane events and Nor'easter winter storms within a 7-day window centered around the peak WL (Figure 2e). Sensitivity analysis reveals that this time window effectively accounts for the onset and dissipation of EWLs, whereas shorter windows fail to capture the full evolution of EWLs across all stations, as they primarily emphasize the peak WL. This approach is further validated by the Bloemendaal et al.'s (2019) study on tropical cyclone storm surge modeling. We then conduct TL to predict the evolution of EWLs at selected target stations by leveraging "gained knowledge" from the closest training stations (Figure 2f). Such knowledge includes hidden sequential patterns and nonlinear associations among input and target data features that are stored as model weights (Muñoz et al., 2021; Zhao et al., 2021). Note that most of the target stations are located in-between two training stations; except Montauk, NY and Sandy Hook, NJ as well as Lake Worth Pier, FL and Virginia Key, FL that are close to a single training station.

Here, the TL approach consists in leveraging all available Bi-LSTM models (e.g., 32 models) at the target stations in order to predict the evolution of EWLs within the predefined 7-day window. Among these models, we identify "transferable" Bi-LSTM models based on the criteria that both KGE and NSE are above a threshold value of 0.70 at the target stations. This threshold ensures that each transferable model adequately accounts for the magnitude and timing of EWLs while also keeping its inherent pattern recognition capabilities on new input data that has not yet been observed at the target stations (e.g., those associated with future extreme events). Lastly, we evaluate model's performance using several metrics that are recommended for models predicting WL dynamics (Abbaszadeh et al., 2020; Lee et al., 2021b; Muñoz, Abbaszadeh, et al., 2022; Nearing et al., 2024). Those include the coefficient of determination (R^2), Mean Bias Error (MBE), Root Mean Square Error (RMSE), Kling-Gupta Efficiency (KGE) (Gupta et al., 2009), and Nash-Sutcliffe Efficiency (NSE) (Nash & Sutcliffe, 1970). Based on the evaluation metrics, there might be a set of "transferable" Bi-LSTM models among the suitable ones from the

training station for which inherent pattern recognition capabilities would be adequate for the target stations (Section 3.4).

4. Results

4.1. Assessment of the Bidirectional LSTM Models

We first assess the performance of Bi-LSTM models at the training stations with and without the attention mechanism layer incorporated in the model architecture. For this, we consider the models' ability to capture: (a) peak and timing of EWLs and (b) evolution of historic TCs (hurricanes) and ETCs (Nor'easter winter storms).

4.1.1. Extreme Water Level Prediction

Based on the sequential split, we observe that the testing set at the training station at Fort Pulaski (FP), GA, contains the highest extreme values. This station has been reporting complete consecutive hourly data from 1975 to the present (Figures 3a and 3b), providing a solid basis for evaluating the performance between models trained with and without the attention mechanism. Predictions of EWLs and associated RMSE, average peak error, and KGE metrics suggest that Bi-LSTM models with the attention layer (hereafter referred to as Bi-LSTM-ATT) can capture the magnitude and timing of peak WLs with a higher accuracy than those without this layer (Figure 3b). Note that this includes the two most EWL events contained in the testing set (e.g., Hurricane Matthew (2016) and Hurricane Irma (2017)). More than half of Bi-LSTM models achieve a median RMSE, absolute peak error, and mean bias of 0.18 m, 0.68 m, and −0.11 m, respectively (Figures 3c–3f and Table S3 in Supporting Information S1). Also, these models achieve very low to moderate performances with a median KGE and NSE of 0.50 and 0.15, respectively (Table S3 in Supporting Information S1). In contrast, the models' performance substantially improves after integrating the attention mechanism layer in the model architecture. In that regard, half of Bi-LSTM-ATT models show a reduction in the median RMSE, absolute peak error, and mean bias by 27% (0.13 m), 36% (0.44 m), and 55% (−0.05 m) with respect to the Bi-LSTM models only (Figures 3g–3j and Table 2). Also, these models achieve a median KGE and NSE of 0.67 and 0.56, respectively (Table 2).

For convenience, we report only the EWL predictions from the Bi-LSTM-ATT models (Table 2) for the remaining training stations. These models outperform the Bi-LSTM models at all stations (Figures 3c–3g), except at Sewells Point (SW), VA, where performance is comparable. The models at Atlantic City (AC), NJ station achieve satisfactory model performance with a median KGE of 0.88 despite the relatively low median NSE of 0.60. The median RMSE and mean bias in this training station are 0.06 and 0.01 m, respectively. Training station Sewells Point (SW), VA has the best performing models with a relatively high median KGE and NSE of 0.92 and 0.80, respectively. These models achieve a median RMSE and mean bias of 0.05 and −0.02 m, respectively. Similarly, results at Beaufort (BF), NC station perform satisfactorily with median KGE and NSE of 0.86 and 0.69, respectively. Also, this training station shows a median RMSE and mean bias of 0.05 and −0.01 m, respectively. Lastly, the models of training station Trident Pier (TD), FL achieve a moderate to satisfactory performance with median KGE and NSE of 0.78 and 0.51, respectively. The models show a median RMSE and mean bias of −0.07 and −0.03 m, respectively.

Regarding the peak time difference (Figure 3h) of the most extreme event per station, all Bi-LSTM-ATT models developed for AC and BF stations show a 1-hr lead difference and a perfect match with respect to the observed peak, respectively. Six models of SW station show a 1-hr lag difference whereas two models of TP station show a 1-hr lead difference with respect to the observed peak. Thirteen and 10 models developed for FP stations show a 1-hr lag difference and 1-hr lead difference with respect to the observed peak, respectively. Overall, the time difference between observed and predicted peak WL is ±1 hr for all trained models.

4.2. Impact of Attention Mechanism on Transfer Learning Performance

After evaluating the models' performance, we proceed with the assessment of transferable models from the closest training stations to target stations. For example, we transfer the trained Bi-LSTM and Bi-LSTM-ATT models and their associated model weights from Sewells Point, VA to the target station at Duck, NC (Figure 2e). Since most of the extreme events at Sewells Point, VA, including Hurricane Isabel (2003), Irene (2011) and Sandy (2012), are in the training set, the performance of all Bi-LSTM models are satisfactory, with metrics comparable to Bi-LSTM-ATT models (Table S3 in Supporting Information S1). Here, the goal is to predict the evolution of

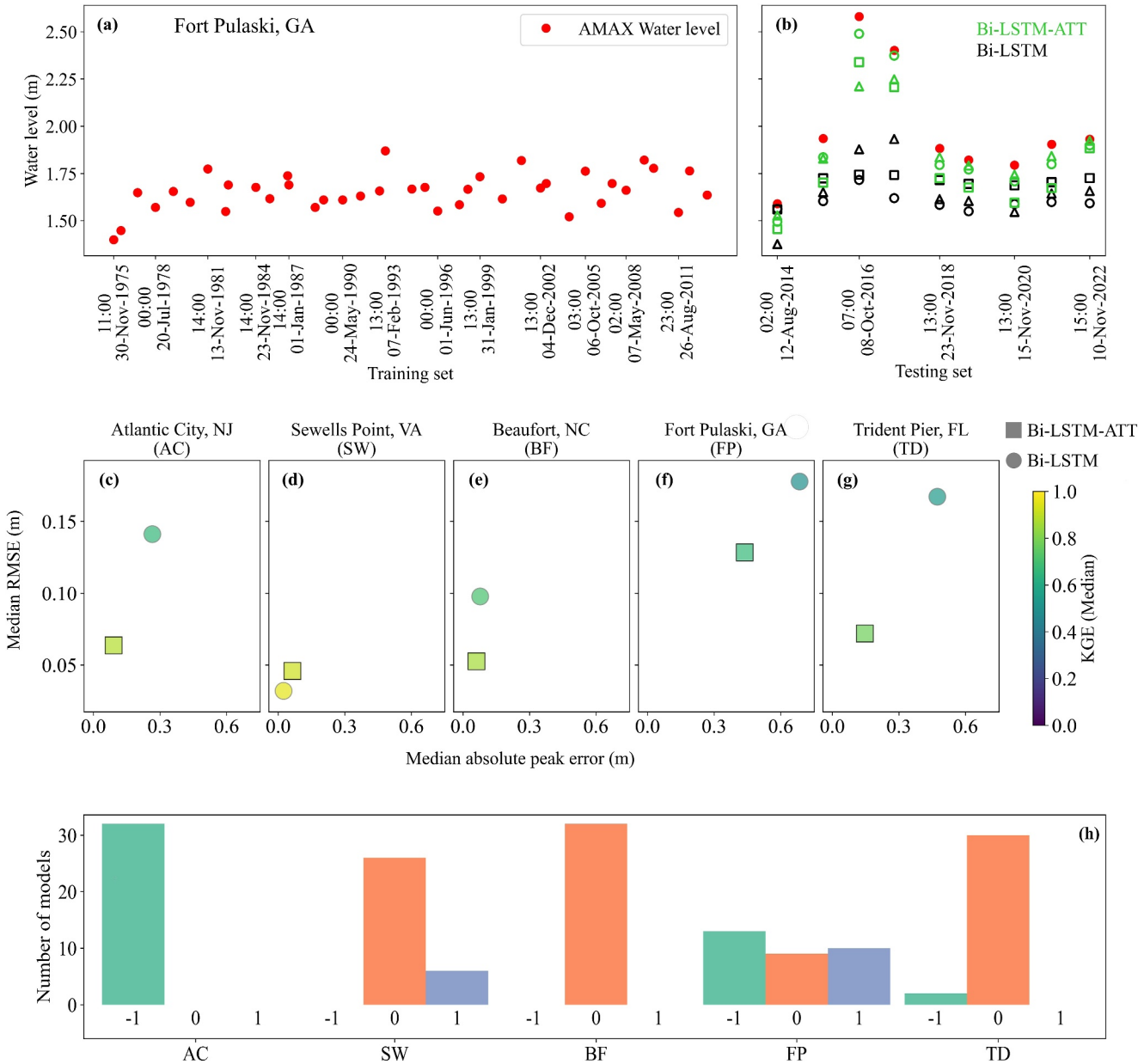

Figure 3. Assessment of the models' performance in the testing set at Fort Pulaski (FP), GA. (a), (b) Annual maximum water levels in the training set (80%), and testing set (20%) in addition to predictions with and without the attention mechanism layer. The three black and three green markers represent the top three best-performing models developed using Bi-LSTM and Bi-LSTM-ATT architectures, respectively. Comparison of overall performance for Bi-LSTM and Bi-LSTM-ATT models at (c) Atlantic City, NJ, (d) Sewells Point, VA, (e) Beaufort, NC, (f) Fort Pulaski, GA and (g) Trident Pier, FL training stations. (h) Observed and predicted peak time differences among the 32 Bi-LSTM-ATT models for each of the five training stations.

EWLs for relevant extreme events such as Hurricane Isabel (2003) and Dorian (2019) (Figure 4). Based on the threshold value of 0.70 (Section 3.4), there are no transferable Bi-LSTM models that can predict the evolution of both storm events. For Hurricane Isabel, the two best models achieve low KGE of −0.98 and −1.03 and NSE of −0.50 and −0.54 (Figure 4a). Although these two models show a better performance for Hurricane Dorian, they still achieve moderate KGE of 0.55 and 0.54 and NSE of 0.43 and 0.41 (Figure 4d). In contrast, the top-two transferable Bi-LSTM-ATT models achieve high KGE of 0.94 and 0.93 and NSE of 0.97 and 0.97 when predicting EWLs triggered by Hurricane Isabel (Figure 4b). For Hurricane Dorian, the top-two transferable Bi-LSTM-ATT models achieve a relatively high KGE of 0.72 and 0.74 as well as NSE of 0.93 and 0.93 (Figure 4e).

Table 2
Transferable Models From Training to Neighboring Target Stations in the U.S. Atlantic Coast.

Target stations	Number of transferable models	Average KGE	Average NSE	Average absolute peak error	Average RMSE	Average median bias	Median KGE	Median NSE	Median absolute peak error	Median RMSE	Median mean bias
Montauk, NY	3	0.81	0.93	0.11	0.11	0.06	0.75	0.92	0.11	0.12	0.09
Sandy Hook, NJ	28	0.88	0.97	0.22	0.13	0.03	0.91	0.97	0.18	0.13	0.04
Lewes, DE	1	0.78	0.97	0.28	0.11	−0.04	0.78	0.97	0.28	0.11	−0.04
Ocean City, MD	18	0.92	0.95	0.08	0.09	0	0.92	0.95	0.07	0.09	0
Kiptopeke, VA	21	0.88	0.96	0.13	0.10	0.01	0.89	0.96	0.12	0.10	0.02
Duck, NC	14	0.84	0.94	0.07	0.10	−0.03	0.85	0.94	0.04	0.09	−0.03
Oregon Inlet, NC	1[a]	0.32	0.60	0.14	0.21	0.15	0.32	0.60	0.14	0.21	0.15
USCG Station Hatteras, NC	1[a]	0.63	0.48	0.28	0.21	0.15	0.63	0.48	0.28	0.21	0.15
Wrightsville Beach, NC	20	0.89	0.94	0.30	0.13	0.02	0.89	0.94	0.19	0.13	0.03
Springmaid Pier, SC	39	0.83	0.96	0.18	0.17	−0.03	0.84	0.96	0.17	0.17	−0.03
Charleston, SC	2	0.84	0.93	0.37	0.18	−0.03	0.85	0.96	0.36	0.16	−0.03
Mayport, FL	4	0.87	0.92	0.20	0.14	−0.04	0.88	0.92	0.21	0.14	−0.03
Lake Worth Pier, FL	14	0.89	0.90	0.12	0.11	0	0.89	0.89	0.07	0.12	0
Virginia Key, FL	1[a]	0.67	0.85	0.06	0.11	0.02	0.67	0.85	0.06	0.11	0.02

[a]Refer to the optimal models with the highest KGE and NSE scores at the target station but do not meet the transferable model criteria, where both KGE and NSE are above 0.70.

Furthermore, we assess the models' performance in terms of R^2 and RMSE for both extreme events and compare model predictions from the best transferable Bi-LSTM and Bi-LSTM-ATT models using a one-to-one plot (Figures 4c and 4f). Bi-LSTM models have poor generalization of EWLs with low predictive accuracy ($R^2 < 0.50$) and high error (RMSE > 0.30). In contrast, the Bi-LSTM-ATT models can predict the evolution of EWLs with a high predictive accuracy ($R^2 > 0.95$) and low error within an acceptable range (RMSE < 0.15 m). In general, RMSEs below 0.20 m are desirable for hurricane storm surge modeling (Muis et al., 2016). Following this analysis, we hereafter present the results derived from Bi-LSTM-ATT models only.

4.3. Performance of Transferred Models at the Target Stations

Once the Bi-LSTM-ATT models and TL approach have been assessed at selected training and target stations (Section 4.1 and 4.2), we introduce the LSTM-SAM framework to accurately predict the evolution of EWLs at target stations in the U.S. Atlantic Coast (Figure 2d). As mentioned before, we set a time window of 7-day centered around the peak to characterize the evolution of EWLs at the target stations and leverage the 32 Bi-LSTM-ATT models developed at each training station. Note that instead of considering the top-two transferable Bi-LSTM models at the target stations (Figure 4), the LSTM-SAM framework identifies the sets of transferable models from all trained models, for which KGE and NSE are above 0.70 when evaluated with respect to TC or ETC events (Figure 5). For practical flood prediction purposes and decision-making support, Bi-LSTM-ATT models achieving the smallest peak deviation among the transferable models are considered as the optimal ones at each target station (e.g., models with the closest prediction to the peak WL).

For instance, there are 3 transferable models from training station AC to the target station in Montauk, NY that accurately predict EWL evolution of The Perfect Storm (1991) and Hurricane Sandy (2012) (Figures 5a and 5b). These models achieve satisfactory performances such as an average KGE and NSE above 0.75. Likewise, there are 21 transferable models from AC and SW that predict EWL evolution of Hurricane Isabel (2003) and Sandy

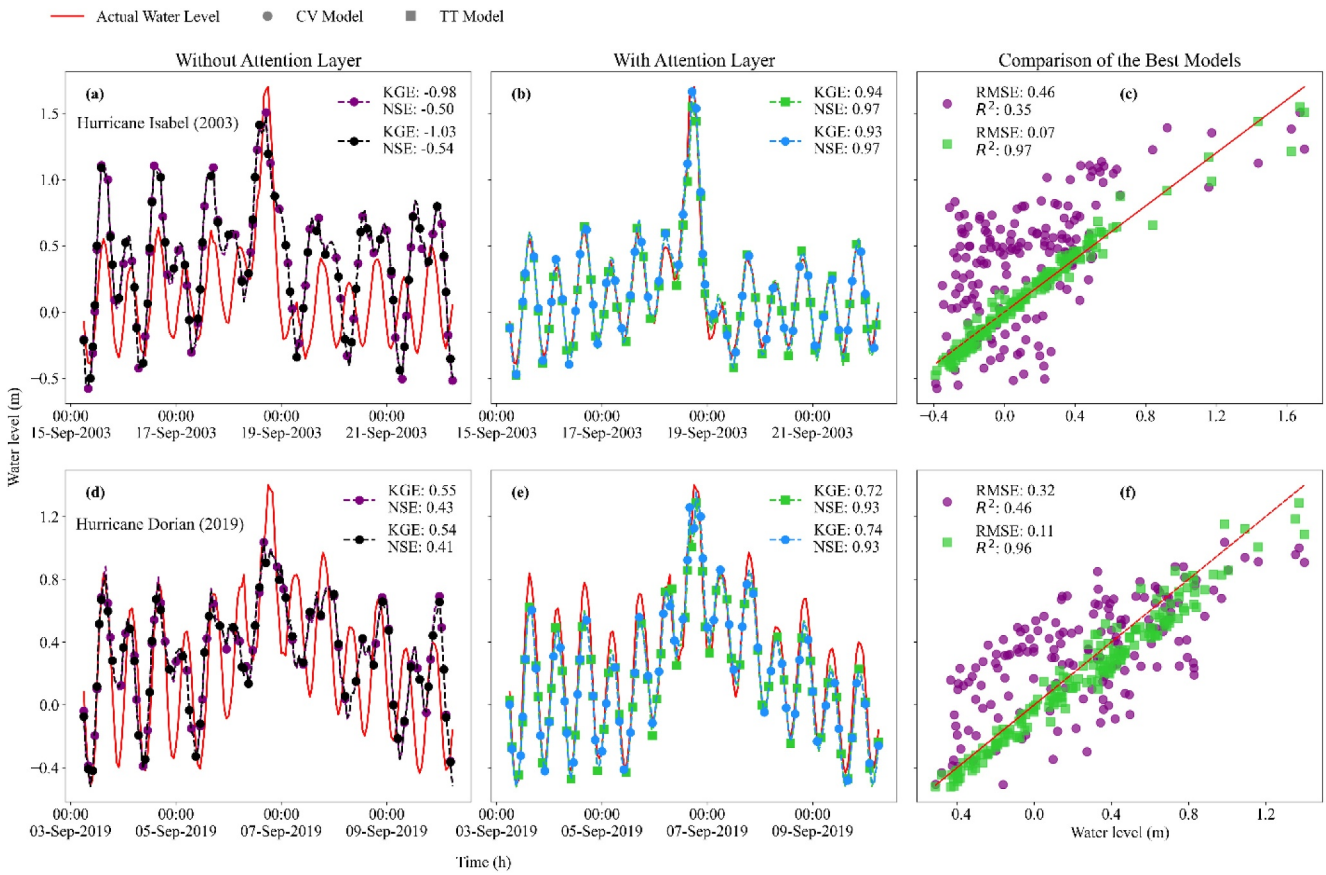

Figure 4. Assessment of transfer learning approach from Sewells Point, VA (training station) to Duck, NC (target station). Prediction of extreme water level evolution using the top two (a), (d) Bi-LSTM, and (b), (e) Bi-LSTM-ATT models for Hurricane Isabel (2003) and Dorian (2019). (c), (f) One-to-one comparison of top two model predictive capabilities based on the hurricane events.

(2012) at the target station in Kiptopeke, VA (Figures 5c and 6d). These models achieve an average KGE and NSE above 0.80. At the target station located in Duck, NC, LSTM-SAM identifies 14 transferable models from SW and BF that predict EWL evolution of Hurricane Isabel (2003) and Dorian (2019) (Figures 5e and 6f). These models show satisfactory performances with average KGE and NSE above 0.75. Similarly, the framework identifies 14 transferable models from TD to the target station in Lake Worth Pier, FL that accurately predict EWL evolution of both Hurricane Dorian (2019) and Nicole (2022) (Figures 5g and 6h). The models achieve average KGE and NSE scores above 0.85.

Results of the remaining target stations show average KGE and NSE ranging between 0.70 and 0.99 (Figure S4 and Table S3 in Supporting Information S1). It is worth noting that we leverage WL data from the training station AC to predict the complete evolution of Hurricane Sandy (2012) at the target station in Sandy Hook, NJ (Figure S4b in Supporting Information S1). This demonstrates the transfer model capability to predict EWL evolution even when tide-gauges fail or become inoperative. However, there are three target stations for which no Bi-LSTM-ATT models are completely transferable given the criteria that both NSE and KGE should be above 0.70 within the 7-day window (Section 3.4). Specifically, the LSTM-SAM framework does not identify any transferable models from SW and BF stations to the target station in Oregon Inlet, NC that can accurately capture the evolution of EWLs of Hurricane Floyd (1999) and Irene (2011) (Figures 6a and 6b). Similarly, the framework does not identify transferable models from SW and BF stations to USCG Station Hatteras, NC for Hurricane Matthew (2016) and Dorian (2019) (Figures 6c and 6d). Lastly, there are no transferable models from TD station to Virginia Key, Florida for Hurricane Irma (2017) and Nicole (2022) (Figures 6e and 6f). However, note that some Bi-LSTM-ATT models can effectively capture the peak WL within a shorter time window centered around the peak (e.g., ±1 day); hence, the relatively low KGE and NSE at the target stations are explained by an overprediction of WLs occurring before and after the peak WL. Therefore, the LSTM-SAM framework considers

10.1029/2024WR039054

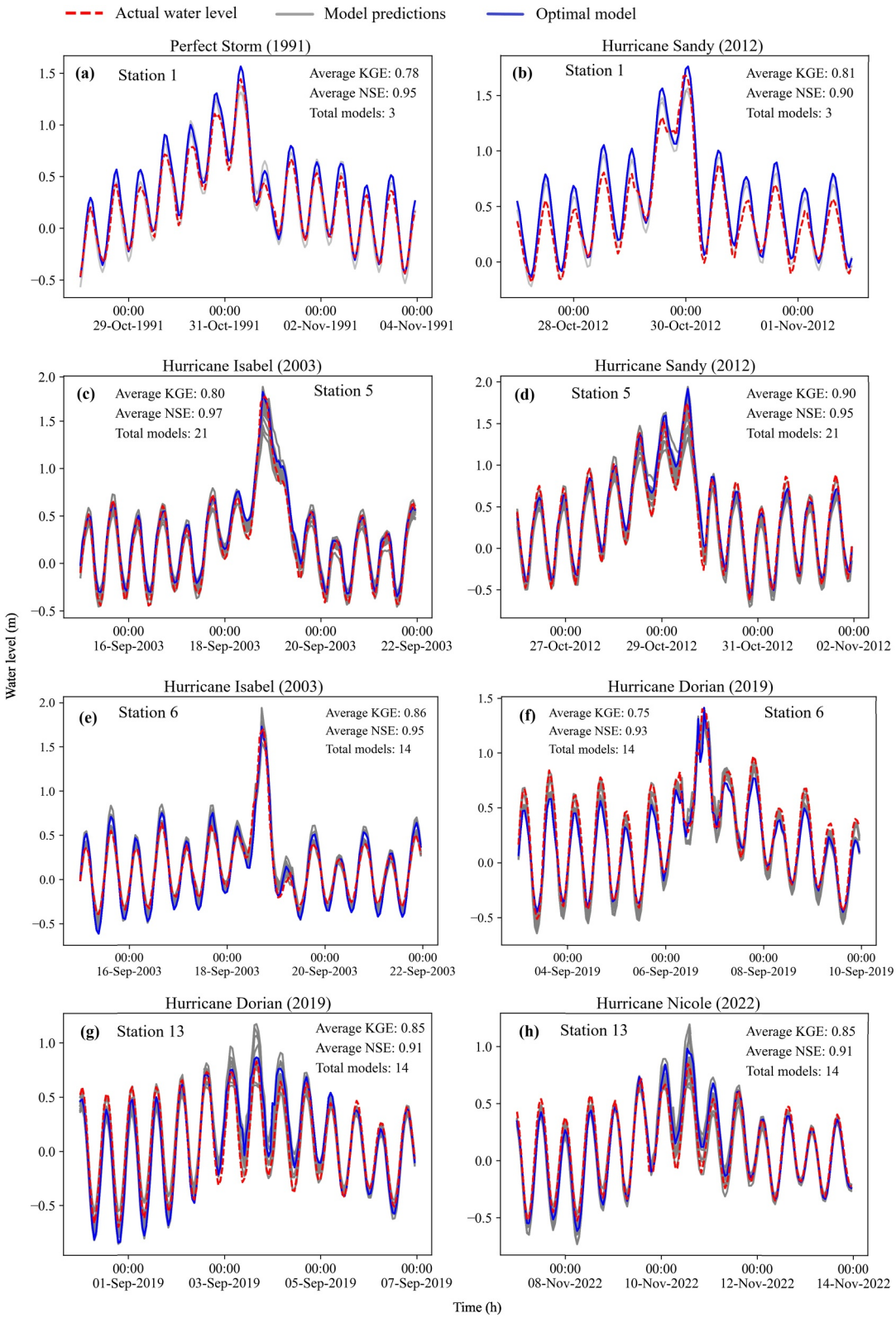

Figure 5. Extreme water level prediction for relevant hurricanes and Nor'easter winter storms in the U.S. Atlantic Coast. Each row panel shows two extreme events at the target stations, the number of transferable models, and their average performance in terms of KGE and NSE. These stations are (a), (b) Montauk, NY (c), (d) Kiptopeke, VA (e), (f) Duck, NC, and (g), (h) Lake Worth Pier, FL. The dashed red, blue, and gray lines represent observed water levels, optimal model, and water level predictions from all transferable models.

the model with the highest KGE and NSE as the optimal models for those three target stations (Figure S4 in Supporting Information S1).

5. Discussion

Physically based models can accurately predict EWLs; however, they are site-specific and not transferable to other domains, even with similar characteristics, due to their need for detailed topographic and bathymetric (topobathy) data (Bates, 2022; Santiago-Collazo et al., 2019). A feasible alternative to overcome this limitation consists of leveraging *state-of-the-art* deep learning models, such as the Bi-LSTM networks, given their effectiveness for learning dynamic and/or sequential data, including nonlinear interactions and hidden patterns from hydrometeorological input features (Tedesco et al., 2023). Conveniently, Bi-LSTM networks enable time-series prediction even in the absence of geographical information or catchment characteristics that may remain quasi-invariant for a relatively long time (e.g., average slope, length, width, catchment size, among other input features). Although model transferability is still challenging (Kratzert et al., 2024; Zhao et al., 2021), adequate feature engineering procedures for selecting inputs (Merizalde et al., 2023) and improving the Bi-LSTM models architecture (e.g., incorporating attention mechanisms) can increase the effectiveness of TL approaches over untrained (target) sites. However, the selection of input features and modifications to the NN architecture should be guided by the physical processes influencing the target variable. This in turn will help advance flood prediction efforts in large scale domains with high accuracy and less computational time (Ding et al., 2020; Li et al., 2021; Nearing et al., 2024).

5.1. Application of the LSTM-SAM Framework

The proposed LSTM-SAM framework is trained on sequential WL data from tide-gage stations and follows a time-series decomposition to obtain hidden patterns and nonlinear associations such as seasonality, trend, harmonic tides, and NTR components (Section 2.3). Moreover, these input features are complemented with hydrometeorological data from reanalysis data sets. As a result, the framework is capable of applying learned knowledge to selected target stations and predicting the evolution of EWLs (Section 3.1). We ensure that Bi-LSTM networks are robustly calibrated through hyperparameter tuning, which resulted in 32 distinctive models across five training stations and a unique set of hyperparameters per model (Section 3.2). We argue that this method preserves inherent pattern recognition capabilities for each model (e.g., model weights) and increases the chances for identifying effective transferable models to target stations.

Conventional Bi-LSTM models show poor performance to predict EWLs at the training station particularly when there exist more extreme events in the test set (Figures 3a and 3b). This is partly due to a sequential training (80%) and test (20%) split, which may align with climate change influences on the frequency and magnitude of EWLs (Bloemendaal et al., 2022; Boumis et al., 2023; Santiago-Collazo et al., 2019). In contrast, leveraging randomly selected subsamples in each batch during the calibration process facilitates a quicker model convergence (De la Fuente et al., 2024) and prevents anomalies in the training data. Nevertheless, training Bi-LSTM models on sequential data ensures that temporal relationships are fully considered in the learning process. Since these models are in general most effective in capturing changing trends of cyclic patterns (Wang et al., 2023), they do ignore some nuanced information of rare and abrupt changes during the training process. Therefore, there is a higher chance of incorrect estimation of equally rare but more extreme data values in testing sets which limits the model's ability to accurately predict EWLs.

Following this, we introduce a custom attention layer in the Bi-LSTM architecture (Section 3.3), which significantly improves the models' ability to capture TC and ETC events (Figure 3b). By amplifying the top 10% of the attention scores, the model's ability to internally focus on the most relevant time steps is substantially improved and leads to more accurate EWL predictions. Moreover, the attention mechanism layer allows the models to perform identical operations consistently beyond the training set and therefore generalizes unseen data with similar characteristics in the test set. Since Bi-LSTM-ATT outperforms conventional Bi-LSTM models (Figures 3f–3h), the proposed LSTM-SAM framework demonstrates the ability to effectively predict the evolution of EWLs even when higher WLs attributed to more frequent TCs and ETCs are expected.

We also observe satifactory performance for most of Bi-LSTM-ATT models developed using specified look-back, batch size, loss function, and data training strategy combination. Models developed with the same

10.1029/2024WR039054

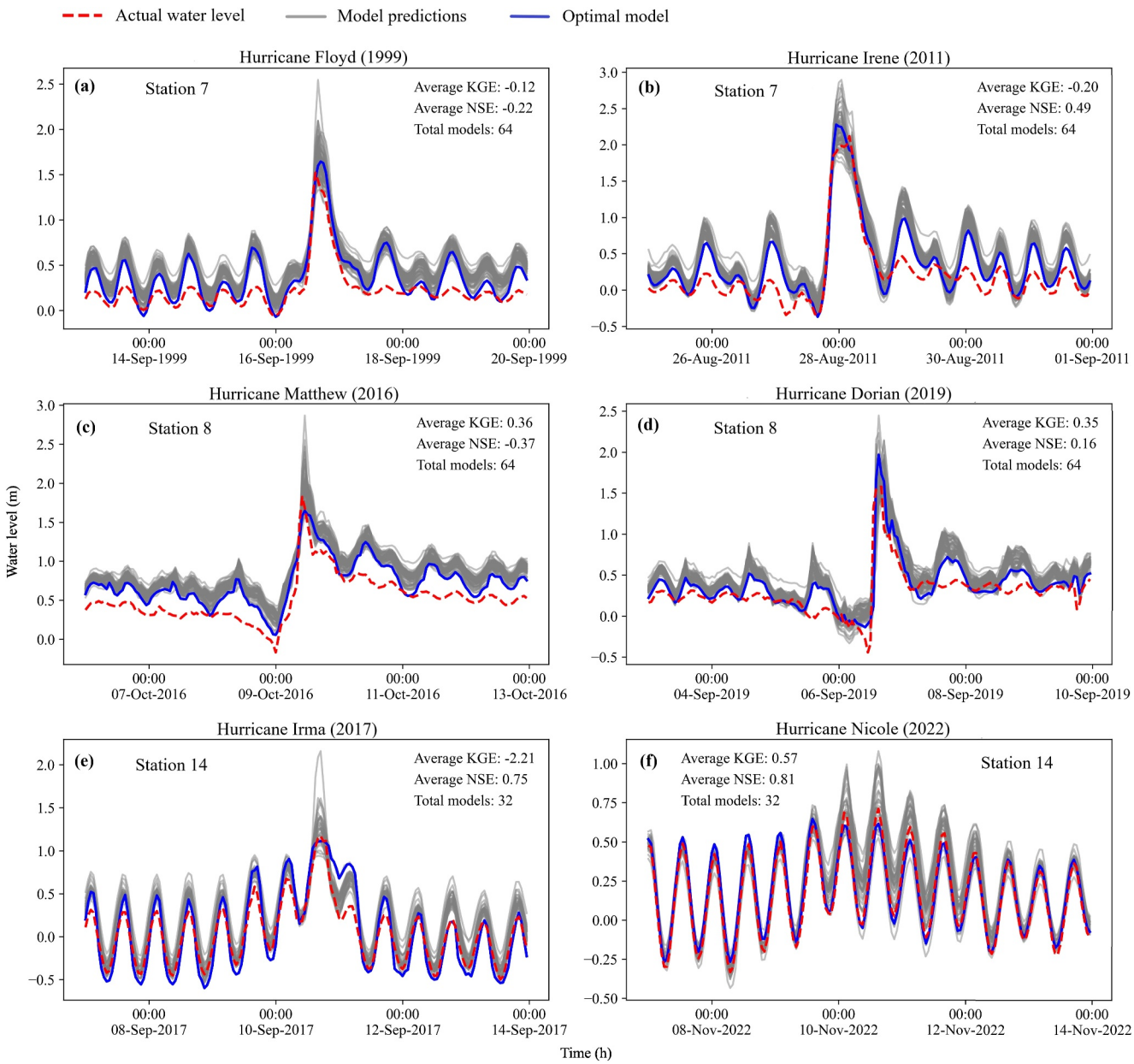

Figure 6. Extreme water level prediction for relevant hurricane events in the U.S. Atlantic Coast. Each row panel shows two extreme events at the target stations, total number of models, and their average performance in terms of KGE and NSE. These stations are (a), (b) Oregon Inlet Marina, NC (c), (d) USCG Station Hatteras, and (e), (f) Virginia Key, FL. The dashed red, blue, and gray lines represent observed water levels, optimal model, and water level predictions from all available Bi-LSTM-ATT models developed at the corresponding training stations.

hyperparameter combinations across all stations achieve consistently more accurate magnitude capture, as reflected in higher KGE scores (Figure S2 in Supporting Information S1). However, their timing accuracy, represented by NSE scores, shows significant variability among the stations (Figure S3 in Supporting Information S1). In addition, average performance for models developed with 24-hr lookback (KGE = 0.91 and NSE = 0.82) compared to 6-hr lookback (KGE = 0.94 and NSE = 0.88) suggests that considering more previous time steps does not necessarily improve the model's predictive accuracy. Notably, models with a batch size of 128 sample points show consistently satistisfactory performance in terms of KGE and NSE for all training stations when processed in 6-hr sequences using MSE as loss function (Figures S2c-S2d and S3c-S3d in Supporting Information S1). This suggests that these configurations effectively capture both the peak and timing of EWLs with higher accuracy.

Although most of the trained Bi-LSTM-ATT models can predict EWLs at the nearby target station, we observe that models with KGE and NSE above 0.70 demonstrate robust generalization capabilities of EWL evolution from onset, peak, to dissipation (Figure 4). Nash-Sutcliffe Efficiency shows a higher accuracy for predictions with correct timing despite the over- or underprediction of WLs (Figures 6e and 6f). In addition, an extended window size renders minor discrepancies in the timing of peak WLs such as 1-hr lead or 1-hr lag negligible in the overall prediction performance. On the other hand, KGE improves when the magnitude of predictions closely aligns with actual events, even if the prediction timing is inconsistent with actual WL observations (Figures 6c and 6d).

The low performance at target stations located at Oregon Inlet Marina, NC (NOAA ID: 8652587) and USCG Station Hatteras, NC (NOAA ID: 8654467) might be attributed to the geographic location of the tide-gauges (Figure 2d). Unlike other target stations that are directly exposed to the Atlantic Ocean, these stations are located behind Bodie's and Cape Hatteras' islands of the Outer Banks barrier island chain. Coastal areas surrounding the target stations experience about 1 m of mean tidal range on the ocean side and 0.30 m behind the island (Velasquez-Montoya et al., 2020). In addition, these areas benefit from vast coastal wetlands and protection infrastructure such as the Herbert C. Bonner bridge that alters tidal dynamics and attenuates storm surges and waves (Velasquez-Montoya et al., 2021, 2022). These stations are sheltered and would have different seabed or bottom roughness conditions, resulting in site-specific water level variability that differ from those at nearby training stations, including Sewells Point, VA (NOAA ID: 8638610) and Beaufort, NC (NOAA ID: 8656483). Excluding wave contributions as input features when developing TL models might improve the model performance at these stations.

Nevertheless, Bi-LSTM-ATT models have the potential to capture the peak WL which is crucial for supporting flood emergency response and decision-making. In fact, the optimal models for those target stations correctly capture the peaks of Hurricane Floyd (1999), Irene (2011), Matthew (2016), and Dorian (2019) (Figures 6a–6d). Similarly, the evolution of EWLs at Virginia Key, Florida (NOAA ID: 8723214) is overestimated (Figures 3e and 3f). Nevertheless, the models can effectively capture the peaks of Hurricane Irma (2017) and Nicole (2022). The tide-gage is located at the entrance of Biscayne Bay close to the Bear Cut bridge, which is most likely responsible for WL being less representative of extreme events. The relatively lower water surface elevation of this station during Hurricane Irma compared to Trident Pier, FL (NOAA ID: 8721604) has also been noted in another study (Alarcon et al., 2022).

In the present study, we highlight that poor TL performance in target domains sharing similar characteristics to the training stations is related to the model's hyperparameter combinations and pattern recognition ability after training. While efforts to optimize hyperparameters through tuning have proven beneficial in training domains (Wendi Li et al., 2021), there are still limitations associated with obtaining the “best possible model” for TL to target domains. For instance, setting predefined ranges for hyperparameter tuning constrains the search space during tuning to values that ensure faster convergence. Changing these ranges could yield different sets of hyperparameters each time. Similarly, if some hyperparameters have fixed values, the other hyperparameters identified through tuning will likely vary each time the fixed values change. It is important to note that the different combinations of hyperparameters affect how models learn and capture patterns (Choudhury et al., 2021). These unique combinations can lead to models having similar levels of high accuracy in the training domain, with at least one combination achieving satisfactory performance in the target domain.

Overall, we developed 32 models for each of the five training stations through hyperparameter tuning (i.e., 160 models in total). The optimal combination of hyperparameters for each station varies, making it challenging to identify a unique combination that performs well across all stations (training and target). Exploring the variability in hyperparameter combinations only increases the likelihood of finding the most suitable TL models with satisfactory performance at the target stations. Note that hyperparameter tuning primarily focuses on identifying the optimal set of model parameters for achieving the most accurate prediction on the training set. However, there is no guarantee that this optimal set will result in improved performance on unseen (test) sets (Tran et al., 2020). This is evident with the inability of the models trained at Fort Pulaski, GA, to accurately predict its test data. Therefore, it is ultimately the model's ability to effectively recognize EWL patterns, which differ from those captured during training, in unseen cyclonic data set that improves its TL capabilities—a feature attributed to the attention mechanism incorporated within the framework.

5.2. Limitations and Future Work

The waves and atmospheric variables obtained from ERA5 have a spatial resolution of approximately 31 km. This level of resolution may not correctly account for local variability; hence higher resolution data might improve the performance of the LSTM-SAM framework. There are instances where wave components for some target stations could not be directly derived from ERA-5, like Lewes, DE, and Charleston, SC. To overcome this challenge, we leveraged wave data from the U.S. Army Corps of Engineers (USACE)'s Wave Information Studies (WIS) data archive (https://wisportal.erdc.dren.mil/#). The WIS portal provides consistent, hourly, and long-term wave climatology along the U.S. coastlines. As a result, some errors could have been introduced in the input features, reducing the accuracy of EWL predictions. Interestingly, these target stations have the lowest number of transferable models compared to other stations (Figure S4 in Supporting Information S1). For the target stations of Oregon Inlet Marina, NC, and USCG Station Hatteras, NC, an in-depth exploration of the model's sensitivity to variations in input features would aid in understanding their impact on the model's predictions. Quantifying the local and global contributions of each feature, including understanding how features interact with each other and how these interactions affect the model's predictions, is a promising direction for future improvements. While KGE and NSE provide robust validation for model performance, future work could benefit from incorporating other statistical tests for validating metaheuristics results to highlight the significance of performance differences among multiple models (Derrac et al., 2011).

We plan to extend the LSTM-SAM framework to inland target stations by taking into account the contribution of river discharge for accurate prediction of total WLs in coastal to inland transition zones (Bilskie & Hagen, 2018; Muñoz, Yin, et al., 2022; Serafin et al., 2017). More advanced deep learning models like Transformers have built-in self-attention mechanisms (Boussioux et al., 2022) and could be a worth-exploring alternative to the proposed Bi-LSTM-ATT models for predicting EWL evolution in coastal areas. Future work should focus on predicting spatiotemporal WL variability and flood inundation extent by combining the LSTM-SAM framework with NN architectures that can accurately account for spatial information (e.g., topography and bathymetry), such as Convolutional Neural Networks (Gavahi et al., 2021; Shahabi & Tahvildari, 2024), by using architectures like Generative Adversarial Networks to integrate conditional inputs, including soil moisture and land area index (Foroumandi et al., 2024). While the simulation of complex processes and their nonlinear interactions is best suited to physically based models like ADCIRC, Delft3D, and HEC-RAS among others, interpretable machine learning could help track nonlinear associations among input features (e.g., time-series of hydrometeorological and oceanic variables) and EWL variability as the target variable. There exist relatively new packages, such as SHAP (SHapley Additive exPlanations; https://shap.readthedocs.io/en/latest/), that aid in the interpretation of machine learning architectures and provide insight into how these features are contributing to the models.

6. Conclusion

We characterize the evolution of extreme water levels (EWLs) at tide-gage stations distributed along the U.S. Atlantic Coast. To achieve this, we identify 5 training stations that were hit by historic hurricane events and contain complete consecutive hourly data spanning at least 40 years. Then, we leverage available WL and hydrometeorological time-series data to train bidirectional Long Short-Term Memory (Bi-LSTM) network models for each training station. Furthermore, we incorporate an attention mechanism layer in the model architecture and a transfer learning (TL) approach with the goal effectively predicting the evolution of EWLs at target (tide-gage) stations. The collection of models with the attention mechanism layer and TL approach is referred to as the LSTM-Station Approximated Models (LSTM-SAM) framework and is effectively applied to 14 target stations. The LSTM-SAM framework predicts the onset, peak, and dissipation of multiple EWL events emerging from tropical cyclones (hurricanes) and extratropical cyclones (Nor'easter storms) with high accuracy. For this, the framework identifies “transferable” models based on KGE and NSE above 0.70 to ensure an accurate generalization of EWLs. Under these criteria, the LSTM-SAM framework demonstrates satisfactory performance with transferable models achieving average KGE, NSE, and RMSE ranging from 0.78 to 0.92, 0.90 to 0.97, and 0.09 to 0.18 at the target stations, respectively.

Following these results, we conclude that the technique aimed at improving the model's ability to effectively identify EWL patterns beyond those observed in the training phase should result in satisfactory model performance in target stations. Such technique enables the detailed capture of nuanced information and low frequency water levels often overlooked by most NN models, as their training is typically based on identifying repetitive

patterns in the training data. Incorporating this technique into the model architecture will sustain its operation in both the training and testing phases, while enhancing the model's pattern-capturing ability. Although hyperparameter combinations significantly influence model performance, the priority should be to enhance the model's ability to accurately understand low frequency periods that corresponds to EWL patterns, rather than over-optimizing the architecture for training accuracy. The LSTM-SAM framework demonstrates the effectiveness of such technique (i.e., attention mechanism) and accurately predicts not only EWLs but also their evolution over time. This capacity could support in large-scale operational flood predictions like the National Water Model (NWM) or Coastal Emergency Risk Assessment (CERA). Future work will focus on predicting spatiotemporal WL variability and flood inundation extent, for example, combining Bi-LSTM-ATT and Convolutional Neural Networks.

Data Availability Statement

The source code of Bi-LSTM networks and data used are publicly available. Links to data repositories and archives have been provided throughout the manuscript. Bi-LSTM on Tensorflow: https://www.tensorflow.org/api_docs/python/tf/keras/layers/Bidirectional. Water level data: https://tidesandcurrents.noaa.gov/map/index.html. Meteorological and wave data: https://cds.climate.copernicus.eu/ and https://wisportal.erdc.dren.mil/#.

Acknowledgments
Partial financial support for this study is provided by the National Science Foundation, CAS-Climate Program (Award # 480948) and the Virginia Sea Grant Fellowship (Award # 464069).

References

Abbaszadeh, P., Gavahi, K., & Moradkhani, H. (2020). Multivariate remotely sensed and in-situ data assimilation for enhancing community WRF-Hydro model forecasting. *Advances in Water Resources*, *145*, 103721. https://doi.org/10.1016/j.advwatres.2020.103721

Alarcon, V. J., Linhoss, A. C., Kelble, C. R., Mickle, P. F., Sanchez-Banda, G. F., Mardonez-Meza, F. E., et al. (2022). Coastal inundation under concurrent mean and extreme sea-level rise in Coral Gables, Florida, USA. *Natural Hazards*, *111*(3), 2933–2962. https://doi.org/10.1007/s11069-021-05163-0

Alipour, A., Jafarzadegan, K., & Moradkhani, H. (2022). Global sensitivity analysis in hydrodynamic modeling and flood inundation mapping. *Environmental Modelling and Software*, *152*, 105398. https://doi.org/10.1016/j.envsoft.2022.105398

Altunkaynak, A., & Kartal, E. (2021). Transfer sea level learning in the Bosphorus Strait by wavelet based machine learning methods. *Ocean Engineering*, *233*, 109116. https://doi.org/10.1016/j.oceaneng.2021.109116

Anderson, D. L., Ruggiero, P., Mendez, F. J., Barnard, P. L., Erikson, L. H., O'Neill, A. C., et al. (2021). Projecting climate dependent coastal flood risk with a hybrid statistical dynamical model. *Earth's Future*, *9*(12), e2021EF002285. https://doi.org/10.1029/2021EF002285

Bai, L.-H., & Xu, H. (2021). Accurate estimation of tidal level using bidirectional long short-term memory recurrent neural network. *Ocean Engineering*, *235*, 108765. https://doi.org/10.1016/j.oceaneng.2021.108765

Bandara, K., Bergmeir, C., & Smyl, S. (2020). Forecasting across time series databases using recurrent neural networks on groups of similar series: A clustering approach. *Expert Systems with Applications*, *140*, 112896. https://doi.org/10.1016/j.eswa.2019.112896

Bates, P. (2023a). Fundamental limits to flood inundation modelling. *Nat Water*, *1*(7), 566–567. https://doi.org/10.1038/s44221-023-00106-4

Bates, P. (2023b). Uneven burden of urban flooding. *Nature Sustainability*, *6*(1), 9–10. https://doi.org/10.1038/s41893-022-01000-9

Bates, P. D. (2022). Flood inundation prediction. *Annual Review of Fluid Mechanics*, *54*(1), 287–315. https://doi.org/10.1146/annurev-fluid-030121-113138

Bentivoglio, R., Isufi, E., Jonkman, S. N., & Taormina, R. (2022). Deep learning methods for flood mapping: A review of existing applications and future research directions. *Hydrology and Earth System Sciences*, *26*(16), 4345–4378. https://doi.org/10.5194/hess-26-4345-2022

Bevacqua, E., Maraun, D., Vousdoukas, M. I., Voukouvalas, E., Vrac, M., Mentaschi, L., & Widmann, M. (2019). Higher probability of compound flooding from precipitation and storm surge in Europe under anthropogenic climate change. *Science Advances*, *5*(9), eaaw5531. https://doi.org/10.1126/sciadv.aaw5531

Bian, G.-F., Nie, G.-Z., & Qiu, X. (2021). How well is outer tropical cyclone size represented in the ERA5 reanalysis dataset? *Atmospheric Research*, *249*, 105339. https://doi.org/10.1016/j.atmosres.2020.105339

Bilskie, M. V., & Hagen, S. C. (2018). Defining flood zone transitions in low-gradient coastal regions. *Geophysical Research Letters*, *45*(6), 2761–2770. https://doi.org/10.1002/2018GL077524

Bilskie, M. V., Zhao, H., Resio, D., Atkinson, J., Cobell, Z., & Hagen, S. C. (2021). Enhancing flood hazard assessments in coastal Louisiana through coupled hydrologic and surge processes. *Front. Water*, *3*. https://doi.org/10.3389/frwa.2021.609231

Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., et al. (2023). Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *WIREs Data Mining and Knowledge Discovery*, *13*(2), e1484. https://doi.org/10.1002/widm.1484

Bloemendaal, N., de Moel, H., Martinez, A. B., Muis, S., Haigh, I. D., van der Wiel, K., et al. (2022). A globally consistent local-scale assessment of future tropical cyclone risk. *Science Advances*, *8*(17), eabm8438. https://doi.org/10.1126/sciadv.abm8438

Bloemendaal, N., Muis, S., Haarsma, R. J., Verlaan, M., Irazoqui Apecechea, M., de Moel, H., et al. (2019). Global modeling of tropical cyclone storm surges using high-resolution forecasts. *Climate Dynamics*, *52*(7–8), 5031–5044. https://doi.org/10.1007/s00382-018-4430-x

Boumis, G., Moftakhari, H. R., & Moradkhani, H. (2023). Coevolution of extreme sea levels and sea-level rise under global warming. *Earth's Future*, *11*(7), e2023EF003649. https://doi.org/10.1029/2023EF003649

Boussioux, L., Zeng, C., Guénais, T., & Bertsimas, D. (2022). Hurricane forecasting: A novel multimodal machine learning framework. *Weather and Forecasting*, *37*(6), 817–831. https://doi.org/10.1175/WAF-D-21-0091.1

Bruneau, N., Polton, J., Williams, J., & Holt, J. (2020). Estimation of global coastal sea level extremes using neural networks. *Environmental Research Letters*, *15*(7), 074030. https://doi.org/10.1088/1748-9326/ab89d6

Chen, D., Zhang, J., & Jiang, S. (2020). Forecasting the short-term metro ridership with seasonal and trend decomposition using loess and LSTM neural networks. *IEEE Access*, *8*, 91181–91187. https://doi.org/10.1109/ACCESS.2020.2995044

Chen, Z., Xu, H., Jiang, P., Yu, S., Lin, G., Bychkov, I., et al. (2021). A transfer Learning-Based LSTM strategy for imputing Large-Scale consecutive missing data and its application in a water quality prediction system. *Journal of Hydrology*, *602*, 126573. https://doi.org/10.1016/j.jhydrol.2021.126573

Choudhury, P., Allen, R. T., & Endres, M. G. (2021). Machine learning for pattern discovery in management research. *Strategic Management Journal*, *42*(1), 30–57. https://doi.org/10.1002/smj.3215

Cleveland, R. B., Cleveland, W. S., & Terpenning, I. (1990). STL: A seasonal-trend decomposition procedure based on loess 3.

Codiga, D. (2011). Unified tidal analysis and prediction using the UTide Matlab functions. https://doi.org/10.13140/RG.2.1.3761.2008

de Amorim, L. B. V., Cavalcanti, G. D. C., & Cruz, R. M. O. (2023). The choice of scaling technique matters for classification performance. *Applied Soft Computing*, *133*, 109924. https://doi.org/10.1016/j.asoc.2022.109924

De la Fuente, L. A., Ehsani, M. R., Gupta, H. V., & Condon, L. E. (2024). Toward interpretable LSTM-based modeling of hydrological systems. *Hydrology and Earth System Sciences*, *28*(4), 945–971. https://doi.org/10.5194/hess-28-945-2024

Derrac, J., García, S., Molina, D., & Herrera, F. (2011). A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. *Swarm and Evolutionary Computation*, *1*, 3–18. https://doi.org/10.1016/j.swevo.2011.02.002

Ding, Y., Zhu, Y., Feng, J., Zhang, P., & Cheng, Z. (2020). Interpretable spatio-temporal attention LSTM model for flood forecasting. *Neurocomputing*, *403*, 348–359. https://doi.org/10.1016/j.neucom.2020.04.110

Dong, Y., Zhang, Y., Liu, F., & Cheng, X. (2021). Reservoir production prediction model based on a stacked LSTM network and transfer learning. *ACS Omega*, *6*(50), 34700–34711. https://doi.org/10.1021/acsomega.1c05132

Fang, Z., Wang, Y., Peng, L., & Hong, H. (2021). Predicting flood susceptibility using LSTM neural networks. *Journal of Hydrology*, *594*, 125734. https://doi.org/10.1016/j.jhydrol.2020.125734

Foroumandi, E., Gavahi, K., & Moradkhani, H. (2024). Generative adversarial network for real-time flash drought monitoring: A deep learning study. *Water Resources Research*, *60*(5), e2023WR035600. https://doi.org/10.1029/2023WR035600

Fraehr, N., Wang, Q. J., Wu, W., & Nathan, R. (2022). Upskilling low-fidelity hydrodynamic models of flood inundation through spatial analysis and Gaussian process learning. *Water Resources Research*, *58*(8), e2022WR032248. https://doi.org/10.1029/2022WR032248

Gavahi, K., Abbaszadeh, P., & Moradkhani, H. (2021). DeepYield: A combined convolutional neural network with long short-term memory for crop yield forecasting. *Expert Systems with Applications*, *184*, 115511. https://doi.org/10.1016/j.eswa.2021.115511

Ghanbari, M., Arabi, M., Kao, S.-C., Obeysekera, J., & Sweet, W. (2021). Climate change and changes in compound coastal-riverine flooding hazard along the U.S. Coasts. *Earth's Future*, *9*(5), e2021EF002055. https://doi.org/10.1029/2021EF002055

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics. Presented at the proceedings of the thirteenth international conference on artificial intelligence and statistics* (pp. 249–256). JMLR Workshop and Conference Proceedings.

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, *377*(1–2), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Hamitouche, M., & Molina, J.-L. (2022). A review of AI methods for the prediction of high-flow extremal hydrology. *Water Resources Management*, *36*(10), 3859–3876. https://doi.org/10.1007/s11269-022-03240-y

Hewamalage, H., Bergmeir, C., & Bandara, K. (2021). Recurrent neural networks for time series forecasting: Current status and future directions. *International Journal of Forecasting*, *37*(1), 388–427. https://doi.org/10.1016/j.ijforecast.2020.06.008

Hino, M., & Nance, E. (2021). Five ways to ensure flood-risk research helps the most vulnerable. *Nature*, *595*(7865), 27–29. https://doi.org/10.1038/d41586-021-01750-0

Huang, C., Zhang, J., Cao, L., Wang, L., Luo, X., Wang, J.-H., & Bensoussan, A. (2020). Robust forecasting of river-flow based on convolutional neural network. *IEEE Transactions on Sustainable Computing*, *5*(4), 594–600. https://doi.org/10.1109/TSUSC.2020.2983097

Jafarzadegan, K., Alipour, A., Gavahi, K., Moftakhari, H., & Moradkhani, H. (2021). Toward improved river boundary conditioning for simulation of extreme floods. *Advances in Water Resources*, *158*, 104059. https://doi.org/10.1016/j.advwatres.2021.104059

Jay, D. A., & Flinchem, E. P. (1999). A comparison of methods for analysis of tidal records containing multi-scale non-tidal background energy. *Continental Shelf Research*, *19*(13), 1695–1732. https://doi.org/10.1016/S0278-4343(99)00036-9

Kardhana, H., Valerian, J. R., Rohmat, F. I. W., & Kusuma, M. S. B. (2022). Improving Jakarta's katulampa barrage extreme water level prediction using satellite-based long short-term memory (LSTM) neural networks. *Water*, *14*(9), 1469. https://doi.org/10.3390/w14091469

Khojasteh, D., Glamore, W., Heimhuber, V., & Felder, S. (2021). Sea level rise impacts on estuarine dynamics: A review. *Science of the Total Environment*, *780*, 146470. https://doi.org/10.1016/j.scitotenv.2021.146470

Kingphai, K., & Moshfeghi, Y. (2023). On time series cross-validation for deep learning classification model of mental workload levels based on EEG signals. In G. Nicosia, V. Ojha, E. La Malfa, G. La Malfa, P. Pardalos, G. Di Fatta, et al. (Eds.), *Machine learning, optimization, and data science* (pp. 402–416). Springer. https://doi.org/10.1007/978-3-031-25891-6_30

Kratzert, F., Gauch, M., Klotz, D., & Nearing, G. (2024). HESS Opinions: Never train an LSTM on a single basin. *Hydrology and Earth System Sciences Discussions*, 1–19. https://doi.org/10.5194/hess-2023-275

Lee, J.-W., Irish, J. L., Bensi, M. T., & Marcy, D. C. (2021a). Rapid prediction of peak storm surge from tropical cyclone track time series using machine learning. *Coastal Engineering*, *170*, 104024. https://doi.org/10.1016/j.coastaleng.2021.104024

Lee, J.-W., Irish, J. L., Bensi, M. T., & Marcy, D. C. (2021b). Rapid prediction of peak storm surge from tropical cyclone track time series using machine learning. *Coastal Engineering*, *170*, 104024. https://doi.org/10.1016/j.coastaleng.2021.104024

Li, W., Kiaghadi, A., & Dawson, C. (2021a). Exploring the best sequence LSTM modeling architecture for flood prediction. *Neural Computing & Applications*, *33*(11), 5571–5580. https://doi.org/10.1007/s00521-020-05334-3

Li, W., Wendi, Y., Ng, W. W., Wang, T., Pelillo, M., & Kwong, S. (2021b). HELP: An LSTM-based approach to hyperparameter exploration in neural network learning. *Neurocomputing*, *442*, 161–172. https://doi.org/10.1016/j.neucom.2020.12.133

Liu, Y., Yang, Y., Chin, R. J., Wang, C., & Wang, C. (2023). Long short-term memory (LSTM) based model for flood forecasting in Xiangjiang river. *KSCE Journal of Civil Engineering*, *27*(11), 5030–5040. https://doi.org/10.1007/s12205-023-2469-7

Ma, J., Cheng, J. C. P., Jiang, F., Chen, W., Wang, M., & Zhai, C. (2020). A bi-directional missing data imputation scheme based on LSTM and transfer learning for building energy data. *Energy and Buildings*, *216*, 109941. https://doi.org/10.1016/j.enbuild.2020.109941

Mahmoudi, S., Moftakhari, H., Muñoz, D. F., Sweet, W., & Moradkhani, H. (2024). Establishing flood thresholds for sea level rise impact communication. *Nature Communications*, *15*(1), 4251. https://doi.org/10.1038/s41467-024-48545-1

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The M4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, *M4*(Competition 36), 54–74. https://doi.org/10.1016/j.ijforecast.2019.04.014

Marco, R., Ahmad, S. S. S., & Ahmad, S. (2022). Bayesian hyperparameter optimization and ensemble learning for machine learning models on software effort estimation. *International Journal of Advanced Computer Science and Applications*, *13*(3). https://doi.org/10.14569/IJACSA.2022.0130351

Marsooli, R., & Wang, Y. (2020). Quantifying tidal phase effects on coastal flooding induced by hurricane sandy in Manhattan, New York using a micro-scale hydrodynamic model. *Front. Built Environ.*, *6*. https://doi.org/10.3389/fbuil.2020.00149

Merizalde, M. J., Muñoz, P., Corzo, G., Muñoz, D. F., Samaniego, E., & Célleri, R. (2023). Integrating geographic data and the SCS-CN method with LSTM networks for enhanced runoff forecasting in a complex mountain basin. *Front. Water*, *5*. https://doi.org/10.3389/frwa.2023.1233899

Moftakhari, H., Muñoz, D. F., Akbari Asanjan, A., AghaKouchak, A., Moradkhani, H., & Jay, D. A. (2024). Nonlinear interactions of sea-level rise and storm tide alter extreme coastal water levels: How and why? *AGU Advances*, *5*(2), e2023AV000996. https://doi.org/10.1029/2023AV000996

Moftakhari, H., Schubert, J. E., AghaKouchak, A., Matthew, R. A., & Sanders, B. F. (2019). Linking statistical and hydrodynamic modeling for compound flood hazard assessment in tidal channels and estuaries. *Advances in Water Resources*, *128*, 28–38. https://doi.org/10.1016/j.advwatres.2019.04.009

Moftakhari, H. R., Jay, D. A., Talke, S. A., Kukulka, T., & Bromirski, P. D. (2013). A novel approach to flow estimation in tidal rivers. *Water Resources Research*, *49*(8), 4817–4832. https://doi.org/10.1002/wrcr.20363

Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V., & Herrera, F. (2012). A unifying view on dataset shift in classification. *Pattern Recognition*, *45*(1), 521–530. https://doi.org/10.1016/j.patcog.2011.06.019

Muis, S., Lin, N., Verlaan, M., Winsemius, H. C., Ward, P. J., & Aerts, J. C. J. H. (2019). Spatiotemporal patterns of extreme sea levels along the western North-Atlantic coasts. *Scientific Reports*, *9*(1), 3391. https://doi.org/10.1038/s41598-019-40157-w

Muis, S., Verlaan, M., Winsemius, H. C., Aerts, J. C. J. H., & Ward, P. J. (2016). A global reanalysis of storm surges and extreme sea levels. *Nature Communications*, *7*(1), 11969. https://doi.org/10.1038/ncomms11969

Muñoz, D. F., Abbaszadeh, P., Moftakhari, H., & Moradkhani, H. (2022). Accounting for uncertainties in compound flood hazard assessment: The value of data assimilation. *Coastal Engineering*, *171*, 104057. https://doi.org/10.1016/j.coastaleng.2021.104057

Muñoz, D. F., Moftakhari, H., & Moradkhani, H. (2020). Compound effects of flood drivers and wetland elevation correction on coastal flood hazard assessment. *Water Resources Research*, *56*(7), e2020WR027544. https://doi.org/10.1029/2020WR027544

Muñoz, D. F., Muñoz, P., Moftakhari, H., & Moradkhani, H. (2021). From local to regional compound flood mapping with deep learning and data fusion techniques. *Science of the Total Environment*, *782*, 146927. https://doi.org/10.1016/j.scitotenv.2021.146927

Muñoz, D. F., Yin, D., Bakhtyar, R., Moftakhari, H., Xue, Z., Mandli, K., & Ferreira, C. (2022). Inter-model comparison of Delft3D-FM and 2D HEC-RAS for total water level prediction in coastal to inland transition zones. *JAWRA Journal of the American Water Resources Association*, *58*(1), 34–49. https://doi.org/10.1111/1752-1688.12952

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I—A discussion of principles. *Journal of Hydrology*, *10*(3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6

Nearing, G., Cohen, D., Dube, V., Gauch, M., Gilon, O., Harrigan, S., et al. (2024). Global prediction of extreme floods in ungauged watersheds. *Nature*, *627*(8004), 559–563. https://doi.org/10.1038/s41586-024-07145-1

Niu, Z., Zhong, G., & Yu, H. (2021). A review on the attention mechanism of deep learning. *Neurocomputing*, *452*, 48–62. https://doi.org/10.1016/j.neucom.2021.03.091

NOAA-NCEI. (2024). U.S. Billion-Dollar weather and climate disasters, 1980 - Present (NCEI accession 0209268) [WWW document]. Retrieved from https://www.ncei.noaa.gov/archive/accession/0209268. accessed 15 7 2024.

Obara, Y., & Nakamura, R. (2022). Transfer learning of long short-term memory analysis in significant wave height prediction off the coast of western Tohoku, Japan. *Ocean Engineering*, *266*, 113048. https://doi.org/10.1016/j.oceaneng.2022.113048

Office for Coastal Management. (2024). Economics and demographics [WWW document]. Retrieved from https://coast.noaa.gov/states/fast-facts/economics-and-demographics.html. accessed 4 1 2024.

Parker, K., Erikson, L., Thomas, J., Nederhoff, K., Barnard, P., & Muis, S. (2023). Relative contributions of water-level components to extreme water levels along the US Southeast Atlantic Coast from a regional-scale water-level hindcast. *Natural Hazards*, *117*(3), 2219–2248. https://doi.org/10.1007/s11069-023-05939-6

Peng, L., Wu, H., Gao, M., Yi, H., Xiong, Q., Yang, L., & Cheng, S. (2022). Tlt: Recurrent fine-tuning transfer learning for water quality long-term prediction. *Water Research*, *225*, 119171. https://doi.org/10.1016/j.watres.2022.119171

Rainey, J. L., Brody, S. D., Galloway, G. E., & Highfield, W. E. (2021). Assessment of the growing threat of urban flooding: A case study of a national survey. *Urban Water Journal*, *18*(5), 375–381. https://doi.org/10.1080/1573062x.2021.1893356

Rashid, M. M., Moftakhari, H., & Moradkhani, H. (2024). Stochastic simulation of storm surge extremes along the contiguous United States coastlines using the max-stable process. *Commun Earth Environ*, *5*, 1–10. https://doi.org/10.1038/s43247-024-01206-z

Rithani, M., Kumar, R. P., & Doss, S. (2023). A review on big data based on deep neural network approaches. *Artificial Intelligence Review*, *56*(12), 14765–14801. https://doi.org/10.1007/s10462-023-10512-5

Saksena, S., & Merwade, V. (2015). Incorporating the effect of DEM resolution and accuracy for improved flood inundation mapping. *Journal of Hydrology*, *530*, 180–194. https://doi.org/10.1016/j.jhydrol.2015.09.069

Sanders, B. F., Schubert, J. E., Kahl, D. T., Mach, K. J., Brady, D., AghaKouchak, A., et al. (2023). Large and inequitable flood risks in Los Angeles, California. *Nature Sustainability*, *6*(1), 47–57. https://doi.org/10.1038/s41893-022-00977-7

Santiago-Collazo, F. L., Bilskie, M. V., & Hagen, S. C. (2019). A comprehensive review of compound inundation models in low-gradient coastal watersheds. *Environmental Modelling and Software*, *119*, 166–181. https://doi.org/10.1016/j.envsoft.2019.06.002

Seleem, O., Ayzel, G., Bronstert, A., & Heistermann, M. (2023). Transferability of data-driven models to predict urban pluvial flood water depth in Berlin, Germany. *Natural Hazards and Earth System Sciences*, *23*(2), 809–822. https://doi.org/10.5194/nhess-23-809-2023

Serafin, K. A., & Ruggiero, P. (2014). Simulating extreme total water levels using a time-dependent, extreme value approach. *Journal of Geophysical Research: Oceans*, *119*(9), 6305–6329. https://doi.org/10.1002/2014JC010093

Serafin, K. A., Ruggiero, P., & Stockdon, H. F. (2017). The relative contribution of waves, tides, and nontidal residuals to extreme total water levels on U.S. West Coast sandy beaches. *Geophysical Research Letters*, *44*(4), 1839–1847. https://doi.org/10.1002/2016GL071020

Shahabi, A., & Tahvildari, N. (2024). A deep-learning model for rapid spatiotemporal prediction of coastal water levels. *Coastal Engineering*, *190*, 104504. https://doi.org/10.1016/j.coastaleng.2024.104504

Shen, C. (2018). A transdisciplinary review of deep learning research and its relevance for water resources scientists. *Water Resources Research*, *54*(11), 8558–8593. https://doi.org/10.1029/2018WR022643

Siami-Namini, S., Tavakoli, N., & Namin, A. S. (2019). The performance of LSTM and BiLSTM in forecasting time series. In *2019 IEEE international conference on big data (big data). Presented at the 2019 IEEE international conference on big data (big data)* (pp. 3285–3292). https://doi.org/10.1109/BigData47090.2019.9005997

Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., & Liu, C. (2018). A survey on deep transfer learning. In V. Kůrková, Y. Manolopoulos, B. Hammer, L. Iliadis, & I. Maglogiannis (Eds.), *Artificial neural networks and machine learning – ICANN 2018* (pp. 270–279). Springer International Publishing. https://doi.org/10.1007/978-3-030-01424-7_27

Tedesco, P., Rabault, J., Sætra, M. L., Kristensen, N. M., Aarnes, O. J., Breivik, Ø., et al. (2023). Bias correction of operational storm surge forecasts using neural networks. https://doi.org/10.48550/arXiv.2301.00892

Thomas, A., Dietrich, J., Asher, T., Bell, M., Blanton, B., Copeland, J., et al. (2019). Influence of storm timing and forward speed on tides and storm surge during Hurricane Matthew. *Ocean Modelling*, *137*, 1–19. https://doi.org/10.1016/j.ocemod.2019.03.004

Tiggeloven, T., Couasnon, A., van Straaten, C., Muis, S., & Ward, P. J. (2021). Exploring deep learning capabilities for surge predictions in coastal areas. *Scientific Reports*, *11*(1), 17224. https://doi.org/10.1038/s41598-021-96674-0

Tran, N., Schneider, J.-G., Weber, I., & Qin, A. K. (2020). Hyper-parameter optimization in classification: To-do or not-to-do. *Pattern Recognition*, *103*, 107245. https://doi.org/10.1016/j.patcog.2020.107245

Velasquez-Montoya, L., Overton, M. F., & Sciaudone, E. J. (2020). Natural and anthropogenic-induced changes in a tidal inlet: Morphological evolution of Oregon Inlet. *Geomorphology*, *350*, 106871. https://doi.org/10.1016/j.geomorph.2019.106871

Velasquez-Montoya, L., Sciaudone, E. J., Smyre, E., & Overton, M. F. (2021). Vulnerability indicators for coastal roadways based on barrier island morphology and shoreline change predictions. *Natural Hazards Review*, *22*(2), 04021003. https://doi.org/10.1061/(ASCE)NH.1527-6996.0000441

Velasquez-Montoya, L., Wargula, A., Nangle, J., Sciaudone, E., Smyre, E., & Tomiczek, T. (2022). Hydrodynamics of a tidal inlet under gray to green coastal protection interventions. *Frontiers in Earth Science*, *10*. https://doi.org/10.3389/feart.2022.991667

Wahl, T., Haigh, I. D., Nicholls, R. J., Arns, A., Dangendorf, S., Hinkel, J., & Slangen, A. B. A. (2017). Understanding extreme sea levels for broad-scale coastal impact and adaptation analysis. *Nature Communications*, *8*(1), 16075. https://doi.org/10.1038/ncomms16075

Wahl, T., Jain, S., Bender, J., Meyers, S. D., & Luther, M. E. (2015). Increasing risk of compound flooding from storm surge and rainfall for major US cities. *Nature Climate Change*, *5*(12), 1093–1097. https://doi.org/10.1038/nclimate2736

Wang, X., Jin, Y., Schmitt, S., & Olhofer, M. (2023). Recent advances in Bayesian optimization. *ACM Computing Surveys*, *55*(13s), 1–36. https://doi.org/10.1145/3582078

Xu, Y., Lin, K., Hu, C., Wang, S., Wu, Q., Zhang, L., & Ran, G. (2023). Deep transfer learning based on transformer for flood forecasting in data-sparse basins. *Journal of Hydrology*, *625*, 129956. https://doi.org/10.1016/j.jhydrol.2023.129956

Zhang, Y., Ragettli, S., Molnar, P., Fink, O., & Peleg, N. (2022). Generalization of an Encoder-Decoder LSTM model for flood prediction in ungauged catchments. *Journal of Hydrology*, *614*, 128577. https://doi.org/10.1016/j.jhydrol.2022.128577

Zhao, G., Pang, B., Xu, Z., Cui, L., Wang, J., Zuo, D., & Peng, D. (2021). Improving urban flood susceptibility mapping using transfer learning. *Journal of Hydrology*, *602*, 126777. https://doi.org/10.1016/j.jhydrol.2021.126777

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., et al. (2021). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, *109*(1), 43–76. https://doi.org/10.1109/JPROC.2020.3004555

Zrira, N., Kamal-Idrissi, A., Farssi, R., & Khan, H. A. (2024). Time series prediction of sea surface temperature based on BiLSTM model with attention mechanism. *Journal of Sea Research*, *198*, 102472. https://doi.org/10.1016/j.seares.2024.102472

Zscheischler, J., Martius, O., Westra, S., Bevacqua, E., Raymond, C., Horton, R. M., et al. (2020). A typology of compound weather and climate events. *Nature Reviews Earth and Environment*, *1*(7), 333–347. https://doi.org/10.1038/s43017-020-0060-z