Ecosystem Models in a Bayesian State Space Framework

John William Smith Jr.

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

Doctor of Philosophy
in
Statistics

Leah R. Johnson, Chair
R. Quinn Thomas, Co-chair
Christopher T. Franck
Eric P. Smith

May 10, 2022
Blacksburg, Virginia

Keywords: Bayesian analysis, data cloning, ecological forecasting, MCMC

Ecosystem Models in a Bayesian State Space Framework

John William Smith Jr.

(ABSTRACT)

Bayesian approaches are increasingly being used to embed mechanistic process models used into statistical state space frameworks for environmental prediction and forecasting applications. In this study, I focus on Bayesian State Space Models (SSMs) for modeling the temporal dynamics of carbon in terrestrial ecosystems. In Chapter 1, I provide an introduction to Ecological Forecasting, State Space Models, and the challenges of using State Space Models for Ecosystems. In Chapter 2, we provide a brief background on State Space Models and common methods of parameter estimation. In Chapter 3, we simulate data from an example model (DALECev) using driver data from the Talladega National Ecosystem Observatory Network (NEON) site and perform a simulation study to investigate its performance under varying frequencies of observation data. We show that as observation frequency decreases, the effective sample size of our precision estimates becomes too small to reliably make inference. We introduce a method of tuning the time resolution of the latent process, so that we can still use high-frequency flux data, and show that this helps to increase sampling efficiency of the precision parameters. Finally, we show that data cloning is a suitable method for assessing the identifiability of parameters in ecosystem models. In Chapter 4, we introduce a method for embedding positive process models into lognormal SSMs.

Our approach, based off of moment matching, allows practitioners to embed process models with arbitrary variance structures into lognormally distributed stochastic process and observation components of a state space model. We compare and contrast the interpretations of our lognormal models to two existing approaches, the Gompertz and Moran-Ricker SSMs. We use our method to create four state space models based off the Gompertz and Moran-Ricker process models, two with a density dependent variance structure for the process and observations and two with a constant variance structure for the process and observations. We design and conduct a simulation study to compare the forecast performance of our four models to their counterparts under model mis-specification. We find that when the observation precision is estimated, the Gompertz model and its density dependent moment matching counterpart have the best forecasting performance under model mis-specification when measured by the average Ignorance score (IGN) and Continuous Ranked Probability Score (CRPS), even performing better than the true generating model across thirty different synthetic datasets. When observation precisions were fixed, all models except for the Gompertz displayed a significant improvement in forecasting performance for IGN, CRPS, or both. Our method was then tested on data from the NOAA Dengue Forecasting Challenge, where we found that our novel constant variance lognormal models had the best performance measured by CRPS, and also had the best performance for both CRPS and IGN for one and two week forecast horizons. This shows the importance of having a flexible method to embed sensible dynamics, as constant variance lognormal SSMs are not frequently used but perform better than the density dependent models here. In Chapter 5, we apply our lognormal moment matching method to embed the DALEC2 ecosystem model into the process component of a state space model using NEON data from University of Notre Dame Environmental Research Center (UNDE). Two different fitting methods are considered for our difficult problem: the updated Iterated Filtering algorithm (IF2) and the Particle Marginal Metropolis Hastings (PMMH) algorithm. We find that the IF2 algorithm is a more efficient algorithm than PMMH for our problem. Our IF2 global search finds candidate parameter values in

thirty hours, while PMMH takes 82 hours and accepts only .12% of proposed samples. The parameter values obtained from our IF2 global search show good potential for out of sample prediction for Leaf Area Index and Net Ecosystem Exchange, although both have room for improvement in future work. Overall, the work done here helps to inform the application of state space models to ecological forecasting applications where data are not available for all stocks and transfers at the operational timestep for the ecosystem model, where large numbers of process parameters and long time series provide computational challenges, and where process uncertainty estimation is desired.

Ecosystem Models in a Bayesian State Space Framework

John William Smith Jr.

(GENERAL AUDIENCE ABSTRACT)

With ecosystem carbon uptake expected to play a large role in climate change projections, it is important that we make our forecasts as informed as possible and account for as many sources of variation as we can. In this dissertation, we examine a statistical modeling framework called the State Space Model (SSM), and apply it to models of terrestrial ecosystem carbon. The SSM helps to capture numerous sources of variability that can contribute to the overall predictability of a physical process. We discuss challenges of using this framework for ecosystem models, and provide solutions to a number of problems that may arise when using SSMs. We develop methodology for ensuring that these models mimic the well defined upper and lower bounds of the physical processes that we are interested in. We use both real and synthetic data to test that our methods perform as desired, and provide key insights about their performance.

Acknowledgments

In the opening chapter of his famous novel “The Great Gatsby”, F Scott Fitzgerald writes “just remember that all the people in this world haven’t had the advantages that you’ve had”. In this vain, I am forever indebted to my family – especially my parents and sister, John, Lisa, and Nicole, who put their faith in me when I left home ten years ago to pursue my studies. They have continued to be my number 1 fans, and have been a guiding light in a life of uncertainty. Thank you for everything. A big thank you to Chris Kappelmeier, my high school physics and calculus teacher, for motivating my interest in science and encouraging me to pursue what I love. This thesis would not have been possible without the support of my committee members. To my chair and co-chair, Leah and Quinn, I would like to thank you for all of the mentorship that you provided me. For me, graduate school was a place where I did a lot of self growth, and I am so appreciative of your patience with me as I found my footing. To name every person who has had an impact on me during graduate school would be impossible. I would like to thank Will, Austin, Pocket, Connor, Nick, Luke, and Thomas for always checking in to make sure that I hadn’t misplaced my sanity. To Nathan and Chris - thank you for allowing me to use BROCE, the computational workhouse behind a large portion of this dissertation, and for being the best roommates that I could have asked for. To KNN, TLC, and CIF, thank you for providing a place where I can take my mind off of things and laugh for hours upon end. To Kulyash, thank you for all of the lovely walks, endless book recommendations, and stimulating conversations. Last but certainly not least, I would like to thank my cat Nebs. The Achilles to my Tortoise, the Mrs Ramsay to my Lily Briscoe, and the Leopold to my Stephen. You are a constant reminder of the role that youth and curiosity play in our every day lives, and the most loyal companion that I have ever had. You will finally get to put that fluffy coat to good use in Montana.

Contents

Chapter 1

Introduction

In their seminal paper on ecological forecasting, Clark et al. [24] define ecological forecasts as "projections of the state of ecosystems, ecosystem services, and natural capital, with fully specified uncertainties". Since that publication, the importance of fully specifying different sources of uncertainty in ecological applications has been echoed by many others [see 88, 35, for two examples]. Thorough quantification of different components that contribute to the overall uncertainty has become a necessity to help understand the systems that we study, bridge knowledge gaps between our hypotheses about their behavior and what we observe in practice, and increase the confidence and trust in our predictions from stakeholders and policy makers [35].

To account for multiple sources of uncertainty, State Space Models (SSMs) are becoming increasingly popular in ecology, where they have been used in applications such as animal tracking [111], fisheries [98], disease ecology [62], population dynamics [33] and terrestrial ecosystem modeling [145]. SSM frameworks treat the underlying process as latent (unobserved), and use noisy observations to make inference about the latent process and its parameters. This allows SSMs to explicitly distinguish between the process stochasticity that is propagated through time and the measurement error in the observations that are available. Additionally, we can fit these models from a Bayesian perspective to provide robust estimates of parameter uncertainty alongside the estimates of process stochasticity, measurement error, and initial condition uncertainty that SSMs offer. The Bayesian paradigm offers a number of fitting methods for SSMs based on Markov Chain Monte Carlo [MCMC; 126],

including Gibbs sampling [22] and particle filter based MCMC [3, 42]. The use of Bayesian methods also allow us to carefully incorporate our knowledge about the system using prior distributions [9], and provides a natural setting for us to make probabilistic forecasts with fully specificed uncertainties.

In this work, I focus on state space modeling of ecosystems. Ecosystems models are typically fit using mechanistic process based models that describe interactions between different components of the ecosystem. Our study focuses on modeling the storage and transfer of carbon in terrestrial ecosystems using two process based models, the DALECev model [154, 43] and the DALEC2 model [12]. While statistical frameworks using process based models have been widely used to model terrestrial ecosystem carbon [see 143, 93, 117, 13, 43, 139, 154, 155, 70, for some examples], they are not frequently fit as SSMs. Incorporating these process based models into SSM frameworks come with their own set of challenges. First, they both have a large number of parameters that govern the process. The absence of proper data to constrain all of these parameters can lead to problems of identifiability [129]. Non-identifiable parameters can result in poor forecast performance, and requires practitioners to have access to methods to assess identifiability of their process parameters. Second, many ecosystem processes have well defined biological upper and lower bounds. A number of analytic fitting methods such as the Kalman filter [71], some Gibbs sampling approaches [22], and approximations to the Kalman filter [e.g. 69, 23] rely in-part on a normality assumption, but these well defined biological bounds are not usually captured by the Gaussian distribution. This requires us to seek alternative fitting methods, and can lead to trade-offs between ecological realism and computational complexity. The final challenge we discuss here is when measurements are available at resolutions that are much more coarse than the operational timestep of the process model. Combined with the fact that many process based ecosystem models run on a daily timestep, these infrequent measurements of the carbon stocks can make it difficult to find suitable fitting methods. To investigate the consequences of these challenges and find sensible ways to overcome them, this work uses a combination of simulation studies,

field data from the National Ecosystem Observatory Network (NEON; [28]), and two process based models: the DALECev model [154, 43] and the DALEC2 model [12]

The layout of the remainder of this dissertation is summarized as follows. In Chapter 2 I give background information on SSMs. In Chapter 3 I introduce methods for Gaussian error structures through the lens of ecosystem carbon models, using meteorological driver data from the Talladega National Forest NEON site. I discuss details of data cloning and how to change the process time resolution. In Chapter 4 I introduce methodology for embedding process and observation models into non-linear, non-Gaussian state space models using log-normal stochastic evolution and observation functions. In Chapter 5 I return to ecosystem carbon models, leveraging data available from NEON to create a carbon budget at University of Notre Dame Environmental Research Center using the lognormal SSMs from Chapter 4. In Chapter 6 I summarize the methods proposed here and provides additional discussion. In Chapter 7 (Appendix A) I provide additional mathematical details and derivations.

Chapter 2

Background

2.1 State Space Models

State Space Models (SSMs), sometimes referred to as Dynamic Linear Models (DLMs) or Hidden Markov Models (HMMs), are a broad class of models used to track the states of a system (some unobserved process $x_{1:T}$) through a set of observations $y_{1:T}$ [53]. An SSM has 3 components – the state component, the observation component, and additional parameters. The state component consists of an unobserved Markov process, $x_{1:T}$, where the state at time $t+1$ is generated by an "evolution function", $f(x_{t+1}|x_t, \Theta)$, that uses the previous state (x_t) and all of the parameters to output a value for the state process at the next time. The implication of $x_{1:T}$ following a Markov process is that the past and the future are independent conditional on the current timepoint x_t [137]. The observation component consists of noisy observations of the latent process, $y_{1:T}$, that are related to the latent states through an observation density function, $g(y_t|x_t, \Theta)$. These observations are assumed to be independent of one another conditional on the latent states $x_{1:T}$. The parameter component, Θ, contains the parameters that govern the evolution function and observation function. A visualization of an SSM can be seen in Figure 2.1, where a given latent state at time t is only related to the observation at time t and the latent states at time $t-1$ and $t+1$.

The flexibility of State Space Models is one of their greatest strengths, and they have been applied to an array of problems ranging from chemical production scheduling [144] to animal population dynamics [111, 115, 135] to American Football scores [47].

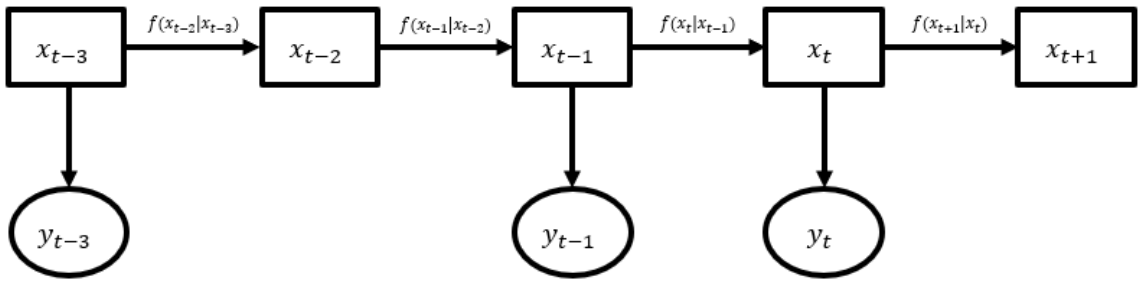

Figure 2.1: Visualization of a state space model, with boxes (x_t's) representing latent states and ovals (y_t's) representing observations. Latent states are moved through time by an evolution function, $f(\cdot)$. When observations are available, they provide information about the latent states.

.

2.2 Parameter Estimation in SSMs

When performing inference in applications involving SSMs, we are usually interested in two things: estimating parameters of the process and observation models, and reconstruction of the unobserved latent states. In the case of the Linear Gaussian State Space Model, Kalman Filtering techniques [71] and Gibbs sampling techniques [22] can be used to estimate latent states. These methods are preferred as they give exact samples from the target latent state distribution and are computationally efficient. These latent state estimation methods can be augmented with parameter estimation methods to simultaneously estimate both parameters and latent states. For example, the Expectation Maximization algorithm [EM; 30] can be used in Kalman filtering applications to estimate unknown parameters [see 58, for a thorough treatment], and Metropolis-Hastings steps [99, 54] can be used to estimate parameters in Bayesian frameworks that use Gibbs sampling to update the latent states.

Extensions of the Kalman filter have been used in an effort to relax assumptions of linearity, increase scalability, or increase robustness [e.g. 150, 63, 69, 41, 23] while still providing a framework for estimation of parameters and latent states. The Ensemble Kalman Filter (EnKF) of Evensen [41] is a recursive filter that uses empirical estimates of the covariance matrices between different ensemble members in place of the Kalman gain matrix. This

makes it better suited problems with large latent state dimensionality such as numerical weather prediction and other geophysical applications, where it has been previously used [148]. Arguably the most restrictive assumption of the Kalman filter is that the system is linear. Approaches such as the Extended Kalman filter [EKF; 69] and the Unscented Kalman filter [UKF; 150] have been applied in an attempt to capture non-linear dynamics while still using a Gaussian distribution. While these extensions provide approximate samples of the latent states and are computationally efficient, they are generally inflexible because they require the noise and support to be accurately captured by a Gaussian distribution.

For non-linear, non-Gaussian SSMs, analytic approaches like the Kalman filter are not available in general. Instead, many application rely on a set of techniques called Sequential Monte Carlo [SMC; 36]. Sequential Monte Carlo methods and particle filtering [3, 37, 21] provide flexible frameworks for estimating latent states and parameters. Some examples of these methods include maximum likelihood based approaches using iterated filtering [61, 62, 60] and Bayesian approaches using Markov Chain Monte Carlo [e.g. 3, 42, 138]. Particle filtering methods involve generating an ensemble of members (called particles) that evolve stochastically using a simulator for the forwards evolution and weighted using the observations. The weights of these particles are then used to compute an estimate of the marginal log likelihood [37]. The use of a forward simulator instead of an analytic Markov transition density, a property sometimes called "plug-and-play" [16, 55], allows particle filtering methods to be used for inference on problems without analytically tractable likelihoods. The trade off for this flexibility comes in the form of computational intensity associated with simulating Monte Carlo iterates of the particles.

While there are many existing methods for estimating latent states, these methods need to be tailored for the nuances of ecosystem models. Long data gaps between observations make latent state estimation difficult, and low data availability can make it infeasible to constrain all of the process parameters that govern the evolution through time. The methods detailed in this dissertation attempt to make these techniques accessible for ecosystem models.

Chapter 3

Methods for Gaussian Error Structures

3.1 Introduction

Many ecological prediction and forecasting applications use mechanistic process-based models to simulate the dynamics of ecosystems [95]. These process models are typically discretizations of ordinary or partial differential equations describing how the system dynamics evolve over time and space. Further they are often linearly conditioned on the value at the previous time step. These models are especially important in applications where available data limit the use of empirical models or predictions are being made for novel conditions not captured in existing data. However, process-based models can be challenging to calibrate due to the large numbers of parameters, and thus robust uncertainty estimation is also difficult [96].

While the importance of quantifying uncertainty in process models is recognized [35], it can be challenging to fully account for all important sources of uncertainty. Research to date has largely focused on estimating and reducing uncertainty that arises from initial conditions, parameters, and observational noise, usually through data integration techniques [see 65, 152, 10, for examples]. Estimating and propagating these sources of uncertainty without process uncertainty assumes that the process model perfectly describes the temporal evolution of the ecosystem up to error in the collection of observations. To estimate process uncertainty,

state space modeling frameworks [53, 40] are increasingly used to account for stochastic elements in the system evolution. Since many ecosystem process models already account for initial condition uncertainty, parameter uncertainty, and observational uncertainty, it is straightforward to convert them into a state space framework by adding an uncertainty structure to the underlying process model.

The Bayesian state space paradigm is a well-suited approach to estimate distributions of parameters and latent states (model states that are not directly observed) in ecosystem models using observations [6]. As a result it has seen a growing use in ecological forecasting applications [see 145, 39, for examples]. State space models treat all forecast terms as probability distributions and allow for more effective quantification, partitioning, and propagation of uncertainty in models. Prior distributions on parameters allow ecosystem scientists to enforce strict upper and lower bounds on parameters and incorporate biological information into the modeling process in a principled way. This focus on uncertainty and process precision estimation prompted us to choose a Bayesian framework over a point estimate focused method. The parameter and latent state posterior distributions in Bayesian state space models are often estimated using Markov Chain Monte Carlo (MCMC).

The added flexibility of the state space model does come with drawbacks. Analyzing ecosystem models as state space models increases the number of parameters that require estimation (i.e., parameters describing the distribution of the process uncertainty) and adds the requirement to estimate latent states for each ecosystem model state. The addition of latent state estimation can add anywhere from tens to hundreds of thousands of additional parameters to estimate because ecosystem models commonly use a daily time step that requires an additional parameter per state for each day of the simulation. Additionally, a parameter is required for describing the process variance for each of the states in the model. Finally, data for ecosystem models may only be available at timescales that are less frequent than the model timestep (i.e., annual or greater). These large gaps between observed data may present a challenge to constraining both the latent states and process variances.

Furthermore, identifiability (or equifinality [96]) is a common concern when estimating parameters in ecosystem process models, which often have many highly correlated parameters. Although problems with parameter identifiability do not necessarily impair latent state or process variance estimation, these parameters correspond to properties of ecosystems and often have important physical interpretation. Therefore, it is crucial to ensure that they can be successfully identified and estimated.

Data cloning (DC) has been used to assess identifiability of parameters in phylogenetic models [122] and for estimation in ecological models [84]. DC is done by applying Bayesian inference to a dataset that is constructed by duplicating the initial dataset r times. As r increases, the resulting posterior parameter estimates approach the maximum likelihood estimate. Data cloning can used to determine whether parameters are non-estimable or unidentifiable through a visual investigation of posterior plots with increasing values of r [122]. Parameters are said to be non-estimable when there exist different parameter values $\theta_1, \theta_2, \ldots, \theta_n$ such that $\mathcal{L}(\theta_1|X) = \mathcal{L}(\theta_2|X) = ... = \mathcal{L}(\theta_n|X) = \text{argmax}_\Theta \mathcal{L}(\theta|X)$, i.e. there are multiple sets of parameter values that maximize the likelihood function [85, 129, 122, 26]. Ponciano et al. [122] introduce terminology for various situations where we have parameters in the model that are not estimable: non-separability, lack of information, nonidentifiable, and identifiable but not estimable. We use Ponciano et al. [122]'s definitions for these terms throughout the remainder of this paper and so we introduce them here. Non-separability occurs when the model is structured such that it is not possible to separate parameters from one another, and may be due to parameter redundancy [see also 26]. Parameters that are non-estimable due to non-separability are referred to as nonidentifiable (NI). Lack of information occurs when the dataset does not contain sufficient information about the parameters to estimate them, resulting in wide posterior distributions that have not been properly constrained by observed data. Parameters that are nonestimable due to lack of information are referred to as INE (identifiable but not estimable). A combination of data cloning and effective sample size (ESS) estimation leads to a thorough analysis procedure

for ecosystem state space models that has not commonly been applied.

To address challenges estimating latent states and parameters when applying Bayesian state space modeling frameworks to ecosystem models, we present a simulation study using a forest ecosystem state space model that predicts carbon cycling among multiple states (a.k.a. "stocks" in the carbon cycling model). Using synthetic data with introduced data gaps, we address four questions focused on estimating precision parameters, latent states, and process parameters. 1) Under what temporal gaps in data does it become difficult to estimate process precisions? 2) Is it possible to recover true latent states under different temporal gaps in data and under scenarios for availability of data describing the transfers between states (i.e., fluxes)? 3) Can we increase the sampling efficiency for the precision of latent states in these cases by disconnecting the time step of the latent states from the time step of the model? and 4) Can we determine when model parameters are identifiable using data cloning? Our study is designed to help inform the application of ecosystem state space models to ecological forecasting applications where process uncertainty estimation is desired, and where data are not available for all stocks and transfers at the ecosystem model operational time-step (i.e., a daily time-step), such as data collected through the U.S. National Ecological Observatory Network (NEON).

3.2 Methods

3.2.1 Process Model

We used the Data Assimilation Linked Ecosystem Carbon model designed for simulating forests composed of evergreen trees [DALECev, 154]. It is a simple model describing carbon dynamics (Figure 3.1) and is similar to other ecosystem models used in carbon stock forecasting applications, for example PnET [1], 3PG [82], and TECOS [155]. The model can be written as a set of equations that are approximately linear and autoregressive in time.

While DALECev has been widely used [154, 139, 43, 12] it is not traditionally fit as a state space model as we do here.

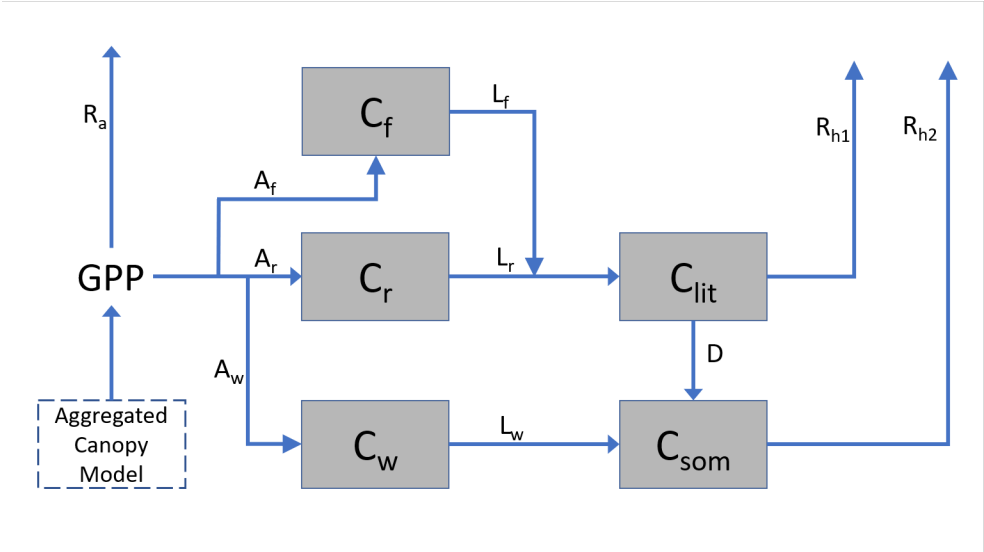

Figure 3.1: A schematic of the DALECev model with boxes denoting stocks of carbons and arrows denoting fluxes of carbon.

DALECev models the amount of carbon stored in five components within an evergreen forest ecosystem at a daily time step, t. These five components, called stocks, include: carbon stored in foliage, $C_f^{(t)}$; carbon stored in woody stems and coarse roots, $C_w^{(t)}$; carbon stored in fine roots, $C_r^{(t)}$; carbon stored in litter, $C_{lit}^{(t)}$; and carbon stored in soil organic matter $C_{som}^{(t)}$. The DALECev model includes 11 process parameters, p_i, for $i \in 1 \dots 11$, each representing the daily rate of an ecological process (e.g., turnover, decomposition, or soil organic matter mineralization), an allocation of a particular flux (transfer rate between stocks), or a parameter used in the calculation of a flux. Information on the physical interpretations of the process parameters, bounds, units, and values used during simulations is in Table 7.1.

Fluxes represent a number of physical processes that move carbon through the ecosystem, including respiration (R), photosynthetic allocation (A), turnover (L), and transfer to another stock (D). The model uses a submodel, the Aggregated Canopy Model (ACM) from [153], to simulate the input of carbon through gross photosynthetic production (GPP; G in Eqns. 5.14

Param.	Description	Lower	Upper	Units	Sim. Value
p_1	Scaled daily decomposition rate	1.1e-05	0.11	day^{-1}	.002
p_2	Fraction of GPP respired	0.2	0.7	unitless	.27
p_3	Fraction of NPP allocated to foliage	0.01	0.5	unitless	.15
p_4	Fraction of NPP after foliage allocation allocated to roots	0.01	0.5	unitless	.33
p_5	Daily turnover rate of foliage	1e-04	0.1	day^{-1}	.00137
p_6	Daily turnover rate of wood	1e-06	0.01	day^{-1}	1.1e-04
p_7	Daily turnover rate of roots	1e-06	0.01	day^{-1}	.00137
p_8	Scaled daily mineralisation rate of litter	1e-06	1	day^{-1}	.1
p_9	Daily mineralisation rate of soil organic matter	1e-06	0.01	day^{-1}	1.096e-05
p_{10}	Parameter for temperature dependent rate parameter	0.05	0.2	$^{\circ}\text{C}^{-1}$	.1725
p_{11}	Nitrogen use efficiency parameter	1	20	$(\text{gN gC}^{-1}\ \text{d}^{-1})^{-1}$	3

Table 3.1: Information on the physical interpretations, upper and lower bounds, units, and values used for simulation for each of the eleven process parameters in DALECev. Parameter descriptions are taken from the REFLEX project supplemental material [43]. Simulation values were chosen such that stock data, leaf area index (LAI) and NEE were reflective of what we expect at the simulation site.

– 5.15). Following Fox et al. [43], all parameters in the ACM submodel were fixed except for p_{11}. Thus, G is a function of p_{11} and meteorological driver inputs $\mathbf{D^{(t)}}$. Drivers include daily maximum and minimum temperatures, radiation, and atmospheric carbon dioxide. For the DALECev model, a given carbon stock C_s at time t can be generically expressed as the carbon at time $t-1$ minus turnover and respiration (carbon lost from the system or transferred to another stock), plus carbon gained by through the allocation of photosynthesis (carbon gained from outside the system) or transfer from another stock:

$$C_s^{(t)} = C_s^{(t-1)} - L_s^{(t-1)} - R_s^{(t-1)} + A_s^{(t-1)} \pm D_s^{(t-1)} + \epsilon_{t-1,s},\ \epsilon_{t,s} \sim N(0, \phi_s). \tag{3.1}$$

Here $\epsilon_{t-1,s}$ is process variation that allows normally-distributed stochastic deviations from the mean system behavior. Thus, the system of equations for stock evolution through time,

the deterministic skeleton are:

$$\begin{aligned}
\mathbb{E}[C_f^{(t)}] &= C_f^{(t-1)} - L_f^{(t-1)} + A_f^{(t-1)} \\
&= C_f^{(t-1)} - p_5 C_f^{(t-1)} + G(\mathbf{D^{(t)}}, p_{11})(1-p_2)p_3 \qquad (3.2) \\
\mathbb{E}[C_r^{(t)}] &= C_r^{(t-1)} - L_r^{(t-1)} + A_r^{(t-1)} \\
&= C_r^{(t-1)} - p_7 C_r^{(t-1)} + G(\mathbf{D^{(t)}}, p_{11})(1-p_2)(1-p_3)p_4 \qquad (3.3) \\
\mathbb{E}[C_w^{(t)}] &= C_w^{(t-1)} - L_w^{(t-1)} + A_w^{(t-1)} \\
&= C_w^{(t-1)} - p_6 C_w^{(t-1)} + G(\mathbf{D^{(t)}}, p_{11})(1-p_2)(1-p_3)(1-p_4) \qquad (3.4) \\
\mathbb{E}[C_{lit}^{(t)}] &= C_{lit}^{(t-1)} - R_{lit}^{(t-1)} - D_{lit}^{(t-1)} + A_{lit}^{(t-1)} \\
&= C_{lit}^{(t-1)} - \Big[\frac{p_1 p_8 \exp(p_{10}\bar{T}^{(t)})}{2} - \frac{p_1(1-p_8)\exp(p_{10}\bar{T}^{(t)})}{2}\Big] C_{lit}^{(t-1)} + \Big[p_5 C_f^{(t-1)} + p_7 C_r^{(t-1)}\Big] \qquad (3.5) \\
\mathbb{E}[C_{som}^{(t)}] &= C_{som}^{(t-1)} - R_{som}^{(t-1)} + D_{som}^{(t-1)} + A_{som}^{(t-1)} \\
&= C_{som}^{(t-1)} - \frac{p_9}{2}\exp(p_{10}\bar{T}^{(t)}) C_{som}^{(t-1)} + \frac{(p_1 - p_1 p_8)}{2}\exp(p_{10}\bar{T}^{(t)}) C_{lit}^{(t-1)} + p_6 C_w^{(t-1)}, \qquad (3.6)
\end{aligned}$$

where $\bar{T}^{(t)}$ is the average temperature for day t. These updates are referred to as the process model, $f(\cdot)$. For any carbon stock C the process model can be written in the form

$$\mathbb{E}[C^{(t)}|C^{(t-1)}] = f(C^{(t-1)}) = A_t C^{(t-1)} + b_t, \qquad (3.7)$$

where A_t, b_t are coefficients that can vary with time. Any stocks that cannot be written in this way can be approximately written in this form using a linearization.

The fluxes, the building blocks for Eqns. (5.14) – (5.18) in the DALECev model, are:

$$R_a^{(t)} = G(\mathbf{D^{(t)}}, p_{11})p_2 \tag{3.8}$$

$$A_f^{(t)} = G(\mathbf{D^{(t)}}, p_{11})(1-p_2)p_3 \tag{3.9}$$

$$A_r^{(t)} = G(\mathbf{D^{(t)}}, p_{11})(1-p_2)(1-p_3)p_4 \tag{3.10}$$

$$A_w^{(t)} = G(\mathbf{D^{(t)}}, p_{11})(1-p_2)(1-p_3)(1-p_4) \tag{3.11}$$

$$L_f^{(t)} = p_5 C_f^{(t-1)} \tag{3.12}$$

$$L_w^{(t)} = p_6 C_w^{(t-1)} \tag{3.13}$$

$$L_r^{(t)} = p_7 C_r^{(t-1)} \tag{3.14}$$

$$R_{lit}^{(t)} = .5\exp(p_{10}\bar{T}^{(t)})p_1 p_8 C_{lit}^{(t-1)} \tag{3.15}$$

$$R_{som}^{(t)} = .5\exp(p_{10}\bar{T}^{(t)})p_9 C_{som}^{(t-1)} \tag{3.16}$$

$$D_{lit}^{(t)} = .5\exp(p_{10}\bar{T}^{(t)})p_1(1-p_8) C_{lit}^{(t-1)} \tag{3.17}$$

These fluxes can be combined to form net fluxes. Since some of the individual fluxes, for example R_{lit}, are unlikely to be measured directly, net fluxes can provide us with information on these processes at the cost of having to isolate them from the net process. One important net flux is net ecosystem exchange (NEE) that is measured using eddy covariance techniques [7] and deployed in many ecological observation networks [100]. NEE is the net of G and Eqns. 5.1, 5.10, and 5.11 and is given by:

$$\begin{aligned} NEE^{(t)} =& R_{lit}^{(t)} + R_{som}^{(t)} + R_a^{(t)} - G(\mathbf{D^{(t)}}, p_{11})(1-p_2) \\ =& .5\exp(p_{10}\bar{T}^{(t)})p_1 p_8 C_{lit}^{(t)} + .5\exp(p_{10}\bar{T}^{(t)})p_9 C_{som}^{(t)} - G(\mathbf{D^{(t)}}, p_{11})(1-p_2) \end{aligned} \tag{3.18}$$

Additionally, soil respiration, S_r, is a net flux that is commonly measured in ecosystem studies. S_r is the net of autotrophic respiration by roots (a component of Equation 5.1) and

heterotrophic respiration by soil micro-organisms (Eqns. 5.10 and 5.11):

$$S_r^{(t)} = .5\exp(p_{10}\bar{T}^{(t)})p_1p_8C_{lit}^{(t)} + .5\exp(p_{10}\bar{T}^{(t)})p_9C_{som}^{(t)} - cG(\mathbf{D^{(t)}}, p_{11})p_2,\ c \in [0,1] \quad (3.19)$$

For this study we have fixed the value of c to be 0.3. In practice c must either be specified or given a very strong prior, as it can be challenging to constrain by other available data.

3.2.2 State Space Model

We estimate the stocks for the DALECev model using a state space model [53, 116, 40, 6]. In the state space framework, we treat the five carbon stocks as components of the state vector, and the additional flux data collected on respiration, photosynthetic allocation, turnover, and transfers as operations on the state vector. Let $\mathbf{C}$ denote the vector of carbon stocks from the model and $\mathbf{C}_{obs}$ denote the observations of the stock, with observations at a subset of time points $I \subset \{1, ..., T\}$. Then, Eqn. (5.14) - (5.18) can be written using matrix notation as:

$$\mathbb{E}[\mathbf{C}^{(t)}] = M_t\mathbf{C}^{(t-1)} + \mathbf{P}^{(t)},\ \text{where} \quad (3.20)$$

$$\mathbf{C}^{(t)} = \begin{bmatrix} C_f^{(t)} \\ C_r^{(t)} \\ C_w^{(t)} \\ C_{lit}^{(t)} \\ C_{som}^{(t)} \end{bmatrix}, M_t = \begin{bmatrix} (1-p_5) & 0 & 0 & 0 & 0 \\ 0 & (1-p_7) & 0 & 0 & 0 \\ 0 & 0 & (1-p_6) & 0 & 0 \\ p_5 & p_7 & 0 & (1-\frac{p_1}{2}Q^{(t)}) & 0 \\ 0 & 0 & 0 & \frac{p_1(1-p_8)}{2}Q^{(t)} & (1-\frac{p_9}{2}Q^{(t)}) \end{bmatrix}, \mathbf{P}^{(t)} = \begin{bmatrix} G(\mathbf{D^{(t)}}, p_{11})\psi_1 \\ G(\mathbf{D^{(t)}}, p_{11})\psi_2 \\ G(\mathbf{D^{(t)}}, p_{11})\psi_3 \\ 0 \\ 0 \end{bmatrix},$$

with $Q^{(t)} = \exp(p_{10}\bar{T}^{(t)})$, $\psi_1 = (1-p_2)p_3$, $\psi_2 = (1-p_2)(1-p_3)p_4$, $\psi_3 = (1-p_2)(1-p_3)(1-p_4)$.

To relate our observations for an arbitrary carbon stock $C_{s,obs}^{(t)}$ to the latent carbon stock

$C_s^{(t)}$, we assume the following relationship:

$$C_{s,obs}^{(t)} = C_s^{(t)} + \epsilon_{obs,t},\ t \in I,\ \epsilon_{obs,t} \sim \mathcal{N}(0, \tau_s)$$

In an ecological context, we are assuming that our observed carbon stock is normally distributed and unbiased [5], with a center at the true (latent) carbon stock, and a fixed precision τ_s. Similar to how adding the process variation term for the DALECev model was an acknowledgement of imperfect process knowledge, adding an error term for the observations is an acknowledgement of measurement error in the data that we observe.

The state space model has two key assumptions: the state process is first order Markov, and the observations is independent conditional on the latent states. Using normally distributed error terms for the process model and the observations, we can write these assumptions using the matrix notation above as:

$$\mathbf{C}^{(t)}|\mathbf{C}^{(t-1)} \sim \text{MVN}(M_t\mathbf{C}^{(t-1)} + \mathbf{P}^{(t)}, \Phi), t = 1, \ldots, T$$

$$\mathbf{C}_{obs}^{(t)}|\mathbf{C}^{(t)} \sim \text{MVN}(\mathbf{C}^{(t)}, \tau), t \in I,\ \text{where}$$

$$\Phi = \begin{bmatrix} 1/\phi_{Cf} & 0 & 0 & 0 & 0 \\ 0 & 1/\phi_{Cr} & 0 & 0 & 0 \\ 0 & 0 & 1/\phi_{Cw} & 0 & 0 \\ 0 & 0 & 0 & 1/\phi_{Clit} & 0 \\ 0 & 0 & 0 & 0 & 1/\phi_{Csom} \end{bmatrix}, \tau = \begin{bmatrix} 1/\tau_{Cf} & 0 & 0 & 0 & 0 \\ 0 & 1/\tau_{Cr} & 0 & 0 & 0 \\ 0 & 0 & 1/\tau_{Cw} & 0 & 0 \\ 0 & 0 & 0 & 1/\tau_{Clit} & 0 \\ 0 & 0 & 0 & 0 & 1/\tau_{Csom} \end{bmatrix}$$

with all τ parameters assumed to be known. This assumption is not uncommon in terrestrial carbon models, as the measurement error is generally well understood. Fixing the measurement error can also lead to better estimation of other precisions, process parameters, and states [5]. The combination of a linear process model, normally distributed process error, and normally distributed measurement error means that we are fitting DALECev as a Normal

Dynamic Linear Model (NDLM) [151].

The fluxes (eq'ns 5.1 - 5.12) are modelled with an observation model. For a given flux F_j, with flux data collected at a subset $I_j \subset \{1, ..., T\}$ and observation $F_{j,obs}$, we assume the relationship

$$F_{j,obs}^{(t)} | F_j^{(t)} \sim \mathcal{N}(F_j^{(t)}, \delta_j), \ t \in I_j, \tag{3.21}$$

where δ_j is a known precision. Fluxes are assumed to follow the functional forms given in Eqns. (5.1) - (5.12), for example $R_a^{(t)} | R_{a,obs}^{(t)} \sim \mathcal{N}(G(\mathbf{D^{(t)}}, p_{11})p_2, \delta_{R_a})$. This specification for the fluxes assumes that flux observations are unbiased but contain measurement error. Each flux has different observation time points I_j to account for the fact that fluxes are measured using different methods, and the methods may work on different timescales, as well as to give flexibility in the case of data collection failure.

With models assigned for our physical process, observations, and fluxes, we can write the complete data likelihood for parameters $\Phi, p_{1:11}$, and $\mathbf{C}^{1:T}$ as:

$$\begin{aligned} &\mathscr{L}(\mathbf{C}^{(1:T)}, \Phi, p_{1:11} | \mathbf{D}^{(1:T)}, \tau, \delta_{1:J}, \mathbf{C}_{obs}, \mathbf{F}_{obs}) \\ &\qquad \propto \prod_{t=1}^{T} \text{MVN}(M_t \mathbf{C}^{(t)} + \mathbf{P}^{(t)}, \Phi) \times \quad \prod_{t \in I} \text{MVN}(\mathbf{C}^{(t)}, \tau) \times \left(\prod_{j=1}^{J} \prod_{t \in I_j} \mathcal{N}(F_j^{(t)}, \delta_j) \right) \end{aligned} \tag{3.22}$$

Many prior studies using terrestrial carbon models include observational uncertainty, but does not include process variation [e.g. 65]. The state space approach used here is designed specifically to help isolate the process uncertainty from observational and parameter uncertainty. This partitioning of uncertainty is critical in understanding the system because no one source or type of uncertainty is likely to dominate total model uncertainty across all of ecological applications [35] and these uncertainties influence forecasts in different ways, e.g.,

process uncertainty propagates from one timestep to another while observation uncertainty does not. The Bayesian state space paradigm outlined here allows for quantification of multiple sources of uncertainty (process, initial condition, observation, and parameters) in the context of temporal gaps in observations, and the state space model gives a natural setting to leverage observed data with process based models.

3.2.3 Inference for parameters and latent states

We estimate the stocks, process precisions, and process parameters for the DALECev model using a Bayesian state space model [124, 45]. Parameters and latent states were estimated using MCMC [126]. MCMC is a flexible method that uses Markov chains to generate samples of the parameters from their posterior distribution. Parameter uncertainty is usually inherently included in MCMC methods, and the samples from the posterior can be used to calculate credible intervals for where the true values of the parameters lie. In addition to the likelihood (Eqn. 3.22), we need to specify prior distributions for process parameters, process precisions, and initial conditions for model states. We assume uniform priors for process parameters with limits informed by the range of values gathered from expert opinion in the Reflex project supplemental material [43] and adjusted to approximate a site in Talladega National Forest (see description below). The values for $p^{(L)}$ and $p^{(U)}$ can be found in Table 7.1. Each process precision was given a univariate conjugate Jeffreys prior [64], to allow for closed form Gibbs sampling of the process precision parameters. Thus the priors are given by

$$p_i \sim \text{Unif}(p_i^{(L)}, p_i^{(U)}), i = 1, \ldots, 11 \tag{3.23}$$

$$C_k^{(0)} \sim \mathcal{N}(\mu_k^{(0)}, \phi_k^{(0)}), k \in \{f, r, w, lit, som\} \tag{3.24}$$

$$\pi(\phi_k) \propto \frac{1}{\phi_k}, k \in \{f, r, w, lit, som\} \tag{3.25}$$

We can derive the full conditional distributions for all latent stocks and precision parameters from these likelihood and priors. The full condition distribution for latent carbon stocks at interior (between the initial and final) time steps with observed data are:

$$\pi(C_k^{(t)}|C_{-k}^{(t)}, C_{k,obs}^{(t)}, \cdot) \sim \mathcal{N}\Big(\frac{\phi_k(A_t C_k^{(t-1)} + b_t + A_{t+1}(C_k^{(t+1)} - b_{t+1})) + \tau_k C_{k,obs}^{(t)}}{\phi_k(1 + A_{t+1}^2) + \tau_k}, \phi_k(1 + A_{t+1}^2) + \tau_k\Big). \tag{3.26}$$

The full condition distributions for latent carbon stocks at interior time steps without observed data at those time points can be written as:

$$\pi(C_k^{(t)}|C_{-k}^{(t)}, \cdot) \sim \mathcal{N}\Big(\frac{\phi_k(A_t C_k^{(t-1)} + b_t + A_{t+1}(C_k^{(t+1)} - b_{t+1}))}{\phi_k(1 + A_{t+1}^2)}, \phi_k(1 + A_{t+1}^2)\Big). \tag{3.27}$$

The full conditional distributions for the initial latent state and final latent states are:

$$\pi(C_k^{(0)}|\cdot) \sim \mathcal{N}\left(\frac{\phi_k(A_1 C_k^{(1)} - A_1 b_1 + \phi_k^{(0)}\mu_k^{(0)})}{\phi_k A_1^2 + \phi_k^{(0)}}, \phi_k A_1^2 + \phi_k^{(0)}\right) \tag{3.28}$$

$$\pi(C_k^{(T)}|\cdot) \sim \mathcal{N}\left(\frac{\phi_k(A_T C_k^{(T-1)} + b_t) + \mathbb{1}_{T\in I}(\tau_k C_{k,obs}^{(T)})}{\phi_k + \mathbb{1}_{T\in I}(\tau_k)}, \phi_k + \mathbb{1}_{T\in I}(\tau_k)\right), \tag{3.29}$$

where $\mathbb{1}_{T\in I}$ is an indicator function that is 1 if there is an observation for C_k at the final time point, and 0 otherwise. Finally, the full conditional distributions for the precisions are:

$$\pi(\phi_k|\cdot) \sim \Gamma\left(\frac{T}{2}, \frac{1}{2}\sum_{t=1}^{T}(C_k^{(t)} - f(C_k^{(t-1)}))^2\right), \tag{3.30}$$

where Γ is the univariate gamma distribution using the rate parameterization.

We estimated the posterior distributions of the parameters using MCMC in the R programming language [123], specifically the Random Walk Metropolis Hastings (RWMH) ([99]; [54]) with a block updating algorithm. We sampled process parameters using a univariate proposal distribution during the initial burn-in phase (2,000 iterations). After burn in, we sampled

parameters for 500 iterations where we jointly sampled highly correlated process parameters using a truncated normal proposal that accounts for their covariance. We recalculated the empirical covariances used in the block updates every 500 iterations. We updated the analytic full conditional distributions given in Eqns. (3.26) - (3.30) for the latent states, process precisions, and initial conditions using a Gibbs sampler [46]. Process parameters were updated using the RWMH. Including burn-in, 10,000 total posterior samples were collected.

Initial latent state estimates were generated using piece-wise linear interpolation. More involved methods for latent state initialization were considered and tested, but our MCMC routine did not give evidence of being sensitive to the choice of initial latent state estimates. Thus the piecewise linear interpolation method was preferred for its simplicity.

3.2.4 Simulation study design

We designed a simulation study to evaluate the ability of standard MCMC methods to estimate process precisions, latent states, and process parameters for the DALECev model, and to identify and address potential problems that may arise when using these methods for the types of (sometimes sparse) data that are available. More specifically, we had three primary objectives. The first was to look at when we can expect precision estimates to have estimation difficulties, and whether changing the time resolution of the latent states can help to alleviate issues with estimation. The second was to examine how well the model recovers the latent carbon stocks with only annual observed data, which is the type of data that we can expect to collect to parameterize models like DALECev. Third we wanted to assess parameter identifiability (via data cloning) when fitting the models to different data that are available and use this information to help inform data collection schemes.

We began by generating a set of synthetic datasets for use in our analysis. Our simulation study was created to emulate conditions at the Talladega National Forest in Alabama, U.S.A (32.95046° N, -87.39327° W). We chose this site for two reasons. First, the site has a

canopy dominated by evergreen tree species (longleaf pine (*Pinus palustris*), loblolly pine (*Pinus taeda*), and slash pine (*Pinus elliotti*)) that matches the canopy type expected by the DALECev model. Second, the site is part of the National Ecological Observatory Network (NEON) and thus has ongoing data collection that can be used in future applications of the methods described here. For the synthetic data set, initial conditions and driver data for the carbon stocks were derived from NEON data [103], with specified initial mean and initial uncertainty. Process parameter values for simulations were chosen such that carbon stock data, leaf area index (LAI), and NEE were reflective of what would be expected at Talladega. Exact values for the process parameters used in the simulations can be found in Table 7.1. Random initial conditions to generate the simulations were drawn from their respective prior distributions. At each time step process noise is added to the states, with observational noise added to the latent states at the end of the model run to create a dataset of observations. Data gaps for synthetic datasets were created by removing observations that are not in the analysis time step.

3.2.5 Effects on estimates of precision and the latent states of varying the state process time resolution

The ability to estimate process precision is crucial in ecosystem models, as that is often the main source of uncertainty [35, 145]. Poor estimation of process precision may lead to more uncertainty in estimates of process parameters and of latent states, which can then affect forecasts and make them unreliable. For models like DALECev, gaps between observations of the states are commonly greater than 1 year, a much slower time scale than the assumed process dynamics, resulting in many unobserved states. In order to reliably apply DALECev in practice, process precision and latent state estimation should be robust to annual or longer data gaps.

To determine when the estimation of process precision becomes difficult, we examined three

different observation scenarios: daily state observations, monthly state observations, and yearly state observations, each with daily flux observations. Initial conditions were drawn from the prior distributions, and driver data from the site was used to run each model for two years. We repeated the generation of the synthetic dataset 15 times. For each dataset, observations were removed to introduce synthetic data gaps that matched the different observation scenarios.

We evaluate the performance of precision estimation using the effective sample size (ESS) [127] of the Markov chains for the precision parameters as a metric. Samples from Markov chains are auto-correlated, meaning that they are correlated with previous samples in the chain. The ESS gives us a metric for determining how efficiently our chain has estimated parameters/states, with low ESS informing us that our samples are highly correlated and to proceed cautiously with using the posterior samples for prediction or interpretation.

To evaluate ESS under different gaps in data, we used MCMC (as described above) to estimate posterior parameter distributions for each dataset and analyzed the process precision MCMC chains using ESS as a measure of sampling efficiency. We identified the data gaps where a large degradation in sampling efficiency occurred.

It also can be difficult to obtain information about latent states when there are large gaps between observed data points. We explored whether changing the latent state time resolution is a solution for alleviating problems with low effective sample sizes without introducing estimation problems for process parameters. These problems may arise from differences in the flux data and observation data likelihoods having different time steps. To analyze these differences, the model was run with a daily latent state resolution and was compared to the model being run with a monthly latent state resolution as follows. Consider an NDLM with daily process resolution for carbon stock C, with process model:

$$C^{(t)} = A_t C^{(t-1)} + b_t. \tag{3.31}$$

Let $T^* = \{t_i, \quad i = 1, .., I\}$ be a proper subset of the timesteps of the model. For an NDLM with a process model of the form in Equation (3.31), state transitions can be rewritten as:

$$C^{(t_i)}|C^{(t_{i-1})} \sim \mathcal{N}\left(\Big(\prod_{j=t_{i-1}+1}^{t_i} A_j\Big)C^{(t_{i-1})} + \sum_{k=t_{i-1}+1}^{t_i}\Big(\prod_{m=k}^{t_i} A_m\Big)b_m + b_{t_i}, \frac{\phi_s}{1 + \sum_{k=t_{i-1}+1}^{t_i}\Big(\prod_{m=k}^{t_i} A_m\Big)}\right). \tag{3.32}$$

This allows the stocks of the model to operate on a different time step than the fluxes, so that daily flux information can be used without requiring estimation of a large number of latent states that have very little data to constrain them. It also gives the model more flexibility, allowing models to change time steps for inference purposes, to decrease computational costs, and to allow for varying time steps across the stocks themselves. For values for A that are constant or similar through time, like we have here, and approximately uniformly spaced entries of T^*, we may treat the precision in Equation (3.32) as a fixed value $\phi_{monthly}$. Applying this approach, we follow the advice given in [6], who say "make simplifying assumptions when data are limited".

To examine how sampling efficiency is influenced by changing the time step at which latent states are estimated, we used two different synthetic datasets. The first synthetic dataset is a set of 30 synthetic time-series that have monthly carbon stock observations and daily flux observations. The second synthetic dataset is a set of 30 synthetic time-series that have annual carbon stock observations and daily flux observations. Each synthetic dataset was analyzed using both a daily time step and a monthly time step in the state process model for the two different observation time resolutions. We compared the effective sample sizes of process precisions between the different observation regimes to assess whether changing the time step of latent state estimation increases sampling efficiency of process precisions.

3.2.6 Impact of observation gaps on estimation of latent states

While efficient estimates of process precisions are important, the primary goal of our ecological state space model is to accurately track the evolution of the latent states through time. This involves both latent state estimation and estimation of process parameters. Here we are especially interested in annual carbon stock availability, as it is the most common case when working with actual data.

Latent state estimation was assessed by creating synthetic datasets with daily, monthly, and annual carbon stock data availability. Synthetic datasets were analyzed using both a daily model time step and the monthly model time step for the latent states. Posterior estimates of the latent states were generated using MCMC and compared to their true (synthetic) values. Coverage rates of latent state posterior distributions were assessed to confirm that they reach roughly the nominal rate by checking if the true latent state values were contained in the 95% highest posterior density intervals, and taking the average for each stock over all time steps. We also assessed convergence of process parameters using the Gelman Rubin Brooks test statistic [17] across four MCMC chains (with different starting conditions) for each parameter for each synthetic dataset.

3.2.7 Parameter identifiability under differing data availability

Identifiability of parameters was assessed using data cloning [84, 85] to analyze three synthetic datasets with annual carbon stock observations. Each dataset had different levels of flux data observations: 1) all fluxes observed; 2) only fluxes available from NEON with GPP data ($G, L_f, L_w, A_f, A_w, S_r$); 3) only fluxes available from NEON with NEE data ($NEE, L_f, L_w, A_f,$

A_w, S_r). These were chosen so that we could compare the ideal case to data that would be more commonly available for terrestrial carbon models. Our MCMC inference proceedure

was performed on each of the synthetic datasets with $r = 1, 5, 25$ data cloning replicates, and posterior distributions of p_1, p_3, p_{11} were analyzed across datasets and data cloning replicates.

Revisiting Eqns. (5.2) - (5.4), we see that A_r, A_f and A_w give additional information for $p_2, p_3,$ and p_4, parameters that are highly correlated due to their entanglement in the carbon update equations. The absence of one or more of these fluxes, like when using NEON data only, may make these parameters difficult to identify. For scientists, a data cloning analysis can serve more purposes than just assessing identifiability of parameters. Simulated data can be used a priori to determine what data is most important to collect in their experiments.

3.3 Results

3.3.1 Estimating process precision under data gaps

As would be expected, we found that increasing the time between observations leads to a noticeable decrease in MCMC sampling efficiency (Figure 3.2). The criteria we used to assess whether or not the posterior samples size was large enough for for inference was an ESS of 200 or higher. By this metric, both monthly and annual observation data give ESS's that are too low to make credible posterior inference. In particular, yearly observed data produces precision estimates with very low effective sample sizes. This is particularly concerning, as annual observations are the most realistic of the three data availability scenarios we explored. This finding motivates the need to alter the sampling methods in order to increase the sampling efficiency of these precision parameters.

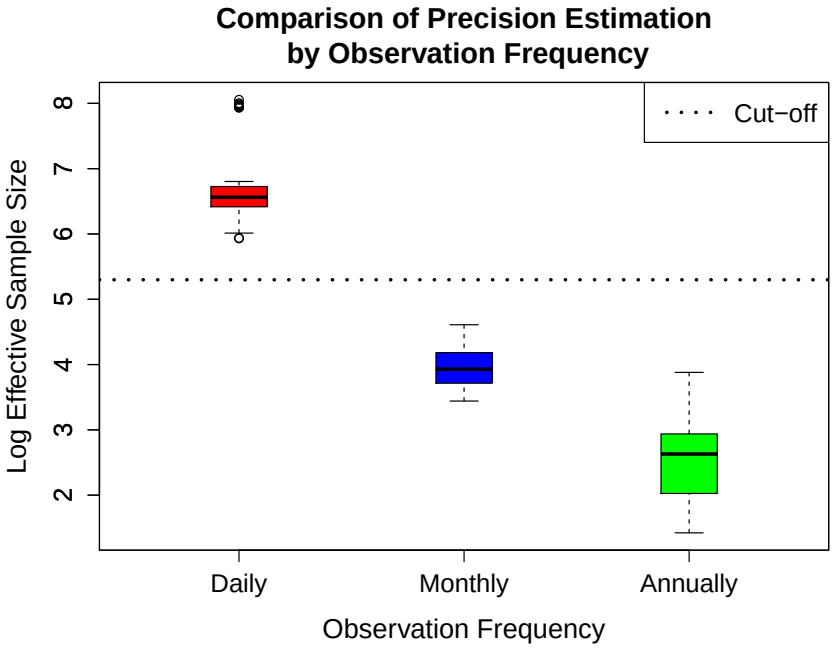

Figure 3.2: A plot of log average ESS of estimated precision parameters versus observation frequency. In each case we used the daily time step model and averaged over fifteen MCMC runs of 10,000 iterations of each model configuration. The five precisions, $\phi_{Cf}, \phi_{Cw}, \phi_{Cr}, \phi_{Clit}, \phi_{Csom}$ had their log ESS pooled. The dotted line indicates the cut-off of a log ESS of 200 (5.30).

3.3.2 Effects on estimates of precision and the latent states of varying the state process time resolution

The effective sample size for the process precision parameters decreased as the gap between observations increased. In Figure 3.3 we compare the effective sample size of process precisions, averaged over thirty simulations with different latent time resolutions. We found that changing the model timestep from daily to monthly improves the estimation efficiency of process precision for both monthly and annual data gaps, quantified by ESS. The monthly latent time resolution model performed substantially better than the daily latent time resolution model when observations were made monthly or annually.

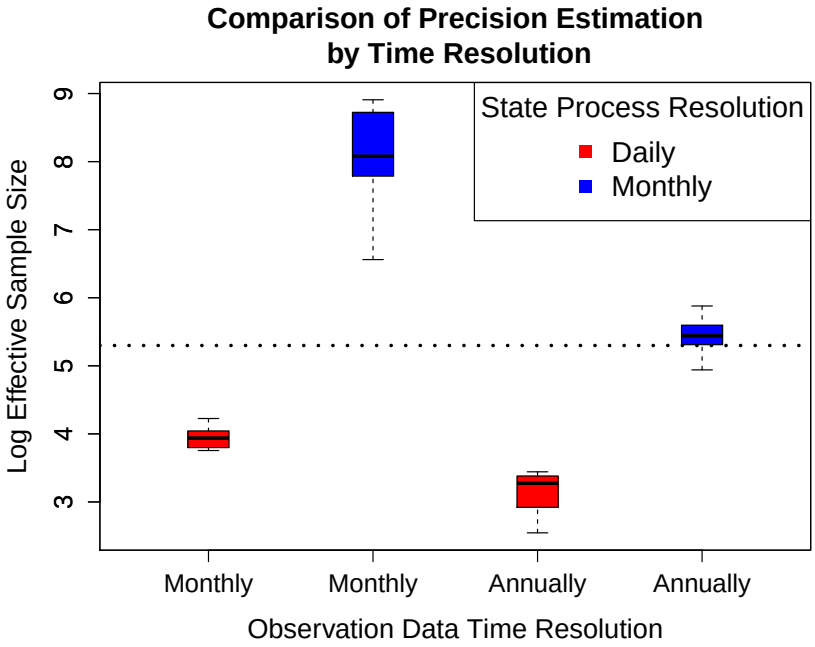

Figure 3.3: A plot of log average effective sample sizes for the estimated precision parameters (inverse variance) vs. observation data time resolution for different observation collection regimes using the daily step and the monthly step models. In each case we averaged over fifteen MCMC runs of each model configuration. The five precisions, $\phi_{Cf}, \phi_{Cw}, \phi_{Cr}, \phi_{Clit}, \phi_{Csom}$ The dotted line indicates the cut-off of a log ESS of 200 (5.30).

An important goal motivating the analysis of ecosystem process models using a state space framework is to track and predict the evolution of latent states through time. In Figure 3.4 we show posterior latent state estimates for carbon stock data observed annually, with all flux data observed daily and the model running on a monthly latent time resolution. Nearly all of the marginal posterior distributions of the process precisions in Figure 3.4 show that the process precisions are being under-estimated. This may come from a trade-off between model precisions and process parameters, where a particular process parameter being estimated may be larger than expected because it is includes some of the process uncertainty, leading to a precision parameter being under-estimated.

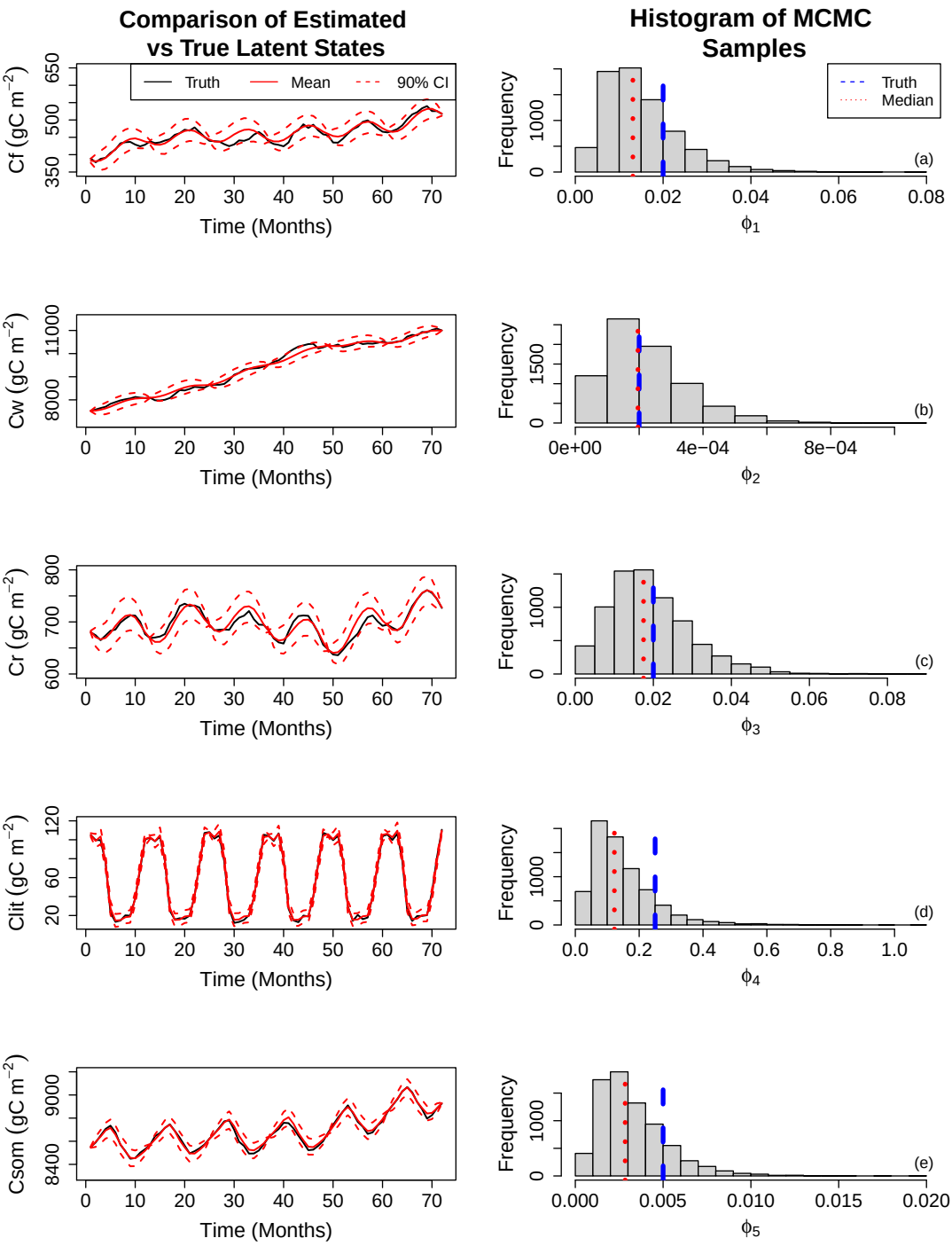

Figure 3.4: (a-e) Estimates for carbon density latent states for a monthly time resolution with annual data observations (Left) along with histograms of corresponding marginal estimated process precision post burn-in (Right). Panels corresponds to a particular carbon stock: C_f represents foliage carbon, C_w represents wood carbon, C_r represents root carbon, C_{lit} represents litter carbon, and C_{som} represents soil organic matter carbon. The posterior medians ($\mathcal{M}$), HPD intervals, and true (simulation) values are: $\mathcal{M}_{\phi_{Cf}} = .013$, $\phi_{Cf} \in (.00137, .0311)$, $\phi_{Cf} = .02$; $\mathcal{M}_{\phi_{Cw}} = 1.96\text{e-}04$, $\phi_{Cw} \in (2.42\text{e-}05, 4.65\text{e-}04)$, $\phi_{Cw} = 2\text{e-}04$; $\mathcal{M}_{\phi_{Cr}} = .0175$, $\phi_{Cr} \in (.00164, .039)$, $\phi_{Cr} = .02$; $\mathcal{M}_{\phi_{Clit}} = .122$, $\phi_{Clit} \in (.0176, .337)$, $\phi_{Clit} = .25$; $\mathcal{M}_{\phi_{Csom}} = .00283$, $\phi_{Csom} \in (3\text{e-}04, .00709)$, $\phi_{Csom} = .005$.

Stock variable	Avg. Coverage	Precision (%)
$\mathbf{C_f}$	.966	100
$\mathbf{C_w}$	.935	100
$\mathbf{C_r}$	.950	100
$\mathbf{C_{lit}}$	.990	92.9
$\mathbf{C_{som}}$	.987	100

Table 3.2: Average empirical coverage and percent of precisions contained in the 95% Highest Posterior Density (HPD) credible interval, averaged over 30 model runs

The coverage rate of posterior credible intervals for the latent states varied by latent state. In Table 3.2 we show the average performance of the monthly latent time resolution model over 30 different model runs. Coverage rate is used to quantify how well latent states are being tracked. For each latent variable, the average probability that the true process precision lies within the 95% credible interval is also given. C_f achieves above the nominal coverage rate, with C_w achieving below, C_r achieving exactly the nominal coverage rate, and C_{lit} and C_{som} achieving well above the nominal coverage rate. While it is difficult to discern why particular stocks are performing in this manner, the tendency of the stocks to perform well above the coverage rate may stem from the systematic underestimation of the precisions, leading to wider credible intervals on the latent processes.

3.3.3 Parameter identifiability under differing data availability

Our analysis of the data cloning results involves three primary considerations. First, for identifiable parameters, we expect that as r increases the variance of the resulting estimate decreases. This can be seen when the resulting posteriors grow tighter around the mean as values of r get larger. Second, identifiable but non-estimable parameters (INE) are parameters that may be identifiable, but do not have a necessary amount of data to estimate the precise values. These are characterized by relatively flat posterior distributions [122]. Third,

parameters that are non-identifiable (NI) tend to have multi-modal posterior distributions, with several values of the parameter that produce high values of the likelihood. Functions of multiple non identifiable parameters can be estimable, but the individual parameters themselves are not [for a simple example see 122].

We found that data cloning served as an effective way to assess identifiability of parameters. However, the results of our data cloning analysis demonstrate that NEON flux data with NEE will require additional flux observations in order to estimate four of the model parameters: p_2, p_3, p_4 and p_{11}. In Figure 4.3 (top row) we show that the posterior is bimodal as r (the number of data cloning replicates) increases for parameter p_2 for the NEE case – that is, it is non-identifiable with the observed flux data for the NEE case. For the GPP case, the posterior distribution for p_2 gets more narrow as r increases, indicating that it is identifiable. For parameter p_3, the posterior distributions for both cases narrow as r increases, indicating that p_3 is identifiable in both cases, with posterior estimates falling near the true simulated value for both NEE and GPP flux data. The parameter p_4 is similarly identifiable in both the NEE and GPP cases. However, in the NEE case, estimates for p_4 are not near the true simulation value, hovering close to the lower boundary, even for the highest r examined ($r = 25$). Similarly p_{11} was identifiable for both the NEE and GPP cases, but in the NEE case our estimates were not near the true simulation value. The poor estimates of p_2, p_4 and p_{11} may be related for the NEE case, as they appear together in Eqns. (5.3) and (5.4) and there is insufficient data to identify p_2 or estimate p_{11} well without GPP flux data available. Our findings illustrate one of the shortfalls of using data cloning: though we are able to determine whether parameters are estimable or identifiable, we cannot be sure our analysis is producing good estimates of the parameters.

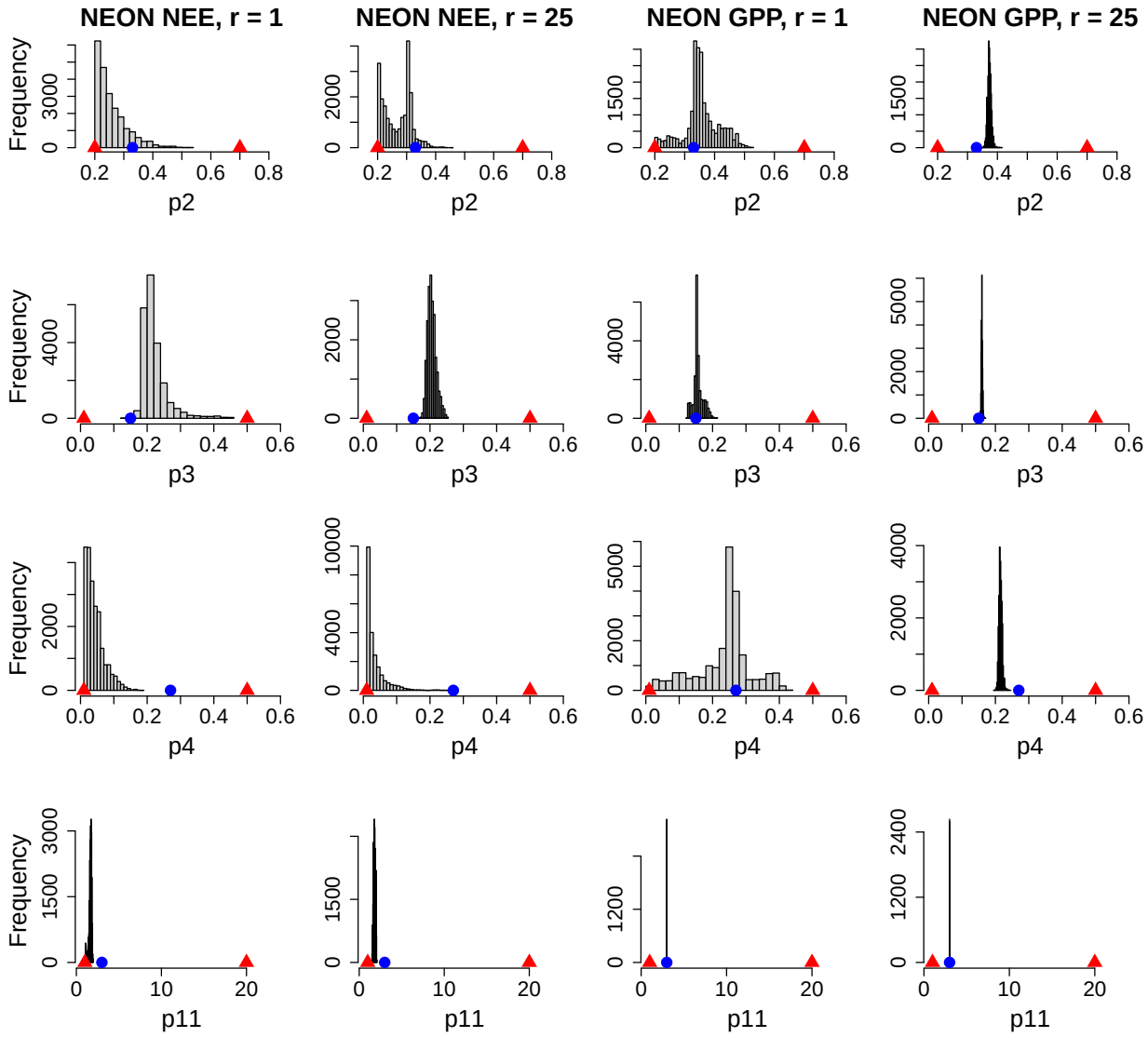

Figure 3.5: Marginal posterior distributions of data cloning for selected process parameters under the two NEON data flux scenarios, with 1 data cloning replicate ($r = 1$) and 25 data cloning replicates ($r = 25$). Blue circular marks on the x axis denote simulation (true) value of the parameter. Red triangular marks denote the upper and lower bounds of the uniform priors given to the process parameters.

3.4 Discussion

Estimating the posterior distributions of parameters in multi-state state space models can be challenging when observations of the states are not readily available. This is especially apparent in ecological models where states are often sampled at more coarse intervals than the model time-step, resulting in many latent states without direct constraint from corresponding observations. Here we introduce a method for changing the time resolution used for generating latent states in a process model so that it is coarser than the operational timestep of the process model (i.e., timestep of the difference equations). Our analysis revealed a large increase in the efficiency of estimation for the posterior distributions of the process precision parameters when using a coarser time-step of latent state estimation, while still achieving nominal coverage rates. One strength of this approach is that it preserves the operational time of the process model used to simulate the ecosystem dynamics. As a result, no adjustments were required to the process model. Another strength is that the equations used to change the time resolution of the latent states do not require the new time steps to be equally spaced, giving the flexibility to allow the latent states of the model to operate on any time scale. However, changes to the latent state time resolution do influence the interpretation of the process uncertainty parameters because they represent the distribution of process error that propagates from one latent state to the next – longer time-intervals between latent states will likely lead to process error distributions with larger variance.

Beyond data gaps in time, gaps in data where particular states and/or fluxes are never observed presents challenges in the ability to estimate the posterior distributions of parameters (identifiablility). To examine identifiablility in DALECev, we confirmed that we were able to successfully recover process parameters and process precisions when all states and fluxes were observed at all time steps and then in the case where there were annual temporal gaps in the observations of states. This indicates that under ideal data collection the parameters are identifiable using the approach presented here. However, a lack of identifiability oc-

curred when a subset of the flux data were not available to constrain model parameters, as is the case in applications using real observations. In this case, our approach had difficulty recovering multiple process model parameters in the DALECev model. In particular, p_2 was non-identifiable with NEE data, and p_3, and p_4 were difficult to estimate without all of the related fluxes used to constrain them. These parameters govern the proportional allocation of photosynthesis (GPP) to respiration, foliage, and roots (Eqns. 5.14 – 5.16), thus requiring observations of their individual production in order to constrain the individual parameters.

Our inference about identifiability of process parameters was based on the application of data cloning (DC). Other methods of assessing identifiability were considered, including Hessian methods [149, 89] and symbolic algebra methods [25, 27]. Here we consider long time series, which would lead to problems with numerical stability when using the Hessian method [20] and lack of computational resources to perform the symbolic algebra calculations in MAPLE [25]. Identifiability is a problem that has long plagued ecological and biological modeling [96], and DC is a simple method that can be used with simulated data prior to the design of an experiment to assist the design of data collection schemes that mitigate identifiability challenges, encouraging scientists to elicit data rather than eliciting priors. Our simulation study used DC with observed flux data that would be available from NEON, and showed that additional flux data are required to constrain a subset of model parameters. The types of data measured at a NEON site are not atypical for a terrestrial ecosystem study, particularly those in the Ameriflux and Fluxnet networks, therefore the results are not specific to a NEON site. Our analysis also illustrated that while some parameters are shown to be identifiable through data cloning, they are not necessarily unbiased. With simulation results showing that some parameters are not identifiable or identifiable but non-estimable, it is crucial that scientists have access to methods to help them assess whether they can trust the results they obtain from their modeling framework.

Data cloning has other uses aside from assessing identifiability and aiding in experimental data collection for ecosystem modelers. A well documented problem with soliciting prior

distributions for parameters in Bayesian analyses is that non-informative prior distributions on one scale may become highly informative prior distributions when transformed (see [86] and references within for a thorough treatment). These falsely non-informative priors can lead to biases in parameter estimation and prediction, in turn leading to incorrect decisions made by stakeholders and policy makers [86]. Data cloning methods can be used to expose accidental biases introduced through using these priors that are non-informative on one scale, as the data cloning posterior will approach the maximum likelihood solution as r increases, and maximum likelihood estimation is invariant to re-parameterizations.

Our study focused on the development and evaluation of methods, and sets the foundation for future work. First, while simulation studies with data synthesized from the DALEC model is necessary to test our methods, it is important to test the performance of these methods with real observations. Second, the results for latent state updates discussed here were for the univariate case where covariance between states is not considered, though the states can be updated *en bloc* with multivariate normal Gibbs updates. Multivariate latent state updates could be complemented with conjugate Gibbs updates for the covariance matrix, allowing full estimation of the covariance structure and (potentially) better latent state estimates. The methods presented here could potentially be extended to non-linear Markovian process models by linearizing the update equations. Third, the Gibbs updates shown in this paper are only applicable to state space models where both the observation and process model errors are normally distributed. While this is a common assumption in terrestrial carbon models [145], other applications may have error structures that do not meet this assumption. For example, error structures may be needed to maintain ecological realism, such as positivity of a particular latent state, and thus require non-normal error structures for values near zero. More complex error structures or model dynamics require alternative fitting methods, such as particle methods [see 72, for a thorough review] or Gaussian Process Regression [147]. Finally, we have shown that it may not be possible to identify or recover process parameters for the DALECev model under yearly data gaps when using only data available

from NEON. However, it is likely that integrating additional data not observed by NEON, such as satellite-derived leaf area index (e.g., MODIS LAI), and incorporating stronger priors that reflect general ecological principles will help to constrain model parameters further [12].

In conclusion, to address the growing popularity of state space modeling in ecological forecasting research [6], we propose methods that help to assess and fix problems with process precision estimation and identifiability of process parameters that frequently arise in ecosystem state space modeling when observations are scarce. The state space framework augmented with data cloning to assess identifiability of parameters presented here is flexible enough to be adapted for a broad range of problems including non normal-normal error structures, non-linear process models, and spatio-temporal models. The methods discussed here will allow practitioners to more effectively and efficiently address and overcome common suites of problems that arise when using state space models.

Chapter 4

Lognormal embeddings for positive process models

4.1 Introduction

Process based models are mathematical representations of the evolution of biological or physical systems [18]. These models are often comprised of systems of ordinary or partial differential equations in time and space, or are discretizations of such systems. Because they encapsulate known or hypothesized mechanisms of physical or biological systems, process based modeling approaches have advantages over empirical or phenomenological modeling approaches, particularly when low data availability limit the ability of empirical models to accurately represent complex processes or require extrapolation beyond conditions captured in training data. As a result, process based models remain common in ecological forecasting applications [87]. However process-based modeling approaches have their own sets of challenges, particularly the quantification of uncertainty in model dynamics and structure or over-parameterization [96].

Quantifying uncertainty in process based models is one of the most challenging tasks when using them for forecasting applications that involve uncertainty propagation. Uncertainty comes from a wide variety of sources, including but not limited to: process, measurement, initial conditions, driver data/boundary conditions, and estimated parameters [35]. Process uncertainty (or process stochasticity) is particularly important to address, as it acknowl-

edges that the modeling framework may contain unknown errors that are best represented stochastically or contain elements that are known to be non-deterministic and that affect the dynamics of the biological process. For this reason, the state space modeling (SSM) framework [53, 40, 116, 6] has been used increasingly in ecological applications [see 145, 39, 33, for examples]. State space models provide a flexible framework, being able to handle missing data and partition multiple sources of uncertainty [35]. Many existing fitting methods for ecosystem process based models already estimate initial condition and parameter uncertainty while accounting for observation uncertainty [see 70, 12, 155, for examples]. By adding stochastic elements to the process based model they can be analyzed as state space models.

One of the nuances of including uncertainty/stochasticity in the process model is the trade-off between simplicity and ecological realism. For example, in forest carbon modeling, Gaussian error with its positive to negative infinity bounds is commonly assumed for carbon stocks (pools) [141, 145, 65], despite the biophysical impossibility of an ecosystem having a negative amount of carbon. In general, biological processes may have well defined lower bounds (and potentially upper bounds) that are not accurately captured by error structures that have support over the entire real line. Consequently, models that have biophysically unrealistic error structures can produce nonsensical predictions as the states approach these well defined bounds, or if the model has large measurement error that can allow observations with negative values as the modeled states approach the lower bounds. This is especially relevant in forecasting applications, where the uncertainty compounds as the forecast horizon increases.

A wealth of probability distributions are available for modeling non-negative processes. The lognormal distribution has a rich history in ecology [32], and is frequently used in state space modeling of populations [e.g., 19, 33, 78] and species abundance [e.g, 98, 102]. Two common formulations for lognormal state space models (SSMs) are the stochastic Gompertz SSM [49] and the stochastic Moran-Ricker SSM [125, 33]. In their simplest forms, these models include substantial assumptions that are disguised by writing the models as Gaussian in log

space - namely the assumptions of density dependent variance and systematic bias in the process evolution and observation functions. When these assumptions are not feasible for the problem at hand, we need a mechanism to insert more sensible assumptions about the process evolution, observation, and variance dynamics. To address challenges with incorporating ecologically realistic error structures into state space models, we propose a novel lognormal moment matching approach, to allow users to specify the mean and variance of their process evolution and observation models.

We fit our lognormal moment matching models from a Bayesian perspective using Markov Chain Monte Carlo (MCMC). The Bayesian paradigm provides a flexible framework for fitting complex models, and the moment matching approach we introduce here offers a closed form Markov transition density, which facilitates the use of MCMC instead of alternative methods such as the particle filter [21, 37]. MCMC also provides a rigorous framework for quantifying parameter uncertainty and assessing forecast performance. MCMC methods generate samples from the posterior distribution, allowing practitioners to generate parameter estimates and build empirical density functions for their forecasts [80]. We can then validate our models by combining out of sample forecast observations with MCMC output and evaluating them with proper scoring rules [48], such as the Continuous Ranked Probability Score [CRPS; 97] and the Ignorance score [IGN; 51, 130].

We created four different models using the lognormal moment matching technique presented here. The four models were all based off of the Gompertz and Moran-Ricker models [49, 33, 125], and were embedded to have unbiased process evolution. We explored two different variance structures: a density dependent variance [33] for the evolution and observations, and a constant variance for the evolution and observations. First, we discuss interpretations of the Gompertz and Moran-Ricker SSMs and contrast them with interpretation of the models we present here. Next, we design and conduct simulation studies to compare forecast performance under model mis-specification, and assess estimability of precision parameters. Finally, we fit and evaluate models using time-series data on dengue fever hospitalizations

in San Juan, Puerto Rico from [66]. Our analysis of the transformed dengue hospitalization data was designed to assess the forecasting performance of the six models considered here, and to identify strengths, weaknesses, and similarities between all of the models fit.

4.2 Motivating Examples

In this section, we motivate the use of biologically realistic error structures in applications by demonstrating problems that may occur when modeling positive processes with error distributions that have mass over the entire real line. In particular we demonstrate how process models produce nonsensical predictions as they approach or exceed well defined biophysical bounds. First, we consider the following toy dynamical system in time:

$$X_t|X_{t-1} \sim \log\mathcal{N}\left(\mu^* = \log\left(\frac{X_{t-1}^2}{\sqrt{X_{t-1}^2+\sigma^2}}\right), \sigma^{2*} = \log\left(1+\frac{\sigma^2}{X_{t-1}^2}\right)\right). \tag{4.1}$$

Since the stochastic evolution function is lognormally distributed, the process is positive, i.e. $X_t \in (0,\infty), \forall t = 1,\ldots,T$. Further the conditional mean and variance are $\mathbb{E}[X_t|X_{t-1}] = X_{t-1}$ and $\mathbb{V}[X_t|X_{t-1}] = \sigma^2$.

Suppose that the positivity of the dynamical system is ignored and instead the process is modeled as simple a Gaussian random walk:

$$X_t|X_{t-1} \sim \mathcal{N}(\mu^* = X_{t-1}, \sigma^{2*} = \sigma^2). \tag{4.2}$$

When the system starts sufficiently far from zero the two models are indistinguishable from each other, producing nearly identical forecast distributions in terms of median estimates and credible intervals (Figure 4.3, top two panels). When the starting value X_0 is chosen to be close to the biological lower bound, the differences in forecasts become more apparent. The Lognormal model forecast intervals are now asymmetric and bounded below at zero,

and quickly reach zero as the forecast horizon increases. The random walk model forecast intervals retain their symmetry and continue to put considerable forecast mass below zero as the forecast horizon increases (Figure 4.3, bottom two panels).

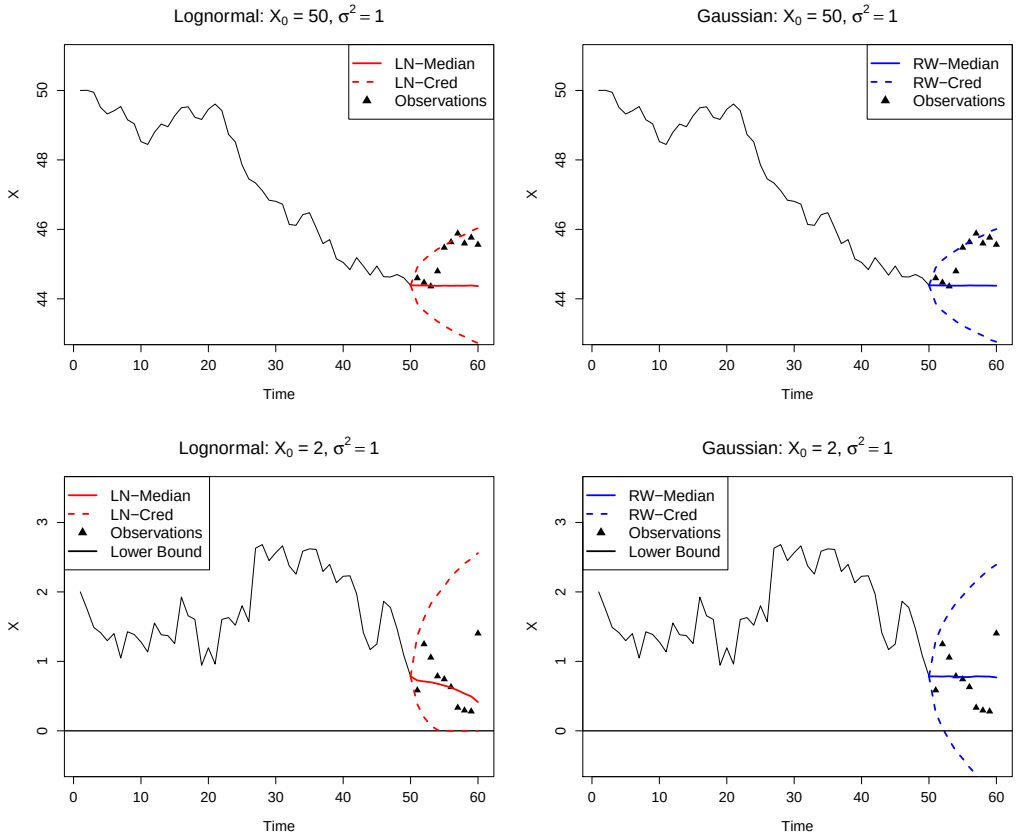

Figure 4.1: Two sample trajectories from the Lognormal dynamical system in Equation 4.1, the top panels start from $X_0 = 50$ and the bottom panels start from $X_0 = 2$. 95 % credible intervals for 10 day forecast horizons (dotted lines) are created using the Lognormal equations and the Gaussian equations to generate 10,000 sample prediction trajectories. In the bottom panels, the Lognormal forecasts (bottom left) are bounded below by zero, the biological lower bound. In contrast, the Gaussian random walk forecasts (bottom right) put considerable forecast probability mass below the biological bound.

In this simple example, the negative forecasts are not causing any issues in the process model, thus only producing unrealistic forecasts. To illustrate an example where issues arise in the process model, consider the following dynamical system:

$$X_t|X_{t-1} \sim \log\mathcal{N}\left(\mu^* = \log(|X_{t-1}|) - a, \sigma^{2*} = \log\left(1 + \frac{\sigma^2}{|X_{t-1}|}\right)^2\right). \tag{4.3}$$

We note that, formally, absolute values are unnecessary for the evolution of the true model (since X_t is positive for all t). However they will become necessary when we attempt to find an analogous Gaussian model with which to forecast.

Computing the conditional mean and variance of the lognormal distribution with the given values of μ^* and σ^{2*} and using them as the conditional mean and variance of a Gaussian distribution results in:

$$X_t|X_{t-1} \sim \mathcal{N}(\exp\left(\mu^* + \frac{\sigma^{2*}}{2}\right), (\exp(\sigma^{2*}) - 1)\exp(2\mu^* + \sigma^{2*})). \tag{4.4}$$

For values sufficiently far from zero, the forecasts from the Lognormal model are nearly identical to the forecasts from the Gaussian model (Figure 4.2, top two panels). The Lognormal forecasts (bottom left panel, Figure 2) enforce the lower bound, and produce forecast distributions with heavy upper tails. However, the Gaussian forecasts when the system approaches zero (bottom right panel, Figure 2) produce forecast distributions with large upper credible bounds that march away from zero, and lower credible bounds that first increase and then move back towards zero. In working on specific applications, predictions like these may be challenging to analyze because it is easy to associate the dynamics with an inadequate process model, when in reality the problem lies in having an error structure that can put probability mass on regions where it should be biologically impossible.

4.3 Methods

4.3.1 Lognormal SSMs

The Gompertz [49] and Moran-Ricker [125] SSMs are simple discrete density-dependent state space models with lognormally distributed process and measurement error [33]. The Gompertz and Moran-Ricker SSMs are popular choices for lognormal state space models

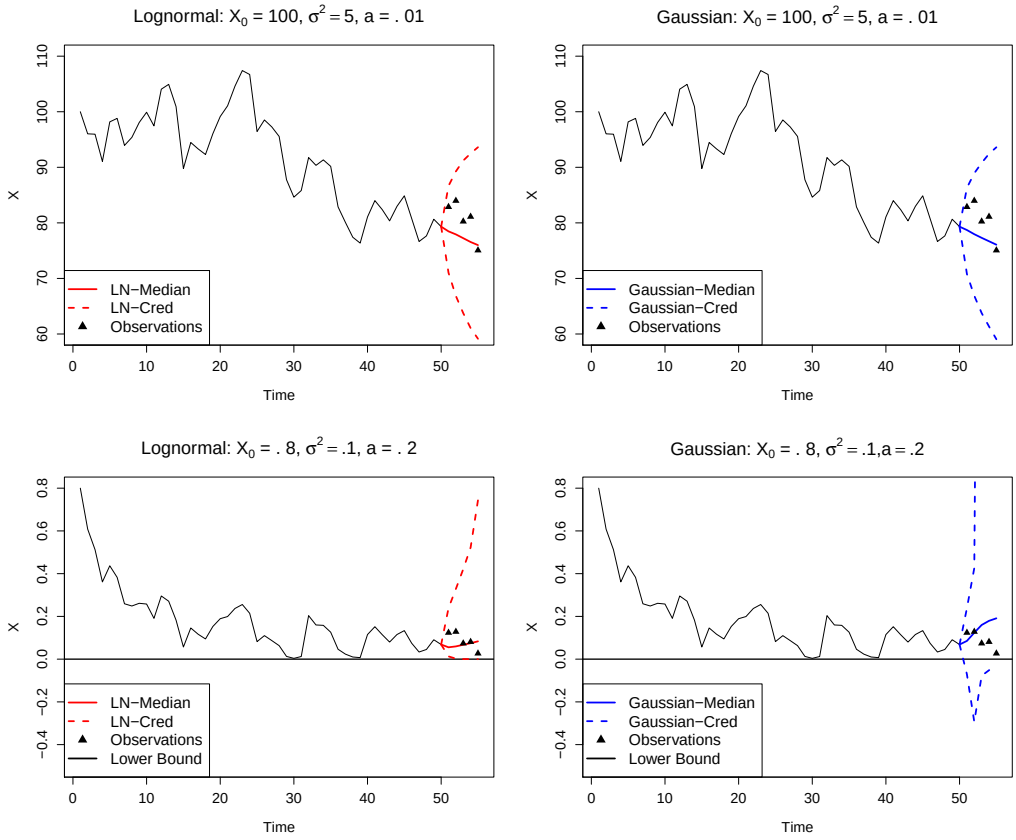

Figure 4.2: Two sample trajectories from the Lognormal dynamical system in Equation 4.3, the top panels starting from $X_0 = 50$ and the bottom panels starting from $X_0 = 2$. 95% credible intervals for 10 day forecast horizons (dotted lines) are created using the Lognormal equations and the Gaussian equations to generate 10,000 sample prediction trajectories. In the bottom panels, the Lognormal forecasts (bottom left) are bounded below by zero, the biological lower bound. In contrast, the Gaussian random walk forecasts (bottom right) put considerable forecast probability mass below the biological bound.

because they can be easily transformed into Normal Dynamic Models (NDMs), and can then take advantage of a suite of well studied fitting methods including Kalman filtering [71], extended Kalman filtering [69], and Gibbs sampling [46, 22, 141], easing computational difficulties associated with fitting SSMs. The latent process models for the Gompertz model can be written as:

$$X_t = X_{t-1} \exp(a + b \log(X_{t-1}) + \epsilon_t), \;\; \epsilon_t \sim \mathcal{N}(0, \phi) \tag{4.5}$$

The latent process model for the Moran-Ricker model can be written as:

$$X_t = X_{t-1}\exp(a + bX_{t-1} + \epsilon_t),\ \epsilon_t \sim \mathcal{N}(0, \phi) \tag{4.6}$$

Letting $A = \exp(a)$, Equations (4.5) and (4.6), the Gompertz and Moran-Ricker process equations, can be rewritten in the form

$$X_t = A(X_{t-1})^{b+1}\exp(\epsilon_t),\ \epsilon_t \sim \mathcal{N}(0, \phi) \tag{4.7}$$

$$X_t = AX_{t-1}\exp(bX_{t-1})\exp(\epsilon_t),\ \epsilon_t \sim \mathcal{N}(0, \phi) \tag{4.8}$$

More generally, we can think of both the Gompertz and Moran-Ricker process models as belonging to a class of lognormally distributed process models that have the form:

$$X_t = f^*(X_{t-1}|\Theta)\exp(\epsilon_t),\ \epsilon_t \sim \mathcal{N}(0, \phi),\ f^*(X_{t-1}|\Theta) > 0, \tag{4.9}$$

where $f^*(X_{t-1}|\theta)$ is the model of choice for non-negative physical process, and the process error is governed by a multiplicative zero mean lognormal distribution. Key statistical properties (conditional mean, conditional variance, and conditional median) of this class of process models are:

$$\begin{aligned}
\mathbb{E}[X_t|X_{t-1}] &= f^*(X_{t-1}|\Theta)\exp((2\phi)^{-1}),\\
\mathbb{V}[X_t|X_{t-1}] &= f^*(X_{t-1}|\Theta)^2\exp(\phi^{-1}),\left(\exp(\phi^{-1}) - 1\right)\\
\mathscr{M}[X_t|X_{t-1}] &= f^*(X_{t-1}|\Theta).
\end{aligned}$$

When looking at the conditional expected value, we see that this class of models is unbiased in terms of median process evolution, but biased in terms of mean process evolution since $\exp((2\phi)^{-1}) > 1$ for any $\phi > 0$. For this class of process models, the variance of the latent state is controlled by the value of the process model $f^*(\cdot)$, with larger values of the

process assumed to have larger variation than smaller ones. Therefore, in the class of process models that describe the the Gompertz and Moran-Ricker models, the density dependent relationship of the latent process variance is assumed to scale with the predicted value of the process model.

The generalization of the Gompertz and Moran-Ricker process models also hold true for the observation model. For example, in Dennis et al. [33], the assumed relationship between an arbitrary observation Y_t and the corresponding latent state X_t for the Gompertz model is given by:

$$Y_t = X_t \exp(\epsilon_{obs,t}),\ \epsilon_{obs,t} \sim \mathcal{N}(0, \tau) \tag{4.10}$$

Broadly, the Gompertz and Moran-Ricker observation models belong to the same class of lognormally distributed observation models in Equation 4.10 that have the form:

$$Y_t = g^*(X_t|\Theta) \exp(\epsilon_{obs,t}),\ \epsilon_{obs,t} \sim \mathcal{N}(0, \tau),\ g^*(X_t|\Theta) > 0, \tag{4.11}$$

where $g^*(X_t|\Theta)$ is now interpreted as the model used to link our observations to their respective latent states, and the measurement error is dictated by a multiplicative zero mean lognormal distribution. The conditional mean, conditional variance, and conditional median of this class of observation models are given by:

$$\begin{aligned}
\mathbb{E}[Y_t|X_t] &= g^*(X_t|\Theta) \exp((2\tau)^{-1}), \\
\mathbb{V}[Y_t|X_t] &= g^*(X_t|\Theta)^2 \exp(\tau^{-1}) \Big(\exp(\tau^{-1}) - 1 \Big), \\
\mathscr{M}[Y_t|X_t] &= g^*(X_t|\Theta).
\end{aligned}$$

Importantly, while this class of observation models assumes that observations are unbiased in log space, they assume there is systematic observation bias in their measurement process, which is given by $B(Y_t|X_t) = g^*(X_t|\Theta)(1 - \exp((2\tau)^{-1}))$.

Though we have only examined the Gompertz and Moran-Ricker models when discussing the class of lognormal process and observation models presented here, there are many examples of lognormal modeling frameworks in the animal population modeling literature [see 19, 78, for examples] and fisheries modeling literature [see 98, 102, for examples] that fall into the classes we discuss here. Including a term to correct the bias is a common method used in applications to rectify assumptions of biased process evolution and biased observations [78, 98, 102] but this bias correction term changes the variance structure, as the lognormal variance is a function of both μ and σ^2.

4.3.2 Lognormal Moment Matching Models

Instead of assuming an unbiased median process evolution and a density dependent variance structure, a modeler developing a lognormal SSM for their application may want to specify the mean evolution and the variance structure of the stochastic lognormal process model. That is, we desire a lognormally distributed stochastic evolution function such that $\mathbb{E}[X_t|X_{t-1}] = f^*(X_{t-1}|\Theta)$ and $\mathbb{V}[X_t|X_{t-1}] = \phi_t^{-1}$. We can create a *process* model with these characteristics by using a moment matching transformation on the mean and precision, specifically:

$$\mu_t^* = \log\left(\frac{f^*(X_{t-1}|\Theta)^2}{\sqrt{f^*(X_{t-1}|\Theta)^2 + \phi_t^{-1}}}\right), \tag{4.12}$$

$$\phi_t^* = \log\left(1 + (f^*(X_{t-1}|\Theta)^2\phi_t)^{-1}\right)^{-1}. \tag{4.13}$$

This approach provides a stochastic lognormal process model with the desired properties (see Appendix A for derivation):

$$\begin{aligned} X_t|X_{t-1} &\sim \text{Lognormal}(\mu_t^*, \phi_t^*), \\ \mathbb{E}[X_t|X_{t-1}] &= f^*(X_{t-1}|\Theta), \\ \mathbb{V}[X_t|X_{t-1}] &= \phi_t^{-1}. \end{aligned}$$

Thus, there is a framework for embedding the process model $f^*(X_{t-1}|\Theta)$ such that the temporal evolution is unbiased, and that allows for a flexible way to model the variance of the process through time. For example, we can choose $\phi_t = \phi$ to model constant variance through time, $\phi_t = \phi / f^*(X_{t-1}|\Theta)^2$ to model density dependent variance, or $\phi_t = \phi \exp(-t^{-1})$ to model variance that dissipates over time.

The lognormal moment matching approach can be similarly applied to the *observation* model to produce a lognormally distributed observation density with the properties $\mathbb{E}[Y_t|X_t] = g^*(X_t|\Theta)$, $\mathbb{V}[Y_t|X_t] = \tau_t$ by using the transformation:

$$\mu_{t,obs}^* = \log\Big(\frac{g^*(X_t|\Theta)^2}{\sqrt{g^*(X_t|\Theta)^2 + \tau_t^{-1}}}\Big), \tag{4.14}$$

$$\tau_t^* = \log\Big(1 + (g^*(X_t|\Theta)^2\tau_t)^{-1}\Big)^{-1}. \tag{4.15}$$

With the forms of both the process model and observation model fully specified in this, it is possible to write a likelihood equation for a model that includes both components, that we call the Lognormal Moment Matching model (LNM3). Suppose that we have observations

$\mathbf{Y}$ at a subset of time points $I \subset \{1, \ldots, T\}$. Then the likelihood for the LNM3 is given by

$$\begin{aligned}\mathscr{L}(X_{1:T}, \Theta|\mathbf{Y}) = &\prod_{t=1}^{T} \log\mathcal{N}\left(\log\left(\frac{f^*(X_{t-1}|\Theta)^2}{\sqrt{f^*(X_{t-1}|\Theta)^2 + \phi_t^{-1}}}\right), \log\left(1 + (f^*(X_{t-1}|\Theta)^2\phi_t)^{-1}\right)^{-1}\right) \\ &\times \prod_{i \in I} \log\mathcal{N}\left(\log\left(\frac{g^*(X_i|\Theta)^2}{\sqrt{g^*(X_i|\Theta)^2 + \tau_i^{-1}}}\right), \log\left(1 + (g^*(X_i|\Theta)^2\tau_i)^{-1}\right)^{-1}\right)\end{aligned} \tag{4.16}$$

The moment matching method can be used to embed any positive process or observation model and any variance structure into a stochastic lognormal model, and thus helps to maintain biophysical realism when modeling positive processes. Fitting these models as state space models further allows them to easily handle missing data and partition between process and measurement error. Thus the LNM3 approach provides a framework that is flexible, biophysically realistic, and statistically coherent.

4.3.3 Model Fitting

Our analysis focused on comparing and contrasting the LNM3 approach to the Gompertz and Moran-Ricker approaches, two common lognormal SSMs. We estimated the latent states, precisions, and model parameters for six different models presented here using Bayesian state space models (SSMs) [53, 40, 116]. The six different models include the Gompertz SSM, the Moran-Ricker SSM, and four different SSM formulations using the Lognormal Moment Matching (LNM3) method presented in Equations 4.12 – 4.15. Model parameters, latent states, and precisions were estimated using Markov Chain Monte Carlo (MCMC) [126]. MCMC is an estimation method that generates samples of parameters from their posterior distributions using Markov chains. Parameter uncertainty can be quantified using posterior samples from these Markov chains. All models were fit using `JAGS` [119] using the `rjags` package [120] in `R` version 4-10 [123].

Gompertz SSM Fitting

We fit the Gompertz model in log space, by taking the logarithm of the latent states ($X_{1:T}$) and observations ($Y_{t\in I}$), $D_t = \log(X_t)$ $F_i = \log(Y_i)$. Under this transformation, the process model and observation model can be written as a Normal Dynamic Linear Model (NDLM) [151]:

$$D_t = a + (1+b)D_{t-1} + \epsilon_t,\ \epsilon_t \sim \mathcal{N}(0,\phi) \tag{4.17}$$

$$Y_i = D_i + \epsilon_{obs,i},\ \epsilon_{obs,i} \sim \mathcal{N}(0,\tau) \tag{4.18}$$

If the log observations ($\mathbf{F}$) are available at a subset of time points, $I \subset \{1,\ldots,T\}$, the likelihood for the Gompertz model can be written as:

$$\begin{aligned}\mathscr{L}(D_{1:T}, a, b, \tau, \phi|\mathbf{F}) \propto &\prod_{t=1}^{T} \sqrt{\phi}\exp\left(-\frac{\phi}{2}(D_t - a - (1+b)D_{t-1})^2\right) \\ &\times \prod_{i\in I}\sqrt{\tau}\exp\left(-\frac{\tau}{2}(F_i - D_i)^2\right).\end{aligned} \tag{4.19}$$

Based on the likelihood for the Gompertz SSM, we need prior distributions for a, b, ϕ, τ, and the initial latent state X_0. Given the interpretation of a and b as the multiplicative constant and the growth rate in the Gompertz model (Equation 4.7), we used Uniform$(-10, 10)$ priors for both a and b. There are important considerations when choosing the prior distributions for ϕ and τ. Though the likelihood in Equation 4.19 can exploit conjugacy and use improper Jeffreys priors [64] for both ϕ and τ, Gelman [44] show that the posterior is sensitive to the choice of ϵ when using a Gamma(ϵ, ϵ) prior in JAGS to emulate the improper Jeffreys prior. Following the advice of Gelman [44] and Polson and Scott [121], we use a central half Cauchy prior distribution for the precision parameters. We note that though these citations recommend central half Cauchy priors on variance parameters, under this choice of prior the inverse variance (precision) is also implied to have a central half Cauchy prior (see Appendix A for proof). We chose the initial condition prior, $\pi(D_0)$, to be normally distributed, with

an initial mean of μ_0, and an initial precision of ϕ_0. Thus the priors are given by:

$$a \sim \text{Uniform}(-10, 10), \tag{4.20}$$

$$b \sim \text{Uniform}(-10, 10), \tag{4.21}$$

$$\phi \sim \text{HalfCauchy}(\gamma = 100), \tag{4.22}$$

$$\tau \sim \text{HalfCauchy}(\gamma = 100), \tag{4.23}$$

$$D_0 \sim \mathcal{N}(\mu_0, \phi_0). \tag{4.24}$$

The full conditional distributions for $D_{1:T}$ in the Gompertz model are analytically tractable, allowing for Gibbs sampling [46]. For interior latent states ($k = 1, 2, \ldots, T-1$) the Gibbs updates are given by:

$$\pi(D_k|\cdot) \sim N\left(\frac{\phi((1+b)D_{k-1} + a + (1+b)(D_{k+1} - a)) + \tau F_k \mathbb{1}_{k\in I}}{\phi(1 + (1+b)^2) + \tau \mathbb{1}_{k\in I}}, \phi(1 + (1+b)^2) + \tau \mathbb{1}_{k\in I}\right). \tag{4.25}$$

The Gibbs updates for the initial latent state and the final latent state are given by

$$\pi(D_T|\cdot) \sim N\left(\frac{\phi(1 + (1+b)D_{T-1} + a) + \tau F_T \mathbb{1}_{T\in I})}{\phi + \tau \mathbb{1}_{T\in I}}, \phi + \tau \mathbb{1}_{T\in I}\right), \tag{4.26}$$

$$\pi(D_0|\cdot) \sim N\left(\frac{\mu_0\phi_0 + \phi(1+b)(D_1 - a)}{\phi(1 + (1+b)^2) + \phi_0}, \phi(1 + (1+b)^2) + \phi_0\right), \tag{4.27}$$

where the function $\mathbb{1}_{i\in I}$ is an indicator function that is checking if an observation F_i is available at time i.

We ran the MC for the Gompertz model for a total of 10,000 iterations, with a burn-in of 2,000 iterations, and an adaptation period (`n.adapt` in `JAGS`) of 1,000.

Moran-Ricker SSM Fitting

We also fit Moran-Ricker model in log space by taking the logarithm of the latent states ($X_{1:T}$) and observations ($Y_{t\in I}$), $D_t = \log(X_t)$, $F_i = \log(Y_i)$. Under this log transformation, the process model and observation model can be written as:

$$D_t = a + D_{t-1} + \exp(bD_{t-1}) + \epsilon_t,\ \ \epsilon_t \sim \mathcal{N}(0, \phi), \tag{4.28}$$

$$Y_i = D_i + \epsilon_{obs,i},\ \ \epsilon_{obs,i} \sim \mathcal{N}(0, \tau). \tag{4.29}$$

Assuming that log observations ($\mathbf{F}$) are available at a subset of time points, $I \subset \{1, \ldots, T\}$, the likelihood for the Moran-Ricker model fit in log space is given by:

$$\begin{aligned}\mathscr{L}(X_{1:T}, a, b, \tau, \phi | Y_I) \propto \prod_{t=1}^{T} \sqrt{\phi} \exp\left(-\frac{\phi}{2}(D_t - a - D_{t-1} - \exp(bD_{t-1}))^2\right), \\ \prod_{i\in I} \sqrt{\tau} \exp\left(-\frac{\tau}{2}(F_i - D_i)^2\right).\end{aligned} \tag{4.30}$$

Prior distributions for the Moran-Ricker are identical to those chosen for the Gompertz model (Equations 4.20 – 4.24). a and b were given Uniform$(-10, 10)$, ϕ and τ were given diffuse half Cauchy priors following recommendations by [44] and [121], and the initial latent state D_0 was given a normal prior centered at an initial mean μ_0 with initial precision ϕ_0. Altogether, the prior equations are:

$$\begin{aligned} a &\sim \text{Uniform}(-10, 10), \\ b &\sim \text{Uniform}(-10, 10), \\ \phi &\sim \text{HalfCauchy}(\gamma = 100), \\ \tau &\sim \text{HalfCauchy}(\gamma = 100), \\ D_0 &\sim \mathcal{N}(\mu_0, \phi_0). \end{aligned}$$

We ran the MC for Moran-Ricker model for a total of 10,000 iterations, with a burn-in of 2,000 iterations and an adaptation period (`n.adapt` in `JAGS`) of 1,000.

Lognormal Moment Match SSM Fitting

We explored four different lognormal models using the moment matching approach. These models include: the Gompertz process model embedded for unbiased mean process evolution and unbiased observation model with constant variances (LGC), the Moran-Ricker process model embedded for unbiased mean process evolution and unbiased observation model with constant variances (LMRC), the Gompertz process model embedded for unbiased mean process evolution and unbiased observation model with density dependent variances (LGD), and the Moran-Ricker process model embedded for unbiased mean process evolution and unbiased observation model with density dependent variances (LMRD). For the remainder of this paper, we will be referring to the first two models (LGC, LMRC) as the constant variance models. The term density dependent models will be used to describe the second two models (LGD, LMRD) as well as the classical Gompertz and Moran-Ricker models. We carefully chose these four models to embed common assumptions for different SSM formulations. The unbiased mean process evolution, unbiased observation density function, and constant variance models (LGC, LMRC) were chosen to mimic the assumptions of homoskedastic Gaussian SSMs, which are not frequently used in lognormal SSMs. The embeddings for the LGD and LMRD models were chosen to mimic the variance structure of the Gompertz and Moran-Ricker models while maintaining an unbiased process evolution function and unbiased observation density function.

The likelihood for the constant variance models is obtained by substituting the values for $f^*(X_{t-1}|\Theta), \phi_t, g^*(X_i|\Theta)$ and τ_i from Table 4.1 into the LNM3 likelihood (Equation 4.16). For the constant variance models, two of the prior choices were modified from the priors used for the Gompertz and Moran-Ricker models. First, the prior distribution on a was modified to a

Model Name	$f^*(X_{t-1}\|\Theta)$	ϕ_t	$g^*(X_i\|\Theta)$	τ_i
LGC	$\exp(a)(X_{t-1})^{b+1}$	ϕ	X_i	τ
LMRC	$X_{t-1}\exp(a+bX_{t-1})$	ϕ	X_i	τ
LGD	$\exp(a)(X_{t-1})^{b+1}$	$\phi(\exp(a)(X_{t-1})^{b+1})^{-2}$	X_i	τX_i^{-2}
LMRD	$X_{t-1}\exp(a+bX_{t-1})$	$\phi(X_{t-1}\exp(a+bX_{t-1}))^{-2}$	X_i	τX_i^{-2}

Table 4.1: Process evolution functions, process error structure, observation density function, and observation error structure for the four types of Lognormal Moment Matching SSMs used. LGC and LMRC represent the Gompertz and Moran-Ricker process functions and observation functions with constant process and measurement variance. LGD and LMRD represent the Gompertz and Moran-Ricker process functions and observation functions with density dependent process and measurement variance.

strictly positive uniform distribution, to reflect the non-negativity of the latent process $X_{1:T}$. Second, the prior on the initial condition, X_0, was changed from $\mathcal{N}(\mu_0, \phi_0)$ to $\text{Log}\mathcal{N}(\mu_0, \phi_0)$ since the constant variance models are not being fit in log space. Priors for b, ϕ, and τ (Equations 4.21 - 4.23) were not changed. Altogether, the priors for the constant variance models are

$$a \sim \text{Uniform}(0, 10), \tag{4.31}$$

$$b \sim \text{Uniform}(-10, 10), \tag{4.32}$$

$$\phi \sim \text{HalfCauchy}(\gamma = 100), \tag{4.33}$$

$$\tau \sim \text{HalfCauchy}(\gamma = 100), \tag{4.34}$$

$$D_0 \sim \text{Lognormal}(\mu_0, \phi_0). \tag{4.35}$$

The density dependent moment matching models were fit in log space, by taking the log of the latent states and observations; $D_t = \log(X_t), F_i = \log(Y_i)$. The model likelihood in log

space for the LGD model can be written as:

$$\mathscr{L}(D_{1:T}, a, b, \tau, \phi|\mathbf{F}) \propto \prod_{t=1}^{T} \frac{\exp\left(-\frac{\log(1+\phi^{-1})^{-1}}{2}(D_t - (a + (1+b)D_{t-1} - .5\log(1+\phi^{-1})))^2\right)}{\sqrt{\log(1+\phi^{-1})}} \prod_{i\in I} \frac{\exp\left(-\frac{\log(1+\tau^{-1})^{-1}}{2}(F_i - (D_i - .5\log(1+\tau^{-1})))^2\right)}{\sqrt{\log(1+\tau^{-1})}}. \tag{4.36}$$

Similarly, the model likelihood in log space for the LMRD model can be written as:

$$\mathscr{L}(D_{1:T}, a, b, \tau, \phi|\mathbf{F}) \propto \prod_{t=1}^{T} \frac{\exp\left(-\frac{\log(1+\phi^{-1})^{-1}}{2}(D_t - (a + D_{t-1} + b\exp(D_{t-1}) - .5\log(1+\phi^{-1})))^2\right)}{\sqrt{\log(1+\phi^{-1})}} \prod_{i\in I} \frac{\exp\left(-\frac{\log(1+\tau^{-1})^{-1}}{2}(F_i - (D_i - .5\log(1+\tau^{-1})))^2\right)}{\sqrt{\log(1+\tau^{-1})}}. \tag{4.37}$$

Priors for these two density dependent moment matching models were chosen to be identical to those chosen for the Gompertz and Moran-Ricker models (Equations 4.20 – 4.24).

The specification of the density dependent variance as $\phi_t = \phi f^*(X_{t-1}|\Theta)^2$ and the log-linear nature of the Gompertz curve conveniently lead to closed form full conditional distributions for the latent states in the LGD model. Similar to the Gompertz model, this allows for Gibbs sampling to be used to update the latent states, which helps to decrease computation time. The derivation of the full conditional distributions is involved, and can be found in the Gibbs Sampling Updates for Certain LNM3 Structures section of the Appendix A.

We ran the MC for each of the four moment matching models for a total of 10,000 iterations, with a burn-in period of 2,000 iterations, and an adaptation period (`n.adapt` in `JAGS`) of 1,000.

4.3.4 Simulation Study

We designed a simulation study to assess the forecasting performance of the six models presented here, both under the cases where they are the true generating models and where the models are mis-specified. Our three primary objectives for the simulation study were: 1) assess forecasting performance of each model when it is the true generating model and when it is misspecified, when observation precision is being estimated 2) assess forecasting performance of each model when it is the true generating model and when it is misspecified, with fixed observation precision; 3) analyze the estimability of precisions for the models considered in this manuscript.

To assess the forecast performance of our models, we used proper scoring rules [48]. Broadly, proper scoring rules use information about the predictive distribution coupled with observations to assign a measure of agreement of the forecast and the observations [80]. Specifically, Gneiting and Raftery [48] define a scoring rule to be proper if the expected value of the score is maximized by a draw from the true forecast distribution, and show that both the Continuous Ranked Probability Score (CRPS) and the Ignorance (IGN) score are proper scoring rules. For our simulation studies, we consider the ignorance score [51, 130] and the Continuous Ranked Probability Score [97]. Given a probability density function $f(\cdot)$ and corresponding cumulative distribution function $F(\cdot)$ for our forecast, and an observation y, the IGN and CRPS may be written:

$$\text{IGN}(y) = -\log(f(y)); \tag{4.38}$$

$$\text{CRPS}(y) = \int_{\mathbb{R}} (F(z) - \mathbb{1}_{z \geq y})^2 dz. \tag{4.39}$$

We use the IGN and CRPS scores to quantitatively compare forecasts, with lower scores within a scoring rule indicating a better performance. CRPS has support over the positive real line, $[0, \infty)$, while IGN can take values between $[-\log(f(y^*)), \infty)$, where $y^* =$

$\text{argmax}_{y \in \mathbb{R}} f(y)$.

To analyze the estimability of precisions in each of our six models, we used the empirical coverage rate of the 95% highest posterior density credible intervals for the precision parameters. Auger-Méthé et al. [5] show that SSMs can have difficulties recovering the process and observation precisions even when the models are linear and Gaussian. To perform a thorough analysis on the estimability of the precision parameters, we fit each model in two different ways. The first model fixed the observation precision and estimated the process precision. The second model estimated both the observation precision and the process precision. This approach allowed us to assess the estimability by looking at the increase in empirical coverage rate for the process precision when the observation precision is fixed compared to when the observation precision is estimated. By choosing to use coverage and proper scoring rules to quantify our model performance we follow a common approach used in the literature – using frequentist concepts to assess Bayesian models [15, 131, 91, 90].

Param.	Gomp	MR	LGC	LMRC	LGD	LMRD
$\exp(a)$	.82	1.26	1.21	1.11	1.21	1.11
b	-.658	-.034	-.099	-.014	-.099	-.014
ϕ	70.2	51.9	4	4	70.2	70.2
τ	188.7	188.7	4	4	188.7	188.7

Table 4.2: Parameter values used for the six different model formulations to create thirty synthetic datasets for each generating model.

To quantify our three objectives, the simulation study was performed as follows: for each of the six models, thirty different synthetic datasets of length 575 were generated by simulating from the underlying process model and observation model, with parameter values taken from Table 4.2. Parameter values for each model were chosen so that the systems had similar mean dynamics for each generating model. Each synthetic dataset was fit by each the six models

discussed here. Models were initially fit with the first 365 days of data, and then forecasts were computed for days 366 – 372, a seven day forecast horizon. The average IGN and CRPS for the seven day forecasts horizons were computed using the `logs_sample` and `crps_sample` function from the `scoringRules` R package [68]. Highest posterior density credible intervals for the precisions being estimated were also saved. We then re-fit the models with the first 372 days of data, and forecasts were computed for days 373–379. This process was repeated until the synthetic data was exhausted. Overall, each individual synthetic data set was fit a total of 12 times – once by each model with observation precision known and once by each model with observation precision estimated, for a total of 2,160 simulations.

This simulation study design helps us to assess our three objectives. Our first two objectives were to assess the forecasting performance of each model under model mis-specification; both in the case of estimated observation precision and in the case of fixed observation precision. Our simulation study had each model fit each dataset twice: once with observation precision fixed and once with observation precision estimated. The forecasts for each precision scenario were evaluated using the CRPS and IGN scores. By using the same synthetic datasets, we were able to isolate the differences in forecast performance for each precision scenario and quantify the difference using paired t-tests with a Holm-Bonferroni adjustment [57]. With thirty different synthetic datasets generated by each model and each synthetic dataset evaluated over thirty forecast horizons by each of the six models, we were able to get a comprehensive picture of how each model performed under model mis-specification for each of the precision scenarios. Our third objective was to analyze the estimability of precisions for the models considered in this manuscript. We used differences between the scores of the forecasts and coverage rates of the precision credible intervals to determine the severity of underlying precision estimation problems. In the absence of estimation problems, we expect that the empirical coverage rates of the precision parameters should achieve close to the nominal coverage rate of 95%. CRPS and IGN are proper scoring rules, and their expected scores are maximized by draws from the true forecast distribution [48]. Thus we expect that

across a large number of simulations, the true generating model should score best in terms of CRPS and IGN in the absence of estimation problems.

4.3.5 Case Study: San Juan Dengue Fever Hospitalizations

The objective of our case study was to assess the forecasting performance and identify strengths, weaknesses, and similarities of the six models presented here using real data. To do this, we used dengue fever hospitalization data collected in San Juan, Puerto Rico taken from the NOAA dengue forecasting challenge [112, 66]. This dataset contains a time series of weekly hospitalizations due to dengue fever from 1990 to 2013, for a total of 1196 time points. Our model fitting and evaluation was done on a transformed dataset,

$$\mathbf{Y}^* = \sqrt{\mathbf{Y} + 1}, \tag{4.40}$$

where $\mathbf{Y}$ represent the San Juan dengue data on the original scale and $\mathbf{Y}^*$ represent the transformed data. The square root helps to reduce large peaks that would be difficult to capture even with a density dependent variance, while the 1 is added under the square root to push values away from zero, as our lognormal models assume a strictly positive population and the dataset contains multiple zeros.

We fit the transformed dengue hospitalization data using the same formulations of the Gompertz and Moran-Ricker models used for the simulation study. The Gompertz curve has previously been used for disease population modeling [see 11, 79, 94, for examples], and the Moran-Ricker model has seen use in modeling risk of vector-borne disease tranmission by mosquitoes [133]. The six models fit were identical to the ones used in our simulation studies: the Gompertz model, the Moran-Ricker model, and the four LNM3 models (LGD, LMRD, LGC, LMRC). Given that the hospitalization data were confirmed, post-processed dengue cases, we assumed that the observation precision was large and fixed to allow us

to still reasonably account for the possibility of false positives and false negatives. For the density dependent variance models $\tau \approx 4167$ was used and for the constant variance models $\tau \approx 104.5$ was used, giving the density dependent models and the constant variance models similar observation precision on average. Each model used the same priors for a, b, ϕ, τ, and D_0 as were used in the simulation study.

Our case study analysis was designed to emulate monthly dengue hospitalization forecasts to be used by hospitals so that they can use previous observations to inform their plans for resource allocation. Models were initially fit using the first 396 weeks of data. Forecast distributions were then generated for the next four weeks, and are used alongside observations to compute CRPS and IGN scores. The model was then re-fit using the new observations and forecast distributions are generated for the next four weeks. This iterative process was repeated 200 times in total, until the entire 1196 observation time series was exhausted. We compared models based on CRPS and IGN performance for one, two, three, and four week horizons.

4.4 Results

4.4.1 Simulation Study

The first objective of our simulation study was to assess the forecasting performance of each model under model mis-specification when observation precision is estimated. We expected that the true generating model would perform best on average for its thirty synthetic datasets, using the average IGN and CRPS scores over each seven day forecast horizon. For the case where both the process precision, ϕ, and the observation precision, τ, were estimated, the Gompertz model (Gomp in Table 4.3) had the best forecasting performance among the density dependent models (MR, Gomp, LMRD, LGD) for both CRPS and IGN. The unbiased moment matching analog to the Gompertz, LGD, also had a strong perfor-

mance, scoring second highest for IGN on all four density dependent variance models, and scoring second highest for CRPS on three out of the four. For the constant variance models (LMRC, LGC), the constant variance Moran-Ricker model (LMRC) had the best forecasting performance for both CRPS and IGN.

Generating Model						
Model	MR	Gomp	LMRC	LGC	LMRD	LGD
Average CRPS						
MR	.8587	.8027	.6496	.6362	.8752	.8970
Gomp	**.8344**	**.7773**	.5997	.5952	**.8279**	**.8500**
LMRC	*.8356*	.7810	**.5913**	**.5889**	.8411	.8672
LGC	.8375	.7809	*.5927*	*.5900*	.8440	.8672
LMRD	.8380	.7872	.6181	.6100	.8357	.8585
LGD	.8357	*.7789*	.6080	.6016	*.8310*	*.8517*
Average IGN Score						
MR	1.816	1.744	1.595	1.557	1.806	1.830
Gomp	**1.785**	**1.710**	1.514	1.494	**1.751**	**1.774**
LMRC	1.812	1.757	**1.456**	**1.452**	1.858	1.915
LGC	1.806	1.750	*1.458*	*1.454*	1.840	1.882
LMRD	1.793	1.723	1.537	1.507	1.762	1.790
LGD	*1.792*	*1.714*	1.530	1.500	*1.755*	*1.778*

Table 4.3: Average CRPS and IGN scores for the simulations where both τ and ϕ are estimated. Columns represent the generating model for the synthetic datasets and rows represent the models used to fit the datasets. Scores are averaged over thirty different synthetic datasets and thirty different 7 day forecast horizon for each combination of generating model and model used to fit the data. Bolded entries represent the lowest score for a given generating model, and italicized entries represent the second lowest score.

The second objective of our simulation study was to assess the forecasting performance of

each model under model mis-specification when observation precision was fixed. For the simulations where the observation precision, τ, was fixed and the process precision, ϕ, was estimated, we found higher consistency between our scoring rules and the generating model. For the simulations that estimated both precision parameters (Table 4.3), the Gomp and LGD models consistently outperformed the other density dependent variance models for both scoring rules. In contrast, the simulations where τ was fixed and ϕ was estimated (Table 4.4) showed a more equal representation among the density dependent models. For the four density dependent generating models, the Moran-Ricker and its unbiased moment matching counterpart scored the best for CRPS, with each model producing the lowest score twice. When measuring performance with IGN, the Moran-Ricker and Gompertz models each scored lowest twice. The density dependent models all scored similarly, agnostic to the choice of true generating model, both for CRPS (max difference $\approx .0052$) and for IGN (max difference $\approx .006$).

Generating Model						
Model	MR	Gomp	LMRC	LGC	LMRD	LGD
Average CRPS						
MR	*.8332*	*.7770*	.5968	.5945	**.8230**	**.8492**
Gomp	.8358	.7774	.6000	.5954	.8276	.8508
LMRC	.8357	.7813	**.5911**	**.5891**	.8337	.8619
LGC	.8362	.7805	*.5929*	*.5897*	.8349	.8603
LMRD	**.8332**	**.7766**	.5971	.5937	*.8233*	*.8495*
LGD	.8359	.7777	.5999	.5959	.8282	.8510
Average IGN Score						
MR	**1.785**	1.711	1.502	1.487	**1.748**	*1.775*
Gomp	1.788	**1.710**	1.515	1.494	1.752	**1.774**
LMRC	1.819	1.764	**1.455**	**1.453**	1.848	1.905
LGC	1.813	1.750	*1.458*	*1.453*	1.834	1.887
LMRD	*1.786*	1.712	1.505	1.486	*1.752*	1.780
LGD	1.788	*1.710*	1.513	1.494	1.752	1.776

Table 4.4: Average CRPS and IGN scores for the simulations where ϕ is estimated and τ is known. Columns represent the generating model for the synthetic datasets and rows represent the models used to fit the datasets. Scores are averaged over thirty different synthetic datasets and thirty different 7 day forecast horizon for each combination of generating model and model used to fit the data. Bolded entries represent the lowest score for a given generating model, and italicized entries represent the second lowest score.

Our third objective was to analyze the estimability of precisions for each of the six models we used. Our first investigation for the estimability of the precision parameters was to quantify the differences in IGN and CRPS scores between the simulations where the observation, τ was fixed and the simulations where τ was estimated. This investigation is also intimately related to the performance of the models under model mis-specification, and therefore provides

insight for all three of our objectives. To evaluate the impacts statistically, we used a paired t-test. To do this, we took the average IGN and CRPS scores over each seven day forecast horizon for the two different scenarios, and treating them as "before" and "after". We justify this by noting that each precision scenario was fit using identical synthetic datasets, with the only difference being whether τ was estimated or not. Unsurprisingly, we found that fixing the observation precision helped to improve forecasting performance for nearly all models. The LGC and LMRC models performed significantly better in terms of CRPS when τ was fixed (LGC: $p = 4.7\text{e-}05$; LMRC: $p = .0032$), but did not perform significantly better in terms of IGN (LGC: $p = .68$; LMRC: $p = .36$). The Moran-Ricker, LMRD, and LGD models all had statistically significant decreases in IGN (MR: $p < 2\text{e-}16$; LMRD: $p = 2.1\text{e-}09$; LGD: 2.8e-05) as well as CRPS (MR: $p < 2\text{e-}16$; LMRD: $p < 2\text{e-}16$; LGD: 1.4e-05) when τ was fixed. Fixing the observation precision did not significantly impact the performance of the Gompertz model for IGN ($p = .84$) or CRPS ($p = .82$).

Our second investigation for the estimability of the precision parameters was to use the empirical coverage rate of the HPD intervals. The empirical coverage of the highest posterior density (HPD) credible intervals for ϕ unanimously increased for the simulations where τ was fixed, and all six models had empirical coverage that falls close to the nominal rate. For the simulations where both ϕ and τ are estimated, Table 4.5 shows that the Gompertz model and LGD models produce the best empirical coverage for the precision parameters. This is likely to be related to their excellent performance in these simulations, where other models struggled to consistently produce precision estimates that contained the ground truth in their 95% HPD intervals. The empirical undercoverage of the Moran-Ricker (MR) model for both ϕ and τ may also explain its poor performance in the forecast results from Table 4.3, where it came in last place for CRPS for all six generating models and last place for IGN for three out of six generating models, including the case where it was itself the true generating model.

Generating Model						
Coverage	MR	Gomp	LMRC	LGC	LMRD	LGD
ϕ Cvg	72.3%	89.1%	89.2%	86.9%	85.8%	89%
τ Cvg	74.7%	96.4%	92.1%	91.8%	91.7%	95.3%
ϕ Cvg, τ fixed	95.2%	94.8%	94.1%	93.6%	94.2%	94.9%

Table 4.5: Average empirical HPD credible interval coverage for the precisions of each generating model, under the scenarios where both ϕ and τ are estimated and when τ is fixed. Coverage rates are averaged over all thirty synthetic datasets and all thirty forecast horizons, for a total of 900 samples.

4.4.2 Case Study: San Juan Dengue Fever Hospitalizations

We found clear differences in forecasting performance between the models that had constant variance (LMRC and LGC in Table 4.6) and the models that had density dependent variance (MR, Gomp, LMRD, LGD in Table 4.6). The constant variance models had better performance for one and two week forecast horizons, both when measured by CRPS and by IGN. For three week forecast horizons, the constant variance models perform similarly to the density dependent variance models in terms of CRPS, but perform comparatively poorly in terms of IGN. At the four week forecast horizon, the constant variance models are the worst performing of the six for both CRPS and IGN.

Similarities were also found in forecasting performance between groups of models. The Gompertz model (Gomp) and the Gompertz density dependent moment matching model (LGD) performed very similarly for all forecast horizons, quantified by both CRPS and IGN. The Moran-Ricker model (MR) and the Moran-Ricker density dependent moment matching model (LMRD) also displayed similarity in CRPS and IGN forecast performance across all forecast horizons. These results are not surprising though, as the moment matching models are structurally similar to their traditional counterparts, with slight modifications to the

process and observation models to remove bias.

We found that quantifying the overall best performing forecast models across all horizons is dependent on the metric that we choose – CRPS or IGN. Using CRPS as our metric, the constant variance Moran-Ricker and constant variance Gompertz (LMRC and LGC) perform the best overall. Choosing to use IGN as the metric instead, the Gompertz model and the unbiased density dependent Gompertz model (LGD) have the best overall performance. Indeed, we see that although both scoring rules used are proper and therefore favor the true distribution in expectation, we have no guarantee of agreement for finite sample sizes.

Fitting Model						
Horizon	MR	Gomp	LMRC	LGC	LMRD	LGD
Average CRPS						
One Week	.5247	.5197	*.4985*	**.4964**	.5244	.5205
Two Week	.5976	.5961	*.5837*	**.5818**	.5972	.5954
Three Week	**.7213**	.7317	.7247	.7256	*.7219*	.7321
Four Week	*.8334*	.8446	.8546	.8518	**.8330**	.8439
Overall	.6692	.6730	*.6654*	**.6639**	.6691	.6730
Average IGN Score						
One Week	1.491	1.472	*1.305*	**1.304**	1.490	1.467
Two Week	1.502	1.495	*1.491*	**1.481**	1.497	1.495
Three Week	1.631	*1.624*	1.751	1.719	1.630	**1.624**
Four Week	1.762	**1.752**	1.898	1.877	1.764	*1.755*
Overall	1.597	*1.586*	1.611	1.595	1.595	**1.585**

Table 4.6: Average CRPS and IGN scores for models fit on the transformed dengue data. Scores were computed for one, two, three, and four week horizons, and the entries in the table are averaged across a total of 200 forecast horizons. Bolded entries represent the lowest score for a given horizon, and italicized entries represent the second lowest score.

4.5 Discussion

Although the Gaussian distribution is a common choice in modeling applications, many ecological processes have strict lower bounds that are not accurately captured by using Gaussian models. These problems are especially exacerbated in forecasting applications, where process uncertainty grows as the forecast horizon increases. Embedding assumptions about the evolution of the latent process and variance dynamics into non-Gaussian SSM frameworks can be challenging. To remedy this, we proposed a method for embedding non-negative process and observation models with arbitrary variance structures into lognormal Bayesian SSMs using moment matching (LNM3). The primary advantage of our method is flexibility: it allows practitioners to create stochastic lognormal distributions for process and measurement components that are unbiased in terms of their mean evolution and observations, have a flexible variance that can change through time, and offer a closed form Markov transition density that allows models to be fit with MCMC software such as `JAGS` [119].

We used a computationally intensive Monte Carlo approach to assess the forecasting performance of the six models discussed here, using a total of 180 synthetic datasets that were fit twelve times each: once by each model with the observation precision estimated, and once by each model with the observation precision fixed. We found that the forecasting performance of our models under mis-specification was heavily dependent on whether or not the observation precision was fixed, and also dependent on the metric used for evaluation: CRPS or IGN. With the observation precision estimated, the Gompertz model had the best average CRPS and IGN scores across all of the synthetic dataset for four out of the six generating models. With the observation precisions fixed at the true values, we found that every model except for the Gompertz model had a significant increase in forecast performance when measured by average CRPS or average IGN. For these simulations, no one model dominated the others in terms of forecast performance for either metric. The Gompertz model outperforming the true generating models identifies one of the difficulties

of using proper scoring rules to evaluate forecast performance. Although the CRPS and IGN scores should favor the true generating model in expectation, we are unlikely to have access to the ground truth parameter values and instead have to rely on estimates of parameter values from our MCMC.

Even simple linear Gaussian SSMs can be prone to estimation problems, especially for parameters that govern the variance structure of the process and observations [5]. We found that our models were no exception to this: when both the process precision and observation precision were estimated, no model came within 5% of the nominal 95% coverage rate of the process precision HPDs. We also found that estimation of the precision parameters was closely related to forecasting performance. For the simulation where both observation and process precision parameters were estimated, the two models with the best performance (Gompertz and LGD) were also the models that had empirical coverage rates closest to 95% for the precision parameters. Similarly, the Moran-Ricker model was 20% below the nominal coverage rate for both the observation and process precision parameters, and had the worst average CRPS scores for every generating model, including itself. For the simulation studies where the observation precision was fixed, we found that coverage rates for each of the six models were close to the nominal 95% coverage rate, with the empirical coverage rates ranging from 93.6% to 95.2%. This supports the findings from Auger-Méthé et al. [5], who show that fixing the measurement error in linear Gaussian SSMs can help to alleviate estimation problems.

The six models fit here showed similar overall performance when used to generate iterative four week forecasts of dengue fever hospitalizations using data from the NOAA Dengue Forecasting Challenge [112, 66]. For the 200 four week forecasts, the mean CRPS scores averaged over all forecast horizons differed by .009 between the best and worst performing model, and the mean IGN scores over all forecast horizons differed by .026. We found that the constant variance models had the best performance for one and two week forecast horizons in terms of average CRPS and IGN scores, and also had the best overall performance for

CRPS averaged across all forecast horizons. The strong near term forecasting performance of the constant variance LNM3 models lends credence to the power of being able to embed sensible model dynamics into lognormal SSM frameworks, as although an unbiased process evolution, observation function, and constant variance is a common structure for Gaussian SSMs, it is rare to fit lognormal SSMs with these assumptions

Though the methods presented here use the lognormal distribution to represent stochasticity, the moment matching approach broadly applies to other distributions as well, and provides opportunities for future directions. For example, the gamma distribution has been considered for state space modeling [142] and stochastic differential equation modeling [31] applications, has non-negative support, can be parameterized in terms of its mean and variance using a moment matching approach, and has lighter tail behavior than the lognormal distribution. The beta distribution, which is beginning to be used in SSMs [see 113, 34, for examples], has a useful support for modeling proportions and can also be parameterized in terms of its mean and variance to allow for moment matching approaches.

In conclusion, to address biological non-realism in models of physical systems, we proposed a novel lognormal state space modeling framework that preserves positivity of the latent process and observations. The methods presented here allow practitioners to embed complex process models and error dynamics into state space models while ensuring that the forecasts they get out of the model agree with the constraints of the system. The flexibility of the moment matching method for representing complex systems along with the variance partitioning of the state space model provide a coherent statistical framework for forecasting, in terms of biophysical realism, forecast assessment, and uncertainty quantification.

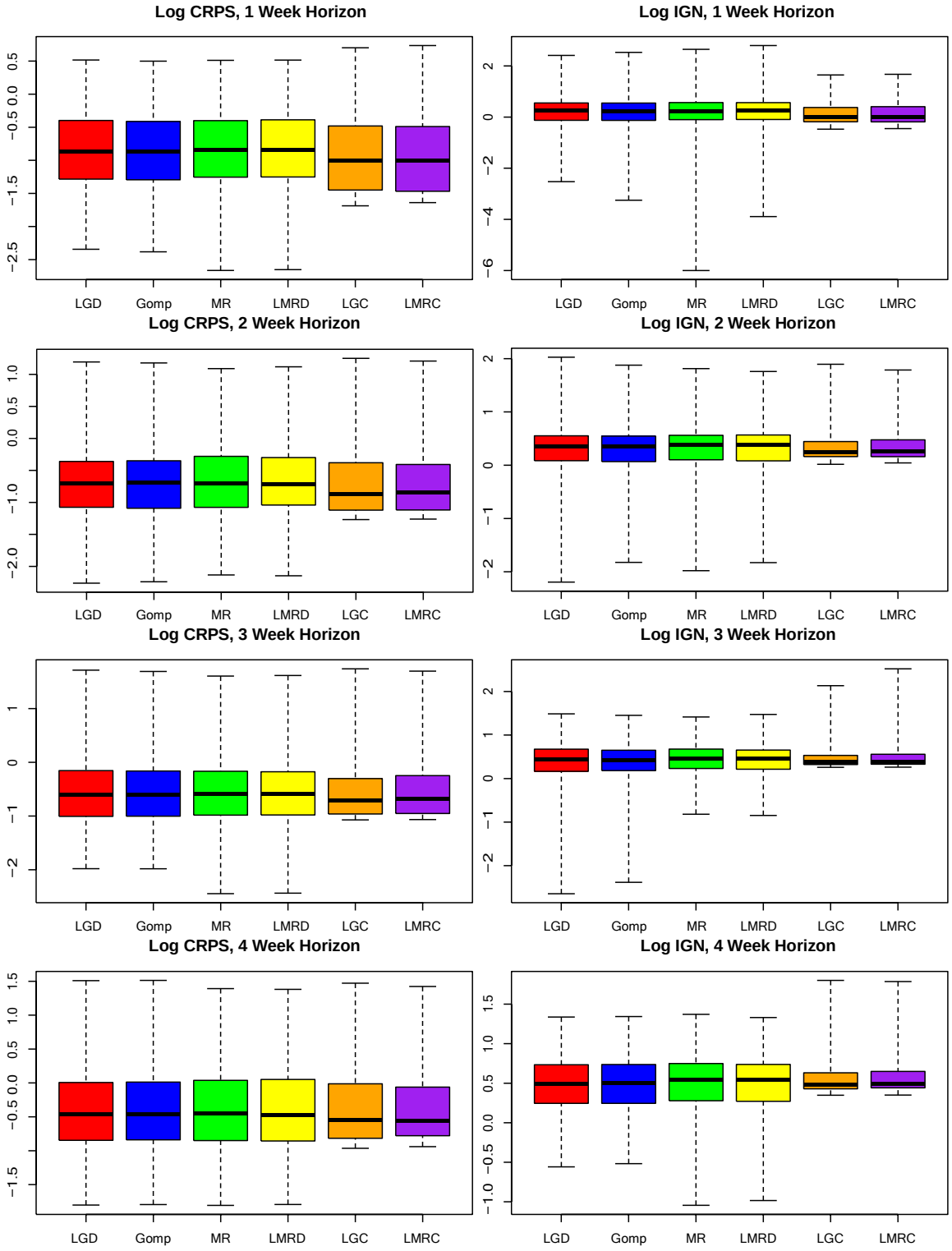

Figure 4.3: Boxplots of log CRPS and log IGN over all forecast horizons for each of the six models used to fit the transformed dengue hospitalization data. The log transform was used for visualization, as its monotonicity preserves the minimum, maximum, and median information displayed by the boxplots.

Chapter 5

Ecosystem Carbon Budgets using Density Dependent Lognormal SSMs

5.1 Introduction

State space models (SSMs) are powerful tools that have been used in a variety of ecological applications, including population modeling and dynamics [78, 77], animal movement [111, 115, 98] and disease ecology [61, 59]. SSMs allow for uncertainty to be partitioned between the error in the underlying dynamical system, measurement error, and initial condition error. This partitioning and propagation of the uncertainty makes SSM frameworks a strong choice for ecological forecasting applications, due to the importance of quantifying the stochasticity of numerous sources that contribute to the predictability of a process [24, 35].

While SSMs have seen growing popularity in other ecological applications, they still remain largely unused in the ecosystem carbon modeling literature. Instead, ecosystem carbon models are often fit using a Bayesian model data fusion framework (MDF) to combine model output with observations [see 70, 12, 13, 155, for examples]. There are advantages to these methods: they are computationally cheap compared to fitting SSMs, and they still provide quantifications of parameter uncertainty and measurement error. One disadvantage of these methods is that they do not quantify the uncertainty in the process model itself, instead assuming that the underlying process model is a perfect simulator of the system dynamics. With carbon uptake and release by terrestrial ecosystems expected to play a major role

in climate change projections [75, 132, 83], uncertainty in our process models needs to be handled meticulously to ensure that our predictions are reliable [74]. In this regard, the state space modeling approach can help to provide a cohesive statistical framework for uncertainty quantification in ecosystem carbon forecasts.

When ecosystem carbon models are fit as SSMs, they are typically fit as linear Gaussian SSMs [see 145, 146, as well as this work, Chapter 3 for examples]. In this paper, we propose a novel non-linear, non-Gaussian SSM based on the lognormal distribution. Our lognormal SSM embeds the DALEC2 process model [12] for the unbiased stochastic evolution function, an unbiased observation model to relate the measurements to the latent states, and uses a density dependent variance structure [33] for both the latent process and observation models. Using a lognormal distribution to model the latent process and the observations allows us to enforce the positivity constraint of the latent states, while the density dependent variance structure allows the heteroskedasticity of to vary with the magnitude of the latent process.

One of the challenges of choosing to use a non-linear, non-Gaussian SSM is that it denies us access to analytic approaches such as Gibbs sampling [46, 22], Kalman filtering [71], and extensions of the Kalman filter [2, 23, 41]. These complex SSMs instead have to rely on a set of techniques called Sequential Monte Carlo methods [SMC; 36, 37]. Among the most popular of these SMC techniques is the particle filter [52, 21]. Particle filtering offers a number of fitting methods for SSMs, including both maximum likelihood based approaches [61, 62] and Bayesian approaches [42, 3]. The robust parameter uncertainty quantification and simultaneous estimation of parameters and latent states from Bayesian particle filtering methods does not come without trade-offs. While results exist for tuning and adapting proposal distributions [e.g. 4, 128] and theoretical results exist for choosing the number of particles [38, 136, 118], particle based Markov Chain Monte Carlo (MCMC) methods can remain difficult to tune in practice. In this paper, we carefully consider the trade-offs of these types of fitting methods for our difficult problem: a six dimensional dynamical system with sparse observations and 22 free parameters, run on a daily time step for 1950 days.

A historical roadblock for constraining complex mechanistic process based modeling approaches is data: both availability and quality. Long term observation networks such as the National Ecosystem Observatory Network (NEON) help to ameliorate both of these issues by providing regular measurements of meteorological driver variables, carbon stocks, and carbon fluxes with a standardized collection procedure. With 47 different terrestrial NEON sites sampled over a variety of ecosystem and climate type, long term time series of NEON data will help to provide insight on the terrestrial carbon budget in the US. In this study, we fit our SSM using data from the University of Notre Dame Environmental Research Center (UNDE) NEON site. We use NEON data products to extract a carbon budget for UNDE, and incorporate two additional data sources: MODIS LAI [29] and atmospheric carbon measurements from the Scripps project [73].

The study we present here had two objectives. Our primary objective was to identify suitable methods to fit our data-limited SSM using NEON data from UNDE, and our secondary objective was to assess the out of sample prediction for Leaf Area Index (LAI) and Net Ecosystem Exchange (NEE) using our model. We considered two fitting methods, the Iterated Filtering algorithm of Ionides et al. [62] (IF2) and the Particle Marginal Metropolis Hastings algorithm of Andrieu et al. [3] (PMMH). Both of these algorithms have shown promising results in other studies involving complex state space models. For example, iterated filtering algorithms such as IF2 have facilitated the use of computational likelihood-based methods for complex SSMs in disease ecology [see 60, and references therein], while PMMH has previously been used to estimate latent states and parameters for applications that embed mechanistic process based models into Bayesian hierarchical modeling frameworks [see 67, 114, for examples]. To assess the out of sample prediction for LAI and NEE we used two strictly proper scoring rules, the Continuous Ranked Probability Score [CRPS; 97] and the Ignorance score [IGN 51]. Strictly proper scoring rules provide a unified framework for comparing probabilistic predictions, and have been used in a wide variety of fields to assess predictive performance of modeling approaches [see 48, and references therein for many examples].

The layout of the remainder of this Chapter is as follows: in Section 2 we discuss the data extraction process, details of the process and state space models, and background on SMC and the fitting methods used here. In Section 3 we present results for our SSM estimates of six different carbon pools, and evaluate out of sample prediction. In Section 4 we provide a discussion of findings of our study, places for improvement, and future work.

5.2 Methods

5.2.1 Data Collection

We chose the University of Notre Dame Environmental Research Center (UNDE) NEON site in Wisconsin, U.S.A. (46.23391° N, -89.537254° W) as the location for our carbon budget analysis. We used a total of seven NEON data products to create our carbon budget. All NEON data products were downloaded using the `neonstore` CRAN package [14].

The process model for our analysis, DALEC2 [12], requires four daily meteorological driver variables: minimum temperature, maximum temperature, shortwave radiation, and atmospheric carbon. Minimum and maximum temperatures were extracted from NEON's triple aspirated air temperature measurements [DP1.00003.001; 109] by taking the minimum and maximum temperatures from daily subsets of the data. We computed daily shortwave radiation by aggregating 30 minute shortwave radiation measurements from NEON's shortwave radiation data product [107] up to a daily timescale, and converting the units to $\mathrm{MJm^{-2}d^{-1}}$. Missing daily temperature and daily shortwave radiation data were imputed using a linear interpolation. We obtained a daily time series of atmospheric CO2 using monthly CO2 data from the Scripps project [73] and assigning each day within a given month its monthly atmospheric CO2 value.

To extract biomass carbon, we first downloaded vegetation structure data from NEON

[DP1.10098.001; 110]. NEON vegetation structure has data available from tower plots, which are plots located in the measurement footprint of the flux tower that measures carbon flux and meteorology, and are usually placed in the most prevalent vegetation type. We found that three tower plots (UNDE037, UNDE043, UNDE065) had annual biomass measurements for the past 6 years. Then, we subset the vegetation structure data to consider only these plots. The subsetted data was cleaned to ensure that: individuals classified as alive were not previously classified as dead; individuals with genus and species information for some years that were missing for other years were given the correct genus and species information; missing stem diameter measurements were linearly interpolated between observations; plants that go from alive to dead were given turnover flags; plants whose stem diameter measurements increase from one year to the next were given growth flags; plants whose measurement heights increase to over 130cm for the first time were given ingrowth flags; standing dead plants were given a standing dead flag. We used the cleaned measurements of stem diameter, genus, and species information to calculate above-ground biomass using the `allodb` R package [50]. Above-ground biomass was converted to above-ground biomass carbon (BiomassC) by multiplying by %C, the number of grams of carbon per gram of Biomass. For our study, %C was given a value of .5. To calculate yearly growth, ingrowth, and turnover rates for each plot, for a given year i and a plant survey set K we used the following formulas:

$$\begin{aligned}
\text{Growth}_i &= \sum_{k \in \text{Survey}} \mathbf{1}(\text{GrowthFlag}_k) \cdot (\text{BiomassC}_i(k) - \text{BiomassC}_{i-1}(k)), \\
\text{Ingrowth}_i &= \sum_{k \in \text{Survey}} \mathbf{1}(\text{IngrowthFlag}_k) \cdot \text{BiomassC}_i(k), \\
\text{Turnover}_i &= \sum_{k \in \text{Survey}} \mathbf{1}(\text{TurnoverFlag}_k) \cdot \text{BiomassC}_i(k).
\end{aligned}$$

For each combination of years i and j, $j > i$, we checked that the following equation holds

for each of our plots, to ensure that there was no biomass carbon that was unaccounted for:

$$\Delta\text{BiomassC}_i^j = \sum_i^j (\text{Growth} + \text{Ingrowth} - \text{Turnover}).$$

We extracted yearly biomass measurements, yearly standing dead measurements, and yearly growth, ingrowth, and turnover rates by taking the average over the three plots.

To extract soil carbon measurements, we downloaded the Megapit soil physical and chemical properties dataset from NEON [DP1.00096.001; 108]. We computed an estimate of total soil carbon by summing the product of the layer width, layer bulk density number (ρ), and %Carbon over each megapit soil horizon:

$$\text{Soil Carbon} \approx \sum_{h \in \text{Horzion}} (\text{TopDepth}(h) - \text{BottomDepth}(h)) \cdot \rho(h) \cdot \%\text{C}.$$

For root carbon estimates, we downloaded the Root biomass and chemistry data product from NEON [DP1.10067.001; 106]. We subset the data to only include the UNDE037, UNDE043, and UNDE065 tower plots. We obtained root carbon measurements by summing the root mass divided by its sample area, and then taking the average over the three plots.

For NEE measurements, we first downloaded the bundled eddy covariance data product from NEON [DP4.00200.001; 104]. To make a time series of daily NEE measurements, we extracted the 30-minute turbulent CO2 fluxes, and checked to see if at least 50% of measurements passed NEON's QAQC for a specific date. If they did, we aggregated them up to a daily timescale, and converted the units to gCm^{-2}. Otherwise, NEE measurements for that date were treated as missing.

To obtain data on litterfall, we downloaded the litterfall and fine woody debris production and chemistry dataset from NEON [DP1.10033.001; 105]. This data product contains information on samples collected using litter traps. Litter traps are fixed area containers left

out for a period of time to accumulate litter. Information is available on: dates the traps were set and collected, area of the trap, amount of litter collected, and location (plotID). To extract daily rates of litterfall for our three plots of interest, we first subset the litterfall dataset to contain only litter traps from UNDE037, UNDE043, and UNDE065. Next, for a given date t, we further subset the litterfall dataset to look at litter traps that were present during date t. For each plot, we computed:

$$\text{Litterfall}(t) = \sum_{i \in \text{Traps}} \frac{\text{Dry Mass}(i)}{\text{Duration}(i)\text{TrapArea}(i)}.$$

Then, we took the average over the plots gave us a time series of approximate daily measurements of the litterfall flux across all littertraps for each of our three tower plots. We note that these measurements were approximate because they were obtained by taking the average across each duration of the littertraps in the field.

We obtained a semi-continuous time series of four day NASA MODIS leaf area index (LAI) estimates at UNDE from a repository of Oak Ridge National Laboratory using their fixed subset tool [101, 29]. We performed QAQC on the MODIS LAI data by only considering periods where 80% or more of the measurements were non-missing. Average measurements for these periods were computed by taking the mean of the non-missing observations.

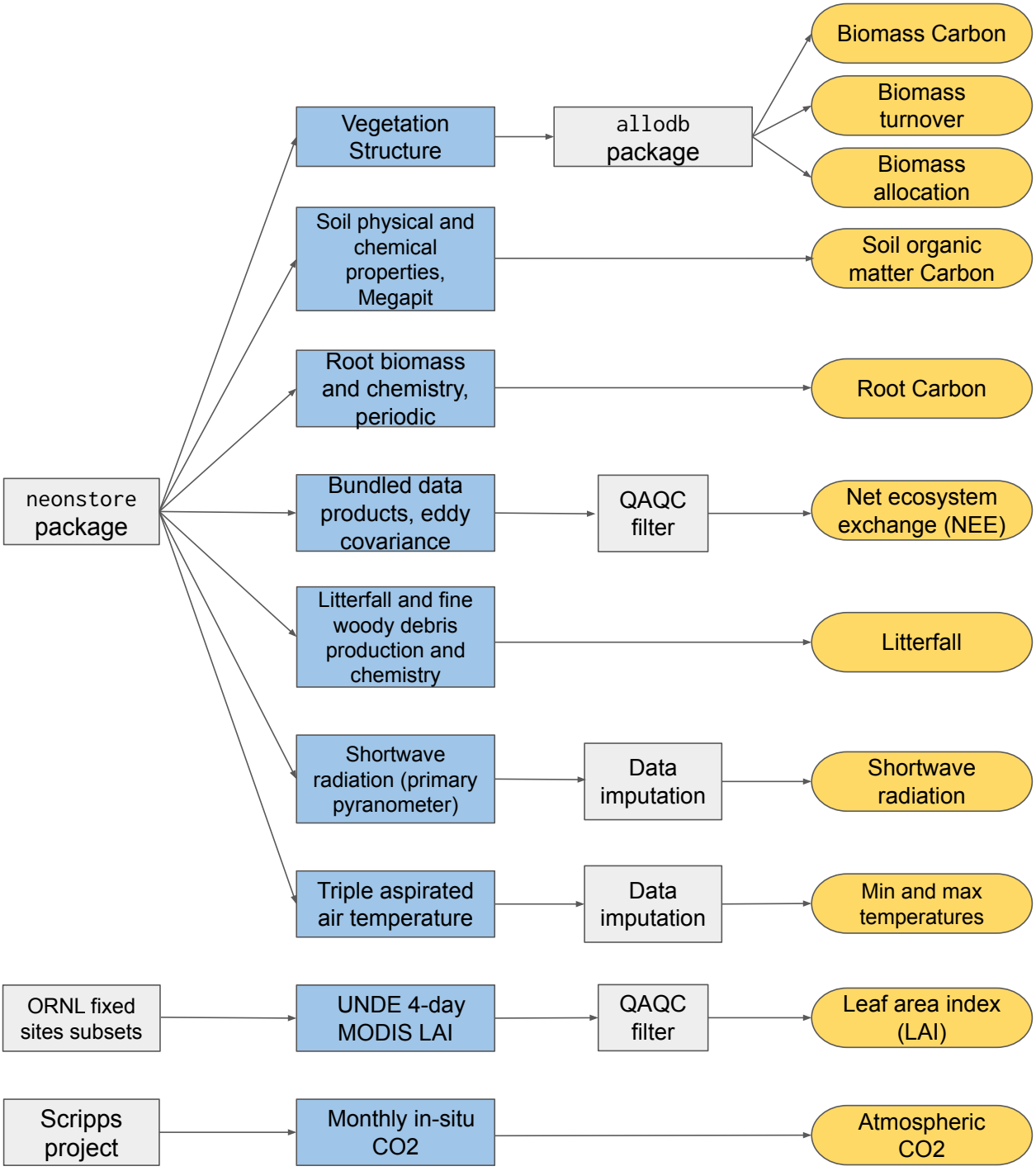

Figure 5.1: Diagram describing the different sources of data and tools used to create our carbon budget at UNDE. Grey boxes denote different data sources and tools that were used, blue boxes denote downloaded data, and orange ovals denote final curated measurements of carbon stocks, fluxes, and meteorological driver variables.

5.2.2 DALEC2 Process Model

To model the carbon dynamics at UNDE, we used the Data Assimilation Linked Ecosystem Carbon model version two (DALEC2) [12]. DALEC2 combines and generalizes its predecessors, DALEC evergreen (DALECev) and DALEC deciduous (DALECde) [see 43, 154, for references] to allow for modeling of mixed forest types, which is what we expect to have at UNDE. In contrast to many previous approaches that use DALEC2 in a model data fusion (MDF) framework [see 143, 93, 117, 13, for examples], we fit DALEC2 as a state space model by adding an error structure to the process model equations. This approach allows us to partition between stochasticity in the evolution of carbon through time and measurement error associated with field observations.

DALEC2 models the evolution of carbon stored in six different pools of a forest at a daily time resolution. These pools, also referred to as stocks , include: foliage carbon ($C_f^{(t)}$); above ground woody stem and coarse root carbon ($C_w^{(t)}$); fine root carbon ($C_r^{(t)}$); litter carbon ($C_{lit}^{(t)}$); long term soil organic matter carbon ($C_{som}^{(t)}$); and rapid turnover soil organic matter carbon ($C_{lab}^{(t)}$). The evolution of these six carbon stocks are governed by 19 different process parameters. We note that our model differs slightly from the traditional DALEC2 model in the sense that we have reparameterized components of the soil and litter carbon. We added an ag_2bg parameter to separate the above ground and below ground plant biomass models, so that above ground biomass can be directly compared to observations. The notation, interpretations, and units of the parameters are detailed in Table 7.1. In addition to the process parameters, DALEC2 also requires daily driver variables ($\mathbf{D}^{(t)}$). These driver variables include daily minimum and maximum temperatures, daily shortwave radiation, and daily atmospheric CO_2. To estimate the gross photosynthetic production (GPP; denoted by $G(\cdot)$ in subsequent equations), DALEC2 uses the Aggregrated Canopy Model [ACM; 153] with the daily meteorological drivers. For our ACM calculations we followed the procedure of [43], and fixed all of the internal ACM parameters except for c_{eff}. This approach differs

Param.	Description	Units
$f_a^\dagger$	Proportion of GPP allocated to autotrophic respiration	unitless
$f_{lab}^\dagger$	Proportion of GPP allocated to labile carbon	unitless
$f_f^\dagger$	Proportion of GPP allocated to foliage carbon	unitless
$f_r^\dagger$	Proportion of GPP allocated to fine root carbon	unitless
$f_w^\dagger$	Proportion of GPP allocated to wood carbon	unitless
θ_w	Daily turnover rate of wood	day^{-1}
θ_r	Daily turnover rate of roots	day^{-1}
θ_{som}	Daily turnover rate of soil organic matter	day^{-1}
θ_{tot}	Sum of daily litter turnover and mineralization	day^{-1}
θ_p	Proportion of θ_{tot} allocated to litter turnover	unitless
Θ	Parameter for temperature dependent rate parameter	$^\circ\text{C}^{-1}$
d_o	Start day of leaf regrowth onset	unitless
d_f	Start day of leaf fall	unitless
c_{eff}	Canopy efficiency	unitless
c_{lma}	Leaf mass per area	gCm^{-2}
c_{lf}	Proportion of leaves lost annually	unitless
c_{ro}	Length of labile carbon release period	day
c_{rf}	Length of leaf fall period	day
ag_2bg	Proportion of woody carbon above ground	unitless

Table 5.1: Information on the notation, physical interpretations, and units for each of the nineteen process parameters in our reparameterization of DALEC2. Parameter descriptions for our reparameterization of DALEC2 are adapted from [12]. † – these parameters obey a simplex constraint, i.e. $f_a + f_{lab} + f_f + f_r + f_w = 1$, with each component non-negative.

slightly from the ACM approach in [12] which estimates both c_{lma} and c_{eff}. This was due to large mismatches in data streams to constrain foliage carbon. We fixed c_{lma} at a value of 75, based on empirical results from [134] that were calibrated at UNDE.

In addition to modeling the stocks of carbon components within a forest ecosystem, DALEC2 also models fluxes. Stocks at time t represent the total accumulation of carbon for a component, while fluxes at time t represent different ecosystem processes that move carbon into the system, out of the system, and between different stocks within the system. In DALEC2, the processes that are modeled by fluxes include photosynthetic allocation (A), transfer to another stock (D), turnover (L), and respiration (R). The functional forms of the fluxes included in DALEC2 are:

$$R_a^{(t)} = G(\mathbf{D^{(t)}}, c_{lma}, c_{eff}) f_a \tag{5.1}$$

$$A_f^{(t)} = G(\mathbf{D^{(t)}}, c_{lma}, c_{eff}) f_f \tag{5.2}$$

$$A_r^{(t)} = G(\mathbf{D^{(t)}}, c_{lma}, c_{eff}) f_r \tag{5.3}$$

$$A_w^{(t)} = G(\mathbf{D^{(t)}}, c_{lma}, c_{eff}) f_w \cdot ag_2bg \tag{5.4}$$

$$A_{lab}^{(t)} = G(\mathbf{D^{(t)}}, c_{lma}, c_{eff}) f_{lab} \tag{5.5}$$

$$L_f^{(t)} = \Phi_f(t, d_f, c_{rf}, c_{lf}) C_f^{(t-1)} \tag{5.6}$$

$$L_w^{(t)} = \theta_w C_w^{(t-1)} \tag{5.7}$$

$$L_r^{(t)} = \theta_r C_r^{(t-1)} \tag{5.8}$$

$$L_{lab}^{(t)} = \Phi_o(t, d_o, c_{ro}) C_{lab}^{(t-1)} \tag{5.9}$$

$$R_{lit}^{(t)} = \theta_{tot}\theta_p \exp(\Theta \bar{T}^{(t)}) C_{lit}^{(t-1)} \tag{5.10}$$

$$R_{som}^{(t)} = \exp(\Theta \bar{T}^{(t)}) \theta_{som} C_{som}^{(t-1)} \tag{5.11}$$

$$D_{lit}^{(t)} = \theta_{tot}(1-\theta_p) \exp(\Theta \bar{T}^{(t)}) C_{lit}^{(t-1)}. \tag{5.12}$$

$\bar{T}^{(t)}$ is defined as the average temperature for day t. $\Phi_o(\cdot)$ and $\Phi_f(\cdot)$ are functions that govern the transfer of labile carbon to foliage carbon during the period of leaf regrowth ($\Phi_o(\cdot)$) and the transfer of foliage carbon to litter carbon during the leaf fall period ($\Phi_f(\cdot)$). Their definitions are given by:

$$\Phi_f(t, d_f, c_r, c_{lf}) = \sqrt{\frac{2}{\pi}} \cdot \frac{-\log(1-c_{lf})}{c_{rf}} \exp\left(-\left(\sin\left(\frac{t - d_f + \psi_f}{s} \right) \frac{\sqrt{2}s}{c_{rf}} \right)^2 \right),$$

$$\Phi_o(t, d_o, c_{ro}) = \sqrt{\frac{2}{\pi}} \left(\frac{6.9088}{c_{ro}} \right) \exp\left(-\left(\sin\left(\frac{t - d_o + .6245 c_{ro}}{s} \right) \frac{\sqrt{2}s}{c_{ro}} \right)^2 \right),$$

where $s = 365.25/\pi$ and $\psi_f = -\sqrt{W_0\left((2\pi \log(1-c_{lf})^2)^{-1}\right)}/\sqrt{2}$, where W_0 is the principal branch of the Lambert W function [81].

With the functional forms of the fluxes defined (Equations 5.1 – 5.12), we can construct the update equations for the DALEC2 process model using a first principles approach. A given carbon stock at time t, $C_s^{(t)}$, $s \in \{f, w, r, lit, som, lab\}$ can be written as its value at time $t-1$ minus the carbon lost from respiration (carbon leaving the system) or the carbon transferred to another stock from stock s (carbon remaining in the system but moving to a different location), plus the carbon gained from photosynthetic allocation (carbon entering the system) or carbon being transferred to stock s from another stock. This can be written mathematically as:

$$\mathbb{E}[C_s^{(t)}|\mathbf{C}^{(t-1)}] = C_s^{(t-1)} - L_s^{(t-1)} - R_s^{(t-1)} + A_s^{(t-1)} \pm D_s^{(t-1)}. \tag{5.13}$$

Using this process model skeleton combined with Equations 5.1 – 5.12, the set of DALEC2

update equations that specify the evolution of carbon through time can be written as:

$$\begin{aligned}\mathbb{E}[C_f^{(t)}] &= C_f^{(t-1)} - L_f^{(t-1)} + A_f^{(t-1)} \\ &= C_f^{(t-1)} - \Phi_f(t, d_f, c_{rf}, c_{lf})C_f^{(t-1)} + \Phi_o(t, d_o, c_{ro})C_{lab}^{(t-1)} + G(\mathbf{D^{(t)}}, c_{lma}, c_{eff})f_f \end{aligned} \tag{5.14}$$

$$\begin{aligned}\mathbb{E}[C_r^{(t)}] &= C_r^{(t-1)} - L_r^{(t-1)} + A_r^{(t-1)} \\ &= C_r^{(t-1)} - \theta_r C_r^{(t-1)} + G(\mathbf{D^{(t)}}, c_{lma}, c_{eff})f_r \end{aligned} \tag{5.15}$$

$$\begin{aligned}\mathbb{E}[C_w^{(t)}] &= C_w^{(t-1)} - L_w^{(t-1)} + A_w^{(t-1)} \\ &= C_w^{(t-1)} - \theta_w C_w^{(t-1)} + ag_2bg \cdot G(\mathbf{D^{(t)}}, c_{lma}, c_{eff})f_w \end{aligned} \tag{5.16}$$

$$\begin{aligned}\mathbb{E}[C_{lit}^{(t)}] &= C_{lit}^{(t-1)} - R_{lit}^{(t-1)} - D_{lit}^{(t-1)} + A_{lit}^{(t-1)} \\ &= C_{lit}^{(t-1)} - \left[\theta_{tot}\theta_p \exp(\Theta \bar{T}^{(t)}) + \theta_{tot}(1-\theta_p)\exp(\Theta \bar{T}^{(t)})\right] C_{lit}^{(t-1)} \\ &+ \left[\Phi_f(t, d_f, c_{rf}, c_{lf})C_f^{(t-1)} + \theta_r C_r^{(t-1)}\right] \end{aligned} \tag{5.17}$$

$$\begin{aligned}\mathbb{E}[C_{som}^{(t)}] &= C_{som}^{(t-1)} - R_{som}^{(t-1)} + D_{som}^{(t-1)} + A_{som}^{(t-1)} \\ &= C_{som}^{(t-1)} - \theta_{som}\exp(\Theta \bar{T}^{(t)} C_{som}^{(t-1)} + \frac{\theta_w C_w^{(t-1)})}{ag_2bg} + \theta_{tot}\theta_p \exp(\Theta \bar{T}^{(t)})C_{lit}^{(t-1)} \end{aligned} \tag{5.18}$$

$$\begin{aligned}\mathbb{E}[C_{lab}^{(t)}] &= C_{lab}^{(t-1)} - L_{lab}^{(t-1)} + A_{lab}^{(t-1)} \\ &= C_{lab}^{(t-1)} - \Phi_o(t, d_o, c_{ro})C_{lab}^{(t-1)} + G(\mathbf{D^{(t)}}, c_{lma}, c_{eff})f_{lab}. \end{aligned} \tag{5.19}$$

There are two additional quantities of importance that we will now define using the stocks and fluxes from DALEC2. The first quantity is net ecosystem exchange (NEE), which is the difference of autotrophic respiration and NPP. NEE is typically measured using eddy covariance techniques on data collected from flux towers [8, 7]. Since flux towers are taking measurements of ecosystem exchange processes multiple times per day, these measurements can be aggregated up to daily measurements. This makes NEE a frequently available data product for NEON and other ecosystem observation networks [100]. The second quanity is the leaf area index (LAI). LAI is a dimensionless quantity that measures the amount of

foliage in the canopy at a particular time. Moderate Resolution Imaging Spectroradiometers (MODIS) provides semi-continuous time series of reflectance measurements using satellite imaging. These measurements of reflectance are then used as model input to generate estimates of LAI at four and eight day intervals at 1km resolution [156]. The functional forms of NEE and LAI are written as:

$$\text{LAI}^{(t)} = \frac{C_f^{(t)}}{c_{lma}} \tag{5.20}$$

$$\text{NEE}^{(t)} = (f_a - 1)G(\mathbf{D}^{(\mathbf{t})}, c_{lma}, c_{eff}) + (\theta_{som}C_{som}^{(t)} + \theta_{tot}(1-\theta_p)C_{lit}^{(t)})\exp(\Theta\bar{T}^{(t)}). \tag{5.21}$$

NEE and LAI are important because their observations come at a much higher frequency than direct field measurements of the carbon stocks. Time series data of LAI help to constrain C_f, the most dynamic stock in DALEC2, and also provide additional information on the parameters that govern the C_f update equation (Equation 5.14). NEE measurements are a function of the dynamics of all carbon pools [12], but isolating the net process from the individual components corresponding to these processes is its own challenge.

5.2.3 State Space Model

The six carbon stocks in DALEC2 are fit as a state space model [6, 116, 40]. Let $\mathbf{C}^{(t)} = [C_f^{(t)}, C_r^{(t)}, C_w^{(t)}, C_{lit}^{(t)}, C_{som}^{(t)}, C_{lab}^{(t)}]^T$ denote the vector of carbon stocks at time t. Then, we can rewrite the DALEC2 process model equations (Equations 5.14 - 5.19) from the previous section using matrix notation as:

$$\mathbb{E}[\mathbf{C}^{(t)}|\mathbf{C}^{(t-1)}] = M_t\mathbf{C}^{(t-1)} + \mathbf{P}^{(t)}, \text{ where} \tag{5.22}$$

$$M_t = \begin{bmatrix} (1-\Phi_f^{(t)}) & 0 & 0 & 0 & 0 & \Phi_o^{(t)} \\ 0 & (1-\theta_r) & 0 & 0 & 0 & 0 \\ 0 & 0 & (1-\theta_w) & 0 & 0 & 0 \\ \Phi_f^{(t)} & \theta_r & 0 & (1-\beta_1^{(t)}) & 0 & 0 \\ 0 & 0 & \theta_w a g_2 b g^{-1} & \beta_1^{(t)} & (1-\beta_2^{(t)}) & 0 \\ 0 & 0 & 0 & 0 & 0 & (1-\Phi_o^{(t)}) \end{bmatrix}, \mathbf{P}^{(t)} = \begin{bmatrix} G(\mathbf{D^{(t)}}, c_{lma}, c_{eff}) f_f \\ G(\mathbf{D^{(t)}}, c_{lma}, c_{eff}) f_r \\ G(\mathbf{D^{(t)}}, c_{lma}, c_{eff}) f_w \cdot a g_2 b g \\ 0 \\ 0 \\ G(\mathbf{D^{(t)}}, c_{lma}, c_{eff}) f_{lab} \end{bmatrix}, \tag{5.23}$$

where $\beta_1^{(t)} = \theta_{tot}\theta_p \exp(\Theta \bar{T}^{(t)}) + \theta_{tot}(1-\theta_p)\exp(\Theta \bar{T}^{(t)})$ and $\beta_2 = \theta_{som}\exp(\Theta \bar{T}^{(t)}$.

The first component of our SSM is the initial condition distribution, $p_\theta(\mathbf{C}^{(0)})$. The initial condition distribution represents a probability distribution for the starting values of our process model. We chose a multivariate normal distribution, given by:

$$\begin{aligned} \mathbf{C}^{(0)} &\sim \text{MVN}(\mu_0, \Sigma_0), \\ \mu_0 &= [C_f^{(0)}, C_r^{(0)}, C_w^{(0)}, C_{lit}^{(0)}, C_{som}^{(0)}, C_{lab}^{(0)}]^T \\ \Sigma_0 &= \text{Diag}(\sigma_{f,0}^2 (C_f^{(0)})^2, \sigma_{r,0}^2 (C_r^{(0)})^2, \sigma_{w,0}^2 (C_w^{(0)})^2, \sigma_{lit,0}^2 (C_{lit}^{(0)})^2, \sigma_{som,0}^2 (C_{som}^{(0)})^2, \sigma_{lab,0}^2 (C_{lab}^{(0)})^2). \end{aligned} \tag{5.24}$$

Four of the entries of μ_0 were fixed: $C_f^{(0)}, C_r^{(0)}, C_w^{(0)}$, and $C_{som}^{(0)}$. This is because data from NEON was available to provide good initial guesses for their starting values. The remaining two entries, $C_{lit}^{(0)}$ and $C_{lab}^{(0)}$, were estimated. $\sigma_{s,0}^2$, where $s \in \{f, r, w, lit, som, lab\}$ are fixed parameters that govern the variance of the initial condition distribution.

The second component that we need to build our SSM is a stochastic evolution function $f_\theta(\mathbf{C}^{(t)}|\mathbf{C}^{(t-1)})$ that moves the latent carbon stocks through time given the set of process parameters (θ). A key assumption about SSMs is that this process evolution must be first order Markov; the value of the current state should depend only on the value of the state that directly precedes it. We model the stochastic evolution of our latent carbon states through time using a multivariate lognormal moment matching approach to embed the DALEC2

process model (this work, Chapter 4).

$$f(\mathbf{C}^{(t)}|\mathbf{C}^{(t-1)}) \sim \text{MVLN}\left(\log\left(M_t\mathbf{C}^{(t-1)} + \mathbf{P}^{(t)}\right) - .5\log(\mathbf{1}\mathbf{1}^T + \Omega)\mathbf{1}, \log(\mathbf{1}\mathbf{1}^T + \Omega)\right), \quad (5.25)$$

where M_t and $\mathbf{P}^{(t)}$ take their values from Equation 5.23, and $\mathbf{1}$ denotes the 6×1 one vector, and $\Omega = \text{Diag}(\omega_f^2, \omega_r^2, \omega_w^2, \omega_{lit}^2, \omega_{som}^2, \omega_{lab}^2)$ is a matrix that governs the density dependent variance of the process evolution function. Each diagonal component of Ω is estimated. The resulting distribution has a conditional expectation and conditional variance given by:

$$\mathbb{E}[f(\mathbf{C}^{(t)}|\mathbf{C}^{(t-1)})] = M_t\mathbf{C}^{(t-1)} + \mathbf{P}^{(t)} \quad (5.26)$$

$$\mathbb{V}[f(\mathbf{C}^{(t)}|\mathbf{C}^{(t-1)})] = (M_t\mathbf{C}^{(t-1)} + \mathbf{P}^{(t)})^T \cdot \Omega \cdot (M_t\mathbf{C}^{(t-1)} + \mathbf{P}^{(t)}). \quad (5.27)$$

The lognormal moment matching embedding of DALEC2 used here was designed such that the stochastic evolution is unbiased at each timestep (Equation 5.26) and the density dependent evolution variance scales quadratically with the DALEC2 process model predictions (Equation 5.27).

The third component that we need to build our SSM is an observation density function $g_C(\cdot)$ to relate our latent carbon stocks to their respective field measurements. Let $\mathbf{C}_{obs}^{(t)}$ be a (possibly incomplete) vector of observations, $\mathbf{C}_{obs}^{(t)} = [C_{f,obs}^{(t)}, C_{r,obs}^{(t)}, C_{w,obs}^{(t)}, C_{lit,obs}^{(t)}, C_{som,obs}^{(t)}, C_{lab,obs}^{(t)}]^T$. We construct the observation density function using a multivariate lognormal moment matching approach, which gives us:

$$g_C(\mathbf{C}_{obs}^{(t)}|\mathbf{C}^{(t)}) \sim \text{MVLN}\left(\log\left(\mathbf{C}^{(t)}\right) - .5\log(\mathbf{1}\mathbf{1}^T + \Sigma)\mathbf{1}, \log(\mathbf{1}\mathbf{1}^T + \Sigma)\right), \quad (5.28)$$

where $\Sigma = \text{Diag}(\sigma_f^2, \sigma_r^2, \sigma_w^2, \sigma_{lit}^2, \sigma_{som}^2, \sigma_{lab}^2)$ is a fixed matrix that governs the measurement error of the carbon observations. The conditional expectation and variance of $g_C(\cdot)$ are given

by:

$$\mathbb{E}[g_C(\mathbf{C}_{obs}^{(t)}|\mathbf{C}^{(t)})] = \mathbf{C}^{(t)} \tag{5.29}$$

$$\mathbb{V}[g_C(\mathbf{C}_{obs}^{(t)}|\mathbf{C}^{(t)})] = (\mathbf{C}^{(t)})^T \cdot \Sigma \cdot (\mathbf{C}^{(t)}). \tag{5.30}$$

Similar to the evolution function, $f_\theta(\cdot)$, our observation model was embedded into $g_\theta(\cdot)$ such that the carbon measurements are unbiased conditioned on the latent states, with a density dependent measurement error variance that scales quadratically with the value of the latent states.

The final component of our SSM is defining auxiliary observation functions to relate the flux observations to the latent states. For LAI, NEE, and L_f, we assume the following relationships for their observation functions:

$$\mathrm{LAI}^{(t)} \sim \mathcal{N}\left(\frac{C_f^{(t)}}{c_{lma}}, \sigma_f^2 \frac{C_f^{(t)2}}{c_{lma}^2}\right) \tag{5.31}$$

$$L_f^{(t)} \sim \mathcal{N}(\Phi_f(t, d_f, c_{rf}, c_{lf}) C_f^{(t-1)}, \beta_0 + \sigma_{lf}^2 (\Phi_f(t, d_f, c_{rf}, c_{lf}) C_f^{(t-1)})^2) \tag{5.32}$$

$$\mathrm{NEE}^{(t)} \sim \mathcal{N}(\mu_{nee}, \tau_{nee}^2), \text{ with} \tag{5.33}$$

$$\mu_{nee} = (f_a - 1) G(\mathbf{D}^{(t)}, c_{lma}, c_{eff}) + \exp(\Theta \bar{T}^{(t)})(\theta_{tot}(1 - \theta_p) C_{lit}^{(t)} + \theta_{som} C_{som}^{(t)})$$

$$\tau_{nee}^2 = (.1 + \sigma_{nee}|\mu_{nee}|)^2,$$

where $\sigma_f^2, \beta_0, \sigma_{lf}^2,$ and σ_{nee} are fixed parameters that control the measurement error of LAI, litterfall, and NEE.

In order to incorporate yearly wood carbon turnover (L_w) and yearly above ground wood carbon photosynthetic allocation rates (A_w) extracted from NEON data, we introduce two

accumulator variables so that the first order Markov assumption of the SSM is not violated:

$$SL_w^{(t)} = SL_w^{(t-1)} + \theta_w C_w^{(t-1)},\ SL_w^{(0)} = 0$$
$$SA_w^{(t)} = SA_w^{(t-1)} + G(\mathbf{D}^{(t)}, c_{lma}, c_{eff}) f_w \cdot ag_2 bg,\ SA_w^{(0)} = 0.$$

Then, we assume the following relationship between the latent accumulations and the observed accumulations of turnover and allocation:

$$SL_{w,obs}^{(t)} \sim \mathcal{N}(SL_w^{(t)}, \omega_{SL}^2 (SL_w^{(t)})^2) \tag{5.34}$$
$$SA_{w,obs}^{(t)} \sim \mathcal{N}(SA_w^{(t)}, \omega_{SA}^2 (SA_w^{(t)})^2), \tag{5.35}$$

where ω_{SL}^2 and ω_{SA}^2 are fixed parameters that govern the measurement error of the observations of the turnover and allocation accumulation processes. Then, our full observation function used to relate the latent carbon stocks to their measurements, $g_\theta(\cdot)$, is a probability distribution proportional to the product of Equations 5.29, 5.31, 5.32, 5.33, 5.34, and 5.35.

5.2.4 Sequential Monte Carlo and Particle Filtering

To fit the SSM presented here, we use two different approaches based on Sequential Monte Carlo techniques [SMC; 36, 21, 37]. In the context of our problem, we use SMC methods to approximate the probability distribution of the latent carbon stocks during the first t days given the observations during the first t days. This probability distribution can be written as $p_\theta(\mathbf{C}^{(1:t)}|\mathbf{C}_{obs}^{(1:t)})$, and is referred to as a marginal filtering distribution. The θ subscript here is used as a reminder that the probability distribution of the latent states given the observations is dependent on additional process parameters. The sequential aspect comes into play by using a forwards recursion relation given by:

$$p_\theta(\mathbf{C}^{(1:t+1)}|\mathbf{C}_{obs}^{(1:t+1)}) \propto p_\theta(\mathbf{C}^{(1:t)}|\mathbf{C}_{obs}^{(1:t)})\ f_\theta(\mathbf{C}^{(t+1)}|\mathbf{C}^{(t)})\ g_\theta(\mathbf{C}_{obs}^{(t+1)}|\mathbf{C}^{(t+1)}). \tag{5.36}$$

The distributions in Equation 5.36 do not omit a closed form solution in general. However, we can use a bootstrap particle filter [52, 21] to generate samples from $p_\theta(\mathbf{C}^{(t)}|\mathbf{C}_{obs}^{(1:t)})$. These samples give us an approximation to the marginal likelihood for a fixed parameter set θ, $\mathscr{L}_\theta(\mathbf{C}_{obs}^{(1:T)})$. An outline and description for a simple particle filter algorithm for our problem is given below:

1. Take n_p draws from the initial condition distribution (Equation 5.24). These draws, $\mathbf{C}_k^{(0)}$, $k = \{1, 2, ..., n_p\}$, will be referred to as particles for the remainder of the algorithm. These particles become the starting values for the stocks in the DALEC2 process model.

2. Repeat the following process for $t = 0 : T$:

   (a) If an observation $\mathbf{C}_{obs}^{(t)}$ is available, compute the weights for each particle using the observation density function $g_\theta(\cdot)$ (Equation 5.29) using the formula:

   $$w_{k,t} = g_\theta(\mathbf{C}_{obs}^{(t)}|\mathbf{C}_k^{(t)}).$$

   These weights give us a measure of how likely the carbon measurements are given each of the n_p different particles. Next we create a vector of normalized weights, given by:

   $$p_{k,t} = \frac{w_{k,t}}{\sum_{k=1}^{n_p} w_{k,t}}.$$

   If an observation is not available at time t, particles are given uniform weights of $p_{k,t} = \frac{1}{n_p}$ and step (b) becomes optional.

   (b) Use the vector of normalized weights at time t, $p_{1:n_p,t}$, to perform a weighted bootstrap resampling of the particles $\mathbf{C}_{1:n_p}^{(t)}$. The resampled particles are denoted $\mathbf{C}_{1:n_p}^{(t)^*}$. Gordon et al. [52] show that $\mathbf{C}_{1:n_p}^{(t)^*} \sim p_\theta(\mathbf{C}^{(t)}|\mathbf{C}_{obs}^{(1:t)})$ using a result from Smith and Gelfand [140].

(c) Generate particles for time $t+1$ by sampling from the stochastic evolution function $f_\theta(\cdot)$ (Equation 5.23) using the resampled particles, $\mathbf{C}_{1:n_p}^{(t)^*}$,

$$\mathbf{C}_k^{(t+1)} \sim f_\theta(\mathbf{C}^{(t+1)}|\mathbf{C}_k^{(t)^*}),\ k = \{1, 2, ..., n_p\}.$$

These particles, $\mathbf{C}_{1:n_p}^{(t+1)}$, are samples from the distribution $p(\mathbf{C}^{(t+1)}|\mathbf{C}_{obs}^{(1:t)})$ [52].

Then, we can compute an approximation to the marginal likelihood by factoring it as $\mathscr{L}_\theta(\mathbf{C}_{obs}^{(1:T)}) = p_\theta(\mathbf{C}_{obs}^{(0)}) \prod_{t=1}^{T} p(\mathbf{C}_{obs}^{(t)}|\mathbf{C}_{obs}^{(t-1)})$ [37]. The individual components of the product can be rewritten as:

$$\begin{aligned} p_\theta(\mathbf{C}_{obs}^{(t)}|\mathbf{C}_{obs}^{(1:t-1)}) &= \int g_\theta(\mathbf{C}_{obs}^{(t)}|\mathbf{C}^{(t)}) f_\theta(\mathbf{C}^{(t)}|\mathbf{C}^{(t-1)}) p_\theta(\mathbf{C}^{(t-1)}|\mathbf{C}_{obs}^{(1:t-1)})\ d\mathbf{C}^{(t-1)} \\ &= \int g_\theta(\mathbf{C}_{obs}^{(t)}|\mathbf{C}^{(t)}) p_\theta(\mathbf{C}^{(t)}|\mathbf{C}_{obs}^{(1:t-1)})\ d\mathbf{C}^{(t)} \\ &= \mathbb{E}[g_\theta(\mathbf{C}_{obs}^{(t)}|\mathbf{C}^{(t)})] \text{ w.r.t. the measure } p_\theta(\mathbf{C}^{(t)}|\mathbf{C}_{obs}^{(1:t-1)}). \end{aligned}$$

Then, we can use the fact that the resampled particles for time t, $\mathbf{C}_{1:n_p}^{(t)^*}$, are approximate draws from $p_\theta(\mathbf{C}^{(t)}|\mathbf{C}_{obs}^{(1:t)})$. Thus we can approximate the marginal likelihood using the particle weights as:

$$\hat{\mathscr{L}}_\theta(\mathbf{C}_{obs}^{(1:T)}) = \prod_{t=1}^{T} \left(\sum_{k=1}^{n_p} \frac{w_{k,t}}{n_p} \right). \tag{5.37}$$

Our final quantities of interest from SMC methods are the smoothing distributions: $p_\theta(\mathbf{C}^{(t)}|\mathbf{C}_{obs}^{(1:T)})$. In contrast to the marginal filtering distribution, the smoothing distributions represent the density of the latent carbon stocks at time t given the entire set of measurements. We can sample from the smoothing distributions entirely using output from the particle filter [see 37, for several algorithms], which allows us to build up full empirical distributions of the latent state trajectories as we run the particle filter at little additional cost.

5.2.5 Latent State and Parameter Estimation

The first method that we used to fit our SSM was the IF2 algorithm from Ionides et al. [62]. The IF2 algorithm uses parameter perturbations alongside an iterative Bayes map to maximize a surrogate likelihood. We give details of the algorithm below, but first need to introduce a few pieces of notation borrowed from Ionides et al. [62]. Let $h_t(\theta|\sigma), t = 0, ..., T$ be a collection of densities used to perturb the parameter set θ, and let $\sigma_{1:n_{iter}}$ be the set of perturbations.

1. Generate an initial parameter vector for each of the n_p particles, $\theta^{(0)}_{1:n_p}$.

2. Repeat the following for $i = 1 : n_{iter}$

   (a) For $k = 1 : n_p$, sample a perturbed initial parameter vector from the initial perturbation density,

   $$\theta^{(0)}_{i,k} \sim h_0(\theta|\theta^{(0)}_{i-1,k}, \sigma_i).$$

   For $k = 1 : n_p$, draw samples from the initial condition distribution (Equation 5.24) conditional on $\theta^{(0)}_{i,k}$:

   $$\mathbf{C}^{(0)}_{i,k} \sim p_{\theta_{i,k}}(\mathbf{C}^{(0)}).$$

   (b) Repeat the following for $t = 1 : T$:

      i. Draw a sample from the t^{th} perturbation density, $\theta^{(t)*}_{i,k} \sim h_t(\theta|\theta^{(t-1)}_{i,k}, \sigma_m)$, for each particle.

      ii. Using these new parameter perturbations, evolve each particle using its evolution function to obtain, $\mathbf{C}^{(t)*}_{i,k} \sim f_{\theta^{(t)}_{i,k}}(\mathbf{C}^{(t)}|\mathbf{C}^{(t-1)}_{i,k})$.

      iii. If an observation is available, compute the weight of each particle using the

observation density:

$$w_{k,t} = g_{\theta_{i,k}^{(t)}}(\mathbf{C}_{obs}^{(t)} | \mathbf{C}_{i,k}^{(t)*}).$$

iv. Compute the normalized weights of each particle, given by:

$$p_{k,t} = \frac{w_{k,t}}{\sum_{k=1}^{n_p} w_{k,t}}.$$

v. Perform a weighted bootstrap resample of size n_p of the pairs $\{\theta_{i,k}^{(t)*}, \mathbf{C}_{i,k}^{(t)}*\}$ with weights $w_{k,t}$. These weighted resamples are referred to as $\theta_{i,1:n_p}^{(t)}$ and $\mathbf{C}_{i,1:n_p}^{(t)}$, respectively.

(c) Set the parameter vectors for each particle to be their value at the final timestep:

$$\theta_{i,k} = \theta_{i,k}^{(T)}.$$

3. Final result: $\theta_{1:n_{iter},1:n_p}$

The output of the IF2 algorithm is a particle swarm of the parameter vector θ, which [62] show is concentrated around the maximum likelihood estimate for sufficiently small perturbations.

To perform a global optimization using the IF2 algorithm, initial parameter values were randomly drawn from a hypercube of acceptable values. IF2 was then run for 150 iterations, using 2,000 particles, and a multivariate normal random walk perturbation density with a geometric cooling fraction of 0.5. Parameter values at the final IF2 iteration were then used to compute five particle filter approximations to the marginal log-likelihood, using 10,000 particles each. This process was repeated a total of 600 times, and initial starting values, time series of the 150 IF2 iterations, and the five marginal likelihood approximations were saved for each run.

The second method that we used to fit our SSM was Particle Markov Chain Monte Carlo

[PMCMC; 42, 3], in particular we use the Particle Marginal Metropolis Hastings algorithm (PMMH) from Andrieu et al. [3]. The PMMH algorithm uses a particle filter to jointly propose a candidate latent state and parameter set, and then uses a Metropolis-Hastings step [99, 54] where the evaluation of the likelihood is replaced with the particle filter approximation of the marginal likelihood (Equation 5.37). An outline of the PMMH algorithm is given below. We introduce the notation $q(\cdot)$ as the proposal distribution, $p(\theta)$ as the prior distribution over the parameter set θ, and n_{iter} as the number of PMCMC iterations.

1. Initialize θ_0, a vector of starting values for the parameter set θ that will be estimated. Use an SMC algorithm with n_p particles to generate a sample $\mathbf{C}_0^{(1:T)}$. Compute the initial marginal likelihood estimate for θ_0, $\mathscr{L}_{\theta_0}(\mathbf{C}_{obs}^{(1:T)})$.

2. Repeat the following for $i = 1 : n_{iter}$

   (a) Generate a proposed value by sampling $\theta^* \sim q(\cdot|\theta_{(i-1)})$.

   (b) Use an SMC algorithm with n_p particles to generate a sample $\mathbf{C}_*^{(1:T)}$ and its marginal likelihood approximation, $\mathscr{L}_{\theta^*}(\mathbf{C}_{obs}^{(1:T)})$.

   (c) Sample $Y \sim \text{Unif}(0, 1)$. If $Y \leq \frac{p(\theta^*)\mathscr{L}_{\theta^*}(\mathbf{C}_{obs}^{(1:T)})}{p(\theta_{(i-1)})\mathscr{L}_{\theta_{(i-1)}}(\mathbf{C}_{obs}^{(1:T)})} \cdot \frac{q(\theta_{(i-1)}|\theta^*)}{q(\theta^*|\theta_{(i-1)})}$, set $\theta_{(i)} = \theta^*$ and $\mathbf{C}_{(i)}^{(1:T)} = \mathbf{C}_*^{(1:T)}$. Otherwise, set $\theta_{(i)} = \theta_{(i-1)}$ and $\mathbf{C}_{(i)}^{(1:T)} = \mathbf{C}_{(i-1)}^{(1:T)}$.

Andrieu et al. [3] show that this algorithm converges to sampling from the posterior distribution independent of the number of particles used.

In contrast to standard MCMC methods, which aim to tune the proposal distribution to increase the efficiency of the algorithm, the PMMH has two items that require tuning: the proposal distribution and the number of particles. While theoretical results for choosing the number of particles exist in the literature [e.g. 38, 118, 136], many of them carry heavy assumptions about the behaviour of $\hat{\mathscr{L}}_\theta$ and do not work well in practice. For our analysis, we ran ten chains of PMCMC for 10,000 iterations each, with an adaptive normal random

walk proposal [128], with 4,000 particles. PMMH can be slow to mix for the amount of computation power it requires [62]. To rectify this, we follow the approach of Ionides et al. [62], and start our PMCMC chains at the best parameter set obtained from the 600 IF2 runs. This ensures that our chains start in an area of high posterior density, so that we can make efficient use of computational resources.

The SSM presented here was fit using the `pomp` package [76] for `R` [123], using the `mif2` and `pmcmc` functions. To decrease computation time as much as possible, all models were implemented in `C` using `pomp`'s `Csnippet` functionality.

5.2.6 Prediction Evaluation

To test the predictive performance of our SSM, our model was fit to 1800 days of data from October 3rd, 2016 to September 6th, 2021. The endpoint, September 6th, was chosen because it was the day that the final datapoint on Woody carbon was measured. We held out 150 day time series data of LAI and NEE from September 7th, 2021 to February 2nd, 2022 for prediction validation.

To evaluate our model predictions, we used strictly proper scoring rules [48]. Generally, scoring rules measure the accuracy of probabilistic forecasts, where lower scores within the same scoring rule indicate better forecast performance. Scoring rules are proper if their expected values are maximized by draws from the true predictive distribution, and are said to be strictly proper if this maximum is unique [48]. We assessed our forecasts with two different strictly proper scoring rules: the Continuous Ranked Probability Score [CRPS; 97] and the Ignorance score [IGN; 51]. To define CRPS and IGN, we first introduce some notation. Let $f(\cdot)$ be the probability density function for a forecast event, let $F(\cdot)$ be the cumulative density function associated with $f(\cdot)$, and let y be a realization of the event.

Then, the forms of CRPS and IGN are given by:

$$\text{CRPS}(y) = \int_{\mathbb{R}} (F(z) - \mathbb{1}_{z \geq y})^2 dz; \tag{5.38}$$

$$\text{IGN}(y) = -\log(f(y)). \tag{5.39}$$

To compute CRPS and IGN scores, we used Monte Carlo output of the smoothing distributions of the latent states [80], and used the `crps_sample` and `logs_sample` functions from the `scoringRules` R package [68].

5.3 Results

We performed an IF2 [62] global search to find suitable parameter values. These parameter values were used to generate estimates of the latent states using a Monte Carlo approach with the particle filter. We show these results, out of sample predictive validation, and a discussion of our findings. We found that the PMMH algorithm [3] did not work well for our problem, and did not generate parameter estimates with a large enough effective sample size to be used for inference and out of sample predictive validation. We provide a dicussion of these findings, and why the PMMH algorithm may have performed poorly for our problem.

5.3.1 IF2 Results

To perform a global search on our parameter space, we ran the IF2 algorithm 600 times from diffuse starting values. To compute Monte Carlo confidence intervals for the six latent carbon stocks, we took 1,000 draws from the particle smoothing distribution with 10,000 particles using the best set of parameters found during our global IF2 search.

We found that the woody carbon (C_w) (Figure 5.2, top left panel) was relatively well con-

strained with six observations, where the estimates of the latent above-ground woody biomass all agreed with observations. The woody carbon is also an excellent demonstration of the accumulation of uncertainty in between observations. In our density dependent SSM, the ω parameters that control the process stochasticity have the interpretation of governing the percent error of the process model at each time step. Despite the IF2 estimate of ω_w being relatively low ($\hat{\omega}_w = .013$), and all of the observations being over 7000 gCm^{-2}, the 90% confidence intervals of the smoothing distributions still predict that woody carbon can drop as low as 4,500 gCm^{-2}. This accumulation of process variance can also be seen to a larger in the soil organic matter carbon pool (Figure 5.2, bottom right panel). The soil carbon uncertainty accumulates and grows over the fitting period despite having regular measurements of NEE to help to constrain it.

While our SSM did a reasonable job of capturing the relationship between foliage carbon and LAI dynamics, it also predicted large spikes in the latent foliage carbon (Figure 5.2, top right panel). These spikes occur during the middle of every year, and conincide with the start of the leaf regrowth period, where Labile carbon (Figure 5.2, bottom right panel) is transferred to foliage. This is likely because we have no direct data sources to constrain the labile carbon – it is only indirectly constrained through measurements of LAI, litterfall, and NEE. The labile carbon (C_{lab}) latent state estimates (Figure 5.2, bottom right panel) also illustrate the importance of using a density dependent lognormal SSM. The labile carbon shows a large uncertainty in the fall and winter months, and very little uncertainty in the summer and spring months. This type of dynamic would not be possible with a constant variance Gaussian error structure, which either require the process variance to be small in the fall and winter months, or to be large in the spring and summer months, which runs the risk producing negative latent state estimates for the labile carbon. Indeed, even Gaussian density dependent models may produce latent state estimates below zero in the spring and summer months, depending on the magnitude of the labile carbon process variance parameters.

Litter carbon uncertainty (Figure 5.2, middle left panel) rapidly increased from the initial

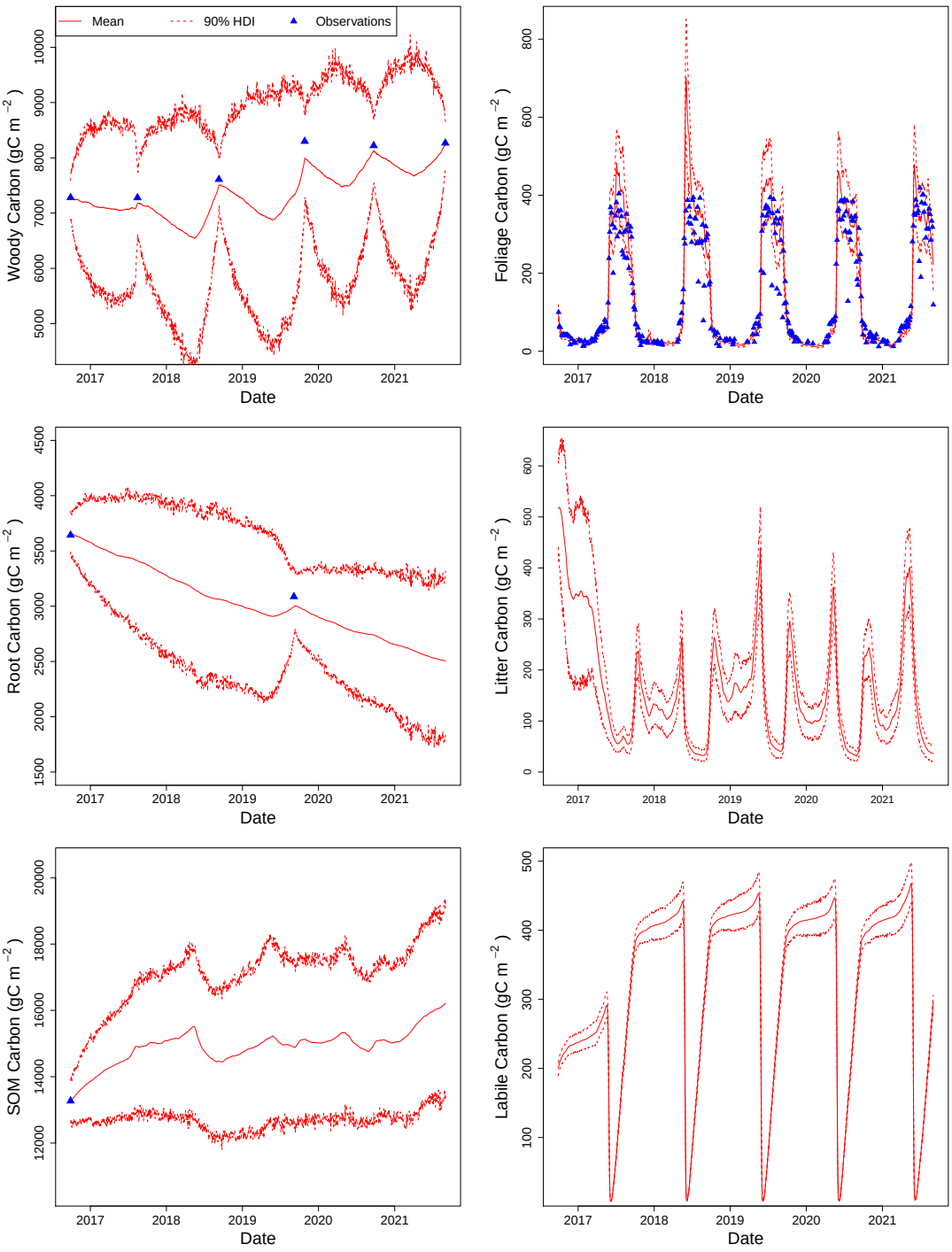

Figure 5.2: Time series plots of the six latent states from our state space model. Means and 90% highest density intervals were estimated using a Monte Carlo approach, by running 1,000 particle filters using the parameter values obtained from our global IF2 search using 10,000 particles. Blue triangles denote observed measurements of the carbon stocks from NEON data and MODIS LAI.

conditions until NEE data streams became available to help constrain it. Once the NEE data came online, it was able to decrease the uncertainty in the estimates for the latent litter carbon, despite the high amounts of noise in the NEE observations. The root carbon (C_r) latent state estimates (Figure 5.2, middle left panel) had a relatively small uncertainty, but show an unrealistic downwards trajectory. This is not surprising to us, as the root carbon only had two observation: one used for the initial condition, and one in 2019 that was almost 600 gCm^{-2} smaller than the initial condition we used, and no additional sources of data to constrain it. We expect that incorporating additional NEON root carbon measurements in the future will help improve the unrealistic trajectory that we see here.

A table of final parameter values from the model IF2 global search is included in Appendix A.

5.3.2 IF2 Prediction Evaluation

The secondary objective of our study was to assess the forecasting performance of our model for two quantities: LAI and NEE. To do this, we fit the model to 1800 days of data using the IF2 algorithm, and held out 150 day time series of LAI and NEE for out of sample validation. Using the parameter estimates obtained from our IF2 global search, we ran the model forwards using a particle filter and used the observation models to create Monte Carlo confidence intervals for our predictions.

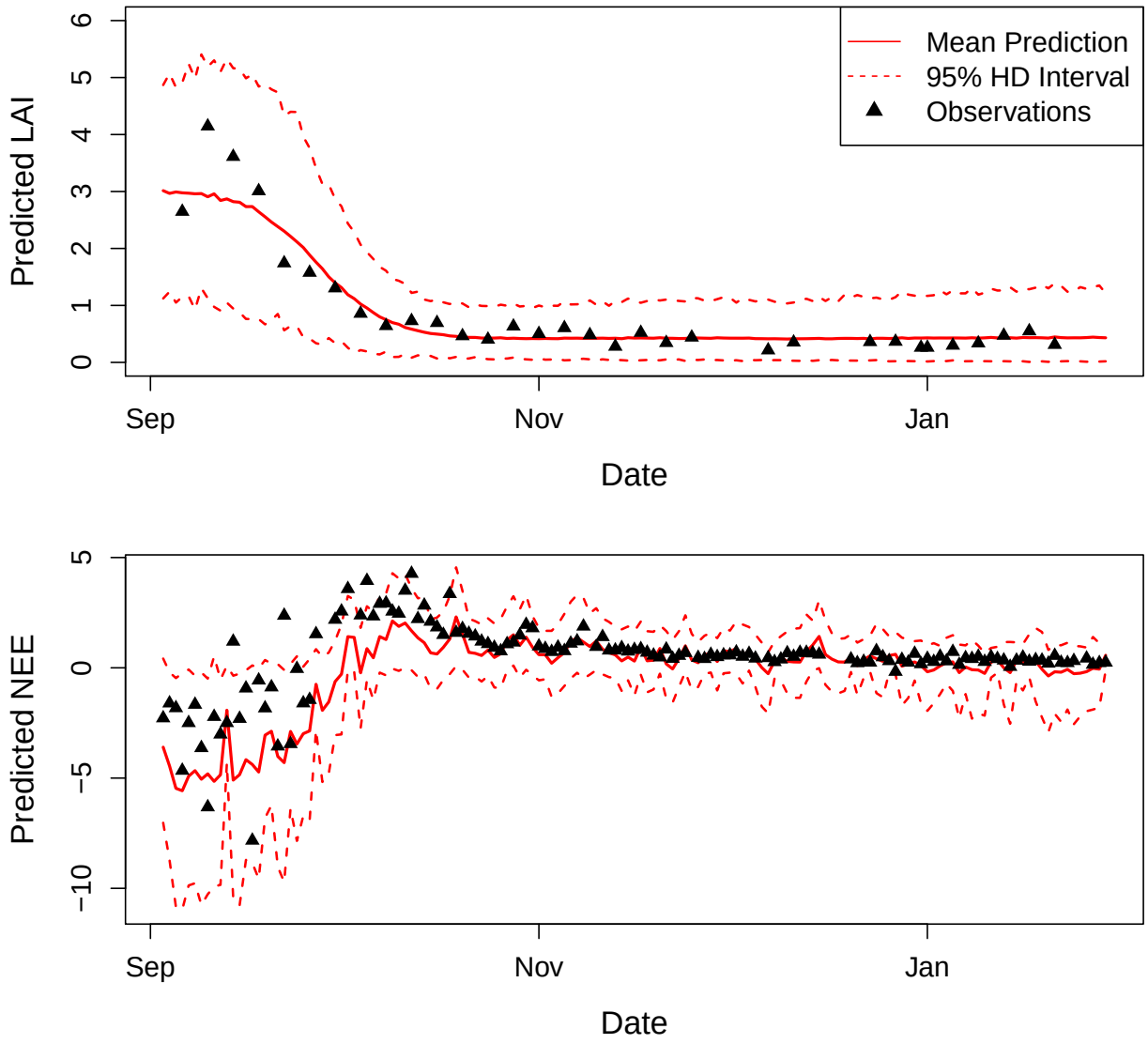

Figure 5.3: Predictions of LAI (top) and NEE (bottom) generated by our SSM using the best parameter values from the IF2 global search. The solid red lines denote the mean from 1,000 draws of the predictive distribution using 10,000 particles, and dashed red lines denote the bounds of the 95% highest density intervals. Black triangles denote observations of the held out LAI and NEE data. Prediction dates range from September 7th, 2021 to February 2nd, 2022

We found that our model had good predictive performance on the held out time series of LAI. The mean predictions of our Monte Carlo estimates are far from the LAI observations in September, but have large 95% confidence interval bounds due to the large measurement error associated with the LAI observations. Our mean predictions fall close to the downwards

trajectory seen in October, and accurately capture the leveling off of the LAI in the winter months. This is reflected in the evaluation for our predictions (Figure 5.4, top panels), where the CRPS and IGN scores are large at first and then decrease and level off as the forecast horizon increases. Our density dependent variance structure for the latent process and observation models allow the uncertainty of our LAI predictions to grow slowly in the winter months, where the DALEC2 process model is relatively confident that there will not be large increases in foliage carbon due to leaf regrowth.

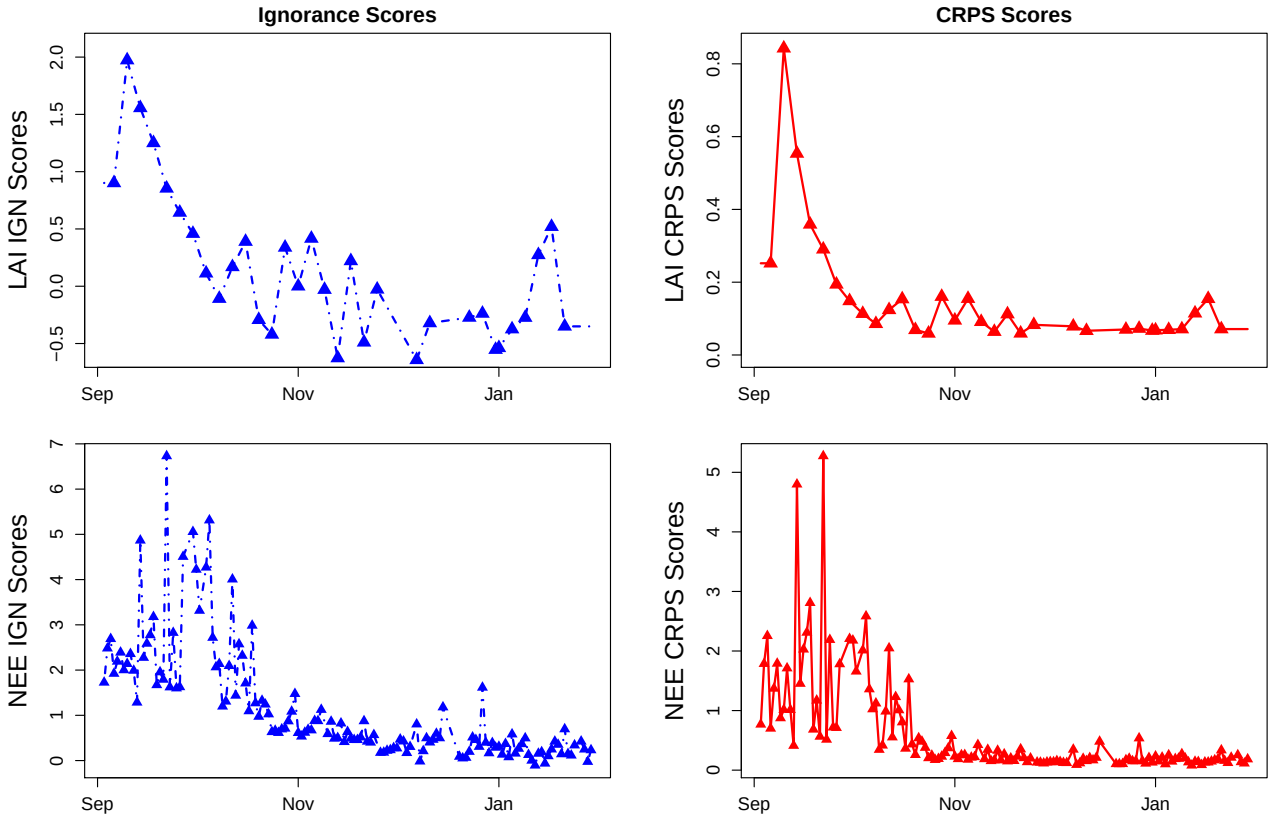

Figure 5.4: Ignorance scores (triangles; left side) and CRPS scores (triangles; right side) for the LAI predictions (top) and NEE predictions (bottom). The dashed and solid lines were used to connect between points in the absence of observations, and were obtained by using a linear interpolation between realizations of IGN and CRPS scores.

We found that out model had a difficult time capturing the NEE dynamics at the start of the prediction period. From the start of the prediction period in September until mid-way through October, our model consistently underestimated the out of sample NEE data.

However, our model was able to accurately capture the stabilization of the out of sample NEE data from November through January. We believe that the underestimation of our model at the start of the time period may be due to biases in the process parameters. In the winter months the NEE tends to stay close to zero, since it is a function of the average temperature. Our observation model for NEE, Equation 5.33, has a variance that is a linear function of the DALEC2 NEE predictions. This may be encouraging the model to fit the $\theta_{som}, \theta_{tot},$ and θ_p parameters to the NEE data in the winter months, where deviations are more harshly discouraged by our density dependent variance model.

5.3.3 PMMH Results

We found that the PMMH algorithm did not perform well for our problem. We ran ten chains of PMMH total, each for 10,000 iterations using 4,000 particles. Each chain took around 82 hours to run, and our total acceptance rate across the ten chains was .12%. This left us with effective sample sizes that were too small to reliably use for inference on the parameters and latent states. However, we acknowledge that this is a very difficult problem for PMMH. By using the PMMH formulation of particle MCMC, we are effectively proposing 11,722 parameters at once: 1950 latent states for each of the six carbon stocks, and the 22 free parameters for the stochastic formulation of the DALEC2 process model. This type of high dimensional proposal complexity, combined with our sparse observations and non-linear, non-Gaussian SSM will likely be a difficult task for any Bayesian MCMC based method.

5.4 Discussion

In ecological applications, uncertainty can be introduced from a wide variety of sources. A few of these include initial condition uncertainty, parameter uncertainty, measurement error, and process stochasticity. Different systems exhibit different levels of sensitivity to

these sources of uncertainty, with no one source of uncertainty carrying the majority of the weight across all applications [35]. This makes partitioning uncertainty between different sources a necessary endeavor to learn about our system and make informed predictions about its future behavior. In this study, we embedded the DALEC2 process model into a novel non-linear, non-Gaussian state space model that accounted for initial condition uncertainty, measurement error, and process stochasticity. Our acknowledgement and quantification of process stochasticity sets us apart from the most common technique in the terrestrial ecosystem carbon modeling literature: Bayesian MDFs [70, 12, 155].

Our study compared two different approaches for fitting our complex, computationally intensive SSM: PMMH and the IF2 algorithm. We allocated roughly the same amount of computational resources to PMMH and IF2, with 600 IF2 runs totaling roughly 900 core hours and ten chains of PMMH with 10,000 iterations totaling approximately 820 core hours. However, this is not a fair comparison in practice for two reasons. The first reason is that 600 IF2 runs took 30 hours when distributed across 30 cores, while PMMH took 82 hours to run ten chains in parallel. The second reason is that the PMMH chains were started at the values obtained from our global IF2 search. If we were to start each chain from diffuse initializations of the parameters, it is likely that these chains would need to be run for much longer than four days. These findings support the claims of Ionides et al. [62] that PMMH can be computationally inefficient compared to the IF2 algorithm. We also found that PMMH did not produce usable estimates of parameter uncertainty for our problem, accepting only .12% of proposed moves with 4,000 particles. While increasing the number of particles would help to decrease the variance of the particle filter approximation of the marginal likelihood and potentially lead to better mixing, the computation time necessary scales linearly with the number of particles used and would increase our computation time from the order of days to the order of weeks or months. With that being said, for applications that hinge on the parameter uncertainty quantification that PMMH offers, several weeks may not be unreasonable at all. Though we did not apply them here, applications

that require parameter fully Bayesian parameter uncertainty quantification with limited time may consider single iteration methods such as the Liu-West filter [92] or single simulation methods such as Approximate Bayesian Computation [ABC; 138].

We found that while our model had a strong performance for predicting LAI data, there is room for improvement in predicting NEE fluxes. Our model consistently underestimated the out of sample NEE data at the start of the prediction period before eventually leveling off and providing good predictions in the late autumn and early winter months. We hypothesized that this may be due to parameter biases for $\theta_{som}, \theta_{tot}$, and θ_p, but this claim is difficult to test without additional measurements of soil organic matter carbon to help to constrain the soil turnover parameter. It is also possible that the underestimation of the out of sample NEE data is an indication that we need to revisit our observation model. NEE was the most difficult component to build an observation model for. Our NEE data showed large measurement error, with differences between successive observations having differences as large as ± 15 gCm^{-2} in the spring and summer months, but tame behavior in the winter months when the temperature is low. In an attempt to capture this, we used a density dependent normal distribution that was a linear function of the DALEC2 NEE predictions, to still reasonably account for measurement error in the winter months. In future work we plan to analyze the effects of different NEE observation functions more rigorously, such as the Laplace distribution approach of Hollinger and Richardson [56], and see if different choices help to alleviate the underestimation problems that we saw here.

The fitting methods that we applied in this study provide interesting directions for future work. Both of the fitting methods used in this study relied on the particle filter. One of the strengths of using a particle filter to approximate the likelihood is that it does not require evaluation of the evolution function $f_\theta(\cdot)$, it only requires that we have a mechanism for simulating from it. While this property, sometimes referred to as "plug-and-play" [16, 55], is not directly beneficial to the work here, it allows for inference of carbon dynamics using models with higher complexity than DALEC2 in future work. The fact that we have

embedded the DALEC2 process model using a lognormal evolution function also means that we have a closed form Markov transition density, and thus we can seek alternative fitting methods to the particle filter. For example, since our likelihood can be written in closed form, we can use standard MCMC algorithms to estimate the parameters and latent states. This eliminates the number of particles as a tuning parameter, but introduces the need to create and tune proposal distributions for sampling the latent states.

With ecosystem carbon uptake projected to be a major player in the evolution of climate in coming decades, we propose the use of non-linear non-Gaussian state space models to enforce biologically realistic lower bounds and quantify the different sources of uncertainty in predictions of process based terrestrial ecosystem models. We identified suitable fitting methods for the sparse, nuanced data that are used to constrain these models, and demonstrate their efficacy in predicting out of sample data. With regular data streams becoming available from a number of long term observatory networks across the globe, these models will become better constrained and help us to bridge gaps in our understanding of carbon dynamics in terrestrial ecosystems. The methods we discuss here provide a flexible framework for embedding any stochastic carbon model that has a simulative mechanism using the open source `pomp` R package, making them accessible to practitioners from a variety of different backgrounds and interests.

Chapter 6

Discussion

The primary focus of this dissertation was to provide a thorough investigation into the difficulties associated with embedding mechanistic process based ecosystem models into statistical state space modeling frameworks. In this work, three of the problems that we considered were: identifiability of process parameters; trade-offs between ecological realism/model complexity and computational requirements; and finding suitable fitting methods for the limited amount of data that we have to constrain these types of models. The SSM framework was a major thread through out the three main chapters of this dissertation. This was due to the importance of quantifying the uncertainty in our process based models, which are usually treated as deterministic. As the importance of quantifying process stochasticity continues to be acknowledged [35, 88], this work can provide insight and solutions to the practical difficulties of using ecosystem state space models to quantify multiple sources of uncertainty.

In Chapter 3, we embedded the DALECev terrestrial ecosystem carbon model [43, 154] as the process model in a Gaussian SSM. This followed the approach of Thomas et al. [145], who embed a different process model into a Gaussian SSM [the 3PG model of 82]. Using DALECev, we generated synthetic datasets with varying levels of observation frequency and flux data available to constrain the processes. We used a simulation study based approach to diagnose potential problems with using the DALECev process model in an SSM framework. This simulation based approach gave us access to the ground truth parameter values that were used generate the synthetic datasets, so that we could compare model estimates of the latent states and parameters to the results obtained from our SSM. One of our objectives for

the work done in this chapter was to find a suitable method of assessing the identifiability of the process parameters. Many existing approaches to assessing identifiability can struggle with long time series (e.g. the Hessian approaches of Bulla and Berzel [20] or the symbolic algebra methods of Cole and McCrea [27]) or with many parameters (such as the profile likelihood methods of Viallefont et al. [149]), and our SSM had both of these. We found that the data cloning methods of Ponciano et al. [122] and Lele et al. [84] provided an approach to assessing identifiability that could handle many parameters, long time series data, and was easy to insert into the Bayesian Markov Chain Monte Carlo framework that we used to estimate the parameters and latent states. Using data cloning, we showed that the process parameters of DALECev may be unidentifiable when only using data that is typically available from the National Ecosystem Observatory Network (NEON).

While Chapter 4 may seem like the odd one out, it provided an important bridge between the third and fifth chapters. The idea behind embedding positive process models into lognormal distributions was largely motivated by work done in Chapter 3. When starting chains from diffuse parameter values in order to test for convergence, we found that some starting parameter values would generate negative latent states. This could create a feedback loop in the process model and cause the latent states to tend towards negative infinity. A search for suitable distributions to model positive processes led us to a wealth of literature on lognormal state space models, such as Dennis et al. [33], Patterson et al. [115], Buckland et al. [19], Mäntyniemi et al. [102], and acquainted us with the Gompertz [49] and Moran-Ricker [125] models that served as comparators in our study. The desire to embed an unbiased process evolution with arbitrary variance structure is what led to the LNM3 method that we present in Chapter 4. This work tied back to our objectives by examining the trade-off between model complexity and computation time. Using lognormal distributions instead of normal distributions allowed us to enforce the positivity constraint of our latent states and observations, making our models more ecologically realistic. However, this also eliminated the option of using Gibbs sampling [46] to update our latent states for all but one of our

LNM3 models. Instead, we had to rely on the random walk Metropolis-Hastings [99, 54] algorithm to update the latent states and parameters, which led to notable increases in the computation time when using `JAGS` [119].

In Chapter 5 we focused on finding suitable fitting methods for real data that is used to constrain these types of process based modeling approaches. We used data available from NEON, NASA, and the Scripps project to extract measurements of carbon stocks, measurements of fluxes, and meteorological driver variables at the University of Notre Dame Environmental Research Center in Wisconsin. We developed a novel lognormal SSM that embeds the DALEC2 process model [12] in a similar manner to our approach from Chapter 3, but with a density dependent variance structure using the approach from Chapter 4. We tested two different fitting methods, the Iterated Filtering algorithm of Ionides et al. [62] (IF2) and the Particle Marginal Metropolis Hastings algorithm of Andrieu et al. [3] (PMMH). We found that parameter set obtained using the IF2 algorithm had good performance for out of sample prediction of Leaf Area Index and Net Ecosystem Exchange, although we believe that there is room for improvement in future work. The PMMH algorithm did not perform well, accepting only .12% of samples across ten chains. One of the difficulties of using the IF2 algorithm is that it does not have the robust parameter uncertainty quantification that Bayesian methods offer. In future work, we plan to continue to seek Bayesian approaches that work for this problem so that we can provide fully specified uncertainties in forecasting applications where we plan to use this model.

The work done in this dissertation aims to find solutions to difficulties that arise when using process based models in state space frameworks, and make the SSM more accessible to practitioners. The work presented here can provide guidance to scientists who desire uncertainty quantification in their process based ecosystem models, which will become increasingly important as we continue to ask questions about the future that hinge on the evolution of our world's ecosystems. Indeed, while the work we present here was very statistically focused, we must remember not to lose sight of the forest ecosystem models for all of the trees. □

Chapter 7

Appendix A

7.1 Derivation of NDLM Gibbs Updates

Suppose that x is from a latent process governed by evolution function

$$x_t|x_{t-1} \sim \mathcal{N}(A_t x_{t-1} + b_t, \phi), t \in \{1, ..., T\}$$

Suppose for $I \subset \{1, ..., T\}$ there exists observation data y_I governed by the observation model

$$y_i|x_i \sim \mathcal{N}(x_i, \tau), i \in I$$

With arbitrary prior on ϕ, a fixed τ, and initial condition prior

$$x_0 \sim \mathcal{N}(\mu_0, \phi_0),$$

there exist tractable full conditionals for x_t that are derived here to be used with Gibbs sampling.

7.1.1 Initial Latent State

$$
\begin{aligned}
\pi(x_0|\cdot) &\propto \exp\Big(-\frac{\phi}{2}(x_1 - A_1x_0 - b_t)^2\Big)\exp\Big(-\frac{\phi_0}{2}(x_0-\mu_0)^2\Big) \\
&\propto \exp\left(-\frac{\phi}{2}\Big(x_1^2 - 2x_1(A_1x_0+b_1) + (A_1x_0+b_1)^2\Big) - \frac{\phi_0}{2}\Big(x_0^2 - 2x_0\mu_0\Big)\right) \\
&\propto \exp\left(-\frac{\phi}{2}\Big(A_1^2x_0^2 - 2x_0(A_1x_1 - A_1b_1)\Big) - \frac{\phi_0}{2}\Big(x_0^2 - 2x_0\mu_0\Big)\right) \\
&\propto \exp\left(-\frac{x_0^2}{2}\Big(\phi A_1^2 + \phi_0\Big) + x_0\Big(\mu_0\phi_0 + \phi A_1(x_1 - b_1)\Big)\right) \\
&\sim \mathcal{N}\left(\mu^* = \frac{\mu_0\phi_0 + \phi A_1(x_1-b_1)}{\phi A_1^2 + \phi_0}, \phi^* = \phi A_1^2 + \phi_0\right)
\end{aligned}
$$

7.1.2 Interior Latent States

No observed data

With no observed data at time t, $t \neq 0, t \neq T$, the full conditional can be written:

$$
\begin{aligned}
\pi(x_t|\cdot) &\propto \exp\Big(\frac{-\phi(x_t - A_tx_{t-1} - b_t)^2}{2}\Big)\exp\Big(\frac{-\phi(x_{t+1} - A_tx_t - b_{t-1})^2}{2}\Big) \\
&\propto \exp\left(\frac{-\phi}{2}\Big(x_t^2 - 2x_t(A_tx_{t-1}+b_t) + x_t^2A_{t+1}^2 - 2x_tA_{t+1}x_{t+1} + 2A_{t+1}b_{t+1}x_t\Big)\right) \\
&\propto \exp\left(\frac{-\phi}{2}\Big(x_t^2(1+A_{t+1}^2) - 2x_t(A_tx_{t-1} + b_t + A_{t+1}(x_{t+1} - b_{t+1}))\Big)\right) \\
&\sim \mathcal{N}\left(\mu^* = \frac{A_tx_{t-1} + b_t + A_{t+1}(x_{t+1} - b_{t+1})}{1 + A_{t+1}^2}, \phi^* = \phi(1 + A_{t+1}^2)\right)
\end{aligned}
$$

With observed data

With observed data at time t, $t \neq 0, t \neq T$, the full conditional can be written:

$$\begin{aligned}
\pi(x_t|\cdot) &\propto \exp\Big(\frac{-\phi(x_t - A_t x_{t-1} - b_t)^2}{2}\Big) \exp\Big(\frac{-\phi(x_{t+1} - A_t x_t - b_{t-1})^2}{2}\Big) \exp\Big(-\frac{\tau}{2}(x_t - y_t)^2\Big) \\
&\propto \exp\left(\frac{-\phi}{2}\Big(x_t^2 - 2x_t(A_t x_{t-1} + b_t) + x_t^2 A_{t+1}^2 - 2x_t A_{t+1} x_{t+1} + 2A_{t+1} b_{t+1} x_t\Big) - \frac{\tau}{2}(x_t^2 - 2x_t y_t)\right) \\
&\propto \exp\left(-\frac{x_t^2}{2}\Big(\phi(1 + A_{t+1}^2) + \tau\Big) + x_t\Big(\phi(A_t x_{t-1} + b_t + A_{t+1}(x_{t+1} - b_{t+1}) + \tau y_t\Big)\right) \\
&\sim \mathcal{N}\left(\mu^* = \frac{\phi(A_t x_{t-1} + b_t + A_{t+1}(x_{t+1} - b_{t+1}) + \tau y_t}{\phi(1 + A_{t+1}^2) + \tau\Big)}, \phi^* = \phi(1 + A_{t+1}^2) + \tau\right)
\end{aligned}$$

7.1.3 Final Latent State

With no observed data

$$\begin{aligned}
\pi(x_T|\cdot) &\propto \exp\Big(-\frac{\phi}{2}(x_T - A_T x_{T-1} - b_t)^2\Big) \\
&\sim \mathcal{N}\Big(\mu^* = A_T x_{T-1} + b_t, \phi^* = \phi\Big)
\end{aligned}$$

With observed data

$$\begin{aligned}
\pi(x_T|\cdot) &\propto \exp\Big(-\frac{\phi}{2}(x_T - A_T x_{T-1} - b_t)^2\Big) \exp\Big(-\frac{\tau}{2}(x_T - y_T)^2\Big) \\
&\propto \exp\left(-\frac{\phi}{2}\Big(x_T^2 - 2x_T(A_T x_{T-1} + b_T)\Big) - \frac{\tau}{2}\Big(x_T^2 - 2x_T y_T\Big)\right) \\
&\propto \exp\left(-\frac{x_T^2}{2}\Big(\phi + \tau\Big) + x_T\Big(\phi A_T x_{T-1} + b_T + \tau y_T\Big)\right) \\
&\sim \mathcal{N}\Big(\mu^* = A_T x_{T-1} + b_t, \phi^* = \phi\Big)
\end{aligned}$$

7.2 Derivation of NDLM Gibbs Updates with Datacloning

Suppose that x is from a latent process governed by evolution function

$$x_t|x_{t-1} \sim \mathcal{N}(A_t x_{t-1} + b_t, \phi), t \in \{1, ..., T\}$$

Suppose for $I \subset \{1, ..., T\}$ there exists r identical observation data points y_I governed by the observation model

$$y_i|x_i \sim \mathcal{N}(x_i, \tau), i \in I$$

With arbitrary prior on ϕ, a fixed τ, and initial condition prior

$$x_0 \sim \mathcal{N}(\mu_0, \phi_0),$$

there exist tractable full conditionals for x_t that are derived here to be used with Gibbs sampling. Note that for the datacloning case, only full conditionals for timepoints with observed data are changed. If there is no observation data, the full conditionals remain unchanged from those derived in Section 8.1.

7.2.1 Interior Latent States

$$
\begin{aligned}
\pi(x_t|\cdot) &\propto \exp\Big(\frac{-\phi(x_t - A_t x_{t-1} - b_t)^2}{2}\Big)\exp\Big(\frac{-\phi(x_{t+1} - A_t x_t - b_{t-1})^2}{2}\Big)\exp\Big(-\frac{\tau}{2}(x_t - y_t)^2\Big)^r \\
&\propto \exp\Big(\frac{-\phi(x_t - A_t x_{t-1} - b_t)^2}{2}\Big)\exp\Big(\frac{-\phi(x_{t+1} - A_t x_t - b_{t-1})^2}{2}\Big)\exp\Big(-\frac{\tau r}{2}(x_t - y_t)^2\Big) \\
&\propto \exp\left(\frac{-\phi}{2}\Big(x_t^2 - 2x_t(A_t x_{t-1} + b_t) + x_t^2 A_{t+1}^2 - 2x_t A_{t+1} x_{t+1} + 2A_{t+1} b_{t+1} x_t\Big) - \frac{\tau r}{2}(x_t^2 - 2x_t y_t)\right) \\
&\propto \exp\left(-\frac{x_t^2}{2}\Big(\phi(1 + A_{t+1}^2) + \tau r\Big) + x_t\Big(\phi(A_t x_{t-1} + b_t + A_{t+1}(x_{t+1} - b_{t+1}) + \tau r y_t\Big)\right) \\
&\sim \mathcal{N}\left(\mu^* = \frac{\phi(A_t x_{t-1} + b_t + A_{t+1}(x_{t+1} - b_{t+1}) + \tau r y_t}{\phi(1 + A_{t+1}^2) + \tau r\Big)}, \phi^* = \phi(1 + A_{t+1}^2) + \tau r\right)
\end{aligned}
$$

7.2.2 Final Latent State

$$
\begin{aligned}
\pi(x_T|\cdot) &\propto \exp\Big(-\frac{\phi}{2}(x_T - A_T x_{T-1} - b_t)^2\Big)\exp\Big(-\frac{\tau}{2}(x_T - y_T)^2\Big)^r \\
&\propto \exp\left(-\frac{\phi}{2}\Big(x_T^2 - 2x_T(A_T x_{T-1} + b_T)\Big) - \frac{\tau r}{2}\Big(x_T^2 - 2x_T y_T\Big)\right) \\
&\propto \exp\left(-\frac{x_T^2}{2}\Big(\phi + \tau r\Big) + x_T\Big(\phi A_T x_{T-1} + b_T + \tau r y_T\Big)\right) \\
&\sim \mathcal{N}\Big(\mu^* = A_T x_{T-1} + b_t, \phi^* = \phi\Big)
\end{aligned}
$$

7.3 Derivation of Moment Matching Transformations

7.3.1 Lognormal Moment Matching

Suppose that we are interested in finding a transformation for random variable X, $X \sim$ Lognormal(μ^*, σ^{*2}) such that $\mathbb{E}[X] = \mu, \mathbb{V}[X] = \sigma^2$, with pdf given by:

$$f(x|\mu,\sigma) = \frac{1}{\mu\sigma\sqrt{2\pi}} \exp\left(- \frac{(\log(x) - \mu)^2}{2\sigma^2} \right), \mu \in \mathbb{R}, \sigma > 0, x \in (0, \infty)$$

Then, we have:

$$\begin{aligned}
\mathbb{E}[X] &= \mu = \exp\left(\mu^* + \frac{\sigma^{*2}}{2}\right) \\
\mathbb{V}[X] &= \sigma^2 = \Big(\exp(\sigma^{*2}) - 1\Big)\exp(2\mu^* + \sigma^{*2})
\end{aligned}$$

Looking at the first equality, we see that $\sigma^{*2} = 2(\log(\mu) - \mu^*)$. We can substitute this into the second equality to get μ^* written in terms of known constants,

$$\begin{aligned}
\sigma^2 &= \Big(\exp(2\log(\mu) - 2\mu^*) - 1\Big)\exp(2\mu^* + 2\log(\mu) - 2\mu^*) \\
&= \Big(\exp(2\log(\mu) - 2\mu^*) - 1\Big)\exp(2\log(\mu)) \\
&= (\mu^2\exp(-2\mu^*) - 1)\mu^2
\end{aligned}$$

Then, rewriting this equation in terms of $\exp(-2\mu^*)$,

$$\begin{aligned}
\exp(-2\mu^*) &= \frac{\sigma^2 + \mu^2}{\mu^4} \\
-2\mu^* &= \log\Big(\frac{\sigma^2 + \mu^2}{\mu^4}\Big) \\
\mu^* &= -\frac{1}{2}\log\Big(\frac{\sigma^2 + \mu^2}{\mu^4}\Big) \\
\mu^* &= \log\Big(\frac{\mu^2}{\sqrt{\mu^2 + \sigma^2}}\Big)
\end{aligned}$$

Substituting our value for μ^* back into the relation $\sigma^{*2} = 2(\log(\mu) - \mu^*)$, we have:

$$
\begin{aligned}
\sigma^{*2} &= 2(\log(\mu) - \mu^*) \\
&= 2\Big(\log(\mu) - \log\Big(\frac{\mu^2}{\sqrt{\mu^2+\sigma^2}}\Big)\Big) \\
&= 2\log\Big(\frac{\sqrt{\mu^2+\sigma^2}}{\mu}\Big) \\
&= \log\Big(\frac{\mu^2+\sigma^2}{\mu^2}\Big) \\
&= \log\Big(1+\frac{\sigma^2}{\mu^2}\Big)
\end{aligned}
$$

Thus our desired transform is $\mu^* = \log\Big(\frac{\mu^2}{\sqrt{\mu^2+\sigma^2}}\Big)$ and $\sigma^{*2} = \log\Big(1+\frac{\sigma^2}{\mu^2}\Big)$.

7.4 Gibbs Sampling Updates for Certain LNM3 Structures

Suppose that we have an LNM3 model such that our desired process model and variance are given by:

$$f(x_t|x_{t-1}, a^*, b^*) = a^*(x_{t-1})^{b^*} \tag{7.1}$$

$$\phi = (f(x_t|x_{t-1}, a^*, b^*)^2\sigma^2)^{-1} \tag{7.2}$$

Similarly, suppose that the desired observation model is given by:

$$g(y_t|x_t) = x_t \tag{7.3}$$

$$\tau = (x_t\omega^2)^{-1} \tag{7.4}$$

By taking the logarithm of the latent states ($X_{1:T}$) and observations ($Y_{t\in I}$), $D_t = \log(X_t)$, $F_i = \log(Y_i)$, these have a form identical to the NDLM Gibbs updates of Section 7.1

7.5 IF2 Parameter Values

Param.	Value	Fixed	Param.	Value	Fixed
f_a	.2443	No	f_{lab}	.3976	No
f_f	.0279	No	f_r	.0402	No
f_w	.2898	No	θ_w	4.488e-04	No
θ_r	3.45e-04	No	θ_{som}	3.49e-05	No
θ_{tot}	.0103	No	θ_p	.6306	No
Θ	.0533	No	d_o	148.1	No
d_f	275.3	No	c_{eff}	38.60	No
c_{lma}	75	Yes	c_{lf}	.8573	No
c_{ro}	10.04	No	c_{rf}	20.11	No
ag_2bg	.792	Yes	ω_w	.0135	No
ω_f	.0716	No	ω_{lab}	2.31e-03	No
ω_r	6.37e-03	No	ω_{lit}	.047	No
ω_{som}	.005	No	σ_w	.03	Yes
σ_f	.25	Yes	σ_r	.03	Yes
σ_{lit}	.1	Yes	ω_{som}	.0036	No
ω_{SL}	.33	Yes	ω_{SA}	.33	Yes
Cf_0	100.2	Yes	Cr_0	3643	Yes
Cw_0	7280	Yes	$Clit_0$	513.6	No
$Clab_0$	199.9	No	$Csom_0$	13270	Yes
σ_{nee}	.45	Yes	σ_{lf}	1	Yes
β_0	5	Yes	NA	NA	NA

Table 7.1: Information on the parameter estimates from the IF2 algorithm, including the parameter name, value from the IF2 global search, and whether or not the parameter was fixed.

Bibliography

[1] Aber, J. D. and Federer, C. A. (1992). A generalized, lumped-parameter model of photosynthesis, evapotranspiration and net primary production in temperate and boreal forest ecosystems. *Oecologia*, 92(4):463–474.

[2] Andrieu, C. and Doucet, A. (2002). Particle filtering for partially observed gaussian state space models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 64(4):827–836.

[3] Andrieu, C., Doucet, A., and Holenstein, R. (2010). Particle markov chain monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342.

[4] Andrieu, C. and Thoms, J. (2008). A tutorial on adaptive mcmc. *Statistics and Computing*, 18:343–373.

[5] Auger-Méthé, M., Field, C., Albertsen, C., Derocher, A., Lewis, M., Jonsen, I., and Flemming, J. (2016). State-space models' dirty little secrets: Even simple linear gaussian models can have estimation problems. *Scientific Reports*, 6:26677.

[6] Auger-Méthé, M., Newman, K., Cole, D., Empacher, F., Gryba, R., King, A. A., Leos-Barajas, V., Mills Flemming, J., Nielsen, A., Petris, G., and Thomas, L. (2021). A guide to state–space modeling of ecological time series. *Ecological Monographs*, 91(4):e01470.

[7] Baldocchi, D. (2014). Measuring fluxes of trace gases and energy between ecosystems and the atmosphere – the state and future of the eddy covariance method. *Global Change Biology*, 20(12):3600–3609.

[8] Baldocchi, D. D. (2003). Assessing the eddy covariance technique for evaluating carbon

dioxide exchange rates of ecosystems: past, present and future. *Global Change Biology*, 9(4):479–492.

[9] Banner, K. M., Irvine, K. M., and Rodhouse, T. J. (2020). The use of bayesian priors in ecology: The good, the bad and the not great. *Methods in Ecology and Evolution*, 11(8):882–889.

[10] Baracchini, T., Wuest, A., and Bouffard, D. (2020). Meteolakes: An operational online three-dimensional forecasting platform for lake hydrodynamics. *Water Research*, 172:115529.

[11] Berger, R. D. (1981). Comparison of the gompertz and logistic equations to describe plant disease progress. *Phytopathology*, 71:716.

[12] Bloom, A. and Williams, M. (2015). Constraining ecosystem carbon dynamics in a data-limited world: integrating ecological "common sense" in a model–data fusion framework. *Biogeosciences*, 12(5):1299–1315.

[13] Bloom, A. A., Exbrayat, J.-F., van der Velde, I. R., Feng, L., and Williams, M. (2016). The decadal state of the terrestrial carbon cycle: Global retrievals of terrestrial carbon allocation, pools, and residence times. *Proceedings of the National Academy of Sciences*, 113(5):1285–1290.

[14] Boettiger, C., Thomas, Q., Laney, C., and Lunch, C. (2022). *neonstore: NEON Data Store*. R package version 0.5.0.

[15] Box, G. E. P. (1980). Sampling and bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society. Series A (General)*, 143(4):383–430.

[16] Breto, C., He, D., Ionides, E., and King, A. (2008). Time series analysis via mechanistic models. *The Annals of Applied Statistics*, 3:319–348.

[17] Brooks, S. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *J. Comput. Graphi. Stat.*, 7:434–455.

[18] Buck-Sorlin, G. (2013). Process-based model. In Dubitzky, W., Wolkenhauer, O., Cho, K.-H., and Yokota, H., editors, *Encyclopedia of Systems Biology*, pages 1755–1755. Springer New York, New York, NY.

[19] Buckland, S., Newman, K., Thomas, L., and Koesters, N. (2004). State-space models for the dynamics of wild animal populations. *Ecological Modelling*, 171(1):157–175.

[20] Bulla, J. and Berzel, A. (2008). Computational issues in parameter estimation for stationary hidden markov models. *Computational Statistics*, 23:1–18.

[21] Cappe, O., Godsill, S. J., and Moulines, E. (2007). An overview of existing methods and recent advances in sequential monte carlo. *Proceedings of the IEEE*, 95(5):899–924.

[22] Carter, C. K. and Kohn, R. (1994). On Gibbs sampling for state space models. *Biometrika*, 81(3):541–553.

[23] Chen, R. and Liu, J. S. (2000). Mixture kalman filters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(3):493–508.

[24] Clark, J. S., Carpenter, S. R., Barber, M., Collins, S., Dobson, A., Foley, J. A., Lodge, D. M., Pascual, M., Pielke, R., Pizer, W., Pringle, C., Reid, W. V., Rose, K. A., Sala, O., Schlesinger, W. H., Wall, D. H., and Wear, D. (2001). Ecological forecasts: An emerging imperative. *Science*, 293(5530):657–660.

[25] Cole, D. J. (2019). Parameter redundancy and identifiability in hidden markov models. *Metron*, 77(2):105–118.

[26] Cole, D. J. (2020). *Parameter Redundancy and Identifiability*. Chapman and Hall/CRC.

[27] Cole, D. J. and McCrea, R. S. (2016). Parameter redundancy in discrete state-space and integrated models. *Biometrical Journal*, 58(5):1071–1090.

[28] Council, N. R. (2004). *Neon: Addressing the Nation's Environmental Challenges*. The National Academies Press, Washington, DC.

[29] DAAC, O. (2018). Fixed sites subsetting and visualization tool. ORNL DAAC, Oak Ridge, Tennessee, USA. Accessed March 23, 2022. Subset obtained for MCD15A3H product at site id 'us_wisconsin_neon_unde'.

[30] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

[31] Dennis, B. and Patil, G. (1984). The gamma distribution and weighted multimodal gamma distributions as models of population abundance. *Mathematical Biosciences*, 68(2):187–212.

[32] Dennis, B. and Patil, G. P. (1988). Applications in ecology. In Crow, E. L. and Shimizu, K., editors, *Lognormal Distributions*, chapter 12, pages 303–330. Taylor and Francis, Routledge.

[33] Dennis, B., Ponciano, J. M., Lele, S. R., Taper, M. L., and Staples, D. F. (2006). Estimating density dependence, process noise, and observation error. *Ecological Monographs*, 76(3):323–341.

[34] Deo, V. and Grover, G. (2021). A new extension of state-space sir model to account for underreporting – an application to the covid-19 transmission in california and florida. *Results in Physics*, 24:104182.

[35] Dietze, M. C., Fox, A., Beck-Johnson, L. M., Betancourt, J. L., Hooten, M. B., Jarnevich, C. S., Keitt, T. H., Kenney, M. A., Laney, C. M., Larsen, L. G., Loescher, H. W., Lunch, C. K., Pijanowski, B. C., Randerson, J. T., Read, E. K., Tredennick, A. T., Vargas, R., Weathers, K. C., and White, E. P. (2018). Iterative near-term ecological

forecasting: Needs, opportunities, and challenges. *Proceedings of the National Academy of Sciences*, 115(7):1424–1432.

[36] Doucet, A., Freitas, N., and Gordon, N. (2001). *Sequential Monte-Carlo Methods in Practice*, volume 1.

[37] Doucet, A. and Johansen, A. (2011). A tutorial on particle filtering and smoothing: Fifteen years later. In Crisan, D. and Rozovskii, B., editors, *The Oxford handbook of nonlinear filtering*, pages 656–705. Oxford University Press, Oxford ; N.Y.

[38] Doucet, A., Pitt, M. K., Deligiannidis, G., and Kohn, R. (2015). Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. *Biometrika*, 102(2):295–313.

[39] Dowd, M. and Meyer, R. (2003). A bayesian approach to the ecosystem inverse problem. *Ecological Modelling*, 168(1):39 – 55.

[40] Durbin, J. and Koopman, S. (2012). *Time Series analysis by state space methods*. Number Second Edition. Oxford University press.

[41] Evensen, G. (2003). The ensemble kalman filter: Theoretical formulation and practical implementation. *Ocean Dynamics*, 53:343–367.

[42] Fernandez-Villaverde, J. and Rubio-Ramirez, J. (2006). Estimating macroeconomic models: A likelihood approach. *Review of Economic Studies*, 74:1059–1087.

[43] Fox, A., Williams, M., Richardson, A., Cameron, D., Gove, J., Quaife, T., Ricciuto, D., Reichstein, M., Tomelleri, E., Trudinger, C., and Wijk, M. V. (2009). The reflex project: Comparing different algorithms and implementations for the inversion of a terrestrial ecosystem model against eddy covariance data. *Agricultural and Forest Meteorology*, 149:1597–1615.

[44] Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1(3):515 – 534.

[45] Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Statistical Methods, Third Edition*. Chapman and Hall/CRC, New York.

[46] Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741.

[47] Glickman, M. E. and Stern, H. S. (1998). A state-space model for national football league scores. *Journal of the American Statistical Association*, 93(441):25–35.

[48] Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

[49] Gompertz, B. (1825). Xxiv. on the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. in a letter to francis baily, esq. f. r. s. &c. *Philosophical Transactions of the Royal Society of London*, 115:513–583.

[50] Gonzalez-Akre, E., Piponiot, C., Lepore, M., Herrmann, V., Lutz, J., Baltzer, J., Dick, C., Gilbert, G., He, F., Heym, M., Huerta, A., Jansen, P., Johnson, D., Knapp, N., Král, K., Lin, D., Malhi, Y., McMahon, S., Myers, J., and Anderson-Teixeira, K. (2022). allodb: An r package for biomass estimation at globally distributed extratropical forest plots. *Methods in Ecology and Evolution*, 13:330–338.

[51] Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 14(1):107–114.

[52] Gordon, N. J., Salmond, D. J., and Smith, A. F. M. (1993). Novel approach to

nonlinear/non-gaussian bayesian state estimation. *IEE Proceedings F - Radar and Signal Processing*, 140(2):107–113.

[53] Hamilton, J. (1994). *Time series analysis*. Princeton Univ. Press, Princeton, NJ.

[54] Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.

[55] He, D., Ionides, E. L., and King, A. A. (2010). Plug-and-play inference for disease dynamics: measles in large and small populations as a case study. *Journal of The Royal Society Interface*, 7(43):271–283.

[56] Hollinger, D. Y. and Richardson, A. D. (2005). Uncertainty in eddy covariance measurements and its application to physiological models. *Tree Physiology*, 25(7):873–885.

[57] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70.

[58] Holmes, E. E. (2013). Derivation of an em algorithm for constrained and unconstrained multivariate autoregressive state-space (marss) models.

[59] Hooker, G., Ellner, S. P., Roditi, L. D. V., and Earn, D. J. D. (2011). Parameterizing state–space models for infectious disease dynamics by generalized profiling: measles in ontario. *Journal of The Royal Society Interface*, 8(60):961–974.

[60] Ionides, E. L., Bhadra, A., Atchadé, Y., and King, A. (2011). Iterated filtering. *The Annals of Statistics*, 39(3):1776 – 1802.

[61] Ionides, E. L., Bretó, C., and King, A. A. (2006). Inference for nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 103(49):18438–18443.

[62] Ionides, E. L., Nguyen, D., Atchadé, Y., Stoev, S., and King, A. A. (2015). Inference for dynamic and latent variable models via iterated, perturbed bayes maps. *Proceedings of the National Academy of Sciences*, 112(3):719–724.

[63] Ishihara, J., Terra, M., and Campos, J. (2006). Robust kalman filter for descriptor systems. *IEEE Transactions on Automatic Control*, 51(8):1354–1354.

[64] Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461.

[65] Jiang, J., Huang, Y., Ma, S., Stacy, M., Shi, Z., Ricciuto, D. M., Hanson, P. J., and Luo, Y. (2018). Forecasting responses of a northern peatland carbon cycle to elevated co2 and a gradient of experimental warming. *Journal of Geophysical Research: Biogeosciences*, 123(3):1057–1071.

[66] Johnson, L. R., Gramacy, R. B., Cohen, J., Mordecai, E., Murdock, C., Rohr, J., Ryan, S. J., Stewart-Ibarra, A. M., and Weikel, D. (2018). Phenomenological forecasting of disease incidence using heteroskedastic Gaussian processes: A dengue case study. *The Annals of Applied Statistics*, 12(1):27 – 66.

[67] Jones, E., Parslow, J., and Murray, L. (2010). A bayesian approach to state and parameter estimation in a phytoplankton-zooplankton model. *Australian Meteorological and Oceanographic Journal*, 59:7–16.

[68] Jordan, A., Krueger, F., and Lerch, S. (2017). Evaluating probabilistic forecasts with the r package scoringrules. *Journal of Statistical Software*, 90:1–37.

[69] Julier, S. and Uhlmann, J. (1997). New extension of the kalman filter to nonlinear systems. In *Defense, Security, and Sensing*.

[70] Jung, C. G. and Hararuk, O. (2022). Assimilation of neon observations into a process-based carbon cycle model reveals divergent mechanisms of carbon dynamics in temperate deciduous forests. *Journal of Geophysical Research: Biogeosciences*, 127(3).

[71] Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45.

[72] Kantas, N., Doucet, A., Singh, S., Maciejowski, J., and Chopin, N. (2015). On particle methods for parameter estimation in general state-space models. *Statistical Science*, (3):328–351.

[73] Keeling, C. D., Piper, S. C., Bacastow, R. B., Wahlen, M., Whorf, T. P., Heimann, M., and Meijer, H. A. (2005). *Atmospheric CO2 and 13CO2 Exchange with the Terrestrial Biosphere and Oceans from 1978 to 2000: Observations and Carbon Cycle Implications*, pages 83–113. Springer New York, New York, NY.

[74] Keenan, T., Carbone, M., Reichstein, M., and Richardson, A. (2011). The model-data fusion pitfall: Assuming certainty in an uncertain world. *Oecologia*, 167:587–597.

[75] Keenan, T. and Williams, C. (2018). The terrestrial carbon sink. *Annual Review of Environment and Resources*, 43:219–243.

[76] King, A. A., Nguyen, D., and Ionides, E. L. (2016). Statistical inference for partially observed markov processes via the r package pomp. *Journal of Statistical Software*, 69(12):1–43.

[77] Knape, J. and de Valpine, P. (2012). Fitting complex population models by combining particle filters with markov chain monte carlo. *Ecology*, 93(2):256–263.

[78] Knape, J., Jonzén, N., and Sköld, M. (2011). On observation distributions for state space models of population survey data. *The Journal of animal ecology*, 80:1269–77.

[79] Koopman, J. J. E., Rozing, M. P., Kramer, A., de Jager, D. J., Ansell, D., De Meester, J. M. J., Prütz, K. G., Finne, P., Heaf, J. G., Palsson, R., Kramar, R., Jager, K. J., Dekker, F. W., and Westendorp, R. G. J. (2011). Senescence rates in patients with end-stage renal disease: a critical appraisal of the gompertz model. *Aging Cell*, 10(2):233–238.

[80] Krüger, F., Lerch, S., Thorarinsdottir, T., and Gneiting, T. (2021). Predictive inference based on markov chain monte carlo output. *International Statistical Review*, 89(2):274–301.

[81] Lambert, J. H. (1758). Observationes variae in mathesin puram. *Acta Helvetica, physico-mathematico-anatomico-botanico-medica*, pages 128–168.

[82] Landsberg, J. and Waring, R. (1997). A generalised model of forest productivity using simplified concepts of radiation-use efficiency, carbon balance and partitioning. *Forest Ecology and Management*, 95(3):209 – 228.

[83] Le Quéré, C., Raupach, M., Canadell, J., Marland, G., Bopp, L., Ciais, P., Conway, T., Doney, S., Feely, R., Foster, P., Friedlingstein, P., Gurney, K., Houghton, R., House, J., Huntingford, C., Levy, P., Lomas, M., Majkut, J., Metzl, N., and Woodward, I. (2009). Trends in the sources and sinks of carbon dioxide. *Nature Geoscience*, 2:831–836.

[84] Lele, S., Dennis, B., and Lutscher, F. (2007). Data cloning: Easy maximum likelihood estimation for complex ecological models using bayesian markov chain monte carlo methods. *Ecology letters*, 10:551–63.

[85] Lele, S., Nadeem, K., and Schmuland, B. (2010). Estimability and likelihood inference for generalized linear mixed models using data cloning. *Journal of the American Statistical Association*, 105(492):1617–1625.

[86] Lele, S. R. (2020). Consequences of lack of parameterization invariance of non-informative bayesian analysis for wildlife management: Survival of san joaquin kit fox and declines in amphibian populations. *Frontiers in Ecology and Evolution*, 7:501.

[87] Lewis, A. S. L., Woelmer, W. M., Wander, H. L., Howard, D. W., Smith, J. W., McClure, R. P., Lofton, M. E., Hammond, N. W., Corrigan, R. S., Thomas, R. Q., and Carey, C. C. (2022). Increased adoption of best practices in ecological forecasting enables comparisons of forecastability. *Ecological Applications*, 32(2):e02500.

[88] Link, J., Ihde, T., Harvey, C., Gaichas, S., Field, J., Brodziak, J., Townsend, H., and Peterman, R. (2012). Dealing with uncertainty in ecosystem models: The paradox of use for living marine resource management. *Progress in Oceanography*, 102:102–114.

[89] Little, M., Heidenreich, W., and Li, G. (2009). Parameter identifiability and redundancy in a general class of stochastic carcinogenesis models. *PloS one*, 4:e8520.

[90] Little, R. (2012). Calibrated bayes, an alternative inferential paradigm for official statistics. *Journal of Official Statistics*, 28:309–334.

[91] Little, R. J. (2006). Calibrated bayes. *The American Statistician*, 60(3):213–223.

[92] Liu, J. and West, M. (2001). Combined parameter and state estimation in simulation-based filtering. In Doucet, A., de Freitas, N., and Gordon, N., editors, *Sequential Monte Carlo Methods in Practice*, pages 197–223. Springer New York, New York, NY.

[93] López-Blanco, E., Exbrayat, J.-F., Lund, M., Christensen, T. R., Tamstorf, M. P., Slevin, D., Hugelius, G., Bloom, A. A., and Williams, M. (2019). Evaluation of terrestrial pan-arctic carbon cycling using a data-assimilation system. *Earth System Dynamics*, 10(2):233–255.

[94] Lounis, M., Torrealba-Rodriguez, O., and Conde-Gutiérrez, R. (2021). Predictive models for covid-19 cases, deaths and recoveries in algeria. *Results in Physics*, 30:104845.

[95] Luo, Y., Ogle, K., Tucker, C., Fei, S., Gao, C., LaDeau, S. L., Clark, J. S., and Schimel, D. S. (2011). Ecological forecasting and data assimilation in a data-rich era. *Ecological applications : a publication of the Ecological Society of America*, 21 5:1429–42.

[96] Luo, Y., Weng, E., Wu, X., Gao, C., Zhou, X., and Zhang, L. (2009). Parameter identifiability, constraint, and equifinality in data assimilation with ecosystem models. *Ecological Applications*, 19(3):571–574.

[97] Matheson, J. E. and Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management Science*, 22(10):1087–1096.

[98] Maunder, M. N., Deriso, R. B., and Hanson, C. H. (2015). Use of state-space population dynamics models in hypothesis testing: advantages over simple log-linear regressions for modeling survival, illustrated with application to longfin smelt (spirinchus thaleichthys). *Fisheries Research*, 164:102–111.

[99] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

[100] Metzger, S., Ayres, E., Durden, D., Florian, C., Lee, R., Lunch, C., Luo, H., Pingintha-Durden, N., Roberti, J., Sanclements, M., Sturtevant, C., Xu, K., and Zulueta, R. (2019). From neon field sites to data portal: a community resource for surface-atmosphere research comes online. *Bulletin of the American Meteorological Society*, 100:2305–2325.

[101] Myneni, R., Knyazikhin, Y., and Park, T. (2015). Mcd15a3h modis/terra+aqua leaf area index/fpar 4-day l4 global 500m sin grid v006. *NASA EOSDIS Land Processes DAAC*.

[102] Mäntyniemi, S. H. P., Whitlock, R. E., Perälä, T. A., Blomstedt, P. A., Vanhatalo, J. P., Rincón, M. M., Kuparinen, A. K., Pulkkinen, H. P., and Kuikka, O. S. (2015). General state-space population dynamics model for Bayesian stock assessment. *ICES Journal of Marine Science*, 72(8):2209–2222.

[103] National Ecological Observatory Network (2020). Woody plant vegetation structure, Data Product DP1.10098.001, Provisional data downloaded from `http://data.neonscience.org` on April 21, 2020.

[104] National Ecological Observatory Network (NEON) (2022a). Bundled data products - eddy covariance (dp4.00200.001).

[105] National Ecological Observatory Network (NEON) (2022b). Litterfall and fine woody debris production and chemistry (dp1.10033.001).

[106] National Ecological Observatory Network (NEON) (2022c). Root biomass and chemistry, periodic (dp1.10067.001).

[107] National Ecological Observatory Network (NEON) (2022d). Shortwave radiation (primary pyranometer) (dp1.00022.001).

[108] National Ecological Observatory Network (NEON) (2022e). Soil physical and chemical properties, megapit (dp1.00096.001).

[109] National Ecological Observatory Network (NEON) (2022f). Triple aspirated air temperature (dp1.00003.001).

[110] National Ecological Observatory Network (NEON) (2022g). Vegetation structure (dp1.10098.001).

[111] Nielsen, A. and Sibert, J. R. (2007). State–space model for light-based tracking of marine animals. *Canadian Journal of Fisheries and Aquatic Sciences*, 64(8):1055–1068.

[112] NOAA (2015). Dengue forecasting. https://dengueforecasting.noaa.gov/. Accessed: 2021-09-30.

[113] Osthus, D., Hickmann, K. S., Caragea, P. C., Higdon, D., and Del Valle, S. Y. (2017). Forecasting seasonal influenza with a state-space sir model. *The Annals of Applied Statistics*, 11(1).

[114] Parslow, J., Cressie, N., Campbell, E. P., Jones, E., and Murray, L. (2013). Bayesian learning and predictability in a stochastic nonlinear dynamical model. *Ecological Applications*, 23(4):679–698.

[115] Patterson, T. A., Thomas, L., Wilcox, C., Ovaskainen, O., and Matthiopoulos, J. (2008). State–space models of individual animal movement. *Trends in Ecology Evolution*, 23(2):87–94.

[116] Petris, G., Petrone, S., and Campagnoli, P. (2009). *Dynamic Linear Models with R*, volume 38, pages 31–84.

[117] Pinnington, E. M., Casella, E., Dance, S. L., Lawless, A. S., Morison, J. I. L., Nichols, N. K., Wilkinson, M., and Quaife, T. L. (2017). Understanding the effect of disturbance from selective felling on the carbon dynamics of a managed woodland by combining observations with model predictions. *Journal of Geophysical Research: Biogeosciences*, 122(4):886–902.

[118] Pitt, M. K., Silva, R. d. S., Giordani, P., and Kohn, R. (2012). On some properties of Markov chain Monte Carlo simulation methods based on the particle filter. *Journal of Econometrics*, 171(2):134–151.

[119] Plummer, M. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling.

[120] Plummer, M. (2019). *rjags: Bayesian Graphical Models using MCMC*. R package version 4-10.

[121] Polson, N. G. and Scott, J. G. (2012). On the Half-Cauchy Prior for a Global Scale Parameter. *Bayesian Analysis*, 7(4):887 – 902.

[122] Ponciano, J. M., Burleigh, J. G., Braun, E. L., and Taper, M. L. (2012). Assessing parameter identifiability in phylogenetic models using data cloning. *Systematic Biology*, 61(6):955 – 972.

[123] R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[124] Reich, B. J. and Ghosh, S. K. (2019). *Bayesian Statistical Methods, First Edition*. Chapman and Hall/CRC, New York.

[125] Ricker, W. E. (1954). Stock and recruitment. *Journal of the Fisheries Research Board of Canada*, 11(5):559–623.

[126] Robert, C. P. and Casella, G. (2005). *Monte Carlo Statistical Methods (Springer Texts in Statistics)*. Springer-Verlag, Berlin, Heidelberg.

[127] Robert, C. P. and Casella, G. (2009). *Introducing Monte Carlo Methods with R (Use R)*. Springer-Verlag, Berlin, Heidelberg, 1st edition.

[128] Rosenthal, S. (2009). Optimal proposal distributions and adaptive mcmc. *Handbook of Markov Chain Monte Carlo*.

[129] Rothenberg, T. J. (1971). Identification in parametric models. *Econometrica*, 39(3):577–591.

[130] Roulston, M. S. and Smith, L. A. (2002). Evaluating probabilistic forecasts using information theory. *Monthly Weather Review*, 130(6):1653 – 1660.

[131] Rubin, D. B. (1984). Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician. *The Annals of Statistics*, 12(4):1151 – 1172.

[132] Schimel, D., Stephens, B. B., and Fisher, J. B. (2015). Effect of increasing co_2 on the terrestrial carbon cycle. *Proceedings of the National Academy of Sciences*, 112(2):436–441.

[133] Sedda, L., Taylor, B. M., Eiras, A. E., Marques, J. T., and Dillon, R. J. (2020). Using the intrinsic growth rate of the mosquito population improves spatio-temporal dengue risk estimation. *Acta Tropica*, 208:105519.

[134] Serbin, S. P., Wu, J., Ely, K. S., Kruger, E. L., Townsend, P. A., Meng, R., Wolfe, B. T., Chlus, A., Wang, Z., and Rogers, A. (2019). From the arctic to the tropics: multibiome prediction of leaf mass per area using leaf reflectance. *New Phytologist*, 224(4):1557–1568.

[135] Servanty, S., Choquet, R., Baubet, E., Brandt, S., Gaillard, J.-M., Schaub, M., Toïgo, C., Lebreton, J.-D., Buoro, M., and Gimenez, O. (2010). Assessing whether mortality is additive using marked animals: A bayesian state - space modeling approach. *Ecology*, 91:1916–23.

[136] Sherlock, C., Thiery, A., Roberts, G., and Rosenthal, J. (2015). On the efficiency of pseudo-marginal random walk metropolis algorithms. *The Annals of Statistics*, 43:238–275.

[137] Shumway, R. and Stoffer, D. (2011). *Time Series Analysis and Its Applications With R Examples*, volume 9.

[138] Sisson, S. A., Fan, Y., and Tanaka, M. M. (2007). Sequential monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 104(6):1760–1765.

[139] Smallman, T. L., Exbrayat, J.-F., Mencuccini, M., Bloom, A. A., and Williams, M. (2017). Assimilation of repeated woody biomass observations constrains decadal ecosystem carbon cycle uncertainty in aggrading forests. *Journal of Geophysical Research: Biogeosciences*, 122(3):528–545.

[140] Smith, A. F. M. and Gelfand, A. E. (1992). Bayesian statistics without tears: A sampling-resampling perspective. *The American Statistician*, 46(2):84–88.

[141] Smith, J. W., Johnson, L. R., and Thomas, Q. (2021). Assessing ecosystem state space models: Identifiability and estimation.

[142] Smith, R. L. and Miller, J. E. (1986). A non-gaussian state space model and application to prediction of records. *Journal of the Royal Statistical Society. Series B (Methodological)*, 48(1):79–88.

[143] Stettz, S. G., Parazoo, N. C., Bloom, A. A., Blanken, P. D., Bowling, D. R., Burns, S. P., Bacour, C., Maignan, F., Raczka, B., Norton, A. J., Baker, I., Williams, M., Shi,

M., Zhang, Y., and Qiu, B. (2022). Resolving temperature limitation on spring productivity in an evergreen conifer forest using a model–data fusion framework. *Biogeosciences*, 19(2):541–558.

[144] Subramanian, K., Maravelias, C. T., and Rawlings, J. B. (2012). A state-space model for chemical production scheduling. *Computers Chemical Engineering*, 47:97–110.

[145] Thomas, Q., Brooks, E., Jersild, A., Ward, E., Wynne, R., Albaugh, T., Dinon-Aldridge, H., Burkhart, H., Domec, J.-C., Fox, T., Gonzalez-Benecke, C., Martin, T., Noormets, A., Sampson, D., and Teskey, R. (2017). Leveraging 35 years of pinus taeda research in the southeastern us to constrain forest carbon cycle predictions: Regional data assimilation using ecosystem experiments. *Biogeosciences*, 14:3525–3547.

[146] Thomas, R. Q., Jersild, A. L., Brooks, E. B., Thomas, V. A., and Wynne, R. H. (2018). A mid-century ecological forecast with partitioned uncertainty predicts increases in loblolly pine forest productivity. *Ecological Applications*, 28(6):1503–1519.

[147] Turner, R. and Deisenroth, M. (2010). State-space inference and learning with gaussian processes. *Journal of Machine Learning Research - Proceedings Track*, 9:868–875.

[148] van Leeuwen, P. J. (2009). Particle filtering in geophysical systems. *Monthly Weather Review*, 137(12):4089 – 4114.

[149] Viallefont, A., Lebreton, J.-D., Reboulet, A.-M., and Gory, G. (1998). Parameter identifiability and model selection in capture-recapture models: A numerical approach. *Biometrical Journal*, 40:313 – 325.

[150] Wan, E. and Van Der Merwe, R. (2000). The unscented kalman filter for nonlinear estimation. In *Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium (Cat. No.00EX373)*, pages 153–158.

[151] West, M. and Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models (2nd Ed.)*. Springer-Verlag, Berlin, Heidelberg.

[152] White, E. P., Yenni, G. M., Taylor, S. D., Christensen, E. M., Bledsoe, E. K., Simonis, J. L., and Ernest, S. K. M. (2019). Developing an automated iterative near-term forecasting system for an ecological study. *Methods in Ecology and Evolution*, 10(3):332–344.

[153] Williams, M., Rastetter, E. B., Fernandes, D. N., Goulden, M. L., Shaver, G. R., and Johnson, L. C. (1997). Predicting gross primary productivity in terrestrial ecosystems. *Ecological Applications*, 7(3):882–894.

[154] Williams, M., Schwarz, P. A., Law, B. E., Irvine, J., and Kurpius, M. R. (2005). An improved analysis of forest carbon dynamics using data assimilation. *Global Change Biology*, 11(1):89–105.

[155] Xu, T., White, L., Hui, D., and Luo, Y. (2006). Probabilistic inversion of a terrestrial ecosystem model: Analysis of uncertainty in parameter estimation and model prediction. *Global Biogeochemical Cycles*, 20(2):GB2007.

[156] Yang, W., Tan, B., Huang, D., Rautiainen, M., Shabanov, N., Wang, Y., Privette, J., Huemmrich, K., Fensholt, R., Sandholt, I., Weiss, M., Ahl, D., Gower, S., Nemani, R., Knyazikhin, Y., and Myneni, R. (2006). Modis leaf area index products: from validation to algorithm improvement. *IEEE Transactions on Geoscience and Remote Sensing*, 44(7):1885–1898.