Assessing Ecosystem State Space Models: Identifiability and Estimation

J. W. SMITH Jr., L. R. JOHNSON, and R. Q. THOMAS

Hierarchical probability models are being used more often than non-hierarchical deterministic process models in environmental prediction and forecasting, and Bayesian approaches to fitting such models are becoming increasingly popular. In particular, models describing ecosystem dynamics with multiple states that are autoregressive at each step in time can be treated as statistical state space models (SSMs). In this paper, we examine this subset of ecosystem models, embed a process-based ecosystem model into an SSM, and give closed form Gibbs sampling updates for latent states and process precision parameters when process and observation errors are normally distributed. Here, we use simulated data from an example model (DALECev) and study the effects changing the temporal resolution of observations on the states (observation data gaps), the temporal resolution of the state process (model time step), and the level of aggregation of observations on fluxes (measurements of transfer rates on the state process). We show that parameter estimates become unreliable as temporal gaps between observed state data increase. To improve parameter estimates, we introduce a method of tuning the time resolution of the latent states while still using higher-frequency driver information and show that this helps to improve estimates. Further, we show that data cloning is a suitable method for assessing parameter identifiability in this class of models. Overall, our study helps inform the application of state space models to ecological forecasting applications where (1) data are not available for all states and transfers at the operational time step for the ecosystem model and (2) process uncertainty estimation is desired.

Key Words: Bayesian analysis; Data cloning; Ecological forecasting; MCMC.

1. INTRODUCTION

Many ecological prediction and forecasting applications use mechanistic process-based models to simulate the dynamics of ecosystems (Luo 2011). These process models are

J. W. Smith Jr. (✉) Department of Mathematical Sciences, Montana State University, Bozeman, MT, USA (E-mail: *john.smith20@montana.edu*). J. W. Smith Jr. · L. R. Johnson, Department of Statistics, Virginia Tech, Blacksburg, VA, USA. R. Q. Thomas, Department of Forest Resources and Environmental Conversation, Virginia Tech, Blacksburg, VA, USA. L. R. Johnson · R. Q. Thomas, Global Change Center, Virginia Tech, Blacksburg, VA, USA. L. R. Johnson · R. Q. Thomas, Department of Biological Sciences, Virginia Tech, Blacksburg, VA, USA
.

Journal of Agricultural, Biological, and Environmental Statistics, Volume 28, Number 3, Pages 442–465
https://doi.org/10.1007/s13253-023-00531-8

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 (Luo 2009).

While the importance of quantifying uncertainty in process models is recognized (Dietze 2018), 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 Jiang (2018), White (2019), Baracchini et al. (2020), 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 (Petris et al. 2009; Durbin and Koopman 2012) 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 (Auger-Méthé 2021). As a result, it has seen a growing use in ecological forecasting applications [see Thomas (2017), Dowd and Meyer (2003), 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 either estimate latent states for each ecosystem model state, or to integrate them out. 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 time step (i.e., annual or greater). These large gaps between observed data may present a challenge to constraining both the latent states and process precisions.

Furthermore, identifiability [or equifinality (Luo 2009)] 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 (Ponciano et al. 2012) and for estimation in ecological models (Lele et al. 2007). DC is done by applying Bayesian inference to a dataset that is constructed by duplicating the initial dataset and treating them as r independent experiment results [as described in Ponciano et al. (2012)]. 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 (Ponciano et al. 2012). Parameters are said to be non-estimable when there exist different parameter values $\theta_1, \theta_2, \ldots, \theta_n$ such that $\mathcal{L}(\theta_1|X) = \mathcal{L}(\theta_2|X) = \cdots = \mathcal{L}(\theta_n|X) = \text{argmax}_{\Theta}\mathcal{L}(\theta|X)$, i.e., there are multiple sets of parameter values that maximize the likelihood function (Lele et al. 2010; Rothenberg 1971; Ponciano et al. 2012; Cole 2020). Ponciano et al. (2012) introduce terminology for various situations where we have parameters in the model that are not estimable: non-separability, lack of information, non-identifiable, and identifiable but not estimable. We use Ponciano et al. (2012)'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 Cole (2020)]. Parameters that are non-estimable due to non-separability are referred to as non-identifiable (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 non-estimable due to lack of information are referred to as INE (identifiable but not estimable). A combination of data cloning and comparison of summaries of posterior samples under different temporal resolutions leads to a thorough analysis procedure for ecosystem state space models that has not commonly been applied.

One strength of the data that we use to fit ecosystem models is that observations are available on both the stocks (components of the state vector of the latent process) and the fluxes (transfer rates between components of the latent process). There has been much work done for state space models with multiple data streams in the population modeling literature [e.g., integrated population models (Copyright 2022)] and fisheries literature [e.g., Nielsen and Berg (2014)]. Integrating multiple data streams can help to make parameters identifiable that would otherwise be unidentifiable with only one of the component data sources (Riecke 2019).

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 three questions focused on different temporal resolutions of the state

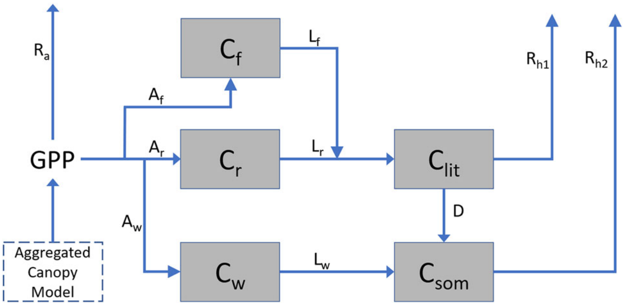

Figure 1. A schematic of the DALECev model with boxes denoting stocks of carbons and arrows denoting fluxes of carbon.

process, temporal resolutions of observations on the states, and the level of aggregation of observations on the fluxes and their impact on estimation of precision parameters, latent states, and process parameters. (1) How does varying the observation time resolution change estimates of process parameters precisions, process precisions, and latent states with a daily state process resolution and all flux data available? (2) How does changing the temporal resolution of the state process from daily to monthly change estimates of process precisions, process parameters, and latent states? and 3) Can we determine when model parameters are identifiable under different levels of aggregation of flux data 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).

2. METHODS

2.1. PROCESS MODEL

We used the Data Assimilation Linked Ecosystem Carbon model designed for simulating forests composed of evergreen trees [DALECev (Williams et al. 2005)]. It is a simple model describing carbon dynamics (Fig. 1) and is similar to other ecosystem models used in carbon stock forecasting applications, for example PnET (Aber and Federer 1992), 3PG (Landsberg and Waring 1997), and TECOS (Xu et al. 2006). The model can be written as a set of equations that are approximately linear and autoregressive in time. While DALECev has been widely used (Williams et al. 2005; Smallman et al. 2017; Fox et al. 2009; Bloom and Williams 2015), it is not traditionally fit as a state space model as we do here.

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

Table 1. Information on the physical interpretations, upper and lower bounds, units, and values used for simulation for the eleven process parameters, five process precisions, and five observation precisions used to generate our synthetic data using DALECev

Param.	Description	Lower	Upper	Units	Sim. value
p_1	Scaled daily decomposition rate	1.1e−05	0.11	day^{-1}	0.002
p_2	Fraction of GPP respired	0.2	0.7	unitless	0.27
p_3	Fraction of NPP allocated to foliage	0.01	0.5	unitless	0.15
p_4	Fraction of NPP after foliage allocation allocated to roots	0.01	0.5	unitless	0.33
p_5	Daily turnover rate of foliage	1e−04	0.1	day^{-1}	0.00137
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}	0.00137
p_8	Scaled daily mineralization rate of litter	1e−06	1	day^{-1}	0.1
p_9	Daily mineralization 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	°C^{-1}	0.1725
p_{11}	Nitrogen use efficiency parameter	1	20	(gN gC^{-1} d^{-1})$^{-1}$	3
ϕ_f	C_f process precision	0	∞	m^4 gC^{-2}	0.2975
ϕ_w	C_w process precision	0	∞	m^4 gC^{-2}	0.00618
ϕ_r	C_r process precision	0	∞	m^4 gC^{-2}	0.06
ϕ_{lit}	C_{lit} process precision	0	∞	m^4 gC^{-2}	3.625
ϕ_{som}	C_{som} process precision	0	∞	m^4 gC^{-2}	0.075
τ_f	C_f observation precision	0	∞	m^4 gC^{-2}	$0.25^{\dagger}$
τ_w	C_w observation precision	0	∞	m^4 gC^{-2}	$0.0625^{\dagger}$
τ_r	C_r observation precision	0	∞	m^4 gC^{-2}	$1^{\dagger}$
τ_{lit}	C_{lit} observation precision	0	∞	m^4 gC^{-2}	$1^{\dagger}$
τ_{som}	C_{som} observation precision	0	∞	m^4 gC^{-2}	$0.04^{\dagger}$

Parameter descriptions are taken from the REFLEX project supplemental material (Fox et al. 2009). Simulation values were chosen such that stock data, leaf area index (LAI), and NEE were reflective of what we expect at the simulation site

†Is used to denote parameters that are known a priori and not estimated during modeling

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 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 Williams et al. (1997), to simulate the input of carbon through gross photosynthetic production [GPP; G in Eqs. (2)–(3))]. Following Fox et al. (2009), 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 the turnover and respiration (carbon lost from the system) or transfer to another stock, plus carbon gained by through the allocation of photosynthesis (carbon gained from outside the system) or transfer from another stock:

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

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 the expected stock evolution through time, the deterministic skeleton, are:

$$\begin{aligned}
\mathbb{E}[C_f^{(t)}] &= C_f^{(t-1)} - L_f^{(t-1)} + A_f^{(t-1)} \\
&= C_f^{(t-1)} - p_5 C_f^{(t-1)} + G(\mathbf{D^{(t)}}, p_{11})(1-p_2)p_3 && (2)\\
\mathbb{E}[C_r^{(t)}] &= C_r^{(t-1)} - L_r^{(t-1)} + A_r^{(t-1)} \\
&= C_r^{(t-1)} - p_7 C_r^{(t-1)} + G(\mathbf{D^{(t)}}, p_{11})(1-p_2)(1-p_3)p_4 && (3)\\
\mathbb{E}[C_w^{(t)}] &= C_w^{(t-1)} - L_w^{(t-1)} + A_w^{(t-1)} \\
&= C_w^{(t-1)} - p_6 C_w^{(t-1)} + G(\mathbf{D^{(t)}}, p_{11})(1-p_2)(1-p_3)(1-p_4) && (4)\\
\mathbb{E}[C_{lit}^{(t)}] &= C_{lit}^{(t-1)} - R_{lit}^{(t-1)} - D_{lit}^{(t-1)} + A_{lit}^{(t-1)} \\
&= C_{lit}^{(t-1)} - \left[\frac{p_1 p_8 \exp(p_{10}\bar{T}^{(t)})}{2} - \frac{p_1(1-p_8)\exp(p_{10}\bar{T}^{(t)})}{2}\right] C_{lit}^{(t-1)} \\
&\quad + \left[p_5 C_f^{(t-1)} + p_7 C_r^{(t-1)}\right] && (5)\\
\mathbb{E}[C_{som}^{(t)}] &= C_{som}^{(t-1)} - R_{som}^{(t-1)} + D_{som}^{(t-1)} + A_{som}^{(t-1)} \\
&= C_{som}^{(t-1)} - \frac{p_9}{2}\exp(p_{10}\bar{T}^{(t)}) C_{som}^{(t-1)} \\
&\quad + \frac{(p_1 - p_1 p_8)}{2}\exp(p_{10}\bar{T}^{(t)}) C_{lit}^{(t-1)} + p_6 C_w^{(t-1)}, && (6)
\end{aligned}$$

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

$$\mathbb{E}[C^{(t)}|C^{(t-1)}] = A_t C^{(t-1)} + b_t, \tag{7}$$

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 Eqs. (2)–(6) in the DALECev model, are:

$$\begin{aligned}
R_{lit}^{(t)} &= .5\exp(p_{10}\bar{T}^{(t)}) p_1 p_8 C_{lit}^{(t-1)} && (8)\\
R_{som}^{(t)} &= .5\exp(p_{10}\bar{T}^{(t)}) p_9 C_{som}^{(t-1)} && (9)\\
R_a^{(t)} &= G(\mathbf{D^{(t)}}, p_{11}) p_2 && (10)\\
A_f^{(t)} &= G(\mathbf{D^{(t)}}, p_{11})(1-p_2)p_3 && (11)\\
A_r^{(t)} &= G(\mathbf{D^{(t)}}, p_{11})(1-p_2)(1-p_3)p_4 && (12)\\
A_w^{(t)} &= G(\mathbf{D^{(t)}}, p_{11})(1-p_2)(1-p_3)(1-p_4) && (13)\\
L_f^{(t)} &= p_5 C_f^{(t-1)} && (14)\\
L_w^{(t)} &= p_6 C_w^{(t-1)} && (15)\\
L_r^{(t)} &= p_7 C_r^{(t-1)} && (16)\\
D_{lit}^{(t)} &= .5\exp(p_{10}\bar{T}^{(t)}) p_1(1-p_8) C_{lit}^{(t-1)}, && (17)
\end{aligned}$$

where subscripts pertain to different carbon stocks, and a represents autotrophic. 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 (Baldocchi 2014) and deployed in many ecological observation networks (Metzger and Others 2019). NEE is the net of G and Eqs. (8)–(10) and is given by:

$$\begin{aligned} NEE^{(t)} &= R_{som}^{(t)} + R_{lit}^{(t)} + R_a^{(t)} - G(\mathbf{D}^{(\mathbf{t})}, p_{11}) \\ &= .5\exp(p_{10}\bar{T}^{(t)})p_1 p_8 C_{lit}^{(t)} + .5\exp(p_{10}\bar{T}^{(t)})p_9 C_{som}^{(t)} - G(\mathbf{D}^{(\mathbf{t})}, p_{11})(1-p_2) \end{aligned} \tag{18}$$

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 Eq. (10)] and heterotrophic respiration by soil micro-organisms [Eqs. (8), (9)):

$$\begin{aligned} S_r^{(t)} &= R_{som}^{(t)} + R_{lit}^{(t)} - cR_a^{(t)} \\ &= .5\exp(p_{10}\bar{T}^{(t)})p_1 p_8 C_{lit}^{(t)} + .5\exp(p_{10}\bar{T}^{(t)})p_9 C_{som}^{(t)} - cG(\mathbf{D}^{(\mathbf{t})}, p_{11})p_2, \;\; c \in [0,1] \end{aligned} \tag{19}$$

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.

2.2. State Space Model

We estimate the stocks for the DALECev model using a state space model (Hamilton 1994; Petris et al. 2009; Durbin and Koopman 2012; Auger-Méthé 2021). 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, Eq. [(2)–(6)] can be written using matrix notation as:

$$\mathbb{E}[\mathbf{C}^{(t)}] = M_t \mathbf{C}^{(t-1)} + \mathbf{P}^{(t)}, \quad \text{where} \tag{20}$$

$$\mathbf{C}^{(t)} = \begin{bmatrix} C_f^{(t)} \\ C_r^{(t)} \\ C_w^{(t)} \\ C_{lit}^{(t)} \\ C_{som}^{(t)} \end{bmatrix}, \quad M_t = \begin{bmatrix} (1-p_5) & 0 & 0 & 0 & 0 \\ 0 & (1-p_7) & 0 & 0 & 0 \\ 0 & 0 & (1-p_6) & 0 & 0 \\ p_5 & p_7 & 0 & (1-\frac{p_1}{2}Q^{(t)}) & 0 \\ 0 & 0 & 0 & \frac{p_1(1-p_8)}{2}Q^{(t)} & \left(1-\frac{p_9}{2}Q^{(t)}\right) \end{bmatrix},$$

$$\mathbf{P}^{(t)} = \begin{bmatrix} G(\mathbf{D^{(t)}}, p_{11})\psi_1 \\ G(\mathbf{D^{(t)}}, p_{11})\psi_2 \\ G(\mathbf{D^{(t)}}, p_{11})\psi_3 \\ 0 \\ 0 \end{bmatrix},$$

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:

$$C_{s,obs}^{(t)} = C_s^{(t)} + \epsilon_{obs,t}, \ t \in I, \ \epsilon_{obs,t} \sim \mathcal{N}(0, \tau_s) \tag{21}$$

In an ecological context, we are assuming that our observed carbon stock is normally distributed and unbiased (Auger-Méthé 2016), 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 are 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:

$$\begin{aligned} &\mathbf{C}^{(t)}|\mathbf{C}^{(t-1)} \sim \mathrm{MVN}(M_t\mathbf{C}^{(t-1)} + \mathbf{P}^{(t)}, \Phi), \quad t = 1, \dots, T \\ &\mathbf{C}_{obs}^{(t)}|\mathbf{C}^{(t)} \sim \mathrm{MVN}(\mathbf{C}^{(t)}, \tau), t \in I, \quad \text{where} \\ &\Phi = \begin{bmatrix} 1/\phi_{Cf} & 0 & 0 & 0 & 0 \\ 0 & 1/\phi_{Cr} & 0 & 0 & 0 \\ 0 & 0 & 1/\phi_{Cw} & 0 & 0 \\ 0 & 0 & 0 & 1/\phi_{Clit} & 0 \\ 0 & 0 & 0 & 0 & 1/\phi_{Csom} \end{bmatrix}, \\ &\tau = \begin{bmatrix} 1/\tau_{Cf} & 0 & 0 & 0 & 0 \\ 0 & 1/\tau_{Cr} & 0 & 0 & 0 \\ 0 & 0 & 1/\tau_{Cw} & 0 & 0 \\ 0 & 0 & 0 & 1/\tau_{Clit} & 0 \\ 0 & 0 & 0 & 0 & 1/\tau_{Csom} \end{bmatrix} \end{aligned}$$

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 (Auger-Méthé 2016). 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) (West and Harrison 1997).

The fluxes [Eqs. (10)–(17)] are modeled 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

$$F_{j,obs}^{(t)}|F_j^{(t)} \sim \mathcal{N}(F_j^{(t)}, \delta_j), \quad t \in I_j, \tag{22}$$

where δ_j is a known precision. Fluxes are assumed to follow the functional forms given in Eqs. (10)–(17), for example $R_a^{(t)}|R_{a,obs}^{(t)} \sim \mathcal{N}(G(\mathbf{D}^{(\mathbf{t})}, p_{11})p_2, \delta_{R_a})$. 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 Φ, $p_{1:11}$ and the latent states $\mathbf{C}^{(1:T)}$:

$$\begin{aligned} &\mathscr{L}(\mathbf{C}^{(1:T)}, \Phi, p_{1:11}|\mathbf{D}^{(1:T)}, \tau, \delta_{1:J}, \mathbf{C}_{obs}, \mathbf{F}_{obs}) \\ &\propto \prod_{t=1}^{T} \mathrm{MVN}(M_t\mathbf{C}^{(t)} + \mathbf{P}^{(t)}, \Phi) \times \prod_{t\in I} \mathrm{MVN}(\mathbf{C}^{(t)}, \tau) \times \left(\prod_{j=1}^{J} \prod_{t\in I_j} \mathcal{N}(F_j^{(t)}, \delta_j) \right) \end{aligned} \tag{23}$$

Many prior studies using terrestrial carbon models include observational uncertainty, but do not include process variation, e.g., Jiang (2018). 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 (Dietze 2018) and these uncertainties influence the forecast in different ways, e.g., process uncertainty propagates from one time step to another while observation uncertainty does not. The Bayesian state space paradigm outlined here allows for quantification of multiple sources of uncertainty (process, initial conditions, observations, and parameters) in the context of temporal gaps in observations, and the state space model gives a natural setting to leverage multiple data streams with process based models.

2.3. Inference for Parameters and Latent States

We estimate the stocks and process parameters for the DALECev model using a Bayesian state space model (Reich and Ghosh 2019). Process parameters and latent states were estimated with MCMC (Robert and Casella 2005). 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 parameters. In addition to the likelihood [Eq. (23)], 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 (Fox et al. 2009) 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 1. Each process precision was given a univariate conjugate Jeffreys prior (Jeffreys

1946), to allow for closed form Gibbs sampling of the process precision parameters. Thus, the priors are given by

$$p_i \sim \text{Unif}(p_i^{(L)}, p_i^{(U)}), \quad i = 1, \dots, 11 \tag{24}$$

$$C_k^{(0)} \sim \mathcal{N}(\mu_k^{(0)}, \phi_k^{(0)}), \quad k \in \{f, r, w, lit, som\} \tag{25}$$

$$\pi(\phi_k) \propto \frac{1}{\phi_k}, \quad k \in \{f, r, w, lit, som\} \tag{26}$$

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

$$C_k^{(t)} | C_{-k}^{(t)}, C_{k,obs}^{(t)}, \cdot \sim \mathcal{N}\Big(\frac{\phi_k(A_t C_k^{(t-1)} + b_t + A_{t+1}(C_k^{(t+1)} - b_{t+1})) + \tau_k C_{k,obs}^{(t)}}{\phi_k(1 + A_{t+1}^2) + \tau_k}, \phi_k(1 + A_{t+1}^2) + \tau_k\Big). \tag{27}$$

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

$$C_k^{(t)} | C_{-k}^{(t)}, \cdot \sim \mathcal{N}\Big(\frac{\phi_k(A_t C_k^{(t-1)} + b_t + A_{t+1}(C_k^{(t+1)} - b_{t+1}))}{\phi_k(1 + A_{t+1}^2)}, \phi_k(1 + A_{t+1}^2)\Big). \tag{28}$$

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

$$C_k^{(0)} | C_{-k}^{(0)} \cdot \sim \mathcal{N}\left(\frac{\phi_k(A_1 C_k^{(1)} - A_1 b_1 + \phi_k^{(0)} \mu_k^{(0)})}{\phi_k A_1^2 + \phi_k^{(0)}}, \phi_k A_1^2 + \phi_k^{(0)}\right) \tag{29}$$

$$C_k^{(T)} | C_{-k}^{(T)}, \cdot \sim \mathcal{N}\left(\frac{\phi_k(A_T C_k^{(T-1)} + b_t) + \mathbb{1}_{T \in I}(\tau_k C_{k,obs}^{(T)})}{\phi_k + \mathbb{1}_{T \in I}(\tau_k)}, \phi_k + \mathbb{1}_{T \in I}(\tau_k)\right), \tag{30}$$

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:

$$\phi_k | \cdot \sim \Gamma\left(\frac{T}{2}, \frac{1}{2}\sum_{t=1}^{T}(C_k^{(t)} - f(C_k^{(t-1)}))^2\right), \tag{31}$$

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

We estimated the posterior distributions of latent states, process parameters ($p_{1:11}$), and process parameters using MCMC (Reich and Ghosh 2019) in the R programming language (R Core Team 2016). 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 latent states using their Gibbs sampling updates given in Eqs. (27)–(31), the process precisions using their Gibbs sampling, and initial conditions using a Gibbs sampler (Geman and Geman 1984). We updated process parameters updated using Random Walk Metropolis-Hastings. Including burn-in, 20,000 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.

2.4. Simulation study

We use simulations 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 how changing the observation time resolutions (gaps between observations of the stocks/states) impacts parameter estimates, and whether we can successfully recover parameters under extreme (annual) observation time resolution. The second was to examine how changing the state process resolution (time step of the process model) changes parameter estimates, with a particular focus on an annual observation resolution. 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, USA (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 elliottii*)) that matches the canopy type expected by the DALECev model. Second, the site is part of 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 (National Ecological Observatory Network 2020), 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. The chosen parameter values for the simulations are shown in Table 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.

2.5. Impact of varying observation data resolution on estimation of parameters and latent states

The ability to estimate process precision is crucial in ecosystem models, as that is often the main source of uncertainty (Dietze 2018; Thomas 2017). 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, estimation of model parameters and latent state should be robust to annual or longer data gaps.

To analyze the effect of observation gaps on estimation of parameters (process parameters, process precisions, and latent states), we examined three different observation scenarios: daily state observations, monthly state observations, and yearly state observations, each with daily flux observations. In the context of our study objectives, we are analyzing the effects of varying the data observation resolution while fixing the state process resolution at a daily time step. We drew initial conditions from the prior distributions, and used driver data from Talladega to run each model for six 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 impact of varying the data observation resolution with a fixed daily state process resolution on estimates of process parameters by looking at summaries of their marginal posterior distributions. In particular, we look at the percent bias of the parameters ($100(\cdot\mathbb{E}[\theta]-\theta)/\theta$) and at visualizations of the posterior variance of process precision parameters. In an ideal situation, we would expect to see little bias in our process parameters, with the variance of the marginal posterior distributions increasing as gaps between observations increase.

To evaluate the quality of parameter estimation under different gaps in data, we used MCMC (as described above) to estimate posterior parameter distributions for each synthetic dataset and analyzed the bias and variance of the resulting marginal posterior distributions. We identified the data gaps where a large degradation in parameter estimation occurred.

2.6. Effects of changing the state process resolution on estimation of parameters and latent states

It can be difficult to obtain information about parameters and 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 estimation of parameters and latent states. These problems may arise from differences in the flux data and observation data likelihoods having different time steps. To analyze these differences, we generated data using DALECev with daily latent state resolution and analyzed the synthetic data using a simplified model being run with a monthly latent state resolution. Consider an NDLM with daily process resolution for carbon stock C, with a process model that takes the form of Eq. (7). Let $T^* = \{t_i, \quad i = 1, .., I\}$ be a proper subset of the time steps of the model. For an NDLM with a process model of the form in Eq. (7), state transitions can be rewritten as:

$$C^{(t_i)}|C^{(t_{i-1})} \sim \mathcal{N}\Bigg(\Big(\prod_{j=t_{i-1}+1}^{t_i} A_j\Big)C^{(t_{i-1})} + \sum_{k=t_{i-1}+1}^{t_i-1} b_k\Big(\prod_{m=k+1}^{t_i} A_m\Big) + b_{t_i}, \frac{\phi_s}{1+\sum_{k=t_{i-1}+1}^{t_i}\left(\prod_{m=k+1}^{t_i} A_m^2\right)}\Bigg). \tag{32}$$

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 Eq. (32) as a fixed value $\phi_{monthly}$. Applying this approach, we follow the advice given in Auger-Méthé (2021), who say “make simplifying assumptions when data are limited.”

To examine how parameter and latent state estimates are influenced by changing the state process resolution, we used two different synthetic datasets. The first synthetic dataset is a set of 15 synthetic time-series that have monthly carbon stock observations and daily flux observations. The second synthetic dataset is a set of 15 synthetic time-series that have annual carbon stock observations and daily flux observations. We analyzed each of these data sets using MCMC with our simplified monthly time step model and looked at the percent bias of the process parameters and latent states. Here, we were especially interested in annual carbon stock availability, as it is the most common case when working with actual data.

2.7. Parameter identifiability under different flux data availability

Identifiability of parameters was assessed using data cloning 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 procedure was performed on each of the synthetic datasets with, once with no additional replication, and then with $r = 25$ data cloning replicates (Lele et al. 2007, 2010). Posterior distributions of p_2, p_3, p_4, p_{11} were analyzed across datasets and levels of replication.

Revisiting Eqs. (11)–(13), 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 are most important to collect in their experiments.

3. RESULTS

3.1. Impact of Varying Observation Data Resolution on Estimation of Parameters and Latent States

We expected that as observation data become less frequent, the variance in our posterior estimates of parameters would increase. While this was the case for some parameters (e.g., ϕ_f and p_{11}), we also found other outcomes. Parameters that receive much of their constraint from flux data, such as p_3, had very similar posterior variance independent of the frequency of observations on the stocks. This was not surprising to us, as while p_3 appears in multiple state update equations, there is little additional information that is not contained in the daily flux data. Another outcome that we observed was that parameters such as p_5 had similar posterior variance for daily and monthly observations, but had substantially larger posterior variance for yearly observations. In ϕ_{lit}, we found a very interesting result. As the observation frequency went from daily to monthly, the posterior variance increased and the direction of the bias went from ϕ_{lit} being over-estimated to being under-estimated. Further, as the observation frequency went from monthly to yearly, the variance decreased but the posterior mean moved very far from the true value of $\phi_{lit} = 3.625$. Lastly, we found that some parameters, such as p_9, have increasing posterior variance and small bias as the observation frequency went from daily to monthly. As the observation frequency became yearly the posterior variance continued to increase, but the bias also increased considerably. Indeed, we find that it is very challenging to estimate process parameters and precisions with a daily state resolution and yearly observations. Visualizations of these posterior variances as boxplots can be found in Fig. 2.

Ideally, we would hope that as the observation resolution goes from daily to monthly to yearly, there would be negligible bias in estimates of the process parameters, process precisions, and latent states. While this was true for a number of process parameters (e.g., p_2, p_5, p_7, p_{11}), process precisions (ϕ_f, ϕ_r), and latent states (C_f, C_w, C_r, C_{som}), there were notable exceptions. In particular, for the yearly observation synthetic datasets, there were large percent biases in the estimates of p_1, p_8, ϕ_{lit}, and the C_{lit} latent states. One possible explanation is that C_{lit} is the most dynamic latent process in DALECev. Since direct observations are only available annually, the model needs to rely on the available flux observations to constrain it. These flux observations are a function of three different parameters and the temperature, which may make it difficult to capture C_{lit} dynamics. This would likely influence estimates of process precision and the related parameters as well. Overall we found that the percent bias in estimates of process parameters, process precisions, and latent states was small for the daily observation resolution and the monthly observation resolution analyses, but not necessarily for the yearly observation resolution analyses. Percent biases that were averaged across each of the 15 synthetic datasets can be found in Table 2.

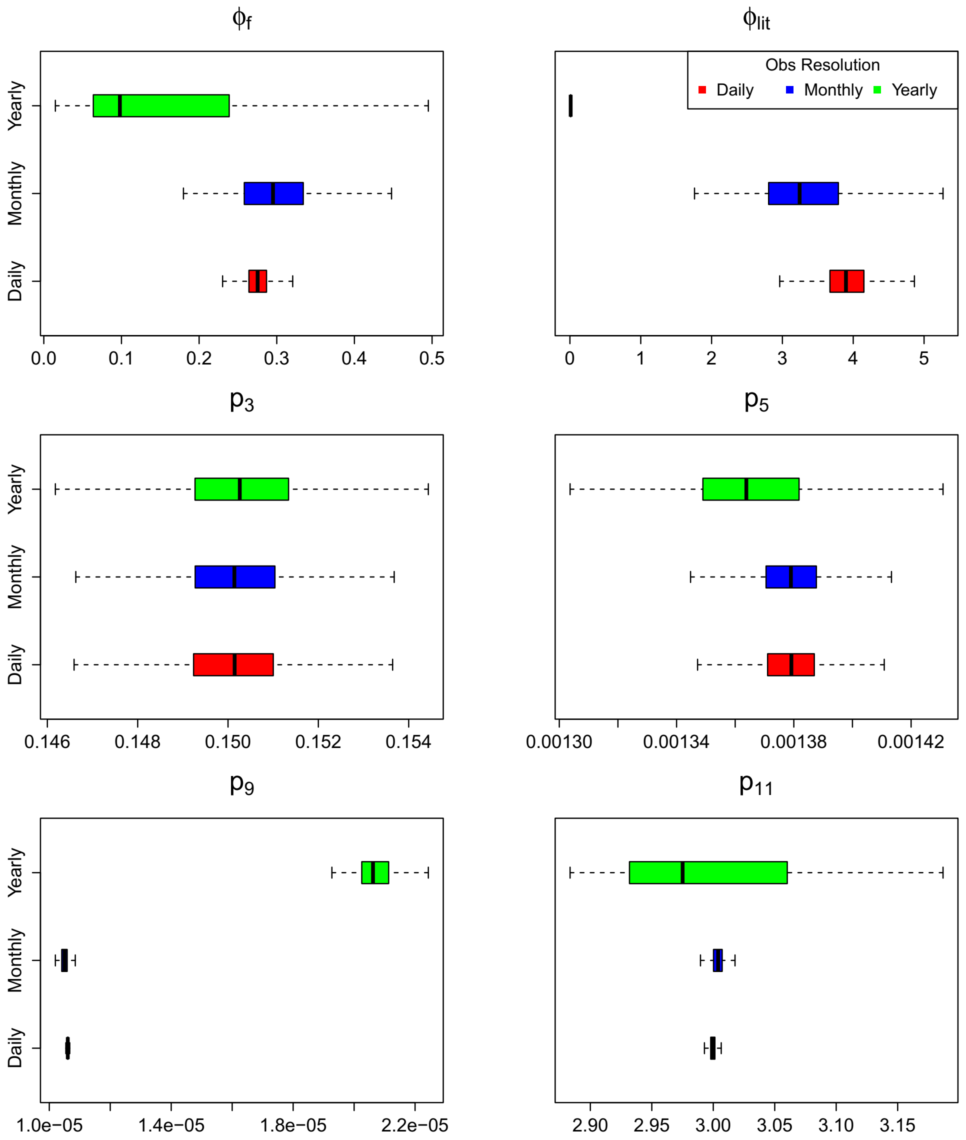

Figure 2. Boxplots of post-burn-in MCMC chains of six different process parameters and precisions, with the observation resolution varying between daily observations, monthly observations, and annual observations. Red (bottom most boxplots) indicates daily observations; blue (middle boxplots) indicates monthly observations; green (upper most boxplots) indicates yearly observations (Color figure online)..

3.2. Effects of Changing the State Process Resolution on Estimation of Parameters and Latent States

We found that by changing from a daily state process resolution model to a monthly state process resolution model, we were able to improve the estimation of our process parameters ($p_{1:11}$) and latent states, particularly for the case of a yearly observation resolution. When using a daily state process resolution, we found considerable biases in estimates of p_1,

Table 2. Percent bias for each parameter for daily, monthly, and yearly observation resolutions

Param.	Daily obs	Monthly obs	Yearly obs
p_1	−1.17	−2.06	−3.16e01
p_2	−5.28e−02	−5.27e−02	−3.88e−02
p_3	−8.03e−02	−7.74e−02	−1.55e−03
p_4	−1.15e−01	−1.13e−01	−5.91e−02
p_5	5.76e−01	4.89e−01	−1.17
p_6	1.711e−02	−5.25e−03	−2.94e−02
p_7	3.18e−03	−1.31e−02	9.55e−02
p_8	−3.45	−3.619	−1.66e01
p_9	−3.69	−5.40	1.11e02
p_{10}	8.33e−01	1.23	−1.57e01
p_{11}	−8.76e−03	−9.42e−03	−1.61
ϕ_f	−1.90	4.19	3.67e01
ϕ_w	−1.22	7.67	−8.04e01
ϕ_r	4.51e−01	10.4	1.34e01
ϕ_{lit}	2.08	−1.92	−9.74e01
ϕ_{som}	−1.40	6.65	−9.73e01
C_f	−2.04e−03	−7.95e−03	1.33
C_w	−3.62e−04	1.28e−02	−2.62e−01
C_r	−3.61e−04	1.58e−02	−1.24-01
C_{lit}	5.43e−03	−5.95e−01	1.65e02
C_{som}	−3.09e−04	5.49e−04	−2.31e−01

Percent bias is computed by taking the posterior means of the MCMC chains after burn-in, subtracting the true value used for data simulation, and then dividing by the true value. For each of the five latent states, the average is taken across all 2190 days. Results are averaged across 15 synthetic datasets

p_9, and C_{lit}. For our monthly state process resolution model, these biases were negligible, but there was significant bias introduced into estimates of p_5, which was not present when using the daily state process resolution model. Biases in parameter estimates obtained from the monthly state process resolution model were similar for the case of monthly observations and yearly observations, and were much better overall than the parameter and latent state estimates that we received from our daily state process resolution model with yearly observations (Table 3).

In addition to estimating the process parameters, 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 Fig. 3, 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 state process resolution. Our monthly state process evolution model was able to accurately capture the dynamics of each of the carbon stocks in our model, even with only yearly observations on the states. Overall, we found that our monthly time resolution model was able to estimate process parameters and latent states with less bias than its daily time resolution counterpart.

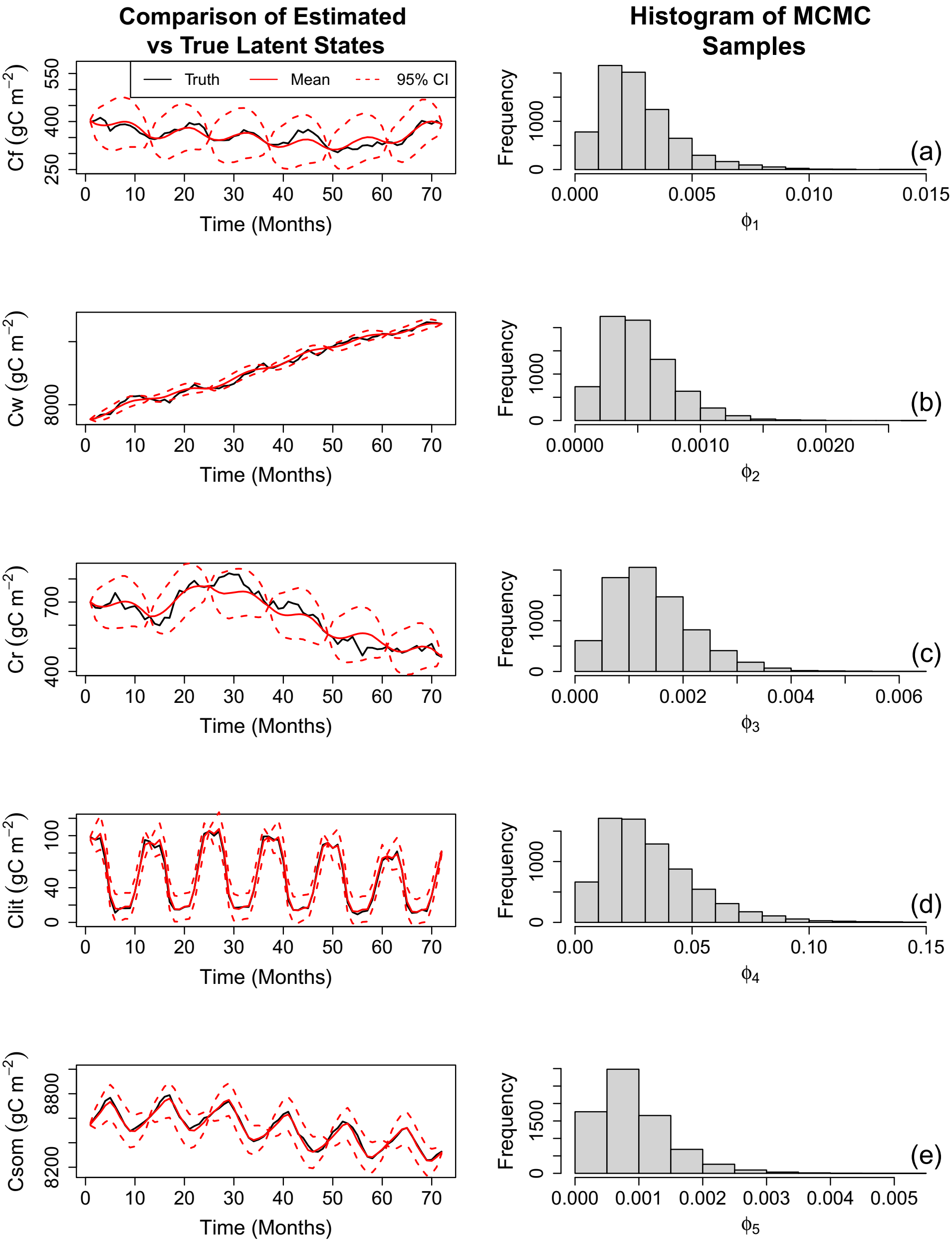

Figure 3. **a**–**e** Estimates for carbon density latent states for a monthly time resolution with annual data observations (left) along with histograms of corresponding marginal estimated process precision post-burn-in (right). Each panel corresponds to a particular carbon stock: C_f represents foliage carbon, C_w represents wood carbon, C_r represents root carbon, C_{lit} represents litter carbon, and C_{som} represents soil organic matter carbon .

Table 3. Percent bias for each parameter and latent state estimate for monthly and yearly observation resolutions, estimated using our monthly state process resolution model

Param.	Monthly obs	Yearly obs
p_1	−4.25e−01	2.92
p_2	−1.04e−01	−1.02e−01
p_3	−2.32e−01	−3.23e−01
p_4	−2.20e−01	−9.22e−02
p_5	−2.77e01	−2.86e01
p_6	−2.23e−02	−4.95e−02
p_7	3.50e−02	2.42e−01
p_8	−5.58e−01	2.12
p_9	−2.27	5.46
p_{10}	5.16e−01	−1.13
p_{11}	−7.57e−03	−7.57e−03
C_f	−2.69e−02	6.72e−01
C_w	1.94e−03	4.71e−02
C_r	2.41e−03	−2.25e−01
C_{lit}	−1.794e−01	2.08
C_{som}	−5.23e−04	3.28e−02

Results are averaged across 15 different synthetic datasets. Percent bias is computed by taking the posterior means of the MCMC chains after burn-in, subtracting the true value used for data simulation, and then dividing by the true value. For each of the five latent states, the average is taken across all 72 latent states. Precision estimates are excluded, as the data was generated under a misspecified model and thus there are no "true" precision values

3.3. PARAMETER IDENTIFIABILITY UNDER DIFFERING FLUX 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 (Ponciano et al. 2012). 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 Ponciano et al. (2012)].

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 Fig. 4 (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

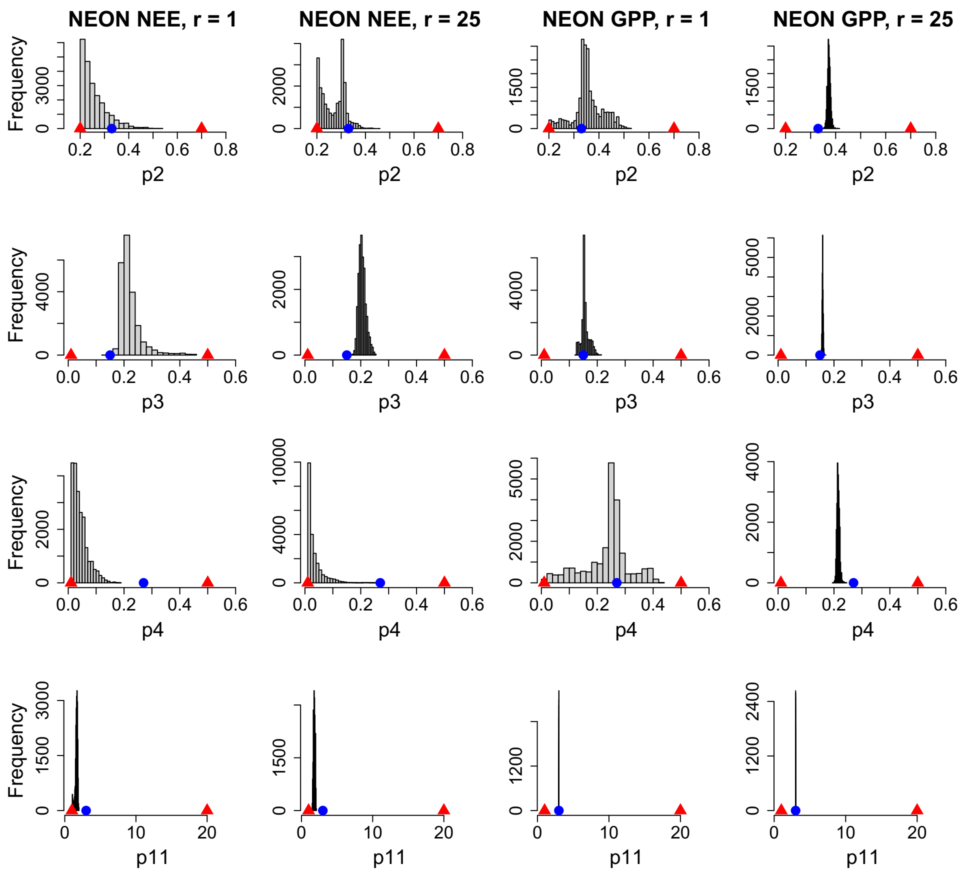

Figure 4. Marginal posterior distributions of data cloning for selected process parameters under the two NEON data flux scenarios, with 1 data cloning replicate ($r = 1$) and 25 data cloning replicates ($r = 25$). Blue circular marks on the x axis denote simulation (true) value of the parameter. Red triangular marks denote the upper and lower bounds of the uniform priors given to the process parameters .

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 Eqs. (12) and (13) and there is insufficient data to identify p_2 or estimate p_{11} well without GPP flux data available. In the NEE scenario, p_2, p_4, and p_{11} all exhibit extreme bias. 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.

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 observations on the states have a coarser temporal resolution than the states, 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 time step of the process model (i.e., time step of the difference equations). Our analysis revealed a large increase in the quality of process parameter estimates, while still capturing the dynamics of the latent states. 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 (identifiability). To examine identifiability 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 occurred 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 [Eqs. (2)–(4)], 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 (Viallefont et al. 1998; Little et al. 2009) and symbolic algebra methods (Cole 2019; Cole and McCrea 2016). Here, we consider long time series, which would lead to problems with numerical stability when using the Hessian method (Bulla and Berzel 2008) and lack of computational resources to perform the symbolic algebra calculations in MAPLE (Cole 2019). Identifiability is a problem that has long plagued ecological and biological modeling (Luo 2009), 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 Lele (2020) 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 (Lele 2020). 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. Third, the Gibbs updates shown in this paper are 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 (Thomas 2017), other applications may have error structures that do not meet this assumption and may not have access to Gibbs sampling. 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. Some possible fitting methods include (but are not limited to): particle methods [see Kantas et al. (2014), for a thorough review], Gaussian process regression (Turner and Deisenroth 2010), hybridizations of MCMC and particle methods (Chopin and Papaspiliopoulos 2020), iterated filtering methods (Ionides et al. 2011), and Laplace approximation (Auger-Méthé et al. 2017). 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 (Bloom and Williams 2015).

In conclusion, to address the growing popularity of state space modeling in ecological forecasting research (Auger-Méthé 2021), 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, nonlinear process models, and spatiotemporal 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.

ACKNOWLEDGEMENTS

We would like to thank the National Ecosystem Observatory Network (NEON) for the data used to create simulations at Talladega National Forest and the REFLEX Project (Fox et al. 2009) for providing several different sets of driver data to test our methodology.

Funding This work was supported by the National Science Foundation Grants DBI # 2016264, DMS/DEB # 1750113, and DEB # 1926388.

Declarations

Conflict of interest The authors declare no conflict of interest.

Availability of data and materials All relevant data and code used to produce the analyses used here are available at https://github.com/johnwilliamsmithjr/EcoSS.

Human and Animal Subjects No human or animal subjects were used in the data used for this research.

Consent to publication All authors consent to the publication of the research presented here.

[Received March 2021. Revised January 2023. Accepted January 2023. Published Online March 2023.]

REFERENCES

Aber JD, Federer CA (1992) A generalized, lumped-parameter model of photosynthesis, evapotranspiration and net primary production in temperate and boreal forest ecosystems. Oecologia 92(4):463–474

Auger-Méthé M et al (2017) Spatiotemporal modelling of marine movement data using Template Model Builder. Mar Ecol Prog Ser 565:237–249. https://doi.org/10.3354/meps12019

Auger-Méthé M et al (2016) State-space models' dirty little secrets: even simple linear Gaussian models can have estimation problems. Sci Rep 6:26677. https://doi.org/10.1038/srep26677

Auger-Méthé M et al (2021) An introduction to state-space modeling of ecological time series. Ecol Monogr 91:e01470

Baldocchi D (2014) Measuring fluxes of trace gases and energy between ecosystems and the atmosphere—the state and future of the eddy covariance method. Glob Change Biol 20(12):3600–3609. https://doi.org/10.1111/gcb.12649

Baracchini T, Wuest A, Bouffard D (2020) Meteolakes: an operational online three-dimensional forecasting platform for lake hydrodynamics. Water Res 172:115529. https://doi.org/10.1016/j.watres.2020.115529

Bloom A, Williams M (2015) Constraining ecosystem carbon dynamics in a data-limited world: integrating ecological common sense in a model-data fusion framework. English. Biogeosciences 12(5):1299–1315. https://doi.org/10.5194/bg-12-1299-2015

Bulla J, Berzel A (2008) Computational issues in parameter estimation for stationary hidden Markov models. Comput Stat 23:1–18

Chopin N, Papaspiliopoulos O (2020) An introduction to sequential Monte Carlo. ISBN: 978-3-030-47844-5

Cole DJ (2019) Parameter redundancy and identifiability in hidden Markov Models. Metron 77(2):105–118

Cole DJ (2020) Parameter redundancy and identifiability. Chapman and Hall/CRC

Cole DJ, McCrea RS (2016) Parameter redundancy in discrete state-space and integrated models. Biom J 58(5):1071–1090. https://doi.org/10.1002/bimj.201400239

Copyright (2022). Integrated population models. In: Schaub M, Kery M (eds). Academic Press, pp 1-622. ISBN: 978-0-323-90810-8. https://doi.org/10.1016/B978-0-12-820564-8.12001-9

Dietze MC et al (2018) Iterative near-term ecological forecasting: needs, opportunities, and challenges. Proc Natl Acad Sci 115(7):1424–1432. https://doi.org/10.1073/pnas.1710231115

Dowd M, Meyer R (2003) A Bayesian approach to the ecosystem inverse problem. Ecol Model 168(1):39–55. https://doi.org/10.1016/S0304-3800(03)00186-8

Durbin J, Koopman S (2012) Time Series analysis by state space methods, 2nd edn. Oxford University Press, English

Fox A et al (2009) The REFLEX project: comparing different algorithms and implementations for the inversion of a terrestrial ecosystem model against eddy covariance data. Agric For Meteorol 149:1597–1615. https://doi.org/10.1016/j.agrformet.2009.05.002

Geman S, Geman D (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans Pattern Anal Mach Intell PAMI 6(6):721–741

Hamilton J (1994) Time series analysis, vol XIV. Princeton University Press, Princeton, p 799. ISBN: 0691042896

Ionides EL, Bhadra A, Atchadé Y, King A (2011) Iterated filtering. Ann Stat 39(3):1776–1802. https://doi.org/10.1214/11-AOS886

Jeffreys H (1946) An invariant form for the prior probability in estimation problems. Proc R Soc Lond. Ser A. Math. Phys. Sci. 186(1007):453–461. https://doi.org/10.1098/rspa.1946.0056

Jiang J et al (2018) Forecasting responses of a Northern Peatland carbon cycle to elevated CO2 and a gradient of experimental warming. J Geophys Res Biogeosci 123(3):1057–1071. https://doi.org/10.1002/2017JG004040

Kantas N, Doucet A, Singh S, Maciejowski J, Chopin N (2014) On particle methods for parameter estimation in general state-space models. Stat Sci—Accept Publ. https://doi.org/10.1214/14-STS511

Landsberg J, Waring R (1997) A generalised model of forest productivity using simplified concepts of radiation-use efficiency, carbon balance and partitioning. For Ecol Manage 95(3):209–228. https://doi.org/10.1016/S0378-1127(97)00026-1

Lele SR (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. Front Ecol Evolut 7:501. https://doi.org/10.3389/fevo.2019.00501

Lele S, Dennis B, Lutscher F (2007) Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov Chain Monte Carlo methods. Ecol Lett 10:551–63. https://doi.org/10.1111/j.1461-0248.2007.01047.x

Lele S, Nadeem K, Schmuland B (2010) Estimability and likelihood inference for generalized linear mixed models using data cloning. J Am Stat Assoc 105(492):1617–1625. https://doi.org/10.1198/jasa.2010.tm09757

Little M, Heidenreich W, Li G (2009) Parameter identifiability and redundancy in a general class of stochastic carcinogenesis models. PLoS One 4:e8520. https://doi.org/10.1371/journal.pone.0008520

Luo Y et al (2009) Parameter identifiability, constraint, and equifinality in data assimilation with ecosystem models. Ecol Appl 19(3):571–574. https://doi.org/10.1890/08-0561.1

Luo Y et al (2011) Ecological forecasting and data assimilation in a data-rich era. Ecol Appl: A Publ Ecol Soc Am 21(5):142942

Metzger S, Others, (2019) From NEON field sites to data portal: a community resource for surface-atmosphere research comes online. Bull Am Meteorol Soc. https://doi.org/10.1175/BAMS-D-17-0307.1

National Ecological Observatory Network (2020) Woody plant vegetation structure, Data Product DP1.10098.001, Provisional data downloaded from http://data.neonscience.org. Accessed 21 Apr 2020

Nielsen A, Berg CW (2014) Estimation of time-varying selectivity in stock assessments using state-space models. Fish Res 158:96–101. https://doi.org/10.1016/j.fishres.2014.01.014

Petris G, Petrone S, Campagnoli P (2009) Dynamic Linear Models with R, vol 38. Springer, pp 31–84. ISBN: 978-0-387-77237-0. https://doi.org/10.1007/b135794_2

Ponciano JM, Burleigh JG, Braun EL, Taper ML (2012) Assessing parameter identifiability in phylogenetic models using data cloning. Syst Biol 61(6):955–972. https://doi.org/10.1093/sysbio/sys055

R Core Team (2016) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna

Reich BJ, Ghosh SK (2019) Bayesian statistical methods, 1st edn. Chapman and Hall/CRC, New York

Riecke TV et al (2019) Integrated population models: model assumptions and inference. Methods Ecol Evol 10(7):1072–1082. https://doi.org/10.1111/2041-210X.13195

Robert CP, Casella G (2005) Monte Carlo statistical methods (Springer texts in statistics). Springer, Berlin, Heidelberg

Rothenberg TJ (1971) Identification in parametric models. Econometrica 39(3):577591

Smallman TL, Exbrayat J-F, Mencuccini M, Bloom AA, Williams M (2017) Assimilation of repeated woody biomass observations constrains decadal ecosystem carbon cycle uncertainty in aggrading forests. J Geophys Res Biogeosci 122(3):528–545. https://doi.org/10.1002/2016JG003520

Thomas RQ et al (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. https://doi.org/10.5194/bg-14-3525-2017

Turner R, Deisenroth M (2010) State-space inference and learning with Gaussian processes. J Mach Learn Res—Proc Track 9:868–875

Viallefont A, Lebreton J-D, Reboulet A-M, Gory G (1998) Parameter identifiability and model selection in capture-recapture models: a numerical approach. Biometr J 40:313–325. https://doi.org/10.1002/(SICI)1521-4036(199807)40:3<313::AID-BIMJ313>3.0.CO;2-2

West M, Harrison J (1997) Bayesian forecasting and dynamic models, 2nd edn. Springer, New York. English

White EP et al (2019) Developing an automated iterative near-term forecasting system for an ecological study. Methods Ecol Evol 10(3):332–344. https://doi.org/10.1111/2041-210X.13104

Williams M et al (1997) Predicting gross primary productivity in terrestrial ecosystems. Ecol Appl 7(3):882–894. https://doi.org/10.1890/1051-0761(1997)007[0882:PGPPIT]2.0.CO;2

Williams M, Schwarz PA, Law BE, Irvine J, Kurpius MR (2005) An improved analysis of forest carbon dynamics using data assimilation. Glob Change Biol 11(1):89–105. https://doi.org/10.1111/j.1365-2486.2004.00891.x

Xu T, White L, Hui D, Luo Y (2006) Probabilistic inversion of a terrestrial ecosystem model: analysis of uncertainty in parameter estimation and model prediction. Global Biogeochem Cycles 20:2. https://doi.org/10.1029/2005GB002468