UNIVERSITY OF CALIFORNIA,
IRVINE

Methods for Bayesian Inference and Data Assimilation of Soil Biogeochemical Models

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Mathematical, Computational, and Systems Biology

by

Hua Wally Xie

Dissertation Committee:
Professor Steven D. Allison, Chair
Professor Elizabeth L. Read
Professor Michele Guindani

2022

DEDICATION

To my wife for being a bedrock of love, a shoulder to lean on, an ear for venting, and a steadfast companion in struggle.

To my parents for their love and filial support and for serving as examples of perseverence with their enduring life experience.

To my daughter, for making finishing up this dissertation that much more intense and exciting and being a reminder that there is no steady state.

To the entirety of feisty hummingbirdkind for the delight and inspiration they offered during the pandemic - may they survive past the anthropocene...

To friends near and far.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

ACKNOWLEDGMENTS

It takes a village to raise a child. It takes a village to build a dissertation!

I thank my advisor and committee chair Steven D. Allison for patiently shepherding me through the pursuit of this degree and introducing me to interdisciplinary research. I had long heard of the importance and necessity of interdisciplinary research for advancing human knowledge, but under Steve's mentorship, I was able taste firsthand the bitter difficulty and savory reward of the endeavor. Through this adventure, I am fortunate to have been exposed a breadth of fields ranging from dynamical systems to soil science. I've been able to encounter (if not fully absorb) wondrous findings sourced from the hard work of scholars past and present, near and far on an almost daily basis. But even more important than the knowledge of scientific results, Steve has guided me to a deep appreciation for team science and collaborative striving. Demonstrating consummate professionalism, respect, and humility in his interactions with people from all walks of life, Steve set an example for me of how collaborations should be built and cultivated. On a related note, I am grateful to Steve for helping me work through my snags and divergences resulting from my exploration of non-academic careers. I would not have come anywhere near the finish line without Steve's focus on collaboration development and all-around humanism.

I thank my commitee members Michele Guindani and Elizabeth L. Read for kindly volunteering their time and energy to emails and committee meetings and providing apt feedback at many turns. Additionally, I thank Michele for introducing me to the broad fields of Bayesian statistics and statistics at large.

Through Michele, I met Erik Sudderth and Debora Sujono. I thank Erik and Debora for their much needed contributions that powered a pivot in my dissertation research when it became clear that I had to adjust from my earlier dissertation plan. Erik's deep statistical and machine learning expertise pointed me to cutting edge variational inference methods that I could explore under computational resource constraints. Debora then spent countless hours in the coding trenches with me, applying, engineering, and debugging myriad variations of approaches. Debora was a brilliant and magnificent collaborator, and I could not have seen the finish line without her help.

I similarly thank the Novocastrian Tom Ryder for rejuvenating and refreshing my dissertation research from across the Atlantic with his deep learning and stochastic differential equation expertise. Tom is a kind, humble soul who volunteered so much of his time punctually providing code and feedback while busy at his day job. Tom embodies the spirit of international friendship.

I thank the members of the Allison Lab for their insights and emotional support. In particular, I am grateful to the maternal Kelly Ramin for volunteering her time as a de-facto PhD life coach and counselor.

I thank Roy Zhao for his companionship and camaraderie in our trials by fire, for being a brother-in-arms in navigating research and life.

I thank my friends Kameran Kolahi, Nikhil Addleman, Lingge Li, Vy Duong, and Daniel Ramirez for their support.

I thank Shashaank Vattikuti and Carson C. Chow for introducing me to research and setting me on a path to graduate school.

I thank Karen Martin for being a meticulous and diligent administrator for the Mathematical, Computational, and Systems Biology PhD program that helped me tame bureaucratic dragons.

I gratefully acknowledge funding from the the U.S. National Institute of General Medical Sciences (grant P50-GM076516) U.S. National Institute of Biomedical Imaging and Bioengineering (grant T32-EB009418), the U.S. National Science Foundation (grant DEB-1900885), and the U.S. Department of Energy Office of Science (grants DE-SC0014374 and DE-SC0020382),

VITA

Hua Wally Xie

EDUCATION

PhD in Mathematical, Computational, and Systems Biology — **2022**
University of California — *Irvine, CA*

BS in Biological Sciences and Radio/Television/Film — **2012**
Northwestern University — *Evanston, IL*

RESEARCH EXPERIENCE

Graduate Research Assistant — **2016–2022**
University of California, Irvine — *Irvine, CA*

Intramural Research Training Award Fellow — **2013–2015**
National Institute of Diabetes and Digestive and Kidney Diseases — *Bethesda, MD*

REFEREED JOURNAL PUBLICATIONS

Hua W. Xie, Adriana L. Romero-Olivares, Michele Guindani, and Steven D. Allison. 2020. "A Bayesian approach to evaluation of soil biogeochemical models." *Biogeosciences.* 17:4043 - 4057.

Shashaank Vattikuti, Phyllis Thangaraj, **Hua W. Xie**, Stephen J. Gotts, Alex Martin, and Carson C. Chow. 2016. "Canonical cortical circuit model explains rivalry, intermittent rivalry, and rivalry memory." *PLoS Comput Bio.* 12:e1004903

REFEREED CONFERENCE WORKSHOPS

Debora Sujono, **Hua W. Xie**, Erik Sudderth, and Steven D. Allison. 2022. "Variational Inference for Soil Biogeochemical Models." *ICML 2022 2nd AI for Science Workshop.*

PREPRINTS

Hua W. Xie, Debora Sujono, Tom Ryder, Erik B. Sudderth, and Steven D. Allison. 2022. "A framework for variational inference and data assimilation of soil biogeochemical models using state space approximations and normalizing flows." *ESSOAr.*

ABSTRACT OF THE DISSERTATION

Methods for Bayesian Inference and Data Assimilation of Soil Biogeochemical Models

By

Hua Wally Xie

Doctor of Philosophy in Mathematical, Computational, and Systems Biology

University of California, Irvine, 2022

Professor Steven D. Allison, Chair

Improving mechanistic understanding and prediction capabilities of long-term organic soil system dynamics is a high priority for biogeochemists, soil scientists, and climate policy researchers who aim to reduce uncertainty regarding changes in the global trajectory of soil carbon sequestration and emissions. While popular "black box" machine learning and classical statistics approaches including XGBoost, LSTM, and ARIMA have been demonstrated to be effective and efficient for time series forecasting, they are not designed to inform on the physical processes underlying a data generating process. Instead, we can turn to *soil biogeochemical models*, also known as soil carbon models, to jointly predict and falsify soil dynamics. Soil biogeochemical models are formulated to simulate the microbe-driven movement of organic elements between terrestrial pools of soil organic matter. As dynamical systems, they provide an avenue to mathematically translate and formalize hypotheses about soil system mechanics into parameterized differential equations.

If we assume that soil biogeochemical models superior at describing empirical soil measurements more closely represent the actual data generating processes, we can surmise that models better at fitting data under biologically realistic parameter regimes are more useful for forecasting purposes; mismatch to data can suggest a need to reparameterize or restructure a model. However, the determination of statistical frameworks that can rigorously assess the capability of models to assimilate observations under compute time and resource limitations remains an open and

unsettled issue. On this note, we will first expound on soil biogeochemical models in greater detail and motivate the use of Bayesian statistical methods as a means of model fitting and parameter inference while incorporating expert uncertainty and beliefs across the first two chapters of this interdisciplinary dissertation. Subsequently, we will demonstrate the use of a contemporary inference algorithm to assimilate two models with the same data set and then compare their goodness-of-fit quantified with Bayesian information criteria and cross-validation metrics. Finally, in the remaining chapters, we will trial the ability of two novel Bayesian soil biogeochemical model inference schemes offering improved computational efficiency to recover observations and parameter values of known synthetic data generating processes and evidence algorithm functionality worthy of future exploration.

CHAPTER 1

An Overview of Soil Carbon Emissions and Soil Biogeochemical Models

1.1 General background

Without further international policy intervention, the Intergovernmental Panel on Climate Change (IPCC) estimates in their landmark Fifth Assessment Report that under their "Business-as-Usual" emissions scenario, called Representative Concentration Pathway 8.5 (RCP8.5), global mean temperature could rise by almost 5 °C over pre-Industrial baselines by 2100 (Pachauri et al., 2014). More extreme scenarios assuming further breakdowns in negative climate feedbacks posit even larger temperature increases over 8 °C (Schneider et al., 2019; Steffen et al., 2018). Warming substantially beyond a debated "safer limit" of 2 °C by the end of the century, which is highly likely (Armstrong McKay et al., 2022; Kemp et al., 2022; Lenton et al., 2019; Raftery et al., 2017), would accelerate the ongoing biodiversity extinction crises to a point where even human existence teeters in the face of societal collapse (Ehrlich and Ehrlich, 2013; Ord, 2020; Richards et al., 2021).

As demonstrated by the recent passage of the Inflation Reduction Act of 2022 in the United States that provisioned funds for climate initiatives, there have been modest shifts in various national policies away from RCP8.5 and "Business-as-Usual" assumptions (Pielke and Ritchie, 2021). However, governmental progress on climate mitigation remains far too incremental to correspond to immediate emissions improvements (Berrang-Ford et al., 2021; Nascimento et al., 2022; Peters et al., 2020; Steffen et al., 2018), and in the face of limitations in data availability and forecasting methods, it is uncertain how much these modest improvements can ultimately stem warming below catastrophic thresholds (Kemp et al., 2022; Meng et al., 2021; Otto et al., 2015; Palmer and Stevens, 2019; Pedersen et al., 2020; Sognnaes et al., 2021).

Part of the extensive uncertainty in temperature increase outcomes relates to the unknown of how climate feedbacks influenced by soil microbes will respond to policy shifts, land-use change, acidification, and warming itself, among other factors (Armstrong McKay et al., 2022; Beillouin et al., 2022; Bradford et al., 2016; Friedlingstein et al., 2014; Holden et al., 2018; Luke and Cox, 2011; Raza et al., 2021; Schaphoff et al., 2013; Shi et al., 2018; Varney et al., 2020). Soil organic carbon (SOC) constitutes the largest terrestrial pool of carbon (C) (Scharlemann et al., 2014), and a shift in soil microbe populations and activity could switch the pool from a C sink to a massive C source contributing to a runaway "Hothouse" climate change regime (Armstrong McKay et al., 2022; Crowther et al., 2016; Hicks Pries et al., 2017; Steffen et al., 2018; Zomer et al., 2017). It is thereby of utmost importance to climate scientists and policy makers that we can model and predict soil microbe responses and C sequestration capacity with more certainty.

1.2 Soil biogeochemical models offer interpretable mechanistic prediction of soil system dynamics

Accurate prediction and mechanistic understanding of biological soil systems are both critical for guiding the implementation of strategies targeting the preservation of SOC as a C sink. Non-interpretable and non-dynamical machine learning models solely optimized for prediction typically offer limited insights regarding system function. Hence, we turn to a class of models called soil biogeochemical models (SBMs)—also known as soil C models—to generate falsifiable and interpretable claims about organic soil system mechanics influenced by microbial activity and predictions about SOC sequestration and heterotrophic soil CO_2 efflux in tandem (Manzoni and Porporato, 2009).

SBMs are *dynamical systems*. Dynamical systems broadly describe the evolution of one or more *state variables* in response to changes in at least one *domain variable*. A tuple of state variable values associated with a domain value can be referred to as a single multi-dimensional state or point. If a dynamical system describing a physical or biological process only has one domain variable, that system is described as being *first-order* in domain and its domain variable is typically time (Irwin and Wang, 2017; Strogatz, 2018). Some more complex dynamical systems model state variable dependence on space in addition to time (or even more domains than that). However, since SBMs by and large have been formulated solely with time dependency (Manzoni and Porporato, 2009; Sierra and Müller, 2015), we will only discuss first-order dynamical systems within the scope of this thesis (and not engage with the rare SBMs described by partial differential equations).

Differential equations are frequently used as the type of function relating states to time (Strogatz, 2018). They are effective tools for describing dynamical systems with their organization of state variable input into functions on the right hand side and the output of state variable derivatives

with respect to the continuous progression of time on the left hand side. The constrained translation and codification of system mechanics into the form a finite, precise set of differential equations to formulate dynamical systems sources the interpretability of dynamical systems (Zhao and Park, 2016). With specific regard to SBMs, differential equations additionally facilitate the clear mathematical formalization and enforcement of five related biophysical properties associated with organic soil systems (Sierra and Müller, 2015). These five properties are:

1. decomposition rate heterogeneity of different soil organic matter (SOM) types,
2. the tendency of SOM toward transformation through biochemical reactions and impermanence,
3. substrate dependence of those reactions,
4. mass balance of those reactions,
5. and the existence of variable environmental influences external to the soil systen (e.g. changing air temperatures and matter injections from aboveground vegetation).

Various notational conventions have been traditionally used to render state derivatives. Throughout this thesis, we adhere to the derivative notation set forth by Gottfried Leibniz, where the derivative of a variable x with regard to time is written as $\frac{\mathrm{d}x}{\mathrm{d}t}$. But, other notational schemes commonly associated with dynamical systems include that proposed by Isaac Newton, where the derivative of x is instead written as $\dot{x}$.

The state variables of SBMs are frequently assigned as C densities or masses held in SOM *pools* (Manzoni and Porporato, 2009). Pools are a simplifying biogeochemistry modeling notion assuming the presence of separate collections of soil organic matter in soil systems that can grouped by homogeneous qualities such as equivalent decomposition rates. In empirical practice, orderly isolation of separate pools within soil samples has been difficult to demonstrate, so the conception of pools is controversial (Crow et al., 2007; Lajtha et al., 2014; Moni et al., 2012;

Schmidt et al., 2011). Nonetheless, pool-based SBMs have been able to accurately replicate observed soil system dynamics and phenomena (Manzoni and Porporato, 2009; Sierra and Müller, 2015).

While single pool models can reproduce empirical dynamics surprisingly well (Todd-Brown et al., 2018), most SBMs have been formulated to be multi-dimensional in state to afford greater capacity and flexibility in the ability of models to maintain the principles of decomposition rate heterogeneity and environmental variability (P. Smith et al., 2020). The addition of more pools and parameters enables the simulation of wider ranges of temporal responses and behaviors (Manzoni and Porporato, 2009; Sierra and Müller, 2015). Furthermore, some SBMs have been established to simulate the densities or masses of other non-C elements that are highly represented in organic material including nitrogen (N) and phosphorous (P), but these models are also outside the scope of this thesis, which will focus on C-only state variable models.

Up until the 2000s, almost all SBMs were parameterized with ordinary differential equation (ODE) dynamical systems maintaining linear decay and mass transfer kinetics across all pools (Feller and Bernoux, 2008; Manzoni and Porporato, 2009; Sierra and Müller, 2015). Under linear decay assumptions, pool decomposition is determined by singular rate parameters invariant to pool densities (Allison et al., 2010; Berardi et al., 2020). Prominent linear SBMs include the CENTURY model along with its variants (Lawrence et al., 2019; Parton, 1996; Parton et al., 1987), one of which is a constituent part of the fifth version of the coupled Community Earth System Model. Other established examples of linear SBMs include the DNDC (C. Li et al., 1992) and RothC models (Jenkinson and Rayner, 1977). A simplified treatment of linear decay SBMs is presented in Allison et al. (2010) as the so-called conventional model (CON).

Individual differential equations describing a dynamical system belonging to the class of linear

decay SBMs can be customarily written as

$$\frac{\mathrm{d}C_1}{\mathrm{d}t} = \mathcal{I}_1(t) - k_1(t) \cdot C_1 + \sum_{i \in \mathcal{D}-\{1\}} (\mathscr{T}_{1,i}(t) \cdot k_i(t) \cdot C_i) \tag{1.1}$$

$$\frac{\mathrm{d}C_2}{\mathrm{d}t} = \mathcal{I}_2(t) - k_2(t) \cdot C_2 + \sum_{i \in \mathcal{D}-\{2\}} (\mathscr{T}_{2,i}(t) \cdot k_i(t) \cdot C_i) \tag{1.2}$$

$$\vdots \tag{1.3}$$

$$\frac{\mathrm{d}C_d}{\mathrm{d}t} = \mathcal{I}_d(t) - k_d(t) \cdot C_d + \sum_{i \in \mathcal{D}-\{d\}} (\mathscr{T}_{d,i}(t) \cdot k_i(t) \cdot C_i) \tag{1.4}$$

$\mathcal{D} = \{1, 2, ..., d\}$ is the set of integers indexing the SOM pool C density state variables of an SBM, where d denotes the final state dimension. We use the notation, $\mathcal{D} - \{j\}$, where $j \in \mathcal{D}$, to signify the subtraction of element j from the set. $\mathcal{I}_j$ in each equation indicates optional exogenous C input into each pool (e.g. C from surface vegetation litter decomposition external to the soil system) and can be specified as a constant or the output of a time-variant function.

The negative $k_j \cdot C_j$ term is the key expression that mathematically prescribes the linear decomposition (and ensuing element mass transfer) of state C_j. In establishing decomposition, the equations enshrine adherence of an SBM to the properties of SOM impermanence, decay rate heterogeneity, substrate dependence, and mass balance. $k_{j \in \mathcal{D}}$ are the decay rate parameters, which can be modeled as constants, but are typically defined to be the output of time-dependent, season-dependent, temperature-dependent, moisture-dependent or other-factor-dependent functions modifying baseline "reference" parameter values to then honor the environmental variability principle.

The summation term in every equation encompasses endogenous (i.e. system-internal) C input into a pool drawn from C exiting other pools. It is biochemically unrealistic and inaccurate to assume complete absorption and recycling of decomposed matter into news pools. Partial input is executed in SBMs through the deployment of $\mathscr{T}$ *transfer fractions*, which are constant model parameters or function outputs on the interval $[0, 1]$ that ration out pool uptake through

multiplication. $\mathscr{T}$ are also referred to as *partition fractions* or *transfer coefficients*.

The above individual equations can be consolidated into matrix and vector notation as

$$\frac{\mathrm{d}C}{\mathrm{d}t} = \mathcal{I}(t) + \mathscr{T}(t) \cdot K(t) \cdot C \tag{1.5}$$

where $C = \left[C_1 \; C_2 \; ... \; C_d \right]^\intercal$ is now a multidimensional state vector and $\mathcal{I}(t) = \left[\mathcal{I}_1(t) \; \mathcal{I}_2(t) \; ... \; \mathcal{I}_d(t) \right]^\intercal$. $\mathcal{T}$ is a $d \times d$ matrix in which the index $(i, i) \; \forall \; i \in \mathcal{D}$ elements down the matrix diagonal are -1 to indicate complete loss of decomposed C from the pool (Sierra and Müller, 2015). The non-diagonal elements of each $\mathcal{T}$ row are the transfer fraction constants or function outputs on $[0, 1]$ depicting the proportion of decomposed elements being absorbed by the pool corresponding to that row; the appearance of a 0 element signals the inability of a pool to accept input from some source. Finally, K is normally a diagonal-only matrix in which the (i, i) element denotes the constant or environmentally-influenced decomposition rate of the ith pool.

As regulation of the decomposition and mass transfer of individual organic C pools is abstracted and simplified into singular function outputs or parameters in these linear decay SBMs, they are also commonly referred to in the field of biogeochemistry as "microbial-implicit" models (Schimel, 2001; Wieder, Allison, et al., 2015). Meanwhile, SBMs that mechanistically describe C decomposition with more biologically realistic and mathematically sophisticated rate-limited functions acknowledging enzyme catalyses kinetics are called "microbial-explicit" models (Allison et al., 2010; Schimel, 2001; Tang, 2015; Tang and Riley, 2013; Wieder, Grandy, Kallenbach, Taylor, et al., 2015).

These microbial-explicit models represent the decomposition of at least one pool and resulting endogenous C uptake into one or more other pools with nonlinear, state-variant functions which, through the present, have typically been multiplicative or Michaelis-Menten in nature (Allison et al., 2010; Blagodatsky et al., 2010; J. Li et al., 2014; Manzoni and Porporato, 2007; Schimel and Weintraub, 2003; Shi et al., 2018; Sinsabaugh and Follstad Shah, 2012; Sulman

et al., 2014; G. Wang et al., 2013; Y. P. Wang et al., 2014; Wieder, Grandy, Kallenbach, and Bonan, 2014; Wieder, Grandy, Kallenbach, Taylor, et al., 2015; Zelenev et al., 2000). Michaelis-Menten functions can in turn be subcategorized into expressions of original Michaelis-Menten, reverse Michaelis-Menten, equilibrium chemistry approximation, or quadratic kinetics assumptions (Tang, 2015; B. Wang and Allison, 2019). To notate SBMs in a manner that accommodates heterogeneous systems featuring both linear and nonlinear decomposition, we can substitute $K(t)$ in (1.5) with the less specific multivariate functional operator $F(C, t)$ such that

$$\frac{\mathrm{d}C}{\mathrm{d}t} = \mathscr{I}(t) + \mathscr{T}(t) \cdot F(C, t) \cdot C \tag{1.6}$$

In the subsequent thesis chapters, we will introduce relevant SBM systems in individual uncompressed equation rather than consolidated matrix form for readability and interpretability purposes. The dynamical systems we work with are small enough in dimensionality such that the verbosity cost is tractable and manageable. The highest-dimensioned systems we will interact with are from the nonlinear, microbial-explicit AWB family of systems, which have four state variables.

1.3 Employing Bayesian inference in model refinement loops accounting for mechanistic uncertainty

The establishment of mechanistic models to account for the biophysics and kinetics underpinning complex biological data-generating processes like soil systems brings substantial *inductive uncertainty*, which is uncertainty derived from incomplete knowledge regarding system operations (Pawitan, 2001). In contrast, in the unlikely chance that a model correctly and fully explains the persistent dynamics governing a data-generating process, model uncertainty is entirely derived from the expression of nonstandard, incidental influences during data-generation constituting *stochastic uncertainty* (Pawitan, 2001). One common source of inductive uncertainty arises from

the proposal of dynamics which are inconsistent with unelucidated environmental reality. Another likely source is associated with the practical needs of modelers to make simplifying assumptions in the conceptualization of model structure that promote computational tractability and feasibility but diverge the model from the complexities of the true data-generating process. Parameters introduced to ensure stability and constrain mathematical behavior of a model may either collapse multiple biophysical processes into one or not correspond to any simple biological indicators. As a result, model parameters frequently cannot be directly measured empirically, even if they are interpretable with clear definitions.

Thus, parameter values most often cannot be directly measured and instead must be inferred and calibrated based on information from data that can be practically observed. In the case of SBMs, their parameters can ideally be inferred based on the comparison of observations of soil pool and CO_2 efflux measurements to model output. Recalling that SBM state variables are themselves pool densities, CO_2 is then typically defined for SBMs as a pool-dependent *explicit algebraic variable*, which is a variable that is a function of one or more states. To clarify, explicit algebraic variables (also termed as state-dependent observations in machine learning literature) are not states themselves and their time derivatives do not correspond to the differential equations of a system. Formulation of the CO_2 function is frequently part of the SBM conceptualization process (Manzoni and Porporato, 2009).

To collate an empirical data set that is suitable for parameter inference, the inclusion of CO_2 soil efflux observations is usually essential. It is hard and labor-intensive to physically isolate and directly measure SOC and harder still to measure the densities of other model SOM pools that are even less tightly delimited than SOC, so applicable pool observations from individual longitudinal soil experiments are normally in short supply (P. Smith et al., 2020). On the other hand, being a specific compound rather than an amorphously classified pool of molecules, CO_2 is are easier to quantify and record. An extensive set of literature and experimental techniques exists that documents procedures for separating CO_2 from soil gas emissions and then filtering captured

CO_2 by respiration source (e.g. autotrophic versus heterotrophic respiration, mycorrhizal versus free-living microbe respiration) (Lankreijer et al., 2003; Moore et al., 2015; P. Smith et al., 2008; P. Smith et al., 2010). Thus, supplementing sparse pool measurements with denser soil CO_2 flux observations can be instrumental for inferring and estimating parameters.

So, once a data set of soil respiration and pool observations has been compiled, how do we algorithmically perform data assimilation (also called model-data integration, as in Bradford et al. (2016)) to condition an SBM on that data and arrive at inference results that indicate the likelihood of the model for serving as the data-generating process? For an approach that accommodates varying *a priori* certainty and biological realism constraints surrounding parameter value ranges as a reflection on model beliefs and mechanistic inductive uncertainty, we can apply and develop methods from the Bayesian paradigm of statistics.

The foundations of Bayesian statistics originated in large part from the seminal contributions of James Bernoulli and Pierre-Simon Laplace to the philosophy and theories of probability from the 16th and 18th centuries (Jaynes and Justice, 1986). Pierre-Simon Laplace is credited by statisticians and historians with the formalization of what is now commonly known as Bayes' theorem (Jaynes and Justice, 1986; Stigler, 1990), the all-important rock of Bayesianism that anchors its methods and interpretations of probability. In adherence to Stigler's Law of Eponymy, which posits that scientific discoveries are rarely named after their originator (Stigler, 1980), we note obviously that neither the name of the theorem nor the field bear Laplace's name. Instead, the name follows Thomas Bayes, who posthumously documented a special case of Bayes' theorem before Laplace's independent rediscovery (Stigler, 1983).

Of course, Bayesian statisticians maintain a range of opinions and beliefs regarding appropriate applications of Bayes' theorem and do not unanimously agree on probability interpretations (Good, 1971). However, to preface with a non-rigorous summary of Bayesian methods and practioners, they assume that model parameters are random variables distributed in some manner such that models are inherently random and uncertain. Fixed data are used to update beliefs about the

parameter distributions by matching model output to the data as closely as possible. This procedure of matching is varyingly synonymized across different subfields and disciplines with the verbs 'fitting,' 'conditioning,' 'assimilating,' 'calibrating,' 'harmonizing,' 'training,' 'integrating,' and more, so some of these terms may accordingly be more familiar than others. We use these verbs intercheageably in this interdisciplinary thesis. The initially proposed parameter distributions are known colloquially as *prior distributions* or *priors*. The updated distributions are known as *posterior distributions* or *posteriors*.

Hence, the Bayesian perspective of persistent model and model parameter randomness is convenient for SBM comparison and validation. It accommodates mechanistic and stochastic uncertainty, which are in high supply for SBMs, as biogeochemical modelers are highly uncertain about the correspondence between true data-generating and proposed model dynamics and the potential influence of external environmental factors on top. In contrast, techniques and devotees associated with the other dominant probabilistic viewpoint in statistics called *frequentism* by default hold that data is a random variable sampled from a distribution, while parameters are fixed and can be deduced in the long-run from repeated experiments and data samples. Under frequentist probability interpretations, representation of model uncertainty is less straightforward. Additionally, the time and labor costs of most soil experiments preclude repeatability in alignment with long-run assumptions. Of course, this paragraph offers a pithy, simplified encapsulation of the difference between Bayesianism and frequentism; one can start with Jaynes and Justice (1986) and Stigler (1990) for deeper treatments on the debates inciting the two camps.

Figure 1.1 illustrates the basic nodes of a Bayesian SBM refinement and validation loop for a single model stepping from dynamical system equation proposal to prior specification to posterior estimation and so forth. Each node of this workflow is associated with a presently unresolved cluster of questions and issues in the disciplines of biogeochemistry and soil science relating to the construction of SBM evaluation testbeds to benchmark, troubleshoot, compare, and select SBMs for predictive and mechanistic precision (Luo et al., 2016; Wieder, Allison, et al., 2015).

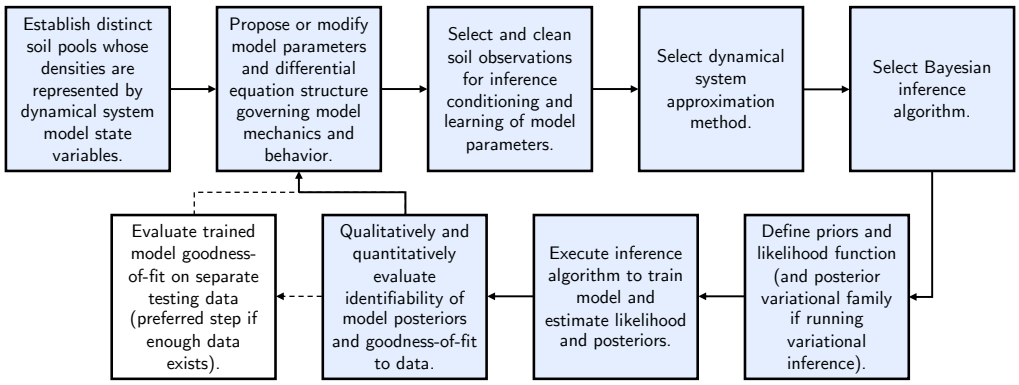

Figure 1.1: A schematic depicting the SBM validation and refinement loop for a single model. We assume the use of a vetted Bayesian inference algorithm; otherwise, the loop must be expanded to acknowledge validation of the algorithm itself. In an ideal situation, enough empirical observations exist for segmentation of the data non-overlapping training and testing sets. In practice, due to the financial, temporal, and labor costs of conducting longitudinal soil manipulation experiments and maintaining regular collection of measurements, data is limited and must be conserved for effective model training (if even enough for that). With respect to the planning and design of future experiments, we recommend that redundant data collection practices be prioritized as much as resources allow to better facilitate train-test splits in data from experiment outset.

This dissertation makes a contribution to the exploration, informing, and motivation of best practices linked to execution of the Figure 1.1 steps. In the next chapter, we survey recent developments in the quantification and formalization of soil ecosystems with mathematical models including SBMs and establish areas of research need in the advancement of SBM-data integration. In chapter three, we novelly demonstrate the use of a prominent Bayesian inference algorithm called the No-U-Turn sampler to assimilate two different SBMs to a simple data set and compare their predictive accuracy with contemporary Bayesian goodness-of-fit metrics. Finally, in chapters four and five, we describe the implementation and validation of two novel SBM inference algorithms drawn from the variational inference area of Bayesian statistics that leverage ma-

chine learning approximation techniques for computational efficiency gain. This work can benefit biogeochemists, soil scientists, climate modelers, and environmental policy-makers who seek to effectively assess SBMs, construct compatible model-training data sets, evaluate predictions of soil C sequestration and emissions, or design long-term soil experiments with SBM inference in mind from the outset.

CHAPTER 2

Advancing Quantitative Models of Soil Microbiology, Ecology, and Biochemistry

2.1 Introduction

Soils host diverse biological communities, including plants, animals, and microbes. Together, these communities provide benefits essential for ecosystem functioning and human well-being. Decomposition of organic matter–primarily driven by microbes–regenerates nutrients that support plant growth in agricultural and unmanaged systems. In turn, plant growth and microbial transformations of organic matter enhance soil carbon (C) sequestration that mitigates greenhouse gas emissions from human activities.

At the same time, the biological services provided by soils are vulnerable to human-caused environmental change (Cavicchioli et al., 2019; Jansson and Hofmockel, 2020). For example, there is concern that global warming will stimulate metabolic activity in soils, weakening C sequestration and potentially turning soils into a net source of greenhouse gases (Davidson and Janssens, 2006). Given these concerns, soil microbes and biological processes are topics of intense research

interest.

Improvements in sequencing technologies and other approaches for probing biological diversity and functioning have led to rapid advances in fundamental knowledge of soil ecology (Bahram et al., 2018). In parallel with these empirical advances, mathematical models of soil systems have blossomed recently (Allison, 2017; Wieder, Allison, et al., 2015. Foundational models of soil biogeochemistry developed during the 1980s and 1990s have been joined by a new generation of biologically-inspired models starting in the early 2000s. Since then, these models have increased in scale and complexity.

Still, there is room for additional model improvement and intellectual development. Large-scale models struggle to replicate fundamental patterns in soil biogeochemical pools and fluxes (Todd-Brown et al., 2014; Todd-Brown et al., 2013; Wu et al., 2018). Many of the most recent models with updated biological mechanisms have not been tested extensively. The field of soil ecological modeling has come a long way, but the pathway to addressing soil-relevant challenges with models remains uncertain.

In an effort to elevate the relevance and impact of soil modeling, this chapter aims to summarize the current state of the art while providing guidance for next steps to advance the field. We discuss some of the main reasons for engaging in soil modeling and then review selected modeling approaches from molecular to global scales. This review does not attempt to be exhaustive, and we focus our attention primarily on advances from the past 5-10 years, especially since the publication of Parton et al. (2015). The chapter concludes with recommendations for model-data integration and future intellectual development.

2.2 Justification for modeling

As with empirical approaches, soil scientists use models to address a range of different goals and questions. Models play an important role in advancing fundamental understanding of soil processes by representing concepts and mechanisms in a quantitative framework. For instance, the *priming effect* is a common biological mechanism in soil whereby addition of fresh organic matter stimulates, or "primes," the decomposition of existing soil C that may be older and more resistant to decay (Fontaine and Barot, 2005). Soil researchers have developed models that represent this mechanism, thereby quantifying the magnitude and impact of priming effects in soil systems (Guenet et al., 2016).

Models are also useful for generating hypotheses. Koven et al. (2015) used a depth-resolved version of the Community Land Model (CLM4.5BGC) to simulate permafrost thaw and its effects on ecosystem C balance. This version of the model is notable for incorporating fundamental understanding of how soil processes vary with depth, a crucial concept in frozen soils with seasonal changes in active layer thickness. Moreover, CLM4.5 represents nitrogen (N) dynamics which likely play into carbon-climate feedbacks. In response to climate warming, the modeling study of Koven et al. (2015) suggested that the positive effects of N release on C storage would be outweighed by the negative effects of permafrost thaw and increased microbial metabolism with soil warming. This outcome is a testable hypothesis that can be addressed with laboratory, field, and global change experiments (Mack et al., 2004; Xue et al., 2016).

More broadly, models can help guide experimental work. A conceptual paradigm proposed by Blankinship et al. (2018) calls for better integration between theory, models, and measurements. This aim could be partially achieved by aligning modeled mechanisms and outcomes with experimental data. For example, models of soil biogeochemistry include a wide array of pools ranging from largely inert to mineral-associated organic matter and highly dynamic microbial biomass. Aligning these pools with the chemical composition of real soils provides a rationale for exploit-

ing cutting-edge organic matter fractionation and characterization approaches, such as nuclear magnetic resonance (NMR), X-ray microspectroscopy, and pyrolysis mass spectroscopy (Kalbitz et al., 2003; Lehmann et al., 2008; Quideau et al., 2005). Likewise, recent advances in modeling microbial diversity can drive new approaches for analyzing sequencing and other datasets that probe the functioning of microbial communities. Building a model can generate practical guidelines for distilling, organizing, and processing the information contained in complex 'omics datasets.

Scaling is another relevant application of soil models (Allison, 2017; Wieder, Allison, et al., 2015). Nearly all of the grand challenges facing soils at the global scale require knowledge of emergent properties arising from smaller spatial scales and shorter time scales. At the molecular level, cells exchange metabolites, enzymes catalyze reactions, and organic compounds interact with mineral surfaces. At cellular to ecosystem scales, these molecular processes combine into emergent biological properties such as growth and respiration. All the way up to the global scale, biological systems interact with soil physical properties to determine outcomes like C and nutrient balance. Modeling offers a quantitative, rational approach for representing key emergent properties at ever-increasing scales. Nested sets of models can, for example, provide insight on how Michaelis-Menten enzyme kinetics at the molecular level scale up to control organic matter decomposition rates at the community scale (Tang and Riley, 2013; B. Wang and Allison, 2019).

Models are also the primary tool available to scientists for making predictions, particularly in the context of global environmental change (Bradford et al., 2016; Todd-Brown et al., 2011). In many studies, the goal of prediction complements other modeling aims such as advancing fundamental understanding, generating hypotheses, and scaling up. Although predictions remain highly uncertain, soil models offer the potential to apply empirical and theoretical advances to simulate C and nutrient pools at the scale of the entire planet, decades or centuries into the future. Such models can provide answers to scientists and decision makers concerned about the future state of soils, such as their capacity to store C in the face of climate and land use change

(IPCC, 2013). The increasing prominence of model outputs in IPCC reports and policy making emphasizes that prediction as a relevant, if not always singular, goal of model development.

2.3 Modeling approaches

Across scales, including the ecosystem scale, differential equation models are often applied to track soil biogeochemical pools and fluxes. Sierra and Müller (2015) described a general framework for this type of soil model based on first principles of mass balance, substrate dependence, heterogeneity of decomposition rates, chemical transformations, variation in environmental drivers, and interactions among soil pools. Nearly all existing models of soil biogeochemistry fit under this general framework, allowing for rigorous comparison of model stability and mathematical properties.

Differential equation models like RothC and Century emerged in the late 1970s and 1980s (Jenkinson and Rayner, 1977; Parton et al., 1988), embracing the principles of mass balance and substrate dependence as envisioned by Olson (1963) with organic matter decaying in proportion to its concentration. These models further included the principle of heterogeneity by representing different pools of organic matter with different decay rates. Transfers among the pools were allowed, following the principle of chemical transformations, and decay rates were functions of temperature and moisture levels, consistent with the principle of varying environmental drivers. Bosatta and Ågren (1985, 1999) generalized the principle of heterogeneous decomposition in their theory of continuous organic matter quality which was intended to better reflect the complexity and diversity of soil organic compounds.

Models like RothC and Century have some convenient mathematical properties, but they omit the fundamental principle of interacting soil pools in the framework of Sierra and Müller (2015). Commonly known as "linear" or "first-order," differential equation models without complex depen-

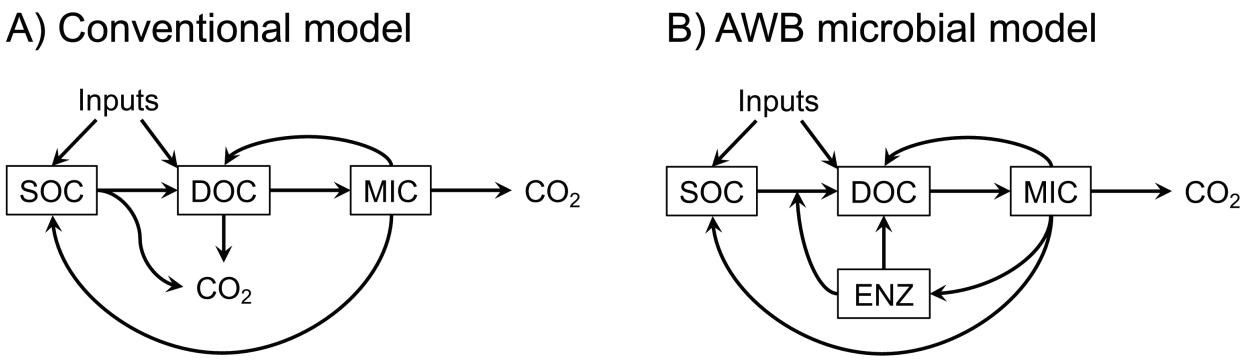

Figure 2.1: **A)** Conventional first-order linear model with microbial implicit transfers among pools. **B)** Allison-Wallenstein-Bradford (AWB) model with microbial explicit interactions among pools of soil organic carbon, dissolved organic carbon, microbial biomass, and extracellular enzymes. Adapted from Allison et al. (2010).

dencies among pools can be readily represented in matrix form and solved analytically (Xia et al., 2013). They also tend to be mathematically stable, meaning that pool sizes and fluxes do not oscillate as the system returns to steady state following perturbation. Despite these advantages, linear models simplify or omit mechanisms of interaction among organic matter pools, such as the enzymatic degradation driven by microbial decomposers (Figure 2.1). Rather, the biological roles of microbes in linear models are assumed to be "implicit" (Schimel, 2001).

An alternative approach to account for the principle of soil pool interactions is to make microbial mechanisms mathematically "explicit." The idea of microbial control over soil biogeochemical processes dates back to at least Waksman (1927). In the late 1970s, O. L. Smith proposed a complex model of soil microbial biogeochemistry that included many of the features described by the general framework of Sierra and Müller (2015), but did not receive much attention (O. L. Smith, 1979a, 1979b). More recently, there has been an explosion of microbially explicit model development and applications (Abramoff et al., 2018; Allison et al., 2010; Fontaine and Barot, 2005; Schimel and Weintraub, 2003). Although they attempt to represent biological mechanisms with higher fidelity, challenges remain with the stability, interpretability, and scaling of microbially-explicit-models (Y. P. Wang et al., 2014). Efforts to analyze microbial processes with models at different scales could help address some of these challenges (Allison, 2012; Kaiser et al., 2014).

Dynamical differential equation models are valuable for representing fundamental processes, but predictive statistical models are a valuable alternative approach. Process-based models with many differential equations require careful parameterization, otherwise they may be mathematically unstable or generate inaccurate predictions. If accurate prediction is the goal, rather than representing mechanisms, statistical models can be very useful, assuming sufficient training data are available. Rapid development of machine learning techniques has made it possible to extrapolate soil properties across time and space based on training data and algorithms such as neural networks and random forest. For example, this approach has been used to determine the global age of soil C based on radiocarbon profiles (Shi et al., 2020) and map soil C stocks across Scotland (Aitkenhead and Coull, 2016).

New approaches have started to combine features of process-based and probabilistic modeling. Rather than representing explicit pools of C, the PROMISE model of Waring et al. (2020) tracks the flow of individual C molecules through a heterogeneous soil system. Molecules undergo transformations and movements based on soil parameters, proximity to microbes and enzymes, and stochastic processes. In this way, molecules with different chemical properties vary in transit time such that the total soil C pool contains a distribution of residence times. This modeling framework requires relatively few assumptions and parameters while replicating emergent properties of soil C more accurately than pool-based models. It also incorporates mass balance and interactions among soil compounds, consistent with the six key principles identified by Sierra and Müller (2015).

2.4 Modeling across scales

2.4.1 Cellular/Molecular

Molecular interactions, both within and outside cells, underlie all soil biotic and abiotic processes. Key interactions include metabolic pathways within microbial cells along with sorption/desorption, enzymatic catalysis, and molecular diffusion outside of cells. Molecular-scale interactions between organic molecules and soil minerals contribute to the physical protection of soil organic matter (Schmidt et al., 2011), whereas extracellular enzyme activity catalyzes decomposition of polymeric molecules (Burns et al., 2013). Many of these interactions are represented in models at larger scales.

Metabolic pathways can be represented with flux balance models that simulate how specific substrates are metabolized in microbial cells. In $^{13}\mathrm{C}$ metabolic flux analysis ($^{13}\mathrm{C}$-MFA), isotopic labeling experiments provide models with information to estimate intracellular metabolic fluxes. Together with $^{13}\mathrm{C}$ fingerprinting to pinpoint central metabolic pathways and RNA-seq to complement the results of $^{13}\mathrm{C}$-MFA, Varman et al. (2016) uncovered the lignin degradation pathway of the bacterium *Sphingobium sp.* SYK-6. Environmental constraints and microbial community interactions must also be considered when modeling microbial metabolism. Jansson and Hofmockel (2018) defined the term metaphenome as the product of microbial functions that are expressed given abiotic and biotic environmental constraints. Flux balance models can be used to determine how microbial metaphenomes will respond to different environmental conditions and perturbations.

Information from molecular mechanisms can be used to quantify and better represent emergent properties in models. Carbon use efficiency (CUE) describes the proportion of C converted to microbial biomass and results from a combination of multiple metabolic processes. Hagerty et al.

(2018) suggested modeling CUE explicitly to account for its dependence on microbial growth and C allocation strategy, including costs from extracellular enzyme production and assimilation efficiency. By representing these additional cellular processes, the accuracy of larger-scale models with static CUE parameters could be improved.

Enzymes are biochemical catalysts involved in many molecular transformations that occur in soil (Burns et al., 2013). Microbes secrete extracellular enzymes outside the cell to obtain resources from complex biopolymers which are abundant in soils and litter. Given their role as bio-catalysts targeting soil organic matter, extracellular enzyme activity represents a mechanism of interaction between soil pools, namely microbial biomass and organic polymers. The Michaelis-Menten (MM) equation describes this activity, which often represents the initial and rate-limiting step of microbial decomposition. The MM equation predicts reaction velocity $\left(\frac{\mathrm{d}C}{\mathrm{d}t}\right)$ as a function substrate concentration (C) based on two parameters: the maximum velocity ($V_{\max}$) at unlimited substrate concentration and the half-saturation constant (K_{M}), which is the substrate concentration at $\frac{1}{2}$ $V_{\max}$:

$$\frac{\mathrm{d}C}{\mathrm{d}t} = \frac{V_{\max} \cdot C}{K_{\mathrm{M}} + C}$$

$V_{\max}$ and K_{M} can be experimentally determined and used to parameterize models. German et al. (2012) used experimental data on MM enzyme kinetics obtained from enzyme assays to build a decomposition model and determine the temperature sensitivity of extracellular enzymes. They found that both $V_{\max}$ and K_{M} are temperature-sensitive and the level of sensitivity is enzyme-specific.

Michaelis-Menten theory was extended in the Dual Arrhenius Michaelis-Menten (DAMM) model (Davidson et al., 2012). DAMM represents the interaction between Arrhenius and MM kinetics at the scale of enzyme active sites to predict CO_2 production from soil. The model accounts for temperature, moisture, and oxygen limitation effects on the metabolism of soluble C substrates.

Model predictions aligned well with laboratory measurements of extracellular enzyme activity at different temperatures and field measurements of soil respiration across seasons. DAMM was later extended to incorporate microsite variation in substrate concentrations and applied to predict not only soil respiration but also CH_4 and N_2O fluxes (Sihi et al., 2020).

The Reverse Michaelis Menten (RMM) and Equilibrium Chemistry Approximation (ECA) equations have also emerged as additional options to explicitly model enzyme kinetics (Moorhead and Weintraub, 2018; Tang, 2015). The RMM equation describes the reaction velocity as a function of enzyme concentration (E) where K_E is the enzyme concentration at $\frac{1}{2}\ V_{\max}$:

$$\frac{\mathrm{d}C}{\mathrm{d}t} = \frac{V_{\max} \cdot E}{K_E + E}$$

This equation is a better fit for situations where substrate available for enzyme binding is limiting. Such situations may be common in soils, and therefore RMM was included in one of the first microbial explicit models of soil C and N dynamics (Weintraub and Schimel, 2003).

The ECA considers both free substrate and enzyme limitations by accounting for mass balance constraints. MM and RMM kinetics are special cases of the ECA (Tang, 2015):

$$\frac{\mathrm{d}C}{\mathrm{d}t} = \frac{k \cdot E \cdot C}{K_{ES} + E + C}$$

where k is a rate constant, and $\frac{1}{K_{ES}}$ is the apparent substrate affinity of the enzyme. The ECA is more widely applicable than the MM and RMM due to its ability to consider a wider range of substrate-to-enzyme ratios. These ratios can shift in soil systems, and the ECA accounts for those changes by converging to either MM or RMM kinetics (B. Wang and Allison, 2019). However, the ECA is more complex and requires additional data for parameterization, so the simpler MM and RMM formulations may be better fits in some environmental contexts.

2.4.2 Population

As microbes consume substrates to obtain energy and nutrients, population size increases, resulting in changes in substrate demand and decomposition ability. Monod growth is an established model used to describe microbial growth given substrate availability (Parton et al., 2015). Analogous to MM kinetics, the specific growth rate (μ') is a function of substrate concentration (S), where $\mu_{\max}$ is the maximum potential growth rate and K_t is the Monod constant, or substrate concentration at $\frac{1}{2}\ \mu_{\max}$:

$$\mu' = \frac{\mu_{\max} \cdot S}{K_t + S}$$

Under the assumption that initial microbial biomass is much greater than initial substrate concentration, the Monod equation can be simplified to the MM equation. The Monod equation does not account for density dependence, so other models such as the logistic equation may be more appropriate if resources limit microbial population growth.

2.4.3 Community

Moving up in scale, multiple models represent interacting populations of microbes, and many of those also include physical features of the environment. Georgiou et al. (2017) found that introducing density-dependent growth of microbial biomass in decomposition models of varying complexity reduced divergence between model predictions and experimental observations. Density-dependent growth accounts for community-level mechanisms, such as competition and spatial limitations, though the exact parameterization may vary across biomes and should be experimentally determined.

Multiple community-scale models have adopted trait-based approaches that focus on the phys-

iological characteristics of microbes. Analogous to some vegetation models, the Guild Decomposition Model (GDM) represents three distinct microbial functional groups involved in litter decomposition: opportunists that process available organic matter, decomposers that break down holocellulose, and miners that degrade more chemically-resistant lignin polymers (Moorhead and Sinsabaugh, 2006). The GDM is a differential equation model with explicit degradation of substrate pools by the microbial functional groups following MM kinetics. The model also includes N which is often a limiting nutrient for fresh litter decomposition. Overall, the GDM successfully simulated decomposition and successional patterns consistent with observations.

The MIMICS model also represents microbial functional groups along with mineral stabilization, making it suitable for application to soil systems (Wieder, Grandy, Kallenbach, and Bonan, 2014). The functional groups in MIMICS distinguish r- versus K-selected life histories, where r-strategists specialize on the degradation of low molecular weight compounds and K-strategists process structural litter and chemically-protected compounds relatively more efficiently. Like the GDM, MIMICS assumes MM kinetics and reproduces observed patterns, including litter decomposition rates and soil response to disturbance.

Building on the idea of functional traits, other community-scale models represent interacting populations and even individuals. The DEMENT model (Allison, 2012) assigns traits at random to tens or hundreds of individual microbial taxa that compete and interact on a spatial grid (Figure 2.2). Rather than assigning taxa to functional groups a priori, taxa with favorable trait combinations for a given set of environmental conditions increase in abundance in the model simulations. The model is individual-based, meaning that it tracks the locations of individual cells or colonies that grow, divide, and disperse according to model assumptions and parameters. DEMENT's unique structure allows for simulation of "virtual microbiome" composition and functioning, including the cycling of C, N, and phosphorus. Once assigned, the traits of individual taxa in DEMENT are fixed, but related models have allowed for trait evolution within taxa (Allison, 2005; Folse and Allison, 2012).

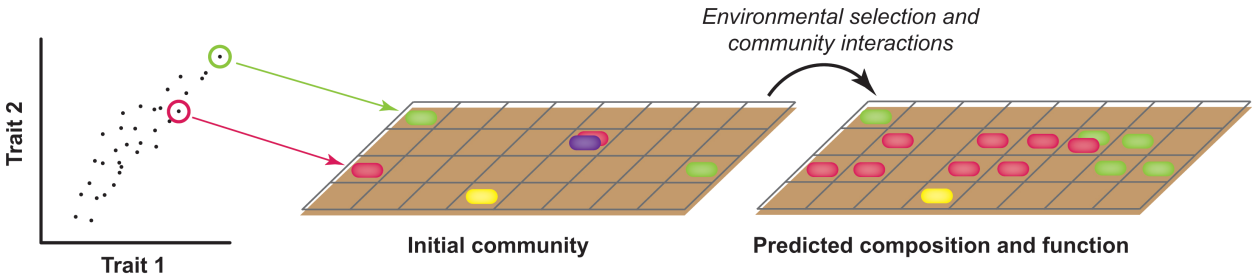

Figure 2.2: Schematic of the Decomposition Model of Enzymatic Traits (DEMENT). Traits are assigned to microbial taxa by drawing at random from empirically-based distributions. Taxa are placed randomly on a spatial grid where they consume substrates, reproduce, disperse, and interact over time. The model predicts community composition and function as taxon abundances change due to environmental selection. Adapted from Allison (2012).

Other models also represent microbial traits at the community scale. For example, an individual-based model with trait-based functional groups interacting on a spatial grid predicted tight cycling of N during litter decomposition, allowing the microbial community to maintain CUE by overcoming stoichiometric imbalances (Kaiser et al., 2014). These findings, along with applications of DEMENT (Allison, 2014), show that community-scale models are essential for predicting emergent, and sometimes unexpected, properties of community functioning. At the same time, challenges remain in translating genomic and physiological datasets into the trait distributions required to parameterize these models.

Spatially-explicit models like DEMENT are designed to represent enzyme kinetics and microbial interactions at appropriately small scales. Simulations with these models have provided insight into the emergent properties of heterogeneous enzyme-substrate interactions occurring at sub-micron scales, which could be useful for refining differential equation models operating at larger scales (B. Wang and Allison, 2019). Similarly, modeling the heterogeneous spatial structure of soil aggregates and associated microbial communities leads to more mechanistic prediction of trace gas fluxes (Ebrahimi and Or, 2016). Like individual-based models, aggregate-based models are useful for determining the scaling rules needed to incorporate heterogeneous soil properties and microbial communities into larger-scale models (B. Wang et al., 2019).

2.4.4 Ecosystem

Ecosystem-scale models of soil microbial and biological processes often include community-level processes as well as inputs and outputs that interact with other ecosystem components such as plants and minerals. Classical models such as RothC and Century have long been applied in an ecosystem context, and now microbial-explicit models are also being used at ecosystem scales. Efforts to integrate these approaches are gathering momentum as well. The Millennial model combines the best of both classical and microbial explicit models, including microbial processes, mineral stabilization, aggregate dynamics, and soil pools that can actually be measured (Abramoff et al., 2018).

Microbial-explicit models represent key microbial traits such as CUE, microbial turnover, and enzyme production that lead to different behaviors and predictions compared to microbial implicit models. The Allison-Wallenstein-Bradford (AWB) model was proposed as a relatively simple microbial explicit model of soil C cycling at the ecosystem scale (Allison et al., 2010). In contrast to the MIMICS model (as described in the Community section), the AWB model does not include functional groups. Instead, it represents average traits of the whole microbial community, such as CUE, enzyme kinetic parameters, and temperature sensitivities. Simulations with AWB showed that the soil C response to 5 °C warming depends on the temperature sensitivity of CUE. Greater temperature sensitivity of CUE results in more stable soil C pools in response to warming due to reductions in the biomass of microbial decomposers.

The Microbial-Enzyme-mediated Decomposition (MEND) model, developed by G. Wang et al. (2013), is similar in structure to AWB but also accounts for mineral stabilization mechanisms. MEND splits soil organic C (SOC) into particulate organic C (POC) and mineral-associated organic C (MOC), both of which are converted into DOC via enzyme activity. DOC can adsorb onto or desorb from MOC. The rate of breakdown into DOC is lower for MOC than POC, representing the physical protection of soil organic matter (Schmidt et al., 2011). Still, MOC and

POC respond similarly to a step increase in temperature, meaning that MEND and AWB end up making comparable predictions of SOC response to warming.

Sulman et al. (2014) proposed the Carbon, Organisms, Rhizosphere, and Protection in the Soil Environment (CORPSE) model, which also explicitly represents microbes but has a somewhat unique structure. Carbon in CORPSE can move between physically protected and unprotected pools, but unlike in MEND, only unprotected C pools can be decomposed. Another difference between CORPSE and MEND is that protected C pool sizes in CORPSE increase with clay content. These differences emphasize a need for additional empirical studies that quantify physical protection and the decomposition rates of protected SOC.

Soil models at the ecosystem scale differ substantially in their responses to plant C inputs. Microbial explicit models like AWB and CORPSE represent the priming effect, or increased turnover of SOC in response to the addition of fresh plant C, documented in many empirical studies (Bernal et al., 2016; Perveen et al., 2019). For example, Sulman et al. (2014) fitted CORPSE to empirical data from free-air CO2 enrichment experiments at Duke Forest and Oak Ridge National Laboratory (ORNL). They found that the priming effect almost completely offset increased litter input at Duke Forest. However, they found that modeled physical protection was stronger at ORNL while the modeled priming effect was much weaker, which corresponds with observations at ORNL showing increased protection of SOC in soil microaggregates.

Ecosystem model development remains a very active area of research. Although there are multiple microbial explicit models available now, many of them still lack key mechanisms such as spatial heterogeneity and cycling of N and other nutrients. When these mechanisms are incorporated, model outcomes may change substantially. For example, the SCAMPS model includes N dynamics and allows for variable C:N within the microbial community (Sistla et al., 2014). This stoichiometric flexibility allows the microbial community to acclimate to warming, resulting in greater losses of soil C through decomposition, especially in winter. The implication is that soil C dynamics likely depend on interactions with nutrients mediated by decomposers and plants.

2.4.5 Earth system

Most earth system models (ESMs) do not explicitly represent microbial communities. Of the 11 ESMs in the 6th Coupled Model Intercomparison Project (CMIP6), only one ESM explicitly represents microbes (Arora et al., 2020). That model—GFDL-ESM4.1 from NOAA's Geophysical Fluid Dynamics Laboratory—represents soil C cycling using CORPSE.

Although they are not fully coupled, there have been efforts to run microbial explicit models on a global grid, forced with output from ESMs. Wieder et al. (2013) created a microbial-explicit version of the Community Land Model (CLM) and compared its outputs with those from the Daily Century (DAYCENT) model and CLM4cn, a version of CLM with N cycling. Compared to microbial implicit CLM4cn and DAYCENT, microbial CLM predicted spatial patterns of steady state soil C that better aligned with global observations. Furthermore, a 20% increase in litter inputs only increased global soil C temporarily due to priming effects in microbial CLM (Figure 2.3). In contrast, soil C steadily increased in the microbial-implicit models CLM4cn and DAYCENT. Global soil C responses to warming were also variable and mediated by the temperature sensitivity of CUE as observed with the AWB model at the ecosystem level.

Hararuk et al. (2015) ran AWB and an ecosystem model by German et al. (2012)—which the study called the GER model—on a global grid. Both models simulated steady state empirical global soil C more accurately than the microbial implicit CLM-CASA model. After calibrating the models using a global soil C database, AWB and GER predicted faster declines in soil C compared to CLM-CASA under the RCP 8.5 climate forcing scenario. The analysis of Hararuk et al. (2015) also quantified the net outcome of decreasing CUE and the priming effect, allowing for key insights into how these opposing processes ultimately influence soil C predictions.

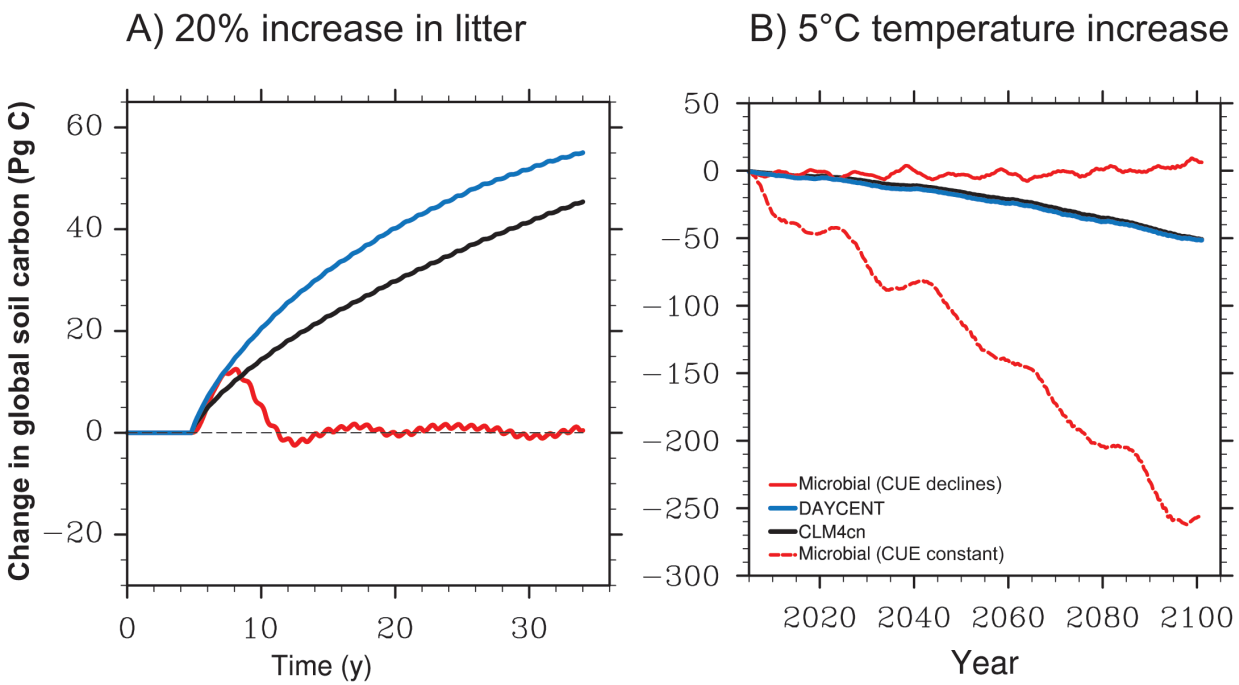

Figure 2.3: Soil carbon response of first-order and microbial explicit models to **A)** increased litter inputs and **B)** warming at the global scale. Warming response in the microbial model depends on whether carbon use efficiency (CUE) declines or remains constant with increasing temperature. Adapted from Wieder et al. (2013).

2.5 Model-data integration

2.5.1 Uncertainty quantification

As soil models continue to advance, they should be evaluated systematically for their effectiveness in achieving research goals (Figure 2.4). The process of reviewing and stress-testing models against observations is termed *model validation* (Marzouk and Willcox, 2015). Uncertainty quantification is a core part of model validation that involves assessment of model variation, biases, limitations, and constraints that lead to deviations between the model and the true, underlying data-generating processes. Uncertainty may arise from unknown values and meanings of system parameters and inputs, potentially because parameters do not correspond to measurable quantities. Related to parameter uncertainty, parametric variability concerns the unknown effects of

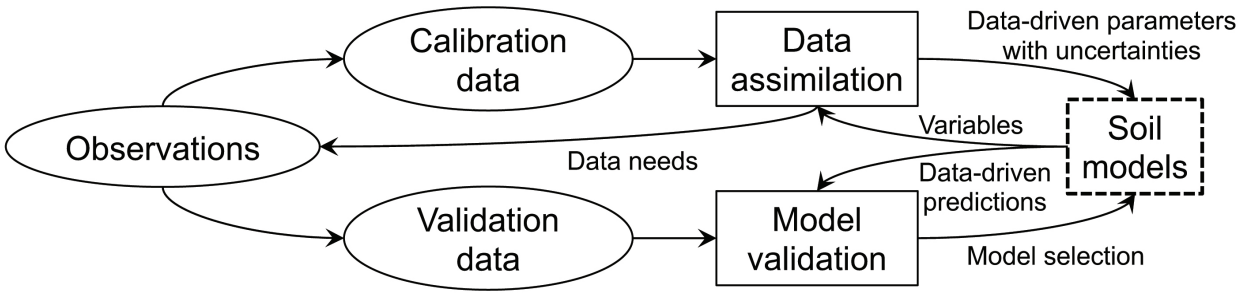

Figure 2.4: Framework for model-data integration. Observations are used for validating model outputs or calibrating model parameters via data assimilation. Bayesian approaches can be used for data assimilation and model validation to obtain posterior parameter distributions and calculate indices of model fit that aid in model selection. Adapted from Wieder, Allison, et al. (2015).

varying conditions on parameter and input values. Uncertainty also stems from model discrepancy, or the intentional and unintentional assumptions and simplifications separating a model from the actual processes it aims to represent.

Parameter uncertainty, parametric variability, and model discrepancy continue to be high for soil biogeochemical models (Shi et al., 2018). Some soil models have parameters that facilitate the functionality of the model, but do not have clear biological interpretations. For instance, the AWB model assumes Arrhenius temperature dependence for SOC transformations, but the associated activation energy parameters are not easy to measure directly (Allison et al., 2010; Xie et al., 2020). Modeling temperature dependence also introduces parametric variability and model discrepancy. Empirical studies confirm that parameters such as CUE and enzyme Vmax and Km are temperature sensitive (Sinsabaugh et al., 2013; Sinsabaugh et al., 2017), but the magnitude and functional form of temperature dependency is still an active area of investigation (Alster et al., 2020; Davidson et al., 2006).

Complex models may have many parameters that may covary, making it difficult to constrain parameter uncertainty (Sierra et al., 2015). Reducing this uncertainty requires that model parameters are identifiable, such that change in parameter value causes an associated change in variables predicted by the model. Sierra et al. (2015) proposed a collinearity index to quantify the identifiability of a model–the higher the index, the lower the identifiability, and the more difficult

it is to find the true parameter values. Increasing the number of datasets used to parameterize a model can increase identifiability of linear models and reduce overfitting, thereby improving predictive accuracy. For microbial explicit models, additional datasets including microbial variables (e.g. soil enzyme activities, microbial biomass) might be needed to increase parameter identifiability and reduce uncertainty.

Bayesian probabilistic frameworks are increasingly applied to interpret uncertainty in soil models. Central to Bayesian uncertainty quantification and model validation are the processes of Bayesian parameter estimation and inference, also known as data assimilation and probabilistic/Bayesian inversion in the geosciences (Lahoz and Schneider, 2014). With these approaches, the likely distribution of model parameter values for a given data set is estimated and characterized. The numerical approximation of parameter distributions and model likelihood estimation is carried out through Markov chain Monte Carlo (MCMC) simulation methods (O. F. Christensen et al., 2006). Although the exact Monte Carlo simulation algorithm may vary, most data assimilation frameworks include the following steps:

1. Choose model types and specific models to evaluate. In the case of soil biogeochemistry, the assimilation of linear and non-linear ordinary differential equation models can be compared, for instance (Xie et al., 2020).
2. Choose a dataset for comparison with model outputs.
3. Establish pre-inference probability density functions of model parameter values (known as the *prior distributions* or *priors*).
4. Iteratively propose model parameter values to generate model outputs for computing model likelihood for the given data set.
5. Approximate the distributions and probability density functions of parameter values that correspond to better model fits to the data set (known as the *posterior distributions* or *posteriors*).

6. Compare model likelihoods conditional on the data set with available and desired goodness-of-fit metrics. The specific Monte Carlo algorithm will dictate the options available for goodness-of-fit metrics.

Asymptotically "exact" Bayesian Monte Carlo schemes aim to comprehensively sample parameter values to compute posterior distributions. These methods include traditional Gaussian random walk Metropolis-Hastings MCMC and Gibbs samplers (McElreath, 2020), adaptive approaches derived from evolutionary optimization algorithms such as differential adaptive evolution Metropolis (Vrugt, 2016), and the physics-inspired, momentum-driven family of Hamiltonian Monte Carlo algorithms (Neal, 2011).

Statisticians have also been investigating "non-exact" Bayesian inference schemes that seek to increase speed through approximation and simplification of parameter spaces. Non-exact approaches include the approximate Bayesian computation (Alahmadi et al., 2020; Csilléry et al., 2010) and variational Bayesian classes of methodologies (Blei et al., 2017; Ryder et al., 2018). Goodness-of-fit methods range from simpler frequentist computations such as coefficient of determination and maximum likelihood estimation to Bayesian metrics including information criteria and cross-validation computations (Gelman et al., 2013). Fully Bayesian goodness-of-fit metrics can be more stable than their frequentist counterparts and provide more diagnostic information about overfitting and inference validity (Vehtari et al., 2017), though there may be higher computational resource demands.

There have been several powerful applications of Bayesian parameter estimation to soil biogeochemical models. Hararuk et al. (2014) integrated global soil C data into the C-only version of the Community Land Model coupled with the Carnegie-Ames-Stanford Approach submodel (CLM-CASA), while Ťupek et al. (2019) integrated respiration data from boreal forests in Finland into the Yasso07, Yasso15, and Century models. Both studies compared model outputs before and after using a Bayesian data assimilation process to constrain model parameters. In all cases, data

integration resulted in model predictions that more closely matched observations.

However, each of these studies has caveats. The soil C database used by Hararuk et al. (2014) did not include time-series data, thereby necessitating a steady-state assumption about C pool sizes. If this assumption is not accurate, estimates of model uncertainty may be difficult to interpret. Ťupek et al. (2019) calibrated models with observed data but did not use an independent dataset to validate model predictions, which can lead to model overfitting. Maintaining separate training and validation datasets, a common practice in machine learning approaches, can help avoid this problem (Botu et al., 2017).

Approaches like Bayesian data assimilation are most effective when extensive, multivariate datasets are available for model calibration and validation across a range of ecosystems. For example, field measurements of dryland soils have improved biogeochemical models of ecosystem-specific C-cycling dynamics (Shen et al., 2016; X. Zhang et al., 2014). Going forward, rapid advancements in remote and in situ environmental sensing tools like light detection and ranging (LiDAR) (Kemppinen et al., 2018) and soil nutrient sensors (Burton et al., 2020) can increase the amount and availability of ecosystem-specific measurements at lower cost, higher resolution, and greater sampling intensity than ever before.

2.5.2 Model intercomparison

Model intercomparison goes hand-in-hand with model selection and data assimilation to evaluate the behaviors and performance of different models relative to one another. For instance, J. Li et al. (2014) compared 3 microbial-explicit models with a conventional first-order model and found that steady-state SOC was much more responsive to varying temperature sensitivity of CUE in the microbial-explicit models. In contrast, SOC stocks were largely independent of microbial CUE in the first-order model. This analysis points toward a need for additional empirical research on how microbial CUE varies with temperature and other factors.

The application of Bayesian approaches to model calibration and selection can readily be extended to model intercomparison. In their global analysis of soil C responses under RCP8.5, Hararuk et al. (2015) used a Bayesian approach to show that the microbial-explicit models AWB and GER better explained the spatial variation of steady-state soil C compared to the CLM-CASA model. However, at least with some parameter values, the microbial-explicit models simulated oscillations in soil C over time, which is an unrealistic behavior at ecosystem to global scales.

Xie et al. (2020) also applied a Bayesian approach to compare AWB with a conventional model (Fig. 16.1). Both models were fit to a meta-analysis dataset on soil respiration response to warming Romero-Olivares et al., 2017 and compared using Bayesian goodness-of-fit metrics such as the widely applicable information criterion (WAIC) and leave-one-out-cross validation (LOO). These metrics account for the posterior distributions of parameter values after model fitting, and LOO is a useful metric when limited data are available for model selection. Both models fit the meta-analysis data reasonably well, but the simpler structure of the conventional model led to slightly better WAIC and LOO scores. These findings emphasize that model selection involves tradeoffs. Simple models with few parameters may be calibrated to match observational datasets with good validation scores, but these models may fall short in capturing the mechanistic details needed to make accurate predictions across a broader range of soil ecosystems.

2.6 Recommendations to advance soil models

Despite recent progress, substantial barriers still prevent the widespread application of models to grand challenges in soil biology. In particular, specialized language, expertise, and skill sets can make it challenging to integrate modeling with other scientific approaches. This specialization can be a barrier to information flow between model and empirical analyses. Such barriers can exacerbate the challenge of collecting data in a form that supports model development, calibration, and validation. In addition, models can be difficult to access and apply if recent versions,

adequate documentation, and user interfaces are not available. Scaling up models, for example to make Earth system predictions, can be limited by insufficient tools for model selection and inter-comparison. Approaches for model validation are still under development and involve specialized knowledge of computational and statistical tools.

Overcoming these barriers would be beneficial. Predictive accuracy would increase for models applied to simulate future soil C stocks, nutrient cycling, and climate change. Given that models have multiple uses beyond prediction, efforts to engage a broader community would also advance fundamental knowledge across the disciplines of soil science, biology, and biochemistry. To reap these benefits, we recommend the following actions:

- Integrate modeling and empirical approaches. Rather than viewing modeling and empirical activities as separate, we recommend co-developing models and empirical research. Operationally, this means reconfiguring scientific teams so that researchers with modeling expertise interact directly and frequently with empirical researchers. From the proposal writing stage through model development and manuscript publication, scientists creating models and collecting data should create spaces to develop a common language and align research goals. By co-creating models and experiments, researchers can ensure that models represent key processes, critical model parameters are measurable, and both model and experimental outcomes are relevant to one another. Such cooperation would be particularly helpful for incorporating complex 'omics datasets into trait-based community models.

- Collect more data. Relatively few time-series datasets are available for some soil variables, such as C stocks, making it difficult to evaluate or avoid the steady-state assumptions often made in biogeochemical models. Sparse data can also limit the possibility of separating data into training versus validation subsets. Better integration between modeling and empirical research could help fill some of these data gaps.

- Cross-train researchers in modeling. To enable the interactions necessary for integration,

researchers should receive training in modeling perspectives and approaches. For example, training activities such as workshops, short courses, and online modules can help students acquire common vocabulary used in modeling. Conversely, students with a modeling background can benefit from training activities focused on theory and empirical work. If designed thoughtfully, seminars and courses can provide opportunities for students to get comfortable communicating and collaborating across the modeling-empirical divide.

- Improve accessibility to model code and analysis tools. The principles of F.A.I.R. data should also apply to model code and outputs: findable, accessible, interoperable, and reusable (Wilkinson et al., 2016). Code repositories such as GitHub and platforms such as the Department of Energy's KBase can host code along with input/output files and user interfaces to make models findable and accessible. For new models, writing and documenting code in widely-used, open-source formats such as R Markdown and Jupyter Notebooks for Python can promote interoperability and reusability. A version control system is also important to ensure analyses from a prior model version can be replicated. Regardless of the model or platform, researchers should always strive to make model code and analyses publicly available with guidelines for reuse so that others may validate, build upon, and broaden applications of published models.

- Plug-and-play models and datasets. Taking the principle of interoperability to another level, we encourage the development of model testbeds that enable mixing and matching of different models and datasets (Wieder et al., 2018). Ideally, such testbeds should allow for modifications of model structure and input datasets. Testbeds can also facilitate standardization of input/output protocols and datasets to enhance interoperability, thereby avoiding tedious data reformatting procedures while also providing guidance on standards that could be adopted by the broader soil science community.

- Develop improved model selection and intercomparison tools. Moving beyond testbeds, the research community would benefit from wider availability of model selection and inter-

comparison resources. For example, the `soilR` package enables users to run simulations with an array of differential equation models, including some that represent soil radiocarbon (Sierra et al., 2012). Global intercomparison initiatives such as the Coupled Model Intercomparison Project (CMIP) have also been tremendously valuable for comparing Earth system models by establishing a standardized set of simulation scenarios and output variables (Arora et al., 2020; Todd-Brown et al., 2013). As new tools for model inference become available, they could be incorporated into intercomparison projects to enable one-stop-shopping for model comparison and selection (Xie et al., 2020).

2.7 Conclusion

Within the last 5-10 years, models of soil systems have advanced substantially. There are now many new approaches for representing microbial and biochemical processes in soil models. As these new models came online, synthesis efforts placed them in the context of broad principles that guide quantitative soil science across scales and ecosystems. We anticipate that these advances will support further integration and unification of soil biological modeling in the next 5-10 years. Still, another modeling renaissance faces some significant challenges. Disciplinary silos as well as difficulties in scaling models from genes to ecosystems must be overcome to maximize the impact of recent model advances. Breaking down these barriers will require better integration of modeling approaches into all branches of soil science. Our recommendations to build computational infrastructure and train a new generation of researchers well-versed in modeling can serve as an initial roadmap for integration. Following our roadmap should help elevate models as valuable tools for tackling soil-related grand challenges facing society, from food security to climate change.

CHAPTER 3

A Bayesian approach to evaluation of soil biogeochemical models using information criteria and cross-validation

3.1 Introduction

Note: We abbreviate soil organic carbon, dissolved organic carbon, microbial biomass carbon, and extracellular enzyme carbon respectively with SOC, DOC, MIC, and ENZ in this chapter, rather than the SOC, DOC, MBC, and EEC acronyms we use in other chapters. This was done to match the manner in which this chapter was originally published as Xie et al. (2020) in *Biogeosciences*.

Coupled Earth system models (ESMs) and constituent soil biogeochemical models (SBMs) are used to simulate global soil organic carbon (SOC) dynamics and storage. As global climate changes, some ESM and SBM simulations suggest that substantial SOC losses could occur, resulting in greater soil CO2 emissions (Crowther et al., 2016). However, there is vast divergence between model predictions. For instance, one ESM predicts a global SOC loss of 72 Pg C over the 21st century, while another predicts a gain of 253 Pg C (Todd-Brown et al., 2014).

Soil biogeochemical models vary greatly in structure (Manzoni and Porporato, 2009) but can be broadly partitioned into two categories: those that implicitly represent soil C dynamics as first-order linear decay processes and those that explicitly represent microbial control over C dynamics with nonlinear Michaelis-Menten functions (Wieder, Allison, et al., 2015). Explicit models typically include more parameters than linear models because multiple microbial parameters are needed for each decay process as opposed to a single rate parameter. The additional parameters allow explicit models to represent microbial mechanisms, but at the expense of greater model complexity.

Rigorous statistical approaches should be applied to investigate how explicit representation of microbial processes affects predictive model performance. ESM and SBM comparisons involving empirical soil C data assimilations have been conducted previously (Allison et al., 2010; J. Li et al., 2014) but few standardized statistical methods for ESM and SBM benchmarking and comparison have been developed that would allow for rigorous model selection. Prior model comparisons have involved graphical qualitative comparisons or use of basic fit metrics, such as the coefficient of determination, R^2, to judge fit quality. However, these simple approaches are insufficient for comparing an increasing number of complex models (Jiang et al., 2015; Luo et al., 2016; Wieder, Allison, et al., 2015).

R^2 on its own provides limited information about goodness-of-fit. In unmodified form, it quantifies the extent to which the variation of just one chosen model outcome — for instance the mean outcome for a range of parameter values — corresponds to the variation in the data set (Gelman et al., 2019). R^2 does not capture model complexity, overfitting, or parameter uncertainty, which is a reason why R^2 by itself is not sufficient for model evaluation (Kvålseth, 1985). Without accounting for model complexity and parameter count, focusing on optimizing fit by R^2 values alone can easily lead to overfitting (Spiess and Neumeyer, 2010).

Encouragingly, a rich toolset to further inform quantitative model evaluation and comparison can be drawn from Bayesian statistics (Hararuk and Luo, 2014; Hararuk et al., 2014; Hararuk et al.,

2018). These tools include information criteria and approximate cross-validation, goodness-of-fit metrics designed for the simultaneous comparison of multiple structurally diverse models. Like R^2, information criteria and cross-validation are quantitative measures that estimate the fit quality of a model to a given data set. Differing from R^2, information criteria and cross-validation are relative rather than absolute measures. These metrics evaluate the extent to which the data set supports particular distributions of parameter values and, in turn, the uncertainty of parameter estimates. Consequently, if the distribution of model A outcomes aligns more closely to the data set than the distribution of model B outcomes, we regard model A as being more likely to explain the data compared to model B. Information criteria and cross-validation metrics also typically include terms penalizing for model complexity and overfitting as part of their computation (Gelman et al., 2014). Hence, information criteria and approximate cross-validation are useful tools for model evaluation because they present a comprehensive summary of model fit to time series data and can estimate model predictive accuracy for unmeasured and out-of-sample data points.

Examples of information criteria popularized by widely used R packages such as `lme4` and `rjags` include the Akaike information criterion (AIC), Bayesian information criterion (BIC), and deviance information criterion (DIC) (Vehtari and Ojanen, 2012). However, these metrics have some limitations. AIC, BIC, and DIC do not use full sampled posterior distributions in their computational processes. AIC and BIC both rely on a pointwise maximum likelihood estimate that cannot be derived from nonuniform Bayesian prior distributions, including normal distributions. AIC and BIC (despite BIC's name) thereby have limited use in Bayesian statistics settings. DIC can accommodate nonuniform priors but is calculated from pointwise simplified posterior means. The compression of full posteriors into pointwise means can prompt DIC to compute an impossible negative effective model parameter count in select situations (Gelman et al., 2014). Consequently, the original forms of AIC, BIC, and DIC are no longer recommended for use in Bayesian model assessment by some statisticians in light of superseding alternatives (Gelman et al., 2014).

Three predictive goodness-of-fit metrics address the limitations and stability issues of AIC, BIC, and DIC by incorporating full, nonuniform posterior distributions in their calculations to better account for overfitting and model size (R. Christensen et al., 2010; Gelman et al., 2014). These metrics include the widely applicable information criterion (WAIC), log pseudomarginal likelihood (LPML), and Pareto-smoothed important sampling leave-one-out cross-validation (PSIS-LOO and hereby referred to as LOO). WAIC, LPML, and LOO can estimate the ability of models to fit unobserved measurements outside of the set of measured data samples (Vehtari et al., 2017). Thus, WAIC, LPML, and LOO can be considered as superior barometers for model predictive accuracy compared to AIC, BIC, and DIC.

The overarching goal of this study was to develop a statistically rigorous and mathematically consistent data assimilation framework for SBM comparison that uses predictive Bayesian goodness-of-fit metrics. We pursued three specific objectives as part of that goal. First, we compared the behaviors of two different SBMs, a linear microbial-implicit model termed the conventional model (CON) and a nonlinear microbial-explicit model called the Allison-Wallenstein-Bradford model (AWB) (Figure 3.1), following data assimilation with soil respiration data sourced from a meta-analysis of soil warming studies (Romero-Olivares et al., 2017). Second, we characterized the parameter spaces of these models using prior probability distributions of parameter values informed by previous studies and expert judgment. Third, we compared specific Bayesian predictive information criteria in WAIC, LPML, and LOO to the coefficient of determination, R^2, for quantifying goodness-of-fit to data. AIC, BIC, and DIC were not analyzed due to their stability limitations, our usage of nonuniform prior distributions, and redundancy with WAIC.

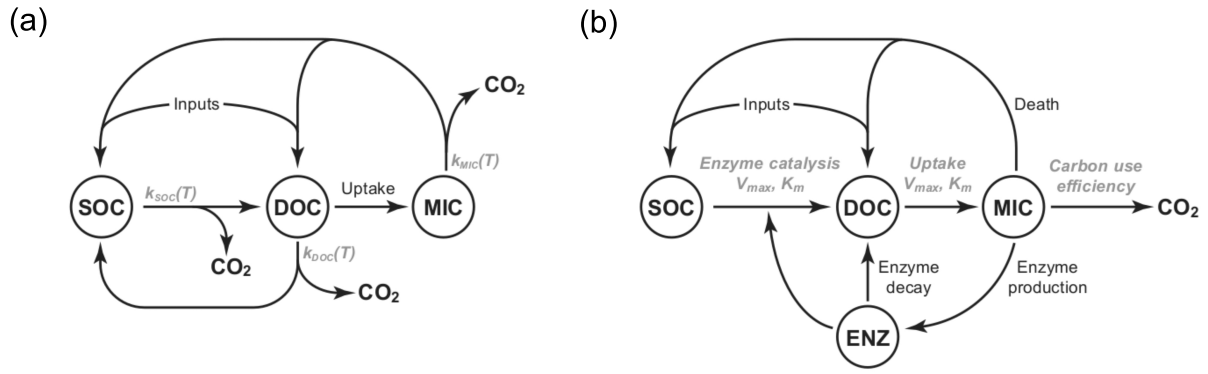

Figure 3.1: Diagrams of the pool structures of the (a) CON model and (b) AWB model drawn from Allison et al. (2010). Pools are shown within circles including soil organic carbon (SOC), dissolved organic carbon (DOC), and microbial (MIC) pools. AWB has SOC, DOC, and MIC pools as in CON but also an extra enzymatic (ENZ) pool. AWB additionally differs from CON in its nonlinear feedbacks and assumption that MIC can influence SOC-to-DOC turnover through the ENZ pool.

3.2 Methods

3.2.1 Model structures

We compared two SBMs, the CON and AWB models (Allison et al., 2010). The models were selected for this study due to their relative equation simplicity, their tractable parameter count, and limited biological data input requirements. The CON system models three separate C pools as state variables including SOC, dissolved organic C (DOC), and microbial biomass C (MIC) pools, while AWB includes SOC, DOC, MIC, and extracellular enzyme biomass C (ENZ) pools (Figure 3.1). Additionally, these models were chosen because they are C-only models without nitrogen (N) pools. The increased complexity of N-accounting SBMs will require future studies with coupled N data sets (Manzoni and Porporato, 2009).

3.2.2 Meta-analysis data

The data set for model fitting was compiled from a recent meta-analysis of 27 soil warming studies that measured CO2 fluxes (Romero-Olivares et al., 2017). The experiments reported between 1 and 13 years of CO2 flux measurements following warming perturbation. The elements of this data set consisted of empirical response ratios calculated by dividing CO_2 fluxes measured in the warming treatments by time-paired CO_2 fluxes measured in the control treatments. We calculated an annual mean response ratio for each experiment (if data were available for that year) after warming treatment began. Using these annual means, we calculated one overall mean response ratio for each year along with pooled variances and standard deviations. Pooled data points were assumed to be "collected" at the halfway point of each year. Because the experiments had variable lengths, the sample size for the pooled annual mean declines with increasing time since warming perturbation. The warming perturbation was 3 °C on average across all the studies, and this average was used as the magnitude of warming in the model simulations.

Model-outputted response ratios were calculated by dividing simulated CO_2 flux following warming perturbation by the CO_2 flux at prewarming steady state. We fit models to flux response ratios rather than raw flux measurements for several reasons (Wieder, Grandy, Kallenbach, Taylor, et al., 2015). First, we eliminate the need to convert flux measurements from different experiments into a common unit. Second, response ratios represent a standardized metric for warming response across disparate ecosystem types with varying climate, soil, and vegetation properties. Finally, fitting a mean response ratio overcomes data gaps present in individual experiments.

3.2.3 Hamiltonian Monte Carlo fitting of differential equation models

CON and AWB ordinary differential equation systems were simulated using the CVODE backward differentiation method (Curtiss and Hirschfelder, 1952) from the SUNDIALS library of equation solvers (Hindmarsh et al., 2005). Differential equation models contain parameters that affect state variables, and model fitting through Markov chain algorithms involves iterating through parameter space one set of parameters at a time. We performed model fitting using a Markov chain algorithm called the Hamiltonian Monte Carlo (HMC), using version 2.18.1 of the RStan interface to the Stan statistical software (Carpenter et al., 2017; Stan Development Team, 2020) and version 3.4.1 of R (R Core Team, 2017). HMC is not a random-walk algorithm and uses Hamiltonian mechanics to determine exploration steps in parameter space. HMC has been theorized to offer more efficient exploration of high-dimensional parameter space than traditional random-walk Metropolis algorithms (Beskos et al., 2013).

Conditional on the meta-analysis data set, the HMC algorithm computed posterior and posterior predictive distributions, from which Bayesian statistical inferences on likely ranges of parameter values were then made. Posterior distributions are the distributions of more likely model parameter values conditional on the data. Posterior predictive distributions are the distributions of more likely values for unobserved data points from the data-generating process conditional on the observations. In the case of this study, the experiments constituting the meta-analysis would be the data-generating process.

For the sake of clarity, it is important to distinguish between the frequentist confidence intervals and Bayesian posterior predictive intervals and distributions we describe in our study. Confidence intervals are calculated from the sample means and standard errors at observed data points and indicate ranges of values that are likely to contain the true data values with repeated sample collections using the same methodology. Posterior predictive intervals and distributions are computed

after estimation of the posterior parameter distributions and represent the likely distributions of unobserved data values conditional on observed data values. Bayesian credible intervals, which we will also discuss in this study, are ranges of values that parameters are likely to take with some probability that are conditional on the observed data. Credible areas indicate the probability densities of parameter values across credible intervals.

We ran four chains for 35000 iterations each for our HMC simulations, with the first 10000 iterations being discarded as burn-in for each chain. Hence, our posterior distributions consisted of 100000 posterior samples per HMC run. In retrospect, because our credible areas displayed sufficient smoothness and Bayesian diagnostics indicated adequate posterior sampling, we could have reduced simulation time without impairing posterior computation by running shorter chains that consisted of 20000 to 30000 iterations. To minimize the presence of divergent energy transitions, which indicate issues with exploring the geometry of the parameter space specified by the prior distributions, we set the adaptation delta to 0.95, the initial step size to 0.1, and maximum tree depth to 12. Those parameters determine how the HMC algorithm proposes new sets of parameters at each step and were set so that the HMC would begin with smaller exploration steps. The algorithm varies the step size from its initial value throughout posterior sampling to maintain a desired acceptance rate; the tuning sensitivity of the step size is governed by the adaptation delta value, with higher values indicating reduced sensitivity.

We further constrained our HMC runs to characterize parameter regimes corresponding to higher biological realism. Normal informative priors were used to initiate the runs, and the prior distribution parameters were chosen based on expert opinion and previous empirical observations (Allison et al., 2010; J. Li et al., 2014). Prior distributions had noninfinite supports; supports were truncated to prevent the HMC from exploring parameter space that was unrealistic.

3.2.4 Model steady-state initialization

Because we were mainly interested in testing model predictions of soil warming response, the models were initiated at steady-state prior to the introduction of warming perturbation to isolate model warming responses from steady-state attraction. We fixed pre-perturbation steady-state soil C densities to prevent HMC runs from exploring parameter regimes corresponding to biologically unrealistic C pool densities and mass ratios.

To set prewarming steady-state soil C densities, we first analytically derived steady-state solutions of the ordinary differential equations of the models. Then, with the assistance of Mathematica version 12, we rearranged the equations by moving the steady-state pool sizes to the left-hand side, such that we could determine the value of parameters dependent on pool sizes while allowing the rest of the parameters to vary for the HMC. Consequently, we could constrain the prewarming pool sizes from reaching unrealistic values in the simulations.

3.2.5 Sensitivity analysis of C pool ratios

Sensitivity analyses examine how the distributions of model input values influence the distributions of model outputs. In our study, we considered prewarming C pool densities as a model input. We performed a sensitivity analysis to observe how the choice of prewarming C pool densities and C pool ratios would affect the model fits and posterior predictive distribution of C pool ratios.

We compared the model outputs and postwarming response behavior of AWB and CON at equivalent C pool densities and ratios. The ratio of soil microbe biomass C (MIC) density to SOC density has been observed to vary approximately from 0.01 to 0.04 (Anderson and Domsch, 1989; Sparling, 1992), so we used those numbers as guidelines for establishing the ranges of the C pool densities and density ratios explored in our simulations. One portion of the analysis involved running HMC simulations in which we set the prewarming MIC density at 2 mg C g^{-1} soil and then

varied the SOC density from 50 to 200 mg C g^{-1} soil in increments of 25, stepping from 0.04 to 0.01 with respect to the MIC-to-SOC ratio. A second portion of the analysis involved observing the effect of varying prewarming MIC from 1 to 8 mg C g^{-1} soil while holding prewarming SOC at 100 mg C g^{-1} soil.

For some combinations of the prior distributions and prewarming steady-state C pool densities, AWB HMC runs wandered into unstable parameter regimes that would prevent the algorithm from reliably running to completion. Consequently, we do not compare simulation results for AWB and CON with prewarming SOC densities below 50 mg C g^{-1} soil. Other combinations of prior distribution and prewarming C pool density choices that were not necessarily biologically realistic allowed stable AWB runs with lower prewarming SOC densities.

3.2.6 Information criteria and cross-validation

In addition to R^2, we used the WAIC, LPML, and LOO Bayesian predictive goodness-of-fit metrics to evaluate models with the meta-analysis warming response data. LPML is an example of cross validation that is calculated similarly to LOO (Gelfand and Dey, 1994; Gelfand et al., 1992; Ibrahim et al., 2001) but differs from LOO in how the importance ratio sampling portion of its computation is handled. For further explanation regarding importance ratios and their role in evaluating approximate cross-validation metrics, refer to the description of the LOO algorithm presented in Vehtari and Ojanen (2012). LOO updates LPML by implementing a smoothing process in which the largest importance ratios are fitted with a Pareto distribution and then replaced by expected values from the distribution, which stabilizes the importance ratio sampling.

We used version 2.0.0 of the `loo` package available for R to calculate our WAIC and LOO values (Vehtari et al., 2019). We specifically employed the LOOIC variant of LOO implemented in `loo` so that LOO could operate on a comparable numerical scale to WAIC. A lower WAIC and LOO and a higher LPML indicate a more likely model for a given data set. LPML can be multipled by

a factor of -2 to occupy a similar scale to LOO.

3.3 Results

3.3.1 Parameter posterior distributions

We obtained distributions of posterior predictive fits to the univariate response ratio data for both AWB and CON across different prewarming MIC-to-SOC ratios. Posterior samples totaled 100000 for each simulation. Sampler diagnostics for the HMC runs indicated that the statistical models were valid at all prewarming steady-state values observed, that model parameter values converged across the four Markov chains, and that the posterior parameter space was effectively sampled and explored to generate enough independent posterior samples for inference. The ratios of effective posterior parameter samples to total samples for parameters were generally satisfactory; across observed MIC-to-SOC ratios, they were all greater than 0.25 and mostly greater than 0.5.

We also tracked divergent transitions, which mark points in chains at which the HMC algorithm was inhibited in its exploration and posterior sampling, potentially due to the parameter space becoming geometrically confined and difficult to navigate. Divergent transitions occurred in the AWB HMC runs, though the ratios of divergent transitions to sampled iterations were relatively low for all runs. The highest divergent transition ratio observed was 0.0217, corresponding to the simulation initiated with prewarming SOC = 200 $\mathrm{mg\,C\,g^{-1}}$ soil. There were no divergent transitions in the CON runs.

3.3.2 Model behaviors

The CON curve monotonically decreases in response ratio over time, whereas the AWB curve displays changes in slope sign (Figure 3.2). The difference in curve shape (Figure 3.3a, b) is in line with CON's linear status and AWB's nonlinear formulation with more parameters (Allison et al., 2010). By 50 years after warming, mean fit curves for AWB and CON return to 1.0 after their initial increase (Figure 3.3c, d), consistent with prior observations and expectations at steady state (Romero-Olivares et al., 2017; van Gestel et al., 2018).

From a cursory visual evaluation, neither of the models clearly outperforms the other across all prewarming steady states. The 95% confidence interval of the first data point at t = 0.5 year (hereby abbreviated as year) does not include the AWB SOC100 posterior predictive mean as it does for the CON SOC100 mean (Figure 3.2), which most likely impaired AWB's quantitative goodness-of-fit metrics. However, the 95% response ratio posterior predictive interval suggests that AWB is able to replicate the response ratio increase in the data from 1.5 to 3.5 years following the warming perturbation, which CON does not. The shape of the AWB posterior predictive interval also fits the data points and confidence intervals occurring 8 years or more after the perturbation more closely than that of CON (Figure 3.3a, b).

For both AWB and CON, increasing the prewarming SOC to higher densities from SOC = 50 to 200 mg C g^{-1} soil (hereby labeled from SOC50 to SOC200) while holding prewarming MIC at 2 mg C g^{-1} soil, DOC at 0.2 mg C g^{-1} soil, and ENZ at 0.1 mg C g^{-1} soil corresponded to lower initial mean response ratios in the first year at the t = 0.5 year time point, which certainly inhibited the quantitative goodness-of-fit (Figure 3.3a, b). For CON, increasing prewarming SOC also reduced the magnitude of the mean fit slope. For AWB, increasing prewarming SOC had no clear effect on the curve slope, but the model needed more time to achieve peak mean response ratio from a lower start, with the peak being reached at t = 1.5 year in the SOC50 case and t = 3.5 year in the SOC 200 case (Figure 3.3b). At the higher prewarming SOC, CON's reduced

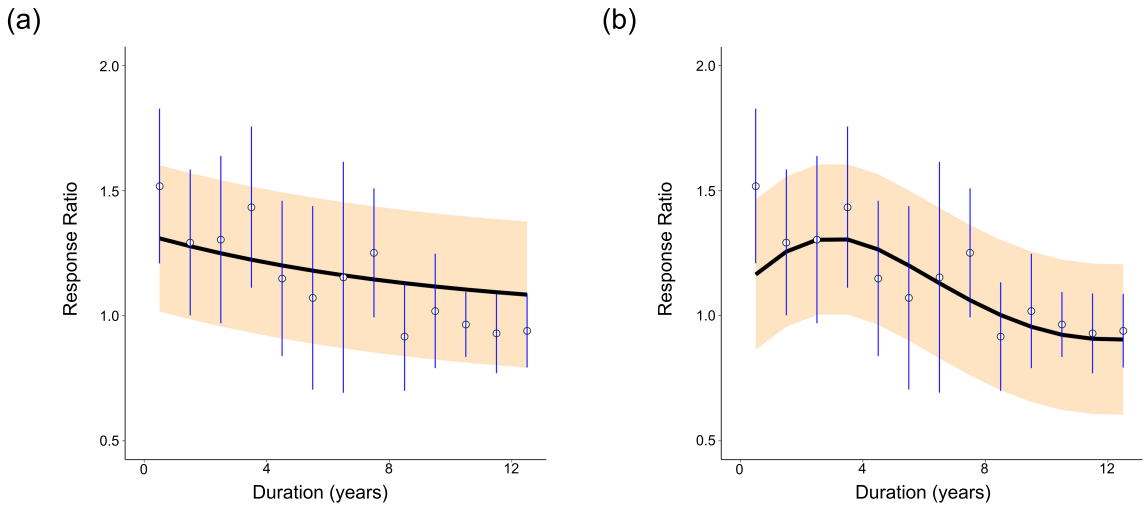

Figure 3.2: Distribution of fits of **(a)** CON and **(b)** AWB to the meta-analysis data from Romero-Olivares et al. (2017). Open circles show the meta-analysis data points. Blue vertical lines mark the 95% confidence interval for each data point calculated from the pooled standard deviation. The black line indicates the mean model response ratio fit. The orange shading marks the 95% posterior predictive interval for the fit. For **(a)**, prewarming steady-state soil C densities were set at SOC = 100 mg C g^{-1} soil, MIC = mg C g^{-1} soil, and DOC = 0.2 mg C g^{-1} soil. For **(b)**, prewarming steady-state soil C densities were set at SOC = 100 mg C g^{-1} soil, MIC = mg C g^{-1} soil, DOC = 0.2 mg C g^{-1} soil, and ENZ = mg C g^{-1} soil.

slope magnitude and AWB's lagging response ratio peak caused both models to exhibit slower returns to the steady-state response ratio of 1.0 (Figure 3.3c, d). On their trajectories back to steady state, the mean SOC 200 CON curve substantially overshoots the data mean after $t = 7.5$ year (Figure 3.3a), whereas the SOC200 AWB curve exceeds the data means at a more moderate extent through the $t = 8.5$, 9.5, 10.5, and 11.5 year time points (Figure 3.3b).

Changing the prewarming MIC-to-SOC steady-state pool size ratio by increasing prewarming MIC from 1 to 8 mg C g^{-1} soil (hereby labeled from MIC1 to MIC8) while holding prewarming SOC at 100 mg C g^{-1} soil had marginal to moderate qualitative effects on the mean response ratio curves for CON and AWB. The CON MIC1 and MIC8 curves were visually indistinguishable, while the AWB MIC1 and MIC8 curves differed with the MIC8 curve displaying more gradual changes in slope and lower slope magnitudes.

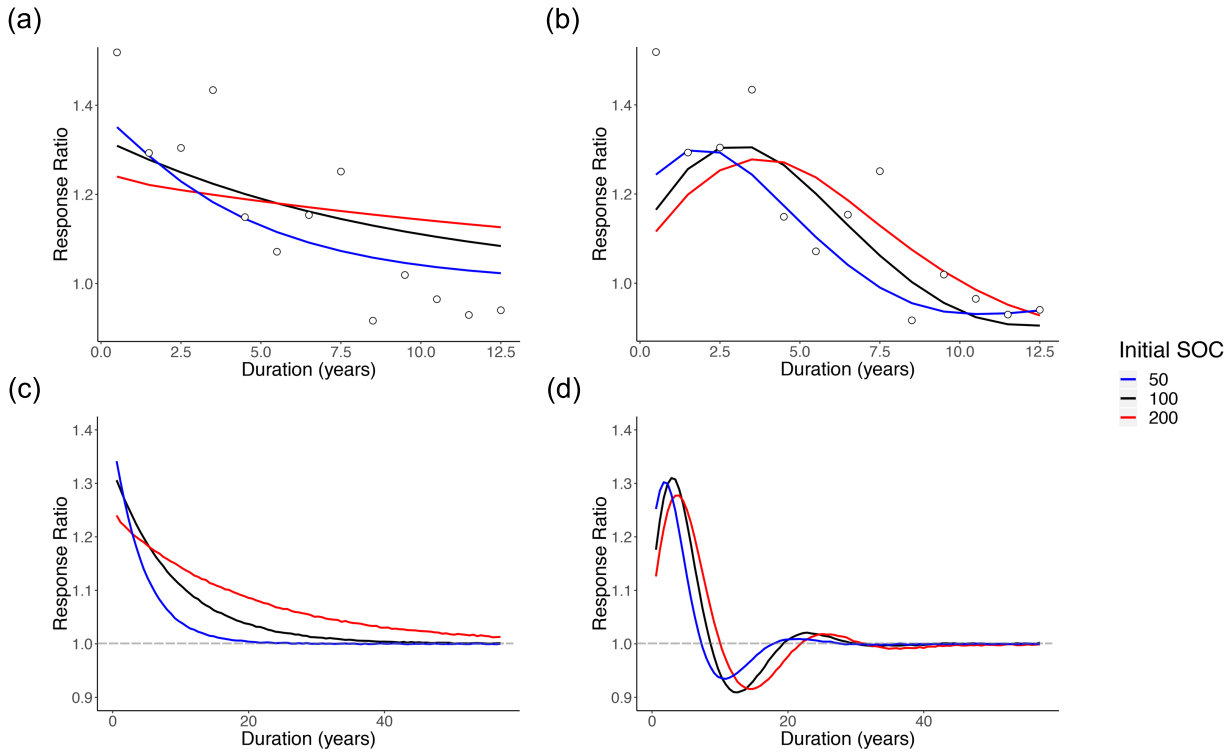

Figure 3.3: Intramodel comparisons of mean posterior predictive response ratio fits for AWB and CON across different MIC-to-SOC ratios. Open circles show the meta-analysis data points for reference. The blue, black, and red lines indicate model mean fits corresponding to different prewarming steady-state SOC values of 50, 100, and 200 mg C g^{-1} soil. The dashed gray line indicates the steady-state expectation at the response ratio of 1.0. Mean fits are plotted in order of **(a)** CON and **(b)** AWB over the time span of the data and **(c)** CON and **(d)** AWB over 57 years.

3.3.3 Sensitivity analysis of parameter distributions to prewarming C pool densities and density ratios

In addition to response ratio fits, we observed the influence of prewarming MIC-to-SOC ratios on model SOC stock response ratios in AWB and CON simulations following warming. Similar to the model flux response ratios, SOC response ratios were calculated by dividing evolved postwarming SOC densities by prewarming densities. The SOC response ratios at 12.5 years for CON and AWB increased as prewarming SOC was raised (and, hence, the MIC-to-SOC ratio decreased) with other prewarming C densities held constant, indicating reduced proportional SOC loss when

SOC stocks were initiated at higher prewarming densities. For CON, SOC loss decreased from 27.1% at SOC50 to 9.2% at SOC200. In a similar trend for AWB, SOC loss decreased from 17.2% at SOC50 to 8.1% at SOC200. In contrast, raising prewarming MIC densities (and, hence, increasing the MIC-to-SOC ratio) with other prewarming C densities held constant did not produce a shared trend for CON and AWB. CON SOC loss decreased from 18.8% at MIC1 to 17.4% at MIC8, while AWB SOC loss increased from 11.3% at MIC1 to 16.3% at MIC8.

Truncation of prior supports, or distribution domains, generally did not prevent posterior densities from retaining normal distribution shapes. Deformation away from Gaussian shapes for the densities of Ea_S from CON was observed at SOC50 and SOC75. For AWB, deformation was observed for the densities of Ea_V, Ea_K, and $E_{C,\mathrm{ref}}$. All CON and AWB parameter posterior densities were otherwise observed to be Gaussian from SOC100 to SOC200. Example posterior densities and means for select model parameters at prewarming SOC100 are presented in Figure 3.4.

3.3.4 Sensitivity analysis of quantitative fit metrics to prewarming C pool densities and density ratios

For both CON and AWB, LOO, WAIC, LPML, and R^2 all worsened as prewarming steady-state SOC density was increased from SOC50 to the less biologically realistic SOC200 (Figure 3.5). CON's LOO and WAIC values increased respectively from -15.704 and -15.818 at SOC50 to -6.891 and -6.966 at SOC200, while AWB's LOO and WAIC values increased respectively from -11.028 and -11.379 at SOC50 to -5.97 and -6.579 at SOC200. Compared to AWB's metrics, CON's goodness-of-fit metrics deteriorated at a faster rate with the increase in prewarming SOC. Nonetheless, CON outperformed AWB in LOO, WAIC, and LPML across all observed prewarming SOC densities. The Bayesian metrics accounted for AWB's larger model size and increased propensity for overfitting as demonstrated by the consistently higher effective parameter counts

associated with AWB.

Varying prewarming steady-state MIC from MIC1 to MIC8 modestly impaired goodness-of-fit across the various metrics. CON's LOO and WAIC values increased respectively from -11.963 and -12.035 at MIC1 to -11.731 and -11.802 at MIC8, while AWB's LOO and WAIC values increased respectively from -8.63 and -9.302 at MIC1 to -8.181 and -8.711 at MIC8. CON did not deteriorate in goodness-of-fit at a faster rate than AWB with respect to increasing prewarming MIC. Increasing prewarming MIC has the opposite effect on the MIC-to-SOC ratio compared to increasing prewarming SOC, but both changes worsened goodness-of-fit across all metrics, indicating that changes to the prewarming MIC-to-SOC ratio did not produce consistent trends.

3.4 Discussion

Our study develops a quantitative, data-driven framework for model comparison that could be applied across different research questions, ecosystems, and scales. We demonstrated the novel deployment of WAIC and LOO, two more recently developed Bayesian goodness-of-fit metrics that estimate model predictive accuracy, to evaluate SBMs using data from longitudinal soil warming experiments. WAIC and LOO improve upon older and more frequently used metrics, such as AIC and DIC, by accounting for model complexity and overfitting of data in a more comprehensive, stable, and accurate fashion. The quantitative agreement between WAIC, LOO, and LPML reinforces the reliability and validity of information criteria and cross-validation metrics to complement use of frequentist R^2.

We constrained the fitting of AWB and CON to bio- logically reasonable parameter space by fixing prewarming steady-state C pool densities and establishing prior distributions informed by expert judgment. We observed that, despite the qualitative difference in the shapes of their mean posterior predictive fit curves, CON and AWB could both potentially account for the

soil warming response in the meta-analysis data set. For both models, posterior predictive fit distributions overlapped with the confidence intervals of the data points (Figure 3.2). However, with respect to the Bayesian goodness-of-fit metrics, CON quantitatively outperformed AWB across all prewarming SOC and MIC densities observed (Figure 3.5) because the Bayesian metrics adjusted for AWB's larger model size and consistently higher effective parameter count. For both models, lower prewarming SOC densities corresponded to better warming response fits (Figure 3.5).

3.5 Model responses to warming over time

After fitting, the response ratio curves of CON and AWB both trended toward the prewarming steady-state response ratio of 1.0 following the soil warming perturbation (Figure 3.3). The settling of the curves to the prewarming model steady states aligns with previous literature which demonstrated that the magnitude of CO2 flux tends to fall after reaching a postwarming maximum (Crowther et al., 2016; Romero-Olivares et al., 2017). In the meta-analysis data set, this peak is reached immediately at the first data point at $t = 0.5$ year (Figure 3.2). CON matched this data pattern in all of our observed simulations in outputting maximum response ratios at the first time point after warming. AWB was unable to output maximum response ratios at the first time point (Figure 3.3b, d) and was therefore penalized in quantitative goodness-of-fit. Examining AWB's system of equations, we surmise that one reason for the later peak was due to the slower growth of MIC in the biologically truncated parameter space that AWB was limited to. MIC is a driving force for the increase in the CO2 flux as a numerator term in the AWB flux equation. Unlike MIC biomass in CON, MIC biomass growth in AWB has two loss terms in its differential equation.

This is not to say that CON was clearly superior from a qualitative standpoint. CON's mean posterior predictive curves were not able to match a subsequent local data maximum in the meta-analysis data set at $t = 3.5$ year, a trend which AWB's curves were able to replicate. The

mean CON curves also substantially overshoot the data at later time points following $t = 7.5$ year (Figure 3.2a and 2.3a, c) because of the inability of first-order linear models such as CON to display oscillatory dynamics (Hale and LaSalle, 1963).

In contrast, AWB displays damped oscillations in its re- sponse ratios following warming due to its nonlinear dynamics (Figure 3.2 and 3.3). AWB was able to match the points after $t = 7.5$ year more closely than CON. The presence of respiration oscillations has been observed in long-term warming experiments, such as the one taking place at Harvard Forest (Melillo et al., 2017). It is possible AWB would be quantitatively rewarded in goodness-of-fit metrics over CON for its ability to replicate biologically realistic oscillations in larger, site-specific data sets such as those from Harvard Forest.

3.6 Sensitivity analyses of C pool densities and density ratios

We performed a goodness-of-fit sensitivity analysis to check whether the response ratio trends stayed consistent, biologically realistic, and interpretable across a range of prewarming, steady-state soil C densities and pool-to-pool density ratios. For instance, we imposed constraints to reflect that MIC-to-SOC density ratios range between 0.01 and 0.04 across various soil types (Anderson and Domsch, 1989; Sparling, 1992). CON and AWB response ratio curves exhibited realistic values and qualitatively consistent shapes across all prewarming SOC and MIC steady-state densities, even at less realistic SOC densities above 100 mg C g^{-1} soil (Figure 3.3). There was enough uncertainty in the data that the 95% posterior predictive intervals for the model output always overlapped with the 95% confidence intervals of each fitted data point (Figure 3.2). In most cases, the posterior mean response ratio curve also fell within the 95% data confidence interval.

We were unable to initiate our prewarming SOC steady-state density below SOC50 with the priors and MIC-to-SOC ratios used for AWB. Under SOC50, AWB HMC runs would not reliably run to conclusion and would terminate due to ODE instabilities. Even at SOC50, we saw a reduction in independent and effective samples for certain parameters, namely Ea_K and $E_{C,\mathrm{ref}}$. We did not drop under SOC50 for CON, as we sought to compare AWB and CON at similar MIC-to-SOC ranges. Our experience underscores the challenge of choosing realistic steady-state soil C densities, density ratios, and prior distributions to obtain valid model comparisons limited to biologically realistic regimes.

The information criteria and cross-validation fit metrics generally indicated higher relative probability and predictive performance at lower prewarming SOC values for AWB and CON (Figure 3.5). The fit results suggest that SOC density of the soil at the sites included in the meta-analysis was likely closer to the lower end of the SOC density ranges examined in our sensitivity analysis. A less pronounced trend toward better fits was observed as prewarming MIC density was decreased while prewarming SOC density was held constant. No clear relationship was observed between MIC-to-SOC ratio and goodness-of-fit in the AWB and CON models.

The worsening IC and CV results at higher SOC densities support the notion that prewarming steady-state SOC densities should not be initialized over SOC100 in AWB and CON when fitting to this meta-analysis data set. Prewarming SOC density was not observed to exceed 50 mg C g^{-1} soil at sites included in the meta-analysis, reaching a maximum of 45 mg SOC g^{-1} soil for the top 20 cm in one study with alpine wetland soil (B. Zhang et al., 2014). The majority of the CO_2 respired by soil microbes is sourced from surface soil (Fang and Moncrieff, 2005), and it is well documented that SOC densities increase toward the soil surface (Jobbágy and Jackson, 2000). 14C measurements of CO_2 fluxes suggest that SOC densities representing the source of most heterotrophic respiration range between 40 and 80 mg SOC g^{-1} soil (Trumbore, 2000), so the effective SOC densities associated with soil respiration at some meta-analysis sites may have been in this range.

Overall, the Bayesian metrics from the goodness-of-fit sensitivity analysis suggest that CON is superior to AWB at explaining the meta-analysis data set when accounting for model parsimony, particularly when the models are initiated in more realistic ranges of prewarming SOC densities under SOC100. However, we caution against using these results to conclude that CON is a comprehensively superior predictive model over AWB without comparisons involving other longitudinal soil warming data sets. And other data aside, we observe that AWB has a useful advantage over CON conditional on the meta-analysis data set alone: AWB was more tolerant of changes in prewarming conditions, displaying less IC and CV than CON as prewarming SOC is increased (Figure 3.5a-c). AWB's compensatory ability stemming from its larger model size could be more quantitatively rewarding in goodness-of-fit sensitivity analyses conducted on data assimilations with larger data sets.

For an additional check on the biological realism and plausibility of our simulations, we conducted a sensitivity analysis examining changes in model SOC stocks following warming. The response ratios of postwarming SOC stocks after 12.5 years, evaluated as the ratio of postwarming to prewarming SOC densities, was computed from observed CON and AWB simulations at the posterior parameter means. SOC losses indicated by the response ratios ranged from 8.13% to 27.1% across both models. These results aligned with a recent comprehensive meta-analysis of 143 soil warming studies. The largest loss of 27.1%, occurring in CON at SOC50, is sizable, but the meta-analysis included seven studies measuring losses greater than 20%, with the maximum loss observed at 54.4% (van Gestel et al., 2018).

Raising prewarming SOC reduced SOC loss after 12.5 years of warming for both models. For CON, SOC loss decreased from 27.1% at SOC50 to 9.2% at SOC200. For AWB, SOC loss decreased from 17.2% at SOC50 to 8.13% at SOC200. Varying prewarming MIC affected the SOC response ratio more substantially for AWB than CON. For AWB, SOC loss increased from 11.4% at MIC1 to 16.3% at MIC8, while SOC loss decreased from 18.8% at MIC1 to 17.4% at MIC8 for CON. The larger effect of increasing MIC on the SOC response ratio in AWB is likely

due to MIC influence on SOC-to-DOC turnover, which is not a feedback accounted for in the equations of the CON model.

The posterior means for the Arrhenius activation energy parameters Ea_i of CON and AWB returned by the HMC simulations across the observed prewarming C densities differed somewhat from the parameter values used in Allison et al. (2010) and J. Li et al. (2014), which were in turn tuned based on activation energies estimated in a prior empirical analysis of enzyme-catalyzed soil organic matter decomposition processes (Trasar-Cepeda et al., 2007). In Allison et al. (2010), CON parameters Ea_S, Ea_D, and Ea_M were respectively set at 47, 40, and 40 kJ mol^{-1}, and AWB parameters Ea_V and Ea_{VU} were both set at 47 kJ mol^{-1}. The AWB Michaelis–Menten K_M terms were not parameterized to have Arrhenius temperature dependence in Allison et al. (2010). In J. Li et al. (2014), CON parameters Ea_S, Ea_D, and Ea_M were set at 47, 47, and 20 kJ mol^{-1}, and AWB parameters Ea_V , Ea_{VU}, Ea_K, and Ea_{KU} were set at 47, 47, 30, and 30 kJ mol^{-1}. These values were in line with the activation energies calculated in Trasar-Cepeda et al. (2007), which ranged from 17.0 to 57.7 kJ mol^{-1}, with the energies corresponding to the decomposition of plant litter and protected organic matter being on the higher end and the energies corresponding to microbial biomass degradation being on the lower.

Our HMC simulations arrived at higher Ea values, with the posterior means of Ea_S, Ea_D, and Ea_M respectively ranging from 51.3 to 77.6, 50.1 to 50.3, and 51.8 to 52.6 kJ mol^{-1} in the prewarming SOC-varied simulations for CON, and the posterior means of Ea_V, Ea_{VU}, Ea_K, and Ea_{KU} respectively ranging from 58.5 to 74.8, 50.2 to 51.1, 25.8 to 42.4, and 49.0 to 49.8 kJ mol^{-1} for AWB. However, these values are still within the ranges of organic matter decomposition activation energies, which have been empirically estimated to exceed 100 kJ mol^{-1} at their highest in the A horizons of temperate soils (Steinweg et al., 2013), suggesting that the Ea posterior means, aided by prior truncation, effectively remained within biologically realistic space across all observed prewarming C densities. The presence of higher Ea_S posterior means also agreed with the empirical trends of higher activation energies for the degradation of SOC-related

organic compounds and lower activation energies for the degradation of material associated with microorganisms.

We found it less useful to compare the posterior means of other fitted parameters including the C pool transfer coefficients, C use efficiency $E_{C,\mathrm{ref}}$ and $V_{\max}$ to empirical estimates for biological benchmarking purposes. Unitless parameters like transfer coefficients and $E_{C,\mathrm{ref}}$ defy straightforward interpretation, measurement, and estimation from experiments (Bradford and Crowther, 2013). Very different values can be found based on whether substrate-specific or substrate-nonspecific assumptions and methods are used (Geyer et al., 2019; Hagerty et al., 2018). $V_{\max}$ parameters are not unitless but display even higher variance than the bounded C transfer and efficiency coefficients. The $V_{\max}$ parameter corresponding to a specific enzyme can vary over orders of magnitude when the sensitivity of the enzyme to an interval of temperatures is considered (Nottingham et al., 2016). The process of consolidating experimental substrate-specific and substrate-nonspecific measurements into a single number to correspond to a model $V_{\max}$ value introduces further complications and uncertainty, rendering comparisons of potentially drastically different $V_{\max}$ values less informative regarding model biological realism.

3.6.1 HMC parameter space exploration

Truncating prior and posterior parameter distributions proved useful for establishing biological constraints and only modestly deformed posterior densities for AWB and CON. From SOC100 to SOC200, CON and AWB posterior densities showed little or no deformation from typical normal distribution shapes. Moderate posterior density deformation was observed for some parameters in both models at SOC50 and SOC75, namely Ea_S for CON and $E_{C,\mathrm{ref}}$ for AWB. Even so, most of the other parameter posterior densities still remained undeformed at those SOC values. Thus, prior truncation generally did not prevent posterior means from falling within biologically realistic intervals, suggesting that priors were appropriately informed and chosen.

A small frequency of divergent transitions was detected in the AWB HMC simulations. Divergent transitions can be thought of as algorithm trajectory errors arising during the HMC's exploration of a convoluted region of parameter space; a more thorough description of the theory, computation, and implications of divergent transitions can be found in literature focusing on the Hamiltonian Monte Carlo algorithm (Betancourt, 2016, 2017). The number of divergent transitions generally increased as the prewarming MIC-to-SOC steady-state ratio was reduced. Prior truncation and the fixing of select parameters to constrain the prewarming steady-state mass values for biological realism could have played a combined role in generating the Markov chain divergences by hindering the smooth exploration of parameter space. We were unable to eliminate divergent transitions by adjusting HMC parameter proposal step size, suggesting that other methods, such as modification of the HMC algorithm itself or introduction of auxiliary parameters to AWB that reduce correlation between existing model parameters, may be more applicable in reducing divergent transitions in our case (Betancourt and Girolami, 2015). Additionally, the interaction between the ranges of values used for the prior distributions and the limited number of observations in the data set could have contributed to the shaping of geometric inefficiencies (Betancourt, 2017).

It is possible that the instability that prevented consistent solving and HMC exploration of AWB under SOC50 could be traced to the forward Michaelis-Menten formulation of decomposition and uptake kinetics used in the present version of the AWB model. We initialized the system with a small DOC density lower than that of MIC at 0.1 $\mathrm{mg\,C\,g^{-1}}$ soil. Since DOC was in the denominator of these decomposition and uptake expressions, those expressions could become larger than tolerable for the system in certain parameter regimes.

Some suggestions for the reparameterization of AWB to improve model stability have been proposed that could reduce or even eliminate divergent transitions by facilitating a smoother and steadier parameter space conducive for HMC exploration. One intermediate possibility would be to modify AWB to use reverse Michaelis-Menten kinetics, which would replace the DOC term in the denominators of the decomposition and uptake expressions with the larger MIC term. The

use of reverse instead of Michaelis-Menten dynamics has been used to stabilize and constrain other SBMs (Sulman et al., 2014; Wieder, Grandy, Kallenbach, Taylor, et al., 2015). A more extensive reformulation involves the replacement of Michaelis-Menten expressions with equilibrium chemistry approximation (ECA) kinetics, which would increase the number of denominator terms in decomposition expressions for further stability. ECA equations have been shown to be more consistent in behavior and robust to parameter regime variation than their Michaelis-Menten counterparts, and they thus have been encouraged as a wholesale replacement for Michaelis-Menten formulations (Tang, 2015; B. Wang and Allison, 2019). These reparameterizations should be implemented and examined in future work that involves sampling and computation of AWB posteriors.

3.7 Outlook and conclusions

Recent SBM comparisons have been unable to demonstrate the superiority of one model over another because the uncertainty boundaries of the data were not sufficient for distinguishing model outcomes (Sulman et al., 2018; Wieder, Boehnert, et al., 2014; Wieder, Grandy, Kallenbach, Taylor, et al., 2015; Wieder et al., 2018). Similar to these previous studies, our results indicate that more data is needed to constrain and differentiate between model posterior predictive distributions. Conditional on the meta-analysis data set, CON demonstrates superior quantitative goodness-of-fit over AWB, but we are not confident that the relative model parsimony of CON and other linear first-order models makes them universally more suitable for predictive use.

Consequently, future SBM comparisons would benefit from additional data collection efforts sourced from long-term ecological research experiments to globally verify the strengths and limitations of linear versus nonlinear SBMs, including CON and AWB, in Earth system modeling. The limited number of longitudinal soil warming studies presents a challenge for facilitating site-specific model comparisons. We addressed this issue by using meta-analysis data to aggregate

warming responses across sites, but this approach does not provide site-specific parameters. Additional data from ongoing and future field warming studies in the vein of the Harvard Forest and Tropical Responses to Altered Climate experiments that demonstrate more varied flux dynamics over time than the meta-analysis data set will be of critical importance for model testing (Melillo et al., 2017; Wood et al., 2019). Model parameters could also be better constrained through the use of multivariate data sets, for example microbial biomass dynamics in addition to soil respiration.

Our approach can be expanded to compare the predictive accuracies of linear microbial-implicit models to those of recently developed nonlinear microbial-explicit SBMs that are much larger than AWB, such as CORPSE (Sulman et al., 2014) and MIMICS (Wieder, Boehnert, et al., 2014). Such comparisons will help broadly determine if inclusion of more detailed microbial dynamics in models offers predictive advantages that can overcome the overfitting burdens associated with an increase in parameter count. With the appropriate data sets, our approach can also be applied to consider the predictive performance of SBMs that describe the cycling of nitrogen (N), phosphorus (P), and other limiting nutrients in addition to C dynamics. Models that represent N and P mineralization have yet to see extensive head-to-head statistical benchmarking against C-only models with respect to predictive use (Manzoni and Porporato, 2009). With models growing ever larger in size and specificity, there is a need to verify whether detailed representation of microbial processes and the cycling of limiting nutrients are worth the increase in variable, parameter, and equation counts. After all, "the tendency of more recent models towards more sophisticated (and generally more mathematically complex) approaches is not always paralleled by improved model performance or ability to interpret observed patterns" (Manzoni and Porporato, 2009).

The data assimilation and posterior sampling of complex models in future work comes with computing performance challenges. Markov chain Monte Carlo algorithms are effective for exploring multidimensional parameter space but are limited by temporal and computational expense, particularly when it comes to fitting nonlinear differential equation models (Calderhead et al., 2009;

Nemeth and Fearnhead, 2021). Time per Markov chain iteration drastically increases with number of parameters and data points. In fact, the present speed limitations of the family of HMC algorithms make it necessary to use a hybrid approach utilizing Monte Carlo and deep learning algorithms for parameter estimation at a global scale; Monte Carlo fitting is used to constrain parameter estimates at a site-based scale before those estimates are tuned globally by deep learning using spatial information derived from satellite maps (Tao et al., 2020). However, Monte Carlo algorithms are still the optimal methods for posterior computation (Duan et al., 2018) and are necessary for Bayesian model comparisons conditional on site-based data. Consequently, recent Monte Carlo algorithm innovations and developments that offer theoretical speed improvements by trading thorough posterior sampling for numerical efficiency have been encouraging and are ripe to be tested in future SBM comparisons involving more complex models and larger data sets. These developments include stochastic gradient Monte Carlo sampling methods, a class of techniques in which a posterior is approximated by fitting to a small subset of data at each iteration rather than estimated through exhaustive sampling (Ma et al., 2015), and Gaussian process acceleration, in which a smooth distribution of likely solutions for a differential equation system is specified and sampled in place of explicitly solving for the state variables during every Markov chain iteration (Calderhead et al., 2009; Dondelinger et al., 2013; Y. Wang and Barber, 2014).

Alongside advances in Monte Carlo algorithms, additional developments in Bayesian cross-validation and information criteria measures are also available for practical trialing in soil biogeochemical data assimilation. Gelman et al. (2019) have proposed a stable Bayesian counterpart of frequentist R^2 defined as "the variance of the predicted values divided by the variance of predicted values plus the expected variance of the errors" that allows for more intuitive and direct comparison to R^2. A Bayesian R^2 distribution provides a signal about the absolute rather than relative goodness-of-fit of an associated posterior predictive distribution to the data. Bürkner et al. (2020) have proposed a leave-future-out (LFO) cross-validation metric which is formulated to estimate relative model predictive accuracy for hypothetical time series data occurring after existing experiment obser-

vations. LFO and LOO are computed similarly, and LOO can also be used for time series data, as we demonstrated in this study. However, the algorithmic differences between LFO and LOO make them better suited for different goals. LOO does not inform about the quality of model fits for hypothetical samples collected after final reported measurements and is more appropriate for estimating out-of-sample model predictive accuracy for hypothetical data samples taken between the interval of observed measurement times (Vehtari et al., 2017).

The development of our formalized, statistically rigorous approach for model comparison and evaluation is a critical step toward the goal of projecting global SOC levels and soil emissions throughout the 21st century. Our initial results indicate promise in continued refinement and expansion of our approach to evaluate the predictive performance of linear and nonlinear SBMs. The future integration of updated Markov chain algorithms and Bayesian predictive accuracy metrics into our framework will expand the ability to efficiently and thoroughly compare differential equation models, even if they vary widely in structure and complexity.

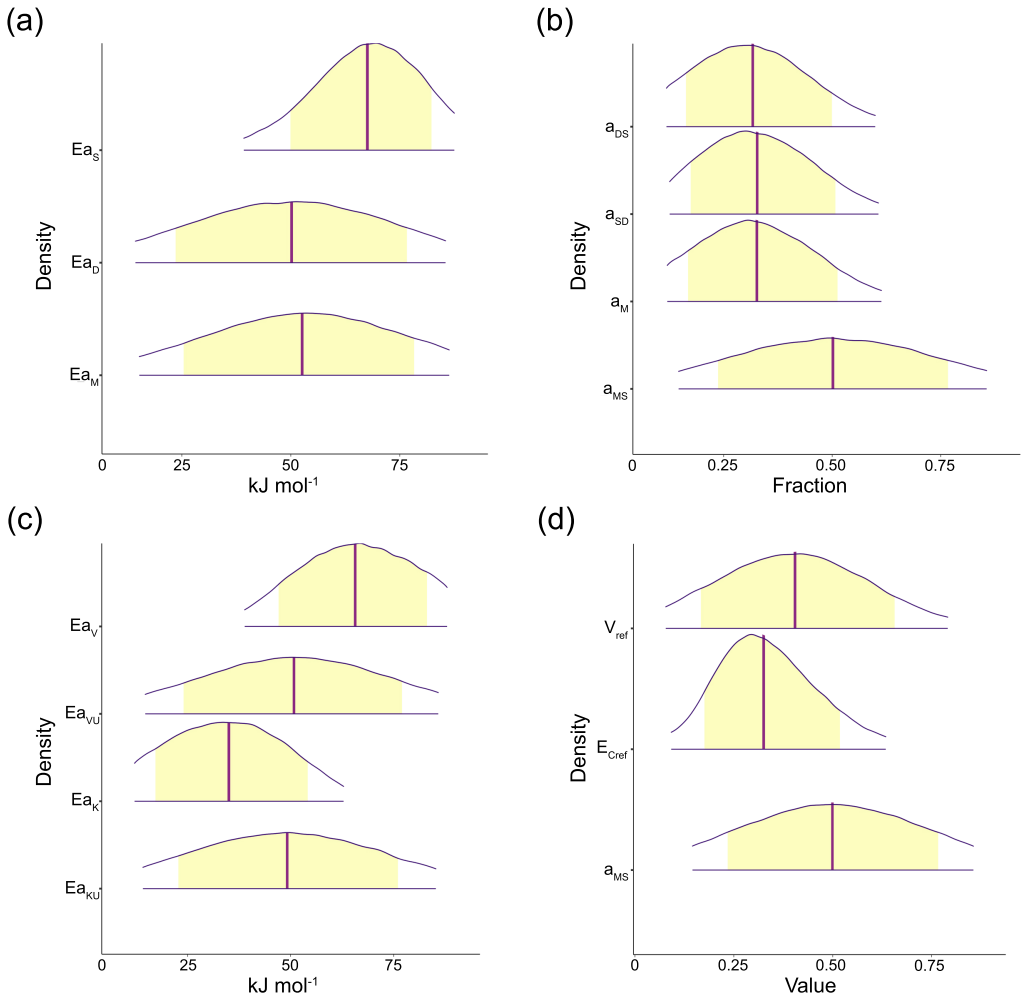

Figure 3.4: The 95% probability density credible areas for model parameters corresponding to pre-warming steady- state SOC = 100 mg C g^{-1} soil, DOC = 0.2 mg C g^{-1} soil, MIC = 2 mg C g^{-1} soil, and (for AWB) ENZ = 0.1 mg C g^{-1} soil. Yellow shaded regions represent the 80% credible areas and vertical purple lines indicate distribution mean. **(a)** CON activation energy parameters Ea_S , Ea_D , and Ea_M ; **(b)** CON C pool partition fraction parameters a_{DS}, a_{SD}, a_M, and a_{MS}; **(c)** AWB activation energy parameters Ea_V , Ea_{VU}, Ea_K, and Ea_{KU}; **(d)** AWB parameters V_{ref}, $E_{C,\text{ref}}$, and a_{MS}. V_{ref} is the SOC $V_{\max}$ at the reference temperature 283.15 K; $E_{C,\text{ref}}$ is the carbon use efficiency fraction at the reference temperature; and, like its CON counterpart, the AWB a_{MS} parameter is the fraction parameter representing the proportion of dead microbial biomass C transferred to the SOC pool.

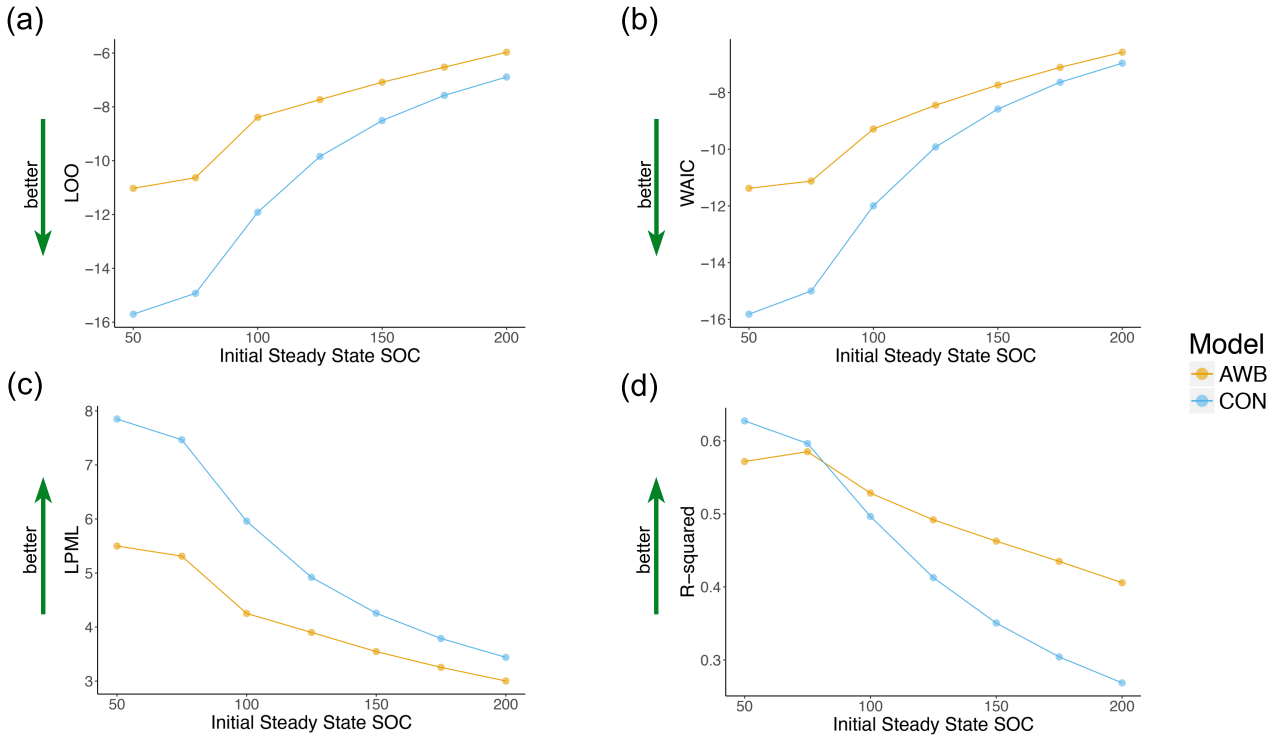

Figure 3.5: Goodness-of-fit metrics plotted against initial steady-state SOC for AWB and CON models for **(a)** LOO, **(b)** WAIC cross validation, **(c)** LPML, and **(d)** R^2 values. Pre-perturbation steady-state MIC, DOC, and ENZ (for AWB) are held constant as pre-perturbation SOC is varied.

CHAPTER 4

A framework for variational inference and data assimilation of soil biogeochemical models using state space approximations and normalizing flows

4.1 Introduction

Soil biogeochemical models (SBMs) are differential equation systems that represent dynamics of organic matter transfer between soil pools, including the soil organic (SOC), dissolved organic (DOC), and microbial biomass carbon (MBC) pools. The state variables of SBMs typically are densities or masses of elements in those pools (Manzoni and Porporato, 2009), and heterotrophic soil CO_2 emissions can be estimated from those state values and microbial parameters (Allison et al., 2010). As soil microbe communities influencing organic mass transfer dynamics evolve and shift under the selection pressures of terrestrial warming, SBMs have become an important tool for soil scientists and biogeochemists to quantify changes in soil system activity and predict

future heterotrophic soil respiration levels (Saifuddin et al., 2021; Sulman et al., 2018).

SBMs offer falsifiability of their dynamics through their depiction of biological soil processes as interpretable mathematical equations governed by model parameters θ. However, the formulation of statistically sound frameworks to assess the dynamical validity and predictive accuracy of SBMs remains an open problem in soil biogeochemistry (Bradford et al., 2021; Georgiou et al., 2021; Luo et al., 2016; Raczka et al., 2021; Xie et al., 2020). One approach for assessing SBM utility involves comparing models by their ability to assimilate soil observation data with suitable θ values, assuming that models that can more accurately describe the past will also be better at predicting the future (Bradford et al., 2021; Wieder, Boehnert, et al., 2014). Past SBM fit evaluations have ranged from visual juxtapositions of manually calibrated model outputs against empirical observations (Sulman et al., 2014; Wieder, Grandy, Kallenbach, Taylor, et al., 2015) to quantitative frequentist comparisons involving correlation coefficients and root-mean-square errors (Todd-Brown et al., 2014; Todd-Brown et al., 2013; Wieder, Boehnert, et al., 2014). In an effort to account for uncertainty in θ values and data observations and encode expert domain beliefs, other comparisons have involved the use of Bayesian Markov chain Monte Carlo (MCMC) inference methods and goodness-of-fit metrics with some success (Hararuk and Luo, 2014; Hararuk et al., 2014; J. Li et al., 2019; Saifuddin et al., 2021; S. Wang et al., 2022; Xie et al., 2020).

MCMC transition sampling methods, such as the Gibbs (Geman and Geman, 1987) and No-U-Turn (NUTS; Hoffman and Gelman, 2014) samplers deployed in widely-used probabilistic programming platforms like JAGS (Plummer, 2003), Stan (Carpenter et al., 2017), and PyMC (Salvatier et al., 2016), are powerful algorithms for inference, but their relative computational cost presently limits their ability to scale for use on model comparisons involving more complex SBM systems conditioned on larger data sets spanning decades (Kucukelbir et al., 2017). Stochastic gradient optimization variational inference (VI) is an alternative approach to Bayesian inference and model-fitting that trades asymptotic exactness and the ability to estimate non-parametric

posterior distributions for increased computational efficiency and simplicity (Blei et al., 2017). It does so by reframing Bayesian inference from a transition sampling problem to an optimization objective of maximizing a metric called the evidence lower bound (ELBO), which corresponds to minimizing the discrepancy between an approximate parametric posterior and true posterior distribution (Salimans et al., 2015).

VI on differential equation models benefits from the use of stochastic differential equation (SDE) over ordinary differential equation (ODE) systems. SDE noise provides a means of adjusting and correcting for proposals of system initial conditions and underlying dynamics that are inconsistent with the true data-generating process sourcing the data observations (Särkkä and Solin, 2019; Whitaker, 2016; Wiqvist et al., 2021). Additionally, noise-driven fluctuation and variation in state trajectories can account for outlier data measurements during inference to reduce optimization pressures that can drive rigid deterministic models into unstable θ regimes. SDE noise thereby improves inference flexibility, stability, and efficiency through the acommodation and mitigation of discrepancies between model outputs and data generation or observation. Furthermore, SDEs offer a more realistic and accurate representation of the stochasticity that is inherent to biological processes across all scales (Abs et al., 2020; Browning et al., 2020; Golightly and Wilkinson, 2011). The ability to effectively fit SDEs is an advantage of VI over many established MCMC methods; off-the-shelf MCMC implementations are frequently not efficient for confronting the noisy likelihood estimates of SDEs (Chen et al., 2014; Fuchs, 2013; Golightly and Wilkinson, 2010).

With the goal of applying VI to SBMs in mind, we formulated SDE versions of the linear deterministic "conventional" (CON) SBM system (Allison et al., 2010; J. Li et al., 2014) to establish an SCON family of models and leverage the versatility of stochastic optimization. As is the case for CON, SCON models have three state variables representing SOC, DOC, and MBC densities in a soil system. We parameterized two SCON variants, "constant diffusion" SCON (SCON-C) and "state-scaling diffusion" SCON (SCON-SS). Diffusion coefficients are model parameters that

govern the noise dynamics of an SDE system. In SCON-C, diffusion, or noise, was set to be independent of time and states, while in SCON-SS, noise was made to depend on and scale with state values.

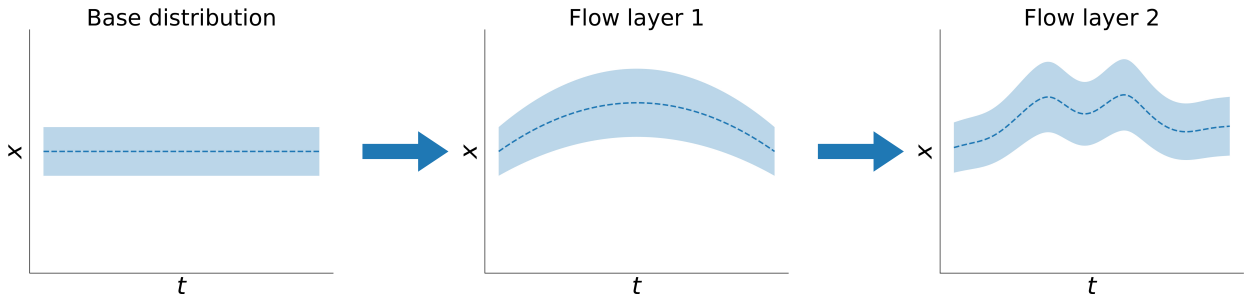

Figure 4.1: In our study, we use normalizing flows to approximate SCON soil biogeochemical model solution trajectories x over time t. The flow operates in a generative direction, mapping a simpler base distribution to a more complex one representing SCON output.

We used a class of methods called *normalizing flows* to approximate SCON models in our inference approach. In simple terms, flows can be thought of as one or more layers of random variable mappings that transform an initial base probability distribution to a new distribution (Papamakarios et al., 2021). When we deploy flows to transform a simpler probability density into a more complex one (Figure 4.1), as we do in our study, it is classified as a *generative* normalizing flow (Kobyzev et al., 2020). The flow approximation refashions SCON from an SDE that depicts state variable dynamics $\frac{\mathrm{d}x}{\mathrm{d}t}$ in continuous time to a probabilistic state space model that specifies distributions of state measurements y_t noisily observed from underlying states x_t in discrete time (Särkkä and Solin, 2019).

The replacement of differential equation solver integration with state space models to approximate dynamical systems offers substantial computational efficiency gains in inference (Ryder et al., 2018; Särkkä and Solin, 2019). At each inference iteration or epoch, rather than sequentially computing state trajectories x one time step at a time with solvers including Euler, Runge-Kutta, and Adams' schemes, as was demonstrated in studies like Xie et al. (2020), we can instead simultaneously sample multiple x in one vectorized draw from a flow-transformed state space distribution object. This increased efficiency allows us to more capably assimilate SBMs with

time series data sets spanning longer periods under computing resource limitations, especially when highly parallelizable graphical processing units (GPUs) can be leveraged.

Drawing from the methodologies of previous work that test various inference approaches (Golightly and Wilkinson, 2006; Ryder et al., 2021; Ryder et al., 2018; Whitaker et al., 2017), our study demonstrates functional stochastic VI of flow-approximated SBMs conditioned on soil observations data y that includes various soil pool and respiration measurements. To support the notion that our VI approach is operational, we show that it can fit model output to y sourced from a known data-generating process and estimate model θ posteriors in line with the true θ values used by that process.

Hence, to begin our study workflow, we generated synthetic y consisting of SOC, DOC, and MBC state and heterotrophic CO_2 respiration rate observations corresponding to SCON-C and SCON-SS data-generating processes. The processes used "true" θ values randomly sampled from constrained data-generating distributions that were chosen to produce faster and more dramatic SOC decay dynamics reminiscent of organic matter decomposition at soil surface, which contrasts with the slower and deeper soil decomposition depicted in J. Li et al. (2014) and Xie et al. (2020). Faster decay provided our inference approach with substantive dynamical information in shorter time series to operate and optimize on. We then conditioned our state space model VI on those synthetic y for estimation of approximate posterior densities $q(\theta)$ that were compared with prior densities $p(\theta)$. Priors were made to be equivalent to our data-generating distributions.

Ultimately, we found that our VI approach allowed us to reasonably fit y. When possible, our fits were checked against solutions from a Kalman smoother algorithm, and we observed that the flow fits were mostly consistent with the Kalman solutions. Crucially, we were also able to recover some of the true θ values used by our data-generating processes against model identifiability limitations that could not be resolved by the extent of information contained in our synthetic data. *Model identifiability* can be summarized as the ability to update prior beliefs about θ and align model to true θ based on available data. Our identifiability issues related to the presence of

ambiguous SCON equation terms involving the multiplication of more than two parameters. Our work offers insights and suggestions for improving the identification of θ, which is of interest for experimentalists and biogeochemists who are interested in building effective data sets for SBM inference.

4.2 Materials and Methods

4.2.1 Inference workflow overview

The general steps constituting our study's SBM data assimilation workflow are outlined in Figure 4.2. We established SCON-C and SCON-SS to serve as known data-generating processes whose true θ values can be compared with the inferred posteriors to test our flow VI method. Discrepancies between the true θ and posterior means inform on the effectiveness of our selected inference algorithm. Differences between the priors and posterior densities further indicate algorithm efficacy and additionally point to the informativeness of the data y for identifying and constraining posteriors.

True SCON-C and SCON-SS θ were sampled from data-generating distributions truncated between lower and upper support bounds to ensure that data-generating processes would remain in parameter regimes with faster state decay corresponding to soil surface decomposition occurring on the order of thousands of hours, rather than tens or hundreds of thousands. This allows us to generate shorter data sets y that enable reduced computational loads and faster turnaround times for testing our inference algorithm while retaining dynamical richness that can inform the algorithm to estimate more certain posteriors. We used logit-normal distributions to handle truncation in our data-generating, prior, and posterior distributions, which we will describe in section 4.2.3. Our inference priors matched our data-generating distributions.

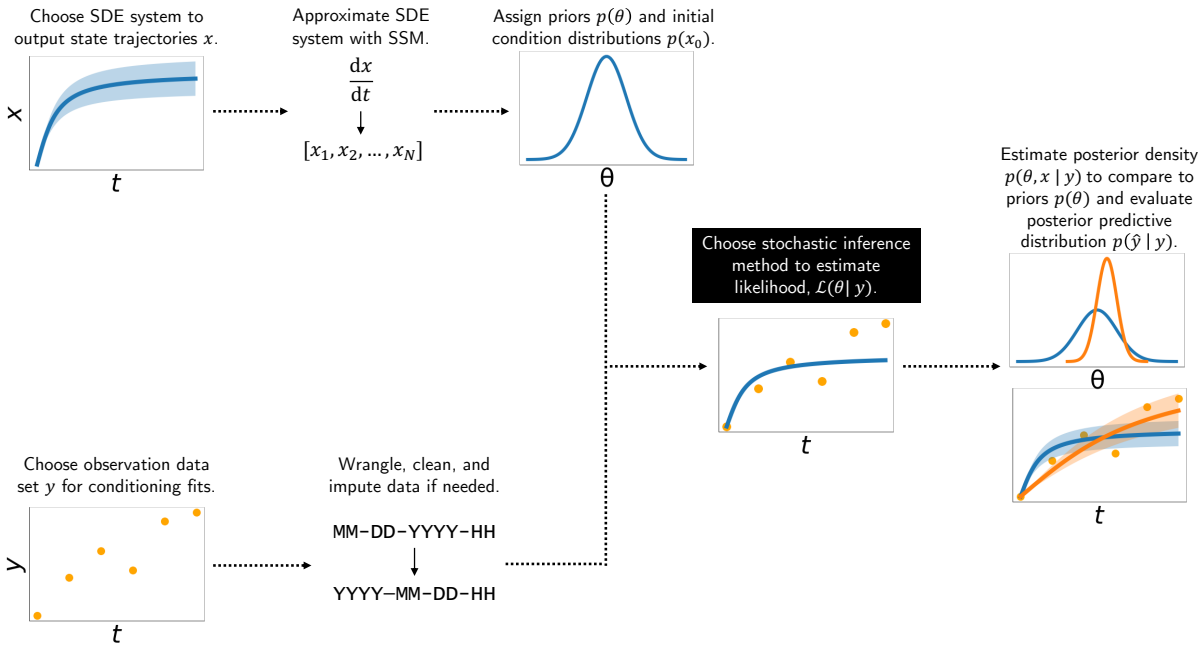

Figure 4.2: A workflow diagram summarizing the steps involved in our study's stochastic variational Bayesian framework. Our workflow efficiently conducts inference and data assimilation on stochastic differential equation (SDE) soil biogeochemical models (SBMs) with their approximation into state space models (SSMs). Our modular workflow is designed to serve as a basis for building future soil biogeochemical model inferences, as the "black box" inference method used can be modified or substituted. Our "black box" inference method of choice was stochastic gradient descent mean-field variational inference. Within the nodes of the diagram, blue lines and shading correspond to prior means and distributions, while orange lines and shading correspond to posterior means and distributions. Orange dots represent observations upon which the inference is conditioned.

Synthetic data y were observed and processed from our data-generating SDE solution trajectories. We parameterized our SCON models based in time units of hours, so observations were collected every 5 hours by default. State space approximation of SDE output, which we will describe in section 4.2.4, requires regular time series discretization (Kalman, 1960), so in an empirical setting, all existing, imputed, or missing observations must coincide with discrete time steps of the state space model in our approach and cannot transpire in between. Different SDE approximation methods would be needed for irregular time discretization.

We selected mean-field stochastic VI as our black box inference method for its mathematical

simplicity and efficiency. Mean-field inference makes the simplifying assumption that model parameters are independently distributed. This aligns with our synthetic data-generating processes, in which our true θ values are sampled from independent logit-normal distributions. VI frames Bayesian inference as an optimization goal of finding the set of mean-field posterior distributions that best describes y. The optimization process takes place over a number of training iterations in which θ values are sampled at each iteration and the likelihood of the resultant model output conditioned on y and θ is evaluated in fulfillment of the objective of the VI algorithm to locate θ corresponding to higher model likelihood. We present an overview of our VI implementation and key algorithm steps in section 4.2.5.

We used normalizing flows to approximate SCON-C and SCON-SS from continuous-time SDEs to time-discretized state space models. These state space approximations then served as our bases for VI optimization. A brief treatment on state space models is given in section 4.2.4. Flow state space approximation increased the computational efficiency of sampling SCON solution trajectories (also referred to as *latent variables*, *states*, or *paths* in machine learning literature) such that multiple trajectories would be simultaneously collected from a flow distribution object rather than sequentially simulated from a differential equation solver at each training iteration. The flow is assembled through deep neural network layers that transform simpler random input into more complex approximated SCON output. The constituent pieces of the machine learning architecture underlying our flow are detailed in section 4.2.6.

Per equations (4.3) and (4.4), SCON-C is a completely linear SDE. Consequently, SCON-C flow-approximated x and its fit of y can be visually benchmarked against output from an instance of the Kalman smoother algorithm summarized in section 4.2.7. Given a known data-generating process and observation error, a Kalman smoother exactly solves the true mean latent path x of the SDE data-generating process sourcing y. We successfully compared SCON-C flow x to the true x solution computed by the smoother, which we describe in section 4.3.1. The smoother algorithm cannot resolve the non-linear diffusion depicted in equation (4.5), so SCON-SS flow

output could not be validated in the same manner.

4.2.2 SCON SDE parameterization and data generation

SDE system equations are frequently written with the state value derivatives $\mathrm{d}x$ on the left-hand side, and consist of a drift coefficient vector, frequently notated as α, and a diffusion coefficient matrix, notated as β, on the right-hand side. For biological SDE models, a square-root diffusion structure is frequently used such that these systems follow the form

$$\mathrm{d}x_t = \alpha(x_t, t, \theta)\mathrm{d}t + \sqrt{\beta(x_t, t, \theta)}\mathrm{d}W_t \tag{4.1}$$

where $\mathrm{d}W_t$ denotes a continuous stochastic Wiener process. Evolution of SDE trajectories x across a simulation duration T in time increments $\mathrm{d}t$ can be interpreted as a series of small steps whose values are independently drawn from a normal distribution with mean $\alpha(x_t, t)\mathrm{d}t$ and variance $\beta(x_t, t)\mathrm{d}t$ (Särkkä and Solin, 2019).

Like the CON model introduced in Allison et al. (2010), SCON has three state dimensions made up of soil organic C (SOC), dissolved organic C (DOC), and microbial biomass C (MBC) densities. We notate total state dimensions with $\mathcal{D}$, so $\mathcal{D} = 3$ for all systems in the SCON family. SOC, DOC, and MBC are respectively notated in the system equations as S, D, and M. Thus, x_t, the solutions of the continuous SCON system at time t, expand to the vector,

$$x_t = \begin{bmatrix} S_t \\ D_t \\ M_t \end{bmatrix} \tag{4.2}$$

and observations of the system y_t are similarly three-dimensional.

We established two SCON versions for inference and data generation use, SCON-C and SCON-SS.

SCON-C and SCON-SS share the same underlying α drift vector, equivalent to the deterministic CON dynamics and following the form:

$$\begin{bmatrix} \mathrm{d}S \\ \mathrm{d}D \\ \mathrm{d}M \end{bmatrix} = \begin{bmatrix} \mathcal{I}_S + a_{DS} \cdot k_D \cdot D + a_M \cdot a_{MSC} \cdot k_M \cdot M - k_S \cdot S \\ \mathcal{I}_D + a_{SD} \cdot k_S \cdot S + a_M \cdot (1 - a_{MSC}) \cdot k_M \cdot M - (u_M + k_D) \cdot D \\ u_M \cdot D - k_M \cdot M \end{bmatrix} \mathrm{d}t + \beta^{0.5} \begin{bmatrix} \mathrm{d}W_S \\ \mathrm{d}W_D \\ \mathrm{d}W_M \end{bmatrix} \tag{4.3}$$

where β now refers to the diffusion matrix component of the SDE and the W_S, W_D, and W_M elements of the Wiener process vector represent random draws from the distribution $\mathcal{N}(0, \sqrt{\mathrm{d}t})$.

For simplification purposes, the β diffusion matrices of both systems were made to be diagonal only and free of covariance diffusion terms. SCON-C diffusion dynamics are given by

$$\beta = \begin{bmatrix} c_S & 0 & 0 \\ 0 & c_D & 0 \\ 0 & 0 & c_M \end{bmatrix} \tag{4.4}$$

while SCON-SS diffusion dynamics are

$$\beta = \begin{bmatrix} s_S \cdot S & 0 & 0 \\ 0 & s_D \cdot D & 0 \\ 0 & 0 & s_M \cdot M \end{bmatrix} \tag{4.5}$$

Thus, SCON-C diffusion noise is *additive*, meaning it is independent of the values of states S, D, and M, and also *stationary*, meaning that is not a function of t. Meanwhile, SCON-SS noise is *multiplicative*, meaning it is dependent on the states. As such, SCON-C is linear in drift and diffusion, while SCON-SS is linear in drift but non-linear in diffusion.

$\mathcal{I}_S$ and $\mathcal{I}_D$ respectively represent the exogenous input of C mass in units of mg C g^{-1} soil h^{-1} into

the SOC and DOC soil pools from litter decay and can be modeled as constants or functions. We used sinusoidal litter input functions with annual periods that assumed litterfall peaking through late summer and early fall in a pattern resembling those observed in tropical forest ecosystems Giweta, 2020. The functions are given by

$$\mathcal{I}_{S,t} = 0.001 + 0.0005 \cdot \sin\left(\frac{2\pi}{365 \cdot 24} t\right) \tag{4.6a}$$

$$\mathcal{I}_{D,t} = 0.0001 + 0.00005 \cdot \sin\left(\frac{2\pi}{365 \cdot 24} t\right) \tag{4.6b}$$

As was previously instituted for CON (Allison et al., 2010; J. Li et al., 2014), the SCON linear first-order decay parameters $k_{i\in\{S,D,M\}}$ remain dependent on temperature. Temperature sensitivity of the $k_{i\in\{S,D,M\}}$ linear first-order decay parameters is enforced by a function derived from the original Arrhenius equation (Arrhenius, 1889),

$$k_{i,t} = k_{i,\,\mathrm{ref}} \exp\left[-\frac{Ea_{k_i}}{R}\left(\frac{1}{\mathrm{temp}_t} - \frac{1}{\mathrm{temp}_{\mathrm{ref}}}\right)\right] \tag{4.7}$$

where R is the ideal gas constant 8.314 J K^{-1} mol^{-1} and $\mathrm{temp}_{\mathrm{ref}}$ specifies a "reference" equilibrium temperature which we set at 283 K.

Through changing values of $k_{i\in\{S,D,M\}}$, SCON systems respond to day-night and seasonal temperature cycles through the composite sinusoid forcing function,

$$\mathrm{temp}_t = \mathrm{temp}_{\mathrm{ref}} + \frac{5t}{80 \cdot 365 \cdot 24} + 10 \cdot \sin\left(\frac{2\pi}{24} t\right) + 10 \cdot \sin\left(\frac{2\pi}{365 \cdot 24} t\right) \tag{4.8}$$

The function assumes a gradual linear increase in mean soil surface temperature by 5 °C over 80 years from the start of the simulation, in line with the upper bound of mean surface temperature increases predicted in emissions scenarios by 2100 (O'Neill et al., 2017).

SDE systems rarely admit tractable analytic solutions. To sample state trajectories accurately

approximating SCON-C and SCON-SS dynamics and construct our synthetic time series data y, we used the long-established and reliable Euler-Maruyama SDE solver (Maruyama, 1955) to numerically integrate solution paths x corresponding to θ randomly sampled from logit-normal distributions. Our solver step size was set to $\mathrm{d}t = 0.1$ hour. We note that we recover the exact SCON-C and SCON-SS processes in continuous time as $\mathrm{d}t$ is decreased to 0.

If inference involved conditioning with CO_2 observations in y in addition to state measurements, model CO_2 respiration rate would be computed from the time-corresponding x state values with the equation

$$\mathrm{CO}_{2,\,t} = (1 - a_{SD}) \cdot k_{S,\,t} \cdot S_t + (1 - a_{DS}) \cdot k_{D,\,t} \cdot D_t + (1 - a_M) \cdot k_{M,\,t} \cdot M_t \tag{4.9}$$

where $\mathrm{CO}_{2,\,t}$ is in units of $\mu g\, g^{-1}\, soil\, h^{-1}$. We then sliced x and CO_2 time series at some regular interval, i.e. every 1 hour or 5 hours, and normally sampled about the sliced values with an observation error vector σ_{obs} in the manner of

$$y_t \sim \mathscr{N}(x_t, \eta_{\mathrm{obs}}) \tag{4.10}$$

to arrive at y. We lower bounded y such that $y \in \mathbb{R}_{\geq 0}$ to preclude nonsense negative state measurements. We used constant η_{obs} that was 10% of the overall state mean such that

$$\eta_{\mathrm{obs}} = 0.1 \odot \begin{bmatrix} \bar{S} \\ \bar{D} \\ \bar{M} \end{bmatrix} \tag{4.11}$$

where $\odot$ indicates elementwise multiplication. This corresponds to an empirical scenario where measurement instruments and processes introduce a stable level of observation noise. CO_2 was similarly observed with noise standard deviation that was 10% of the overall CO_2 mean across the total time span in the sampling of y including CO_2.

θ	Biogeochemical interpretation	Target hyperparameters	Units
u_M	MBC uptake rate	$\mathscr{LN}(0.0016, 0.0004, 0, 1)$	mg C g^{-1} C h^{-1}
a_{DS}	DOC to SOC transfer fraction	$\mathscr{LN}(0.5, 0.125, 0, 1)$	NA
a_{SD}	SOC to DOC transfer fraction	$\mathscr{LN}(0.5, 0.125, 0, 1)$	NA
a_M	MBC to organic C transfer fraction	$\mathscr{LN}(0.5, 0.125, 0, 1)$	NA
a_{MSC}	MBC to SOC transfer fraction	$\mathscr{LN}(0.5, 0.125, 0, 1)$	NA
$k_{S,\mathrm{ref}}$	SOC decomposition rate	$\mathscr{LN}(0.0005, 0.000125, 0, 0.1)$	mg C mg^{-1} C h^{-1}
$k_{D,\mathrm{ref}}$	DOC decomposition rate	$\mathscr{LN}(0.0008, 0.0002, 0, 0.1)$	mg C mg^{-1} C h^{-1}
$k_{M,\mathrm{ref}}$	MBC decomposition rate	$\mathscr{LN}(0.0007, 0.000175, 0, 0.1)$	mg C mg^{-1} C h^{-1}
Ea_S	SOC decomposition activation energy	$\mathscr{LN}(20, 5, 5, 80)$	kJ mol^{-1}
Ea_D	DOC decomposition activation energy	$\mathscr{LN}(20, 5, 5, 80)$	kJ mol^{-1}
Ea_M	MBC decomposition activation energy	$\mathscr{LN}(20, 5, 5, 80)$	kJ mol^{-1}
c_S	SCON-C SOC β constant	$\mathscr{LN}(0.1, 0.025, 0, 0.1)$	mg C g^{-1} soil
c_D	SCON-C DOC β constant	$\mathscr{LN}(0.002, 0.0005, 0, 0.1)$	mg C g^{-1} soil
c_M	SCON-C MBC β constant	$\mathscr{LN}(0.002, 0.0005, 0, 0.1)$	mg C g^{-1} soil
s_S	SCON-SS SOC β factor	$\mathscr{LN}(0.0005, 0.000125, 0, 0.1)$	NA
s_D	SCON-SS DOC β factor	$\mathscr{LN}(0.0005, 0.000125, 0, 0.1)$	NA
s_M	SCON-SS MBC β factor	$\mathscr{LN}(0.0005, 0.000125, 0, 0.1)$	NA

Table 4.1: List of SCON-C and SCON-SS θ and their corresponding marginal data-generating and informed prior hyperparameters. The marginal densities are formatted as $\mathscr{LN}(\mu, \sigma, a, b)$, where μ and σ are the desired target density mean and standard deviation and a and b are the truncated distribution support lower and upper bounds.

We generated and conditioned inferences on synthetic y of up to 5000 hours in total timespan T. Data-generating θ distribution hyperparameters were chosen to produce stable and informative state dynamics in a shorter span of time and minimize the memory footprint of the data set under available computing resources. We used elevated $k_{i,\mathrm{ref}}$ means compared to previous literature values (Allison et al., 2010; J. Li et al., 2014; Xie et al., 2020). Sampled θ values and T scale are thereby reminiscent of an organic decay process occurring at the soil surface, rather than a slower subterranean decomposition. θ data-generating distribution hyperparameters, equivalent to the prior distribution $p(\theta)$ hyperparameters, along with the biogeochemical interpretations associated with each θ, are detailed in Table 4.1.

4.2.3 The generalized univariate logit-normal distribution

We used a univariate logit-normal distribution family for our data-generating, informed prior $p(\theta)$, and mean-field variational posterior $q(\theta|y)$ probability density functions. To avoid being restricted to the standard $[0, 1]$ distribution support that the logit-normal density is typically associated with in statistics, we defined a generalized form of the family whose supports could be enclosed between an arbitrary positive $[a, b]$, where $a, b \in \mathbb{R}_{\geq 0}$ and $b > a$. Generalized logit-normal distributions provide similar utility to truncated normal distributions used previously in SBM inference projects for constraining θ values to finite supports (Xie et al., 2020), but are more stable for backpropagation, as the inverse cumulative distribution function of the truncated normal distribution has inherent stability issues close to support boundaries.

We notate logit-normal distribution parameters in order of desired "target" mean μ, standard deviation σ, support lower bound a, and upper bound b akin to

$$\theta \sim \mathscr{LN}(\mu, \sigma, a, b) \tag{4.12}$$

Via passage through a sigmoid function, logit-normal distributions are transformed from normal distributions $\mathscr{N}(\check{\mu}, \check{\sigma})$, where $\check{\mu}$ and $\check{\sigma}$ are respectively the "parent" mean and standard deviation distribution parameters:

$$\check{\theta} \sim \mathscr{N}(\check{\mu}, \check{\sigma}) \tag{4.13}$$

$$\theta^{\text{mid}} = \frac{1}{1 + \exp(-\check{\theta})} \tag{4.14}$$

$$\theta = (b - a) \cdot \theta^{\text{mid}} + a \tag{4.15}$$

The logit-normal distribution has no closed form probability density function and its probability moments are not analytically resolvable, so no formula can be deduced that allows us to make

random variable transformations between logit-normal and normal distributions. Hence, to arrive at a particular logit-normal density with target μ and σ in each VI optimization iteration to sample from, we must first numerically solve for the parent $\check{\mu}$ and $\check{\sigma}$ of a normal distribution that corresponds to the desired logit-normal properties following the transformations from equations (4.13) to (4.15). We can do this with root-finding algorithms like the bisection method that search for an appropriate $\check{\mu}$ in the truncated support interval between a and b and $\check{\sigma}$ within a provided range of tolerated standard deviation values (Daunizeau, 2017).

4.2.4 State space model approximation of the SDE

Instead of optimizing SCON θ via an iterative SDE solver, we optimized time-discretized state space models approximating the SCON-C and SCON-SS SDEs. State space models describe the distribution of a discrete sequence of observations y sourced from discrete latent states x. They can take the general form

$$x_t \sim p(x_t|x_{t-1}, \theta) \tag{4.16}$$

$$y_t \sim p(y_t|x_t, \theta) \tag{4.17}$$

Equation (4.16) indicates that the transition from x_{t-1} to x_t occurs at a probability density of $p(x_t|x_{t-1}, \theta)$ and that subsequent states of a state space model depend on previous ones, thus indicating that x constitutes a Markov chain. Equation (4.17) specifies that y_t is observed from x_t at a density of $p(y_t|x_t, \theta)$. An initial state x_0 must be nominated to compute x and it can be set as a constant, or informed as a density, $p(x_0)$, which we do in our case.

The state space model θ are the same model parameters as in the SDE counterpart. When accounting for the SDE α drift and β diffusion dynamics, x_t, the latent states of the state space

model at time t can be written as

$$x_t = x_{t-1} + \alpha(x_{t-1}, \theta)\Delta t + \epsilon_t \sqrt{\beta(x_{t-1}, \theta)\Delta t} \tag{4.18}$$

with the same α and β as in (4.1). ϵ_t is a random noise vector of length $\mathcal{D}$ independently sampled via $\epsilon_t \sim \mathcal{N}(0, I_{\mathcal{D}})$. $I_{\mathcal{D}}$ is an identity matrix with number of diagonal elements equal to $\mathcal{D}$. Δt is the state space model time step, not to be confused with SDE solver time step $\mathrm{d}t$. We used $\Delta t = 1.0$ hour for our state space model approximations in contrast to the aforementioned $\mathrm{d}t = 0.1$ for Euler-Maruyama solving of our data generating processes.

There is overlap between SDEs and state space models. Both depict the evolution of state values in a series of steps where future values depend on past ones. Both require the specification of initial conditions x_0 to compute state trajectories.

However, SDEs and state space models treat time differently. A key distinction that makes state space model approximation helpful for inference efficiency is that Δt can be made relatively large versus SDE solver $\mathrm{d}t$. This is helpful for common biological inference and data assimilation situations where y is sparsely observed due to the expense and difficulty of collecting measurements.

Differential equation systems are instead typically numerically integrated and like state space models, are solved in discrete steps, as only smooth analytic solutions can only be obtained from the simplest systems. But, the differential equation integration procedures still assume that states are evolving in continuous time. The integrating solvers almost always require relatively small integration time steps $\mathrm{d}t$ that are as close to 0 as possible; the solvers tend to fail at higher $\mathrm{d}t$.

The divergent handling of time in state space models and SDEs renders them more apt for different objectives. State space models are better suited for estimating observations over long T, whereas SDEs are required for precise analyses of accurately simulated system dynamics. With their differing but related purposes, we can ultimately use state space models to represent discrete

observations from continuous SDEs.

4.2.5 VI optimization

Under a Bayesian statistics framework, the goal of statistical inference broadly consists of estimation of the θ posterior density function for some model, $p(\theta|y)$, conditioned on some data set y via Bayes' rule,

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \tag{4.19}$$

$p(y|\theta)$, also notated as $\ell(\theta|y)$, is the likelihood function, which indicates model goodness-of-fit across various values of individual parameters comprising θ. $p(\theta)$ is the prior probability representing beliefs about θ uncertainty. $p(y)$ is the probability density of the observed data.

The prior density $p(\theta)$ can be specified in an informed fashion, as we did in our workflow with distributions that matched our known data-generating distributions, or with less certainty in the absence of empirical information or domain experience. In most cases, $p(y|\theta)$ is not obtainable in closed analytic form and has to be numerically estimated with methods including Monte Carlo sampling and Laplace approximation (Reid, 2015). Additionally, $p(y)$, sometimes known as the *marginal evidence*, tends to be unresolvable (Gelman et al., 2013; McElreath, 2020). Thus, we take advantage of the proportionality relationship based on (4.19),

$$p(\theta|y) \propto p(y|\theta)p(\theta) = p(y, \theta) \tag{4.20}$$

to estimate $p(\theta|y)$ and practically conduct inference.

For Bayesian inference on state space models, we additionally need to account for the transition and observation densities generalized in equations (4.16) and (4.17), which influence the θ posterior. Estimation of the posterior of θ in state space model inference must occur along with

estimation of the posterior of x, whether in a joint or marginal fashion, in a case such as ours where the transition and observation distributions are not known. We opted for joint estimation. The joint posterior density of θ and x is notated as $p(\theta, x|y)$. We arrive at an expression for $p(\theta, x|y)$ by substituting (4.16) and (4.17) into (4.20):

$$p(\theta, x|y) \propto p(y, \theta, x) \tag{4.21}$$

$$= p(y|\theta, x)p(\theta, x) \tag{4.22}$$

$$= p(y|\theta, x)p(\theta)p(x|\theta) \tag{4.23}$$

$$= p(\theta) \prod_{i \in \mathfrak{N}} p(y_i|x_i, \theta) \prod_{t=1}^{T} p(x_t|x_{t-1}, \theta) \tag{4.24}$$

T denotes the final discretized integer time index of x. Since we set state space model $\Delta t = 1.0$ hour, our final time index matches total synthetic experiment hours and T can signify both. We also use T to represent the set of x state space model discretization indices not including the initial state at $t = 0$. We can then adopt a $\mathfrak{T} \subseteq T$ to indicate the set of y observation time indices not including an initial observation at $t = 0$, which is required in our VI procedure. The total number of x discretizations is $N = T + 1$ when including the $t = 0$ index. $\mathfrak{N} = \{0\} \cup \mathfrak{T}$ notates the full set of y indices.

In stochastic VI on state space models, we optimize a parametric density $q(\theta, x)$ to match the true joint posterior $p(\theta, x|y)$ as closely as possible. Per the probability chain rule, $q(\theta, x)$ expands to,

$$q(\theta, x) = q(x|\theta)q(\theta) \tag{4.25}$$

The density functions we select for our marginalized $q(\theta)$ and $q(x|\theta)$ are known as our *variational families*. As mentioned in section 4.2.3, we chose a mean-field logit-normal variational family for

$q(\theta)$ that assumed independent univariate distributions per θ such that

$$q(\theta) = q(\theta_1, \theta_2, \ldots, \theta_{\mathcal{P}}) = \prod_{j=1}^{\mathcal{P}} q_j(\theta_j) \tag{4.26}$$

where $\mathcal{P}$ is the total number of individual SBM θ and $\mathcal{P} = 14$ for SCON-C and SCON-SS (Table S1). We chose a class of normalizing flow called a *neural moving average flow* in Ryder et al. (2021) for our $q(x|\theta)$ variational family, which we will describe subsequently in section 4.2.6.

We can index individual representatives of our joint variational family by the properties and hyperparameters of the distribution symbolized in aggregate by $\phi_{(\theta,x)}$ such that an instance of $q(\theta, x)$ is notated as $q(\theta, x; \phi_{(\theta,x)})$. $q(\theta, x; \phi_{(\theta,x)})$ can be decomposed into $q(\theta; \phi_\theta)q(x|\theta; \phi_x)$ since $\phi_{(\theta,x)} = (\phi_\theta, \phi_x)$. ϕ are termed *variational parameters*, as they represent the distribution settings that can be varied and tuned to adjust the approximation. For neural network models like flows, variational parameters would include the hidden neural network parameters and weights. A particular distribution can be referred to by its variational parameter index in shorthand.

Our VI framework seeks a set of variational parameters ϕ that minimizes discrepancies between $q(\theta, x; \phi_{(\theta,x)})$ and $p(\theta, x|y)$, the approximate and true posteriors. One measure of distance between distributions customarily employed in statistics and machine learning literature is called the *Kullback-Leibler (KL) divergence*, notated as $D_{\mathrm{KL}}[q(\theta, x; \phi_{(\theta,x)})||p(\theta, x|y)]$ (Joyce, 2011; Kullback and Leibler, 1951; Perez-Cruz, 2008). Targeting the KL divergence for minimization, our optimization objective can then be mathematically stated as,

$$q(\theta, x; \phi^*_{(\theta,x)}) = \mathrm{argmin}_{q(\theta,x;\phi_{(\theta,x)})}(D_{\mathrm{KL}}[q(\theta, x; \phi_{(\theta,x)})||p(\theta, x|y)]) \tag{4.27}$$

where $\phi^*_{(\theta,x)}$ indexes the set of variational parameters that corresponds to the idealized global KL divergence minimum. After several omitted steps that can be referenced in greater detail from

Blei et al. (2017), we proceed from (4.27) to

$$D_{\mathrm{KL}}[q(\theta, x; \phi_{(\theta,x)})||p(\theta, x|y)] = \mathbb{E}_{q(\theta,x;\phi_{(\theta,x)})}[\log q(\theta, x; \phi_{(\theta,x)})] - \mathbb{E}_{q(\theta,x;\phi_{(\theta,x)})}[\log p(y|\theta, x)] \tag{4.28}$$

$$= \mathbb{E}_{q(\theta,x;\phi_{(\theta,x)})}[\log q(\theta, x; \phi_{(\theta,x)})] - \mathbb{E}_{q(\theta,x;\phi_{(\theta,x)})}[\log p(y, \theta, x)] + \log p(y) \tag{4.29}$$

where the expectations $\mathbb{E}$ subscripted with $q(\theta, x; \phi_{(\theta,x)})$ are taken with respect to the density $q(\theta, x; \phi_{(\theta,x)})$.

Reviewing the notion that $p(y)$ and in turn, the log marginal evidence, are typically unavailable (R. Christensen et al., 2010), we then rely on a reduced and rearranged expression that constitutes the ELBO function, notated as $\mathcal{L}$,

$$\mathcal{L}[\phi_{(\theta,x)}] = \mathbb{E}_{q(\theta,x;\phi_{(\theta,x)})}[\log p(y, \theta, x)] - \mathbb{E}_{q(\theta,x;\phi_{(\theta,x)})}[\log q(\theta, x; \phi_{(\theta,x)})] \tag{4.30}$$

$$= \mathbb{E}_{q(\theta,x;\phi_{(\theta,x)})}[\log p(y, \theta, x) - \log q(\theta, x; \phi_{(\theta,x)})] \tag{4.31}$$

$$= \mathbb{E}_{q(\theta,x;\phi_{(\theta,x)})} \langle \log p(y, \theta, x) - \log[q(x|\theta; \phi_x)q(\theta; \phi_\theta)]\rangle \tag{4.32}$$

$$= \mathbb{E}_{q(\theta,x;\phi_{(\theta,x)})} \langle \log p(y, \theta, x) - \log q(x|\theta; \phi_x) - \log q(\theta; \phi_\theta)\rangle \tag{4.33}$$

Substituting in (4.24) for $p(y, \theta, x)$ results in

$$\mathcal{L}[\phi_{(\theta,x)}] = \mathbb{E}_{q(\theta,x;\phi_{(\theta,x)})} \langle \log \left[p(\theta) \prod_{i \in \mathfrak{N}} p(y_i|x_i, \theta) \prod_{t=1}^{T} p(x_t|x_{t-1}, \theta) \right] - \log q(x|\theta; \phi_x) - \log q(\theta; \phi_\theta)\rangle \tag{4.34}$$

which, recalling that the set of total y indices $\mathfrak{N} = \{0\} \cup \mathfrak{T}$, expands into

$$\mathcal{L}[\phi_{(\theta,x)}] = \mathbb{E}_{q(\theta,x;\phi_{(\theta,x)})} \langle \log p(\theta) + \log p(y_0|x_0, \theta) + \sum_{i \in \mathfrak{T}} \log p(y_i|x_i, \theta) + \sum_{t=1}^{T} \log p(x_t|x_{t-1}, \theta) - \log q(x|\theta; \phi_x) - \log q(\theta; \phi_\theta)\rangle \tag{4.35}$$

We will decompose the marginal variational log-density of x, $\log q(x|\theta;\phi_x)$, in more granular detail when we describe the architecture of the neural moving average flow in section 4.2.6.

The ELBO function is called as such because it is the lower bound of the log marginal evidence:

$$\log p(y) = \mathcal{L}[\phi_{(\theta,x)}] + D_{\mathrm{KL}}[q(\theta,x;\phi_{(\theta,x)})||p(\theta,x|y)] \tag{4.36}$$

$$\geq \mathcal{L}[\phi_{(\theta,x)}] \tag{4.37}$$

Maximizing $\mathcal{L}[\phi_{(\theta,x)}]$, or minimizing the negative ELBO $-\mathcal{L}$, as we need to do in machine learning libraries like PyTorch that implement gradient descent rather than ascent, is commensurate to minimizing $D_{\mathrm{KL}}[q(\theta,x;\phi_{(\theta,x)})||p(\theta,x|y)]$. Hence, $\mathcal{L}[\phi_{(\theta,x)}]$ is our optimization objective function.

For pithier description of the ELBO gradient, $\nabla\mathcal{L}$, used to update $\phi_{(\theta,x)}$ via automatic differentiation, we set $\log p(y,\theta,x) - \log q(\theta,x;\phi_{(\theta,x)})$ in (4.31) equal to $\mathscr{R}(\theta,x,y,\phi)$, where $\mathscr{R}$ is a log-density ratio function. This reduces the ELBO equation to

$$\mathcal{L}[\phi_{(\theta,x)}] = \mathbb{E}_{q(\theta,x;\phi_{(\theta,x)})}\left[\mathscr{R}(\theta,x,y,\phi_{(\theta,x)})\right] \tag{4.38}$$

and the ELBO gradient is

$$\nabla\mathcal{L}[\phi_{(\theta,x)}] = \nabla_\phi\left\langle \mathbb{E}_{q(\theta,x;\phi_{(\theta,x)})}\left[\mathscr{R}(\theta,x,y,\phi_{(\theta,x)})\right]\right\rangle \tag{4.39}$$

$$= \nabla_\phi\left[\int_\theta\int_x q(\theta,x;\phi_{(\theta,x)})\mathscr{R}(\theta,x,y,\phi_{(\theta,x)})\mathrm{d}x\mathrm{d}\theta\right] \tag{4.40}$$

$$= \int_\theta\int_x \nabla_\phi\left[q(\theta,x;\phi_{(\theta,x)})\mathscr{R}(\theta,x,y,\phi_{(\theta,x)})\right]\mathrm{d}x\mathrm{d}\theta \tag{4.41}$$

which decomposes to

$$\nabla\mathcal{L}[\phi_{(\theta,x)})] = \int_\theta\int_x q(\theta,x;\phi_{(\theta,x)})\nabla_\phi\left[\mathscr{R}(\theta,x,y,\phi_{(\theta,x)})\right]\mathrm{d}x\mathrm{d}\theta + \int_\theta\int_x \mathscr{R}(\theta,x,y,\phi_{(\theta,x)})\nabla_\phi\left[q(\theta,x;\phi_{(\theta,x)})\right]\mathrm{d}x\mathrm{d}\theta \tag{4.42}$$

Note that the gradients ∇_ϕ are taken with respect to the variational parameters. This presents a complication, as examining the second term of (4.42), we are left with the situation that $\nabla_\phi \left[q(\theta, x; \phi_{(\theta,x)}) \right]$ is by and large unavailable, as q can be sampled from, but is usually not analytically differentiable. Our joint variational family q is no exception to that pattern; our marginal mean-field $q(\theta; \phi_\theta)$ has the straightforward analytic form given in (4.26), but use of the neural moving average flow as the variational family for $q(x|\theta; \phi_x)$ precludes the overall joint density $q(\theta, x; \phi_{(\theta,x)})$ from having an orderly closed form.

To ultimately compute the gradient of an expectation as in (4.39) in numerical fashion, we thereby turn to the *reparameterization trick* set forth in Salimans and Knowles (2013) and Kingma and Welling (2014). The reparameterization trick involves setting (θ, x) as an output of an invertible, deterministic, and differentiable function $g(z, \phi_{(\theta,x)})$, where z is a random vector sampled from some fixed density $q(z)$. This enables us to tractably take a gradient of a simpler fixed distribution whose probability density function is easier to differentiate and not dependent on the variational parameters ϕ (Ruiz et al., 2016).

In our case, z elements are sampled from standard normal distributions and undergo invertible transformations to proceed to x. θ is still directly sampled from its mean-field logit-normal family described in section 4.2.5 as part of the operations of g. Hence, $\mathcal{L}$ can be estimated with each VI training iteration with Monte Carlo sampling of z and θ entries starting with the steps

$$z^{(s)} \sim \mathcal{N}(0, I_N) \tag{4.43}$$

$$\theta^{(s)}, x^{(s)} = g(z^{(s)}, \phi_{(\theta,x)}) \tag{4.44}$$

where I_N is an identity matrix with number of diagonal entries matching the total x discretization indices N. The superscript (s) indexes an individual draw from a distribution. We can then reframe (4.39) from an analytically intractable gradient of an expectation to a numerically assessable

expectation of a gradient with

$$\nabla\mathcal{L}[\phi_{(\theta,x)}] = \nabla_\phi \left\langle \mathbb{E}_{q(z)} \left[\mathscr{R}(\theta, x, y, \phi_{(\theta,x)}) \right] \right\rangle \tag{4.45}$$

$$= \mathbb{E}_{q(z)} \left[\nabla_\phi \left\langle \mathscr{R}(\theta, x, y, \phi_{(\theta,x)}) \right\rangle \right] \tag{4.46}$$

$$\approx \frac{1}{\mathcal{S}} \sum_{s=1}^{\mathcal{S}} \nabla_\phi \left\langle \mathscr{R}(\theta^{(s)}, x^{(s)}, y, \phi_{(\theta,x)}) \right\rangle \tag{4.47}$$

$$\widehat{\nabla\mathcal{L}[\phi_{(\theta,x)}]} = \frac{1}{\mathcal{S}} \sum_{s=1}^{\mathcal{S}} \nabla_\phi \left\langle \mathscr{R} \left[g(z^{(s)}, \phi_{(\theta,x)}), y, \phi_{(\theta,x)}) \right] \right\rangle \tag{4.48}$$

$\mathcal{S}$ denotes the total number of independent θ and z samples drawn per training iteration. $\widehat{\nabla\mathcal{L}[\phi_{(\theta,x)}]}$ specifies the Monte Carlo estimate of $\nabla\mathcal{L}[\phi_{(\theta,x)}]$.

4.2.6 Masked neural moving average flow architecture

Delineating a normalizing flow more formally than in section 4.1, a general flow is a chain of *bijections*, or invertible transformation functions mapping an object in a set one-to-one to an object in another set. Flows can be decomposed into

$$x = g(z) = g_M \circ g_{M-1} \circ \cdots \circ g_m \circ \cdots \circ g_1(z) \tag{4.49}$$

where $\circ$ notates function composition operations and M marks the total number of bijections. In the generative direction, our flow takes us from a random object z to a random object x corresponding to a set of approximated SCON state trajectories.

A generative flow is constructed such that computation of $\log q(x|\theta;\phi_x)$ in (4.35) is available to facilitate the optimization of $q(x|\theta;\phi_x)$. The log-density of x is available through the change of

variables formula:

$$\log q(x) = \sum_{t=1}^{T} \varphi(z_t) - \log |\det J| \tag{4.50}$$

$$\log q(x) = \sum_{t=1}^{T} \varphi(z_t) - \log \prod_{m=1}^{M} |\det J_m| \tag{4.51}$$

$$\log q(x) = \sum_{t=1}^{T} \varphi(z_t) - \sum_{m=1}^{M} \log |\det J_m| \tag{4.52}$$

where J is the Jacobian matrix of the overall transformation and J_m is the Jacobian of bijection g_m with respect to the intermediate transformed variable $g_{m-1} \circ g_{m-2} \circ \cdots \circ g_1(z)$. We use $\varphi(z_t)$ to indicate the log-density of each element of z, z_t. We establish that z here is equivalent to the z introduced in section 4.2.5, so each $\varphi(z_t)$ is then a unit standard normal log-density in our framework. We notate the density function of z with $q(z)$. Since $q(z)$ is the starting distribution before transformations are layered, it is also termed the *base distribution*.

The particular flow we implemented as the marginal variational family for $q(x|\theta)$ was patterned after the original neural moving average flow introduced in Ryder et al. (2021). Neural moving average flows include *affine* bijections (Dinh et al., 2015; Dinh et al., 2017; Kingma et al., 2016; Papamakarios et al., 2017) among the functions constituting g in which an x^{out} is transformed from an x^{in} in the general form of

$$x^{\text{out}} = \mu + \sigma \odot x^{\text{in}} \tag{4.53}$$

where $\odot$ represents elementwise multiplication to denote that μ, σ, and x^{in} can be matrices and vectors in addition to scalars, though our explicit situation involves scalars. μ and σ are respectively known as shift and scale values of the bijection and it is required that $\sigma \in \mathbb{R}_+$. Cumulative μ and σ values of a flow are usually implemented as trained outputs of a neural network and are super- and subscripted to identify the transformation layer and input elements they correspond to. They are notated as such by convention and not to be confused with the

similarly notated target logit-normal mean and standard deviation parameters in section 4.2.3.

These linear affine transformations are basic in structure and consequently are individually not so *expressive*, or able to flexibly transition a base distribution into substantially different distributions of varying complexity. However, when layered repeatedly and stacked, their cumulative expressivity increases and with sufficient layers, composite affine functions can come to embody any distribution that is log-concave and bookended by declining density tails (Lee et al., 2021), which represents a large swath of probability distributions.

Neural moving average flows are specifically distinguished from other flows containing affine layers through their execution of affine bijections with 1-dimensional convolutional neural networks (CNNs). To apply 1-dimensional CNNs rather than 2-dimensional CNNs, we note that for systems with $\mathcal{D} > 1$, like SCON family instances, we must begin with z in a 1-dimensional "melted" form that is $\mathcal{D} \cdot T$ elements in length before reshaping the final transformed x to a $\mathcal{D} \times T$ matrix matching the SDE solution structure demonstrated in (4.2) following the conclusion of g. Thus, in equations (4.43) and (4.52), we replace T with $\mathcal{D} \cdot T$ in our implementation.

Through *masking*, in which inputs to the convolution patch are zeroed out through multiplication by weights, the flow is imbued with an *autoregression* property in which the σ_i and μ_i values producing an ith element of an output vector x does not depend and convolve on any element $z_{j \geq i}$ in the base input vector. This autoregression is critical for the intent of arranging the computation of $\sum_{m=1}^{M} \log |\det J_m|$ in (4.52) to be manageable. The autoregression ensures that J is a diagonal matrix whose non-zero elements are the σ scale parameters underlying the overall transformation, which simplifies calculation of $\det J$ and $\log q(x)$ to

$$\log q(x) = \sum_{t=1}^{T} \varphi(z_t) - \sum_{m=1}^{M} \sum_{t=1}^{T} \log \sigma_t^m \tag{4.54}$$

where σ_t^m is the shift parameter of the bijection producing the tth term of the mth affine layer output of length T.

Figure 4.3b portrays a schematic of the autoregressive convolutions and affine bijections used in our specific neural moving average flow implementation. The operations occur within *residual blocks*, component pieces of deep learning networks consisting of organizations of layers oriented toward the mitigation of training and approximation error that can otherwise snowball with greater network depth. Residual blocks do this with the use of *skip connections*, which preserve and carry over output from previous layers to serve as input to subsequent layers and in doing so prevent noisy degradation of information cascading through the network (He et al., 2016).

In each residual block, we perform two masked 1-dimensional convolutions, Convolution A and Convolution B, that each have a kernel length of 3 elements and a stride length of 1. To enshrine autoregressiveness of the flow, Convolution A applies a kernel masked as $[1, 0, 0]$ that outputs a shift and scale value pair. The Convolution A operation and associated affine bijection can be generally expressed as

$$(\mu_i, \sigma_i) = f_i^{\mathrm{A}}(x_{i-1}^{\mathrm{in}}) \tag{4.55}$$

$$x_i^{\mathrm{out}} = \mu_i + \sigma_i \cdot x_i^{\mathrm{in}} \tag{4.56}$$

where μ_i, σ_i, x_i^{in}, and x_i^{out} are scalar elements of vectors and f_i^{A} is the Convolution A operation. The subsequent Convolution B involves a single stride kernel masked as $[1, 1, 0]$ and it can be expressed together with its associated bijection as

$$(\mu_i, \sigma_i) = f_i^{\mathrm{B}}(x_{i-1}^{\mathrm{in}}, x_i^{\mathrm{in}}) \tag{4.57}$$

$$x_i^{\mathrm{out}} = \mu_i + \sigma_i \cdot x_i^{\mathrm{in}} \tag{4.58}$$

Combined, the two convolutions in sequence have a total receptive field length of 2.

To be able to produce the μ and σ parameters associated with the affine transformation of vector endpoint elements under autoregressive alignment, both convolutions require *zero padding*, in which zero elements are added to either end of the vector. As can be gleaned from Figure 4.3b,

without zero padding, the kernels producing $(\mu_1^{\text{mid}}, \sigma_1^{\text{mid}})$ and $(\mu_1^{\text{out}}, \sigma_1^{\text{out}})$ would lack 1 element to convolve on, and the kernels sourcing $(\mu_N^{\text{mid}}, \sigma_N^{\text{mid}})$ and $(\mu_N^{\text{out}}, \sigma_N^{\text{out}})$ would by overhang their vectors by 1. A zero pad of length 1 was thereby sufficient for our purposes.

Simplified from our actual implementation and not pictured in Figure 4.3b is our expansion of input into many *channels*, which are duplicates of the input vector that are stacked on top of each other in a matrix. At each convolution stage, the same kernel is applied in parallel across all the channels. Enlarging channel depth broadens the space of neural network weight values constituting f_i^{A} and f_i^{B} that can be explored per training iteration. We set the number of channels at 96 for both convolutions and did not experiment further with channel depth. Also not pictured in Figure 4.3b, but implied in Figure 4.3a, is the injection of auxiliary features extracted from y and observation indices $\mathfrak{N}$ in the form of vectors stacked on top of the input channels to inform training of the neural network weights associated with the shift and scale values. Further elaboration on the incorporation of auxiliary information is available in the supplement of Ryder et al. (2021).

In the overall flow procedure, the convolutions and affine bijections in the affine residual block are linked with other transformations that we organize into repeatable sets of layers. The order of transformations for each layer set is outlined in Figure 4.3a. Preceding the affine blocks are *order-reversing permutations*, in which element order of a vector input is flipped such that a vector $[x_1^{\text{in}}, x_2^{\text{in}}, ..., x_N^{\text{in}}]$ becomes $[x_1^{\text{out}}, x_2^{\text{out}}, ..., x_N^{\text{out}}] = [x_N^{\text{in}}, x_{N-1}^{\text{in}}, ..., x_1^{\text{in}}]$. Order-reversing permutations are a method of extending the expressivity and stability of a flow by enabling more complex dependency structures while preserving flow autoregression (Papamakarios et al., 2021). We found that adding order reversals allowed us to modestly boost our ELBO learning rates. The permutations can be seamlessly interspersed between other transformations since their absolute Jacobian determinant is valued at 1, so they do not affect the computation of $\log q(x)$.

Differing from the neural moving average flow of Ryder et al. (2021), our flow follows affine blocks with *batch renormalization* transformations. Batch renormalization is a simple extension

of *batch normalization*, which is a means of normalizing and regularizing our variational samples such that our optimization is less influenced by random fluctuations in neural network weights and sample characteristics from one training iteration to the next (Ioffe and Szegedy, 2015). Similar in intent but not operation to permutations, batch normalization and renormalization are applied to bolster algorithm stability and flexibility with increasing layer depth. They empirically allow VI algorithms to tolerate higher learning rates (Bjorck et al., 2018), poor initialization of variational parameters ϕ (Zhu et al., 2020), and erratic base distribution $z^{(s)}$ draws.

Batch normalization and renormalization overlap in the following steps that compute a *batch mean* $\mu_{\mathcal{S}}$ and *batch standard deviation* $\sigma_{\mathcal{S}}$ from input x^{in} samples, not to be confused with the affine bijection and logit-normal μ and σ:

$$\mu_{\mathcal{S}} = \frac{1}{\mathcal{S}} \sum_{s=1}^{\mathcal{S}} x_s^{\text{in}} \tag{4.59}$$

$$\sigma_{\mathcal{S}} = \sqrt{\varepsilon + \frac{1}{\mathcal{S}} \sum_{s=1}^{\mathcal{S}} (x_s^{\text{in}} - \mu_{\mathcal{S}})^2} \tag{4.60}$$

where ε is a small constant added for stability. $\mu_{\mathcal{S}}$ and $\sigma_{\mathcal{S}}$ are involved in computation of the optimization objective—again, $\mathcal{L}[\phi_{(\theta,x)}]$ for our purposes—during the model training phase. They also update a lagging *running average* $\mu_{\mathcal{R}}$ and *running mean* $\sigma_{\mathcal{R}}$ that are less sensitive to change. $\mu_{\mathcal{R}}$ and $\sigma_{\mathcal{R}}$ are used after training of the model—the joint variational family $q(\theta, x; \phi_{(\theta,x)})$ in this setting—has been halted to estimate the objective metric at the testing stage.

In the testing phase, batch renormalization and normalization are equivalent in transforming input to output:

$$x^{\text{mid}} = \frac{x^{\text{in}} - \mu_{\mathcal{R}}}{\sigma_{\mathcal{R}}} \tag{4.61}$$

$$x^{\text{out}} = \gamma \cdot x^{\text{mid}} + \Upsilon \tag{4.62}$$

The collection of γ_t^m and Υ_t^m parameters in each flow layer set are learned neural network outputs. Batch renormalization diverges from batch normalization during training with the steps

$$r = \frac{\sigma_{\mathcal{S}}}{\sigma_{\mathcal{R}}} \tag{4.63}$$

$$d = \frac{\mu_{\mathcal{S}} - \mu_{\mathcal{R}}}{\sigma_{\mathcal{R}}} \tag{4.64}$$

$$x^{\text{mid}} = \frac{x^{\text{in}} - \mu_{\mathcal{S}}}{\sigma_{\mathcal{S}}} \cdot r + d \tag{4.65}$$

$$x^{\text{out}} = \gamma \cdot x^{\text{mid}} + \Upsilon \tag{4.66}$$

where r and d are variable correction factors. r and d are intended to limit the divergence between batch and running sample characteristics. r is clipped between the interval $[1/r_{\max}, r_{\max}]$, where $r_{\max}$ is gradually increased to 3 over the course of inference, and d is clipped between the interval $[-d_{\max}, d_{\max}]$, where $d_{\max}$ is gradually increased to 5. These intervals were established based on guidelines from previous empirical work (Ioffe, 2017). Batch normalization is a special case of batch renormalization where $r = 1$ and $d = 0$.

Batch renormalization's changes more tightly correlate the batch and running sample characteristics and have been documented to minimize discrepancy between train and test objectives (Ioffe, 2017). We observed this with our ELBO results, where consistent gaps remained between the train and test $\mathcal{L}[\phi_{(\theta,x)}]$ until we swapped batch normalization for renormalization. Batch renormalization also improves training on low batch sizes (Ioffe, 2017; Summers and Dinneen, 2020), and in our position where variational path samples were limited by GPU video memory constraints, renormalization was helpful for decreasing the total number of training iterations we needed for algorithm convergence.

With batch (re)normalization layers, $\log q(x)$ accrues log determinant Jacobian summation terms corresponding to those transformations and develops from (4.54) to become, in the training

phase,

$$\log q(x) = \sum_{t=1}^{T} \varphi(z_t) - \sum_{m=1}^{M} \sum_{t=1}^{T} \left[\log \sigma_t^m - \log r_t^m - \log \gamma_t^m + \log \sigma_{\mathcal{S},t}^m\right] \tag{4.67}$$

or in the testing phase,

$$\log q(x) = \sum_{t=1}^{T} \varphi(z_t) - \sum_{m=1}^{M} \sum_{t=1}^{T} \left[\log \sigma_t^m - \log \gamma_t^m + \log \sigma_{\mathcal{R},t}^m\right] \tag{4.68}$$

where we now take M to mark the total number of layer sets rather than layers as we did before in (4.54). This assumes that each layer set always includes 1 single affine block and 1 batch renormalization layer. Substituting (4.67) or (4.68) into (4.35) for $\log q(x|\theta;\phi_x)$ leads respectively to our fully decomposed train or test $\mathcal{L}[\phi_{(\theta,x)}]$ calculation unless an optional single softplus transformation is used to ensure constraint of flow output to $\mathbb{R}_{\geq 0}$. In that case, the resulting train $\log q(x)$ is

$$\log q(x) = \sum_{t=1}^{T} \varphi(z_t) - \sum_{m=1}^{M} \sum_{t=1}^{T} \left[\log \sigma_t^m - \log r_t^m - \log \gamma_t^m + \log \sigma_{\mathcal{S},t}^m\right] - \sum_{t=1}^{T} \log(-\mathrm{e}^{-x_t} + 1) \tag{4.69}$$

where x is our terminally transformed random variable following softplus constraint. Setting

$$\lambda_t = \varphi(z_t) - \log(-\mathrm{e}^{-x_t} + 1) - \sum_{m=1}^{M} \left(\log \sigma_t^m - \log r_t^m - \log \gamma_t^m + \log \sigma_{\mathcal{S},t}^m\right) \tag{4.70}$$

$$\log q(x) = \sum_{t=1}^{T} \lambda_t \tag{4.71}$$

our fully decomposed train $\mathcal{L}[\phi_{(\theta,x)}]$ calculation that we use in each iteration of VI optimization

(Algorithm 4.1) then consolidates from (4.35) into

$$\mathcal{L}[\phi_{(\theta,x)}] = \mathbb{E}_{q(\theta,x;\phi_{(\theta,x)})}\langle \log p(\theta) + \log p(y_0|x_0,\theta) - \log q(\theta;\phi_\theta) + \sum_{i\in\mathfrak{T}} \log p(y_i|x_i,\theta) + \sum_{t=1}^{T} [\log p(x_t|x_{t-1},\theta) - \lambda_t]\rangle \quad (4.72)$$

with softplus flow termination. The test ELBO equation is equivalent except for use of a different λ_t assignment that lacks the $\log r_t^m$ term and swaps $\sigma_{\mathcal{S},t}^m$ for $\sigma_{\mathcal{R},t}^m$.

Algorithm 4.1 Synopsis of the operations occuring in each iteration of our soil biogeochemical state space model VI framework

Define $q(\theta;\phi_\theta)$ and $q(x|\theta;\phi_x)$;
Initialize (ϕ_θ, ϕ_x);
$N \leftarrow$ total desired training iterations;
for $i \leftarrow 1$ to N **do**
  for $s \leftarrow 1$ to $\mathcal{S}$ **do**
    Draw $\theta^{(s)} \sim q(\theta;\phi_\theta)$;
    Draw $x^{(s)} \sim q(x|\theta;\phi_x)$ transformed from $z^{(s)}$;
  end for
  Compute $\mathcal{L}[\phi_{(\theta,x)}]$ (or $-\mathcal{L}[\phi_{(\theta,x)}]$ for gradient descent) as per (4.72);
  Compute the gradient $\nabla\widehat{\mathcal{L}[\phi_{(\theta,x)}]}$ from (4.48) with automatic differentiation;
  Update variational parameters $\phi_{(\theta,x)}$ based on the gradient;
end for
return $q(\theta, x; \phi_{(\theta,x)})$ corresponding to the $\mathcal{L}[\phi_{(\theta,x)}]$ value at N;

We note that it is not required for the total permutation layers, affine blocks, and batch renormalization layers constituting a neural moving average flow architecture to match in count; we can choose to omit certain layers in a layer set. To slightly reduce the neural network size, we would frequently use 1 less batch renormalization layer than total affine blocks or permutation layers, omitting batch renormalization in the first layer set since we empirically observed little qualitative difference in visual fit quality between running with 3, 4, or 5 batch renormalizations. If the numbers of affine blocks and batch renormalization layers do not match, then the log Jacobian determinant summations in (4.67) to (4.70) need to be adjusted accordingly.

It is apparent that each layer set of our neural moving average flow corresponds to a matrix of

hidden parameters, including affine and batch renormalization parameters, of dimensions $[T, h]$, where h is the count of hidden parameters per layer set. Thus, when conditioning on long, dense T data that is complex in such a manner that would require many layer sets for flow representation, we note that a different choice of marginal variational family for $q(x|\theta)$ aside from the neural moving average flow may be appropriate for minimizing computational expense.

4.2.7 Kalman smoother validation

When a state space model is linear in drift and its diffusion is stationary and additive, as is the state space model approximation of SCON-C, the posterior density $p(x|y)$ can be determined analytically and precisely in closed form with the Kalman smoother algorithm, provided the algorithm is fed the true θ and observation noise (Kalman, 1960; Rauch et al., 1965). Flow VI in contrast can only numerically estimate $p(x|y)$ through a variational approximation, but has the critical advantage of being capable of functioning without exact knowledge of θ given uninformed prior distributions and is able to estimate the joint density $p(x, \theta|y)$ via variational approximations. Thus, comparing a Kalman-derived true $p(x|y)$ to a post-optimization $q(x|\theta; \phi_x)$ can be a revealing means of benchmarking flow approximation performance and accuracy before applying an architecture with confidence to approximation, optimization, and θ inference of models like SCON-SS that cannot be resolved by the smoother.

The Kalman smoother procedure is a two part process consisting of a *forward pass* followed by a *backward pass*. The forward pass computes a "filtering" posterior $p(x_t|y_{0:t})$, which notates the posterior of x_t given observations up to the time indexed by t, going forward in time from $t = \{0, \ldots, T\}$. The backward pass computes a "smoothing" posterior $p(x_t|y)$, which notates the posterior of x_t given all observations, going backward in time from $t = \{T, \ldots, 0\}$. Reconciling the "filtering" and "smoothing" posteriors produces the true $p(x|y)$. A comprehensive explication of Kalman smoothing is available in Särkkä (2013).

4.2.8 Flow neural network training tuning choices

We settled on using 5 layer sets of permutation, affine, and batch renormalization layers for our neural moving average flow. This offered qualitatively superior fits over flow architectures with lower layer set counts. For inferences of duration $T = 5000$ with $\Delta t = 1.0$ with 5 layers, maximum training batch size $\mathcal{S}$ at 16 GB of VRAM was 31, so we set $\mathcal{S} = 31$. For $T = 1000$, we used $\mathcal{S} = 150$, though use of smaller $\mathcal{S}$ also appeared functional. For $T = 5000$ inferences we used 120000 non-warmup ELBO training iterations. For $T = 1000$ inferences we used 60000 non-warmup ELBO training iterations.

With respect to gradient optimizers including AdaMax (Kingma and Ba, 2015), which was the particular optimizer we selected to carry out gradient descent, the *learning rate* is a hyperparameter that scales the objective gradient and in doing so regulates the extent to which neural network weights can updated with each training iteration. The learning rate can be adjusted over the course of training based on a schedule. It is frequently decayed over the course of training to promote convergence of our objective function toward a maximum (for gradient ascent) or minimum (for gradient descent) (You et al., 2019). We chose a step decline schedule for learning rate decay. For our $T = 5000$ inferences, we started with a pre-decay ELBO learning rate of 1×10^{-2} and decayed it by a factor of 0.6 every 10000 iterations. For our $T = 1000$ inferences, we started with a pre-decay learning rate of 4×10^{-3} and decayed it by a factor of 0.6 every 5000 iterations.

We employed *training warmup*, in which we began optimization with a phase of low learning rate at 1×10^{-6} before increasing the rate to its initial pre-decay levels. As has been demonstrated previously (Goyal et al., 2017), we found warmup allowed us to use higher pre-decay learning rates, experience more stable ELBO loss trajectories, and converge to lower average ELBO values over training (Figure 4.4). We found 5000 warmup iterations to be sufficient for those purposes.

4.2.9 Software and hardware

With respect to the computational software and hardware powering the inference operations, our DGP and inference code was developed for a Python 3.9.7 environment distributed by Anaconda ("Anaconda Software Distribution", 2020) and used the Numpy 1.20.3 (Harris et al., 2020) and PyTorch 1.10.2 (Paszke et al., 2019) software libraries. PyTorch 1.10.2 was compiled with the Nvidia CUDA 10.2 toolkit. The inferences were run on one Nvidia Tesla V100 GPU at a time updated to CUDA version 11.4.0 with a maximum of 16 GB of video random access memory and two Intel Xeon Gold 6148 CPU cores clocked at 2.40 GHz on the University of California, Irvine HPC3 cluster. Our flow VI framework code modules, data-generating notebooks, and synthetic data are available via the address https://doi.org/10.5281/zenodo.6836049.

The deterministic CON $p(\theta|y)$ posteriors compared with flow VI $q(\theta;\phi_\theta)$ in Figure 4.7 were estimated using Stan's NUTS algorithm, which is an extension of the Hamiltonian Monte Carlo inference algorithm (Hoffman and Gelman, 2014). Application of Hamiltonian Monte Carlo for data assimilation and inference of SBMs is further described in Xie et al. (2020) and intuition behind the algorithm can be found in Betancourt (2017). The Stan inference was conducted on a 2017 Intel MacBook Pro in an R 4.0.4 environment using Stan 2.29.1 (Carpenter et al., 2017) through the CmdStanR interface (Gabry and Češnovar, 2021). The NUTS simulation ran with 2 chains of 1000 warmup iterations and 5000 sampling iterations each. In our experience, 1000 warmup iterations were sufficient for locating the bulk of the posterior density.

4.3 Results

We generated synthetic y of various lengths, dimensions (i.e. whether CO_2 observations were included in addition to state information), and regular observation densities (i.e. whether we observed measurements from our SCON family data generating processes every 1 or 5 hours).

We explored the validity of our state space model VI approach for data assimilation and posterior identification of model θ with inferences conditioned on those y. Below, our results suggest the neural moving average flow framework was functional for approximating the SCON family of SDE systems as state space models, fitting y, constraining posteriors, and recovering some true θ values. We also demonstrate subsequently that stochastic gradient optimization in our case was more stable, efficient, and capable at θ identification than an MCMC procedure involving deterministic ODE models adapted from Xie et al. (2020) conditioned on the same y.

4.3.1 Flow-approximated SCON-C converges to fit synthetic data

Following optimization, an SCON-C state space model approximated by our neural moving average flow implementation reasonably assimilated a $T = 5000$ hour y produced by an SCON-C data-generating process that included CO_2 observations (Figure 4.5a). The relatively flat $-\mathcal{L}[\phi_{(\theta,x)}]$ trajectory steadily hovering between -1550 and -1600 in the latter half of variational training iterations indicates that our flow VI algorithm converged to a local ELBO minimum (Figure 4.6).

The mean of the marginal posterior density of latent states $q(x|\theta;\phi_x)$ was estimated from 250 x samples drawn from the joint variational density after ELBO training. The mean latent SOC, DOC, and MBC paths and state-derived CO_2 measurements corresponding to the SCON-C flow sit centrally between the y data points and observation noise across the entire time series (Figure 4.5a). The latent means are able to adhere to many of the sharp peaks and valleys in the dynamics of the data and the flow CO_2 mean was able to reproduce the rapid oscillatory behavior of the observed CO_2 time series.

Upon closer qualitative inspection and comparison to the true latent distribution computed by a Kalman smoother (Figure 4.5b), we note the presence of visual discrepancies between the Kalman and flow means and 95% $q(x)$ diffusion distribution intervals. Firstly, the extent of SOC diffusion noise is substantially underestimated by the flow, which is line with documentation in

literature that a mean-field VI approach tends to underestimate posterior uncertainty compared to more complex full-rank approaches (Kucukelbir et al., 2017). For the other two states, DOC and MBC, the extent of diffusion noise is more consistent to that which is observed in the Kalman output, but the flow DOC and MBC densities and means appear noisier and more uneven than the Kalman means.

Still, the flow encouragingly is generally congruous with the true Kalman solution in dynamics. The flow means fall entirely within the bounds of the 95% Kalman diffusion interval from $t = 0$ to 500 as can be seen in Figure 4.5b and we observed for this particular optimization that they almost always remain within those Kalman diffusion bounds through the rest of the time series. Also, we see that the CO_2 mean and distribution calculated from the 250 SCON-C state space model x draws closely matches their Kalman counterparts. The ability of the flow to align with the Kalman smoother in latent state densities improves our confidence in the ability of the neural moving average flow to approximate systems that are non-linear in diffusion, like SCON-SS.

4.3.2 SCON-C flow VI marginal θ posteriors indicate appropriate optimization

Beyond fitting data, we needed to ascertain that proper posterior optimization was occurring for confidence in inference algorithm function. In our setting, we would expect our posterior densities to at least always be as informed and certain about θ values as our prior densities, not less. With a mean-field logit-normal variational family for $q(\theta; \phi_\theta)$, evidence of suitable optimization would come in the form of marginal posterior densities being narrower than priors to indicate greater certainty after the introduction of information from y along with posterior means separating from prior means and approaching the true θ used by the data-generating process.

Figure 4.7 indicates that valid posterior optimization indeed occurred in our SCON-C state space model inference to support the notion that our flow VI framework was functional. Almost all the

marginal posterior densities narrowed compared to the priors with the information learned from y by the algorithm. Moreover, many of the marginal $q(\theta;\phi_\theta)$ means drifted closer to the true θ, including the means of u_M, a_{SD}, and $k_{S,\mathrm{ref}}$.

We contrasted the flow VI parametric $q(\theta;\phi_\theta)$ posterior densities to the non-parametric $p(\theta|y)$ posterior densities estimated with an SBM inference framework conditioned on the same $T = 5000$ SCON-C y that was previously applied in Xie et al. (2020). This prior framework involves Stan's NUTS algorithm and can only infer θ of deterministic models, so the CON system that the SCON family was parameterized from served as the basis for inference in this approach. With the flexibility and stability afforded by the ability of stochastic optimization to adjust for poor initial condition proposals, noisy state path fluctuations, and outlier observations, the flow VI framework expectedly outperformed the deterministic NUTS workflow. The flow VI marginal $q(\theta;\phi_\theta)$ densities were all-around better informed and identified, exemplified by the subplots corresponding to the u_M, a_{SD}, a_M, $k_{S,\mathrm{ref}}$, $k_{D,\mathrm{ref}}$, Ea_S, and Ea_M θ (Figure 4.7). Moreover, some NUTS posterior densities, including those corresponding to the a_{MSC}, a_M, and Ea_S θ, consolidated near their lower or upper support bounds, which points to the deterministic model inference method compensating for its lack of versatility with more extreme θ proposals.

Scrutiny of the posterior for the transfer fraction parameter a_{MSC} brings the issue of θ identifiability limitations to our attention. We see that the SCON-C flow VI marginal a_{MSC} posterior density barely budged from the a_{MSC} $p(\theta)$ density post-optimization (Figure 4.7). For good posterior identifiability, the a_{MSC} posterior should both narrow substantially to signal reduced uncertainty and shift its density peak toward the true a_{MSC} value.

4.3.3 Flow VI can effectively assimilate both full and reduced SCON-SS state space approximations

After visually demonstrating the ability of the flow VI framework to optimize $q(\theta, x; \phi_{(\theta,x)})$ through the fitting of the approximated SCON-C state space model to synthetic SCON-C y and the informing and identification of some marginal $q(\theta; \phi_\theta)$ densities, we proceeded to test if the flow VI approach could similarly function with a moderately more complex model in SCON-SS that is non-linear in diffusion.

Reviewing the fact that the SCON-SS state space model diffusion is not stationary or additive, it was no longer possible for us to validate SCON-SS $q(x|\theta; \phi_x)$ estimated from post-optimization x samples against a true $p(x|y)$ determined by a Kalman smoother. Nonetheless, we observed that flow VI was able to optimize $q(\theta, x; \phi_{(\theta,x)})$ adequately enough to fit the approximated SCON-SS state space model $q(x|\theta; \phi_x)$ means to $T = 5000$ y generated by an SCON-SS SDE. As was the case for the SCON-C flow, the mean latent SOC, DOC, and MBC trajectories of the trained SCON-SS flow traced a central route through the observed state values and diffusion noise (Figure 4.8). The trajectories were able to follow the peaks and valleys of the state dynamics recorded in y, and the flow CO_2 mean derived from the sampled states tightly replicated the y CO_2 oscillations.

SCON-SS $q(\theta; \phi_\theta)$ posterior densities were consistent with proper optimization from information learned in y. Juxtaposed with priors $p(\theta)$, marginal $q(\theta; \phi_\theta)$ densities mostly narrowed and did not move drastically away from their corresponding true θ to inhibit identifiability (Figure 4.9). There were clear exceptions for the state-scaling diffusion θ posteriors due to reasonable flow neural network approximation error that prompts an overestimate of diffusion noise and, again, for the a_{MSC} posterior. The modest shift of some θ posterior means away from the true θ is counterbalanced by movement of other related θ, like in the circumstance of the Ea_D posterior mean being counterbalanced by Ea_M to satisfy equation (4.9).

For one more test to corroborate proper functioning of our flow VI framework, we established a reduced SCON-SS model with all θ fixed in value except for the $k_{i\in S,D,M,\,\mathrm{ref}}$ linear decay and state-scaling β diffusion parameters. Mirroring above procedures, a synthetic $T = 5000$ y was produced by a reduced SCON-SS data-generating process to condition an SCON-SS state space model optimization. For an appropriately behaving inference algorithm, we would expect that removing degrees of freedom should bolster θ identifiability.

We verified that identifiability was indeed clarified and improved in the remaining drift θ (Figure 4.10). The marginal $k_{S,\,\mathrm{ref}}$ $q(\theta;\phi_\theta)$ posterior density was tightly constrained right about the true $k_{S,\,\mathrm{ref}}$ value. The $k_{D,\,\mathrm{ref}}$ and $k_{M,\,\mathrm{ref}}$ posterior means did not align exactly with their true θ, but unambiguously offset each other in a manner that plainly fulfilled (4.3) and (4.9).

We were unable to fix the state-scaling diffusion θ without breaking the flow VI framework, as it became apparent that the algorithm needed to maintain the ability to overestimate diffusion noise to work. This makes intuitive sense as the flow neural network approximation process will always come with some amount of noisy approximation error that adds to the base diffusion of the unapproximated system. The algorithm can no longer work if the flow diffusion noise needs to be fixed at about the same level as it is in the unapproximated system, as it leaves no room for the approximation error to overflow into. So, with the diffusion θ left unfixed during VI training, the algorithm once more overestimated their $q(\theta;\phi_\theta)$ means, but this is not a cause for concern since the discrepancy can be explained by neural network approximation error.

4.3.4 Increasing information in y alters SCON posterior certainty and identifiability

The preceding results all involved y that had CO_2 respiration observations included. Inference conditioned on just state observations is also possible and in our experience was able to fit the data well, but it was much less effective for constraining posteriors and identifying θ (Figure

4.11). Without CO_2 information, marginal $q(\theta;\phi_\theta)$ posterior densities tended to be wider and less informed and density means were frequently farther away from true θ, as exemplified by panels corresponding to the a_{SD} and Ea_S SCON-C θ in Figure 4.11.

We separately observed that increasing the amount of information in y by lengthening duration T of the time series greatly benefitted posterior identifiability (Figure 4.12). Alternatively, θ identifiability was boosted without elongating T by bolstering observation density. Individually, the two actions trade off between improvements. In comparison to densifying observations across $T = 1000$ such that the set of observation indices $\mathfrak{N}$ matches the set of state space model discretization indices N, extending T to 5000 more tightly constrained $q(\theta;\phi_\theta)$ posterior densities for all θ and concentrated a_{SD}, $k_{S,\mathrm{ref}}$, $k_{D,\mathrm{ref}}$, Ea_S, and Ea_M posterior means closer to the true θ.

However, increasing observation density for $T = 1000$ data had the benefit of further constraining posterior densities without also enlarging the divergence in identification of the true SCON-C β diffusion θ, c_S, c_D, and c_M by the means of their corresponding $q(\theta;\phi_\theta)$ densities. The enlarged divergence and uncertainty of the diffusion θ in the $T = 5000$ hour inference compared to the $T = 1000$ inferences is not unexpected. Cumulative approximation error of state space x trajectories compounds for the flow with greater T in a manner typical to approximation methods. Larger accrued approximation error then corresponds to estimation of greater diffusion noise during inference.

4.4 Discussion

We developed a stochastic SBM data assimilation and inference framework that is a versatile, stable, and computationally efficient alternative to MCMC approaches assimilating deterministic ODE systems, especially when GPU hardware is available. The framework involves approximation

of SBMs as state space models whose state trajectories can be sampled at reduced computational and temporal cost in comparison to SDEs.

In our demonstration, we carried out state space model approximation with a class of normalizing flows called neural moving average flows that successively transition random variables from simpler to more complex distributions with the stacking of neural network layers. We applied this framework to fit approximated representatives of the SCON family of SBMs to synthetic data. Conditioning with synthetic rather than empirical data allowed us to visualize discrepancies between estimated posterior densities, data-generating densities, and true θ values used by the data-generating process for an assessment of framework performance.

Flow-approximated SCON-C state trajectories were able to effectively track state and CO_2 observations after variational optimization and graphically align with the true latent state distributions determined by a Kalman smoother. Following Kalman validation of our SCON-C inference, we then successfully assimilated synthetic observations and estimated posteriors with SCON-SS, which is non-linear in diffusion and modestly more complex than SCON-C.

4.4.1 More data promotes model θ identifiability and constraining of posteriors

In terms of implications for experimental work focused on producing data sets suitable for SBM inference and data assimilation, we firstly recommend that CO_2 respiration measurements be collected and included in y. CO_2 information is highly beneficial for informing the posteriors of SBMs like SCON for which CO_2 efflux equations have been established (Figure 4.11). Additionally, the collection of supplemental measurements, such as radiolabeled C densities linked to SBM pool transfer fraction θ, should further constrain and identify θ posteriors. Our results indicate that just the CO_2 and state observations were not enough to effectively identify the marginal posterior of the SCON MBC-to-SOC transfer θ, a_{MSC} (Figure 4.7, 4.9, 4.11, 4.12).

With respect to a_{MSC} posterior identifiability or lack thereof, inspection of the SCON system drift in (4.3) and CO_2 efflux rate equation in (4.9) suggests that the consistent lack of identifiability is not the consequence of a general algorithm issue but instead stems from a dearth of information in y to further constrain the marginal $a_{MSC}\ q(\theta;\phi_\theta)$ density. The a_{MSC} parameter appears in two terms of (4.3) that are each the product of four elements, the $a_M \cdot a_{MSC} \cdot k_M \cdot M$ term in the $\mathrm{d}S$ equation denoting C mass transfer from the MBC to SOC soil pool and the $a_M \cdot (1 - a_{MSC}) \cdot k_M \cdot M$ term in the $\mathrm{d}D$ equation denoting C transfer from the MBC to DOC soil pool. The posterior densities of the a_M and k_M θ in those terms appear in (4.9) and are accordingly better constrained and identified with CO_2 measurements in y. This is not the case for a_{MSC}, which is not present in (4.9). Informing of a_{MSC} can thereby only occur through the state measurements in y, and as only one element in the drift product terms, a_{MSC} can take many values between its $[0, 1]$ support bounds without greatly affecting the products.

Furthermore, our results suggests that raising both study time or data collection frequency would improve posterior estimation accuracy and identifiability in our framework (Figure 4.12). But, under budget and personnel limitations, empiricists creating inference data sets should prioritize one or the other depending on the specific SBMs targeted for data assimilation and their research objectives. For model comparison of naive stochastic SBM parameterizations where the conceived diffusion θ are less biologically meaningful, accumulating approximation error is less of a concern and prioritizing the maximization of T would be reasonable. In scenarios involving SBMs parameterized with more biogeochemically sophisticated β matrices where accurate estimation of system diffusion θ takes precedence and falsification of specific dynamics is the goal, it would be more important to minimize approximation error with denser observations.

4.4.2 Future work and research directions

Having demonstrated functional flow VI on more compact synthetic data sets, we highlight some engineering expansions and modifications to our existing framework that would facilitate efficient SBM inferences conditioned on empirical data sourced from long-term ecological (LTER) experiments like those documented in Melillo et al. (2017) and Wood et al. (2019). Efficiently scaling to these data sets is a key priority to assimilate them into SBMs on the scale of hours rather than weeks, as experienced in Xie et al. (2020), for statistically rigorous head-to-head model comparison and selection.

The T of data sets sourced from LTER experiments can be on the order of 100,000 to 200,000 hours, much larger than the peak $T = 5000$ hour timespan we explored in our study. With the ability of our framework to leverage GPU hardware, our $T = 5000$ inferences typically ran between one to two days to ensure convergence, but even more limiting than time were GPU memory thresholds preventing adequate variational sample sizes with T much longer than 5000. A way forward for conditioning inferences on y with longer T is to avoid simulating state space model x for the entire T at each training iteration, and this can be done in stochastic gradient optimization with the leveraging of the *mini-batching* technique. Under a mini-batching scheme, a y is partitioned into smaller subsections $y_{\tau i}$ during training, where τ merely distinguishes a subsection from the entire sequence and $i \in \mathcal{B}$ where $\mathcal{B} \in \mathbb{N}$ is the set of integers counting total subsections in natural order. In each training iteration, a $y_{\tau i}$ can then be randomly selected for likelihood evaluation such that the SBM only needs to simulate an $x_{\tau i}$ subsection for calculation of the optimization objective. Mini-batching is targeted for future incorporation in our framework, having been demonstrated in recent flow-related machine learning literature including Papamakarios et al. (2021) and Ryder et al. (2021).

LTER data sets tend to have constituent observation vectors whose elements greatly vary in information density and measurement intervals due in part to the varying physical practicality

associated with the sampling and measurement of different observations. Hence, it would also be helpful to engineer our flow to handle irregular and "ragged" observations. Moreover, alterations can be made to the flow network architecture to enable more efficient conditioning on SBM θ values and allow for feature extraction from additional auxiliary information, such as the time elapsed between observations.

Beyond flow engineering and architecture, other relevant research priorities include the study of less naive SCON treatments for inference use. SCON family representatives that explicitly and mechanistically model system diffusion as a function of the underlying system reaction stoichiometry can be formulated (Fuchs, 2013; Golightly and Wilkinson, 2011) and the stability and predictive accuracy associated with different diffusion covariance structures can be compared. Moreover, stochastic parameterizations of SBMs that simulate mass transfer with Michaelis-Menten dynamics and would be non-linear in drift should be investigated so that their predictive accuracy can be compared to those of linear drift models under our VI framework. This would go toward an existing priority in biogeochemistry to examine whether explicit representation of enzyme catalysis in SBMs improves model performance (J. Li et al., 2014; J. Li et al., 2019; Sulman et al., 2014; Wieder, Grandy, Kallenbach, Taylor, et al., 2015; Xie et al., 2020).

Application of our VI approach to head-to-head model comparison and selection begets a need for incorporation of goodness-of-fit quantification into our framework. MCMC has access to metrics like the widely application information criterion (Vehtari et al., 2017), leave-one-out cross-validation, and leave-future-out cross-validation (Bürkner et al., 2020) for Bayesian predictive accuracy quantification, but with their established formulations, these metrics cannot be computed under a VI procedure without prohibitive computational expense (Dao et al., 2022). The development of Bayesian goodness-of-fit metrics for VI is still an open area (Giordano et al., 2018; Yao et al., 2018), but there has been recent work adapting cross-validation for VI that is promising for integration with a state space model inference pipeline (Dao et al., 2022; Dhaka et al., 2020).

4.4.3 Conclusion

Going forward, we recommend that inference approaches involving state space model approximation of stochastic SBMs be used in future biogeochemical data assimilation, fitting, and model comparison research in pursuit of superior computational stability, flexibility, and efficiency. SDE systems are far more robust than ODE systems at accommodating prior density, initial condition, and model structure proposals that are inconsistent with the true data generating process (Whitaker, 2016; Wiqvist et al., 2021). Then, state space approximation greatly reduces the burden of sampling SDE model state trajectories for likelihood evaluation. Rather than integrating an SDE solver $\mathcal{S}$ times at great computational cost with each algorithm training iteration, we can efficiently sample $\mathcal{S}$ paths from the variational approximation in one pass. Additionally, the discrete nature of state space models integrates well with likelihood estimation conditioned on sparsely observed data sets from long term ecological research where fine-grained knowledge of continuous state dynamics of a model are not necessary or useful for the inference algorithm. State space model discretization can be handled much more coarsely, which facilitates more efficient scaling to larger T.

Many of the steps of our data assimilation framework are common to those of other Bayesian inference approaches and hence, a wealth of options exist for modification of this approximated SBM inference framework depending on computational resources and desired posterior estimation accuracy. Different non-variational black box inference methods that are compatible with state space approximation of SDEs can be substituted, such as sequential Monte Carlo algorithms (Golightly and Kypraios, 2018), stochastic gradient Hamiltonian Monte Carlo (Chen et al., 2014), stochastic gradient langevin dynamics (Brosse et al., 2018), and stochastic gradient Markov chain Monte Carlo (Aicher et al., 2019; Nemeth and Fearnhead, 2021).

Resesarchers using variational Bayesian methods for their black box can opt for $q(\theta)$ variational approximations that are more complex than mean-field representation. These include full-rank

multivariate logit-normal families in which θ are not assumed to be independent and covariance is established. The full-rank modeling of covariance mitigates underestimation of $q(\theta)$ uncertainty, which is prevalent in mean-field inference, to correspond to wider marginal $q(\theta)$ densities (Kucukelbir et al., 2017; Sujono et al., 2022).

Additionally, when memory availability inhibits the establishment of larger neural networks to train affine shift and scale values for long T or when another method is known to be faster and more convenient for approximating a particular SBM class, different variational families can be used in place of neural moving average flows for representation and optimization of $q(x|\theta)$. These include multivariate normal distributions with specialized covariance structures (Archer et al., 2015), automatic differention VI (Kucukelbir et al., 2017) with Gauss-Markov distributions (Sujono et al., 2022), neural stochastic differential equations (Jia and Benson, 2019; X. Li et al., 2020; Tzen and Raginsky, 2019), and recurrent neural networks (Krishnan et al., 2017; Ryder et al., 2018), among others. Thus, our framework is flexible and can be repurposed as needed for assimilation of different SBMs or Earth system models that vary in complexity and simulation requirements.

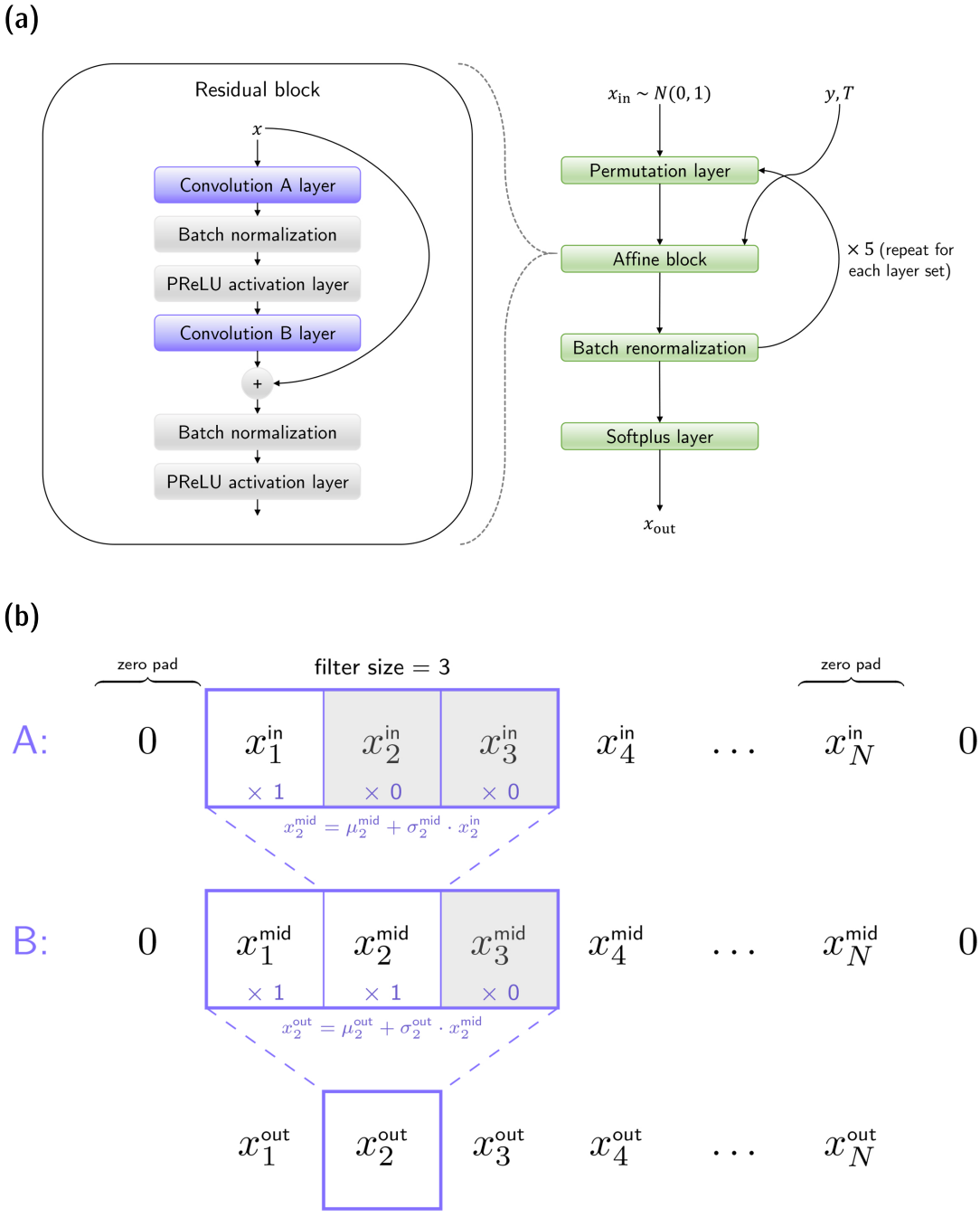

Figure 4.3: Architecture blueprint of the neural moving average flow used as the marginal variational family for $q(x|\theta)$. **(a)** outlines the sequence of layers and operations. The affine block is a residual block in which the autoregressive convolution operations that distinguish neural moving average flows occur. **(b)** illustrates the two bijections, Convolution A and Convolution B, that link three hypothetical layers x^{in}, x^{mid}, and x^{out} together in each instance of an affine layer in our particular flow. Convolution A applies a $[1, 0, 0]$ mask, while Convolution B applies a $[1, 1, 0]$ mask. The example affine μ and σ parameters are indexed by superscripts and subscripts respectively identifying the layer and element they are associated with.

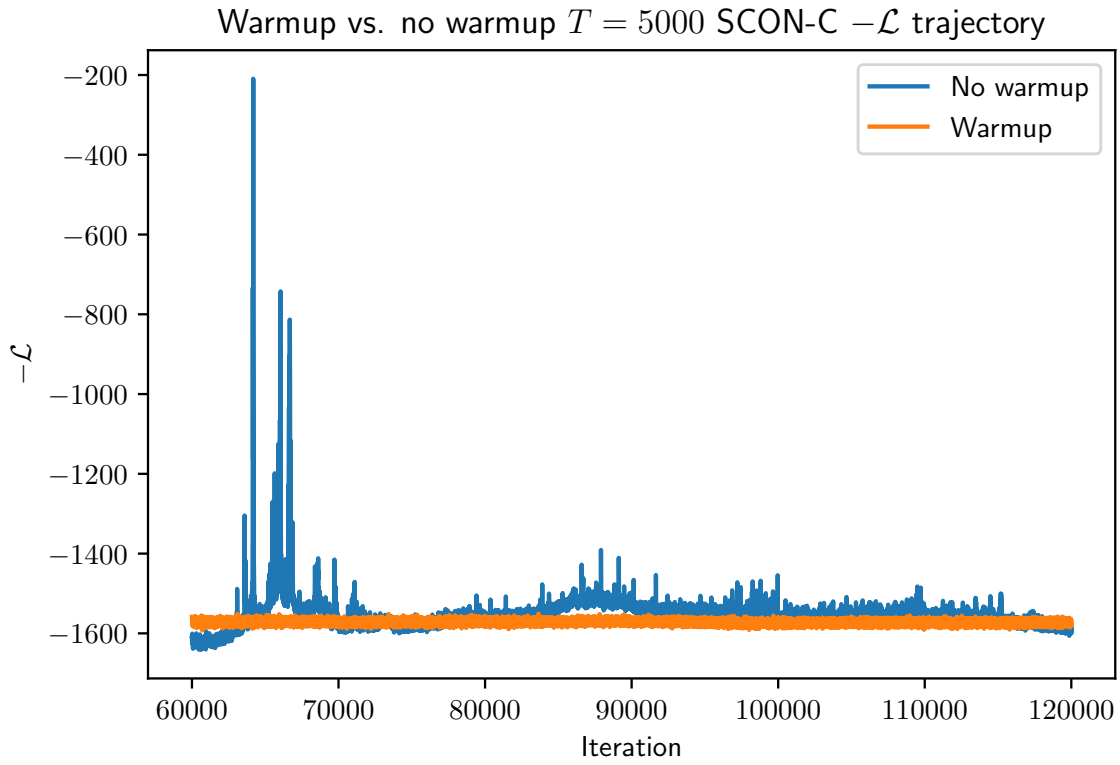

Figure 4.4: Comparison of $-\mathcal{L}$ trajectories from the latter halves of $T = 5000$ hour SCON-C flow trainings without (blue) and with (orange) warmup indicates that warmup helps stabilize training and speed up convergence. The trajectory corresponding to warmup displays much less prominent instability spiking and has flattened more quickly in contrast to the that of the no warmup counterpart.

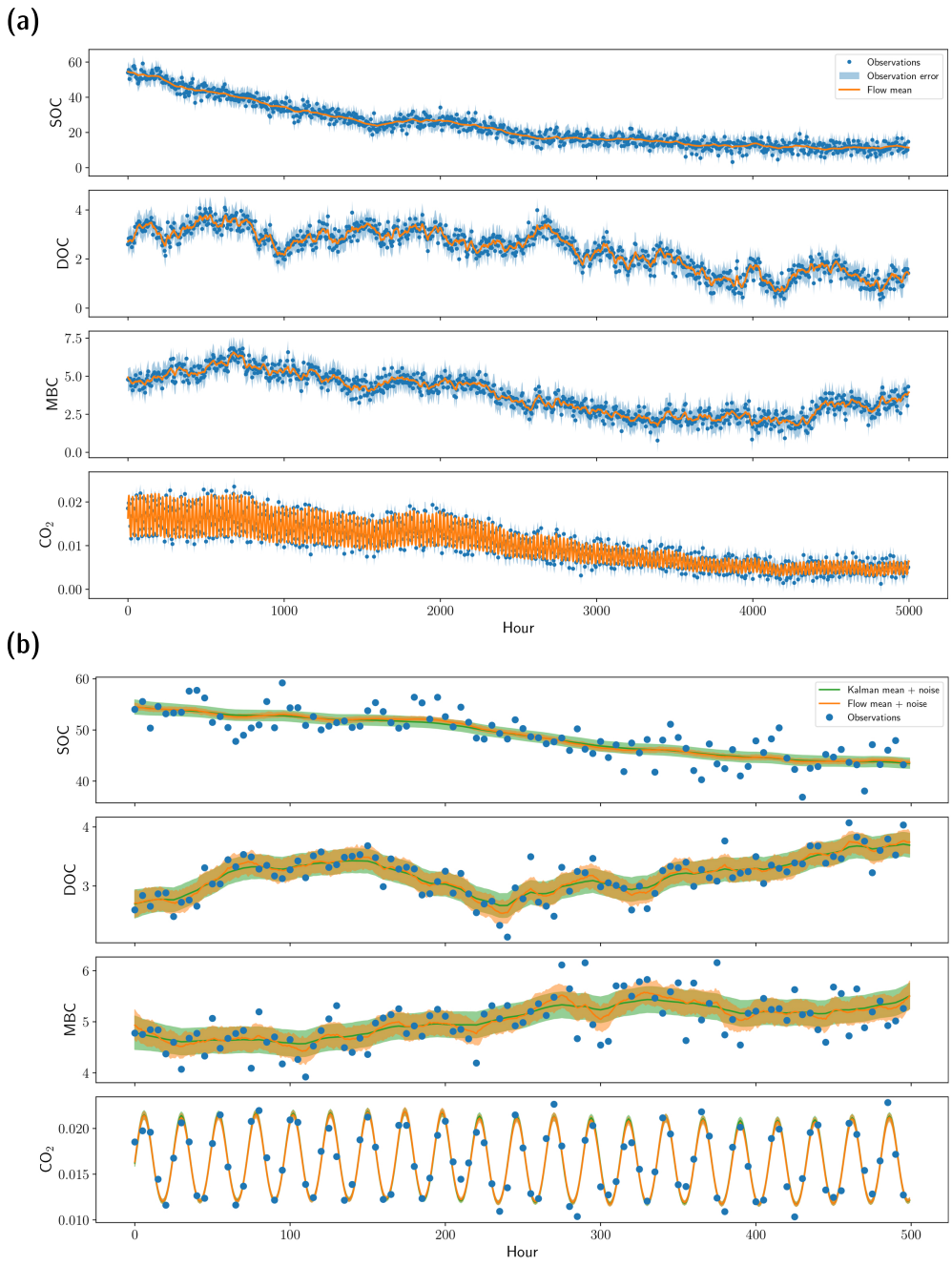

Figure 4.5: Marginal posterior $q(x|\theta;\phi_x)$ soil pool state means (orange lines) of the SCON-C state space model approximated by the neural moving average flow following VI optimization. The means are estimated from 250 x samples drawn from the optimized joint density $q(\theta, x; \phi_{(\theta,x)})$. The states are in units of mg C g^{-1} soil. In **(a)**, the trajectories of flow-approximated state means are compared to the synthetic observations an SCON-C $T = 5000$ hour y backgrounded by the 95% interval of the observation noise (blue dots over blue shading). In **(b)**, we zoom into a subset of the above plot from $t = 0$ to 500 hour and additionally compare the state means and 95% interval of the diffusion distribution of the optimized model to the true posterior means and 95% diffusion noise computed by a Kalman smoother with knowledge of the true θ values.

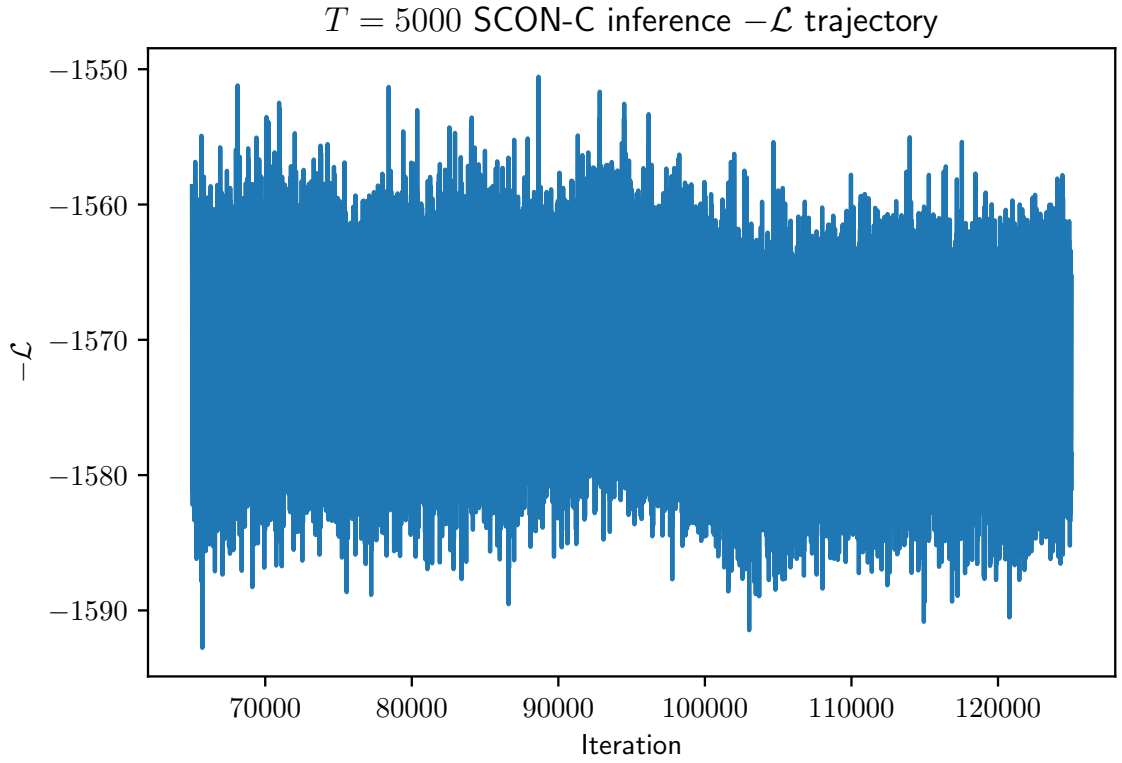

Figure 4.6: Comparison of $-\mathcal{L}$ trajectories from the latter halves of $T = 5000$ hour SCON-C flow trainings without (blue) and with (orange) warmup indicates that warmup helps stabilize training and speed up convergence. The trajectory corresponding to warmup displays much less prominent instability spiking and has flattened more quickly in contrast to the that of the no warmup counterpart.

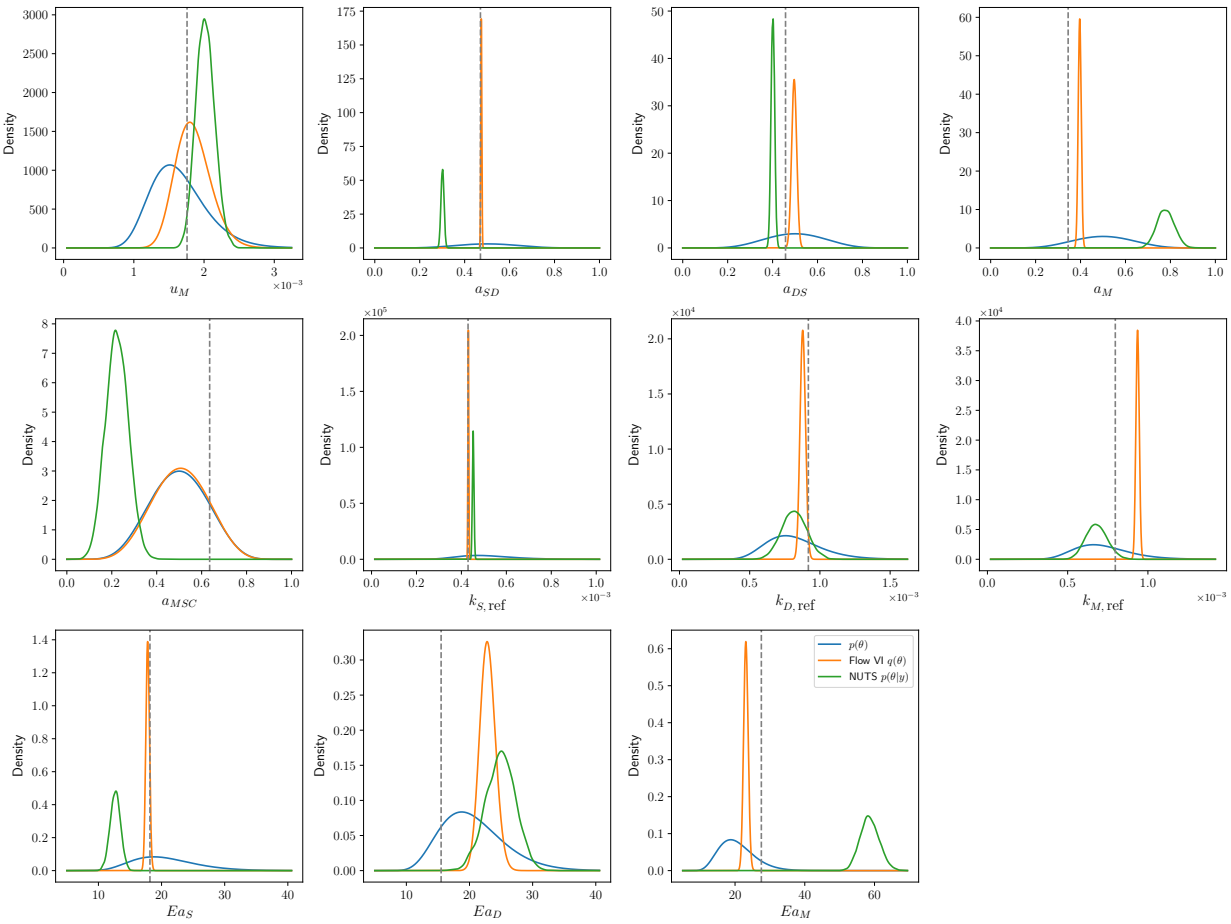

Figure 4.7: SCON-C state space model marginal $q(\theta;\phi_\theta)$ posterior densities following flow VI optimization (orange) compared to mean-field prior $p(\theta)$ densities (blue) and non-parametric CON ODE marginal $p(\theta|y)$ posterior densities estimated with Stan's NUTS algorithm (green). Flow VI and NUTS were conditioned on the same $T = 5000$ hour y generated by an SCON-C SDE. The true θ values sampled during data generation are marked by vertical dashed gray lines. Being a deterministic ODE system, CON does not have β diffusion θ, so subplots portraying the marginal $q(\theta;\phi_\theta)$ densities for the SCON-C state space model c_S, c_D, and c_M θ were not included in this figure due to a lack of comparison.

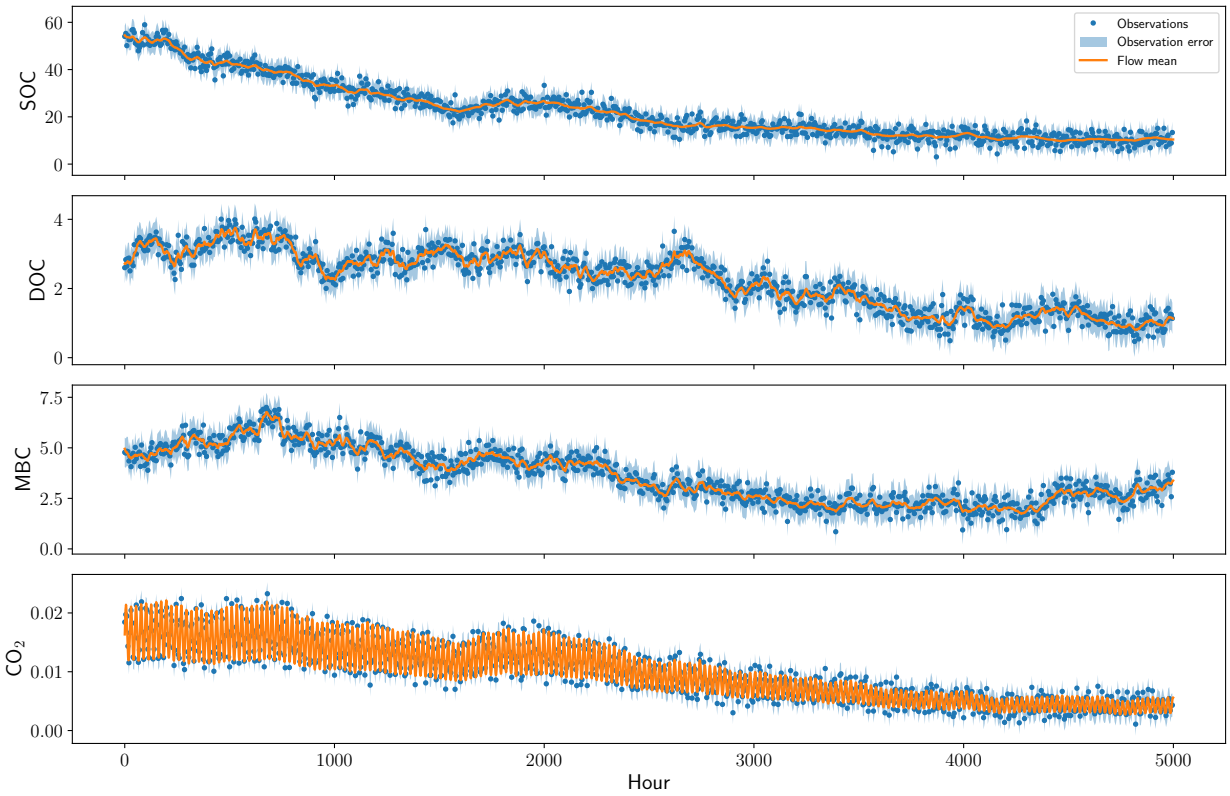

Figure 4.8: Flow-approximated SCON-SS $q(x|\theta;\phi_x)$ latent state and observed CO_2 means conditioned on $T = 5000$ SCON-SS data-generating process y estimated from 250 x paths sampled from the optimized joint variational $q(\theta, x; \phi_{(\theta,x)})$ density.

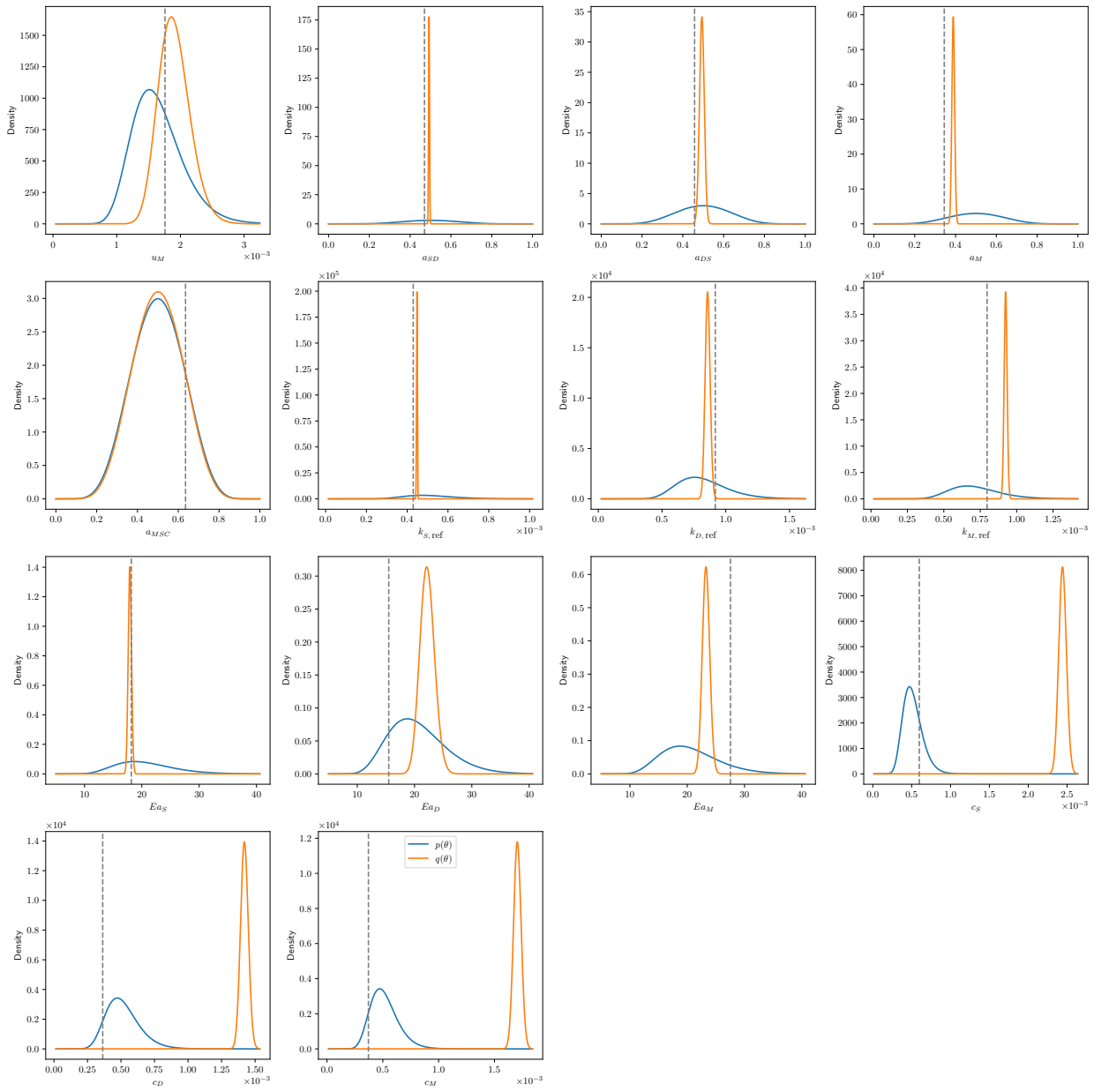

Figure 4.9: Full SCON-SS state space model marginal $q(\theta; \phi_\theta)$ posterior densities (orange) conditioned on $T = 5000$ SCON-SS data-generating process y compared to the prior densities $p(\theta)$ (blue). The true θ values sampled during data generation are marked by vertical dashed gray lines.

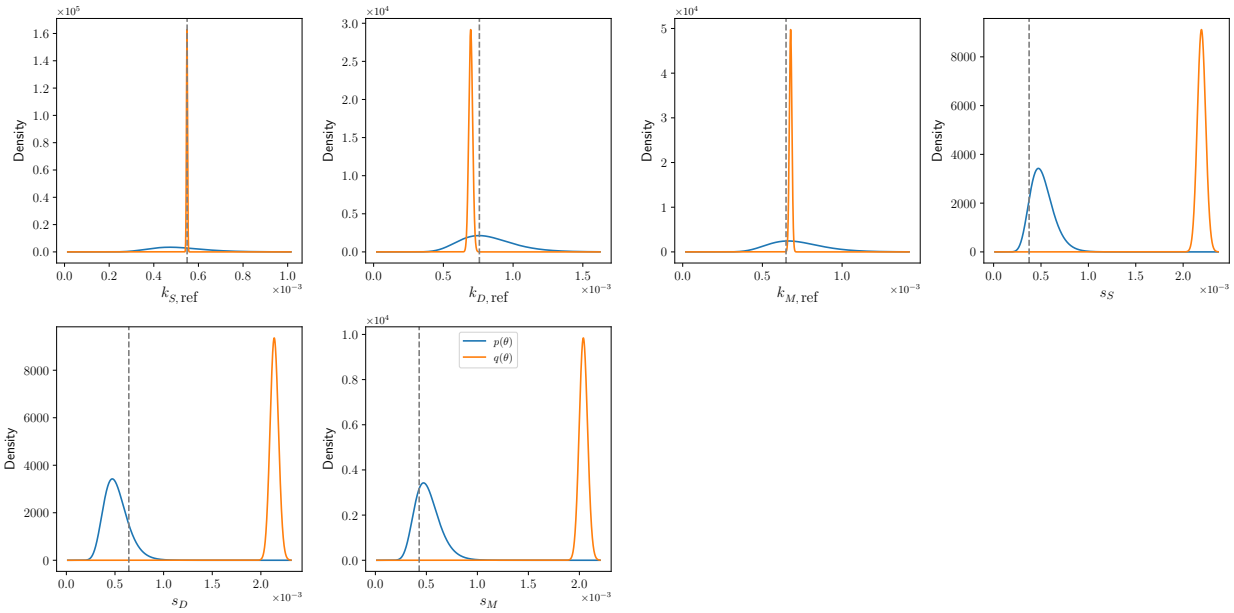

Figure 4.10: Optimized marginal posterior $q(\theta;\phi_\theta)$ densities (orange) of a reduced SCON-SS model with all but the $k_{i\in S,D,M,\mathrm{ref}}$ decay and state-scaling diffusion θ fixed compared to mean-field priors $p(\theta)$ (blue).

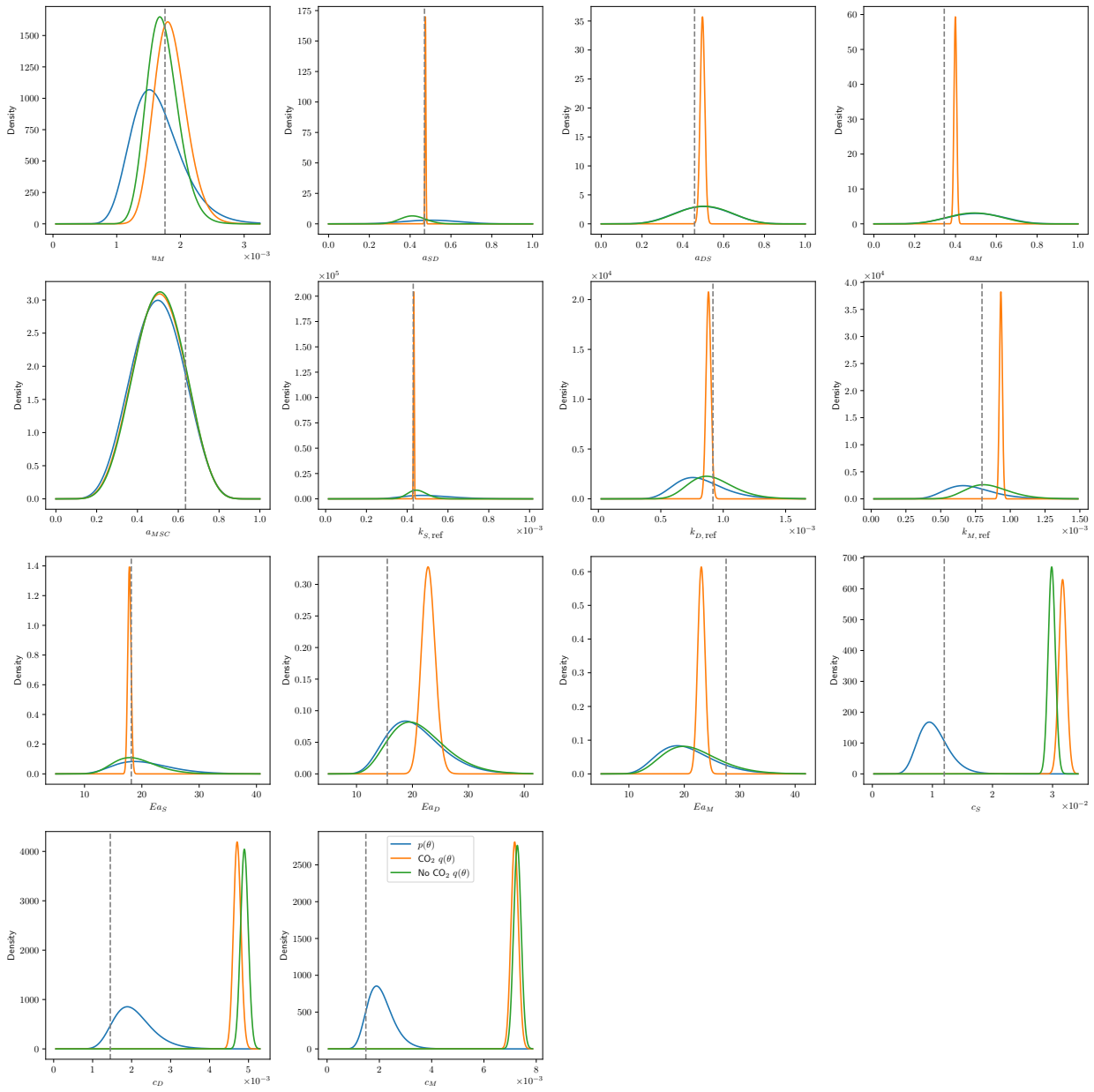

Figure 4.11: Approximate SCON-C state space model marginal $q(\theta; \phi_\theta)$ posterior densities conditioned with (orange) and without (green) CO_2 information in y produced by the same SCON-C data-generating process compared to mean-field prior densities $p(\theta)$ (blue). The true θ values sampled during data generation are marked by vertical dashed gray lines.

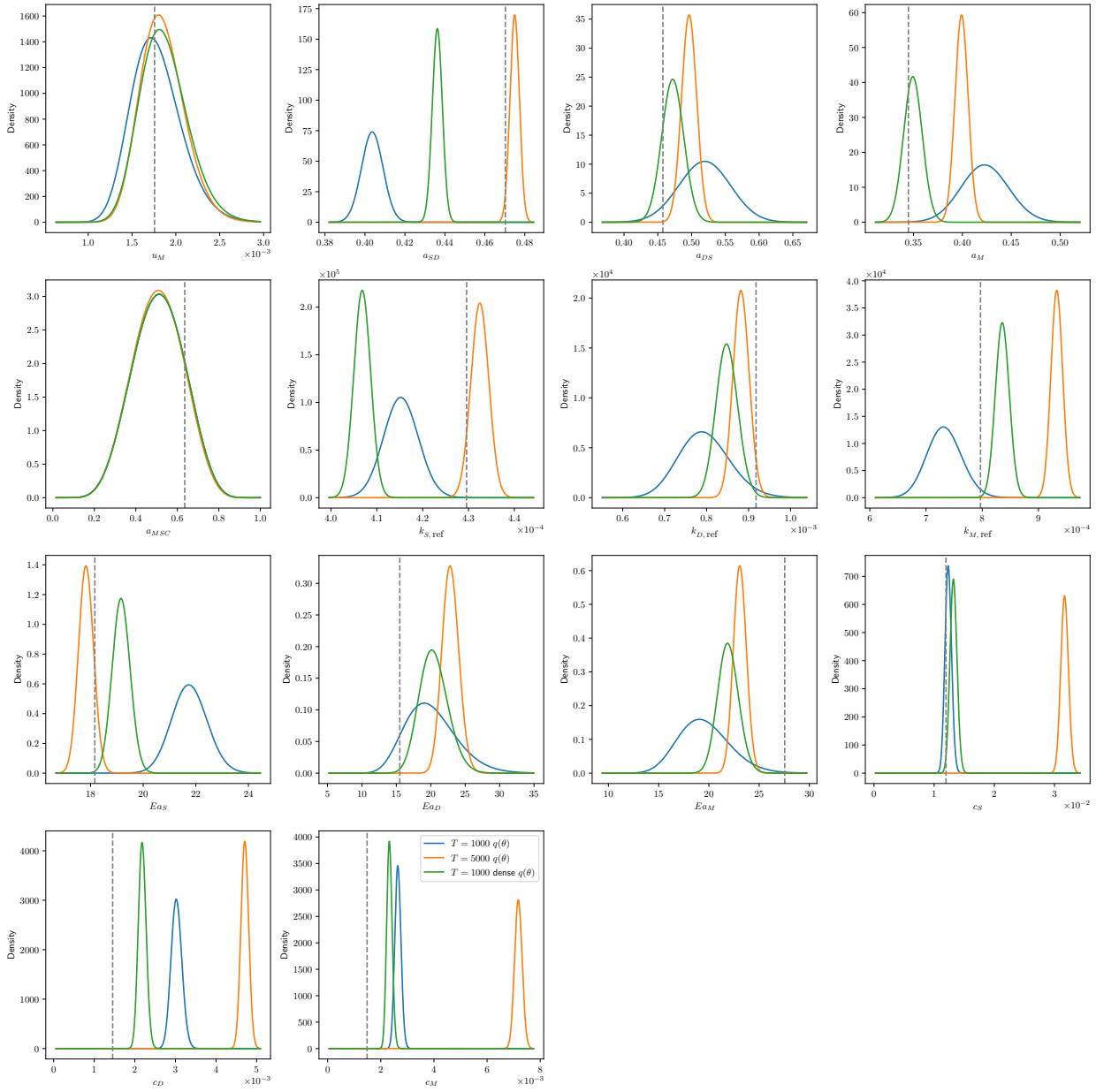

Figure 4.12: Approximate SCON-C state space model marginal $q(\theta; \phi_\theta)$ posterior densities conditioned with $T = 1000$ data observed every 5 hours (blue), $T = 5000$ data observed every 5 hours (orange), and $T = 1000$ data observed every hour (green). All three y share the same SCON-C data-generating process and include CO_2 information. The true θ values sampled during data generation are marked by vertical dashed gray lines.

CHAPTER 5

Black-box data assimilation of a nonlinear soil biogeochemical model with a continuous-time variational inference scheme using mini-batching

5.1 Introduction

5.1.1 Motivation for a separate SBM inference approach

With an ever expanding array of soil biogeochemical models (SBMs) and model variants being proposed in biogeochemistry literature (Luo et al., 2016; Manzoni and Porporato, 2009), and in particular, a surge in the development of nonlinear "microbial-explicit" models starting in the past decade (Wieder, Allison, et al., 2015), we are driven to expand our toolkit of SBM inference frameworks that can effectively assimilate these models to data. Rigorous Bayesian data assimilation facilitates informative validation, stress testing, and benchmarking of these SBMs conditioned on

available data sets. Without reliable validation efforts, it is difficult to assess whether the more heterogeneous and granularized representation of soil microbes in microbial-explicit SBMs like the Microbial-Mineral Carbon Stabilization (MIMICS; Wieder, Grandy, Kallenbach, and Bonan, 2014; Wieder, Grandy, Kallenbach, Taylor, et al., 2015) and Carbon, Organisms, Rhizosphere, and Protection in the Soil Environment (CORPSE; Sulman et al., 2014) models bolsters mechanistic and predictive accuracy to an extent that merits the computational cost of nonlinear and potentially stiff, unstable dynamics.

In incremental fulfillment of this inference framework development goal, we implemented a variational inference (VI) algorithm in the preceding chapter to assimilate variants of the stochastic conventional (SCON) SBM family with data. This algorithm involves establishment of a *variational distribution* of SBM parameters θ and model latent states $\hat{x}$, $q(\theta, \hat{x})$, to estimate the true posterior $p(\theta, \hat{x}|y)$. We selected the so-called neural moving average flow for our marginal $q(\hat{x}|\theta)$ variational family and a mean-field logit-normal variational family for our marginal $q(\theta)$. The variational distribution is characterized by its *variational parameters*, $\phi_{(\theta,\hat{x})}$. We index a particular variational distribution by its corresponding ϕ as $q(\theta, \hat{x}; \phi_{(\theta,\hat{x})})$. Over the course of VI, ϕ is tuned to maximize the evidence lower bound (ELBO) objective, $\mathcal{L}(\phi)$, in gradient descent (or minimize $-\mathcal{L}(\phi)$ in gradient descent), which is commensurate with reducing the gap between $q(\theta, \hat{x})$ and $p(\theta, \hat{x}|y)$.

The efficiency gains are sourced from use of the neural moving average flow. Flow approximation shifts the SBM from a continuous-time dynamical system to a probabilistic discrete-time state space model. In doing so, it absolves the computational expense of solving SBMs sequentially from one time step to the next in each training iteration. Multiple latent state trajectories can be sampled in a single parallelized pass (rather than a sequential process) from the variational distribution constituting the state space model for likelihood evaluation. This parallelization renders the algorithm well-positioned to harness graphical processing unit (GPU) architecture for computation.

Flow approximation does come with trade-offs though, as we observed firsthand. For one, with flows being constituted by neural network layers, large approximation error can accumulate with increased layer depth intended to make the flow more flexible and expressive. Additionally, the number of neural network weights parameterizing the flow scales with the total system simulation time, T, such that inferences conditioned on larger T may correspond to a video random access memory burden that is not compatible under GPU resource limitations. Randomly poor initialization or training draws in the neural network weights can furthermore contribute to inference-killing gradient explosions and approximated system state instability.

A combination of these issues came to a head when we tried to calibrate a particular flow-approximated nonlinear microbial-explicit SBM to synthetic data sampled from the corresponding continuous-time stochastic differential equation (SDE) system. This SBM is called the "state-scaling" stochastic Allison-Wallenstein-Bradford equilibrium chemistry approximation (SAWB-ECA-SS) model. It has four state variables corresponding to soil organic carbon (SOC), dissolved organic carbon (DOC), microbial biomass carbon (MBC), and extracellular enzyme biomass carbon (EEC) pool densities. We will describe the model more extensively in section 5.2.1. Using the same neural moving flow architecture detailed in 4.2.6, even with drastic reductions in optimizer learning rate, we experienced premature termination of approximated SAWB-ECA-SS training due to the appearance of `nan` and `inf` elements in the optimization gradient, ELBO objective, or state path sample tensors.

As we empirically found the neural moving average flow functional for approximation of linear SCON in the last chapter and we encountered simulation hindrances for the original AWB model stemming from state derivative blow-ups in incompatible parameter regimes over the course of our work for Chapter 3 (Xie et al., 2020), our suspicion was that the neural moving average architecture in its existing form could not numerically tolerate the instability and occasionally rapidly changing derivatives of the non-linear carbon (C) transfer dynamics in SAWB-ECA-SS. This suspicion is reinforced by the fact that other nonlinear SBMs like MIMICS have also been

noted to become numerically unstable in certain parameter regimes (Buchkowski et al., 2017; Shi et al., 2018).

5.1.2 Application of a continuous-time SDE optimization scheme

Moderate modifications in the hyperparameters and sequencing of the architecture layers were not sufficient to realize SAWB-ECA-SS flow approximation. Hence, we sought alternative inference approaches. X. Li et al. (2020), Peluchetti and Favaro (2021), and Tzen and Raginsky (2019) outline some elegant options for the approximation of dynamical systems into state space models with *neural SDEs* instead of flows. Analogous to flows, neural SDEs can transform one probability density to another. But, rather than doing so over discrete decomposable bijection layers, as in (4.49), the transformation is carried out by another continuous SDE system selected to return an evolved desired output based on random variable input. In theory, neural SDEs seem preferable to flows due to the potential for mitigation of memory costs. Whereas the number of variational parameters ϕ parameterizing a neural SDE stays constant following the SDE's formulation, increasing flow layer depth for expressiveness accordingly swells the number of ϕ parameterizing all the bijections. However, in practice, the mathematical elegance of neural SDEs has not yet been demonstrated to correspond to consequentially better compute speed or transformation expressiveness over discrete approximations (Xu et al., 2022). Neural SDE implementation also comes with a separate set of challenges under active research, such as transformation SDE choice, complexity, and stability.

Consequently, we considered a simpler option. What about returning to direct inference on the continuous-time dynamical system similar to the Chapter 3 approach, except this time with a non-deterministic SDE and an algorithm more tolerant of noisy likelihoods in VI? In this manner, we could attempt to navigate rapid, "stiff" nonlinear dynamical variation and avert state instability with small solver step sizes under our control.

Of course, solver use in data assimilation, especially with small step sizes lower than discrete model discretizations, recalls persistent issues of compute expense and speed issues. Fortunately, Ryder and Prangle (n.d.) introduces a new GPU-compatible continuous-time VI (CTVI) algorithm that attenuates those numerical burdens by incorporating a prevalent machine learning technique called *mini-batching* into the ELBO gradient optimization process. We quote from the previous chapter for a simplified description of mini-batching: "[under] a mini-batching scheme, a y is partitioned into smaller subsections $y_{\tau i}$ during training, where τ merely distinguishes a subsection from the entire sequence and $i \in \mathcal{B}$ where $\mathcal{B} \in \mathbb{N}$ is the set of integers counting total subsections in natural order. In each training iteration, a $y_{\tau i}$ can then be randomly selected for likelihood evaluation such that the SBM only needs to simulate a [latent $\hat{x}_{\tau i}$] subsection for calculation of the optimization objective." We subsequently apply CTVI in a black-box fashion to assimilate an untrained SAWB-ECA-SS system with synthetic data from a known SAWB-ECA-SS data-generating process and demonstrate the method's functionality for fitting states and recovering parameter values. In doing so, we establish CTVI as a viable choice of inference algorithm in SBM validation frameworks against baseline results obtained from the established No-U-Turn sampler (NUTS) algorithm. Along the way, we test the sensitivity of CTVI to solver step size settings and the availability of CO_2 respiration information in conditioning data.

5.2 Methods

5.2.1 The SAWB-ECA-SS model

The SAWB-ECA-SS SBM is extended from the original Allison-Wallenstein-Bradford (AWB) ordinary differential equation (ODE) model introduced in Allison et al. (2010) and has the same

number of states $\mathcal{D}$ as AWB, with $\mathcal{D} = 4$. Its drift dynamics obey

$$\begin{bmatrix} \mathrm{d}S \\ \mathrm{d}D \\ \mathrm{d}M \\ \mathrm{d}E \end{bmatrix} = \begin{bmatrix} \mathcal{I}_S + a_{MSA} \cdot r_M \cdot M - \frac{V_{\mathrm{DE}} \cdot E \cdot S}{K_{\mathrm{DE}} + E + S} \\ \mathcal{I}_D + (1 - a_{MSA}) \cdot r_M \cdot M + \frac{V_{\mathrm{DE}} \cdot E \cdot S}{K_{\mathrm{DE}} + E + S} + r_L \cdot E - \frac{V_{\mathrm{UE}} \cdot M \cdot D}{K_{\mathrm{UE}} + M + D} \\ u_Q \cdot \frac{V_{\mathrm{UE}} \cdot M \cdot D}{K_{\mathrm{UE}} + M + D} - (r_M + r_E) \cdot M \\ r_E \cdot M - r_L \cdot E \end{bmatrix} \mathrm{d}t + \beta^{0.5} \begin{bmatrix} \mathrm{d}W_S \\ \mathrm{d}W_D \\ \mathrm{d}W_M \\ \mathrm{d}W_E \end{bmatrix} \tag{5.1}$$

where S, D, M, and E are further abbreviations of the SOC, DOC, MBC, and EEC state pools.

The AWB ODE was formulated with standard Michaelis-Menten kinetics for its nonlinear mass transfer functions following the form of $f(A, B) = \frac{V_{\max} \cdot A \cdot B}{K_{\mathrm{M}} + B}$, where A and B are biochemical reactant molecules with A as the enzyme and B as the substrate (Allison et al., 2010; J. Li et al., 2014). In our stochastic extension, we swapped standard Michaelis-Menten for equilibrium chemistry approximation (ECA) kinetics where $f(A, B) = \frac{V_{\max} \cdot A \cdot B}{K_{\mathrm{M}} + A + B}$ due to purported numerical stability boosts assumed from the increase of the denominator by the extra addition term (Tang, 2015; Tang and Riley, 2013).

Following what was introduced in equation (4.5) of section 4.2.2, the naive "state-scaling" diffusion dynamics of SAWB-ECA-SS obey

$$\beta = \begin{bmatrix} s_S \cdot S & 0 & 0 & 0 \\ 0 & s_D \cdot D & 0 & 0 \\ 0 & 0 & s_M \cdot M & 0 \\ 0 & 0 & 0 & s_E \cdot E \end{bmatrix} \tag{5.2}$$

which can be written more compactly as $\beta = \mathrm{diag}\,(s \odot x)$, where diag is the vector diagonalization function, $\odot$ is the Hadamard elementwise multiplication operator, $s = [s_S,\, s_D,\, s_M,\, s_E]$, and $x = [S,\, D,\, M,\, E]$.

SAWB-ECA-SS element	Description	Units
$\mathcal{I}_S$	Exogenous input to SOC pool	mg C g^{-1} soil h^{-1}
$\mathcal{I}_D$	Exogenous input to DOC pool	mg C g^{-1} soil h^{-1}
$u_{Q,\,\mathrm{ref}}$	AWB MBC uptake C use efficiency	NA
Q	AWB CUE temperature sensitivity slope	$^{\circ}C^{-1}$
a_{MSA}	AWB MBC to SOC transfer fraction	NA
K_{DE}	SOC decomposition Michaelis constant	mg C g^{-1} soil
K_{UE}	MBC uptake Michaelis constant	mg C g^{-1} soil
$V_{\mathrm{DE},\,\mathrm{ref}}$	SOC decomposition maximum rate	h^{-1}
$V_{\mathrm{UE},\,\mathrm{ref}}$	MBC uptake maximum rate	h^{-1}
$Ea_{V_{\mathrm{DE}}}$	SOC decomposition activation energy	kJ mol^{-1}
$Ea_{V_{\mathrm{UE}}}$	MBC uptake activation energy	kJ mol^{-1}
r_M	MBC death rate	h^{-1}
r_E	EEC production rate	h^{-1}
r_L	EEC destruction rate	h^{-1}
s_{SOC}	SOC noise scale factor	NA
s_{DOC}	DOC noise scale factor	NA
s_{MBC}	MBC noise scale factor	NA
s_{EEC}	EEC noise scale factor	NA

Table 5.1: List of SAWB-ECA-SS equation elements and their associated biogeochemical or mathematical descriptions and units.

The non-state elements appearing in SAWB-ECA-SS equation terms, their corresponding biogeochemical or mathematical interpretations, and units are listed in Table 5.1. We specified 16 of these 18 elements to be Bayesian random variable model parameters θ whose values are realized through sampling in our synthetic data-generating process. These parameters in turn served as our targets of inference in our CTVI validation experiments. The exogenous pool inputs $\mathcal{I}_S$ and $\mathcal{I}_D$ are not model parameters or constants in this case, but time-dependent function returns. The θ data-generating distribution supports are subsequently detailed in section 5.2.2.

We defined $V_{\mathrm{DE},\,\mathrm{ref}}$ and $V_{\mathrm{UE},\,\mathrm{ref}}$ to output non-reference V_{DE} and V_{UE} through Arrhenius relationships associated with Ea energy of activation θ such that the nonlinear mass transfer functions remained sensitive to external environmental influence to hue to biological realism. V_{DE} and V_{UE}

are given by

$$V_{i\in\{\mathrm{DE},\,\mathrm{UE}\}} = V_{i,\,\mathrm{ref}} \exp\left[-\frac{Ea_i}{R}\left(\frac{1}{\mathrm{temp}_t} - \frac{1}{\mathrm{temp}_{\mathrm{ref}}}\right)\right] \tag{5.3}$$

where R is the ideal gas constant 8.314 $\mathrm{J\,K^{-1}\,mol^{-1}}$, and $\mathrm{temp}_{\mathrm{ref}}$ specifies the "reference" equilibrium temperature which we set at 283 K. As in J. Li et al. (2014) and Xie et al. (2020), u_Q was defined to linearly vary from its reference value based on temperature through

$$u_Q = u_{Q,\,\mathrm{ref}} - Q \cdot (\mathrm{temp} - \mathrm{temp}_{\mathrm{ref}}) \tag{5.4}$$

In contrast to the deterministic AWB model defined in J. Li et al. (2014), we did not establish temperature dependency for the Michaelis constants for this version of SAWB-ECA-SS, K_{DE} and K_{UE}. Since Davidson et al. (2012) found that modeling temperature sensitivity for Michaelis constants in addition to Michaelis-Menten reaction $V_{\max}$ parameters did not substantially improve fits to enzyme assay data sets, we opted for a more parsimonious SBM without K energy of activation θ to reduce overfitting potential.

CO_2 respiration of SAWB-ECA-SS at a given time is defined as a function output dependent on the model states at the corresponding time, D_t and M_t, and θ, K_{UE} and evolved $V_{\mathrm{UE},t}$ (itself a function of t, Ea_{UE}, and the reference temperature due to its Arrehenius temperature dependence). The function is

$$\mathrm{CO}_{2,\,t} = (1 - u_{Q,\,t}) \cdot \frac{V_{\mathrm{UE},\,t} \cdot M_t \cdot D_t}{K_{\mathrm{UE}} + M_t + D_t} \tag{5.5}$$

with CO_2 taking units of $\mathrm{\mu g\,g^{-1}\,soil\,h^{-1}}$.

5.2.2 Synthetic data generation

To assess the validity of CTVI, we mirror the overall approach summarized in section 4.2.1 that was previously used to test the flow-based inference method. First, we observe synthetic time series data y from an SBM with known true θ simulated in continuous-time with a standard SDE solver, like Euler-Maruyama (Maruyama, 1955). Then after assimilating an uncertain model to said data using the algorithm under investigation, we are able to consider how well the algorithm fits states and recovers θ. We can visualize the discrepancy between "truth" and estimates without additional concerns and complications regarding proper empirical data collection, processing, and normalization across disparate experiments.

This time, our y time series was sampled from one seed of an SAWB-ECA-SS SDE system run for a time span T of 2000 hours. We chose 2000 since it was long enough to make mini-batching worthwhile for performance reasons while not being too taxing for relatively recent CPUs from the past decade to simulate. With the noted stability issues of the AWB family (Xie et al., 2020), we could re-run SAWB-ECA-SS reasonably fast enough using different seeds and then filter for y visually featuring state dynamics that were varied, not biologically absurd, and not too volatile.

We used mean-field independently distributed θ as in the prior chapter's data generation for simplicity. However, we used truncated normal (TN) distributions rather than the logit-normal distributions depicted in 4.2.3 this time to handle biological bounding of θ values. To arrive at our TN data-generating distributions, we picked "parent" mean $\check{\mu}$, standard deviation $\check{\sigma}$, and support bounds $[a, b]$ requiring $\check{\mu}, \check{\sigma} \in \mathbb{R}_+$ and $a, b \in \mathbb{R}$ with $b > a$ that modulated random variable transformations from originating normal to TN densities. While the TN distribution is disadvantaged versus the logit-normal in backpropagation stability, it has three ease-of-use advantages including that:

- it has a closed-form probability density function enabling analytic transformations from a

normal distribution,

- the TN location μ and scale σ tend to remain close to the pre-truncation $\check{\mu}$ and $\check{\sigma}$ (with enough distance designated between support bounds) so that post-transformation distribution parameters usually align with targeted specifications,
- and its probability density can be kept symmetric with equidistant truncation on each side relative to the mean.

Refer to Burkardt (2014) for a treatise on TN that steps through the derivation of TN's probability, cumulative, and inverse cumulative density functions from the standard normal probability density function.

We selected data-generating densities corresponding to θ values we qualitatively found to produce prominently varying dynamics and state changes across T for most random seeds. The distribution $\check{\mu}$ and supports are listed in Table 5.2. We set $\check{\sigma} = \frac{\check{\mu}}{4}$. We chose low "state-scaling" noise levels, as we found higher diffusion settings to correspond to rapid "crashing" of SAWB-ECA-SS states toward 0 and drastically diverge from the drift-only dynamics. The system initial conditions x_0 were sampled as $x_0 \sim \mathscr{N}(\tilde{x}_0,\, 0.1{\cdot}\tilde{x}_0)$, where $\tilde{x}_0 = [2{\cdot}S_{\mathrm{ss}}, 4{\cdot}D_{\mathrm{ss}}, 2{\cdot}M_{\mathrm{ss}}, 4{\cdot}E_{\mathrm{ss}}]$ and the ss subscript notates the steady state values of the matching drift-only, deterministic AWB-ECA ODE. We used the deterministic steady state as a non-rigorous means of initializing the system sufficiently far away from the actual stochastic steady state (i.e. stationary solution). The stationary solution that accounts for the SDE noise structure is more difficult to compute. The deterministic steady states were computed from the true θ using the analytic equilibrium equations of the AWB-ECA

ODE that were resolved with the help of computer algebra software. They are

$$A \coloneqq (-K_{\mathrm{DE}} \cdot r_L \cdot (r_E + r_M) \cdot (1 + u_{Q,\mathrm{ref}}) + r_E \cdot u_{Q,\mathrm{ref}} \cdot (\mathcal{I}_S + \mathcal{I}_D)) \cdot (\mathcal{I}_S r_E \cdot (-1 + u_{Q,\mathrm{ref}}) - a_{MSA} \cdot \mathcal{I}_D \cdot r_M \cdot u_{Q,\mathrm{ref}} + \mathcal{I}_S \cdot r_M \cdot (-1 + u_{Q,\mathrm{ref}} - a_{MSA} \cdot u_{Q,\mathrm{ref}})) \tag{5.6}$$

$$B \coloneqq (r_E + r_M) \cdot (-1 + u_{Q,\mathrm{ref}}) \cdot (\mathcal{I}_D \cdot u_{Q,\mathrm{ref}} \cdot (-a_{MSA} \cdot r_L \cdot r_M + r_E \cdot V_{\mathrm{DE}}) + \mathcal{I}_S \cdot (r_E \cdot r_L \cdot (-1 + u_{Q,\mathrm{ref}}) + r_L \cdot r_M \cdot (-1 + u_{Q,\mathrm{ref}} - a_{MSA}) + r_E \cdot u_{Q,\mathrm{ref}} \cdot V_{\mathrm{DE}}) \tag{5.7}$$

$$S_{\mathrm{ss}} = \frac{A}{B} \tag{5.8}$$

$$D_{\mathrm{ss}} = -\frac{K_{\mathrm{UE}} \cdot (r_E + r_M) \cdot (-1 + u_{Q,\mathrm{ref}}) - u_{Q,\mathrm{ref}} \cdot (\mathcal{I}_S + \mathcal{I}_D)}{(-1 + u_{Q,\mathrm{ref}}) \cdot (r_E + r_M - u_{Q,\mathrm{ref}} \cdot V_{\mathrm{UE}})} \tag{5.9}$$

$$M_{\mathrm{ss}} = \frac{u_{Q,\mathrm{ref}} \cdot (\mathcal{I}_S + \mathcal{I}_D)}{(1 - u_{Q,\mathrm{ref}}) \cdot (r_E + r_M)} \tag{5.10}$$

$$E_{\mathrm{ss}} = \frac{r_E \cdot M_{\mathrm{ss}}}{r_L} = \frac{r_E \cdot u_Q \cdot (\mathcal{I}_S + \mathcal{I}_D)}{r_L \cdot (1 - u_{Q,\mathrm{ref}}) \cdot (r_M + r_E)} \tag{5.11}$$

We used the deterministic steady state as a non-rigorous means of initializing the system sufficiently far away from the actual stochastic steady state solutions, also known as the stationary solutions, as the stochastic steady state accounting for noise structure is more difficult to compute. These initial condition values are biologically plausible for soil surface decomposition processes (Xie et al., 2022). On the note of biological plausibility, we used the same litter input and temperature sinusoids featured in equations (4.6) and (4.8) from section 4.2.2 that were evocative of tropical biome conditions to respectively inject external C into the system and force the $V_{i \in \{\mathrm{DE,UE}\},\mathrm{ref}}$ parameters.

In addition to using truncated θ data-generating distributions, we also hard-bounded $\hat{x}$ to be greater than or equal to 1×10^{-5} to avoid 0 and negative values for biological realism in data generation. With our choice of parameters and simulation time span, x rarely met the 1×10^{-5} lower bound in practice. We took care to select a simulation corresponding to a random seed that never encountered the lower bound.

y were densely and noisily observed from the true x every 10 hours with $y_t \sim \mathcal{N}(x_t, \eta_{\mathrm{obs}})$, where

SAWB-ECA-SS θ	μ	Support
$u_{Q,\,\mathrm{ref}}$	0.22	$[0,\,1]$
Q	0.001	$[0,\,0.001]$
a_{MSA}	0.5	$[0,\,1]$
K_{DE}	1000	$[0,\,5000]$
K_{UE}	0.1	$[0,\,1]$
$V_{\mathrm{DE,\,ref}}$	0.04	$[0,\,1]$
$V_{\mathrm{UE,\,ref}}$	0.005	$[0,\,0.1]$
$Ea_{V_{\mathrm{DE}}}$	40	$[5,\,80]$
$Ea_{V_{\mathrm{UE}}}$	30	$[5,\,80]$
r_M	0.00016667	$[0,\,0.1]$
r_E	0.0002	$[0,\,0.1]$
r_L	0.0004	$[0,\,0.1]$
s_{SOC}	0.0001	$[0,\,0.1]$
s_{DOC}	0.0001	$[0,\,0.1]$
s_{MBC}	0.0001	$[0,\,0.1]$
s_{EEC}	0.0001	$[0,\,0.1]$

Table 5.2: Corresponding data-generating density means and supports of of SAWB-ECA-SS θ.

η_{obs} is the noise standard deviation. If CO_2 measurements were included in y (as they were by default unless otherwise specified), CO_2 was computed directly from the latent time-corresponding x (not from the state observations in y) with equation (5.5) and then observed with the same Gaussian noise pattern. The observation error was structured to resemble equation (4.11) in section 4.2.2 with η_{obs} being 10% of the state and CO_2 means across the data-generation time span such that

$$\eta_{\mathrm{obs}} = 0.1 \odot \begin{bmatrix} \bar{S} \\ \bar{D} \\ \bar{M} \\ \bar{E} \\ \overline{\mathrm{CO}_2} \end{bmatrix} \tag{5.12}$$

5.2.3 Continuous-time variational inference model training procedure

Training of an uncertain SAWB-ECA-SS model to attempt fitting of y sampled from the known SAWB-ECA-SS data-generating process and recover the true underlying latent states and θ was conducted with the aforementioned CTVI method. The algorithm will be more formally and rigorously described in a forthcoming manuscript from Ryder and Prangle (n.d.). For now, we loosely outline its steps in Algorithm 5.1 and generally treat it as a black box.

We review, modify, and introduce some more notation and information relevant to the algorithm outline. We partition our observations and latent state trajectories y and X spanning $T = 2000$ into 5 evenly sized subsections that are $\frac{T}{5} = 400$ hours in length. A particular subsection is then subscripted with an index τb, where τ distinguishes any mini-batch from the full time series and $b \in \mathcal{B} \coloneqq \{1 : 5\}$, where $\mathcal{B}$ is the set of natural number indices counting all the subsections. Each subsection starts at $t = (b-1) \cdot 400$ and ends at $t = b \cdot 400$. There is some intermediate overlap between the end and start points, e.g. coverage of $t = 400$ is shared by the end of a $\tau 1$ subsection and start of a $\tau 2$ subsection and coverage of $t = 1600$ is shared by $\tau 4$ and $\tau 5$ subsections.

Before touching touch upon the reason for the mini-batch time overlap, we make some further notational distinctions. Whereas $X_{\tau b}$ marks a distribution of true latent subtrajectories t-spanning $[(b-1) \cdot 400, b \cdot 400]$, $\hat{X}_{\tau b}$ indicates a distribution of subtrajectories for the model being trained. $\hat{x}_{\tau b}$ is then a sampled realization from the $\hat{X}_{\tau b}$ distribution, $q(\hat{x}_{\tau b}|\theta; \phi_{\hat{x}_{\tau b}})$. $\phi_{\hat{x}_{\tau b}}$ denotes the variational parameters that specifically influence the distribution of τb subtrajectories. In this case, $\phi_{\hat{x}_{\tau b}}$ parameterizes the sampling of initial conditions for subtrajectory integration by an SDE solver from $[(b-1) \cdot 400, b \cdot 400]$. Hence, we arrive at the reason for our end and start point overlap; for the subsections located after $\tau 1$, information regarding the distribution of endpoints of the immediately preceding subsection is used to set initial conditions and tune $q(\hat{x}_{\tau b}|\theta; \phi_{\hat{x}_{\tau b}})$ in

the gradient optimization step. With regard to solving of $\tau 1$ subsections, the CTVI algorithm was established to learn a mean initial condition $\hat{x}_0$ at t_0. For simplicity, we initialized optimization with y_0 as the preliminary $\hat{x}_0$ for $\hat{x}_{\tau 1}$ integration.

We integrate $\mathcal{S} = 32$ samples from $q(\hat{x}_{\tau b}|\theta; \phi_{\hat{x}_{\tau b}})$ in each training iteration. Of course, $q(\hat{x}_{\tau b}|\theta; \phi_{\hat{x}_{\tau b}})$ is conditional on a draw of θ from the approximate marginal $q(\theta; \phi_\theta)$. In contrast to the Chapter 4, we used a flexible non-parametric *inverse autoregressive flow* (IAF) for our posterior variational family. We summarize IAFs here as a reversible sequence of discrete one-to-one random variable transformations (i.e. *bijections*) that transforms one probability density to another. The bijections are parameterized by layers of hidden parameters ϕ_θ constituting a neural network. Refer to Kingma et al. (2016) for a detailed exposition of IAFs. Our IAF transforms a basic standard normal distribution into a more complex one, but it is also possible to operate an IAF in an opposite direction going from complex to simpler. Our prior $p(\theta)$ was initialized as an uninformed IAF with widely distributed ϕ_θ. Also in contrast to Chapter 4, in which the logit-normal posterior variational family had two-sided truncation, the IAF was only bounded at one side with a smooth softplus function to have a support of $[0, \infty)$. As we will observe in section 5.3, one-sided bounding was sufficient for constraint of θ.

For our SDE solver, we used a simple Euler-Maruyama scheme (Maruyama, 1955) due to its ease of implementation and low memory cost that was sufficient for our purposes. We experimented with using solver time step sizes of $\mathrm{d}t = 0.05$ and $\mathrm{d}t = 0.2$. While we did not require the use of a more complex solver this time, we note that an advantage of the CTVI method is that we can use arbitrary SDE solvers, including more elaborate schemes with superior error and convergence guarantees, to tackle the inference of stiffer and more convoluted systems.

Over the course of our specified number of training iterations, the overall goal of our VI optimization process is to find the ϕ that best lowers the distance between the approximate and true posteriors. This distance is indicated by the *Kullback-Leibler (KL) divergence* (Kullback and Leibler, 1951). As encapsulated in section 4.2.5, minimization of the KL divergence is commen-

surate to the minimization of the negative ELBO objective $-\mathcal{L}(\phi)$ in stochastic gradient descent. The complete expression for $\mathcal{L}(\phi)$ will be provided in Ryder and Prangle (n.d.). For the time being, Figure 5.1 indicates empirically that the ELBO objective was appropriate for functional optimization. The ELBO gradient $\widehat{\nabla\mathcal{L}(\phi)}$ used to update ϕ in the gradient optimization step was estimated with the assistance of the reparameterization trick.

We chose the Adam optimizer (Kingma and Ba, 2015) to handle stochastic gradient updates with the optimizer learning rate annealed from 5×10^{-4} to 1×10^{-6} over more than 2.4×10^{6} training iterations. Training was conducted on a single Nvidia GPU.

Algorithm 5.1 Synopsis of steps in each CTVI algorithm loop iteration. Notation detailed throughout section 5.2.3.

Define $q(\theta, x; \phi)$;
Initialize ϕ;
$N \leftarrow$ total training iterations;
for $i \leftarrow 1$ to N **do**
    Uniformly sample subsection b from $\mathcal{B} \coloneqq \{1:5\}$;
    Draw $\theta \sim q(\theta; \phi_\theta)$;
    for $s \leftarrow 1$ to $\mathcal{S} = 32$ **do**
        Integrate $\hat{x}^{(s)} \sim q(\hat{x}|\theta; \phi_x)$ with an SDE solver;
        if y includes CO_2 observations **then**
            Compute $\widehat{CO_2}$ from $\hat{x}$ using equation (5.5);
        end if
    end for
    Evaluate $\widehat{\nabla\mathcal{L}(\phi)}$ for stochastic gradient update of ϕ;
end for
return $q(\theta, x; \phi)$ corresponding to the $\mathcal{L}[\phi_{(\theta,x)}]$ value at N;

5.2.4 Stan No-U-Turn sampler comparison

We compared CTVI marginal θ posterior estimates and state fits to results obtained from Markov chain Monte Carlo (MCMC) inference of a deterministic AWB-ECA ODE system conditioned on the same y. The comparison inference followed the "uni-batched" data assimilation framework applied in Xie et al. (2020) and was conducted on a dual-core Intel Core i5 Kaby Lake central pro-

cessing unit (CPU) using the NUTS engine implemented in Stan 2.29.1 (Carpenter et al., 2017) via the CmdStanR interface (Gabry and Češnovar, 2021. For more detail about the NUTS procedure, refer to Hoffman and Gelman (2014) and Betancourt (2017). Simulation was conducted on CPU rather than GPU because Stan's GPU support is currently still limited in development.

The deterministic AWB-ECA dynamics match the SAWB-ECA-SS drift given in equation (5.1). We used AWB-ECA in the comparison rather than SAWB-ECA-SS because Stan only has built-in ODE solvers and no native SDE solver at the time of this writing. Also, we wanted to demonstrate superiority of the SDE CTVI approach over the ODE-NUTS framework of Xie et al. (2020). We opted for Stan's implementation of the Cash-Karp scheme (Cash and Karp, 1990) as our choice of ODE solver due to its reputation as a robust, memory-efficient integrator of nonlinear and relatively stiff systems. For simplicity, we assumed the observed y_0 as the initial condition x_0 for system integration starting from $t_0 = 0$. Stan's ODE solver implementation made it difficult for us to flexibly vary and learn separate initial conditions, as the solver application programming interface required definition of a fixed x_0.

The AWB-ECA priors were defined as independent TN distributions mirroring the data-generating distributions detailed in Table 5.2. We found that the mean-field NUTS posteriors required two-sided truncation for biological constraints, as we had experienced in Xie et al. (2020); the posterior supports matched those of the prior and data-generating densities.

With regard to NUTS configuration, we specified the collection of posterior samples across four ergodic chains of 500 warmup iterations and 1250 sampling iterations each such that a maximum of $4 \times 1250 = 5000$ posterior samples were amassed per θ. Other NUTS engine prameters were left at Stan's defaults, except for the `adapt_delta` argument controlling the target average acceptance probability. We increased `adapt_delta` to 0.95 from the CmdStanR version default of 0.8 to prompt shorter jumps in θ proposals. In our experience with models from the AWB family, longer jumps resulted in more frequent navigation into unstable θ regimes that produced divergent transitions.

5.3 Results

5.3.1 CTVI ELBO convergence

To check for suitable ELBO optimization and gradient descent behavior in the CTVI algorithm, we visualized the negative ELBO trajectories of two CTVI runs conditioned on our synthetic SAWB-ECA-SS y including CO_2 observations (Figure 5.1). These two runs are distinguished by their Euler-Maruyama integrator $\mathrm{d}t$ step size, with one maintaining $\mathrm{d}t = 0.05$ and the other $\mathrm{d}t = 0.2$. The $\mathrm{d}t = 0.05$ run took about three hours, while the $\mathrm{d}t = 0.2$ run took about one hour.

The gradual overall declines in the values and slope magnitudes of both $-\mathcal{L}$ curves with increasing training iterations signals approach toward an approximate local minimum in the ELBO objective and, by proxy, KL divergence. Though the $\mathrm{d}t = 0.05$ trajectory appears to usually track a little lower than its $\mathrm{d}t = 0.2$ counterpart, the two curves ultimately display relatively little separation in terminating value and overall shape over their courses.

5.3.2 Comparing the CTVI method to the ODE-NUTS approach

We visually and qualitatively compared the results of the $\mathrm{d}t = 0.05$ time step SAWB-ECA-SS CTVI optimization to those from the AWB-ECA ODE-NUTS assimilation starting with the state fits to y (that included CO_2 observations). In contrast to the three hours it took to run the $\mathrm{d}t = 0.05$ CTVI run, the AWB-ECA ODE-NUTS inference took about eight hours.

When viewing the complete simulation time span from $t = 0$ to $t = 2000$ (Figure 5.2), we observe that the trajectory of the CTVI SOC posterior distribution and more easily visible posterior mean aligns somewhat more closely with the synthetic y SOC observations and observation noise interval

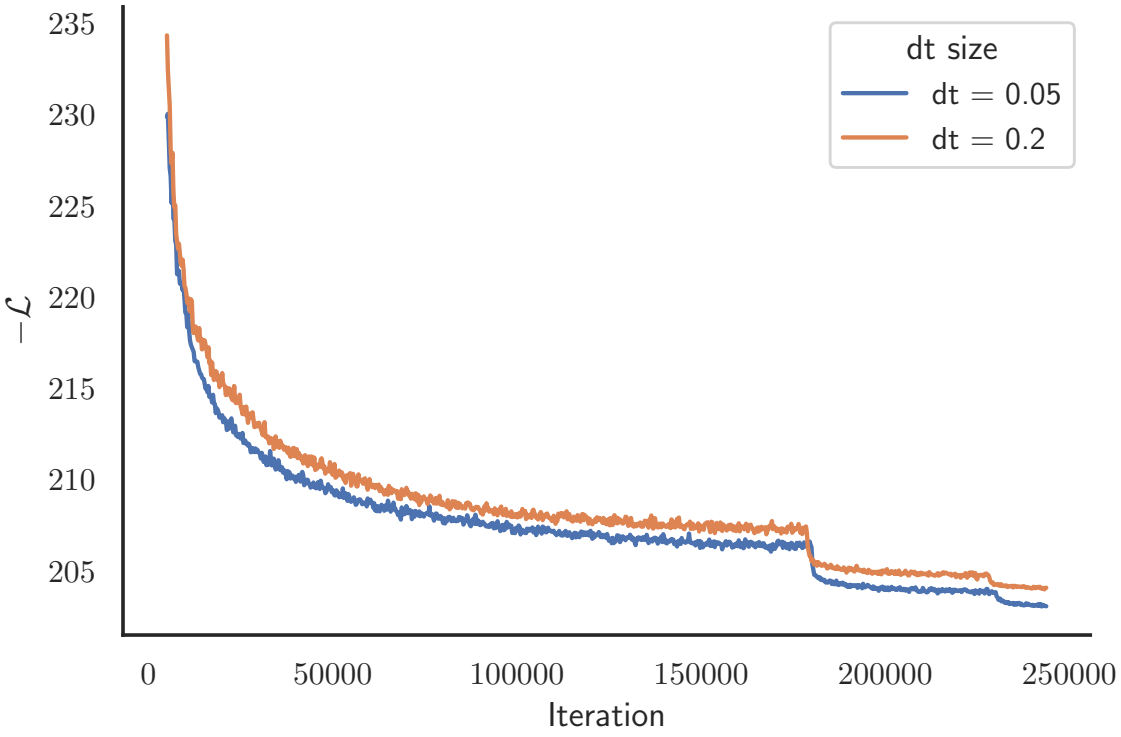

Figure 5.1: The $-\mathcal{L}$ trajectories of two CTVI runs conditioned on y—one operating with an Euler-Maruyama SDE solver time step size of $\mathrm{d}t = 0.05$ (blue) and the other with $\mathrm{d}t = 0.2$ (orange)—traced over the course of the more than 2×10^6 stochastic gradient descent iterations for each run.

than the ODE-NUTS SOC posterior distribution and mean. The ODE-NUTS SOC posterior distribution overshoots the observation noise interval throughout the entire time span and is not able to adjust its trajectory from ODE solver initialization at y_0. It is more difficult to distinguish between the CTVI and ODE-NUTS state posteriors corresponding to DOC, MBC, and EEC, which visually fit y similarly well. Likewise, with regard to CO_2, drastic differences in fit quality were not apparent, though the CTVI CO_2 means were consistently better at staying within the 95% observation interval, such as at $t = 100$ (Figure 5.3) and $t = 1840$ (Figure 5.4). The $\mathrm{d}t = 0.2$ time step CTVI optimization state posteriors were not plotted, as they tracked too closely with the $\mathrm{d}t = 0.05$ results.

We additionally detect in Figure 5.2 increases in the uncertainty of the posterior DOC, MBC,

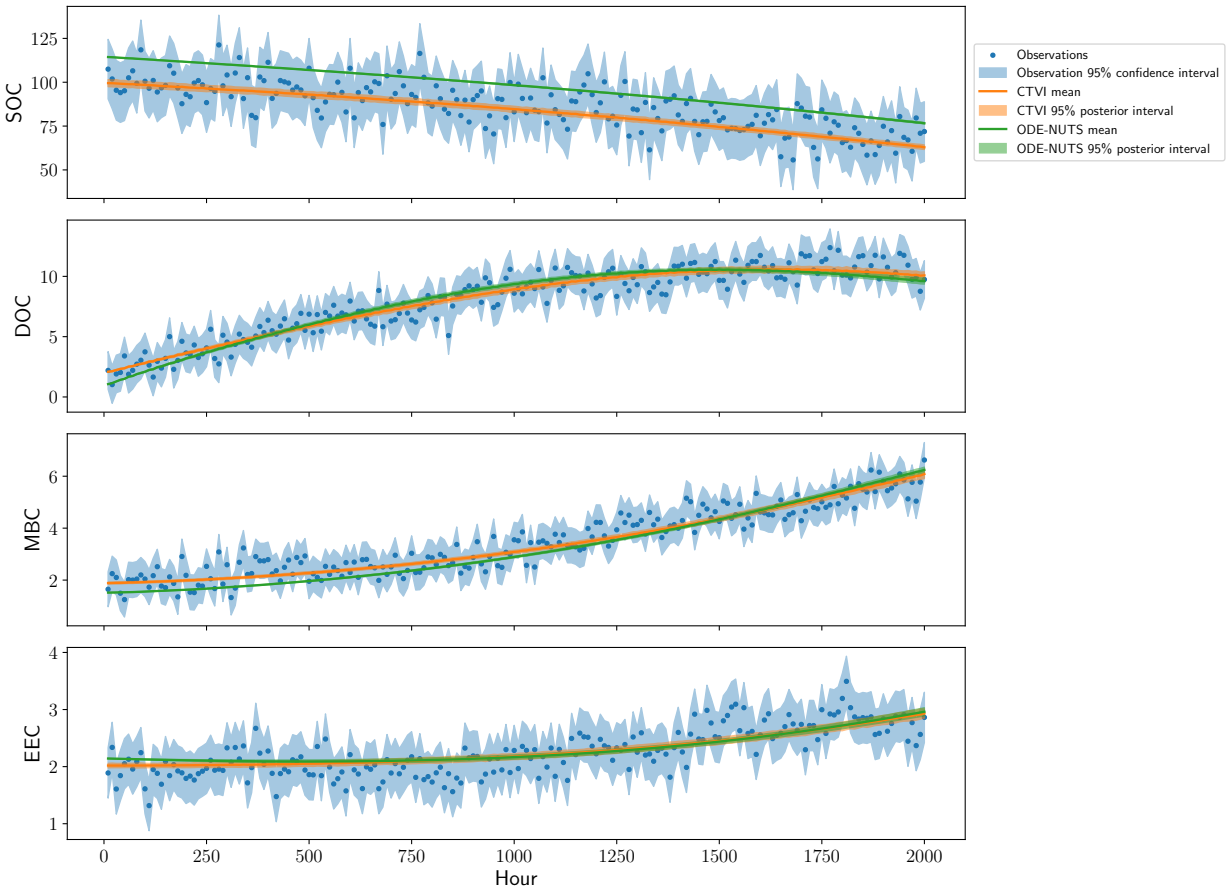

Figure 5.2: Full view of the distributions of the SAWB-ECA-SS CTVI $\mathrm{d}t = 0.05$ and AWB-ECA ODE-NUTS x state fits to y from $t = 0$ to $t = 2000$. CO_2 comparison is not pictured due to poor visibility of points at this time window.

and EEC distributions for both CTVI and ODE-NUTS through the overall time span. This is juxtaposed by a decrease in uncertainty for the SOC CTVI posterior. The changes in uncertainty and state posterior variance are more clearly visible when comparing zoomed-in views of the beginning (Figure 5.3) and end (Figure 5.4) of the time series.

We note that the dynamics of the CTVI trajectories appear quite smooth and deterministic (Figures 5.2, 5.3, 5.4), which points to training toward low marginal state-scaling diffusion θ means in the optimization of approximate $q(\theta)$. This is an accurate reflection of the true data-generating state-scaling θ depicted in Table 5.2.

Before we proceed to describe our interpretations of figures contrasting marginal θ posteriors

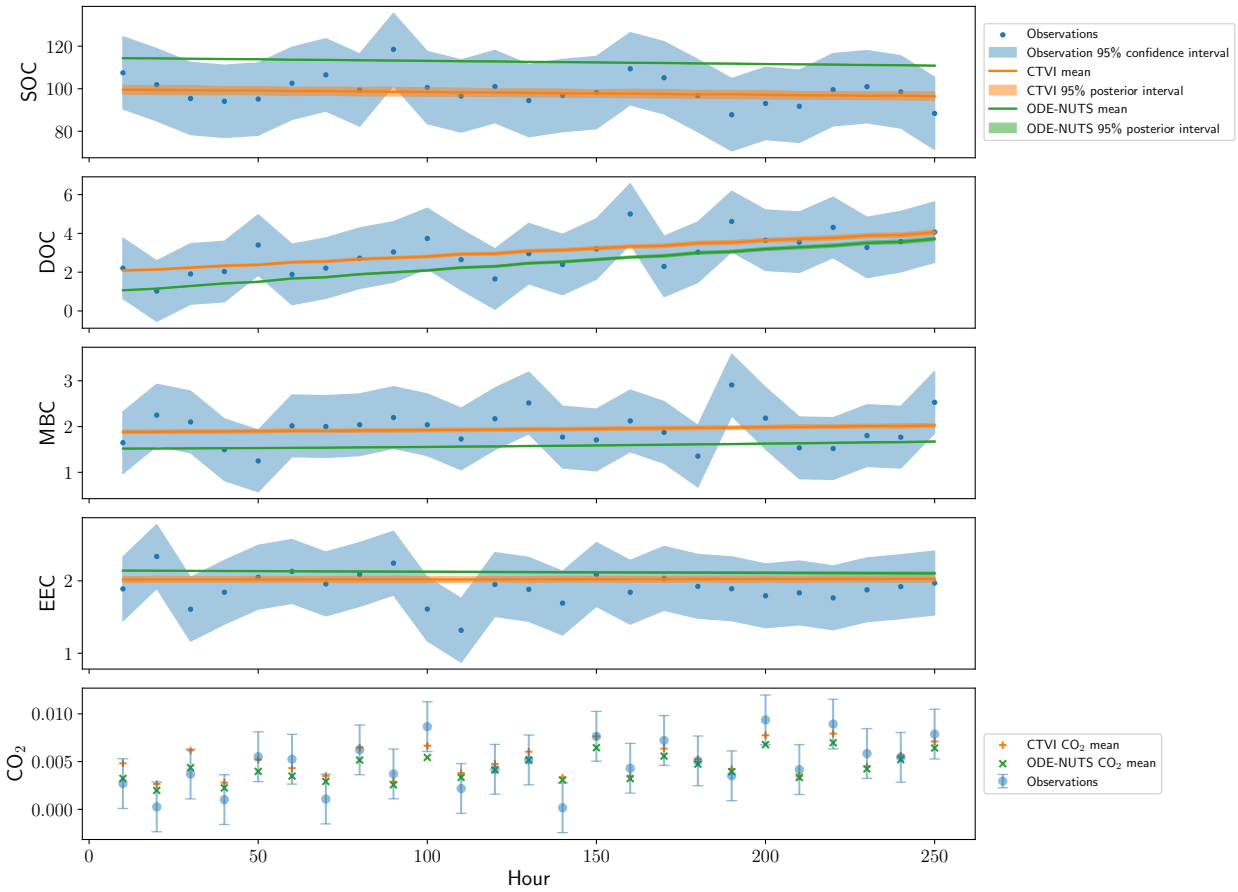

Figure 5.3: Zoomed view of the SAWB-ECA-SS CTVI and AWB-ECA ODE-NUTS x fits from $t = 0$ to $t = 250$.

estimated by various approaches, we state for clarity that we judged the ability of posteriors to "identify" true θ by the horizontal distances separating perceived posterior density peaks[1] from true θ values. We deemed that smaller distances indicated enhanced posterior identifiability. If two posteriors harmonized their peaks over a true θ with similar precision, the tiebreaker for identifiability was narrower and more certain posterior width. As the delineation of model θ identifiability can be amorphous, we settled on a definition of identifiability that we thought was reasonable and allowed for prompt visual assessment.

When contrasting the marginal SAWB-ECA-SS IAF-family posteriors estimated by CTVI and the

[1]Henceforth, we also use density "means" interchangeably with "peaks"' for shorthand, acknowledging that for the non-parametric IAF-family densities, the peaks and means may be close to each other since the IAF-family densities are predominantly Gaussian-shaped but not the exact same.

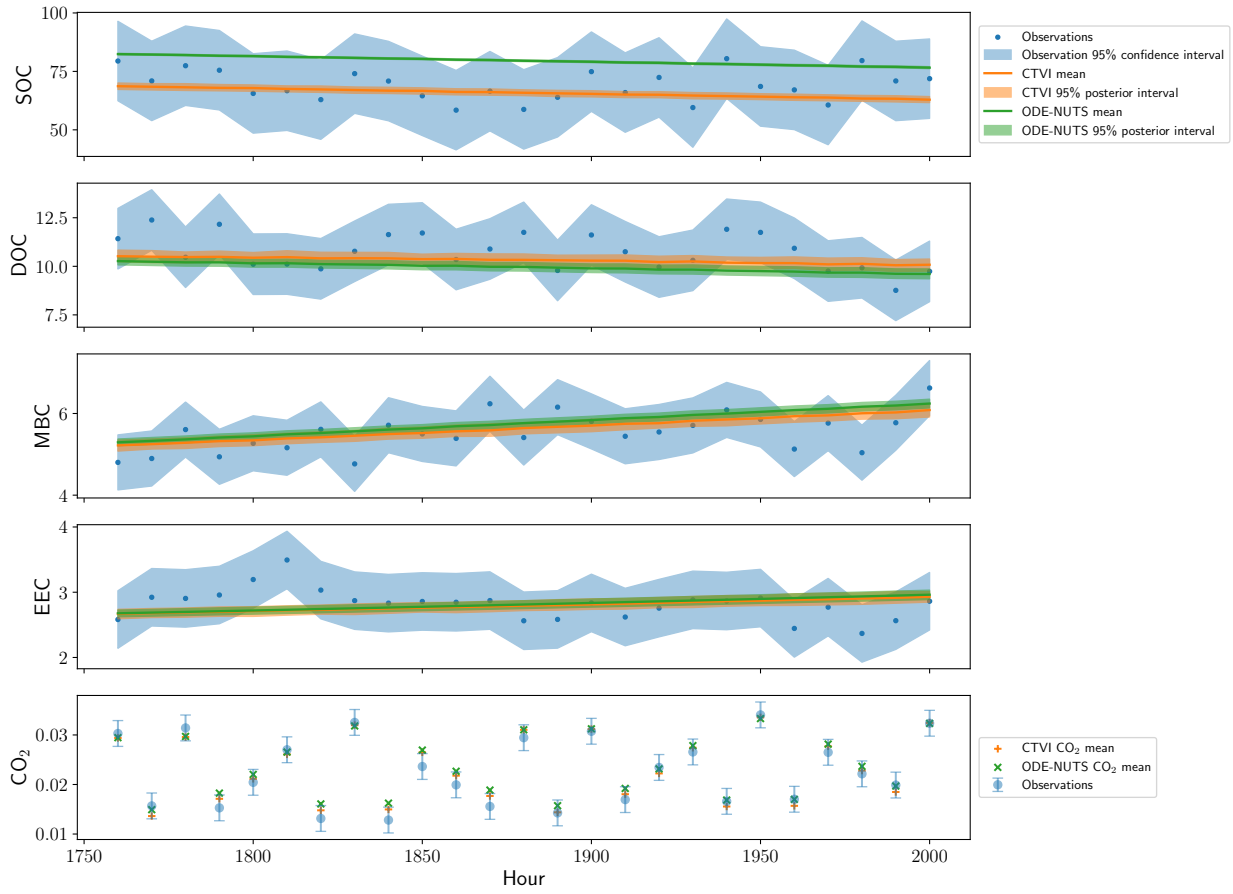

Figure 5.4: Zoomed view of the SAWB-ECA-SS CTVI and AWB-ECA ODE-NUTS x fits from $t = 1750$ to $t = 2000$.

marginal AWB-ECA posteriors sampled by NUTS, we saw that the CTVI-estimated posteriors customarily displayed closer and more certain identification of the true θ (Figure 5.5). This is especially apparent in the $u_{Q,\mathrm{ref}}$, Q, K_{DE}, K_{UE}, $V_{\mathrm{DE,ref}}$, $V_{\mathrm{UE,ref}}$, and $Ea_{V_{\mathrm{DE}}}$ subplots, where the CTVI posterior densities consolidated about means that precisely matched or were exceedingly close to the true θ. Performance was also arguably better for CTVI in the a_{MSA}, $Ea_{V_{\mathrm{UE}}}$, and r_M subplots, where the CTVI posterior means were not located on top the true θ, but were still visibly closer.

The SAWB-ECA-SS IAF-family posterior densities never displayed extreme density consolidation far above any true θ. On the other hand, two of the marginal AWB-ECA posteriors corresponding to Q and r_M in Figure 5.5 appeared massed and compressed against their respective upper and

lower support bounds.

5.3.3 Effect of CTVI SDE solver step size on CTVI-estimated θ posteriors

When comparing marginal $q(\theta)$ densities approximated by CTVI using Euler-Maruyama SDE solver step sizes of $\mathrm{d}t = 0.05$ and $\mathrm{d}t = 0.2$ (Figure 5.6), we found them to be similar in shape and location. The $\mathrm{d}t = 0.05$ and $\mathrm{d}t = 0.2$ runs appear to randomly exchange slight advantages in θ estimation accuracy and identification across the various subplots. Despite its greater numerical computation expense, use of the lower step size did not correspond to consistent or clearly organizable visual advantages in posterior location and certainty.

5.3.4 Effect of CO_2 information exclusion from y on CTVI-estimated θ posteriors

When comparing the marginal $q(\theta)$ densities of $\mathrm{d}t = 0.05$ CTVI runs conditioned with and without CO_2 measurements in y (Figure 5.7), we observed that marginal posterior uncertainty increased for some of the θ. $u_{Q,\mathrm{ref}}$, which is involved in the determination of SAWB-ECA-SS CO_2 respiration as stated in equation (5.5), was one of the θ that saw marginal posterior widening with CO_2-knockout. Some posteriors θ not directly involved with CO_2 computation also saw substantial widening in $V_{\mathrm{DE},\mathrm{ref}}$, $Ea_{V_{\mathrm{DE}}}$, and $Ea_{V_{\mathrm{UE}}}$.

Posterior widening from CO_2 exclusion was not unanimously observed across all θ. The "no CO_2" θ posteriors of K_{DE} and r_L surprisingly both narrowed and became more closely located to the true θ to show improved θ identification. That being said, more marginal $q(\theta)$ still widened (10) than narrowed (6) without CO_2 information based on density peak heights, where higher peaks indicate narrower and more constrained posteriors due to conservation of probability area.

Moreover, the density narrowing did not necessarily correspond to more accurate θ identification; the "no CO_2" $V_{\mathrm{UE,ref}}$ and r_E $q(\theta)$ density peaks shifted further away from the truth. "No CO_2" density narrowing did coincide with the inching of peaks toward true θ in the K_{DE}, r_L, and r_M subplots. In the case of s_{DOC}, the "no CO_2" $q(\theta)$ widened and became more uncertain, but its density peak and apparent mean became more accurately located.

5.4 Discussion

5.4.1 The CTVI algorithm appeared operational as a stochastic gradient optimization approach

Considering the $-\mathcal{L}$ trajectories as diagnostic indicators, Figure 5.1 suggests that our implementation of the CTVI algorithm exhibits appropriate gradient descent behavior and is at least somewhat operational for stochastic gradient optimization purposes. The decreases in $-\mathcal{L}$ value and slope magnitudes suggests a successful approach in both $\mathrm{d}t = 0.05$ and $\mathrm{d}t = 0.05$ runs toward ELBO convergence and triangulation of a ϕ regime that is at least a local minimum for KL divergence.

That the $\mathrm{d}t = 0.05$ run generally tracks at a lower $-\mathcal{L}$ than the $\mathrm{d}t = 0.2$ run to signal that reducing SDE solver step size allows the CTVI algorithm to minimize the KL divergence between the approximate and true posteriors. This is expected, as the obvious consequence of dropping solver step size is the reduction in global accumulation of state approximation error during the integration of states through a mini-batch simulation time span (Maruyama, 1955). Improvement of integration error of course mitigates a barrier to accurate approximation and recapitulation of the true latent states by the untrained model.

However, given the large relative difference between 0.05 and 0.2, we were surprised that the

$-\mathcal{L}$ trajectories tracked so closely with each other. It was our first sign that CTVI could tolerate larger solver step sizes for valuable savings in compute time and cost without showing much degradation in approximation accuracy.

5.4.2 SDE-CTVI compared agreeably with ODE-NUTS at fitting states and identifying true θ

Even at a surface level, VI and NUTS deviate substantially in their assumptions and methodology starting from the operations occuring in each of their iterations. VI approximates posterior densities by iteratively optimizing a predefined posterior family with an objective function, while NUTS, as an MCMC method, handles approximation non-parametrically with the iterative appending of samples to ergodic chains. With their drastic differences, no singular approach presently exists to neatly and properly quantify their goodness-of-fit in a Bayesian mode that retains information about goodness-of-fit distributions and uncertainty. As such, we compared the results of the CTVI and ODE-NUTS approaches in a visual qualitative manner in this study to illustrate the validity of the CTVI algorithm, leaving more quantitative comparisons for future work on Bayesian goodness-of-fit metrics development.

Starting with the fit distributions in Figure 5.2, we see qualitative support for the validity and effective function of the CTVI algorithm. The CTVI-optimized SAWB-ECA-SS state trajectory distributions and means compare agreeably with the trends in the AWB-ECA trajectories learned by Stan's NUTS engine, which is trusted as a "gold standard" benchmark of reliability for comparison of unproven inference methods (Nemeth and Fearnhead, 2021). However, the CTVI run was accomplished with substantially less temporal cost, only taking about 3 hours of wall-clock computation with the leveraging of a GPU. The ODE-NUTS run took about 8 hours on a CPU.

The superiority of using a stochastic model inference approach that can flexibly adapt initial conditions to fit noisy data randomized by data-generating and observation stochasticity is evident

in the Figure 5.2 SOC panel. ODE-NUTS was unable to adjust from SOC outlier observations and a high SOC initial condition assumed when using y_0 as initial ODE values x_0. Hence, ODE-NUTS overshot the SOC trajectory over the large concentrations of the observation interval. The SDE-CTVI framework was able to accommodate a preliminary proposal of initial conditions inconsistent with the true latent data-generating process and settle its SOC trajectory more centrally within the observation interval.

Both the ODE-NUTS and SDE-CTVI state trajectory distributions depict visibly increasing DOC, MBC, and EEC posterior uncertainty at the end of the time series (Figure 5.3) in comparison to the beginning (Figure 5.4). In contrast, we observe little change in state variance for the ODE-NUTS SOC state trajectory, while we see a marked uncertainty decrease for SDE-CTVI SOC. We feel that these observations could be explained in a straightforward manner by the changing state magnitudes and dynamical system structures of SAWB-ECA-SS and AWB-ECA. State variance sourced from θ posterior uncertainty and solver approximation error, which exists in both deterministic and stochastic models, would naturally become easier to see at larger state magnitudes. Indeed, in the fitting of y, the DOC, MBC, and EEC states of SAWB-ECA-SS and AWB-ECA drastically increase many times over from their early values through the simulation time span. Meanwhile, to account for the lack of palpable uncertainty change in AWB-ECA SOC, we note that the SOC trajectories vary less dramatically with gradual decay by approximately half of their starting values by the end of the simulation. The standout decrease in SAWB-ECA-SS SOC posterior variance then can be attributed to the stochastic system's "state-scaling" β diffusion structure set forth in equation (5.2). The $s_{i\in\{SOC,\,DOC,\,MBC,\,EEC\}}$ θ scale the extent of diffusion noise along the state magnitudes such that a drop in SOC ensures a corresponding decline in SAWB-ECA-SS SOC diffusion noise.

On the subject of noise dynamics, we see that the SAWB-ECA-SS trajectories qualitatively appear fairly deterministic and look similar to the smooth curves of AWB-ECA. This reflects the low true data-generating and CTVI-estimated $s_{i\in\{SOC,\,DOC,\,MBC,\,EEC\}}$ state-scaling noise θ. Having

observed SAWB-ECA-SS instability at higher θ levels as described in section 5.2.2, we took abundant caution to reduce noise to a level that would, with certainty, prevent fluctuation of SAWB-ECA-SS state values into the lower bound across $T = 2000$. Admittedly, too much caution was potentially taken, as the low noise settings likely inhibited visual discrimination between the SDE-CTVI and ODE-NUTS fit trajectories, especially for DOC, MBC, EEC, and CO_2. If synthetic y sampled from a noisier data-generating process were used, we hypothesize that this would have demonstrated a clearer state-fitting advantage for the SDE-CTVI method. It is also possible that higher noise y would have facilitated more obvious qualitative differentiation between SAWB-ECA-SS state trajectories estimated using different Euler-Maruyama step sizes. We did not plot the $\mathrm{d}t = 0.2$ CTVI-estimated trajectories because they were essentially on top of the $\mathrm{d}t = 0.05$ results

Superiority of SDE-CTVI approach over the ODE-NUTS framework is more apparent when looking at the comparison of marginal θ posteriors in Figure 5.5. With respect to the identification of true θ, CTVI was able to precisely align SAWB-ECA-SS posterior density peaks with the truth in arguably six cases, $u_{Q,\mathrm{ref}}$, K_{DE}, K_{UE}, $V_{\mathrm{DE},\mathrm{ref}}$, $V_{\mathrm{UE},\mathrm{ref}}$, and $Ea_{V_{\mathrm{DE}}}$. One could make a case of effective CTVI-estimated SAWB-ECA-SS $q(\theta)$ mean alignment for Q as well. Meanwhile, for the ODE-NUTS framework, good posterior peak alignment to true θ was only demonstrated for K_{DE}, $V_{\mathrm{DE},\mathrm{ref}}$, and r_L. And for K_{DE} and $V_{\mathrm{DE},\mathrm{ref}}$, the NUTS-estimated posteriors were still surpassed by their CTVI-estimated counterparts in informed narrowing about the true θ. The ODE-NUTS framework only qualitatively outperformed SDE-CTVI at θ identification in combined posterior certainty and mean location evaluations in one subplot corresponding to Q. SDE-CTVI was arguably superior head-to-head in θ identification and estimation accuracy in the other 11 subplots.

Though lacking two-sided truncation and being only bounded above 0, the marginal SAWB-ECA-SS IAF-family posteriors never showed extreme density consolidation situated either too closely against their lower bounds or too far away to the unbounded right of the true θ values. In contrast,

two AWB-ECA posteriors, those of Q and r_L, evinced compressed density consolidation against support bounds in ODE-NUTS' efforts to accommodate and compensate for noisy and outlier observations in y with deterministic-only dynamics. Without tending to the ODE-NUTS inference with two-sided truncation, it is likely that the Q and r_L posteriors would have consolidated at nonsense value ranges that could pull the rest of the system θ into biologically untenable regimes. That θ truncation could be relaxed in this trial of the CTVI method is another indicator of the algorithm's superiority over the ODE-NUTS approach at tolerating stochastic data and identifying true θ. These results suggest the potential applicability of CTVI to the data assimilation of SBMs with even noisier empirical data sets.

5.4.3 CTVI SDE solver step size did not clearly improve posterior identification of SAWB-ECA-SS θ

Having observed a lack of visual separation between the $dt = 0.2$ and $dt = 0.05$ CTVI-estimated state trajectories, we wanted to discern if reduced integration approximation error sourced from lower step sizes affected or benefitted posterior θ estimation and identification. SAWB-ECA-SS was parameterized with a relatively high θ count (see Table 5.2) and its equations (5.1) include terms that encompass multiple elements which can compensate for each other such as $a_{MSA} \cdot r_M \cdot M$. Because of the high model equifinality to go along with the possible randomness of stochastic gradient descent, we thought it possible that $dt = 0.2$ and $dt = 0.05$ marginal $q(\theta)$ would diverge in an informative manner.

Ultimately, no clear advantages in $q(\theta)$ approximation could be discerned for the $dt = 0.05$ run despite its much heavier compute load (Figure 5.6). Indeed, we were hard-pressed to ascertain any significant or consequential differences between the two posterior sets. The $dt = 0.05$ and $dt = 0.2$ runs seemed to be similarly informed and constrained, randomly exchanging slight advantages in $q(\theta)$ peak locations relative to true θ values across the subplots.

It is possible that probabilistic accounting of model output uncertainty and inference on noisy and *randomly*[2] inaccurately observed data places less stringent demands on solver accuracy. Relative solver inaccuracy can be compensated by large enough observation and diffusion noise. And as is, proposed SBM dynamics are thought to remain far off from faithful representation of empirical data-generating mechanics (Bradford et al., 2016); we would expect the proposal of more more truthful model mechanics be more critical for accurate θ estimation than modest differences in solver accuracy. Following from this, we consider that the substitution of Euler-Maruyama with more numerically complicated and compute-intensive SDE solvers that are theoretically more accurate may not be useful if Euler-Maruyama suffices for managing system instability and nonlinearity.

Of course, much more testing of CTVI with different models, data sets, and noise levels needs to be carried out before robust conclusions about the implications of solver step size on CTVI θ estimation can be made. Nonetheless, for future, we recommend experimenting with higher solver step sizes in future SBM inferences using CTVI for greater compute efficiency on the basis of the Figure 5.6 results.

5.4.4 CO_2 information generally improves SAWB-ECA-SS θ identifiability

The presence of CO_2 observations in y were not as overwhelmingly informative for SAWB-ECA-SS $q(\theta)$ estimation as it was for SCON in Chapter 4. CO_2 knockout resulted in reduced posterior constraint in seven of the nine SCON-C non-diffusion θ[3] (Figure 4.11). In comparison, CO_2 knockout widened and lessened uncertainty in six of the 12 SAWB-ECA-SS non-diffusion θ posteriors (Figure 5.7).

[2]We emphasize "randomly" here because we could see a case where regular, systemic bias in data collection such as consistent overmeasurement has a different effect favoring solver accuracy.

[3]We set discussion of the diffusion θ aside for now, as the flow VI approach was limited to persistent overestimation of diffusion θ due to neural network approximation error in its Xie et al. (2022) implementation.

The diminished influence of CO_2 measurements for calibrating SAWB-ECA-SS θ seems reasonable given the number of θ involved in each system's CO_2 computation, six for SCON-C by equation (4.9) and three for SAWB-ECA-SS by equation (5.5). Certainly, CO_2 information still helped more than it harmed, with 11 of the 16 SAWB-ECA-SS θ posteriors[4] showing reduced distance of peaks to true θ or tighter density consolidation about the truth for better identifiability with CO_2 inclusion. Needless to say, we still recommend the inclusion of CO_2 observations in y in the data assimilation of any SBM for whom a CO_2 respiration equation has been defined. Nonetheless, it was surprising to observe that four of the 16 total "no CO_2" θ posteriors corresponding to K_{DE}, r_M, r_L, and s_{DOC} became arguably better identified. We attribute this to the interaction between model equifinality and black-box randomness in PyTorch's intricate Autograd automatic differentiation engine (Paszke et al., 2019) that manages gradient computations.

5.4.5 Conclusions and future work

Through visualizations of ELBO trajectories (Figure 5.1), state fits (Figure 5.2), and approximate posterior estimates (Figures 5.5, 5.6, and 5.7), results of this cursory study tentatively suggest applicability of the CTVI algorithm as a viable technique for assimilating SBMs with soil pool and CO_2 efflux observations and estimating compatible SBM parameter values. The method shows promise as a potential inference method option to be integrated into SBM testbed frameworks (Wieder et al., 2018) that aim to validate, compare, and subsequently select models for further refinement and development.

In light of the existence of methods that approximate SDEs into probabilitistic discrete state space models to support vectorized latent trajectory sampling (Ryder et al., 2021; Sujono et al., 2022), we harbor some reservations about the ability of the CTVI method relating to scale to higher simulation time spans associated with long-term empirical soil warming experiments (Melillo et al., 2017; Wood et al., 2019) due to the computational expense of sequential SDE

[4]We list the 11 here as $u_{Q,\mathrm{ref}}$, Q, a_{MSA}, K_{UE}, $V_{\mathrm{DE,ref}}$, $V_{\mathrm{UE,ref}}$, $Ea_{V_{\mathrm{DE}}}$, $Ea_{V_{\mathrm{UE}}}$, r_E, s_{MBC}, and s_{EEC}.

solvers. However, we note that the incorporation of mini-batching into CTVI helped address our concerns. For a not insubstantial time span totaling $T = 2000$ hours, mini-batching respectively facilitated encouraging wall-clock run times of about 1 and 3 hours for the $\mathrm{d}t = 0.05$ and $\mathrm{d}t = 0.2$ SAWB-ECA-SS inferences. The absence of a clear qualitative advantage in estimation accuracy for the $\mathrm{d}t = 0.05$ run over its $\mathrm{d}t = 0.2$ counterpart (Figures 5.1 and 5.6) promotes the setting of higher solver time steps as a supplemental means of maximizing CTVI efficiency.

We note in favor of the CTVI algorithm that the $\mathrm{d}t = 0.05$ and $\mathrm{d}t = 0.2$ inferences actually required much shorter durations than the Chapter 4 SCON flow VI inferences, which tended operate on the scale of days with a comparable amount of training iterations[5]. And in addition to run times, two more points in favor of CTVI over flow VI in our experience are:

- that CTVI could navigate a nonlinear SBM (when flow VI could not) and
- that it was not guaranteed to overestimate diffusion θ due to accumulating approximation error, as highlighted by a juxtaposition between Figures 4.9 and 5.6.

CTVI ought to be clearly preferenced over neural moving average flow VI (or at least the Chapter 4 flow VI implementation) if accurate estimation of diffusion θ is a research priority.

This results of this study evokes and connects to a number of other lines of inquiry, exploration, and testing. We will subsequently mention a few those.

First, we stress that CTVI needs to be tested through inference conditioning on much noisier data to have a better shot at teasing out potential effects of algorithm settings such as solver step size. Use of noisier data may also induce more blatant and informative separation between CTVI-estimated results and those estimated via the deterministic inference frameworks traditionally used in biogeochemistry.

[5]Albeit, those flow VI inferences had to contend with a simulation time span of $T = 5000$.

While we can sample synthetic data sets simulated with more diffusion and observation noise, the Holy Grail objective is to be able to assess and juxtapose inference algorithms with a regularly observed, multivariate data set that is dense enough in observations to sustain partitioning into model training and testing portions. Ideally, this data could be sourced from a single experiment, as mixing data from separate experiments, even if located nearby, could obfuscate and garble empirical data-generating processes and distort inference results away from truth. Also, soil measurements in this data set could ideally be associated with pool states of various SBMs with few necessary transformations.

Attainment of such a Holy Grail of a data product obviously remains a distant work in progress and would require a hefty and extended ask of temporal, financial, and personnel resources to satisfy data collection, cleaning, and storage best practices. The cost and effort of systematic soil data collection sources data scarcity issues in the field of biogeochemistry (Moulatlet et al., 2017; Zuquim et al., 2019). Soil experiments running at sites like Harvard Forest (Melillo et al., 2017) and the Luquillo Experimental Forest (Wood et al., 2019) are being designed and adjusted as their funding and workforce allows to incrementally address the demand for inference-ready data products. This progress is propitious, as the construction of standard benchmarking data sets and their integration into modular SBM testbeds enables rigorous and methodical A/B testing of inference algorithms including CTVI to inform the toolbox of the biogeochemical modeler.

Next, we promote the trialing of CTVI to assimilate naive stochastic parameterizations of more elaborate, higher state-dimensioned nonlinear SBMs such as CORPSE (Sulman et al., 2014) and MIMICS (Wieder, Grandy, Kallenbach, and Bonan, 2014; Wieder, Grandy, Kallenbach, Taylor, et al., 2015) starting with synthetic data to be able to identify discrepancies between algorithm results and latent data-generating processes with certainty. It could also be interesting to use CTVI, along with other inference algorithms, to explore less naive, mechanistic stochastic conversions of lower state-dimensioned nonlinear SBMs like AWB and the microbial enzyme-mediated decomposition (MEND) system (G. Wang et al., 2013) that represent the underlying stoichiometry of

the chemical reaction networks assumed by the models (Golightly and Wilkinson, 2011).

Finally, we encourage the continued development and standardization of goodness-of-fit metrics for VI approximations that build on the work of Dhaka et al. (2020), Giordano et al. (2018), and Yao et al. (2018). From this foundation, we see the emergence of diagnostics based on Pareto-smoothed importance sampling (Vehtari et al., 2017) that can equitably contrast and quantify goodness-of-fit and posterior approximation error of data assimilation results achieved using MCMC and VI.

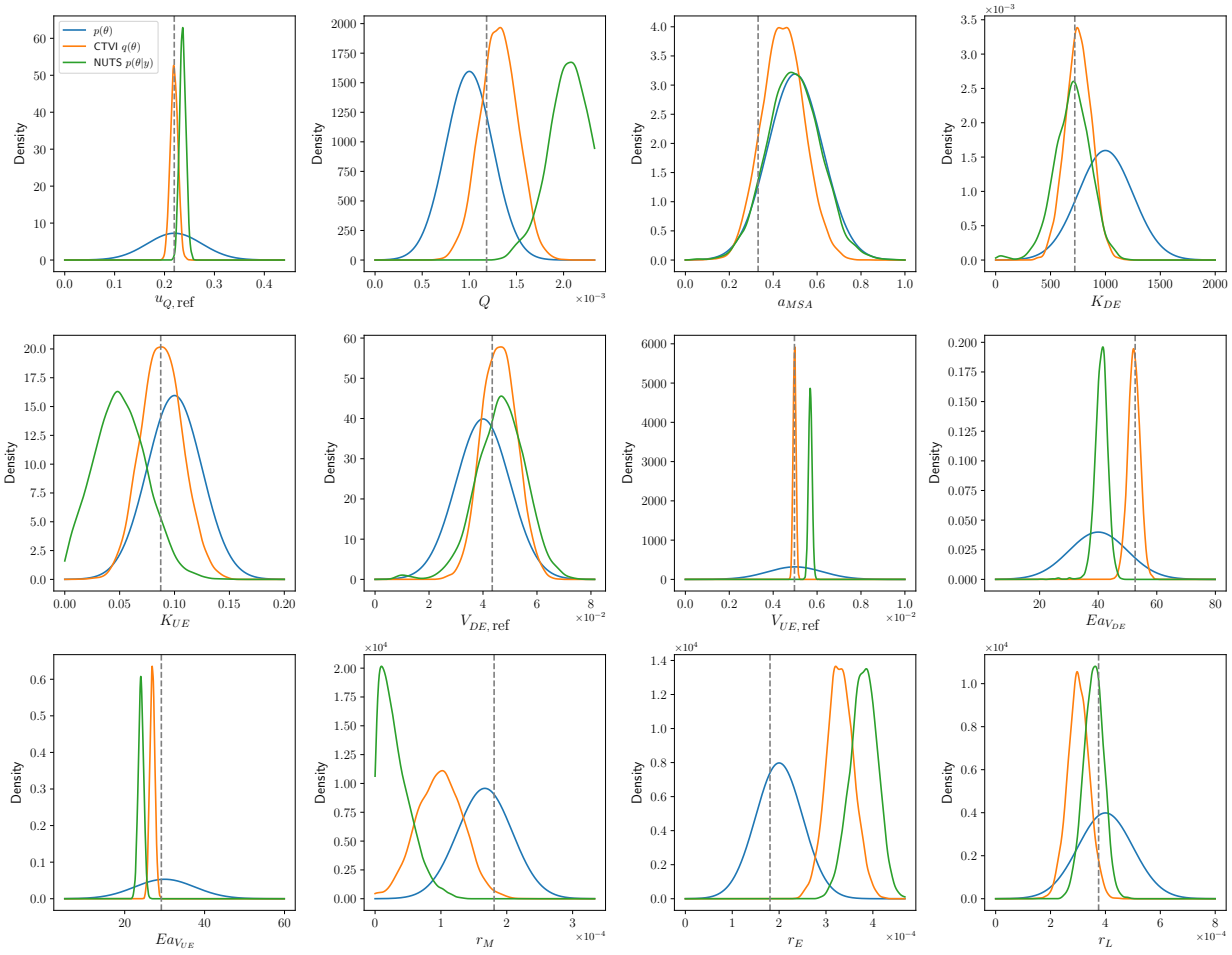

Figure 5.5: SAWB-ECA-SS data-generating densities (blue) juxtaposed against mean-field SAWB-ECA-SS approximate posterior $q(\theta)$ densities estimated by the CTVI method (orange) and mean-field AWB-ECA $p(\theta|y)$ densities estimated by NUTS (green). SAWB-ECA-SS marginal approximate posteriors corresponding to the state-scaling diffusion θ have been omitted from this figure due to a lack of deterministic AWB-ECA counterparts.

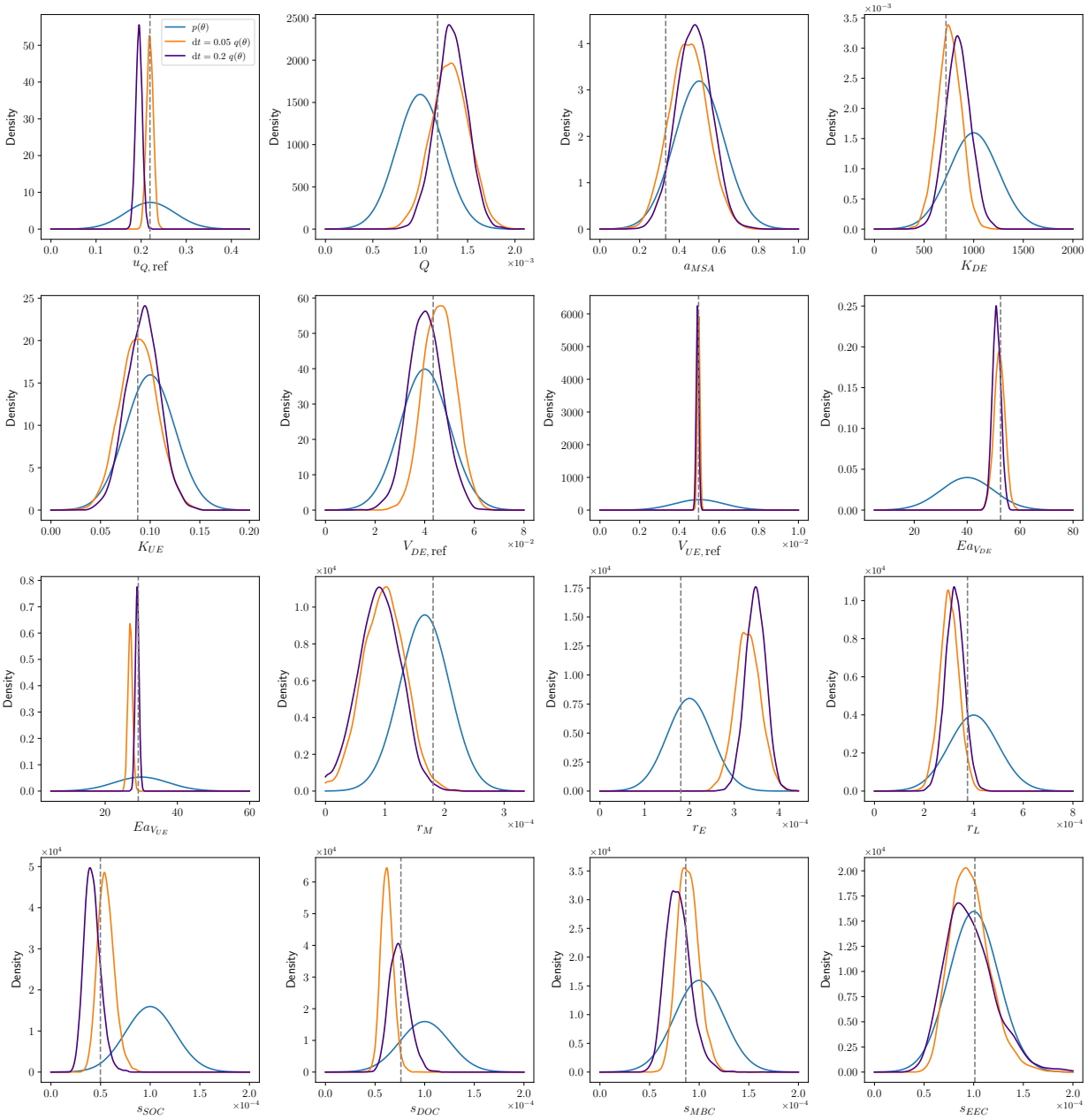

Figure 5.6: SAWB-ECA-SS data-generating densities (blue) juxtaposed against mean-field SAWB-ECA-SS approximate posterior $q(\theta)$ densities estimated using the CTVI algorithm with Euler-Maruyama SDE solver step sizes of $\mathrm{d}t = 0.05$ (orange) and $\mathrm{d}t = 0.2$ (violet).

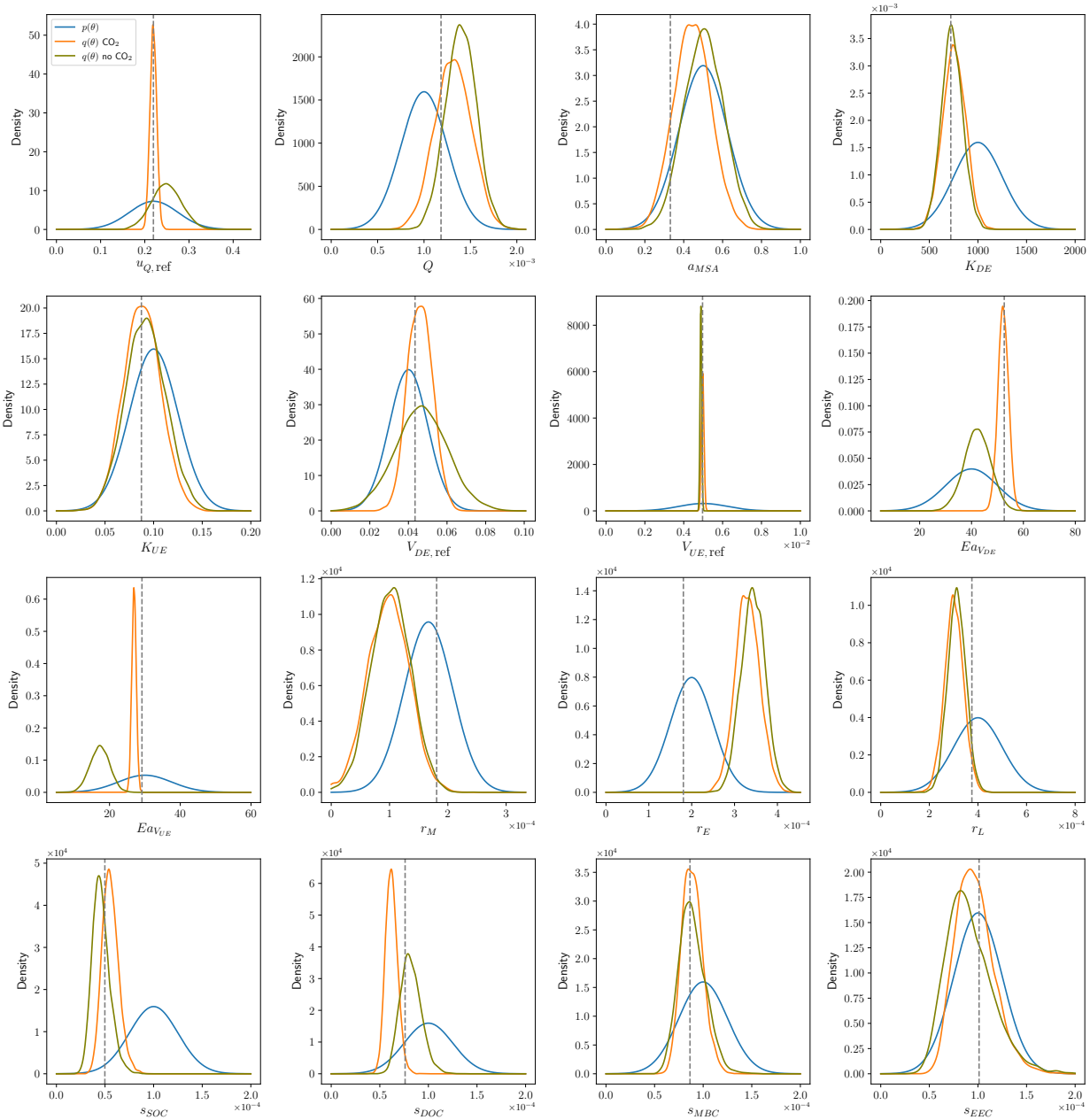

Figure 5.7: SAWB-ECA-SS data-generating densities (blue) juxtaposed against mean-field SAWB-ECA-SS approximate posterior $q(\theta)$ densities optimized using the CTVI algorithm with (orange) and without (olive) CO_2 information.

Bibliography

Abramoff, R., Xu, X., Hartman, M., O'Brien, S., Feng, W., Davidson, E., Finzi, A., Moorhead, D., Schimel, J., Torn, M., & Mayes, M. A. (2018). The Millennial model: in search of measurable pools and transformations for modeling soil carbon in the new century. *Biogeochemistry*, *137*(1), 51–71. https://doi.org/10.1007/s10533-017-0409-7

Abs, E., Leman, H., & Ferrière, R. (2020). A multi-scale eco-evolutionary model of cooperation reveals how microbial adaptation influences soil decomposition. *Communications Biology*, *3*(1). https://doi.org/10.1038/s42003-020-01198-4

Aicher, C., Ma, Y.-A., Foti, N. J., & Fox, E. B. (2019). Stochastic Gradient MCMC for State Space Models. *SIAM Journal on Mathematics of Data Science*, *1*(3). https://doi.org/10.1137/18M1214780

Aitkenhead, M., & Coull, M. (2016). Mapping soil carbon stocks across scotland using a neural network model. *Geoderma*, *262*, 187–198. https://doi.org/10.1016/j.geoderma.2015.08.034

Alahmadi, A. A., Flegg, J. A., Cochrane, D. G., Drovandi, C. C., & Keith, J. M. (2020). A comparison of approximate versus exact techniques for Bayesian parameter inference in nonlinear ordinary differential equation models. *Royal Society Open Science*, *7*(3), 191315. https://doi.org/10.1098/rsos.191315

Allison, S. D. (2005). Cheaters, diffusion and nutrients constrain decomposition by microbial enzymes in spatially structured environments. *Ecology Letters*, *8*(6), 626–635. https://doi.org/10.1111/j.1461-0248.2005.00756.x

Allison, S. D. (2012). A trait-based approach for modelling microbial litter decomposition. *Ecology Letters*, *15*(9), 1058–1070. https://doi.org/10.1111/j.1461-0248.2012.01807.x

Allison, S. D. (2014). Modeling adaptation of carbon use efficiency in microbial communities. *Frontiers in Microbiology*, *5*. https://doi.org/10.3389/fmicb.2014.00571

Allison, S. D. (2017). Building Predictive Models for Diverse Microbial Communities in Soil. *Microbial Biomass* (pp. 141–166). https://doi.org/10.1142/9781786341310_0006

Allison, S. D., Wallenstein, M. D., & Bradford, M. A. (2010). Soil-carbon response to warming dependent on microbial physiology. *Nature Geoscience*, *3*(5). https://doi.org/10.1038/ngeo846

Alster, C. J., von Fischer, J. C., Allison, S. D., & Treseder, K. K. (2020). Embracing a new paradigm for temperature sensitivity of soil microbes. *Global Change Biology*, *26*(6), 3221–3229. https://doi.org/10.1111/gcb.15053

Anaconda Software Distribution (Version 2-2.4.0). (2020). Anaconda Inc. https://docs.anaconda.com/

Anderson, T.-H., & Domsch, K. H. (1989). Ratios of microbial biomass carbon to total organic carbon in arable soils. *Soil Biology and Biochemistry*, *21*(4). https://doi.org/10.1016/0038-0717(89)90117-X

Archer, E., Park, I. M., Buesing, L., Cunningham, J., & Paninski, L. (2015). Black box variational inference for state space models. https://doi.org/10.48550/ARXIV.1511.07367

Armstrong McKay, D. I., Staal, A., Abrams, J. F., Winkelmann, R., Sakschewski, B., Loriani, S., Fetzer, I., Cornell, S. E., Rockström, J., & Lenton, T. M. (2022). Exceeding 1.5 °C global warming could trigger multiple climate tipping points. *Science*, *377*(6611). https://doi.org/10.1126/science.abn7950

Arora, V. K., Katavouta, A., Williams, R. G., Jones, C. D., Brovkin, V., Friedlingstein, P., Schwinger, J., Bopp, L., Boucher, O., Cadule, P., Chamberlain, M. A., Christian, J. R., Delire, C., Fisher, R. A., Hajima, T., Ilyina, T., Joetzjer, E., Kawamiya, M., Koven, C. D., ... Ziehn, T. (2020). Carbon–concentration and carbon–climate feedbacks in CMIP6

models and their comparison to CMIP5 models. *Biogeosciences*, *17*(16), 4173–4222. https://doi.org/10.5194/bg-17-4173-2020

Arrhenius, S. (1889). Über die Dissociationswärme und den Einfluss der Temperatur auf den Dissociationsgrad der Elektrolyte. *Zeitschrift für Physikalische Chemie*, *4U*(1). https://doi.org/10.1515/zpch-1889-0408

Bahram, M., Hildebrand, F., Forslund, S. K., Anderson, J. L., Soudzilovskaia, N. A., Bodegom, P. M., Bengtsson-Palme, J., Anslan, S., Coelho, L. P., Harend, H., Huerta-Cepas, J., Medema, M. H., Maltz, M. R., Mundra, S., Olsson, P. A., Pent, M., Põlme, S., Sunagawa, S., Ryberg, M., ... Bork, P. (2018). Structure and function of the global topsoil microbiome. *Nature*, *560*(7717), 233–237. https://doi.org/10.1038/s41586-018-0386-6

Beillouin, D., Cardinael, R., Berre, D., Boyer, A., Corbeels, M., Fallot, A., Feder, F., & Demenois, J. (2022). A global overview of studies about land management, land-use change, and climate change effects on soil organic carbon. *Global Change Biology*, *28*(4). https://doi.org/10.1111/gcb.15998

Berardi, D., Brzostek, E., Blanc-Betes, E., Davison, B., DeLucia, E. H., Hartman, M. D., Kent, J., Parton, W. J., Saha, D., & Hudiburg, T. W. (2020). 21st-century biogeochemical modeling: Challenges for Century-based models and where do we go from here? *GCB Bioenergy*, *12*(10), 774–788. https://doi.org/10.1111/gcbb.12730

Bernal, B., McKinley, D. C., Hungate, B. A., White, P. M., Mozdzer, T. J., & Megonigal, J. P. (2016). Limits to soil carbon stability; Deep, ancient soil carbon decomposition stimulated by new labile organic inputs. *Soil Biology and Biochemistry*, *98*, 85–94. https://doi.org/10.1016/j.soilbio.2016.04.007

Berrang-Ford, L., Siders, A. R., Lesnikowski, A., Fischer, A. P., Callaghan, M. W., Haddaway, N. R., Mach, K. J., Araos, M., Shah, M. A. R., Wannewitz, M., Doshi, D., Leiter, T., Matavel, C., Musah-Surugu, J. I., Wong-Parodi, G., Antwi-Agyei, P., Ajibade, I., Chauhan, N., Kakenmaster, W., ... Abu, T. Z. (2021). A systematic global stocktake of evidence

on human adaptation to climate change. *Nature Climate Change*, *11*(11). https://doi.org/10.1038/s41558-021-01170-y

Beskos, A., Pillai, N., Roberts, G., Sanz-Serna, J. M., & Stuart, A. (2013). Optimal tuning of the hybrid Monte Carlo algorithm. *Bernoulli*, *19*(5 A). https://doi.org/10.3150/12-BEJ414

Betancourt, M. (2016). Diagnosing Suboptimal Cotangent Disintegrations in Hamiltonian Monte Carlo. http://arxiv.org/abs/1604.00695

Betancourt, M. (2017). A Conceptual Introduction to Hamiltonian Monte Carlo. http://arxiv.org/abs/1701.02434

Betancourt, M., & Girolami, M. (2015). Hamiltonian Monte Carlo for Hierarchical Models. *Current Trends in Bayesian Methodology with Applications*. https://doi.org/10.1201/b18502-5

Bjorck, N., Gomes, C. P., Selman, B., & Weinberger, K. Q. (2018). Understanding Batch Normalization. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems*. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2018/file/36072923bfc3cf47745d704feb489480-Paper.pdf

Blagodatsky, S., Blagodatskaya, E., Yuyukina, T., & Kuzyakov, Y. (2010). Model of apparent and real priming effects: Linking microbial activity with soil organic matter decomposition. *Soil Biology and Biochemistry*, *42*(8), 1275–1283. https://doi.org/10.1016/j.soilbio.2010.04.005

Blankinship, J. C., Berhe, A. A., Crow, S. E., Druhan, J. L., Heckman, K. A., Keiluweit, M., Lawrence, C. R., Marín-Spiotta, E., Plante, A. F., Rasmussen, C., Schädel, C., Schimel, J. P., Sierra, C. A., Thompson, A., Wagai, R., & Wieder, W. R. (2018). Improving understanding of soil organic matter dynamics by triangulating theories, measurements, and models. *Biogeochemistry*, *140*(1), 1–13. https://doi.org/10.1007/s10533-018-0478-2

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, *112*(518). https://doi.org/10.1080/01621459.2017.1285773

Bosatta, E. B., & Ågren, G. I. (1985). Theoretical analysis of decomposition of heterogeneous substrates. *Soil Biology and Biochemistry*, *17*(5), 601–610. https://doi.org/10.1016/0038-0717(85)90035-5

Bosatta, E. B., & Ågren, G. I. (1999). Soil organic matter quality interpreted thermodynamically. *Soil Biology and Biochemistry*, *31*(13), 1889–1891. https://doi.org/10.1016/S0038-0717(99)00105-4

Botu, V., Batra, R., Chapman, J., & Ramprasad, R. (2017). Machine Learning Force Fields: Construction, Validation, and Outlook. *The Journal of Physical Chemistry C*, *121*(1), 511–522. https://doi.org/10.1021/acs.jpcc.6b10908

Bradford, M. A., & Crowther, T. W. (2013). Carbon use efficiency and storage in terrestrial ecosystems. *New Phytologist*, *199*(1). https://doi.org/10.1111/nph.12334

Bradford, M. A., Wieder, W. R., Bonan, G. B., Fierer, N., Raymond, P. A., & Crowther, T. W. (2016). Managing uncertainty in soil carbon feedbacks to climate change. *Nature Climate Change*, *6*(8). https://doi.org/10.1038/nclimate3071

Bradford, M. A., Wood, S. A., Addicott, E. T., Fenichel, E. P., Fields, N., González-Rivero, J., Jevon, F. V., Maynard, D. S., Oldfield, E. E., Polussa, A., Ward, E. B., & Wieder, W. R. (2021). Quantifying microbial control of soil organic matter dynamics at macrosystem scales. *Biogeochemistry*, *156*(1). https://doi.org/10.1007/s10533-021-00789-5

Brosse, N., Durmus, A., & Moulines, E. (2018). The promises and pitfalls of Stochastic Gradient Langevin Dynamics. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems*. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2018/file/335cd1b90bfa4ee70b39d08a4ae0cf2d-Paper.pdf

Browning, A. P., Warne, D. J., Burrage, K., Baker, R. E., & Simpson, M. J. (2020). Identifiability analysis for stochastic differential equation models in systems biology. *Journal of The Royal Society Interface*, *17*(173). https://doi.org/10.1098/rsif.2020.0652

Buchkowski, R. W., Bradford, M. A., Grandy, A. S., Schmitz, O. J., & Wieder, W. R. (2017). Applying population and community ecology theory to advance understanding of belowground biogeochemistry. *Ecology Letters*, *20*(2), 231–245. https://doi.org/10.1111/ele.12712

Burkardt, J. (2014). The Truncated Normal Distribution. https://people.sc.fsu.edu/~jburkardt/presentations/truncated_normal.pdf

Bürkner, P.-C., Gabry, J., & Vehtari, A. (2020). Approximate leave-future-out cross-validation for Bayesian time series models. *Journal of Statistical Computation and Simulation*, *90*(14). https://doi.org/10.1080/00949655.2020.1783262

Burns, R. G., DeForest, J. L., Marxsen, J., Sinsabaugh, R. L., Stromberger, M. E., Wallenstein, M. D., Weintraub, M. N., & Zoppini, A. (2013). Soil enzymes in a changing environment: Current knowledge and future directions. *Soil Biology and Biochemistry*, *58*, 216–234. https://doi.org/10.1016/j.soilbio.2012.11.009

Burton, L., Jayachandran, K., & Bhansali, S. (2020). Review - The "Real-Time" Revolution for In situ Soil Nutrient Sensing. *Journal of The Electrochemical Society*, *167*(3), 037569. https://doi.org/10.1149/1945-7111/ab6f5d

Calderhead, B., Girolami, M., & Lawrence, N. D. (2009). Accelerating Bayesian Inference over Nonlinear Differential Equations with Gaussian Processes. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in Neural Information Processing Systems 21*. Curran Associates, Inc. http://papers.nips.cc/paper/3497-accelerating-bayesian-inference-over-nonlinear-differential-equations-with-gaussian-processes.pdf

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1). https://doi.org/10.18637/jss.v076.i01

Cash, J. R., & Karp, A. H. (1990). A Variable Order Runge-Kutta Method for Initial Value Problems with Rapidly Varying Right-Hand Sides. *ACM Transactions on Mathematical Software*, *16*(3), 201–222. https://doi.org/10.1145/79505.79507

Cavicchioli, R., Ripple, W. J., Timmis, K. N., Azam, F., Bakken, L. R., Baylis, M., Behrenfeld, M. J., Boetius, A., Boyd, P. W., Classen, A. T., Crowther, T. W., Danovaro, R., Foreman, C. M., Huisman, J., Hutchins, D. A., Jansson, J. K., Karl, D. M., Koskella, B., Mark Welch, D. B., ... Webster, N. S. (2019). Scientists' warning to humanity: Microorganisms and climate change. *Nature Reviews Microbiology*, *17*(9), 569–586. https://doi.org/10.1038/s41579-019-0222-5

Chen, T., Fox, E., & Guestrin, C. (2014). Stochastic Gradient Hamiltonian Monte Carlo. In E. P. Xing & T. Jebara (Eds.), *Proceedings of the 31st International Conference on Machine Learning*. PMLR. https://proceedings.mlr.press/v32/cheni14.html

Christensen, O. F., Roberts, G. O., & Sköld, M. (2006). Robust Markov chain Monte Carlo Methods for Spatial Generalized Linear Mixed Models. *Journal of Computational and Graphical Statistics*, *15*(1), 1–17. https://doi.org/10.1198/106186006X100470

Christensen, R., Johnson, W., Branscum, A., & Hanson, T. E. (2010). *Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians* (1st ed.). Taylor & Francis.

Crow, S. E., Swanston, C. W., Lajtha, K., Brooks, J. R., & Keirstead, H. (2007). Density fractionation of forest soils: Methodological questions and interpretation of incubation results and turnover time in an ecosystem context. *Biogeochemistry*, *85*(1), 69–90. https://doi.org/10.1007/s10533-007-9100-8

Crowther, T. W., Todd-Brown, K. E., Rowe, C. W., Wieder, W. R., Carey, J. C., MacHmuller, M. B., Snoek, B. L., Fang, S., Zhou, G., Allison, S. D., Blair, J. M., Bridgham, S. D., Burton, A. J., Carrillo, Y., Reich, P. B., Clark, J. S., Classen, A. T., Dijkstra, F. A., Elberling, B., ... Bradford, M. A. (2016). Quantifying global soil carbon losses in response to warming. *Nature*, *540*(7631). https://doi.org/10.1038/nature20150

Csilléry, K., Blum, M. G., Gaggiotti, O. E., & François, O. (2010). Approximate Bayesian Computation (ABC) in practice. *Trends in Ecology & Evolution*, *25*(7), 410–418. https://doi.org/10.1016/j.tree.2010.04.001

Curtiss, C. F., & Hirschfelder, J. O. (1952). Integration of Stiff Equations. *Proceedings of the National Academy of Sciences of the United States of America*, *38*(3). https://doi.org/10.1073/pnas.38.3.235

Dao, V. H., Gunawan, D., Tran, M.-N., Kohn, R., Hawkins, G. E., & Brown, S. D. (2022). Efficient selection between hierarchical cognitive models: Cross-validation with variational Bayes. *Psychological Methods*. https://doi.org/10.1037/met0000458

Daunizeau, J. (2017). Semi-analytical approximations to statistical moments of sigmoid and softmax mappings of normal variables. https://doi.org/10.48550/ARXIV.1703.00091

Davidson, E. A., Janssens, I. A., & Luo, Y. (2006). On the variability of respiration in terrestrial ecosystems: Moving beyond Q10. *Global Change Biology*, *12*(2), 154–164. https://doi.org/10.1111/j.1365-2486.2005.01065.x

Davidson, E. A., & Janssens, I. A. (2006). Temperature sensitivity of soil carbon decomposition and feedbacks to climate change. *Nature*, *440*(7081), 165–173. https://doi.org/10.1038/nature04514

Davidson, E. A., Samanta, S., Caramori, S. S., & Savage, K. (2012). The Dual Arrhenius and Michaelis–Menten kinetics model for decomposition of soil organic matter at hourly to seasonal time scales. *Global Change Biology*, *18*(1). https://doi.org/10.1111/j.1365-2486.2011.02546.x

Dhaka, A. K., Catalina, A., Andersen, M. R., Magnusson, M., Huggins, J., & Vehtari, A. (2020). Robust, Accurate Stochastic Optimization for Variational Inference. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems* (pp. 10961–10973). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2020/file/7cac11e2f46ed46c339ec3d569853759-Paper.pdf

Dinh, L., Krueger, D., & Bengio, Y. (2015). NICE: Non-linear Independent Components Estimation. In Y. Bengio & Y. LeCun (Eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*. http://arxiv.org/abs/1410.8516

Dinh, L., Sohl-Dickstein, J., & Bengio, S. (2017). Density estimation using Real NVP. *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. https://openreview.net/forum?id=HkpbnH9lx

Dondelinger, F., Husmeier, D., Rogers, S., & Filippone, M. (2013). ODE parameter inference using adaptive gradient matching with Gaussian processes. In C. M. Carvalho & P. Ravikumar (Eds.), *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*. PMLR. http://proceedings.mlr.press/v31/dondelinger13a.html

Duan, L. L., Johndrow, J. E., & Dunson, D. B. (2018). Scaling up Data Augmentation MCMC via Calibration. *Journal of Machine Learning Research*, *19*(1).

Ebrahimi, A., & Or, D. (2016). Microbial community dynamics in soil aggregates shape biogeochemical gas fluxes from soil profiles – upscaling an aggregate biophysical model. *Global Change Biology*, *22*(9), 3141–3156. https://doi.org/10.1111/gcb.13345

Ehrlich, P. R., & Ehrlich, A. H. (2013). Can a collapse of global civilization be avoided? *Proceedings of the Royal Society B: Biological Sciences*, *280*(1754). https://doi.org/10.1098/rspb.2012.2845

Fang, C., & Moncrieff, J. B. (2005). The variation of soil microbial respiration with depth in relation to soil carbon composition. *Plant and Soil*, *268*(1). https://doi.org/10.1007/s11104-004-0278-4

Feller, C., & Bernoux, M. (2008). Historical advances in the study of global terrestrial soil organic carbon sequestration. *Waste Management*, *28*(4), 734–740. https://doi.org/10.1016/j.wasman.2007.09.022

Folse, H., & Allison, S. D. (2012). Cooperation, Competition, and Coalitions in Enzyme-Producing Microbes: Social Evolution and Nutrient Depolymerization Rates. *Frontiers in Microbiology*, *3*. https://doi.org/10.3389/fmicb.2012.00338

Fontaine, S., & Barot, S. (2005). Size and functional diversity of microbe populations control plant persistence and long-term soil carbon accumulation. *Ecology Letters*, *8*(10), 1075–1087. https://doi.org/10.1111/j.1461-0248.2005.00813.x

Friedlingstein, P., Meinshausen, M., Arora, V. K., Jones, C. D., Anav, A., Liddicoat, S. K., & Knutti, R. (2014). Uncertainties in CMIP5 Climate Projections due to Carbon Cycle Feedbacks. *Journal of Climate*, *27*(2). https://doi.org/10.1175/JCLI-D-12-00579.1

Fuchs, C. (2013). *Inference for Diffusion Processes: With Applications in Life Sciences*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-25969-2

Gabry, J., & Češnovar, R. (2021). *CmdStanR* (Version 0.4.0). https://mc-stan.org/cmdstanr/

Gelfand, A. E., & Dey, D. K. (1994). Bayesian Model Choice: Asymptotics and Exact Calculations. *Journal of the Royal Statistical Society. Series B (Methodological)*, *56*(3).

Gelfand, A. E., Dey, D. K., & Chang, H. (1992). Model determination using predictive distributions, with implementation via sampling-based methods (with discussion). In J. Bernardo, J. Berger, A. Dawid, & A. Smith (Eds.), *Bayesian Statistics 4*. Oxford University Press.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., & Rubin, D. (2013). *Bayesian Data Analysis* (3rd ed.). Taylor & Francis. https://doi.org/10.1201/b16018

Gelman, A., Goodrich, B., Gabry, J., & Vehtari, A. (2019). R-squared for Bayesian Regression Models. *The American Statistician*, *73*(3). https://doi.org/10.1080/00031305.2018.1549100

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*(6). https://doi.org/10.1007/s11222-013-9416-2

Geman, S., & Geman, D. (1987). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. In M. A. Fischler & O. Firschein (Eds.), *Readings in Computer Vision*. Morgan Kaufmann. https://doi.org/10.1016/B978-0-08-051581-6.50057-X

Georgiou, K., Abramoff, R. Z., Harte, J., Riley, W. J., & Torn, M. S. (2017). Microbial community-level regulation explains soil carbon responses to long-term litter manipulations. *Nature Communications*, *8*(1), 1223. https://doi.org/10.1038/s41467-017-01116-z

Georgiou, K., Malhotra, A., Wieder, W. R., Ennis, J. H., Hartman, M. D., Sulman, B. N., Berhe, A. A., Grandy, A. S., Kyker-Snowman, E., Lajtha, K., Moore, J. A. M., Pierson, D., &

Jackson, R. B. (2021). Divergent controls of soil organic carbon between observations and process-based models. *Biogeochemistry*. https://doi.org/10.1007/s10533-021-00819-2

German, D. P., Marcelo, K. R. B., Stone, M. M., & Allison, S. D. (2012). The Michaelis–Menten kinetics of soil extracellular enzymes in response to temperature: a cross-latitudinal study. *Global Change Biology*, *18*(4), 1468–1479. https://doi.org/10.1111/j.1365-2486.2011.02615.x

Geyer, K. M., Dijkstra, P., Sinsabaugh, R., & Frey, S. D. (2019). Clarifying the interpretation of carbon use efficiency in soil through methods comparison. *Soil Biology and Biochemistry*, *128*. https://doi.org/10.1016/j.soilbio.2018.09.036

Giordano, R., Broderick, T., & Jordan, M. I. (2018). Covariances, Robustness, and Variational Bayes. *Journal of Machine Learning Research*, *19*(51). http://jmlr.org/papers/v19/17-670.html

Giweta, M. (2020). Role of litter production and its decomposition, and factors affecting the processes in a tropical forest ecosystem: A review. *Journal of Ecology and Environment*, *44*(1). https://doi.org/10.1186/s41610-020-0151-2

Golightly, A., & Kypraios, T. (2018). Efficient SMC^2 schemes for stochastic kinetic models. *Statistics and Computing*, *28*(6). https://doi.org/10.1007/s11222-017-9789-8

Golightly, A., & Wilkinson, D. (2010). Markov chain Monte Carlo algorithms for SDE parameter estimation. In N. D. Lawrence, M. Girolami, M. Rattray, & G. Sanguinetti (Eds.). MIT Press.

Golightly, A., & Wilkinson, D. J. (2006). Bayesian sequential inference for nonlinear multivariate diffusions. *Statistics and Computing*, *16*(4). https://doi.org/10.1007/s11222-006-9392-x

Golightly, A., & Wilkinson, D. J. (2011). Bayesian parameter inference for stochastic biochemical network models using particle Markov chain Monte Carlo (2011/09/29). *Interface Focus*, *1*(6). https://doi.org/10.1098/rsfs.2011.0047

Good, I. J. (1971). 46656 varieties of Bayesians. *American Statistician*, *25*(5), 62.

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., & He, K. (2017). Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour. https://doi.org/10.48550/ARXIV.1706.02677

Guenet, B., Moyano, F. E., Peylin, P., Ciais, P., & Janssens, I. A. (2016). Towards a representation of priming on soil carbon decomposition in the global land biosphere model ORCHIDEE (version 1.9.5.2). *Geoscientific Model Development*, *9*(2), 841–855. https://doi.org/10.5194/gmd-9-841-2016

Hagerty, S. B., Allison, S. D., & Schimel, J. P. (2018). Evaluating soil microbial carbon use efficiency explicitly as a function of cellular processes: implications for measurements and models. *Biogeochemistry*, *140*(3). https://doi.org/10.1007/s10533-018-0489-z

Hale, J. K., & LaSalle, J. P. (1963). Differential Equations: Linearity vs. Nonlinearity. *SIAM Review*, *5*(3). https://doi.org/10.1137/1005068

Hararuk, O., & Luo, Y. (2014). Improvement of global litter turnover rate predictions using a Bayesian MCMC approach. *Ecosphere*, *5*(12). https://doi.org/10.1890/ES14-00092.1

Hararuk, O., Smith, M. J., & Luo, Y. (2015). Microbial models with data-driven parameters predict stronger soil carbon responses to climate change. *Global Change Biology*, *21*(6), 2439–2453. https://doi.org/10.1111/gcb.12827

Hararuk, O., Xia, J., & Luo, Y. (2014). Evaluation and improvement of a global land model against soil carbon data using a Bayesian Markov chain Monte Carlo method. *Journal of Geophysical Research: Biogeosciences*, *119*(3). https://doi.org/10.1002/2013JG002535

Hararuk, O., Zwart, J. A., Jones, S. E., Prairie, Y., & Solomon, C. T. (2018). Model-Data Fusion to Test Hypothesized Drivers of Lake Carbon Cycling Reveals Importance of Physical Controls. *Journal of Geophysical Research: Biogeosciences*, *123*(3). https://doi.org/10.1002/2017JG004084

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., ...

Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825). https://doi.org/10.1038/s41586-020-2649-2

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/CVPR.2016.90

Hicks Pries, C. E., Castanha, C., Porras, R. C., & Torn, M. S. (2017). The whole-soil carbon flux in response to warming. *Science*, *355*(6332). https://doi.org/10.1126/science.aal1319

Hindmarsh, A. C., Brown, P. N., Grant, K. E., Lee, S. L., Serban, R., Shumaker, D. E., & Woodward, C. S. (2005). SUNDIALS: Suite of nonlinear and differential/algebraic equation solvers. *ACM Transactions on Mathematical Software*, *31*(3). https://doi.org/10.1145/1089014.1089020

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*(47). http://jmlr.org/papers/v15/hoffman14a.html

Holden, P. B., Edwards, N. R., Ridgwell, A., Wilkinson, R. D., Fraedrich, K., Lunkeit, F., Pollitt, H., Mercure, J. .-., Salas, P., Lam, A., Knobloch, F., Chewpreecha, U., & Viñuales, J. E. (2018). Climate–carbon cycle uncertainties and the Paris Agreement. *Nature Climate Change*, *8*(7). https://doi.org/10.1038/s41558-018-0197-7

Ibrahim, J. G., Chen, M.-H., & Sinha, D. (2001). *Bayesian Survival Analysis* (1st ed.). Springer-Verlag New York. https://doi.org/10.1007/978-1-4757-3447-8

Ioffe, S. (2017). Batch Renormalization: Towards Reducing Minibatch Dependence in Batch-Normalized Models. *Proceedings of the 31st International Conference on Neural Information Processing Systems*.

Ioffe, S., & Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*.

IPCC. (2013). *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press. https://doi.org/10.1017/CBO9781107415324

Irwin, M., & Wang, Z. (2017). Dynamic Systems Modeling. *The International Encyclopedia of Communication Research Methods* (pp. 1–12). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118901731.iecrm0074

Jansson, J. K., & Hofmockel, K. S. (2018). The soil microbiome - from metagenomics to metaphenomics. *Current Opinion in Microbiology*, *43*, 162–168. https://doi.org/10.1016/j.mib.2018.01.013

Jansson, J. K., & Hofmockel, K. S. (2020). Soil microbiomes and climate change. *Nature Reviews Microbiology*, *18*(1), 35–46. https://doi.org/10.1038/s41579-019-0265-7

Jaynes, E., & Justice, J. H. (1986). Bayesian methods: General background. *Maximum Entropy and Bayesian Methods in Applied Statistics: Proceedings of the Fourth Maximum Entropy Workshop University of Calgary, 1984*, 1–25. https://doi.org/10.1017/CBO9780511569678.003

Jenkinson, D., & Rayner, J. (1977). THE TURNOVER OF SOIL ORGANIC MATTER IN SOME OF THE ROTHAMSTED CLASSICAL EXPERIMENTS. *Soil Science*, *123*(5), 298–305. https://doi.org/10.1097/00010694-197705000-00005

Jia, J., & Benson, A. R. (2019). Neural Jump Stochastic Differential Equations. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Curran Associates Inc.

Jiang, L., Yan, Y., Hararuk, O., Mikle, N., Xia, J., Shi, Z., Tjiputra, J., Wu, T., & Luo, Y. (2015). Scale-dependent performance of CMIP5 earth system models in simulating terrestrial vegetation carbon. *Journal of Climate*, *28*(13). https://doi.org/10.1175/JCLI-D-14-00270.1

Jobbágy, E., & Jackson, R. B. (2000). The Vertical Distribution of Soil Organic Carbon and Its Relation to Climate and Vegetation. *Ecological Applications*, *10*(April). https://doi.org/10.2307/2641104

Joyce, J. M. (2011). Kullback-Leibler Divergence. In M. Lovric (Ed.), *International Encyclopedia of Statistical Science*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-04898-2_327

Kaiser, C., Franklin, O., Dieckmann, U., & Richter, A. (2014). Microbial community dynamics alleviate stoichiometric constraints during litter decay. *Ecology Letters*, *17*(6), 680–690. https://doi.org/10.1111/ele.12269

Kalbitz, K., Schwesig, D., Schmerwitz, J., Kaiser, K., Haumaier, L., Glaser, B., Ellerbrock, R., & Leinweber, P. (2003). Changes in properties of soil-derived dissolved organic matter induced by biodegradation. *Soil Biology and Biochemistry*, *35*(8), 1129–1142. https://doi.org/10.1016/S0038-0717(03)00165-2

Kalman, R. E. (1960). A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, *82*(1). https://doi.org/10.1115/1.3662552

Kemp, L., Xu, C., Depledge, J., Ebi, K. L., Gibbins, G., Kohler, T. A., Rockström, J., Scheffer, M., Schellnhuber, H. J., Steffen, W., & Lenton, T. M. (2022). Climate Endgame: Exploring catastrophic climate change scenarios. *Proceedings of the National Academy of Sciences*, *119*(34). https://doi.org/10.1073/pnas.2108146119

Kemppinen, J., Niittynen, P., Riihimäki, H., & Luoto, M. (2018). Modelling soil moisture in a high-latitude landscape using LiDAR and soil data. *Earth Surface Processes and Landforms*, *43*(5), 1019–1031. https://doi.org/10.1002/esp.4301

Kingma, D. P., & Ba, J. (2015). Adam: A Method for Stochastic Optimization. In Y. Bengio & Y. LeCun (Eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. http://arxiv.org/abs/1412.6980

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., & Welling, M. (2016). Improved Variational Inference with Inverse Autoregressive Flow. *Proceedings of the 30th International Conference on Neural Information Processing Systems*.

Kingma, D. P., & Welling, M. (2014). Auto-Encoding Variational Bayes. In Y. Bengio & Y. LeCun (Eds.), *2nd international conference on learning representations, ICLR 2014, banff, ab, canada, april 14-16, 2014, conference track proceedings*. http://arxiv.org/abs/1312.6114

Kobyzev, I., Prince, S., & Brubaker, M. (2020). Normalizing Flows: An Introduction and Review of Current Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/tpami.2020.2992934

Koven, C. D., Lawrence, D. M., & Riley, W. J. (2015). Permafrost carbon-climate feedback is sensitive to deep soil carbon decomposability but not deep soil nitrogen dynamics. *Proceedings of the National Academy of Sciences*, *112*(12), 3752–3757. https://doi.org/10.1073/pnas.1415123112

Krishnan, R. G., Shalit, U., & Sontag, D. (2017). Structured Inference Networks for Nonlinear State Space Models. *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., & Blei, D. M. (2017). Automatic Differentiation Variational Inference. *Journal of Machine Learning Research*, *18*(1), 430–474.

Kullback, S., & Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*, *22*(1). https://doi.org/10.1214/aoms/1177729694

Kvålseth, T. O. (1985). Cautionary Note about R^2. *The American Statistician*, *39*(4). https://doi.org/10.1080/00031305.1985.10479448
doi: 10.1080/00031305.1985.10479448

Lahoz, W. A., & Schneider, P. (2014). Data assimilation: Making sense of earth observation. *Frontiers in Environmental Science*, *2*. https://doi.org/10.3389/fenvs.2014.00016

Lajtha, K., Bowden, R. D., & Nadelhoffer, K. (2014). Litter and Root Manipulations Provide Insights into Soil Organic Matter Dynamics and Stability. *Soil Science Society of America Journal*, *78*(S1), S261–S269. https://doi.org/10.2136/sssaj2013.08.0370nafsc

Lankreijer, H., Janssens, I. A., Buchmann, N., Longdoz, B., Epron, D., & Dore, S. (2003). Measurement of Soil Respiration. In R. Valentini (Ed.), *Fluxes of Carbon, Water and*

Energy of European Forests (pp. 37–54). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-05171-9_3

Lawrence, D. M., Fisher, R. A., Koven, C. D., Oleson, K. W., Swenson, S. C., Bonan, G., Collier, N., Ghimire, B., van Kampenhout, L., Kennedy, D., Kluzek, E., Lawrence, P. J., Li, F., Li, H., Lombardozzi, D., Riley, W. J., Sacks, W. J., Shi, M., Vertenstein, M., ... Zeng, X. (2019). The Community Land Model Version 5: Description of New Features, Benchmarking, and Impact of Forcing Uncertainty. *Journal of Advances in Modeling Earth Systems*, *11*(12), 4245–4287. https://doi.org/10.1029/2018MS001583

Lee, H., Pabbaraju, C., Sevekari, A., & Risteski, A. (2021). Universal Approximation for Log-concave Distributions using Well-conditioned Normalizing Flows. https://doi.org/10.48550/ARXIV.2107.02951

Lehmann, J., Solomon, D., Kinyangi, J., Dathe, L., Wirick, S., & Jacobsen, C. (2008). Spatial complexity of soil organic matter forms at nanometre scales. *Nature Geoscience*, *1*(4), 238–242. https://doi.org/10.1038/ngeo155

Lenton, T. M., Rockström, J., Gaffney, O., Rahmstorf, S., Richardson, K., Steffen, W., & Schellnhuber, H. J. (2019). Climate tipping points - too risky to bet against. *Nature*, *575*(7784). https://doi.org/10.1038/d41586-019-03595-0

Li, C., Frolking, S., & Frolking, T. A. (1992). A model of nitrous oxide evolution from soil driven by rainfall events: 1. Model structure and sensitivity. *Journal of Geophysical Research: Atmospheres*, *97*(D9), 9759–9776. https://doi.org/10.1029/92JD00509

Li, J., Wang, G., Allison, S. D., Mayes, M. A., & Luo, Y. (2014). Soil carbon sensitivity to temperature and carbon use efficiency compared across microbial-ecosystem models of varying complexity. *Biogeochemistry*, *119*(1-3). https://doi.org/10.1007/s10533-013-9948-8

Li, J., Wang, G., Mayes, M. A., Allison, S. D., Frey, S. D., Shi, Z., Hu, X.-M., Luo, Y., & Melillo, J. M. (2019). Reduced carbon use efficiency and increased microbial turnover with soil warming. *Global Change Biology*, *25*(3). https://doi.org/10.1111/gcb.14517

Li, X., Wong, T.-K. L., Chen, R. T. Q., & Duvenaud, D. (2020). Scalable Gradients for Stochastic Differential Equations. In S. Chiappa & R. Calandra (Eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. PMLR. https://proceedings.mlr.press/v108/li20i.html

Luke, C. M., & Cox, P. M. (2011). Soil carbon and climate change: From the Jenkinson effect to the compost-bomb instability. *European Journal of Soil Science*, *62*(1). https://doi.org/10.1111/j.1365-2389.2010.01312.x

Luo, Y., Ahlström, A., Allison, S. D., Batjes, N. H., Brovkin, V., Carvalhais, N., Chappell, A., Ciais, P., Davidson, E. A., Finzi, A., Georgiou, K., Guenet, B., Hararuk, O., Harden, J. W., He, Y., Hopkins, F., Jiang, L., Koven, C., Jackson, R. B., ... Zhou, T. (2016). Toward more realistic projections of soil carbon dynamics by Earth system models. *Global Biogeochemical Cycles*, *30*(1). https://doi.org/10.1002/2015GB005239

Ma, Y.-A., Chen, T., & Fox, E. B. (2015). A Complete Recipe for Stochastic Gradient MCMC. *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*.

Mack, M. C., Schuur, E. A. G., Bret-Harte, M. S., Shaver, G. R., & Chapin, F. S. (2004). Ecosystem carbon storage in arctic tundra reduced by long-term nutrient fertilization. *Nature*, *431*(7007), 440–443. https://doi.org/10.1038/nature02887

Manzoni, S., & Porporato, A. (2007). A theoretical analysis of nonlinearities and feedbacks in soil carbon and nitrogen cycles. *Soil Biology and Biochemistry*, *39*(7), 1542–1556. https://doi.org/10.1016/j.soilbio.2007.01.006

Manzoni, S., & Porporato, A. (2009). Soil carbon and nitrogen mineralization: Theory and models across scales. *Soil Biology and Biochemistry*, *41*(7). https://doi.org/10.1016/j.soilbio.2009.02.031

Maruyama, G. (1955). Continuous Markov processes and stochastic equations. *Rendiconti del Circolo Matematico di Palermo*, *4*(1). https://doi.org/10.1007/BF02846028

Marzouk, Y. M., & Willcox, K. E. (2015). Uncertainty Quantification. In N. Higham, M. Dennis, P. Glendinning, P. Martin, F. Santosa, & J. Tanner (Eds.), *The Princeton Companion to Applied Mathematics* (pp. 131–134). Princeton University Press.

McElreath, R. (2020). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan, 2nd Edition* (2nd ed.). CRC Press. http://xcelab.net/rm/statistical-rethinking/

Melillo, J. M., Frey, S. D., DeAngelis, K. M., Werner, W. J., Bernard, M. J., Bowles, F. P., Pold, G., Knorr, M. A., & Grandy, A. S. (2017). Long-term pattern and magnitude of soil carbon feedback to the climate system in a warming world. *Science*, *358*(6359). https://doi.org/10.1126/science.aan2874

Meng, J., Way, R., Verdolini, E., & Anadon, L. D. (2021). Comparing expert elicitation and model-based probabilistic technology cost forecasts for the energy transition. *Proceedings of the National Academy of Sciences*, *118*(27). https://doi.org/10.1073/pnas.1917165118

Moni, C., Derrien, D., Hatton, P.-J., Zeller, B., & Kleber, M. (2012). Density fractions versus size separates: Does physical fractionation isolate functional soil compartments? *Biogeosciences*, *9*(12), 5181–5197. https://doi.org/10.5194/bg-9-5181-2012

Moore, J. A. M., Jiang, J., Patterson, C. M., Mayes, M. A., Wang, G., & Classen, A. T. (2015). Interactions among roots, mycorrhizas and free-living microbial communities differentially impact soil carbon processes. *Journal of Ecology*, *103*(6), 1442–1453. https://doi.org/10.1111/1365-2745.12484

Moorhead, D. L., & Sinsabaugh, R. L. (2006). A THEORETICAL MODEL OF LITTER DECAY AND MICROBIAL INTERACTION. *Ecological Monographs*, *76*(2), 151–174. https://doi.org/10.1890/0012-9615(2006)076[0151:ATMOLD]2.0.CO;2

Moorhead, D. L., & Weintraub, M. N. (2018). The evolution and application of the reverse Michaelis-Menten equation. *Soil Biology and Biochemistry*, *125*, 261–262. https://doi.org/10.1016/j.soilbio.2018.07.021

Moulatlet, G. M., Zuquim, G., Figueiredo, F. O. G., Lehtonen, S., Emilio, T., Ruokolainen, K., & Tuomisto, H. (2017). Using digital soil maps to infer edaphic affinities of plant species in

Amazonia: Problems and prospects, journal = Ecology and Evolution. *7*(20), 8463–8477. https://doi.org/10.1002/ece3.3242

Nascimento, L., Kuramochi, T., Iacobuta, G., den Elzen, M., Fekete, H., Weishaupt, M., van Soest, H. L., Roelfsema, M., Vivero-Serrano, G. D., Lui, S., Hans, F., de Villafranca Casas, M. J., & Höhne, N. (2022). Twenty years of climate policy: G20 coverage and gaps. *Climate Policy*, *22*(2). https://doi.org/10.1080/14693062.2021.1993776

Neal, R. M. (2011). MCMC Using Hamiltonian Dynamics. *Handbook of Markov Chain Monte Carlo*. CRC Press. https://doi.org/10.1201/b10905-7

Nemeth, C., & Fearnhead, P. (2021). Stochastic Gradient Markov Chain Monte Carlo. *Journal of the American Statistical Association*, *116*(533). https://doi.org/10.1080/01621459.2020.1847120

Nottingham, A. T., Turner, B. L., Whitaker, J., Ostle, N., Bardgett, R. D., McNamara, N. P., Salinas, N., & Meir, P. (2016). Temperature sensitivity of soil enzymes along an elevation gradient in the Peruvian Andes. *Biogeochemistry*, *127*(2). https://doi.org/10.1007/s10533-015-0176-2

Olson, J. S. (1963). Energy Storage and the Balance of Producers and Decomposers in Ecological Systems. *Ecology*, *44*(2), 322–331. https://doi.org/10.2307/1932179

O'Neill, B. C., Kriegler, E., Ebi, K. L., Kemp-Benedict, E., Riahi, K., Rothman, D. S., van Ruijven, B. J., van Vuuren, D. P., Birkmann, J., Kok, K., Levy, M., & Solecki, W. (2017). The roads ahead: Narratives for shared socioeconomic pathways describing world futures in the 21st century. *Global Environmental Change*, *42*. https://doi.org/10.1016/j.gloenvcha.2015.01.004

Ord, T. (2020). *The precipice: Existential risk and the future of humanity*. Hachette Books.

Otto, F. E. L., Frame, D. J., Otto, A., & Allen, M. R. (2015). Embracing uncertainty in climate change policy. *Nature Climate Change*, *5*(10). https://doi.org/10.1038/nclimate2716

Pachauri, R., Allen, M., Barros, V., Broome, J., Cramer, W., Christ, R., Church, J., Clarke, L., Dahe, Q., Dasqupta, P., Dubash, N., Edenhofer, O., Elgizouli, I., Field, C., Forster, P.,

Friedlingstein, P., Fuglestvedt, J., Gomez-Echeverri, L., Hallegatte, S., ... van Ypersele, J.-P. (2014). *Climate change 2014 synthesis report. contribution of working groups I, II, and III to the fifth assessment report of the Intergovernmental Panel on Climate Change*. IPCC.

Palmer, T., & Stevens, B. (2019). The scientific challenge of understanding and estimating climate change. *Proceedings of the National Academy of Sciences*, *116*(49). https://doi.org/10.1073/pnas.1906691116

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., & Lakshminarayanan, B. (2021). Normalizing Flows for Probabilistic Modeling and Inference. *Journal of Machine Learning Research*, *22*(57). http://jmlr.org/papers/v22/19-1028.html

Papamakarios, G., Pavlakou, T., & Murray, I. (2017). Masked Autoregressive Flow for Density Estimation. *Proceedings of the 31st International Conference on Neural Information Processing Systems*.

Parton, W. J. (1996). The CENTURY model. In D. S. Powlson, P. Smith, & J. U. Smith (Eds.), *Evaluation of Soil Organic Matter Models* (pp. 283–291). Springer Berlin Heidelberg.

Parton, W. J., Del Grosso, S. J., Plante, A. F., Adair, E. C., & Lutz, S. M. (2015). Modeling the Dynamics of Soil Organic Matter and Nutrient Cycling. In E. A. Paul (Ed.), *Soil Microbiology, Ecology and Biochemistry* (Fourth Edition, pp. 505–537). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-415955-6.00017-7

Parton, W. J., Schimel, D. S., Cole, C. V., & Ojima, D. S. (1987). Analysis of Factors Controlling Soil Organic Matter Levels in Great Plains Grasslands. *Soil Science Society of America Journal*, *51*(5), 1173–1179. https://doi.org/10.2136/sssaj1987.03615995005100050015x

Parton, W. J., Stewart, J. W. B., & Cole, C. V. (1988). Dynamics of C, N, P and S in grassland soils: a model. *Biogeochemistry*, *5*(1), 109–131. https://doi.org/10.1007/BF02180320

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., ... Chintala, S. (2019). PyTorch: An Im-

perative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

Pawitan, Y. (2001). In All Likelihood: Statistical Modelling and Inference Using Likelihood.

Pedersen, J. S. T., van Vuuren, D. P., Aparício, B. A., Swart, R., Gupta, J., & Santos, F. D. (2020). Variability in historical emissions trends suggests a need for a wide range of global scenarios and regional analyses. *Communications Earth & Environment*, *1*(1). https://doi.org/10.1038/s43247-020-00045-y

Peluchetti, S., & Favaro, S. (2021). Doubly infinite residual neural networks: a diffusion process approach. *Journal of Machine Learning Research*, *22*(175), 1–48. http://jmlr.org/papers/v22/20-706.html

Perez-Cruz, F. (2008). Kullback-Leibler divergence estimation of continuous distributions. *2008 IEEE International Symposium on Information Theory*. https://doi.org/10.1109/ISIT.2008.4595271

Perveen, N., Barot, S., Maire, V., Cotrufo, M. F., Shahzad, T., Blagodatskaya, E., Stewart, C. E., Ding, W., Siddiq, M. R., Dimassi, B., Mary, B., & Fontaine, S. (2019). Universality of priming effect: An analysis using thirty five soils with contrasted properties sampled from five continents. *Soil Biology and Biochemistry*, *134*, 162–171. https://doi.org/10.1016/j.soilbio.2019.03.027

Peters, G. P., Andrew, R. M., Canadell, J. G., Friedlingstein, P., Jackson, R. B., Korsbakken, J. I., Le Quéré, C., & Peregon, A. (2020). Carbon dioxide emissions continue to grow amidst slowly emerging climate policies. *Nature Climate Change*, *10*(1). https://doi.org/10.1038/s41558-019-0659-6

Pielke, R., & Ritchie, J. (2021). Distorting the view of our climate future: The misuse and abuse of climate pathways and scenarios. *Energy Research & Social Science*, *72*. https://doi.org/10.1016/j.erss.2020.101890

Plummer, M. (2003). JAGS: A Program for Analysis of Bayesian Graphical Models using Gibbs Sampling. In K. Hornik, F. Leisch, & A. Zeileis (Eds.), *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*.

Quideau, S., Graham, R., Oh, S.-W., Hendrix, P., & Wasylishen, R. (2005). Leaf litter decomposition in a chaparral ecosystem, Southern California. *Soil Biology and Biochemistry*, *37*(11), 1988–1998. https://doi.org/10.1016/j.soilbio.2005.01.031

R Core Team. (2017). *R: A Language and Environment for Statistical Computing* (Version 3.4.1). Vienna, Austria. https://www.r-project.org

Raczka, B., Hoar, T. J., Duarte, H. F., Fox, A. M., Anderson, J. L., Bowling, D. R., & Lin, J. C. (2021). Improving CLM5.0 Biomass and Carbon Exchange Across the Western United States Using a Data Assimilation System [e2020MS002421 2020MS002421]. *Journal of Advances in Modeling Earth Systems*, *13*(7). https://doi.org/10.1029/2020MS002421

Raftery, A. E., Zimmer, A., Frierson, D. M. W., Startz, R., & Liu, P. (2017). Less than 2 °c warming by 2100 unlikely. *Nature Climate Change*, *7*(9). https://doi.org/10.1038/nclimate3352

Rauch, H. E., Tung, F., & Striebel, C. T. (1965). Maximum likelihood estimates of linear dynamic systems. *AIAA journal*, *3*(8).

Raza, S., Zamanian, K., Ullah, S., Kuzyakov, Y., Virto, I., & Zhou, J. (2021). Inorganic carbon losses by soil acidification jeopardize global efforts on carbon sequestration and climate change mitigation. *Journal of Cleaner Production*, *315*. https://doi.org/10.1016/j.jclepro.2021.128036

Reid, N. M. (2015). Approximate likelihoods. In L. Guo & Z.-M. Ma (Eds.), *Proceedings of the 8th International Congress on Industrial and Applied Mathematics*. Higher Education Press.

Richards, C. E., Lupton, R. C., & Allwood, J. M. (2021). Re-framing the threat of global warming: An empirical causal loop diagram of climate change, food insecurity and societal collapse. *Climatic Change*, *164*(3). https://doi.org/10.1007/s10584-021-02957-w

Romero-Olivares, A. L., Allison, S. D., & Treseder, K. K. (2017). Soil microbes and their response to experimental warming over time: A meta-analysis of field studies. *Soil Biology and Biochemistry*, *107*. https://doi.org/10.1016/j.soilbio.2016.12.026

Ruiz, F. R., Titsias RC AUEB, M., & Blei, D. (2016). The Generalized Reparameterization Gradient. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems*. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2016/file/f718499c1c8cef6730f9fd03c8125cab-Paper.pdf

Ryder, T., Prangle, D., Golightly, A., & Matthews, I. (2021). The neural moving average model for scalable variational inference of state space models. In C. de Campos & M. H. Maathuis (Eds.), *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence* (pp. 12–22). PMLR. https://proceedings.mlr.press/v161/ryder21a.html

Ryder, T., Golightly, A., McGough, A. S., & Prangle, D. (2018). Black-Box Variational Inference for Stochastic Differential Equations. In J. Dy & A. Krause (Eds.), *Proceedings of the 35th International Conference on Machine Learning*. PMLR. https://proceedings.mlr.press/v80/ryder18a.html

Ryder, T., & Prangle, D. (n.d.). *Accelerating Inference for Neural Stochastic Differential Equations with Checkpoints* [Manuscript in preparation].

Saifuddin, M., Abramoff, R. Z., Davidson, E. A., Dietze, M. C., & Finzi, A. C. (2021). Identifying Data Needed to Reduce Parameter Uncertainty in a Coupled Microbial Soil C and N Decomposition Model. *Journal of Geophysical Research: Biogeosciences*, *126*(12). https://doi.org/10.1029/2021JG006593

Salimans, T., Kingma, D., & Welling, M. (2015). Markov Chain Monte Carlo and Variational Inference: Bridging the Gap. In F. Bach & D. Blei (Eds.), *Proceedings of the 32nd International Conference on Machine Learning*. PMLR. https://proceedings.mlr.press/v37/salimans15.html

Salimans, T., & Knowles, D. A. (2013). Fixed-Form Variational Posterior Approximation through Stochastic Linear Regression. *Bayesian Analysis*, *8*(4). https://doi.org/10.1214/13-BA858

Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016). Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, *2*. https://doi.org/10.7717/peerj-cs.55

Särkkä, S. (2013). *Bayesian Filtering and Smoothing*. Cambridge University Press. https://doi.org/10.1017/CBO9781139344203

Särkkä, S., & Solin, A. (2019). *Applied Stochastic Differential Equations*. Cambridge University Press. https://doi.org/10.1017/9781108186735

Schaphoff, S., Heyder, U., Ostberg, S., Gerten, D., Heinke, J., & Lucht, W. (2013). Contribution of permafrost soils to the global carbon budget. *Environmental Research Letters*, *8*(1). https://doi.org/10.1088/1748-9326/8/1/014026

Scharlemann, J. P., Tanner, E. V., Hiederer, R., & Kapos, V. (2014). Global soil carbon: Understanding and managing the largest terrestrial carbon pool. *Carbon Management*, *5*(1). https://doi.org/10.4155/cmt.13.77

Schimel, J. P. (2001). Biogeochemical Models: Implicit versus Explicit Microbiology. In E.-D. Schulze, M. Heimann, S. Harrison, E. Holland, J. Lloyd, I. C. Prentice, & D. Schimel (Eds.), *Global Biogeochemical Cycles in the Climate System* (pp. 177–183). Academic Press. https://doi.org/10.1016/B978-012631260-7/50015-7

Schimel, J. P., & Weintraub, M. N. (2003). The implications of exoenzyme activity on microbial carbon and nitrogen limitation in soil: A theoretical model. *Soil Biology and Biochemistry*, *35*(4), 549–563. https://doi.org/10.1016/S0038-0717(03)00015-4

Schmidt, M. W. I., Torn, M. S., Abiven, S., Dittmar, T., Guggenberger, G., Janssens, I. A., Kleber, M., Kögel-Knabner, I., Lehmann, J., Manning, D. A. C., Nannipieri, P., Rasse, D. P., Weiner, S., & Trumbore, S. E. (2011). Persistence of soil organic matter as an ecosystem property. *Nature*, *478*(7367), 49–56. https://doi.org/10.1038/nature10386

Schneider, T., Kaul, C. M., & Pressel, K. G. (2019). Possible climate transitions from breakup of stratocumulus decks under greenhouse warming. *Nature Geoscience*, *12*(3). https://doi.org/10.1038/s41561-019-0310-1

Shen, W., Jenerette, G. D., Hui, D., & Scott, R. L. (2016). Precipitation legacy effects on dryland ecosystem carbon fluxes: Direction, magnitude and biogeochemical carryovers. *Biogeosciences*, *13*(2), 425–439. https://doi.org/10.5194/bg-13-425-2016

Shi, Z., Allison, S. D., He, Y., Levine, P. A., Hoyt, A. M., Beem-Miller, J., Zhu, Q., Wieder, W. R., Trumbore, S., & Randerson, J. T. (2020). The age distribution of global soil carbon inferred from radiocarbon measurements. *Nature Geoscience*, *13*(8), 555–559. https://doi.org/10.1038/s41561-020-0596-z

Shi, Z., Crowell, S., Luo, Y., & Moore, B. (2018). Model structures amplify uncertainty in predicted soil carbon responses to climate change. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-018-04526-9

Sierra, C. A., Müller, M., & Trumbore, S. E. (2012). Models of soil organic matter decomposition: The SoilR package, version 1.0. *Geoscientific Model Development*, *5*(4), 1045–1060. https://doi.org/10.5194/gmd-5-1045-2012

Sierra, C. A., Malghani, S., & Müller, M. (2015). Model structure and parameter identification of soil organic matter models. *Soil Biology and Biochemistry*, *90*, 197–203. https://doi.org/10.1016/j.soilbio.2015.08.012

Sierra, C. A., & Müller, M. (2015). A general mathematical framework for representing soil organic matter dynamics. *Ecological Monographs*, *85*(4), 505–524. https://doi.org/10.1890/15-0361.1

Sihi, D., Davidson, E. A., Savage, K. E., & Liang, D. (2020). Simultaneous numerical representation of soil microsite production and consumption of carbon dioxide, methane, and nitrous oxide using probability distribution functions. *Global Change Biology*, *26*(1), 200–218. https://doi.org/10.1111/gcb.14855

Sinsabaugh, R. L., & Follstad Shah, J. J. (2012). Ecoenzymatic Stoichiometry and Ecological Theory. *Annual Review of Ecology, Evolution, and Systematics*, *43*(1), 313–343. https://doi.org/10.1146/annurev-ecolsys-071112-124414

Sinsabaugh, R. L., Manzoni, S., Moorhead, D. L., & Richter, A. (2013). Carbon use efficiency of microbial communities: Stoichiometry, methodology and modelling. *Ecology Letters*, *16*(7), 930–939. https://doi.org/10.1111/ele.12113

Sinsabaugh, R. L., Moorhead, D. L., Xu, X., & Litvak, M. E. (2017). Plant, microbial and ecosystem carbon use efficiencies interact to stabilize microbial growth as a fraction of gross primary production. *New Phytologist*, *214*(4), 1518–1526. https://doi.org/10.1111/nph.14485

Sistla, S. A., Rastetter, E. B., & Schimel, J. P. (2014). Responses of a tundra system to warming using SCAMPS: A stoichiometrically coupled, acclimating microbe–plant–soil model. *Ecological Monographs*, *84*(1), 151–170. https://doi.org/10.1890/12-2119.1

Smith, O. L. (1979a). An analytical model of the decomposition of soil organic matter. *Soil Biology and Biochemistry*, *11*(6), 585–606. https://doi.org/10.1016/0038-0717(79)90027-0

Smith, O. L. (1979b). Application of a model of the decomposition of soil organic matter. *Soil Biology and Biochemistry*, *11*(6), 607–618. https://doi.org/10.1016/0038-0717(79)90028-2

Smith, P., Fang, C., Dawson, J. J. C., & Moncrieff, J. B. (2008). Impact of Global Warming on Soil Organic Carbon. Academic Press. https://doi.org/https://doi.org/10.1016/S0065-2113(07)00001-6

Smith, P., Lanigan, G., Kutsch, W. L., Buchmann, N., Eugster, W., Aubinet, M., Ceschia, E., B., P., Yeluripati, J. B., Osborne, B., Moors, E. J., Brut, A., Wattenbach, M., Saunders, M., & Jones, M. (2010). Measurements necessary for assessing the net ecosystem carbon budget of croplands. *Agriculture, Ecosystems & Environment*, *139*(3), 302–315. https://doi.org/10.1016/j.agee.2010.04.004

Smith, P., Soussana, J.-F., Angers, D., Schipper, L., Chenu, C., Rasse, D. P., Batjes, N. H., van Egmond, F., McNeill, S., Kuhnert, M., Arias-Navarro, C., Olesen, J. E., Chirinda, N., Fornara, D., Wollenberg, E., Álvaro-Fuentes, J., Sanz-Cobena, A., & Klumpp, K. (2020). How to measure, report and verify soil carbon change to realize the potential of soil carbon sequestration for atmospheric greenhouse gas removal. *Global Change Biology*, *26*(1), 219–241. https://doi.org/10.1111/gcb.14815

Sognnaes, I., Gambhir, A., van de Ven, D.-J., Nikas, A., Anger-Kraavi, A., Bui, H., Campagnolo, L., Delpiazzo, E., Doukas, H., Giarola, S., Grant, N., Hawkes, A., Köberle, A. C., Kolpakov, A., Mittal, S., Moreno, J., Perdana, S., Rogelj, J., Vielle, M., & Peters, G. P. (2021). A multi-model analysis of long-term emissions and warming implications of current mitigation efforts. *Nature Climate Change*, *11*(12). https://doi.org/10.1038/s41558-021-01206-3

Sparling, G. P. (1992). Ratio of microbial biomass carbon to soil organic carbon as a sensitive indicator of changes in soil organic matter. *Australian Journal of Soil Research*, *30*(2). https://doi.org/10.1071/SR9920195

Spiess, A. N., & Neumeyer, N. (2010). An evaluation of R^2 as an inadequate measure for nonlinear models in pharmacological and biochemical research: A Monte Carlo approach. *BMC Pharmacology*, *10*(1). https://doi.org/10.1186/1471-2210-10-6

Stan Development Team. (2020). *RStan: the R interface to Stan* (Version 2.18.1). New York City, New York. https://mc-stan.org

Steffen, W., Rockström, J., Richardson, K., Lenton, T. M., Folke, C., Liverman, D., Summerhayes, C. P., Barnosky, A. D., Cornell, S. E., Crucifix, M., Donges, J. F., Fetzer, I., Lade, S. J., Scheffer, M., Winkelmann, R., & Schellnhuber, H. J. (2018). Trajectories of the Earth System in the Anthropocene. *Proceedings of the National Academy of Sciences*, *115*(33). https://doi.org/10.1073/pnas.1810141115

Steinweg, J. M., Jagadamma, S., Frerichs, J., & Mayes, M. A. (2013). Activation Energy of Extracellular Enzymes in Soils from Different Biomes. *PLOS ONE*, *8*(3). https://doi.org/10.1371/journal.pone.0059943

Stigler, S. (1980). STIGLER'S LAW OF EPONYMY. *Transactions of the New York Academy of Sciences*, *39*(1 Series II), 147–157. https://doi.org/10.1111/j.2164-0947.1980.tb02775.x

Stigler, S. (1983). Who Discovered Bayes's Theorem? *The American Statistician*, *37*(4a), 290–296. https://doi.org/10.1080/00031305.1983.10483122

Stigler, S. (1990). *The History of Statistics: The Measurement of Uncertainty before 1900*. Harvard University Press.

Strogatz, S. (2018). *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*. Taylor & Francis. https://doi.org/10.1201/9780429492563

Sujono, D., Xie, H. W., Allison, S., & Sudderth, E. B. (2022). Variational Inference for Soil Biogeochemical Models. *ICML 2022 2nd AI for Science Workshop*. https://openreview.net/forum?id=3_HrwqrPd4U

Sulman, B. N., Moore, J. A., Abramoff, R., Averill, C., Kivlin, S., Georgiou, K., Sridhar, B., Hartman, M. D., Wang, G., Wieder, W. R., Bradford, M. A., Luo, Y., Mayes, M. A., Morrison, E., Riley, W. J., Salazar, A., Schimel, J. P., Tang, J., & Classen, A. T. (2018). Multiple models and experiments underscore large uncertainty in soil carbon dynamics. *Biogeochemistry*, *141*(2). https://doi.org/10.1007/s10533-018-0509-z

Sulman, B. N., Phillips, R. P., Oishi, A. C., Shevliakova, E., & Pacala, S. W. (2014). Microbe-driven turnover offsets mineral-mediated storage of soil carbon under elevated CO_2. *Nature Climate Change*, *4*(12). https://doi.org/10.1038/nclimate2436

Summers, C., & Dinneen, M. J. (2020). Four Things Everyone Should Know to Improve Batch Normalization. *International Conference on Learning Representations*. https://openreview.net/forum?id=HJx8HANFDH

Tang, J. Y. (2015). On the relationships between the Michaelis–Menten kinetics, reverse Michaelis–Menten kinetics, equilibrium chemistry approximation kinetics, and quadratic kinetics. *Geoscientific Model Development*, *8*(12). https://doi.org/10.5194/gmd-8-3823-2015

Tang, J. Y., & Riley, W. J. (2013). A total quasi-steady-state formulation of substrate uptake kinetics in complex networks and an example application to microbial litter decomposition. *Biogeosciences*, *10*(12), 8329–8351. https://doi.org/10.5194/bg-10-8329-2013

Tao, F., Zhou, Z., Huang, Y., Li, Q., Lu, X., Ma, S., Huang, X., Liang, Y., Hugelius, G., Jiang, L., Doughty, R., Ren, Z., & Luo, Y. (2020). Deep Learning Optimizes Data-Driven Representation of Soil Organic Carbon in Earth System Model Over the Conterminous United States. *3*. https://doi.org/10.3389/fdata.2020.00017

Todd-Brown, K. E. O., Hopkins, F. M., Kivlin, S. N., Talbot, J. M., & Allison, S. D. (2011). A framework for representing microbial decomposition in coupled climate models. *Biogeochemistry*, *109*(1-3). https://doi.org/10.1007/s10533-011-9635-6

Todd-Brown, K. E. O., Randerson, J. T., Hopkins, F., Arora, V., Hajima, T., Jones, C., Shevliakova, E., Tjiputra, J., Volodin, E., Wu, T., Zhang, Q., & Allison, S. D. (2014). Changes in soil organic carbon storage predicted by earth system models during the 21st century. *Biogeosciences*, *11*(8), 2341–2356. https://doi.org/10.5194/bg-11-2341-2014

Todd-Brown, K. E. O., Randerson, J. T., Post, W. M., Hoffman, F. M., Tarnocai, C., Schuur, E. A. G., & Allison, S. D. (2013). Causes of variation in soil carbon simulations from CMIP5 Earth system models and comparison with observations. *Biogeosciences*, *10*(3). https://doi.org/10.5194/bg-10-1717-2013

Todd-Brown, K. E. O., Zheng, B., & Crowther, T. W. (2018). Field-warmed soil carbon changes imply high 21st-century modeling uncertainty. *Biogeosciences*, *15*(12), 3659–3671. https://doi.org/10.5194/bg-15-3659-2018

Trasar-Cepeda, C., Gil-Sotres, F., & Leirós, M. C. (2007). Thermodynamic parameters of enzymes in grassland soils from Galicia, NW Spain. *Soil Biology and Biochemistry*, *39*(1). https://doi.org/10.1016/j.soilbio.2006.08.002

Trumbore, S. (2000). Age of soil organic matter and soil respiration: Radiocarbon constraints on belowground C dynamics. *Ecological Applications*, *10*(2). https://doi.org/10.1890/1051-0761(2000)010[0399:AOSOMA]2.0.CO;2

Ťupek, B., Launiainen, S., Peltoniemi, M., Sievänen, R., Perttunen, J., Kulmala, L., Penttilä, T., Lindroos, A.-J., Hashimoto, S., & Lehtonen, A. (2019). Evaluating CENTURY and Yasso soil carbon models for CO2 emissions and organic carbon stocks of boreal forest soil with Bayesian multi-model inference. *European Journal of Soil Science*, *70*(4), 847–858. https://doi.org/10.1111/ejss.12805

Tzen, B., & Raginsky, M. (2019). Neural Stochastic Differential Equations: Deep Latent Gaussian Models in the Diffusion Limit. https://doi.org/10.48550/ARXIV.1905.09883

van Gestel, N., Shi, Z., van Groenigen, K. J., Osenberg, C. W., Andresen, L. C., Dukes, J. S., Hovenden, M. J., Luo, Y., Michelsen, A., Pendall, E., Reich, P. B., Schuur, E. A. G., & Hungate, B. A. (2018). Predicting soil carbon loss with warming. *Nature*, *554*(7693). https://doi.org/10.1038/nature25745

Varman, A. M., He, L., Follenfant, R., Wu, W., Wemmer, S., Wrobel, S. A., Tang, Y. J., & Singh, S. (2016). Decoding how a soil bacterium extracts building blocks and metabolic energy from ligninolysis provides road map for lignin valorization. *Proceedings of the National Academy of Sciences*, *113*(40), E5802–E5811. https://doi.org/10.1073/pnas.1606043113

Varney, R. M., Chadburn, S. E., Friedlingstein, P., Burke, E. J., Koven, C. D., Hugelius, G., & Cox, P. M. (2020). A spatial emergent constraint on the sensitivity of soil carbon turnover to global warming. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-19208-8

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., & Gelman, A. (2019). loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models. https://mc-stan.org/loo

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5). https://doi.org/10.1007/s11222-016-9696-4

Vehtari, A., & Ojanen, J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, *6*(1). https://doi.org/10.1214/12-ss102

Vrugt, J. A. (2016). Markov chain Monte Carlo simulation using the DREAM software package: Theory, concepts, and MATLAB implementation. *Environmental Modelling & Software*, *75*, 273–316. https://doi.org/10.1016/j.envsoft.2015.08.013

Waksman, S. A. (1927). *Principles of Soil Microbiology*. Williams & Wilkins, Co.

Wang, B., & Allison, S. D. (2019). Emergent properties of organic matter decomposition by soil enzymes. *Soil Biology and Biochemistry*, *136*. https://doi.org/10.1016/j.soilbio.2019.107522

Wang, B., Brewer, P. E., Shugart, H. H., Lerdau, M. T., & Allison, S. D. (2019). Building bottom-up aggregate-based models (ABMs) in soil systems with a view of aggregates as biogeochemical reactors. *Global Change Biology*, *25*(8), e6–e8. https://doi.org/10.1111/gcb.14684

Wang, G., Post, W. M., & Mayes, M. A. (2013). Development of microbial-enzyme-mediated decomposition model parameters through steady-state and dynamic analyses. *Ecological Applications*, *23*(1), 255–272. https://doi.org/10.1890/12-0681.1

Wang, S., Luo, Y., & Niu, S. (2022). Reparameterization Required After Model Structure Changes From Carbon Only to Carbon-Nitrogen Coupling. *Journal of Advances in Modeling Earth Systems*, *14*(4). https://doi.org/10.1029/2021MS002798

Wang, Y. P., Chen, B. C., Wieder, W. R., Leite, M., Medlyn, B. E., Rasmussen, M., Smith, M. J., Agusto, F. B., Hoffman, F., & Luo, Y. Q. (2014). Oscillatory behavior of two nonlinear microbial models of soil carbon decomposition. *Biogeosciences*, *11*(7), 1817–1831. https://doi.org/10.5194/bg-11-1817-2014

Wang, Y., & Barber, D. (2014). Gaussian Processes for Bayesian Estimation in Ordinary Differential Equations. *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*.

Waring, B. G., Sulman, B. N., Reed, S., Smith, A. P., Averill, C., Creamer, C. A., Cusack, D. F., Hall, S. J., Jastrow, J. D., Jilling, A., Kemner, K. M., Kleber, M., Liu, X.-J. A., Pett-Ridge, J., & Schulz, M. (2020). From pools to flow: The PROMISE framework for new insights on soil carbon cycling in a changing world. *Global Change Biology*, *26*(12), 6631–6643. https://doi.org/10.1111/gcb.15365

Weintraub, M. N., & Schimel, J. P. (2003). Interactions between Carbon and Nitrogen Mineralization and Soil Organic Matter Chemistry in Arctic Tundra Soils. *Ecosystems*, *6*(2), 0129–0143. https://doi.org/10.1007/s10021-002-0124-6

Whitaker, G. A. (2016). *Bayesian inference for stochastic differential mixed-effects models* (Doctoral dissertation). Newcastle University.

Whitaker, G. A., Golightly, A., Boys, R. J., & Sherlock, C. (2017). Bayesian Inference for Diffusion-Driven Mixed-Effects Models. *Bayesian Analysis*, *12*(2). https://doi.org/10.1214/16-BA1009

Wieder, W. R., Allison, S. D., Davidson, E. A., Georgiou, K., Hararuk, O., He, Y., Hopkins, F., Luo, Y., Smith, M. J., Sulman, B., Todd-Brown, K. E. O., Wang, Y. P., Xia, J., & Xu, X. (2015). Explicitly representing soil microbial processes in Earth system models. *Global Biogeochemical Cycles*, *29*(10). https://doi.org/10.1002/2015GB005188

Wieder, W. R., Boehnert, J., & Bonan, G. B. (2014). Evaluating soil biogeochemistry parameterizations in Earth system models with observations. *Global Biogeochemical Cycles*, *28*(3). https://doi.org/10.1002/2013GB004665

Wieder, W. R., Bonan, G. B., & Allison, S. D. (2013). Global soil carbon projections are improved by modelling microbial processes. *Nature Climate Change*, *3*(10), 909–912. https://doi.org/10.1038/nclimate1951

Wieder, W. R., Grandy, A. S., Kallenbach, C. M., & Bonan, G. B. (2014). Integrating microbial physiology and physio-chemical principles in soils with the MIcrobial-MIneral Carbon Stabilization (MIMICS) model. *Biogeosciences*, *11*(14), 3899–3917. https://doi.org/10.5194/bg-11-3899-2014

Wieder, W. R., Grandy, A. S., Kallenbach, C. M., Taylor, P. G., & Bonan, G. B. (2015). Representing life in the Earth system with soil microbial functional traits in the MIMICS model. *Geoscientific Model Development*, *8*(6). https://doi.org/10.5194/gmd-8-1789-2015

Wieder, W. R., Hartman, M. D., Sulman, B. N., Wang, Y. P., Koven, C. D., & Bonan, G. B. (2018). Carbon cycle confidence and uncertainty: Exploring variation among soil biogeochemical models. *Global Change Biology*, *24*(4). https://doi.org/10.1111/gcb.13979

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

Wiqvist, S., Golightly, A., McLean, A. T., & Picchini, U. (2021). Efficient inference for stochastic differential equation mixed-effects models using correlated particle pseudo-marginal algorithms. *Computational Statistics & Data Analysis*, *157*. https://doi.org/10.1016/j.csda.2020.107151

Wood, T. E., González, G., Silver, W. L., Reed, S. C., & Cavaleri, M. A. (2019). On the shoulders of giants: Continuing the legacy of large-scale ecosystem manipulation experiments in Puerto Rico. *Forests*, *10*(3). https://doi.org/10.3390/f10030210

Wu, D., Piao, S., Liu, Y., Ciais, P., & Yao, Y. (2018). Evaluation of CMIP5 Earth System Models for the Spatial Patterns of Biomass and Soil Carbon Turnover Times and Their Linkage with Climate. *Journal of Climate*, *31*(15), 5947–5960. https://doi.org/10.1175/JCLI-D-17-0380.1

Xia, J., Luo, Y., Wang, Y.-P., & Hararuk, O. (2013). Traceable components of terrestrial carbon storage capacity in biogeochemical models. *Global Change Biology*, *19*(7), 2104–2116. https://doi.org/10.1111/gcb.12172

Xie, H. W., Romero-Olivares, A. L., Guindani, M., & Allison, S. D. (2020). A Bayesian approach to evaluation of soil biogeochemical models. *Biogeosciences*, *17*(15). https://doi.org/10.5194/bg-17-4043-2020

Xie, H. W., Sujono, D., Ryder, T., Sudderth, E. B., & Allison, S. D. (2022). A framework for variational inference and data assimilation of soil biogeochemical models using state space approximations and normalizing flows. https://doi.org/10.1002/essoar.10512103.7

Xu, W., Chen, R. T. Q., Li, X., & Duvenaud, D. (2022). Infinitely Deep Bayesian Neural Networks with Stochastic Differential Equations. In G. Camps-Valls, F. J. R. Ruiz, & I. Valera (Eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics* (pp. 721–738). PMLR. https://proceedings.mlr.press/v151/xu22a.html

Xue, K., M. Yuan, M., J. Shi, Z., Qin, Y., Deng, Y., Cheng, L., Wu, L., He, Z., Van Nostrand, J. D., Bracho, R., Natali, S., Schuur, E. A. G., Luo, C., Konstantinidis, K. T., Wang, Q., Cole, J. R., Tiedje, J. M., Luo, Y., & Zhou, J. (2016). Tundra soil carbon is vulnerable to rapid microbial decomposition under climate warming. *Nature Climate Change*, *6*(6), 595–600. https://doi.org/10.1038/nclimate2940

Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2018). Yes, but Did It Work?: Evaluating Variational Inference. In J. Dy & A. Krause (Eds.), *Proceedings of the 35th International Conference on Machine Learning*. PMLR. https://proceedings.mlr.press/v80/yao18a.html

You, K., Long, M., Wang, J., & Jordan, M. I. (2019). How Does Learning Rate Decay Help Modern Neural Networks? https://doi.org/10.48550/ARXIV.1908.01878

Zelenev, V. V., van Bruggen, A. H. C., & Semenov, A. M. (2000). "BACWAVE," a Spatial–Temporal Model for Traveling Waves of Bacterial Populations in Response to a Moving Carbon Source in Soil. *Microbial Ecology*, *40*(3), 260–272. https://doi.org/10.1007/s002480000029

Zhang, B., Chen, S., He, X., Liu, W., Zhao, Q., Zhao, L., & Tian, C. (2014). Responses of soil microbial communities to experimental warming in alpine grasslands on the Qinghai-Tibet Plateau. *PLOS ONE*, *9*(8). https://doi.org/10.1371/journal.pone.0103859

Zhang, X., Niu, G.-Y., Elshall, A. S., Ye, M., Barron-Gafford, G. A., & Pavao-Zuckerman, M. (2014). Assessing five evolving microbial enzyme models against field measurements from a semiarid savannah - What are the mechanisms of soil respiration pulses? *Geophysical Research Letters*, *41*(18), 6428–6434. https://doi.org/10.1002/2014GL061399

Zhao, Y., & Park, I. M. (2016). Interpretable Nonlinear Dynamic Modeling of Neural Trajectories. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems*. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2016/file/b2531e7bb29bf22e1daae486fae3417a-Paper.pdf

Zhu, Q., Bi, W., Liu, X., Ma, X., Li, X., & Wu, D. (2020). A Batch Normalized Inference Network Keeps the KL Vanishing Away. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. https://doi.org/10.18653/v1/2020.acl-main.235

Zomer, R. J., Bossio, D. A., Sommer, R., & Verchot, L. V. (2017). Global Sequestration Potential of Increased Organic Carbon in Cropland Soils. *Scientific Reports*, *7*(1). https://doi.org/10.1038/s41598-017-15794-8

Zuquim, G., Stropp, J., Moulatlet, G. M., Van doninck, J., Quesada, C. A., Figueiredo, F. O. G., Costa, F. R. C., Ruokolainen, K., & Tuomisto, H. (2019). Making the most of scarce data: Mapping soil gradients in data-poor areas using species occurrence records. *Methods in Ecology and Evolution*, *10*(6), 788–801. https://doi.org/10.1111/2041-210X.13178