Variational Inference for Soil Biogeochemical Models

Debora Sujono [1] **Hua Wally Xie** [2] **Steven Allison** [3] **Erik B. Sudderth** [1]

Abstract

Soil biogeochemical models (SBMs) are an important tool used by Earth scientists to quantify the impact of rising global surface temperatures. SBMs represent the soil carbon and microbial dynamics across time as differential equations, and inference on model parameters is conducted to project changes in parameter values under warming climate conditions. Traditionally, the field has relied on MCMC algorithms for posterior inference, often implemented via probabilistic programming languages like Stan. However, computational cost makes it difficult to scale MCMC methods to more complex SBM models and large-scale datasets. In this paper, we develop variational inference methods for time-discretized SBMs as an alternative to MCMC. We propose an efficient family of variational approximations based on Gauss-Markov distributions that leverages the temporal structure of sequential models, scaling linearly in both time and space with respect to the sequence length. We show in experiments with simulated data and real CO_2 response ratios that our approach converges faster, and recovers posterior that more accurately captures uncertainty than previous variational methods. Our black-box inference approach is designed to integrate with probabilistic programming languages to enable future scientific applications.

1. Introduction

Soil microbes play a crucial role in global carbon cycle and Earth ecosystem functions. Although it is widely known that rising global surface temperatures affect soil microbes, it is hard to quantify the impact. *Soil biogeochemical models* (SBMs) are used to represent the transfer of elements, such as carbon, between types of organic molecules and quantify the response of biological soil systems to global warming (Xie et al., 2020). However, there can be vast differences between the predictions of competing models. To quickly compare different hypotheses we seek fast, accurate, and easy inference methods that also maintain biological interpretability of model parameters and latent variables.

SBMs represent the soil carbon and microbial dynamics across time as differential equations, and inference on model parameters is conducted to project changes in parameter values under warming climate conditions. Partly due to ease of use in *probabilistic programming languages* (PPLs) like Stan (Carpenter et al., 2017), the field has mainly relied on MCMC methods for posterior inference of model parameters, where latent states are approximated by deterministic *ordinary differential equation* (ODE) solvers (Li et al., 2019; Wang et al., 2022; Xie et al., 2020). PPLs allow a user to specify a model in an intuitive modeling language and (mostly) automate the inference process, which is important for scientific applications that require iterating over many competing models. However, computational cost makes it difficult to scale MCMC methods to more complex, non-linear SBMs for larger datasets spanning decades.

Variational inference (VI; Wainwright & Jordan (2008)) has been widely successful as a faster alternative to MCMC in other large scale applications of machine learning (Gan et al., 2015; Gopalan et al., 2016; Ji et al., 2019). VI reframes the task of posterior inference as an optimization problem by minimizing the KL divergence between the true posterior and an approximate variational distribution. For complex non-conjugate models, variational algorithms often require tedious and model-specific manual derivations, and are thus not ideal for scientific applications. Recently, "black box" variational inference methods have been proposed that can be readily applied to general models. For differentiable models with continuous latent variables, *Automatic Differentiation Variational Inference* (ADVI; Kucukelbir et al. (2017)) provides unbiased and low variance gradient estimates through transformation of random variables and the "reparameterization trick" (Kingma & Welling, 2014; Rezende et al., 2014). ADVI has been implemented as the default variational algorithm in Stan.

In this paper, we propose an extension of ADVI that lever-

[1]Department of Computer Science, University of California, Irvine [2]Center for Complex Biological Systems, University of California, Irvine [3]Department of Ecology and Evolutionary Biology, University of California, Irvine. Correspondence to: Debora Sujono <dsujono@uci.edu>.

2nd AI4Science Workshop at the 39^{th} International Conference on Machine Learning (ICML), 2022.

ages the temporal structure of sequential models. Our proposed variational family based on Gauss-Markov distributions captures temporal dependence efficiently and scales linearly in the length of the sequence. We reformulate ODE models from prior work as *stochastic differential equations* (SDEs), as the smoother trajectories of ODEs oversimplify the substantial noise that is inherent to biological systems (Browning et al., 2020; Abs et al., 2020; Golightly & Wilkinson, 2011). Rather than solving the SDEs directly, we use an approximation based on time discretization and express the problem as a sequential latent variable model, where we jointly optimize model parameters and latent states. We show that previous variational methods perform poorly by systematically under or over-estimating variance of the posterior. In experiments with simulated data and real CO_2 response ratios, we show that our method leads to faster convergence to better loss and variational approximations compared to several baseline methods, including standard ADVI and an amortized variational inference approach.

Several authors have examined inference in SDEs and related sequence models. Ryder et al. (2018) proposed an amortized VI algorithm specifically designed for SDEs, where variational distributions are parameterized by neural networks. Other authors have recently proposed continuous-time methods for sequential and time series models (Chen et al., 2018; Rubanova et al., 2019; Kidger et al., 2020; Li et al., 2020; Schirmer et al., 2021). Here, we favor the simpler approach based on ADVI and discrete-time approximations. Our method is broadly applicable to general sequential latent variable models and designed to enable integration with PPLs that allows scientists to easily and quickly design, test, and revise different models.

2. Automatic Differentiation Variational Inference

Given any model with latent variable z and observation y, the goal of Bayesian inference is to infer posterior distribution $p(z|y) = \frac{p(z)p(y|z)}{p(y)}$. Except for very simple models, the normalizing constant $p(y) = \int_z p(z)p(y|z)dz$ is intractable and thus approximation of the posterior is needed. Variational inference seeks a variational distribution $q(z;\lambda)$ parameterized by variational parameter λ by maximizing the *evidence lower bound* (ELBO):

$$\mathcal{L}(\lambda) = E_{q(z;\lambda)}[\log p(z,y) - \log q(z;\lambda)]. \tag{1}$$

Maximizing the ELBO is equivalent to minimizing KL divergence of $q(z;\lambda)$ from the true posterior $p(z|y)$.

Automatic Differentiation Variational Inference (ADVI; Kucukelbir et al. (2017)) is a black-box VI algorithm that can be applied to any differentiable probability models. It automates ELBO optimization of equation (1) via automatic differentiation. It works by first transforming the support of the latent variables z to the real coordinate space $\tilde{z} = T(z) \in \mathbb{R}^K$. The transformed joint density is given by

$$p(y,\tilde{z}) = p(y, T^{-1}(\tilde{z}))|\det J(\tilde{z})|, \tag{2}$$

where $J(\tilde{z})$ is the Jacobian of the inverse transformation T^{-1}. The resulting ELBO in the transformed coordinates is

$$\mathcal{L} = \mathbb{E}_q[\log p(y, T^{-1}(\tilde{z})) + \log|\det J(\tilde{z})|] + \mathbb{H}[q(\tilde{z};\lambda)].$$

$\mathbb{H}[q(\tilde{z};\lambda)]$ is the entropy of the variational distribution.

ADVI employs Gaussian variational approximations for the transformed variable $\tilde{z}$, which may implicitly induce a non-Gaussian variational approximation in the original latent space z. The Gaussian assumption allows the entropy $\mathbb{H}[q(\tilde{z};\lambda)]$ to be computed analytically.

A simple but naive approximation assumes a fully-factorized variational distribution with independent latent variables: $q(\tilde{z};\lambda) = \prod_{k=1}^K \text{Normal}(\tilde{z}_k;\mu_k,\sigma_k^2)$. Because σ must always be positive, we parameterize as $\omega = \log(\sigma)$. The set of variational parameters in the mean-field approximation is therefore $\lambda_k = (\mu_k, \omega_k)$.

Another convenient yet more expressive option for the variational approximation is the multivariate or full-rank Gaussian, $q(\tilde{z};\lambda) = \text{MultivariateNormal}(\tilde{z};\mu,\Sigma)$, where Σ is the covariance matrix. Unlike the mean-field assumption, this allows arbitrary correlation structure between any pair of the latent variables. To ensure Σ is always positive semi-definite, we parameterize the covariance matrix as $\Sigma = LL^T$, where L is a lower triangular matrix with $\frac{K(K+1)}{2}$ unconstrained real-valued nonzero entries. In the full-rank case, the variational parameter becomes $\lambda = (\mu, L)$.

Now that we can freely optimize the ELBO in the real coordinate space, we require its gradient with respect to λ for stochastic gradient optimization. To enable automatic differentiation, we push the gradient operation inside the expectation, applying the "reparameterization trick" (Rezende et al., 2014; Kingma & Welling, 2014). We use the inverse Gaussian standardization S^{-1} to reparameterize $\tilde{z} = S_\lambda^{-1}(\epsilon)$ as a deterministic function of the standard Gaussian noise $\epsilon \sim N(0,1)$ given the variational parameters λ. For example, the standardization in the mean-field case is $\epsilon_k = S_{\lambda_k}(\tilde{z}_k) = \frac{\tilde{z}_k - \mu_k}{\exp(\omega_k)}$. This allows us to approximate the gradient using the Monte Carlo estimator:

$$\nabla_\lambda \mathcal{L} \approx \frac{1}{B}\sum_{b=1}^B \nabla_\lambda f(y,\epsilon^{(b)};\lambda) + \nabla_\lambda \mathbb{H}[q(\tilde{z};\lambda)], \tag{3}$$

$$f(y,\epsilon;\lambda) = \log p(y, T^{-1}(S_\lambda^{-1}(\epsilon))) + \log|\det J(S_\lambda^{-1}(\epsilon))|$$

Note that since the entropy is evaluated analytically, we do not need to reparameterize $\tilde{z}$ in the entropy term, as the gradient can also be computed analytically.

3. Soil Biogeochemical Models

Soil biogeochemical models (SBMs) are differential equation models that represent the dynamics of soil pool densities (Xie et al., 2020). We consider two classes of models: the linear conventional (CON) model and the nonlinear Allison-Wallenstein-Bradford (AWB) model (Allison et al., 2010). State variables x correspond to densities of elements in organic molecules that evolve over time t following an ODE. CO_2 emissions at time t can be estimated as a function of these state variables. Noisy measurements y are collected at potentially irregular time intervals. Model parameters θ correspond to other biological elements that govern the system. We define the joint distribution $p(\theta, x, y) = p(\theta)p(x|\theta)p(y|x, \theta)$.

The CON system consists of three state variables $x_t = (S_t, D_t, M_t)$: soil organic carbon (S) indicates the carbon density of stable organic soil molecules; dissolved organic carbon (D) denotes the carbon density contained in less stable, more decomposed organic molecule types; and microbial biomass carbon (M) describes the carbon density encompassed by the population of soil microbial organisms. The AWB system consists of four state variables $x_t = (S_t, D_t, M_t, E_t)$, which additionally includes extracellular enzyme carbon (E) that signifies the carbon density tied up in the enzymes secreted by the microbial organisms to help break down organic material in the system for consumption. The nonlinearity in AWB model comes from explicitly representing micorbial processes with nonlinear Michaelis-Menten functions (Wieder et al., 2015). The diagrams in figure 1 show the interaction between state variables and the various model parameters in the CON and AWB models.

To more realistically capture the stochasticity inherent to biological systems, we formulate the SDE parameterization of the CON and AWB models by adding noise to the system, which we call the stochastic CON (SCON) and stochastic AWB (SAWB) models. Concretely, we have the SDE

$$dx_t = \alpha(x_t, \theta, t)dt + \beta(x_t, \theta, t)dW_t, \tag{4}$$

where x_t is an M-dimensional vector of random variables, $\alpha(x_t, \theta, t)$ is an M-dimensional *drift vector*, $\beta(x_t, \theta, t)$ is an $M \times M$ *diffusion matrix*, and W_t is a standard Wiener process. The drift and diffusion depend on θ, a D-dimensional vector of unknown parameters.

Our drift vector α corresponds directly to the original ODE model (details in the supplement). We consider two versions of diagonal-only diffusion matrices: constant diffusion $\beta = I\sigma$ (SCON-C and SAWB-C) and state-scaling diffusion $\beta = I(\sigma \odot \sqrt{x})$ (SCON-SS and SAWB-SS), where $\odot$ denotes elementwise multiplication and $\sigma \in \theta$ is a model parameter that controls the noise of the dynamics.

In practice, all variables in the model are constrained to fall within biologically realistic intervals. For example, all latent states must be positive and activation energy parameters must be between 0 and 100.

The SDE trajectory x across a finite timespan can be discretized into a series of T equally spaced steps of length Δ_t. Transition densities between states at successive times are approximated as truncated Gaussian to enforce positivity:

$$p(x_t|x_{t-1}, \theta) = \text{TN}(x_t; x_{t-1} + \alpha(x_{t-1}, \theta)\Delta_t, \beta^2(x_{t-1}, \theta)\Delta_t, 0, \infty) \tag{5}$$

where $\text{TN}(x; \mu, \Sigma, a, b)$ denotes the truncated Gaussian density with mean μ, covariance Σ, lower bound a, and upper bound b. The transition likelihood is $p(x|\theta) = p(x_0|\theta)\prod_{t=1}^{T} p(x_t|x_{t-1}, \theta)$.

We place an informative logit-normal prior on model parameter θ based on expert knowledge, $p(\theta) = \text{LogitNormal}(\theta; \mu_\theta, \Sigma_\theta, a_\theta, b_\theta)$. Here, lower and upper bounds (a_θ, b_θ) vary across different parameters.

Finally, we assume truncated Gaussian observations, $p(y_t|x_t, \theta) = \text{TN}(y_t; \mu(x_t, \theta), \Sigma_y, a_y, b_y)$, where y_t is an N-dimensional vector of observations, $\mu(x_t, \theta)$ is a function of latent states x_t and model parameters θ, and Σ_y can either be a model parameter or a known constant. For example, the observations could be noisy measurements of CO_2, whose means are computed as a function of the states (see equations (17) and (21) in the supplement), or of the state variables themselves. The lower and upper bounds (a_y, b_y) can also vary across different types of observations. Letting $\mathcal{T}$ be the time steps where observations are available, the observation likelihood is given by $p(y|x, \theta) = \prod_{t\in\mathcal{T}} p(y_t|x_t, \theta)$.

4. ADVI for Soil Biogeochemical Models

The goal of SBM inference is to project changes in parameter values under warming climate conditions. We would like to infer the posterior of model parameters conditioned on post-warming treatment observations, $p(\theta|y)$. We can apply ADVI by defining state variables x and model parameters θ as our latent variables, $z = (\theta, x)$. We further assume factorized variational distribution $p(\theta, x|y) \approx q(\theta, x) = q(\theta)q(x)$. We proceed to infer posterior of model parameters θ and latent states of the SDE dynamics x jointly.

To ensure biologically realistic values, θ is constrained between some lower and upper bounds (a_θ, b_θ), while x is constrained to be positive. We apply the rescaled logit function to transform θ to the unconstrained real space: $\tilde{\theta} = T_\theta(\theta) = \text{logit}\left(\frac{\theta - a}{b - a}\right)$. This transformation implicitly induces a logit-normal variational approximation on θ. For x, we apply the inverse softplus transformation that maps positive values to the real space: $\tilde{x} = T_x(x) = \log(\exp(x) - 1)$.

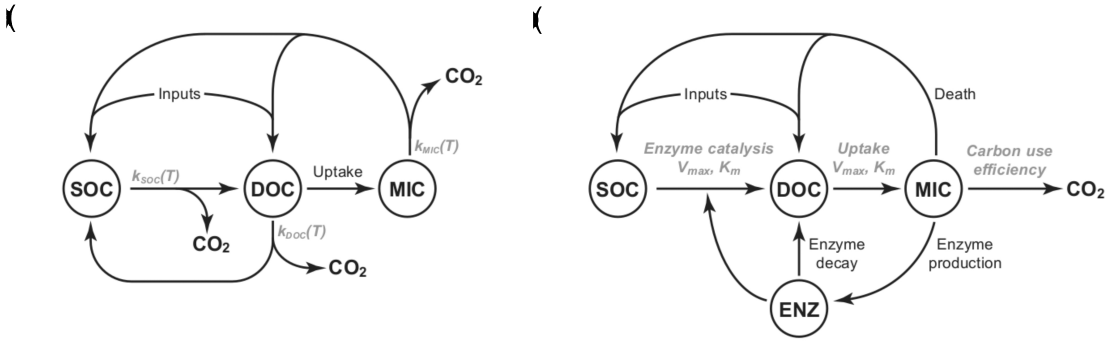


Figure 1. Diagrams of the (a) CON and (b) AWB models drawn from Allison et al. (2010). Latent states are shown within circles. The CON model consists of three latent states: SOC (S), DOC (D), and MIC (M). In addition to these same three states, the AWB model additionally includes ENZ (E).

The ADVI objective of Equation (3) becomes:

$$\begin{aligned} f(y, \epsilon; \lambda) = \log p(y, T_x^{-1}(S_{\lambda_x}^{-1}(\epsilon_x)), T_\theta^{-1}(S_{\lambda_\theta}^{-1}(\epsilon_\theta))) \\ + \log |\det J_x(S_{\lambda_x}^{-1}(\epsilon_x))| \\ + \log |\det J_\theta(S_{\lambda_\theta}^{-1}(\epsilon_\theta))| \end{aligned} \tag{6}$$

In practice, users do not need to derive this equation for new models, since the process is automated. They only need to specify the generative model, any constraints on the support of the latent variables, and the choice of desired variational approximation: mean-field, full-rank, or Gauss-Markov.

4.1. Gauss-Markov ADVI

A natural choice for the variational approximation of θ is the full-rank distribution to allow correlations to be captured between any arbitrary pair of model parameters. The full covariance representation scales quadratically in the number of parameters, but this is fine as both models are limited to a relatively small number of parameters.

However, for the sequential latent states x, neither full-rank nor mean-field is an ideal choice. Mean-field, although computationally very cheap (linear in the sequence length), is too unrealistic since it falsely assumes temporal independence. As we will show in our experiments, this may lead to biased parameter estimates. The full-rank approximation allows arbitrary correlations between any states at any time step, but this requires a covariance matrix of size $MT \times MT$ that scales quadratically in the length of the sequence, which could be very long. In addition to being costly, this also makes optimization sensitive to divergence or local optima due to having too many irrelevant parameters.

In sequential models with first-order Markov priors and observations that are local to single time points, as is the case in SDEs, we are guaranteed that the posterior is also first-order Markov. In this case, parameterizing the full covariance is wasteful as it has provably extraneous parameters. In order to more parsimoniously capture temporal dependencies, we propose to directly parameterize our variational approximation as a Markov chain. This results in a sampling procedure that scales linearly with respect to sequence length in both time and space.

Consider the set of variational parameters $\lambda_t = (\mu_t, A_t, L_t)$ for $t = 1, \ldots, T$, where $\mu_t \in \mathbb{R}^M$ is the mean at time t, $A_t \in \mathbb{R}^{M \times M}$ represents correlations between state variables at times $t-1$ and t, and $L_t \in \mathbb{R}^{M \times M}$ is a lower triangular covariance square-root matrix that represents correlations between different state variables at time t. Letting $\epsilon_t \sim N(0, I_M)$ be standard Gaussian noise of dimension M and $\eta_0 = L_0 \epsilon_0$, then for subsequent $t = 1, ..., T$:

$$\eta_t = A_t \eta_{t-1} + L_t \epsilon_t, \tag{7}$$

$$\tilde{x}_t = \mu_t + \eta_t. \tag{8}$$

The first line adds the temporal dependence, where the total stochasticity of $\tilde{x}_t$ is the sum of two sources: a linear transformation of the state variables from the previous time step $A_t \tilde{x}_{t-1}$, and independent noise at the current time step $L_t \epsilon_t$. The second line sets the mean μ_t. In general, the resulting sample x_t is correlated with all previous time points $x_{s<t}$ with covariance $\text{Cov}(x_s, x_t) = \prod_{i=s+1}^{t} A_i \text{Var}(x_s)$ without having to directly parameterize the full covariance matrix.

Because the states are defined by a Markov chain, the resulting entropy is a sum of conditional entropies, $\mathbb{H}[q(\tilde{x}; \lambda)] = \sum_{t=1}^{T} \mathbb{H}[q(\tilde{x}_t | \tilde{x}_{t-1}; \lambda_t)]$. Furthermore, the conditional entropy at time t is equal to that of a Gaussian with covariance $L_t L_t^T$, so:

$$\begin{aligned} \mathbb{H}[q(\tilde{x}; \lambda)] &= \sum_{t=1}^{T} \mathbb{H}[N(0, \Sigma_t = L_t L_t^T)] \\ &= \sum_{t=1}^{T} \frac{M}{2}(1 + \log(2\pi)) + \frac{1}{2} \log(\det(L_t L_t^T)). \end{aligned} \tag{9}$$

Similar to the inverse standardization, this procedure parameterizes $\tilde{x}$ as a deterministic function of standard Gaussian noise, which allows the gradient to be computed using automatic differentiation of the Monte Carlo estimator in (3).

Intuitively, the Gauss-Markov variational approximation is a middle ground between the mean-field and the full-rank approximations. It is more expressive than the mean-field approximation by allowing arbitrary first-order correlations. It enables more efficient inference than the full-rank approximation by encoding the Markov structure of the true posterior, reducing the effective number of parameters.

4.2. Related Work

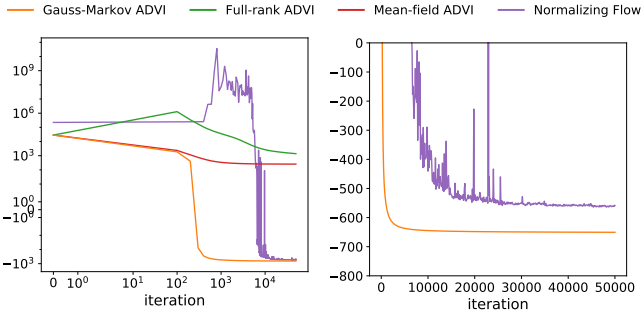
Prior work has developed black-box VI algorithms for other types of state space models with a Gaussian variational posterior that is parameterized by a tridiagonal inverse covariance matrix (Archer et al., 2015; Bamler & Mandt, 2017). It can be shown that our Gauss-Markov posterior also has a block tridiagonal inverse covariance, but we never explicitly construct this matrix. While we use tranformations to implicitly allow non-Gaussian posteriors, prior work assumed Gaussian posteriors. Archer et al. (2015) use amortized inference where the mean and inverse covariance of the Gaussian are determined by the output of a neural network with observations y as its input. The variational approximation is thus parameterized in a way that still requires matrix inversion. Although matrix inversion can be done more efficiently in sparse block-tridiagonal matrices, this is still more expensive than our parameterization that samples via a single forward recursion. To solve the linear system induced by the sparse inverse-covariance, Bamler & Mandt (2017) apply a forward and backward pass of temporal message-passing, while we only require a single forward recursion.

Ryder et al. (2018) proposed an algorithm for black-box variational inference for SDE that similarly works by discretizing the time and reframing the problem as state space models, then jointly optimizing for parameters and latent states. The authors use mean-field approximation on the model parameters θ and applies amortized inference for the latent states x. The variational posterior of the latent states are assumed to be Gaussian with mean and variance parameterized by the output of recurrent neural networks. A more recent follow-up work (Ryder et al., 2021) uses normalizing flows as the variational approximation to allow non-Gaussian posterior, and replaces recurrent neural networks with convolutional networks for faster inference.

5. Experiments

5.1. Simulated Data

We consider two variants of the SCON model: SCON-C and SCON-SS. We define informative priors on model parameters $p(\theta)$ and the initial condition $p(x_0)$ based on expert knowledge, and simulate data from the joint distribution $p(\theta, x, y) = p(\theta)p(x|\theta)p(y|x, \theta)$ with $T = 1000$ time steps. We sample noisy observations of S, D, M, and CO_2 every 5 time steps. We use regular gap between observations in our simulated data for simplicity, but this is not a requirement for our method.



(a) SCON-C (b) SCON-SS

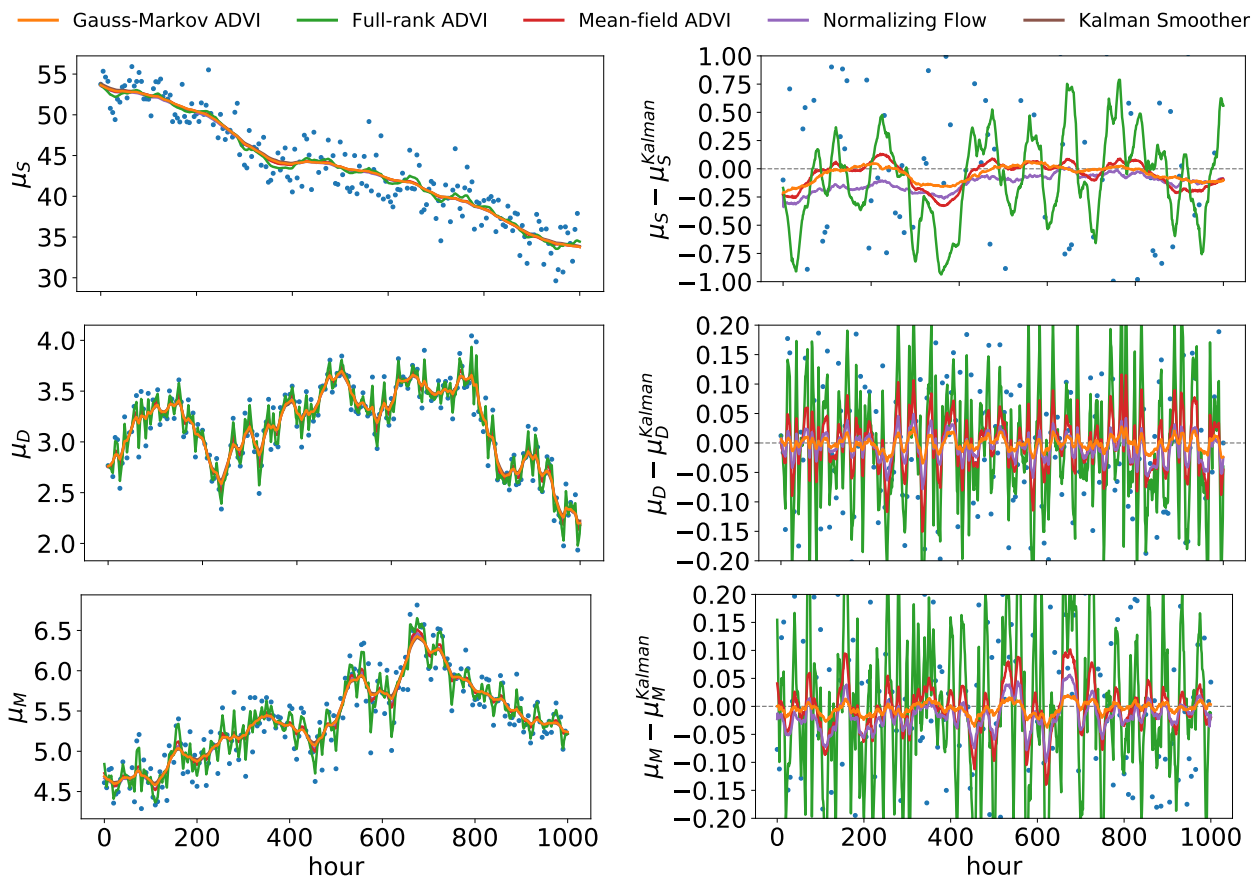
Figure 2. Trace plots of negative ELBO loss, estimated from 100 samples every 100 iterations. MF-ADVI converges to suboptimal loss, while FR-ADVI fails to converge within 50,000 iterations. In (b), we highlight that normalizing flow converges to worse loss than GM-ADVI and is substantially more noisy.

When the system is linear with Gaussian noise that does not depend on the state variables, the Kalman smoother (Kalman, 1960) can exactly recover the true posterior $p(x|y; \theta)$. We first use the constant diffusion model SCON-C to verify that our proposed method can recover the optimal Kalman smoother solution. We also compare Gauss-Markov ADVI (GM-ADVI) to three alternative variational approximations for the latent states: full-rank ADVI (FR-ADVI), mean-field ADVI (MF-ADVI), and an amortized inference method based on normalizing flows (Ryder et al., 2021). For the Kalman smoother, we fix θ using the true values used to generate the data. For the rest of the methods, θ and x are inferred jointly. For all these methods where θ is inferred, we use the full-rank variational approximation on θ with rescaled sigmoid transformation to enforce the (fixed) lower and upper bounds.

For SCON-SS, we drop the comparison with FR-ADVI and MF-ADVI since they perform poorly on the simpler model, and focus on the comparison between GM-ADVI and the normalizing flow. Note that although the SCON-SS model is linear, Kalman smoother cannot be applied here, since the state-scaling diffusion noise depends on state variables.

In all experiments, we fix a budget of 50 samples and 50,000 iterations for all methods. We use AdaGrad with learning rate 0.1 for ADVI and Adam with learning rate 0.01 for the normalizing flows.

GM-ADVI converges more quickly to better variational



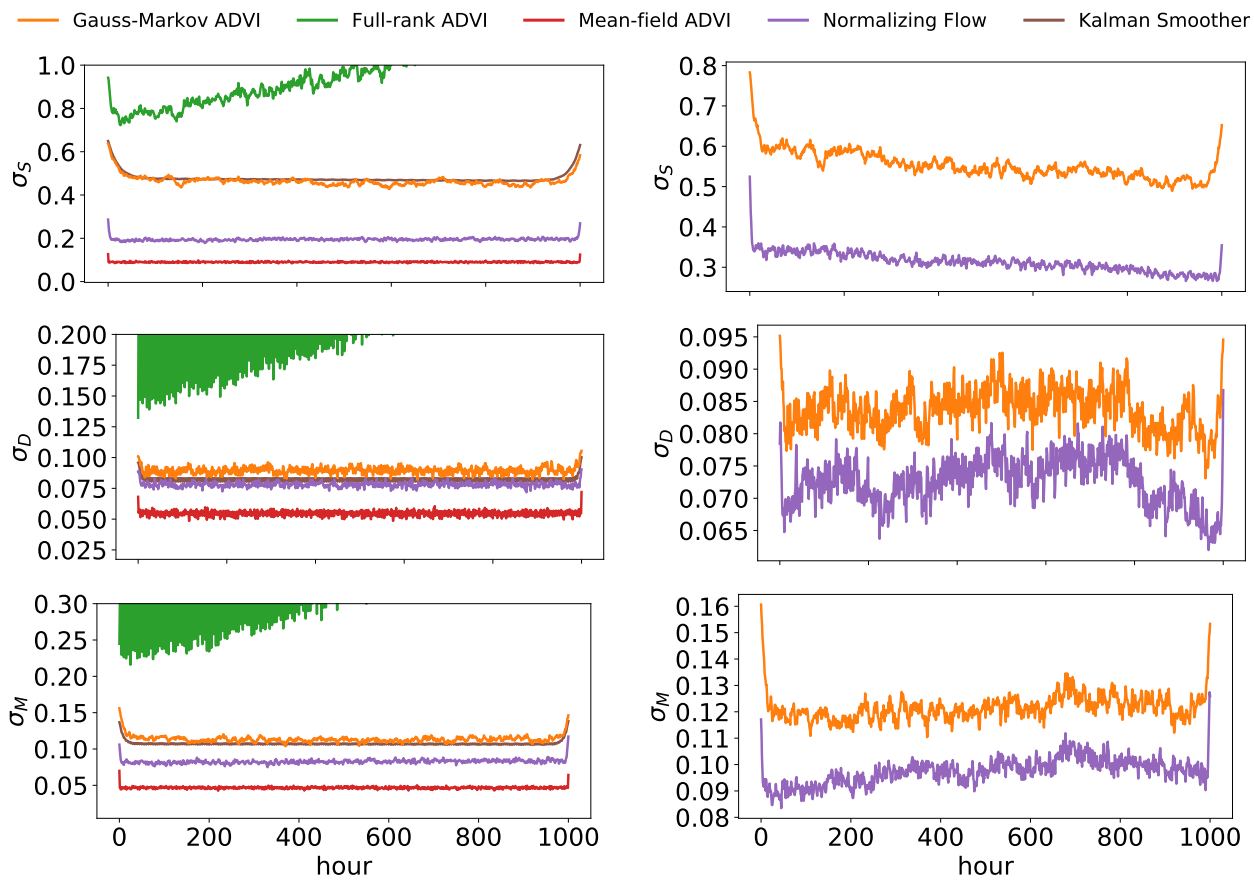
(a) Posterior mean estimates

(b) Errors in posterior mean estimates

Figure 3. Comparing variational mean of each latent state to the optimal Kalman smoother in the SCON-C model. (a) shows the posterior mean, while (b) shows the error in posterior mean, calculated as the difference between the estimated mean and the correct Kalman smoother mean. Blue dots represent observations. Posterior means are computed analytically for Kalman smoother. For variational methods, the means are estimated empirically with 1,000 samples drawn from the variational distribution $q(x)$. The posterior means recovered by Gauss-Markov ADVI align closely with those computed by the Kalman smoother. Model parameters θ are known to the Kalman smoother, but unknown to all variational methods.

bounds. Figure 2 compares the negative ELBO losses of different methods across inference iterations. The ELBOs are estimated via Monte Carlo sampling from 100 samples every 100 iterations. In figure 2(a), both GM-ADVI and the normalizing flow converge to much better (lower) loss compared to FR-ADVI and MF-ADVI. MF-ADVI converges to suboptimal loss due to false temporal independence assumption. Although potentially highly expressive, FR-ADVI fails to converge after 50,000 iterations, likely due to having too many irrelevant parameters that make it susceptible to noise and local optima. Compared to the normalizing flow, Gauss-Markov ADVI is much less noisy and still converges to better variational bounds, as highlighted in figure 2(b).

GM-ADVI more accurately captures the variance of the latent state posterior. Figures 3 and 4 compare the variational posterior of the latent states $q(x)$ recovered by the variational methods against the true posterior $p(x|y;\theta)$ computed by the Kalman smoother on the SCON-C and SCON-SS models. For Kalman smoother, the figure shows analytical means and standard deviations. For the variational methods, we use empirical estimates from 1,000 samples drawn from the variational posterior $q(x)$. Except for FR-ADVI which is highly noisy and clearly overfits the observations, all other methods may seem to perform reasonably well at estimating the posterior mean in figure 3(a). However, figure 3(b) highlights that GM-ADVI mean estimates are still the closest to the true Kalman smoother estimates; other methods are biased by noisy observations and insufficiently smooth. More importantly, GM-ADVI is the only approach that captures the appropriate amount of uncertainty in the posterior distribution, matching the Kalman smoother variance on all three states, as shown in figure 4(a). The full-rank approximation systematically overestimates



(a) SCON-C posterior standard deviation

(b) SCON-SS posterior standard deviation

Figure 4. Posterior standard deviation comparison in the SCON-C (a) and SCON-SS (b) models. Blue dots represent observations. Posterior standard deviations are computed analytically for Kalman smoother. For variational methods, the standard deviations are estimated empirically with 1,000 samples drawn from the variational distribution $q(x)$. The posterior standard deviations recovered by Gauss-Markov ADVI align closely with those computed by the Kalman smoother.

variance, while the mean-field and normalizing flow approximations systematically underestimate variance. Figure 4(b) shows the posterior standard deviation of GM-ADVI and the normalizing flow on the SCON-SS model.

GM-ADVI estimates the posterior of diffusion noise parameters more accurately. Figure 5 shows the marginal distributions of the variational posterior on select SCON-C model parameters θ. For all methods, we use the full-rank variational approximation on θ with rescaled sigmoid transformation to enforce the lower and upper bounds. We expect posteriors to shift away from the priors (blue) toward the true θ value (gray vertical line), but still display uncertainty. For all drift parameters (the first three rows), all methods seem to converge to similar posterior distributions. All methods other than GM-ADVI seem to overestimate diffusion parameters (c_S, c_D, c_M) compared to their true values. Figure 6 shows the marginal distributions of the variational posterior on SCON-SS diffusion parameters (s_S, s_D, s_M). We do not show the drift parameter comparison since both methods infer similar posterior distributions. Here, we see more clearly that the normalizing flow still overestimates diffusion parameters compared to GM-ADVI.

GM-ADVI is more practical for scientific applications. Compared to the normalizing flow, all ADVI methods share equal advantage of being simpler and more easily integrated to PPLs to enable automatic inference. They require fewer tuning hyperparameters, such as number of layers and choice of neural network architectures, among others. Although we show no direct runtime comparison in this paper, all ADVI experiments were run on the CPU, while the normalizing flow experiments required a GPU, which may not be easily accessible for many scientific applications. Compared to other ADVI approximations, Gauss-Markov offers balance between expressivity and efficiency: it is expressive without the extraneous parameters, while maintaining linear time and space.

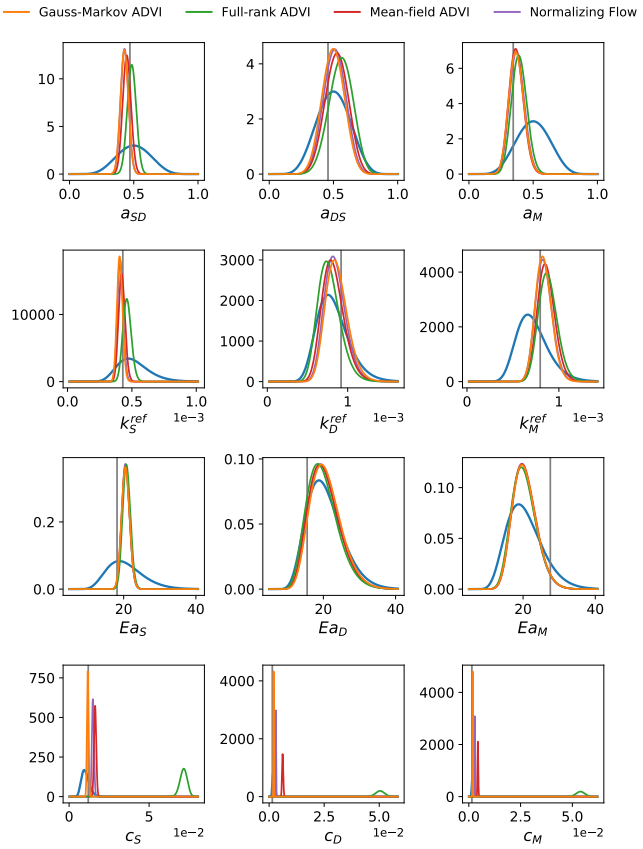


Figure 5. Marginal distributions of the variational posterior on select model parameters θ in the SCON-C model. Blue represents the prior distribution, while gray vertical lines show the true θ value used to simulate the data.

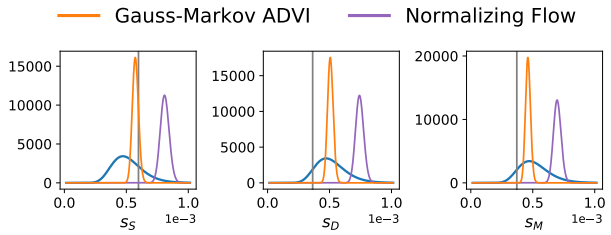
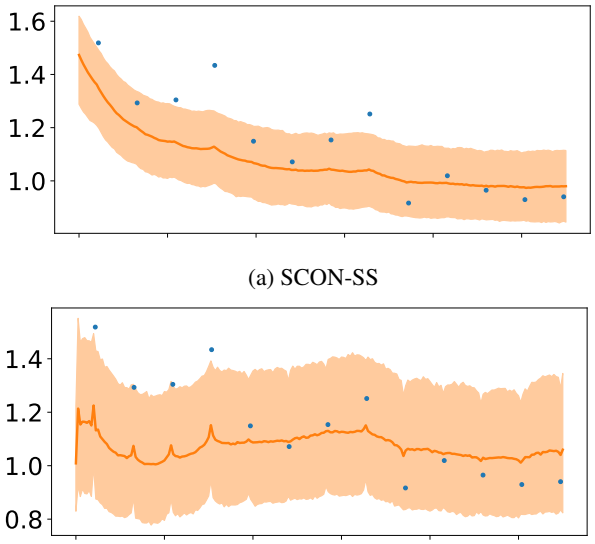


Figure 6. Posterior marginal distribution of SCON-SS diffusion parameters compared between GM-ADVI and the normalizing flow. Blue represents the prior distribution, while gray vertical lines show the true θ value used to simulate the data.



(a) SCON-SS

(b) SAWB-SS

Figure 7. GM-ADVI posterior of CO_2 response ratios on meta-analysis data. Lines show the median, while shaded regions show the 2.5th and 97.5th percentiles.

5.2. Meta-Analysis Data

Finally, we apply the proposed Gauss-Markov ADVI method on meta-analysis data used in (Xie et al., 2020). The dataset was compiled from 27 soil warming studies that measured CO_2. The pooled data consist of annual CO_2 response ratios over a period of 13 years. Each response ratio is calculated by dividing the annual CO_2 mean following warming perturbation by CO_2 measured at prewarming steady state. We consider the stochastic variants of the two models explored by Xie et al. (2020): SCON-SS and SAWB-SS.

We focus on Gauss-Markov ADVI in this case study and drop comparison with the remaining variational methods since they perform poorly on simulated data. We discretize the 13-year timespan into $T = 220$ discrete time steps. Following Xie et al. (2020), annual response ratios are assumed to be "collected" at the halfway point of each year. We modify our observation likelihood to use the CO_2 response ratios as observations (details in the supplement). As before, we infer latent states x and model parameters θ jointly. We run 20,000 inference iterations with 50 samples using AdaGrad with learning rate 0.1. We estimate the posterior of CO_2 response ratios by drawing 5,000 samples of $x^{(b)} \sim q(x)$ and $\theta^{(b)} \sim q(\theta)$, and evaluate the CO_2 response ratio corresponding to each sample b as a function of $x^{(b)}$ and $\theta^{(b)}$. Figure 7 shows the variational posterior of CO_2 response ratios.

With only a total of 13 data points, this meta-analysis dataset is substantially more sparse than the simulated data. However, the original simulation (Xie et al., 2020) using Hamiltonian Monte Carlo (Hoffman et al., 2014) and deterministic ODE solver still took multiple weeks to run. In comparison, our approach took less than an hour on the CPU.

6. Conclusion

As more effort is being made in biogeochemistry and other scientific fields to collect denser observations over longer periods of study, the need for faster and more scalable inference algorithms also becomes more crucial. For example, the ongoing Harvard Forest study (Melillo et al., 2017) that began in 1991 currently contains over 500 CO_2 observations over the span of 30 years and continues to grow. We have developed Gauss-Markov ADVI for sequential latent variable models that leverages the Markov structure of stochastic differential equations to arrive at an algorithm that is both expressive and scales linearly in the sequence length. We demonstrate its effective use on the time discretized soil biogeochemical models. Compared to other variational methods, Gauss-Markov ADVI converges faster and more accurately captures uncertainty in the posterior. It provides a compelling inference method for complex sequence models with interpretable latent variables and meaningful priors. It is designed to enable easy integration with PPLs, and in future work we would like to implement our algorithm in PPLs such as Stan or Pyro, to allow for convenient use by the scientific community. In the meantime, our Python code will be made available online.

Acknowledgements

This research was supported in part by a UC Irvine ICS Exploration Research Award, and by NSF Robust Intelligence Award No. IIS-1816365.

References

Abs, E., Leman, H., and Ferrière, R. A multi-scale eco-evolutionary model of cooperation reveals how microbial adaptation influences soil decomposition. *Communications biology*, 3(1):1–13, 2020.

Allison, S. D., Wallenstein, M. D., and Bradford, M. A. Soil-carbon response to warming dependent on microbial physiology. *Nature Geoscience*, 3(5):336–340, 2010. ISSN 1752-0908. doi: 10.1038/ngeo846. URL https://doi.org/10.1038/ngeo846.

Archer, E., Park, I. M., Buesing, L., Cunningham, J., and Paninski, L. Black box variational inference for state space models. *arXiv preprint arXiv:1511.07367*, 2015.

Bamler, R. and Mandt, S. Structured black box variational inference for latent time series models. *arXiv preprint arXiv:1707.01069*, 2017.

Browning, A. P., Warne, D. J., Burrage, K., Baker, R. E., and Simpson, M. J. Identifiability analysis for stochastic differential equation models in systems biology. *Journal of the Royal Society Interface*, 17(173):20200652, 2020.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1), 2017.

Chen, R. T., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Gan, Z., Henao, R., Carlson, D., and Carin, L. Learning deep sigmoid belief networks with data augmentation. In *Artificial Intelligence and Statistics*, 2015.

Golightly, A. and Wilkinson, D. J. Bayesian parameter inference for stochastic biochemical network models using particle markov chain monte carlo. *Interface focus*, 1(6): 807–820, 2011.

Gopalan, P., Hao, W., Blei, D. M., and Storey, J. D. Scaling probabilistic models of genetic variation to millions of humans. *Nature genetics*, 48(12):1587, 2016.

Hoffman, M. D., Gelman, A., et al. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

Ji, G., Cheng, D., Ning, H., Yuan, C., Zhou, H., Xiong, L., and Sudderth, E. B. Variational training for large-scale noisy-OR Bayesian networks. In *Conference on Uncertainty in Artificial Intelligence*, 2019.

Kalman, R. E. A new approach to linear filtering and prediction problems. *J. Basic Eng.*, 82(1):35–45, 1960.

Kidger, P., Morrill, J., Foster, J., and Lyons, T. Neural controlled differential equations for irregular time series. *Advances in Neural Information Processing Systems*, 33: 6696–6707, 2020.

Kingma, D. P. and Welling, M. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. M. Automatic differentiation variational inference. *Journal of machine learning research*, 2017.

Li, J., Wang, G., Mayes, M. A., Allison, S. D., Frey, S. D., Shi, Z., Hu, X.-M., Luo, Y., and Melillo, J. M. Reduced carbon use efficiency and increased microbial turnover with soil warming. *Global Change Biology*, 25(3):900–910, 2019. doi: https://doi.org/10.1111/gcb.14517. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/gcb.14517.

Li, X., Wong, T.-K. L., Chen, R. T., and Duvenaud, D. Scalable gradients for stochastic differential equations. In *International Conference on Artificial Intelligence and Statistics*, pp. 3870–3882. PMLR, 2020.

Melillo, J. M., Frey, S. D., DeAngelis, K. M., Werner, W. J., Bernard, M. J., Bowles, F. P., Pold, G., Knorr, M. A., and Grandy, A. S. Long-term pattern and magnitude of soil carbon feedback to the climate system in a warming world. *Science*, 358(6359):101–105, 2017.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and variational inference in deep latent gaussian models. In *International conference on machine learning*, volume 2, pp. 2. Citeseer, 2014.

Rubanova, Y., Chen, R. T., and Duvenaud, D. K. Latent ordinary differential equations for irregularly-sampled time series. *Advances in neural information processing systems*, 32, 2019.

Ryder, T., Golightly, A., McGough, A. S., and Prangle, D. Black-box variational inference for stochastic differential equations. In *International Conference on Machine Learning*, pp. 4423–4432. PMLR, 2018.

Ryder, T., Prangle, D., Golightly, A., and Matthews, I. The neural moving average model for scalable variational inference of state space models. In *Uncertainty in Artificial Intelligence*, pp. 12–22. PMLR, 2021.

Schirmer, M., Eltayeb, M., Lessmann, S., and Rudolph, M. Modeling irregular time series with continuous recurrent units. *arXiv preprint arXiv:2111.11344*, 2021.

Wainwright, M. J. and Jordan, M. I. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1:1–305, 2008.

Wang, S., Luo, Y., and Niu, S. Reparameterization required after model structure changes from carbon only to carbon-nitrogen coupling. *Journal of Advances in Modeling Earth Systems*, 14(4):e2021MS002798, 2022. doi: https://doi.org/10.1029/2021MS002798. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021MS002798. e2021MS002798 2021MS002798.

Wieder, W. R., Allison, S. D., Davidson, E. A., Georgiou, K., Hararuk, O., He, Y., Hopkins, F., Luo, Y., Smith, M. J., Sulman, B., et al. Explicitly representing soil microbial processes in earth system models. *Global Biogeochemical Cycles*, 29(10):1782–1800, 2015.

Xie, H. W., Romero-Olivares, A. L., Guindani, M., and Allison, S. D. A bayesian approach to evaluation of soil biogeochemical models. *Biogeosciences*, 17(15):4043–4057, 2020. doi: 10.5194/bg-17-4043-2020. URL https://bg.copernicus.org/articles/17/4043/2020/.

A. Transition Likelihood

The transition likelihod is given by:

$$p(x_t|x_{t-1},\theta) = \text{TN}(x_t; x_{t-1} + \alpha(x_{t-1},\theta)\Delta_t, \beta^2(x_{t-1},\theta)\Delta_t, 0, \infty)$$

where $\text{TN}(\mu, \Sigma, a, b)$ denotes the truncated Gaussian density with mean μ, covariance Σ, lower bound a, and upper bound b.

The diffusion matrix in the constant diffusion model (SCON-C and SAWB-C) is $\beta(x_{t-1},\theta) = I\sigma$. In the state-scaling model (SCON-SS and SAWB-SS), it is $\beta(x_{t-1},\theta) = I(\sigma \odot \sqrt{x})$, where $\odot$ denotes elementwise multiplication and $\sigma \in \theta$ is a model parameter. We define the drift function α below.

A.1. SCON Model

The drift vector in the SCON model obeys the following dynamics:

$$\alpha(x_{t-1},\theta) = \begin{bmatrix} I_{S,t} + a_{DS} \cdot k_{D,t} \cdot D_{t-1} + a_M \cdot a_{MS} \cdot k_{M,t} \cdot M_{t-1} - k_{S,t} \cdot S_{t-1} \\ I_{D,t} + a_{SD} \cdot k_{S,t} \cdot S_{t-1} + a_M \cdot (1 - a_{MS}) \cdot k_{M,t} \cdot M_{t-1} - (u_M + k_{D,t}) \cdot D_{t-1} \\ u_M \cdot D_{t-1} - k_{M,t} \cdot M_{t-1} \end{bmatrix} \tag{10}$$

with latent states $x_t = (S_t, D_t, M_t)$ and external input rate $(I_{S,t}, I_{D,t})$. The model parameters consist of both drift and diffusion parameters $\theta = (u_M, a_{DS}, a_{SD}, a_M, a_{MS}, k_{S,\text{ref}}, k_{D,\text{ref}}, k_{M,\text{ref}}, Ea_S, Ea_D, Ea_M, \sigma_S, \sigma_D, \sigma_M)$.

The k linear first-order decay parameters obey Arrhenius temperature dependence such that:

$$k_{i,t} = k_{i,\text{ref}} \exp\left[-\frac{Ea_i}{R}\left(\frac{1}{T_t} - \frac{1}{T_{\text{ref}}}\right)\right] \tag{11}$$

where R is the ideal gas constant 8.314 $\text{J} \cdot \text{K}^{-1} \cdot \text{mol}^{-1}$, T_t is the temperature at time t, and T_{ref} specifies a "reference" equilibrium temperature which we set at 283 K (for simulated data) or 283.15 K (for meta-analysis data).

A.2. SAWB Model

The drift vector in the SAWB model obeys the following dynamics:

$$\alpha(x_{t-1},\theta) = \begin{bmatrix} I_{S,t} + a_{MS} \cdot r_M \cdot M_{t-1} - \frac{V_{D,t} \cdot E_{t-1} \cdot S_{t-1}}{K_D + S_{t-1}} \\ I_{D,t} + (1 - a_{MS}) \cdot r_M \cdot M_{t-1} + \frac{V_{D,t} \cdot E_{t-1} \cdot S_{t-1}}{K_D + S_{t-1}} + r_L \cdot E_{t-1} - \frac{V_{U,t} \cdot M_{t-1} \cdot D_{t-1}}{K_U + M_{t-1}} \\ u_{Q,t} \cdot \frac{V_{U,t} \cdot M_{t-1} \cdot D_{t-1}}{K_U + M_{t-1}} - (r_M + r_E) \cdot M_{t-1} \\ r_E \cdot M_{t-1} - r_L \cdot E_{t-1} \end{bmatrix} \tag{12}$$

with latent states $x_t = (S_t, D_t, M_t, E_t)$ and external input rate $(I_{S,t}, I_{D,t})$. The model parameters consist of both drift and diffusion parameters $\theta = (u_{Q,ref}, Q, a_{MS}, K_D, K_U, V_{D,\text{ref}}, V_{U,\text{ref}}, Ea_{V_D}, Ea_{V_U}, r_M, r_E, r_L, \sigma_S, \sigma_D, \sigma_M, \sigma_E)$.

The V reaction rate parameters are forced by temperature via Arrhenius temperature dependence such that:

$$V_{i,t} = V_{i,\text{ref}} \exp\left[-\frac{Ea_{V_i}}{R}\left(\frac{1}{T_t} - \frac{1}{T_{\text{ref}}}\right)\right] \tag{13}$$

The u_Q carbon use efficiency fraction follows linear temperature dependence such that:

$$u_{Q,t} = u_{Q,\text{ref}} + Q \cdot (T_t - T_{\text{ref}}) \tag{14}$$

B. Observation Likelihood

The observation likelihood is given by:

$$p(y_t|x_t,\theta) = \text{TN}(y_t; \mu(x_t), \Sigma_y, a_y, b_y)$$

In all experiments, we fix lower and upper bounds (a_y, b_y).

In experiments with simulated data, our observations are noisy measurements of the state variables and CO_2. We fix the observation covariance Σ_y, and the observation mean is computed as a function of the state variables:

$$\mu(x_t, \theta) = \begin{bmatrix} S_t & D_t & M_t & r_{\mathrm{CO}_2,t} \end{bmatrix}^T \tag{15}$$

In the meta-analysis data, we observe only the CO_2 response ratios. We assume prior $\Sigma_y \sim \text{HalfCauchy}(1)$ for the observation covariance and infer its posterior. The observation mean is:

$$\mu(x_t, \theta) = \frac{r_{\mathrm{CO}_2,t}}{\hat{r}_{\mathrm{CO}_2}} \tag{16}$$

where $\hat{r}_{\mathrm{CO}_2} = g(\hat{x}, \theta)$ is the CO_2 computed at pre-warming steady state. We define CO_2 and steady state solutions $\hat{x}$ in SCON and SAWB models below.

B.1. SCON Model

The CO_2 at time t in SCON is given by the equation:

$$r_{CO_2,t} = g(x_t, \theta) = (1 - a_{SD}) \cdot k_{S,t} \cdot S_t + (1 - a_{DS}) \cdot k_{D,t} \cdot D_t + (1 - a_M) \cdot k_{M,t} \cdot M_t \tag{17}$$

The SCON respiration can be thought of as the sum of the fractions of carbon fluxes not transferred into other state variables at t.

The steady state solutions for the state variables in SCON are:

$$\hat{D} = \frac{a_{SD} I_S + I_D}{u_M + k_D + u_M a_M (a_{MS} - a_{MS} a_{SD} - 1) - a_{DS} k_D a_{SD}} \tag{18}$$

$$\hat{S} = \frac{I_S + D_0 (a_{DS} + k_D + u_M a_M a_{MS})}{k_S} \tag{19}$$

$$\hat{M} = \frac{D_0 u_M}{k_M} \tag{20}$$

B.2. SAWB Model

The CO_2 at time t in SAWB is given by the equation:

$$r_{CO_2,t} = g(x_t, \theta) = (1 - u_{Q,t}) \cdot \frac{V_{U,t} \cdot M_t \cdot D_t}{K_U + D_t} \tag{21}$$

The SAWB respiration can be thought of as the fraction of carbon flux out of the DOC pool that is not absorbed by the microbial biomass MBC.

The steady state solutions for the state variables in SAWB are:

$$\hat{S} = \frac{-r_L K (I_S (r_M (1 + E_C (a_{MS} - 1)) + r_E (1 - E_C) + E_C I_D a_{MS} r_M)}{I_S (r_M (r_L (1 + E_C (a_{MS} - 1))) + r_E r_L (1 - E_C) - E_C V)) + E_C I_D (a_{MS} r_M r_L - r_E V)} \tag{22}$$

$$\hat{M} = \frac{E_C (I_D + I_S)}{(1 - E_C)(r_M + r_E)} \tag{23}$$

$$\hat{D} = \frac{-K_U (r_M + r_E)}{r_M + r_E - E_C V_U} \tag{24}$$

$$\hat{E} = \frac{r_E E_C (I_D + I_S)}{r_L (1 - E_C)(r_M + r_E)} \tag{25}$$