



RESEARCH ARTICLE

10.1029/2024MS004542

Key Points:

- We adapt the Bayesian inference approach “calibrate, emulate, sample” to estimate parameters in the land surface model CLM-FATES
- The framework leverages computationally efficient and uncertainty-aware inference in a multidimensional parameter space
- We jointly assimilate water and carbon flux data and improve model agreement with synthetic and real observations at a boreal forest site

Supporting Information:

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

Correspondence to:

L. T. Keetz,
lassekee@gmail.com;
l.t.keetz@geo.uio.no

Citation:

Keetz, L. T., Aalstad, K., Fisher, R. A., Poppe Terán, C., Naz, B., Pirk, N., et al. (2025). Inferring parameters in a complex land surface model by combining data assimilation and machine learning. *Journal of Advances in Modeling Earth Systems*, 17, e2024MS004542. <https://doi.org/10.1029/2024MS004542>

Received 28 JUN 2024

Accepted 22 MAY 2025

© 2025 The Author(s). Journal of Advances in Modeling Earth Systems published by Wiley Periodicals LLC on behalf of American Geophysical Union. This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

Inferring Parameters in a Complex Land Surface Model by Combining Data Assimilation and Machine Learning

L. T. Keetz^{1,2} , K. Aalstad¹ , R. A. Fisher³ , C. Poppe Terán^{4,5} , B. Naz⁴ , N. Pirk¹ , Y. A. Yilmaz¹ , and O. Skarpaas² 

¹Department of Geosciences, University of Oslo, Oslo, Norway, ²Natural History Museum, University of Oslo, Oslo, Norway, ³CICERO Center for International Climate Research, Oslo, Norway, ⁴Institute of Bio- and Geosciences: Agrosphere (IBG-3), Research Centre Jülich, Jülich, Germany, ⁵Faculty of Georesources and Materials Engineering, Rheinisch-Westfälische Technische Hochschule (RWTH), Aachen, Germany

Abstract Complex Land Surface Models (LSMs) rely on a plethora of parameters. These parameters and the associated process formulations are often poorly constrained, which hampers reliable predictions of ecosystem dynamics and climate feedbacks. Robust and uncertainty-aware parameter estimation with observations is complicated by, for example, the high dimensionality of the model parameter space and the computational cost of LSM simulations. Herein, we adapt a novel Bayesian data assimilation (DA) and machine learning framework termed “calibrate, emulate, sample” (CES) to infer parameters in a widely-used LSM coupled with a demographic vegetation model (CLM-FATES). First, an iterative ensemble Kalman smoother provides an initial estimate of the posterior distribution (“calibrate”). Subsequently, a machine-learning-based emulator is trained on the resulting model-observation mismatches to predict outcomes for unseen parameter combinations (“emulate”). Finally, this emulator replaces CLM-FATES simulations in an adaptive Markov Chain Monte Carlo approach enabling computationally feasible posterior sampling with enhanced uncertainty quantification (“sample”). We test our implementation with synthetic and real observations representing a boreal forest site in southern Finland. We estimate a total of six plant-functional-type-specific photosynthetic parameters by assimilating evapotranspiration (ET) and gross primary production (GPP) flux data. CES provided the best estimates of the synthetic truth parameters when compared to data-blind emulator sampling designs while all approaches reduced model-observation errors compared to a default parameter simulation (GPP: −10% to −30%, ET: −4% to −6%). Although errors were also consistently reduced with real data, comparing the emulator designs was less conclusive, which we mainly attribute to equifinality, structural uncertainty within CLM-FATES, and/or unknown errors in the data that are not accounted for.

Plain Language Summary Land surface models (LSMs) are important tools for simulating interactions between climate and terrestrial systems, and we need to improve them for more reliable climate and ecological predictions. LSM equations include adjustable model parameters that can be tuned to better match predictions with available environmental observations. As modeling complex processes involves many uncertainties and gathering sufficient data is challenging, the default parameter values are often poorly constrained. Data assimilation provides a statistical toolbox for estimating model parameters and their uncertainties from available observations while considering measurement inaccuracies. A recently presented approach termed “calibrate, emulate, sample” (CES) combines powerful data assimilation techniques and machine learning, which allows us to tackle some of the main challenges we traditionally face in LSM parameter estimation. For example, CES only requires running the computationally expensive models a few hundred rather than hundreds of thousands of times. Herein, we estimate important photosynthesis parameters to better reproduce land surface carbon and water exchange observations from a boreal forest in southern Finland. Our results with artificial data suggest that the method is suitable for estimating the complex LSM's parameters. However, experiments with real observations were less conclusive. Follow-up work thus needs to better account for real-world complexities.

1. Introduction

Land surface models (LSMs) rely on accurate representations of highly non-linear and spatiotemporally heterogeneous environmental processes. For instance, empirical studies consistently illustrate that carbon fluxes are dependent on available water and energy (Pérez-Ruiz et al., 2021), surface albedo is sensitive to plant type

distribution (Leonardi et al., 2015), carbon dioxide fertilization of ecosystems is limited by nutrient availability and other factors (Terrer et al., 2019), and biogeochemical feedbacks can influence the hydrological cycle (Yang et al., 2021). Although LSMs constitute state-of-the-art tools for comprehensively simulating terrestrial systems and—when coupled to Earth System Models (ESMs)—their climate feedbacks, predictions often remain inconsistent with observational data and long-term projections vary considerably (Fisher & Koven, 2020; Friedlingstein et al., 2022; Koven et al., 2022; Seiler et al., 2022). Persistent modeling challenges include inadequate representations of soil carbon content and vegetation characteristics such as leaf area index (LAI) and plant type distributions (Collier et al., 2018; Horvath et al., 2021; Seiler et al., 2022). Considering the important role the vegetated land surface has played in mitigating anthropogenic greenhouse gas emissions and the large uncertainties associated with future trajectories of terrestrial land carbon sinks, we need to overcome such model shortcomings for more accurate climate and carbon budget projections (Fernández-Martínez et al., 2023; IPCC, 2023; Seiler et al., 2022).

Due to the intricate cross-linkages inherent within the Earth system, many LSMs are highly complex and represent many interacting processes. Consequently, they incorporate a myriad of model parameters, most of which are difficult or impossible to directly observe in the field (Bonan & Doney, 2018; Sawada, 2020). In the case of vegetation parameters, prior knowledge published in the scientific literature and databases (e.g., Kattge et al., 2020) is often limited to a small number of species or locations and can be subject to substantial local and seasonal variations (Massoud et al., 2019; Rogers, 2014). Various approaches exist to calibrate model parameters, that is, to algorithmically adjust their values to minimize the mismatch between independent observations and the model outputs (Tsai et al., 2021). The high computational cost of running ESMs and LSMs, which resolve energy and gas exchange at sub-hourly timescales across the entire globe, means that these models are not especially amenable to many established sampling-based calibration techniques that can require hundreds of thousands of model simulations (Cleary et al., 2021; O'Hagan, 2006). ESM and LSM calibration has traditionally often been conducted via “hand tuning” of key parameters (cf. Helgeson et al., 2021). In practice, it is thus still challenging to formally compare model structures, identify sources of persistent biases, and understand the reasons for model spread in future projections.

The need for more robust and generalizable methods to calibrate LSM parameters is increasingly recognized (Raoult, Beylat, et al., 2024). For example, numerous efforts have been made, or are underway, to quantify model output sensitivities to extensive parameter perturbations, which can help guide LSM uncertainty quantification and calibration (e.g., Dagon, 2020; Massoud et al., 2019; McNeall et al., 2024; Zaehle & Friend, 2010). Fundamentally, robust parameter estimation of complex models such as LSMs requires automated, computationally tractable, and uncertainty-aware methods that can fully leverage data from Earth observations (Kumar et al., 2022). Data assimilation (DA) methods built on the foundation of Bayesian inference are a particularly promising class of such methods (Evensen et al., 2022; Wikle & Berliner, 2007). The high computational cost associated with complex process model simulations has long also been a bottleneck for many Bayesian approaches (Van Oijen, 2017). However, this limitation has increasingly been circumvented with machine learning methods trained to predict model outputs, model-data mismatches, or other summary statistics from the model input parameters as input features (Fer et al., 2018; Massoud, 2019; O'Hagan, 2006). Replacing full LSM runs with these computationally cheap emulators (also called surrogate models) enables parameter estimation and sensitivity analyses not only in Bayesian (Fer et al., 2018; Huang et al., 2016; Sawada, 2020; Watson-Parris et al., 2021), but also in non-Bayesian frameworks (Dagon et al., 2020; Li et al., 2023).

Bayesian inference with machine learning surrogate models has yielded promising results for estimating LSM parameters (Fer et al., 2018; Huang et al., 2016; Sawada, 2020). Nevertheless, finding computationally feasible solutions for reliable global parameter estimation with reasonable uncertainty quantification remains challenging (Raoult, Beylat, et al., 2024). A central prerequisite for emulator-based Bayesian inference is constructing accurate and well-generalizing surrogate models. Although studies regularly report high predictive performances for their LSM emulators (e.g., Dagon et al., 2020; Li et al., 2023), these commonly only incorporate a handful of input features (i.e., adjustable model parameters) where a prior reduction in parameter space dimensionality is typically derived from model sensitivity analyses (Lu et al., 2017; McNeall et al., 2024). Ideally, we want to estimate many more parameters simultaneously through Bayesian inference to fully capture their (non-linear) interactive effects on different modeled outputs. With higher input parameter dimensionality, however, LSM emulation becomes increasingly difficult and inaccurate (Massoud, 2019). To generate training data for an emulator, one needs to run an ensemble of model simulations to generate samples of the parameter-to-model-data

map. Although such simulations are computationally parallelizable, even sophisticated space-filling sampling designs, such as Latin hypercube sampling (LHS) (McKay et al., 1979), still require unfeasibly large numbers of ensemble members to sample high-dimensional parameter spaces effectively (Cleary et al., 2021; Wang et al., 2014). From a Bayesian perspective, we are only interested in the typical set (regions of high probability mass) of the posterior distribution (MacKay, 2003; Murphy, 2023) that can be expected to occupy only a small fraction of this large parameter space, which is why creating samples without considering the data can easily become wasteful and insufficient.

To address many of these challenges, Cleary et al. (2021) introduced the Bayesian DA and machine learning framework “calibrate, emulate, sample” (CES), which provides a strong theoretical basis for uncertainty-aware parameter estimation in non-linear models with high parameter dimensionality. As the name suggests, the workflow sequentially combines three major conceptual components.

The “calibrate” step generates an initial estimation of the constrained parameter space by exploiting the robustness of ensemble Kalman approaches. This family of Bayesian DA methods has demonstrated remarkable empirical success in the geosciences (Carrassi et al., 2018; Katzfuss et al., 2016). Crucially, unlike variational Bayesian techniques, ensemble Kalman methods are gradient-free and thus enable inverse modeling without adjoint models which can be challenging to obtain for complex models like LSMs (Garbuno-Inigo et al., 2020; Mamejtanov et al., 2012). The ensemble Kalman-based approach we use herein, an iterative Ensemble Kalman smoother (IEnKS; Section 2.2.7), serves as a fast yet powerful approximate sampler that can cluster a parameter ensemble near the typical set of the posterior distribution. However, using this method alone has several limitations. Like most ensemble Kalman methods, the IEnKS is based on Gaussian linear assumptions, that is, that the forward model is linear, the prior distribution is Gaussian, and the observation error is Gaussian. Ensemble Kalman methods can still work well in practice even when these assumptions are violated (Carrassi et al., 2018; Evensen et al., 2022). Nonetheless, they will in practical non-linear LSM applications always remain approximations for sampling from the posterior and are limited by the number of ensemble members (i.e., samples) we can feasibly generate with the costly LSMs. Ergo, the IEnKS does not provide systematic uncertainty quantification (Cleary et al., 2021). CES includes the “emulate” and “sample” steps to overcome these drawbacks.

Together, the “emulate” and “sample” steps enable computationally feasible Markov chain Monte Carlo (MCMC) sampling, a gold standard family of algorithms for Bayesian inference of the posterior distribution with comprehensive uncertainty quantification (Helgeson et al., 2021). MCMC enables exact sampling (hence “sample”) from the posterior without requiring an analytical representation of the full normalized posterior distribution. One only needs to be able to determine the point-wise density of the unnormalized posterior for particular parameter sets, that is, calculate the product of the likelihood and prior for a specific ensemble member (Neal, 1993; Van Ravenzwaaij et al., 2018). However, achieving convergence requires many—typically in the order of hundreds of thousands—full forward model runs to evaluate the likelihood of different parameter sets sequentially, with limited potential for computational parallelization. This is not practically feasible for computationally expensive forward models like LSMs. Here, the “emulate” step comes into play: It replaces the forward model simulations with a computationally cheap surrogate model as outlined before (O’Hagan, 2006). Conceptually, the “calibrate” step can therefore be regarded as a data-driven experimental design optimization for obtaining a better emulator. By concentrating the model ensemble closer to the typical set of the true posterior region with the IEnKS (“calibrate”), we aim to focus on and improve LSM emulator predictions for this region (“emulate”) within MCMC sampling (“sample”).

The combination of what Cleary et al. refer to as “emulate” and “sample” has already become a promising workflow for LSM parameter estimation (Fer et al., 2018; Raoult, Beylat, et al., 2024; Sawada, 2020). In our study, we wish to emphasize the importance of an initial “calibrate” step involving DA to help guide the design of the emulator. Such an initial approximate Bayesian inference step could relatively easily be added to existing LSM parameter estimation workflows without a substantial increase in computational cost (cf. Section 2.2.7). We rely on controlled experiments with synthetic data to demonstrate the theoretical added value of such an initial “calibrate” step. The framework is then tested “out-of-the-box” (i.e., with a default model setup) in less controlled real experiments with gap-filled observational data (Section 3.2.2). This simplistic setup may serve as a starting point for research aiming to relax the strong assumptions in our experimental design going forward.

To the best of our knowledge, this work constitutes the first attempt at estimating parameters of a complex LSM with the CES framework, which is increasingly gaining traction (e.g., Dunbar et al., 2024). Following Cleary

et al. (2021), we hypothesize that (a) an iterative ensemble Kalman smoother scheme (“calibrate”) will provide a good initial approximation of samples from the constrained LSM parameter space, (b) based on these samples, a machine learning regression model can learn the relationship between the model input parameters and the resulting likelihood surface to make accurate predictions for unseen parameter combinations in the parameter region of interest (“emulate”), (c) replacing the complex LSM with this emulator will enable an adaptive MCMC approach to provide an improved estimate of the joint posterior distribution allowing for comprehensive uncertainty quantification (“sample”). We first describe our adaptation of the CES framework tailored to estimating six photosynthetic parameters in the CLM-FATES model. Subsequently, we present experiments that scrutinize different framework components by assimilating synthetic and actual water and carbon surface flux observations at a boreal forest site.

2. Materials and Methods

2.1. A Complex Land Surface Model: CLM-FATES

2.1.1. Model Description

We use the vegetation demographics model Functionally Assembled Terrestrial Ecosystem Simulator (FATES; Fisher et al., 2015; Koven et al., 2020) embedded in the LSM Community Land Model, version 5 (CLM; Lawrence et al., 2019) as our example of a complex LSM. CLM is a widely used LSM for global and regional climate modeling. FATES is a comprehensive representation of the vegetation-mediated elements of the terrestrial system. It is a vegetation demographic model that, in its most complex configuration, represents the heterogeneous dynamics of ecosystems by clustering individual plants into bins of cohorts with similar plant functional type (PFT), size class, and disturbance history. It is tightly coupled with CLM. When FATES is activated, CLM performs the surface hydrology, energy and gas exchange, and soil biogeochemistry calculations, while FATES handles all processes related to living vegetation, seeds, fires, and fuels.

2.1.2. Model Configuration

Across all simulations, we utilized CLM’s point-scale (single-site) setting in “land-only” mode. We extracted atmospheric forcing (GSWP3v1; <http://hydro.iis.u-tokyo.ac.jp/GSWP3/>) and land surface characteristics (e.g., soil texture) at the site location (cf. Section 3.2) from CLM5 default data sets (Lawrence et al., 2019). Refer to CLM5’s documentation (Lawrence et al., 2019) and the data generation scripts and surface data file contained in our public code repository for additional details (Keetz & Aalstad, 2024). All experiments use 10-year atmospheric forcing from 2004 to 2013 (cf. Section 2.2.5). Note that we do not explicitly account for the various input data uncertainties in this pilot study, which will be further discussed in the Outlook section.

We ran FATES in a reduced complexity “satellite phenology” mode (FATES-SP). Accordingly, vegetation distributions as per static PFT area fractions, PFT canopy heights, and fixed annual LAI and stem area index cycles (derived from multi-annual averages) are prescribed to grid cells based on satellite observations (Fisher et al., 2023). With FATES-SP, we reduce the degrees of freedom compared to the full-complexity model, which allows us to focus the parameter estimation on the simulation of gas exchange processes while reducing certain potential model error sources such as PFT growth and competition for resources. This strategy has previously been employed in LSM parameter estimation studies, such as Dagon et al. (2020) who used the closely related SP mode in CLM5. Unlike full-complexity FATES, the mode does not explicitly model PFT growth, mortality, disturbance, competition for canopy space, nitrogen limitations, or the full (mass-conserving) carbon cycle. It represents vegetation spatial composition and structure with one distinct patch per PFT and one cohort per patch (Fisher et al., 2023; Lawrence et al., 2019). Figure S1b in Supporting Information S1 depicts the prescribed PFTs and their maximum annual LAI values for the study site in relation to available field observations.

Moreover, we assigned the mean value of the half-hourly flux tower CO₂ concentration observations from 2004 to 2013 (389 ppmv) as the constant ambient atmospheric CO₂ concentration (“CCSM_CO2_PPMV”; Figure S8 in Supporting Information S1; Section 3.2). We further changed the depth to bedrock surface data field (“zbedrock”) to 1.5 m to ensure a more realistic representation of the observed soil depth at the site (Hari & Kulmala, 2005). All simulations were preceded by a 10-year spin-up period using the same GSWP3v1 forcing to start the analyses from a steady model state (cf. Liao et al., 2023). Note that through our use of FATES-SP,

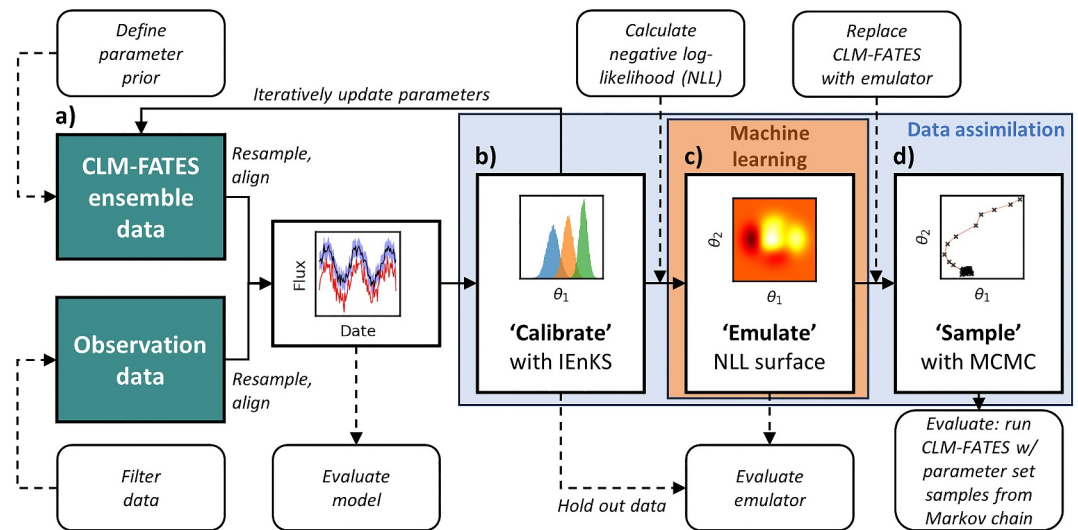


Figure 1. CES implementation overview. (a) A CLM-FATES perturbed parameter ensemble is created using parameter sets sampled from a prior distribution (Section 2.2.4); the observations and the corresponding model predictions are resampled to daily values and aligned regarding dates and units (Section 2.2.5). (b) An iterative ensemble Kalman smoother (IEnKS) gradually updates the model parameters through approximate Bayesian inference (Section 2.2.7). (c) A machine-learning-based emulator is trained on the IEnKS calibration results to predict negative log-likelihoods for any parameter combination within the parameter bounds (Section 2.2.8). (d) The emulator replaces CLM-FATES simulations in an adaptive Markov chain Monte Carlo algorithm to estimate the full joint posterior distribution; a final model ensemble is generated using random subsamples from the parameter sets in the Markov chain for evaluation (Section 2.2.9).

the spin-up is mostly important for state variables related to the soil hydrology, which appeared reasonably stable with the chosen period.

2.1.3. Estimated Parameters

We estimated PFT-specific FATES parameters that are known to substantially and independently impact the variables we assimilate, evapotranspiration (ET) and gross primary productivity (GPP) (Massoud et al., 2019): The maximum carboxylation capacity at 25°C at the canopy top (V_{cmax}) (e.g., Walker et al., 2014), and the slope defining the relationship between stomatal conductance and the Ball Index in the Ball-Berry stomatal conductance model (BB_{slope}) (Ball et al., 1987; Miner & Bauerle, 2017). Accordingly, we infer a total of six parameters for the three PFTs prescribed at the site in default FATES-SP mode (Table S1, Figure S1b in Supporting Information S1): needleleaf evergreen trees (PFT_{needle}), broadleaf cold-deciduous trees (PFT_{broad}), and C3 grass (PFT_{grass}). We refer to the FATES technical documentation for the implementation details (Fisher et al., 2023). In future implementations, both the observations and parameters can be expanded to encompass a wider range of processes.

2.2. “Calibrate, Emulate, Sample” Adaptation

After briefly introducing parameter estimation in the context of Bayesian inference, this section presents the theory behind CES for LSM parameter estimation and our adaptations to integrate the framework with CLM-FATES (Figure 1). We then outline the numerical experiments we used to analyze the individual framework components and provide rough computational runtime estimates.

2.2.1. Bayesian Inference

Let the random vector θ denote the set of N_p model parameters that we are uncertain about. For simplicity, we make the so-called strong constraint (perfect model) assumption that the forward model, CLM-FATES in this case, can map perfectly from parameters θ to noise-free observations y (Evensen et al., 2022). Although this is clearly a strong assumption that ignores many other sources of uncertainty (M. Dietze, 2017), it is commonly adopted in practice as a natural first step toward addressing more general weak constraint problems (Evensen

et al., 2022). It is also becoming clear that parameter uncertainty is an important, though certainly not the only, source of uncertainty in land surface modeling (D. Kennedy et al., 2024; N. Raoult, Beylat, et al., 2024). Since real observations are uncertain due to noise in various forms, the forward mapping can be expressed as follows

$$\mathbf{y} = \mathcal{M}(\boldsymbol{\theta}) + \boldsymbol{\epsilon}, \quad (1)$$

where the vector $\mathbf{y}$ contains the N_o observations, $\mathcal{M}(\cdot)$ is the forward model operator, and $\boldsymbol{\epsilon}$ is the observation error under the strong constraint assumption. Through (1), parameter estimation is now naturally cast as an inverse problem where we seek to invert the forward model to solve for the parameters that generated the observations. As emphasized by Stuart (2010), this inverse problem is ill-posed, in part since it is usually under ($N_p > N_o$) or over ($N_p < N_o$) determined, but more generally due to the presence of observation noise whereby infinitely many choices of $\boldsymbol{\theta}$ are equally compatible with the observations. Thereby, we abandon the ill-posed notion of targeting a single optimal solution and instead adopt a Bayesian approach where the target becomes a probability distribution of possible solutions.

In the Bayesian framework, probability $p(\cdot)$ becomes a general measure of epistemic uncertainty (i.e., belief) rather than a synonym for frequency (MacKay, 2003). Therein, probability can be seen as encoding the beliefs of an agent (Hennig et al., 2022), which we can take to be the forward model. Although this step appears to introduce a superfluous subjective agent, it is actually a probabilistic extension of the model that Makes assumptions transparent, represents uncertainty in the model and observations, and allows us to solve generally ill-posed inverse problems such as parameter estimation. Adopting this framework, the inverse problem Equation 1 for $\boldsymbol{\theta}$ can be solved by updating the prior probability $p(\boldsymbol{\theta})$ (the agent's beliefs before considering the data) using the likelihood $p(\mathbf{y}|\boldsymbol{\theta})$ (what the data tells the agent) to obtain the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$ (the agent's beliefs after considering the data) through Bayes' rule

$$p(\boldsymbol{\theta}|\mathbf{y}) = p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})/Z, \quad (2)$$

where $|$ denotes “given” as used in the context of conditional probability. The model evidence $Z = p(\mathbf{y})$ (MacKay, 2003) is an unknown normalizing constant that we can thankfully, through the design of the assimilation methods, ignore in the context of parameter estimation pursued in this study.

2.2.2. The Prior and the Likelihood

The prior is used to specify the initial beliefs of the agent concerning plausible values for the parameters $\boldsymbol{\theta}$ that can be distilled from expert opinion, physical bounds, so-called objective defaults, and earlier analyses (Banner et al., 2020). Note that although superficially setting a flat prior may be tempting from the slippery notion of objectivity, it has the downsides of requiring the specification of bounds and not being scale invariant. Flat priors are inefficient when used without justification since they can lead to a diffuse posterior by ignoring relevant prior knowledge concerning parameters that are usually available to help specify their typical location and scale (Banner et al., 2020). Usually, it is more instructive to employ a weakly informative prior that encodes what we roughly know about the parameters (bounds, location, and scale), which has the additional benefit of helping to regularize the inverse problem (MacKay, 2003). This is the route we will follow in Section 2.2.4. The important point to note here is that in the context of parameter estimation, we take the prior as given (or rather, as a fixed part of our agent's probabilistic model) in a known functional form that can be expressed analytically. It is possible to relax this assumption and add more levels in the Bayesian hierarchy to infer hyperparameters that control the parameter priors (Gelman et al., 2013; MacKay, 2003; Winkle & Berliner, 2007). However, this incurs a prohibitive increase in computational cost. Thereby, we will not pursue such hierarchical models here but note that promising methods are emerging that can make them tractable (Alsing et al., 2019; Dokoohaki et al., 2022; Fer, Gardella, et al., 2021; Katzfuss et al., 2020).

The likelihood $p(\mathbf{y}|\boldsymbol{\theta})$ is the main computational challenge for parameter estimation in complex non-linear forward models such as CLM-FATES. Specifically, the likelihood is the probability of obtaining the (fixed) observed data $\mathbf{y}$ that we are assimilating given a specific set of parameters $\boldsymbol{\theta}$. Following the likelihood principle in Bayesian inference, where the observations are fixed while $\boldsymbol{\theta}$ is the random variable, the likelihood should be viewed as a function of the uncertain parameters and not the noisy but fixed observations (MacKay, 2003).

Loosely speaking, the likelihood is a model fit term that quantifies how close the forward model is to the observations for a particular set of parameters, with a higher likelihood indicating a closer fit. Unlike simpler models where the likelihood may be written down in closed analytical form, the fact that we are using a complex forward model makes the likelihood term a computational bottleneck. That is, we do not know what the likelihood of a particular parameter set will be before actually evaluating (i.e., running) the forward model for that parameter set and comparing the predicted observations that it generates to the observations. Herein, as is typical in DA (Carrassi et al., 2018), we adopt a Gaussian observation error model (Section 2.2.6) where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$ in which $\mathbf{R}$ is the observation error covariance matrix. For simplicity, in line with the common assumption that observation errors are independent in time, we use a diagonal observation error covariance matrix $\mathbf{R}$ (Evensen et al., 2022). Thereby, the likelihood takes the following Gaussian form

$$p(\mathbf{y}|\boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2}[\mathbf{y} - \hat{\mathbf{y}}]^T \mathbf{R}^{-1} [\mathbf{y} - \hat{\mathbf{y}}]\right), \quad (3)$$

where $\hat{\mathbf{y}} = \mathcal{M}(\boldsymbol{\theta})$ denotes predicted observations that were generated from the forward model for a particular parameter set $\boldsymbol{\theta}$. Due to the strong constraint assumption and the deterministic nature of CLM-FATES, these predicted observations will be noise-free with each parameter set mapping onto a single set of predicted observations. Note that although the form of this likelihood is Gaussian, it is nonetheless highly nonlinear in that it involves the predicted observations $\hat{\mathbf{y}}$ that can only be evaluated exactly by running the computationally expensive forward model.

2.2.3. Data Assimilation

In theory, obtaining the posterior density through Bayesian inference is deceptively simple and only a matter of (up to a normalizing constant) multiplying the prior and likelihood for some $\boldsymbol{\theta}$ in (Equation 2). Naively then, we could approximately infer the posterior through an exhaustive enumeration in a grid search by discretizing the parameter space into a fine grid. However, this approach becomes unfeasible with increasing numbers of adjustable parameters due to the exponential increase of possible combinations, which is one of several challenges collectively known as the “curse of dimensionality” (Chopin & Papaspiliopoulos, 2020; Murphy, 2023). In this study, we are trying to infer six parameters (Section 2.2.4), so even with a coarse discretization of 10 values per parameter, an exhaustive enumeration would require one million evaluations of the likelihood. Accordingly, the challenge of approximating the posterior is somewhat akin to looking for a needle in a haystack: more parameters create a larger haystack and more informative observations create a smaller needle.

To tackle this computational challenge, a plethora of sophisticated and closely related inference algorithms have emerged across several fields (Carrassi et al., 2018; Evensen et al., 2022; Gelman et al., 2013; MacKay, 2003; McElreath, 2020; Murphy, 2023). In the geosciences, the application of such inference algorithms to parameter and/or state estimation in (typically mechanistic) models is often referred to as DA (Evensen et al., 2022). Roughly speaking, for nonlinear forward models, these algorithms can be split into two main schools: Variational methods and Monte Carlo (sampling-based) methods.

As variational methods are gradient-based, which complicates their application to complex mechanistic forward models like CLM-FATES that are implemented as a long piece of (possibly non-differentiable) source code, here we focus our attention on more broadly applicable Monte Carlo methods. Monte Carlo methods, which we take to include ensemble Kalman methods (Carrassi et al., 2018; Evensen et al., 2022), MCMC methods (Gelman et al., 2013; MacKay, 2003), and particle (or sequential Monte Carlo) methods (Chopin & Papaspiliopoulos, 2020), share in common that they all rely in part on sampling distributions using (pseudo) random number generators. While MCMC methods asymptotically provide a gold standard for Bayesian inference (Gelman et al., 2013), they often require prohibitively many sequential steps with relatively limited potential for computational parallelization (Cleary et al., 2021). In practical geoscientific applications, we can often only afford to run the forward model for a few hundred ensemble members (i.e., samples), which means the only valid alternatives are particle and ensemble Kalman methods (Evensen et al., 2022).

In DA problems with moderate (i.e., more than a couple of parameters) to high dimensional problems, ensemble Kalman methods have been shown to be a more robust choice than standard particle methods which tend to suffer from degeneracy problems (Alonso-González et al., 2022; Carrassi et al., 2018; Pirk et al., 2022). Parameter

estimation for land surface modeling falls under this moderate to high dimensional case, which motivates our adoption of ensemble Kalman methods, specifically the IEnKS. The hope is that by combining the IEnKS in an initial calibration step with subsequent machine-learning-based emulation and MCMC sampling, we will also be able to get closer to gold standard Bayesian inference with full MCMC, with a computationally feasible amount of full forward model evaluations (Cleary et al., 2021).

2.2.4. Prior Specification and Transformations

To generate an initial ensemble of prior states from CLM-FATES (Figure 1a), we have to assign a prior distribution that encodes our beliefs concerning the parameters to be calibrated. The location of the priors should represent a central best-guess estimate of the true model parameter configuration with a spread that captures the initial parameter uncertainty (i.e., assumed error statistics). Such weakly informative priors are regularizing and robust to moderate misspecifications in the sense that a bias in their central location does not negatively impact the results considerably, as long as the parameter bounds encompass all plausible parameter values (Banner et al., 2020). Here, our initial best guesses (i.e., the location parameters) are the default FATES parameter values and we assume independent generalized logit-normal distributions for each parameter. The generalized logit-normal distribution is defined in the supplement.

The logit-normal hyperparameters for each of the six CLM-FATES parameters are outlined in Table S1 in Supporting Information S1. We set the bounds for V_{cmax} to be $\pm 75\%$ of the default values and bounded all BB_{slope} parameters to be between 1 and 20. We picked a scale factor $\sigma = 0.3$ for all the parameters, corresponding to a weakly informative prior in that a scale of $\sigma = 1.74$ corresponds to the maximum entropy choice for univariate centered logit-normal distributions. For simplicity, these priors are assumed to be independent of one another, such that the joint prior is given by the product of the marginal prior distributions. Note that it is possible to generate correlated priors, which can be useful for more effective exploration of the parameter space (Pirk et al., 2022), but for simplicity, we decided not to consider this route here. In practice, approximate Bayesian inference (2) is now carried out in the transformed unbounded (i.e., ϕ) parameter space while the model predictions are generated using the corresponding bounded (i.e., θ) parameter space (Aalstad et al., 2018; Dunbar et al., 2021; Guidicelli et al., 2023; Pirk et al., 2022).

2.2.5. Integration Period and Data Alignment

The CLM-FATES framework we use in this study limits the parameter update steps as each IEnKS iteration requires a separate model experiment that depends on the previous iteration (Figure 1b). Each CLM-FATES experiment is submitted as a new job to a high-performance computing cluster, and the subsequent DA procedure requires costly file read and write operations. Conceptually, the integration period that corresponds to the assimilation window for parameter estimation should capture characteristic timescales of the system being modeled (Alonso-González et al., 2022; Dunbar et al., 2021; Pirk et al., 2022). We selected 10 years within the atmospheric forcing coverage where good-quality FI-Hyy eddy-covariance (EC) measurements were available as our DA window (2004–2013) (Mammarella et al., 2019). The goal was to capture interannual flux variability while keeping the computation time and output file sizes manageable. Accordingly, each IEnKS iteration involves running a new model ensemble with updated parameters for the entire DA window.

Moreover, the aggregation period for which the model-data error is minimized can substantially alter the inference of the posterior parameter distribution (Mäkelä et al., 2016; Pirk et al., 2022). We chose to assimilate daily aggregate values: this time scale approaches the half-hourly period at which the relevant processes are modeled and measured while mitigating the impact of potential observation noise (e.g., if the fluxes mildly violated the assumption of statistical stationarity within the 30-min aggregation time window). Our workflow converts the partitioned and gap-filled flux data to the model units, resamples them to mean daily values during the DA window, and adjusts the dates to match the atmospheric forcing data set's time zone and calendar. The corresponding daily model output is compiled into tabular files for each ensemble member, following the same formatting (Keetz & Aalstad, 2024).

2.2.6. Observation Error Model

Bayesian DA involves modeling the observation error process to account for measurement uncertainties. The IEnKS formally requires the observation error model to be Gaussian. For simplicity, it is further often taken to be

independent and homoscedastic (e.g., Alonso-González et al., 2022; Pulido et al., 2018; N. Raoult, Beylat, et al., 2024), which is the route we also pursue in this study while noting potential improvements in the outlook. In reality, eddy-covariance measurement errors due to sensor imperfections tend to increase with the flux magnitude and are not independent. Flux tower measurement random errors typically resemble a symmetric exponential (Laplace) distribution (Richardson et al., 2006). The standard deviation usually lies within 20% of the true flux (Mäkelä et al., 2016; Rannik et al., 2016). However, the gap-filling and storage correction methods applied to the flux data, as well as the carbon flux partitioning models used to derive GPP (Kulmala et al., 2019), can be expected to add additional error sources to the assimilated flux values. Our error model should also consider the representativeness error, which describes the mismatch between the spatio-temporal scale of the observations and the model (Pirk et al., 2022). Thus, following the Gaussian requirements and the central limit theorem (e.g., Chopin & Papaspiliopoulos, 2020), we model the observation errors as follows.

We assume a Gaussian error $\epsilon \sim N(0, \sigma_F)$ with a standard deviation σ_F of 15% of averaged near maximum annual flux measurements. The σ_F assumption was based on expert experience and loosely approximates the 20% attributed to sensor imperfection cited in Mäkelä et al. (2016). The main goal was not to underestimate the observation errors in the summer when the estimated parameters have the highest impact on the modeled fluxes. Conversely, this conservative model comes at the expense of exaggerated errors during winter when measured and true fluxes are approaching zero. For each assimilated variable's valid half-hourly measurement time series, we.

1. Calculated 99% percentile values for each year within the 10-year DA window to eliminate potential outliers,
2. Calculated the arithmetic mean,
3. Multiplied the mean by 0.15.

The error model for the half-hourly GPP observations is hence approximated as $\epsilon_{GPP} \sim N(0, \sigma_F = 3.27)$ in $g\ C\ m^{-2}\ d^{-1}$ and for ET as $\epsilon_{ET} \sim N(0, \sigma_F = 1.13)$ in $kg\ H_2O\ m^{-2}\ d^{-1}$.

During the DA procedure, we aggregate the measurements and model outputs to daily values (cf. Section 2.2.5). Therefore, the estimated daily error standard deviations are determined through the central limit theorem (Pirk et al., 2022):

$$\sigma_{\bar{F}} = \frac{\sigma_F}{\sqrt{N_F}} \quad (4)$$

where $\sigma_{\bar{F}}$ is the observation variable's aggregated (here: daily) flux error standard deviation, σ_F is the corresponding error model's standard deviation for the original (here: Half-hourly) fluxes, and N_F is the averaged number of valid measurements within the aggregation period (here: valid half-hourly measurements in a day, averaged over the DA period).

2.2.7. “Calibrate”—Iterative Ensemble Kalman Smoother

We adopt a batch iterative ensemble Kalman smoother (IEnKS) (Evensen et al., 2022) for parameter estimation in the initial “calibrate” step of CES (Figure 1b). More specifically, we employ an IEnKS known as the ensemble smoother with multiple DA (ES-MDA) (Emerick & Reynolds, 2013). This method is particularly appealing in that it is straightforward to implement and has been shown in practice to yield reasonable approximations of the posterior with only a handful of iterations (Aalstad et al., 2018; Guidicelli et al., 2023; Pirk et al., 2022). The multiple DA iterations impose likelihood tempering by inflating the observation error covariance while maintaining Bayesian consistency (Chopin & Papaspiliopoulos, 2020; Murphy, 2023; Stordal & Elsheim, 2015) so as to make ensemble Kalman methods more robust to the non-linearity of the underlying model (Aalstad et al., 2018; Alonso-González et al., 2022; Emerick & Reynolds, 2013; Groenke et al., 2023; Pirk et al., 2022). In contrast to ensemble Kalman filters (Evensen, 2003) that have also been applied to LSM DA (Shu et al., 2022; Strebel et al., 2022), batch smoothers like the IEnKS do not update the model parameters sequentially after each forward model time step, but instead perform a global update by assimilating all data in an observation window. They are therefore more suitable for static parameter estimation (Emerick & Reynolds, 2013). Hereafter we briefly outline this IEnKS algorithm and refer to the seminal work of Emerick and Reynolds (2013) and the recent book of Evensen et al. (2022) for more details on the implementation and theory.

The IEnKS is initialized by sampling $i = 1, \dots, N_e$ (where N_e is the number of ensemble members) prior parameter vectors $\theta_i^{(0)}$ from the $N_p = 6$ dimensional joint prior parameter distribution described in Table S1 in Supporting Information S1. These prior parameter vectors are then used to generate an ensemble of prior predicted (i.e., modeled) observations from the forward model $\mathcal{M}(\cdot)$ (CLM-FATES) through $\hat{\mathbf{y}}_i^{(0)} = \mathcal{M}(\theta_i^{(0)})$. The predicted observations $\hat{\mathbf{y}}$ contain modeled daily GPP and ET fluxes that correspond to the 10 years of valid daily aggregated observed EC fluxes (Figure 1a). All these valid observations are contained in the observation vector $\mathbf{y}$ with N_o entries. The multiple DA iterations of the IEnKS then proceed as follows for $\ell = 0, \dots, N_a - 1$ iterations while implicitly looping over ensemble members i :

1. Transform the parameters to the unbounded Gaussian space

$$\phi_i^{(\ell)} = \psi(\theta_i^{(\ell)}) \quad (5)$$

where ψ is the generalized logit transform defined in the supplement.

2. Update these transformed parameters through the Kalman analysis step

$$\phi_i^{(\ell+1)} = \phi_i^{(\ell)} + \mathbf{K}^{(\ell)}(\mathbf{y}_i^{(\ell)} - \hat{\mathbf{y}}_i^{(\ell)}) \quad (6)$$

where $\mathbf{K}^{(\ell)}$ is an $N_p \times N_o$ ensemble Kalman gain wherein $\mathbf{R}$ has been inflated by $\alpha^{(\ell)} = N_a$ thereby effectively tempering the likelihood while ensuring consistency through the condition $\sum_{\ell=1}^{N_a} 1/\alpha^{(\ell)} = 1$, and $\mathbf{y}_i^{(\ell)} = \mathbf{y} + \epsilon_i^{(\ell)}$ are the correspondingly inflated perturbed observations using the observation error models presented in Section 2.2.6 (Aalstad et al., 2018; Emerick & Reynolds, 2013).

3. Transform these updated parameters back to the bounded physical space

$$\theta_i^{(\ell+1)} = \psi^{-1}(\phi_i^{(\ell+1)}) \quad (7)$$

where ψ^{-1} is the inverse generalized logit transform defined in the supplement.

4. Run the CLM-FATES forward model to obtain an updated ensemble of predicted observations (and model states) using the updated parameters $\hat{\mathbf{y}}_i^{(\ell+1)} = \mathcal{M}(\theta_i^{(\ell+1)})$.

The approximate posterior from the IEnKS is then given by the ensemble of updated parameters in the final iteration $\theta_i^{(N_a)}$ and the corresponding model states $\mathbf{x}_i^{(N_a)}$ and predicted observations $\hat{\mathbf{y}}_i^{(N_a)}$. In particular, the posterior predicted observations are only a small subset of the full posterior state output by CLM-FATES with updated parameters. Note that all the steps above are implicitly run for all ensemble members ($i = 1, \dots, N_e$).

The number of multiple DA iterations (N_a) and ensemble members (N_e) are hyperparameters that can have significant impacts on the IEnKS performance: too few iterations or too small an ensemble can lead to subpar results while too many iterations or too large an ensemble unnecessarily increase the number of model runs (Aalstad et al., 2018; Lorentzen & Naevdal, 2011). The existing ensemble Kalman literature (e.g., Aalstad et al., 2018; Carrassi et al., 2018; Pirk et al., 2022) suggests that an ensemble size of around $N_e = 100$ is enough for an efficient yet relatively accurate approximation of the posterior in moderate dimensional problems. Therefore, we adopt an ensemble size of $N_e = 128$ (a power of 2) due to technical requirements. Moreover, as a sensitivity analysis of N_a , we compare the model ensemble error metrics and parameter posterior distributions for $N_a = 1, 2$ and 4 iterations (K_1 , K_2 , and K_4 ; Table 1). With each iteration, a new model ensemble with $N_e = 128$ members is generated from the updated parameter matrix, such that the entire IEnKS loop (including the prior) generates a total of $(N_a + 1) \times N_e$ parameter-vector-to-predicted-observation maps that can be used to train an emulator.

2.2.8. “Emulate”—Negative Log-Likelihood Emulator

The MCMC algorithm described in Section 2.2.9 sequentially steps through parameter space and accepts a proposed new step in parameter space with a higher probability if it increases the posterior density compared to

Table 1

Overview of the Experimental Setups and Their Identifiers

Synthetic experiments (Section 3.1)

IEnKS calibration

Test different amounts of ensemble Kalman iterations

K_0	No ensemble Kalman iterations, that is, model ensemble output from the chosen logit-normal priors
K_1	1 ensemble Kalman iteration, that is, the non-iterative EnKS
K_2	2 ensemble Kalman iterations
K_4	4 ensemble Kalman iterations

Emulator configuration

Test adaptive MCMC with different negative log-likelihood emulators, trained on the outputs generated from K_4

CES-GPR	Gaussian Process Regressor with a scikit-learn default Matérn 3/2 Kernel
CES-GPR _{1σ}	CES-GPR prediction inflated by a 1σ (one standard deviation) posterior uncertainty estimate
CES-GPR _{5σ}	CES-GPR prediction inflated by a 5σ (five standard deviations) posterior uncertainty estimate
CES-XGBoost	Tree-based Extreme Gradient Boosting regression with grid search hyperparameter tuning

Emulator-based MCMC data sampling strategy

Test emulator-based MCMC with different strategies for generating the emulator training data

eMCMC-UNI _{ST}	Random sampling from a uniform distribution
eMCMC-LHS _{ST}	Latin hypercube sampling
eMCMC-LON _{ST}	Random sampling from the logit-normal prior
eMCMC-CES _{ST}	Full CES workflow, that is, IEnKS (K_4) output based on a logit-normal prior

Eddy-covariance observations (Section 3.2)

eMCMC-UNI _{EC}	Random sampling from a UNIFORM distribution (UNI)
eMCMC-LHS _{EC}	Latin Hypercube Sampling (LHS)
eMCMC-LON _{EC}	Random sampling from the LOGit-Normal prior (LON)
eMCMC-CES _{EC}	Full CES workflow, that is, IEnKS (K_4) output based on a logit-normal prior

the previous step. As shown in Equation 11, proposed parameter vectors are accepted with a higher probability if they reduce the negative log-likelihood (NLL) obtained from running the forward model with the proposed parameters. In practical terms, we can view the NLL as the sum of the square residuals between observed and predicted (modeled) values of the assimilated variables, scaled by the observation error variance:

$$\mathcal{L}(\phi) = \sum_{k=1}^{N_o} \frac{(y_k - \hat{y}_k(\phi))^2}{2\sigma_{y_k}^2} \quad (8)$$

where $\mathcal{L}(\phi)$ is the NLL based on a model realization with a parameter set in transformed space ϕ , y_k is a specific (here: daily averaged) flux observation, $\hat{y}_k(\phi)$ is the corresponding model prediction that depends on the transformed parameters ϕ , and σ_{y_k} is the specific flux observation's measurement error standard deviation (cf. Section 2.2.6). Note again that y_k , $\hat{y}_k$, and σ_{y_k} contain values for multiple observed variables (here: GPP and ET).

As making a sufficient amount (typically tens of thousands) of sequential forward model predictions $\hat{y}(\phi)$ for MCMC sampling to converge is practically not feasible with CLM-FATES, we instead use a supervised regression model to emulate the NLL surface, trained on a feasible number of initial ensemble runs (Figure 1c). In

short, we aspire to train an NLL emulator, $\hat{\mathcal{L}}(\phi)$, such that $\hat{\mathcal{L}}(\phi) \approx \mathcal{L}(\phi)$ where the true NLL is given by Equation 8 with samples available from an initial ensemble. There are many possible choices for surrogate models that can all be viewed as some form of statistical or machine learning-based regression (curve-fitting) methods (Gramacy, 2020; M. C. Kennedy & O'Hagan, 2001; Murphy, 2023; C. Wang et al., 2014). As the relationships between ϕ and the CLM-FATES model outputs are surely (both the parameter transformations $\theta = \psi^{-1}(\phi)$ and the forward model $\hat{y} = \mathcal{M}(\theta)$ itself) strongly non-linear (Pirk et al., 2022; Strebel et al., 2022), we tested two non-linear and non-parametric machine learning approaches: Gaussian Process Regression (GPR) (Rasmussen & Williams, 2005) and eXtreme Gradient Boosting (XGBoost) (Chen & Guestrin, 2016). Both have previously been applied in the context of LSM emulation (e.g., Fer et al., 2018; Li et al., 2023; Sawada, 2020).

GPR is a probabilistic approach in the family of Bayesian methods and is often well-suited for uncertainty-aware function approximation in noisy and non-linear problems (Gramacy, 2020; Hennig et al., 2022; MacKay, 2003; O'Hagan, 2006; Rasmussen & Williams, 2005; Santner et al., 2018; Särkkä et al., 2013). It builds on the concept of Gaussian processes (GP). Formally, a GP is a collection of random variables, such that any finite number of these (from arbitrary points in the domain) will share a joint Gaussian distribution (Rasmussen & Williams, 2005). In the context of GPR, these random variables represent the (typically univariate) output variable as random (i.e., uncertain) functions evaluated at different points in the continuous input variable domain. Accordingly, the GPR predictions are derived from a probability distribution over many possible functions where each point-wise prediction can be treated as a random variable with a marginal Gaussian distribution characterized by a mean and variance. Importantly, the predictions are more certain (have lower variance) the closer the input variables are to those in the training data. GPR relies on a kernel (covariance) function that determines the shape and smoothness of the underlying GPs, which affects the GP model's flexibility and performance. The kernel prior is controlled by hyperparameters such as the length scale ℓ and smoothness ν . An exact posterior GP can then be obtained analytically while the hyperparameters are optimized following the common practice of using an empirical Bayes method of maximum marginal likelihood estimation (Murphy, 2023; Rasmussen & Williams, 2005). Here, we use the *GaussianProcessRegressor()* implementation in the *scikit-learn* package and a Matérn 3/2 Kernel (i.e., $\nu = 1.5$) initialized with default settings (Pedregosa et al., 2011). Refer to Table S3 in Supporting Information S1 for the initial and optimized hyperparameters.

Initial testing suggested that this default GPR setup can give inaccurate NLL predictions for parameter regions that are not sufficiently represented by samples in the training data. While limited design (training) points are a general challenge for all surrogate model building (Gramacy, 2020; Santner et al., 2018; Wang et al., 2014), this becomes especially relevant for implementations based on non-uniform and thus less-space filling priors such as the joint logit-normal prior we use within CES. It is to be expected that large edge regions do not contain many training samples and that deterministic emulator extrapolations there can thus not be blindly trusted. Therefore, we test how making GPR predictions more conservative affects the posterior approximations within MCMC. In particular, we make use of the built-in uncertainty awareness of GPs and inflate each posterior mean NLL emulator prediction by adding a scaled version of its parameter-dependent posterior standard deviation:

$$\hat{\mathcal{L}}_g(\phi) = \mu_g(\phi) + \beta \sigma_g(\phi) \quad (9)$$

where $\mu_g(\phi)$ and $\sigma_g(\phi)$ are the posterior mean and standard deviation, respectively, of the NLL predictions from the GPR evaluated at ϕ . The scaling term $\beta \geq 0$ is a penalization factor that makes the emulated NLL more conservative or pessimistic (higher NLL implying a lower likelihood) at input locations ϕ where the emulator is more uncertain. Consequently, the larger β , the more unlikely the Markov chain is to accept parameter samples in parameter regions of high GPR uncertainty. On the other hand, it is expected that a too large β can fully prevent the Markov chain from moving to predictions of even low or moderate uncertainty, which would eventually defeat the purpose of emulation. Thus, we test the approach for three different β to investigate the impact on the respective MCMC posterior samples (0, 1, and 5; Table 1).

The second method, XGBoost, is a popular machine learning model for both regression and classification tasks with high predictive accuracy (Chen & Guestrin, 2016). We primarily include it as a machine learning model benchmark for the GPR. The XGBoost implementation employed here uses regression trees as base learners, which are added iteratively to minimize a specified loss function (here: the mean squared error). Newly added trees correct the errors of the existing ensemble until adding more does not improve the predictions any longer, or

until the algorithm reaches a predefined maximum number of iterations. Each tree aims to partition the data into homogeneous target value groups by splitting it at optimal parameter thresholds. The final prediction aggregates the individual predictions from all trees. This ensemble approach, combined with in-built regularization techniques, enables XGBoost to make accurate predictions while mitigating overfitting. However, optimal performance can require meticulous hyperparameter tuning (Bentéjac et al., 2021). Here we used grid-search with 5-fold cross-validation to tune important hyperparameters and selected the configuration that minimized the average mean squared error. Refer to Table S2 in Supporting Information S1 for the hyperparameters, their tested ranges, and the final values.

To test how the different emulators impact the posterior approximation, we evaluate the four models (GPR, GPR_{1σ}, GPR_{5σ}, and XGBoost) with the NLL samples obtained from the K₄ IEnKS experiment (640 total forward model runs). A random selection of 80% (512 forward model runs) of these samples is used for emulator training and 20% (128 forward model runs) of the samples are retained for evaluating the predictive performance. We then use each trained emulator in the adaptive MCMC algorithm described in the next section and pick the configuration whose posterior distribution best represented the synthetic truth parameters (Section 3.1) for the following experiments. In those experiments, the emulators were trained on 100% of the forward model data.

Finally, we calculate SHapley Additive exPlanations (SHAP) values for the chosen emulator to investigate which parameters mainly governed its predictions. SHAP values were developed as a unified mathematical framework for complex machine learning model interpretation based on game-theoretic principles (Li et al., 2023; Lundberg & Lee, 2017). They quantify the contribution of individual input features to the output prediction of any type of machine learning model while also accounting for parameter interactions and dependencies. In essence, each parameter gets assigned an importance value for each individual prediction. These results can then be aggregated to provide insights into the overall importance of parameters within the model. Here, we use an implementation of the Kernel SHAP method (*KernelExplainer()*) that can approximate SHAP values for any regression model, with an especially straightforward implementation for *scikit-learn* models (<https://github.com/shap/shap/tree/v0.42.1>; last accessed 2023-10-20).

2.2.9. “Sample”—Emulator-Based Markov Chain Monte Carlo Sampling

Markov chain Monte Carlo schemes (Gelman et al., 2013; MacKay, 2003; McElreath, 2020; Neal, 1993; Van Ravenzwaaij et al., 2018) involve constructing Markov chains that asymptotically sample from a stationary target distribution. Herein, this target distribution is the (emulated) posterior that we wish to sample from in the final step of CES (Figure 1d). In general, MCMC algorithms sequentially step through a forward model's parameter space and either accept or reject the proposed step, thereby eventually obtaining samples from the Bayesian posterior distribution in (Equation 2). The Metropolis-Hastings algorithm (Hastings, 1970; Metropolis et al., 1953), on which we base our implementation, is arguably the most archetypal MCMC method. In this approach, a new step ϕ_{j+1}^* (new parameter vector value) in the Markov chain is proposed by randomly drawing from a proposal distribution $q(\phi_{j+1}^*|\phi_j)$ that is conditioned (typically by centering) on the current step ϕ_j . This proposed new step is accepted based on the acceptance ratio

$$\alpha_{j+1} = \frac{p(y|\phi_{j+1}^*)p(\phi_{j+1}^*)}{p(y|\phi_j)p(\phi_j)} \quad (10)$$

if the inequality $u_{j+1} \leq \min(\alpha_{j+1}, 1)$ is satisfied where $u_{j+1} \sim U(0, 1)$ is a (pseudo) randomly generated number between 0 and 1, otherwise it is rejected. In the case of acceptance, the Markov chain moves to the proposed point such that $\phi_{j+1} = \phi_{j+1}^*$. In the case of rejection, the Markov chain stays at the current point $\phi_{j+1} = \phi_j$. A proposed point will always be accepted if it has a higher posterior density than the current point since then $\alpha_{j+1} > 1$ and $u_{j+1} \leq 1$ always holds. Furthermore, the proposed point ϕ_{j+1}^* always has a chance to be accepted even if it has a lower posterior density. In this case, the probability of acceptance is the acceptance ratio. This probabilistic acceptance rule ensures that asymptotically, that is as $j \rightarrow \infty$, this Markov chain will sample from the posterior distribution.

By design, the difficult-to-estimate normalizing constant in (Equation 2) cancels out in the acceptance ratio (Equation 10). In general, the acceptance ratio also includes the ratio of the proposal distributions at the current

and new point, but in the typical case considered herein with a symmetric proposal, these too cancel out. Unlike other Monte Carlo methods such as importance sampling, the full (i.e., not thinned) posterior samples from an MCMC algorithm will be auto-correlated rather than independent. The sample means and covariances from the samples in the chain will nonetheless converge to the corresponding posterior expectations provided that the chain is run for long enough, typically on the order of 10^5 iterations. In practice, the initial samples in the chain before it has converged (mixed) are discarded as a burn-in period to avoid biasing estimates of posterior statistics. Instead of using the costly full likelihood in (Equation 10), we adopt the emulated likelihood which allows us to run very long chains in a reasonable amount of time. This is achieved by using the following emulated acceptance ratio:

$$\hat{\alpha}_{j+1} = \frac{\exp(-\hat{\mathcal{L}}(\boldsymbol{\phi}_{j+1}^*))p(\boldsymbol{\phi}_{j+1}^*)}{\exp(-\hat{\mathcal{L}}(\boldsymbol{\phi}_j))p(\boldsymbol{\phi}_j)}, \quad (11)$$

where $\hat{\mathcal{L}}(\boldsymbol{\phi})$ denotes the emulated NLL evaluated for a particular parameter set.

The success of the classic Random Walk Metropolis (RWM) method (Metropolis et al., 1953) with a Gaussian proposal distribution $q(\boldsymbol{\phi}_{j+1}^*|\boldsymbol{\phi}_j) = \mathcal{N}(\boldsymbol{\phi}_{j+1}^*|\boldsymbol{\phi}_j, \Sigma_q)$ hinges on choosing an adequate value for the proposal covariance matrix Σ_q . The burden of having to manually tune the proposal has spurred the development of adaptive MCMC algorithms where the proposal is tuned automatically in an online fashion during the evolution of the Markov chain (Haario et al., 2001). Herein, we use the Robust Adaptive Metropolis (RAM) algorithm proposed by Vihola (2012) which is an extension of the adaptive method of Haario et al. (2001) that is both efficient and simple to implement. This RAM method automatically adapts the proposal covariance, largely during the so-called burn-in period, to be non-isotropic in such a way that the chain nearly achieves an asymptotically optimal acceptance rate of 0.234 (Roberts et al., 1997). We refer the reader to Vihola (2012) for more details on the implementation of RAM, but note in passing that we adopted the default hyperparameters suggested therein and ran single chains with a length of 200,000 and a burn-in of 10%.

2.2.10. Comparing Alternative Emulator Sampling Designs

Emulating the NLL surface based on IEnKS results, the “calibrate” step in the CES framework, involves sequential (i.e., interdependent) ensemble simulations for the iterative parameter updates. Different design strategies exist to generate training data for a forward model emulator (Wang et al., 2014). Simpler data-blind sampling designs do not depend on an initial calibration through sequential iterations and can thus require less computing time. Perhaps the most straightforward approach is to (pseudo) randomly draw parameter samples from uninformative or vague prior distributions, where the most ubiquitous choices are uniform distributions (Banner et al., 2020). A similar, slightly more sophisticated, systematic, and space-filling approach is LHS (McKay et al., 1979), which has also been applied in the context of LSM emulation (Dagon et al., 2020; Li et al., 2023). Simply put, LHS partitions the multidimensional parameter space into N_s regular intervals, with N_s being the desired number of samples. Subsequently, one sample is drawn uniformly at random within each interval (parameter space region). LHS aims to enhance the parameter space representativeness compared to uniform sampling, which becomes especially relevant when the model output is mainly governed by only a small number of parameters in a data set (McKay et al., 1979). Conceptually, these simpler space-filling designs generate emulator training points without considering the data, which from a Bayesian point of view can be regarded as inefficient: going back to the haystack analogy, we are only interested in emulating the needle (the typical posterior region of interest), not the full haystack (the full parameter space).

We thus benchmark the CES performance against data-blind techniques. We run the RAM MCMC algorithm described in Section 2.2.9 with NLL emulators trained on model outputs from uniform sampling (eMCMC-UNI), LHS (eMCMC-LHS), the full CES workflow (eMCMC-CES), and the logit-normal priors we use to initialize CES (eMCMC-LON) (Table 1). We also compare these eMCMC results to the IEnKS results (i.e., “calibrate” without “emulate” and “sample”). Note that due to practical concerns, we generated 256 ensemble members for eMCMC-UNI, eMCMC-LHS, and eMCMC-LON, whereas eMCMC-CES utilizes 128 initial ensemble members and four Kalman iterations, resulting in 640 total model runs. We used 256 randomly drawn samples of the K_4 outputs as the eMCMC-CES emulator design points to increase fairness regarding emulator training data set sizes. The final eMCMC performance evaluations are based on $n = 128$ randomly drawn parameter sets from the respective Markov chains.

2.2.11. Computational Runtime

One major selling point of the CES approach is computational feasibility (Cleary et al., 2021), which we herein briefly put into perspective with runtime estimates of our implementation. The IEnKS we use in the “calibrate” step relies on sequential (non-parallelizable) iterations involving costly forward-model evaluations and input/output operations. It therefore has increased computational demands compared to non-iterative approximate Bayesian inference approaches—with the benefit that it can provide improved posterior estimates (cf. Section 4.1.1). Nevertheless, the computational demands of a handful of IEnKS iterations and subsequent “emulate” and “sample” steps as used in this study remain manageable, especially when compared to gold standard MCMC that relies on the full (i.e., non-emulated) forward model.

We simplistically estimate a fixed wall-clock time of 1.5 hr for creating and running one (parallelized) single-site CLM-FATES ensemble experiment and a subsequent IEnKS iteration with the experimental setup, software, and hardware used in this study. Let us further assume 1 hr for a corresponding single-instance CLM-FATES experiment and subsequent MCMC iteration with the full forward model. Running the IEnKS with $N_a = 4$ Kalman iterations and $N_e = 128$ ensemble members then results in 640 forward model evaluations within 5 non-parallelizable iterations (including the prior run), translating to a total wall-clock time of 7.5 hr. Reasonably assuming one would require 100,000 non-parallelizable full forward model evaluations for somewhat robust MCMC results, this would translate to a total wall-clock time of 100,000 hr (ca. 11.4 years). Training an emulator on the IEnKS results only takes seconds. Finally, emulator-based MCMC with Markov chains of total length 200,000 resulted in a total runtime of roughly 1 hr for the “sample” step.

3. Data

3.1. Synthetic Experiments

We first tested our CES implementation in a so-called identical twin experiment (Arnold & Dey, 1986) by assimilating noisy synthetic observations. These were obtained by perturbing model outputs from a “synthetic truth” (ST) to explore the performance of the respective assimilation algorithms under controlled perfect model assumptions, which can help to investigate the methods' general applicability to our problem. For this purpose, we postulated “true” values for the six calibrated model parameters, reducing default V_{cmax} values by 15% and BB_{slope} values by 20% (Table S1 in Supporting Information S1). Subsequently, we used these true parameters in a single-instance ST simulation with the same model configuration described in Section 2.1.2. We then aggregated GPP and ET outputs for the 10-year study period with the same procedure we apply to the eddy-covariance measurements (cf. Section 3.2). We then added random perturbations to the results to mimic observation errors following the observation error models (cf. Section 2.2.6). Figure S9 in Supporting Information S1 shows the corresponding actual model output and the perturbed data that was assimilated.

3.2. Hyytiälä Field Observations

3.2.1. Site Description

This study's observational data stems from the Hyytiälä SMEAR II station in Juupajoki, southern Finland (FI-Hyy; 61°51'N, 24°17'E; 160–180 m.a.s.l.; Figure S1a in Supporting Information S1). We chose FI-Hyy as our test site because of its comprehensive study record as well as its extensive and openly available data. The boreal forest site is dominated by the conifers Scots pine (*Pinus sylvestris*) and Norway spruce (*Picea abies*), with scattered occurrences of cold-deciduous silver birch (*Betula pendula*). Dwarf shrubs (*Vaccinium* sp.), mosses and other bryophytes, and tree seedlings characterize the understory vegetation. The site is mainly covered by mineral podzols with medium fertility, mostly having less than 1 m vertical extent. The average annual (1981–2010) air temperature is 3.5°C while the mean total annual precipitation is 711 mm (Hari & Kulmala, 2005; Launiainen et al., 2022; Suni et al., 2003).

3.2.2. Eddy-Covariance Measurements

We assimilated gap-filled GPP and ET flux observations collected at the EC tower located in FI-Hyy. We retrieved these data from Mammarella et al. (2019) and used auxiliary meteorological measurements available via the FLUXNET database (Pastorello et al., 2020) for climate forcing comparisons to GSWP3v1 (Figure S2 in Supporting Information S1 ff.). The GPP and ET products are derived from direct CO_2 and H_2O measurements

using the EC micrometeorological technique: a gold standard to measure turbulent vertical exchanges of energy, momentum, and gases between the atmosphere and the land surface (Baldocchi, 2003). The data in this study stems from a tower setup at a nominal height of 23.3 m above the surface providing recordings since 1997. It uses an ultrasonic anemometer and a closed-path infrared gas analyzer to measure CO_2 and H_2O concentrations and fluxes. Refer to the metadata in Mammarella et al. (2019) for more details about the site and instruments.

In the published end-user data set, the raw 10 Hz EC recordings were processed to 30 min fluxes using standard EC methodologies (Mammarella et al., 2016, 2019). Our use of FATES-SP mode, which focuses exclusively on the biophysics, hydrology, and canopy gas exchange processes, precludes comparison with the net ecosystem exchange (NEE) that is natively observed with the EC technique because it does not simulate the soil carbon cycle and its heterotrophic respiration, nor the stem and root components of autotrophic respiration. Hence, we used the gap-filled GPP data, which is derived from an empirical model that partitions the measured NEE into GPP and temperature-dependent respiration while also correcting for below-measurement-height carbon storage changes. Moreover, fluxes measured under low turbulence conditions were discarded and gap-filled from model extrapolations (Kulmala et al., 2019). The gap-filled ET data is based on storage-corrected primary measurements and has no turbulence or stability filtering (Mammarella et al., 2019). Roughly 38% and 15% of the original NEE and ET data, respectively, were gap-filled during our chosen DA window (cf. Section 2.2.5).

4. Results and Discussion

4.1. Synthetic Data Experiments

4.1.1. IEnKS Calibration

The identical twin experiment evaluating the IEnKS calibration with synthetic data suggests that the chosen setup is in principle capable of providing constrained approximate samples from the CLM-FATES posterior distribution. When compared to the prior run, all experiments that involved the IEnKS reduced both the ensemble-averaged temporal RMSE and bias for both ET and GPP (Table 2; Figure 2). Here, the number of ensemble Kalman iterations (N_a) had a noticeable impact on the results: while K_1 (i.e., the more standard non-iterative EnKS) yielded approximately the same RMSE and bias for ET as K_2 and K_4 , more iterations improved the GPP error statistics (Table 2). The same pattern can also be observed for the marginal posteriors, as illustrated by the $\text{PFT}_{\text{needle}}$ parameters in the kernel density plots in Figure 2: while all experiments involving IEnKS calibration cluster the posterior samples around the true values of BB_{slope} and V_{cmax} , K_2 and K_4 exhibit a markedly narrower concentration when compared to K_1 , that is, the maximum a posteriori (MAP) values are closer to the ST value and the standard deviations are smaller (Table 2). Comparing K_2 to K_4 is less conclusive: in the case of GPP, their RMSE and bias values are almost identical and mainly differ in the sign of the bias, with nearly identical absolute values. This may indicate that the assimilation converges and thus that 2 to 4 iterations suffice for optimized results with the given setup. However, the marginal posteriors for the other two PFTs ($\text{PFT}_{\text{broad}}$ and $\text{PFT}_{\text{grass}}$) from K_4 had ensemble-based MAPs closer to the ST values and smaller standard deviations in all cases, suggesting an advantage of additional iterations to constrain the less abundant PFTs' parameters. In general, the marginal posteriors for the $\text{PFT}_{\text{needle}}$ parameters had smaller standard deviations compared to the other two PFTs' parameters across all experiments. Crucially, as shown in Figure 2, the IEnKS with 4 iterations is able to provide highly constrained samples from the approximate posterior distribution that encompass the true values of the parameters for the dominant PFT ($\text{PFT}_{\text{needle}}$) which can help provide more calibrated emulators.

4.1.2. MCMC Emulator Configuration

Both machine learning models used for emulation, GPR and XGBoost, learned to extrapolate the NLL surface with the CLM-FATES input parameters (ϕ) as predictive features within acceptable error margins. Based on the held-out test data, the GPR performance yielded superior results—almost perfect fits—in terms of RMSE and linear regression statistics (Figure 3). GPR also outperformed other surrogate model approaches in other studies (Wang et al., 2014; Watson-Parris et al., 2021). However, it is important to note that due to our experimental design where the training and test data stem from all K_4 iterations (Figure 2c), we did not test extrapolating the NLL surface across all possible parameter combinations in theory supported by the prior. This is acceptable in the context of CES where we assume the IEnKS provides samples in the posterior region of interest, but it does not necessarily mean that we can reliably emulate the NLL of all a-priori possible parameter combinations.

Table 2

IEnKS Parameter Calibration Statistics and Evaluation Metrics in an Identical Twin Experiment Where ‘Synthetic Truth’ (ST) CLM-FATES Parameter Values Generated the Assimilated Flux Data, That Is, Evapotranspiration (ET) and Gross Primary Production (GPP) Outputs

	Default value	ST	K_1	K_2	K_4
V_{cmax}					
PFT _{needle}	62.0	52.7	58.94 (± 4.76)	50.53 (± 2.28)	53.19 (± 2.09)
PFT _{broad}	58.0	49.3	58.96 (± 5.59)	61.01 (± 6.00)	55.54 (± 5.70)
PFT _{grass}	78.0	66.3	73.93 (± 8.27)	76.48 (± 9.32)	67.68 (± 7.26)
BB _{slope}					
PFT _{needle}	8.0	6.4	5.94 (± 0.32)	6.31 (± 0.14)	6.20 (± 0.12)
PFT _{broad}	8.0	6.4	6.89 (± 0.80)	8.08 (± 0.79)	7.72 (± 0.73)
PFT _{grass}	8.0	6.4	7.49 (± 1.10)	5.95 (± 0.89)	6.43 (± 0.77)
GPP	<i>Prior run</i>				
RMSE ^a	0.89 (± 0.24)	–	0.54 (± 0.06)	0.48 (± 0.01)	0.48 (± 0.01)
r^2	0.98 (± 0.00)	–	0.98 (± 0.00)	0.98 (± 0.00)	0.98 (± 0.00)
Bias ^a	0.48 (± 0.22)	–	0.17 (± 0.09)	–0.02 (± 0.06)	0.02 (± 0.05)
ET					
RMSE ^b	0.29 (± 0.09)	–	0.17 (± 0.00)	0.17 (± 0.00)	0.17 (± 0.00)
r^2	0.97 (± 0.01)	–	0.98 (± 0.00)	0.98 (± 0.00)	0.98 (± 0.00)
Bias ^b	0.12 (± 0.06)	–	0.00 (± 0.00)	0.00 (± 0.00)	0.00 (± 0.00)

Note. The upper part of the table compares the model default (generalized-logit prior location) parameters and ST values for V_{cmax} and BB_{slope} along with statistics for the marginal posterior distributions (mean $\pm$ SD) for different numbers of ensemble Kalman iterations with $N_e = 128$ ensemble members. The lower part shows the resulting model error statistics in relation to the perturbed ST data (ensemble mean $\pm$ SD). ^aIn g C m^{–2} d^{–1}. ^bIn kg H₂O m^{–2} d^{–1}.

Fortunately, we are only concerned with areas of parameter space within the typical set (high probability region) of the posterior such that as long as the IEnKS provides representative samples from this region, then the NLL emulator will be able to help provide refined posterior samples in CES.

As GPR yielded outstanding performance on the test data and provides uncertainty quantification on the emulated NLL surface, we proceed with this approach. We note, however, that the machine learning benchmark XGBoost is also capable of emulating CLM-FATES within acceptable error margins, and that a more rigorous approach to hyperparameter tuning may improve XGBoost's performance (Bentéjac et al., 2021), especially considering hyperparameters associated with model regularization (“alpha” and “lambda” in our case (Table S2 in Supporting Information S1; Chen & Guestrin, 2016)).

Figure 4 shows kernel density estimates of the marginal posteriors for all adjusted CLM-FATES parameters based on the pruned Markov chains (10% burn-in samples removed; chain length of 180,000), which resulted from using the four different emulator models within the MCMC sampling algorithm. Comparing the outcome of a GPR configuration without uncertainty penalization (the first row of plots) to the other approaches illustrates the impact of a less conservative emulator setup: Although this regression model yielded near-perfect fits when evaluated against the held-out test NLL data (Figure 3), it provided the worst representations of the ST parameter values when directly used within the MCMC sampling algorithm. All of the resulting marginal posteriors were bimodal and spanned considerably wider parameter ranges. The likely explanation for this pattern is therefore that the GPR's extrapolation resulted in overly optimistic (i.e., low) NLL predictions in unexplored parameter regions that were not well-represented by the training data (Järvenpää & Corander, 2024). While such extrapolations are not always incorrect (e.g., parameter equifinality could lead to bimodal posteriors), here they are likely an artifact of a poorly generalizing default GPR emulator rather than a reflection of the true posterior that would be obtained by running gold standard MCMC. This is evident from the other conservative GPR emulators incorporating uncertainty penalization, which provided posteriors that also encompass the ST values while being more

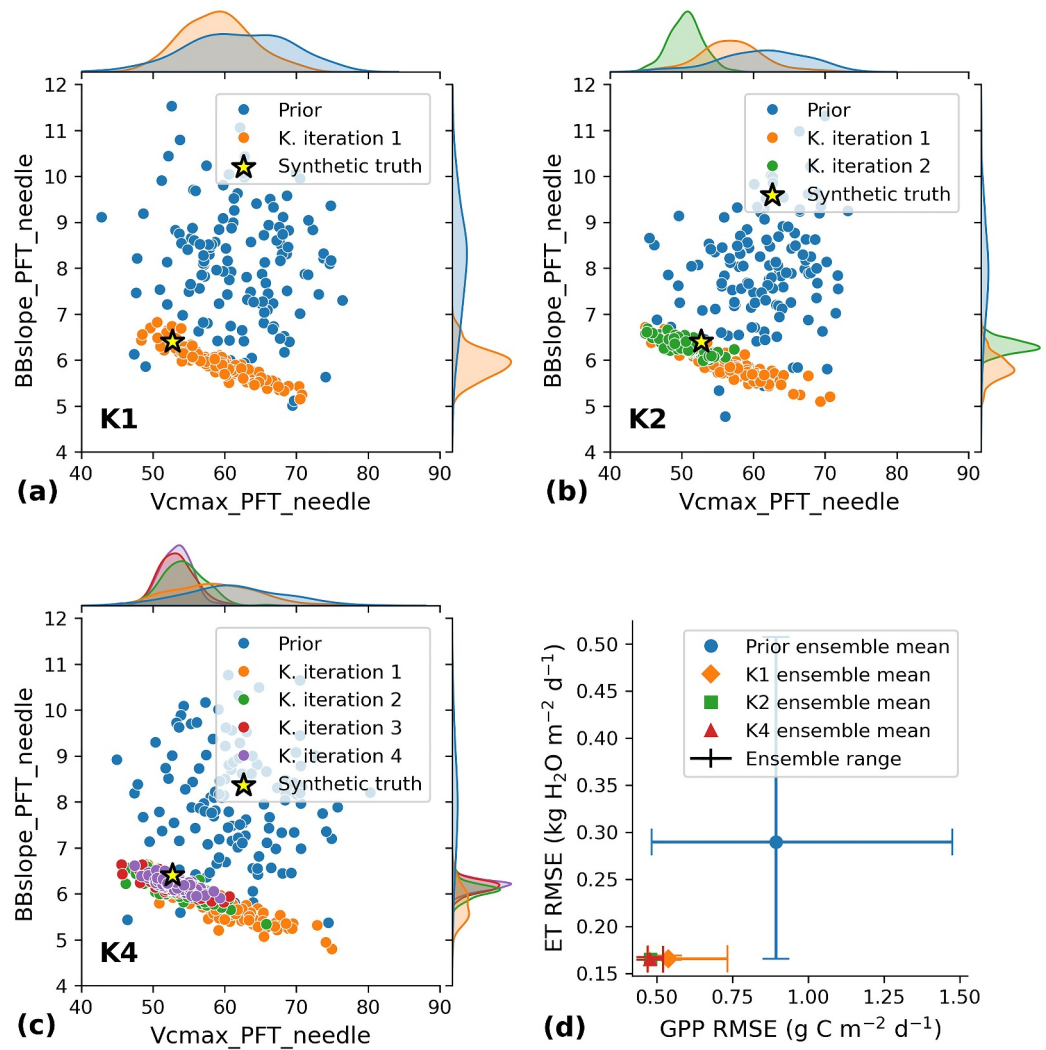


Figure 2. Iterative ensemble Kalman smoother (IEnKS) V_{cmax} and BB_{slope} parameter updates for the dominant plant functional type (PFT), needle-leaf evergreen trees. Panels (a–c) compare parameter updates for a different total number of ensemble Kalman iterations ($N_a = 1, 2$, and 4 , respectively). The “synthetic truth” (ST) parameter values marked with a star produced the CLM-FATES evapotranspiration (ET) and gross primary production (GPP) outputs which were perturbed to generate noisy synthetic observations that were assimilated in these twin experiments (Section 3.1). Panel (d) compares RMSE for GPP and ET against the ST outputs after the final update step of each experiment ($N_e = 128$; the error bars represent ensemble ranges).

efficiently constrained around these values. Recall that XGBoost also employed hyperparameter tuning that affected the in-built regularization.

The results show that even if the IEnKS produces a good but necessarily (due to non-linearity) approximate estimate of the true posterior (Table 2), it is advisable to not blindly accept unsubstantiated GPR likelihood predictions for parameter sets in poorly sampled parameter regions. Inflating the GPR posterior mean predictions with their posterior standard deviations proved successful in this context (Figure 4). When comparing CES-GPR $_{1\sigma}$ to CES-GPR $_{5\sigma}$, a higher penalization factor (β) generally resulted in smaller marginal posterior standard deviations, as expected. Accordingly, penalizing the GPR predictions too severely can also have adverse effects as, for example, illustrated by the BB_{slope} marginal posterior for PFT $_{broad}$ (Figure 4): while CES-GPR $_{1\sigma}$ includes the ST value well within the left tail of the PDF, this is no longer the case for CES-GPR $_{5\sigma}$. The marginal posteriors obtained with the CES-XGBoost emulator are overall rather similar to the ones from CES-GPR $_{1\sigma}$. Nevertheless, fitting CES-XGBoost adequately was generally more computationally cumbersome and the

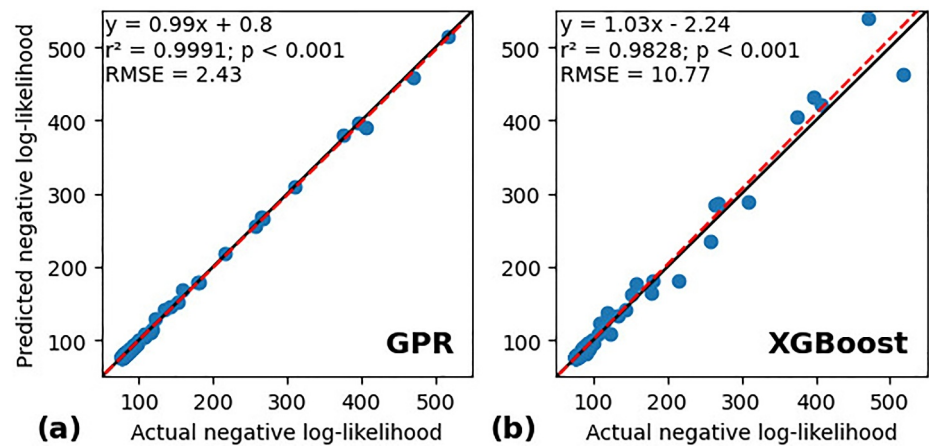


Figure 3. Negative log-likelihood predictions from the Gaussian Process Regression and XGBoost emulators plotted against the corresponding true values in the held-out 20% test data ($n = 128$). Identity lines in black, linear regression dotted lines in red, and associated regression statistics and RMSE in the upper left corners.

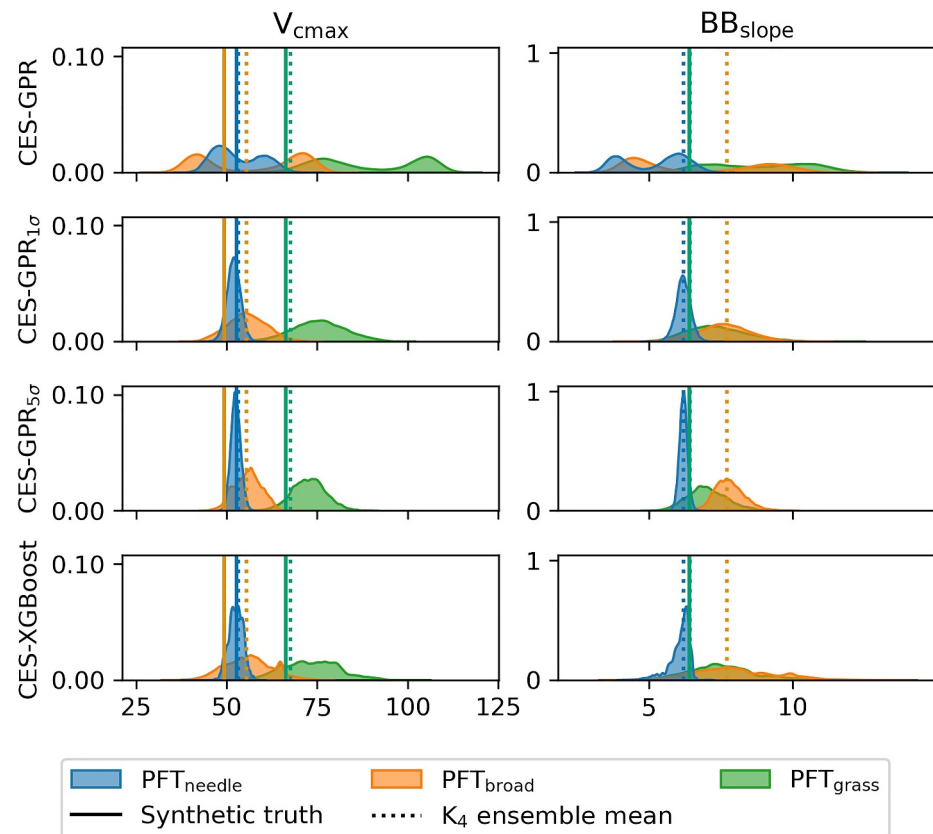


Figure 4. V_{cmax} and BB_{slope} marginal posterior distributions derived from Markov chain Monte Carlo sampling with different negative log-likelihood (NLL) emulators. All emulators were trained on the same data, that is, random 80% of the forward model runs in the K_4 experiment ($n = 512$). The colors represent parameter associations to the different plant functional types (PFT). The colored solid lines show the PFT “synthetic truth” (ST) parameter values while the colored dashed lines depict the ensemble means after the K_4 calibration. Note that the default and hence ST value of BB_{slope} is identical for all PFTs.

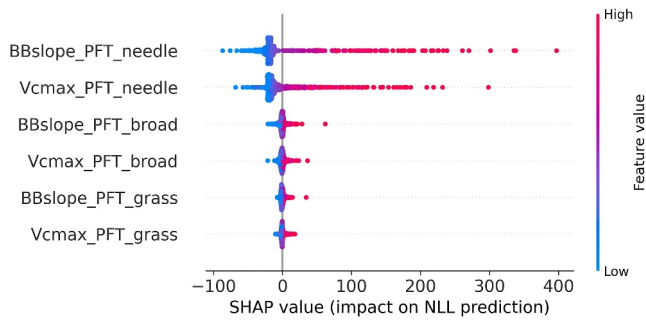


Figure 5. SHAP values for individual negative log-likelihood (NLL) predictions of the CES-GPR emulator ($n = 640$; data from K_4). The higher the *absolute* SHAP value for a parameter, the more relative impact it had on a prediction. The colors indicate the corresponding relative parameter values.

marginal posteriors appeared much less smooth, even though true likelihood surfaces commonly are characterized by some level of smoothness (Fer et al., 2018).

There are theoretical arguments for using GPRs as a surrogate model in addition to our experimental results and the benefit of the in-built GP uncertainty estimation discussed earlier. Unlike many regression models such as XGBoost, GPRs can interpolate and extrapolate the emulated surface by passing exactly through the design (training) points. Therefore, it can be guaranteed that the emulator is exact, that is, $\hat{\mathcal{L}}_g(\phi_d) = \mathcal{L}(\phi_d)$, at these design points ϕ_d where the GP variance smoothly converges to zero (Fer et al., 2018; O'Hagan, 2006). We therefore selected CES-GPR_{1 σ} as the emulator configuration to use in the subsequent eMCMC experiments. However, Cleary et al. (2021) remark that GPRs may not scale well to higher parameter input dimensionality and suggest that neural networks could be preferable in such cases in line with emerging empirical evidence (Alsing

et al., 2019). Future work on LSM emulation with additional adjustable parameters should consider this potential limitation. It should also be noted that because we obtained satisfying results with our setup, we did not test different GPR kernel functions. Rigorously choosing a kernel implies a prior belief in the smoothness and shape of the emulated surface. It may therefore be worthwhile to treat the type of kernel function as an additional tunable hyperparameter in future work (e.g., Duvenaud et al., 2013).

In virtually all test cases, the most well-constrained parameters were V_{cmax} and BB_{slope} of PFT_{needle}. This is unsurprising as it is by far the most abundant PFT at the study site (Figure S1b in Supporting Information S1). Consequently, it is sensible that the parameters for the two less dominant PFTs only provide a minor contribution to the total flux estimates. The fact that it was more difficult to constrain the less dominant PFTs' parameters is reflected both in the wider posterior distributions (Figure 4) and the approximated low absolute SHAP values for the CES-GPR_{1 σ} emulator (Figure 5). The SHAP values also indicate that the emulator predictions were mostly sensitive to both PFT_{needle} parameters, with BB_{slope} overall being the most impactful predictive feature for the NLL. From Figure 5 we can also infer that relatively high values for these two parameters were associated with higher NLL predictions (i.e., lower likelihoods and thus worse fits) and vice versa, which corresponds well to the results we obtained when generating the training data in K_4 (Figures 2c and 2d). Estimating SHAP values, which are gaining increasing popularity in geoscientific fields (Li et al., 2023), therefore appeared worthwhile for an easy-to-implement, theory-driven, and intuitive parameter sensitivity analysis. Note that estimating Kernel SHAP values can be computationally expensive for large data sets (e.g., when dealing with higher parameter dimensionality), but that heuristics exist for fast approximations (Lundberg & Lee, 2017).

4.1.3. CES Benchmarking Experiments

All four CES benchmarking eMCMC experiments reduced the RMSE and bias of modeled ET relative to the ST run by roughly the same order of magnitude, but note that the sign of the bias was both positive and negative (Table 3). The RMSE and bias of GPP, on the other hand, had comparable ranges for UNI, LON, and CES, while the LHS method resulted in worse fits (Table 3). Unsurprisingly, mismatches for both GPP and ET exhibited a seasonal pattern: the different parametrizations mostly affected model mismatches in the warmer months when the vegetation can be expected to photosynthesize more and they approached zero in the winter (Figure S10 in Supporting Information S1). It should also be noted that the IEnKS calibration alone resulted in approximately the same error metrics as the best-performing approaches in the eMCMC experiments (Table 2). This further supports our motivation for using the IEnKS as a suitable initial approximate posterior sampler.

The kernel density plots of the final MCMC chains, on the other hand, revealed striking differences across the different approaches. As a general trend, the two more space-filling design methods UNI and LHS resulted in narrower parameter distributions implying lower estimated posterior uncertainty. These marginal distributions were, however, often located far away from the true parameter values (Figure 6). The highest agreement across all emulator configurations was for the BB_{slope} parameter of PFT_{needle} where all kernel PDFs encompass the ST value.

Table 3

Error Statistics Between a Prior Run (Section 2.2.4) or the Different eMCMC Sampling Strategies (Table 1) Versus the Perturbed Synthetic Truth Data (Ensemble Means $\pm$ SD)

	GPP			ET		
	RMSE ^a	r ²	Bias ^a	RMSE ^b	r ²	Bias ^b
Prior	0.89 (± 0.24)	0.98 (± 0.0)	0.48 (± 0.22)	0.29 (± 0.1)	0.97 (± 0.0)	0.12 (± 0.06)
UNI _{ST}	0.49 (± 0.01)	0.98 (± 0.0)	0.08 (± 0.03)	0.17 (± 0.0)	0.98 (± 0.0)	0.01 (± 0.01)
LHS _{ST}	0.69 (± 0.01)	0.98 (± 0.0)	-0.40 (± 0.01)	0.17 (± 0.0)	0.98 (± 0.0)	0.00 (± 0.0)
LON _{ST}	0.48 (± 0.01)	0.98 (± 0.0)	-0.07 (± 0.05)	0.17 (± 0.0)	0.98 (± 0.0)	0.00 (± 0.01)
CES _{ST}	0.49 (± 0.02)	0.98 (± 0.0)	-0.05 (± 0.07)	0.17 (± 0.0)	0.98 (± 0.0)	-0.01 (± 0.01)

^aIn g C m⁻² d⁻¹. ^bIn kg H₂O m⁻² d⁻¹.

The LON approach achieved the lowest overall error metrics. However, the associated density plots show that it does not perfectly represent all ST parameter values. For instance, the MAP underestimated V_{cmax} for PFT_{needle}. This example indicates that our model setup was subject to equifinality (Beven, 2006). As previously noted, the fluxes are dominated by the most abundant PFT_{needle} and sensitive to both V_{cmax} and BB_{slope} . For LON, the former was underestimated and the latter was slightly overestimated in terms of the MAP versus the ST value. Yet, the error metrics are similar to, for example, CES, which captured V_{cmax} more accurately but slightly underestimated BB_{slope} as per the MAP. Both a higher BB_{slope} and increased V_{cmax} can increase GPP and ET fluxes to varying degrees, for example, depending on plant type. It therefore appears plausible that we have equifinality whereby several parameter combinations could have achieved the same net result of a high likelihood in terms of minimizing GPP and ET residuals.

Notably, CES was the only method where all parameter marginal posteriors encompassed the ST values (Figure 6). Moreover, the marginal posteriors also tended to maintain a higher variance which may indicate a better-calibrated uncertainty in terms of coverage than the other methods which, despite being more confident,

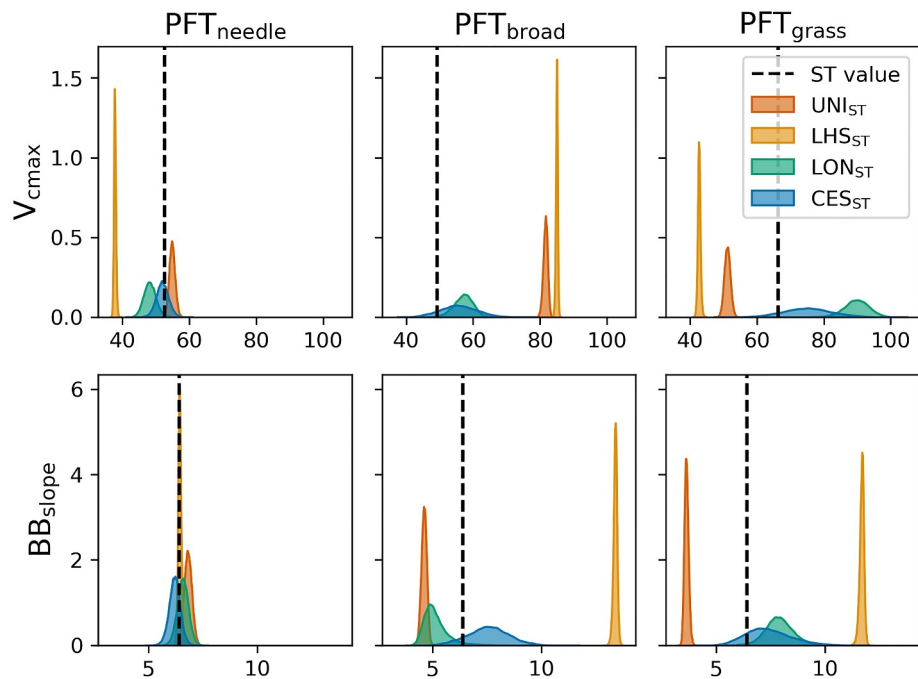


Figure 6. Kernel density plots of the marginal posterior from the pruned Markov chains ($n = 180,000$ after 10% of samples discarded as “burn-in”) derived from the different eMCMC sampling strategies (Table 1). The vertical dashed lines show the synthetic truth (ST) values that generated the assimilated synthetic observations of ET and GPP fluxes.

Table 4
Error Statistics for the Different Experimental Setups Depicted in Figure 7 (Ensemble Means $\pm$ SD)

	GPP			ET		
	RMSE ^a	r ²	Bias ^a	RMSE ^b	r ²	Bias ^b
Default	1.67	0.86	0.78	0.69	0.68	−0.11
UNI _{EC}	1.26 (± 0.00)	0.87 (± 0.00)	−0.05 (± 0.04)	0.65 (± 0.00)	0.67 (± 0.00)	−0.21 (± 0.01)
LHS _{EC}	1.28 (± 0.01)	0.87 (± 0.00)	0.15 (± 0.04)	0.65 (± 0.00)	0.67 (± 0.00)	−0.23 (± 0.01)
LON _{EC}	1.39 (± 0.03)	0.87 (± 0.00)	−0.43 (± 0.05)	0.65 (± 0.00)	0.66 (± 0.00)	−0.25 (± 0.01)
CES _{EC}	1.40 (± 0.03)	0.87 (± 0.00)	−0.43 (± 0.05)	0.66 (± 0.00)	0.65 (± 0.00)	−0.31 (± 0.01)
K ₄	1.55 (± 0.03)	0.86 (± 0.00)	0.68 (± 0.03)	0.65 (± 0.00)	0.67 (± 0.00)	−0.22 (± 0.00)

^aIn g C m^{−2} d^{−1}. ^bIn kg H₂O m^{−2} d^{−1}.

converge to the wrong solutions. Regarding the PFT_{broad} and PFT_{grass} parameters, a lack of constraint can be justified: as these PFTs' relative contributions to the fluxes are low, it can be expected that the available observations are too uninformative to constrain the parameters effectively and that incorporating additional data is needed. The narrow corresponding marginal posteriors provided by the UNI and LHS methods at parameter values substantially diverging from the ST, in turn, show how uncalibrated emulators used within our MCMC implementation can result in misleading emulated posterior samples. Incorporating multiple Markov chains with different starting positions to test if these mix poorly by getting stuck in different local minima, or testing more sophisticated gradient-based MCMC methods (MacKay, 2003; Neal, 1993), could be a possible improvement for a more comprehensive analysis of the sampling step (Gelman et al., 2013; Lu et al., 2017; McElreath, 2020).

A possible explanation for the misleading and overconfident parameter estimation of the UNI and LHS approaches supports our initial hypothesis concerning the inadequacy of data-blind space-filling designs: training the emulators on ensemble runs with too spread-out data points could have resulted in inadequate interpolations of the NLL surface. It seems plausible that the Markov chain could get stuck in local minima contained in the training data as moving to another NLL sample in parameter space that is supported by training data involves more uncertain emulator predictions compared to the more clustered CES samples; one isolated relatively low NLL sample, which might not even stem from the typical region of interest, could dominate the emulator predictions and thus the positioning of the chain. Potential deadlocks were likely aggravated by the penalizing constant (β) we adopted within all experiments as we by design make more conservative GPR predictions in the uncertain underexplored regions of the parameter space. It is therefore questionable if a one-fits-all emulator configuration was the right choice for a fair comparison between the different eMCMC approaches. Developing an enhanced, eMCMC-sampling-design-specific emulator hyperparameter calibration as a part of the workflow could potentially address this issue to some extent, but the underlying problem of potentially not representing the typical posterior region of interest with a sufficient number of samples would likely prevail.

4.2. Eddy-Covariance Data Experiment

Applying the different eMCMC sampling designs to assimilate the real FI-Hyy observations further underlined that the choice of design points can substantially alter the parameter estimation results. First, it should be noted that the default model parameters overall reproduced the GPP and ET observations reasonably well but with room for improvement (Table 4; Figure 7). Both variables' modeled dynamics reasonably captured the (30-day rolling average) seasonal trends, even though there was a tendency to overestimate GPP fluxes during the summer months (Figure 7). Another notable systemic model-data mismatch occurred for the ET fluxes in the winter months: judging by the smoothed 30-day running means, CLM-FATES consistently predicted flux values around or slightly below zero, while the gap-filled observations contained mostly higher, positive values fluctuating approximately between 0 and 1. Note that negative ET fluxes are associated with condensation, that is, dew and frost formation.

Eddy-covariance measurement limitations may partly explain this bias as the method struggles to capture surface fluxes under stably stratified conditions without well-developed turbulence, which are often present during winter months in boreal forests (Aubinet, 2008). On the other hand, the problem could also hint at incomplete process

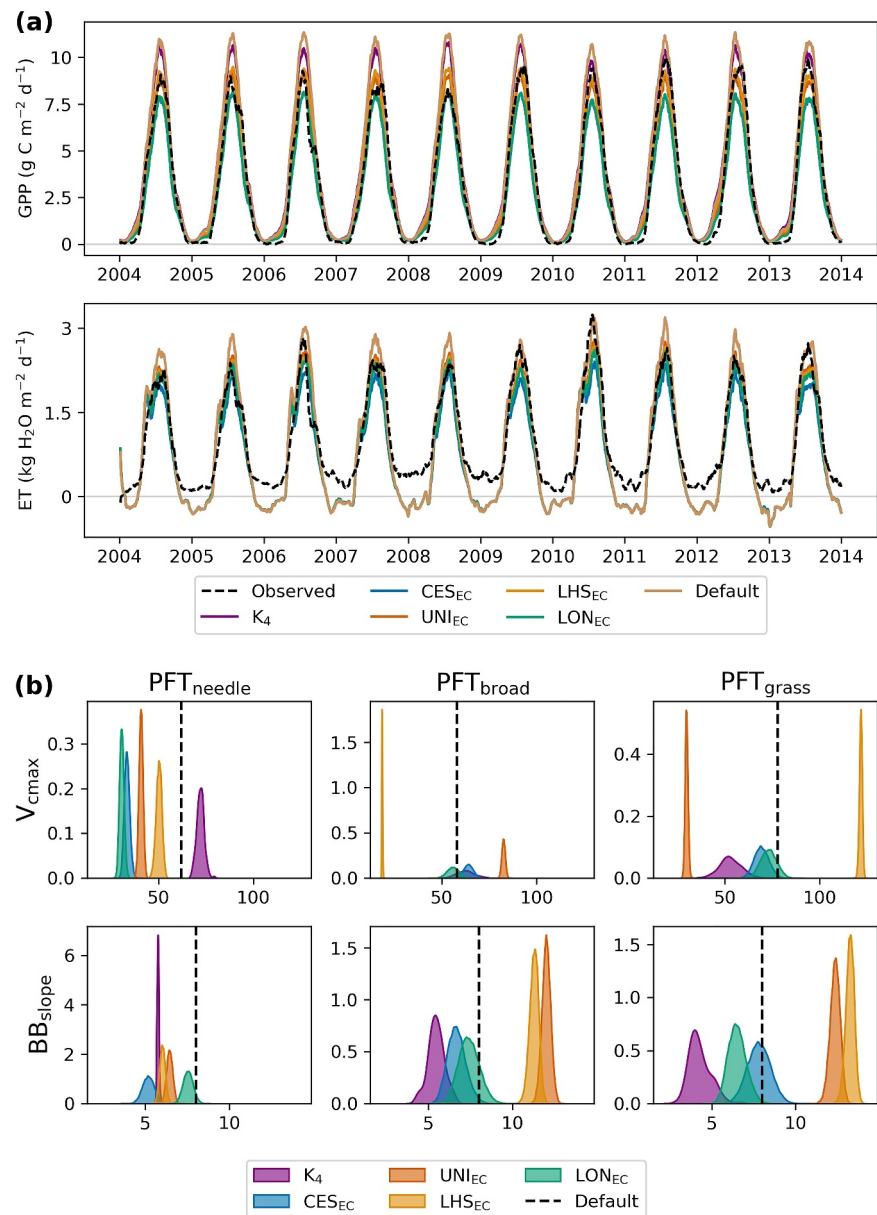


Figure 7. eMCMC approach comparison for assimilating FI-Hyy eddy-covariance (EC) observations (Table 1). Panel (a) shows 30-day rolling averages of the observed and modeled (ensemble means) ET and GPP fluxes. The eMCMC approaches used $n = 128$ ensemble members randomly sampled from the corresponding pruned Markov chains with $n = 180,000$ parameter sets; K₄ shows the last (fourth) iteration of the IEnKS calibration; “Default” represents a single-instance default parameter simulation. Note that uncertainty bands for observations and model ensembles were omitted for readability. Panel (b) shows the corresponding marginal posteriors from the pruned Markov chains (eMCMC), or the final parameter update in the case of K₄, as kernel density plots.

formulations in CLM-FATES. LSMs including CLM were repeatedly shown to inaccurately model ET at the single-site scale, which was, for example, attributed to uncertainties surrounding root water uptake or tree transpiration processes (Launiainen et al., 2019; Strebel et al., 2024). As none of the parameters we adjusted in this work affected the modeled winter ET fluxes, a setup that excludes the winter months in the DA window (e.g., based on satellite snow cover observations or a temperature threshold) could thus have been more sensible to calibrate photosynthesis parameters in this biome. Our use of gap-filled end-user flux data products to infer parameters in these proof-of-concept experiments leaves room for further improvements regarding the observational data handling. EC flux gap-filling methods can introduce biases and uncertainties (e.g., Vekuri

et al., 2023), which could propagate into the CLM-FATES parametrization when included in the DA procedure. Future work may therefore wish to examine the effects of excluding gap-filled observations on the parameter estimates.

A major assumption in our experimental design is the previously alluded to strong constraint formulation of our problem, wherein we assume that all model uncertainty stems from the parameters. In the synthetic experiments this was true by design since the synthetic data were generated corrupting synthetic truth runs from CLM-FATES with noise whose statistics were known. In reality, this assumption always breaks down to a certain extent. In addition to parameter uncertainty, we must also contend with initial condition, input, and process (structural) uncertainty as well as biased observations (Dee, 2005; Dietze, 2017). On the one hand, it is likely generally the case that parameter uncertainty is a major contributor to the overall uncertainty in complex LSMs as evidenced by ongoing parameter perturbation experiments (Kennedy et al., 2024). On the other hand, it is certainly also the case that both the CLM-FATES model and the assimilated observations can be subject to unknown yet systematic errors, which have been shown to obfuscate the parameter estimation exercise (Cameron et al., 2022). As such, it is likely that the differences in performance between the controlled synthetic experiments and the subsequent real experiments can to a considerable extent be attributed to the omission of these sources of uncertainty in the inference exercise.

The site metadata also states that the interannual variability in observed H₂O fluxes might have partly been caused by instrument calibration inaccuracy (Mammarella et al., 2019). Moreover, the forest in FI-Hyy was partially thinned in 2002 and the LAI and stand height have in reality been steadily increasing during our DA window. This did not substantially impact the measured ecosystem carbon and water exchange, which was ascribed to an increase in understory vegetation growth (Launiainen et al., 2022). The LAI and stand heights are fixed values in FATES-SP, however, which means that the model here did not represent the dynamics of this disturbance. Note that the 30-min gap-filled GPP fluxes and the assimilated aggregated daily values contained some small negative values in the winter months. This contradicts the formal GPP definition but can be explained by measurement noise and the statistical partitioning model used to derive the GPP from NEE (cf. Kulmala et al., 2019). It is generally advised to not discard such erroneous gap-filled values during data analysis to avoid introducing biases (Pastorello et al., 2020).

The different eMCMC ensemble parametrizations again mostly influenced the magnitudes of the ET and GPP flux peaks during the summer and, to a small degree, the trajectory of the GPP flux onset during spring (Figure 7). Notably, the different eMCMC model-observation mismatches do not follow consistent patterns throughout the assimilation window: for example, the modeled GPP posterior ensemble means from LON and CES generally showed lower fluxes relative to the other approaches, which represented the EC observations in years such as 2008 well. However, in other years such as 2013, approaches predicting higher fluxes (e.g., K₄) were more consistent with the GPP observations. In addition to the observation data limitations and potentially incomplete process representations discussed earlier, this might also indicate that our DA setup was too simplistic to properly reproduce the more complex dynamics in the real measurements. In particular, the two PFT-specific parameters V_{cmax} and BB_{slope} are only a small subset of the uncertain parameters that influence modeled GPP fluxes (Massoud et al., 2019). Therefore, it is important to recognize that drawing generalizable conclusions from the obtained posteriors is difficult, and that future work should focus on incorporating additional parameters.

When assimilating real observations, model-data mismatches might also be caused by inaccurate climate forcing or static land surface data (cf. Wang et al., 2020), which can introduce biases to parameter estimation. Even though the employed GSWP3v1 data overall matched the site-level meteorological observations reasonably well, considerable inconsistencies especially regarding precipitation and incoming solar radiation remained (Figure S2 in Supporting Information S1 ff.). This might partly explain the mismatches to the modeled GPP and ET fluxes and should be investigated further. We also utilized a constant ambient CO₂ concentration to facilitate the technical model setup (Section 2.1.2), which is inconsistent with the substantial seasonal variations and trends observed in reality (Figure S8 in Supporting Information S1). This could have caused inaccuracies for both GPP and ET calculations, which are both influenced by the ambient CO₂ concentration and performed at a half-hourly time step (cf. Fisher et al., 2023). Future efforts should therefore consider forcing CLM-FATES with the available FLUXNET observations (or other data products) directly or test introducing forcing uncertainty parameters as

part of the Bayesian calibration. These could, for example, account for potential annual or seasonal biases (Alonso-González et al., 2022).

All four eMCMC setups, as well as the IEnKS calibration, reduced the RMSE for both GPP (ca. -10% to -30%) and ET (ca. -5%) relative to a default model run (Table 4; Figure S11 in Supporting Information S1). While the absolute GPP bias also decreased in all cases, the absolute bias in ET increased. This can be explained by the systematic model-data ET mismatches: a general overestimation trend of the default model in most summers partially canceled out the underestimation during winter. The latter winter underestimation of ET persists even after assimilation in all experiments, possibly indicative of a model structural error or at least a bias that cannot be removed through the vegetation-related parameters considered herein. All eMCMC approaches and K_4 reduced the average RMSE of ET by similar margins, and the ensemble spreads in RMSE across all cases are strikingly low; only CES performed slightly worse, but the relative differences are small (Table 4). On the other hand, GPP RMSEs depicted a higher variation across the different approaches. Recall that in all the experiments the model-data mismatch used to update the parameters is encoded in a Gaussian likelihood (Equation 8), which involves a normalized sum of squared residuals. Therein, a normalization is performed using the daily error standard deviations σ_F in (Equation 4) that differ for GPP and ET. These error standard deviations determine the relative weighting of mismatches in ET and GPP used to update from the prior to the posterior parameter distribution. As such, any improvements in the RMSE for ET and GPP after DA should be compared on the relative non-dimensional scale obtained by normalizing with the corresponding observation error standard deviation.

The highest overall reduction in RMSEs was achieved by LHS and UNI (Table 4). At first glance, this seems to contradict our hypothesis that these approaches should be less appropriate for parameter estimation compared to CES. However, here it is important to note that these methods again seemingly led to overconfident (i.e., low variance) approximated marginal posteriors for all parameters, which partly substantially contradict each other (e.g., V_{cmax} for PFT_{grass}; Figure 7). It is therefore highly questionable if either approach approximates the true parameter uncertainty adequately. We also used a relatively large number of ensemble members for the eMCMC experiments ($N_e = 256$). It is thus unsurprising that sampling the six-dimensional parameter space (with four parameters of low sensitivity and which is probably characterized by equifinality) with the space-filling approaches by chance included some samples that reduced the RMSE slightly more effectively. This again motivates changing the experimental setup to test how these results transfer to higher parameter space dimensionality or a lower number of ensemble members. Even though drawing generalizable conclusions is difficult for the many aforementioned reasons, virtually all approaches (except for K_4 in the case of V_{cmax}) decreased the V_{cmax} and BB_{slope} parameters relative to the default value, thereby reducing the overall trend of GPP and ET flux overestimation during the growing seasons.

It is further unlikely that the K_4 IEnKS calibration provided a good approximation of the true posterior. Most notably, the suggested confidence in the most sensitive parameter (BB_{slope} for PFT_{needle}) appears overly optimistic given that UNI and LHS reduced the RMSE for ET and GPP more effectively. This indicates that our positive conclusions for the ST case with perfect model assumptions cannot be directly transferred to the more complex real data problem. This is unfortunate, as the CES workflow relies on the IEnKS to approximate the posterior adequately. Scrutinizing the “calibrate” step through additional experiments is therefore also advised, for example, by testing a higher number of ensemble Kalman iterations for real applications. The marginal posteriors of CES and K_4 further depicted striking dissimilarities and were sometimes not even overlapping, which could be explained by our experimental setup. Unlike Cleary et al. (2021), we included data from all four K_4 iterations rather than just the final iteration. To summarize CES performance for the real data experiment, both the IEnKS calibration and the subsequent emulator-based MCMC sampling successfully reduced the model-observation errors, but it is unlikely that the resulting parameter ensemble adequately captures the true joint posterior.

4.3. Outlook

In line with our hypotheses and the findings in (Cleary et al., 2021), our results illustrate that CES yields promising potential for computationally feasible LSM parameter estimation under idealized assumptions. It is particularly promising that our twin experiments suggest a benefit of data-driven emulator design via approximate Bayesian inference in the initial “calibrate” step over relying on data-blind approaches, already for this relatively low-dimensional problem. Nevertheless, extending this proof-of-concept to more generalizable CLM-FATES

parameter estimation requires additional improvements to the presented workflow. We discuss some promising future avenues in this section.

The employed Gaussian error models lead to unrealistic assumptions in the winter, for example, the perturbed GPP observations can become negative (Section 2.2.6). As we know the assimilated fluxes are positive during the period the parameters have an influence on them, a simple potential improvement to model non-negative observations could be to assume log-normal (Gaussian) errors for the measurements instead, in concert with limiting the DA window to the growing season months as discussed earlier. That way, we could still satisfy the Gaussian requirements of the IEnKS while also introducing error heteroscedasticity (i.e., that the measurement error increases with higher fluxes), which is more in line with real sensor imperfection errors. Following the same reasoning, it may also be worth considering even more complex error models such as the generalized likelihood function introduced by Schoups and Vrugt (2010). In addition, the error hyperparameters that we assumed to be known herein could instead be optimized by model evidence (marginal likelihood) maximization (Pulido et al., 2018) or inferred using a fully hierarchical Bayesian approach (Stordal et al., 2023). Hierarchical Bayesian methods could also be used to infer other hyperparameters (e.g., those related to the prior) in LSMs (cf. Dokoohaki et al., 2022; Fer, Shiklomanov, et al., 2021). In particular, calibrated emulation-based methods such as CES may help further open up the door to performing computationally tractable hierarchical Bayesian inference in LSMs following a recent trend with complex simulations (e.g., Alsing et al., 2019). Future studies may also wish to further investigate the effect of different CES algorithm configurations, particularly the effect of varying the ensemble size and/or increasing the number of Kalman iterations beyond four as well as the length and number of the Markov chains used in sampling.

An intriguing future research direction would be to combine the CES framework, especially the initial ensemble-based calibration with the IEnKS, with the current state-of-the-art methods in emulator-based LSM parameter estimation literature (Fer et al., 2018; Raoult, Beylat, et al., 2024). These typically involve multiple passes or waves of forward model simulations to sequentially refine the emulator design, but avoid this initial ensemble-based calibration. How to best combine such calibration and multiple passes in an active learning approach (Alsing et al., 2019; Murphy, 2023) to robust LSM parameter estimation while optimally balancing accuracy and computational cost remains an open research question. Such an approach would also benefit from an even more direct inclusion of the emulator uncertainty in the sampling step, for example, by building on the work of Fer et al. (2018), and incorporating recent developments on inference using noisy likelihoods from machine learning (Järvenpää & Corander, 2024).

Paying additional attention to the specification of the prior distribution could be another improvement to this work. Even though the posterior approximation should in a relative sense be somewhat insensitive to moderate changes in the prior location parameters, the parameter bounds (a, b) and the prior scale (σ) hyperparameters may be more sensitive and should thus be more carefully evaluated and informed by the literature (cf. Dagon et al., 2020; Kattge et al., 2020; Rogers, 2014). Poorly informed priors can aggravate problems regarding equifinality or result in better model-data fits for biologically implausible reasons (LeBauer et al., 2013). Particularly, our choice to set a fixed $\sigma = 0.3$ logit-normal prior hyperparameter could have resulted in a too-narrow (overly informative) prior, thereby constraining the possible posterior estimation too severely. As a remedy to this, data-driven trait-based priors and hierarchical modeling may help to avoid both equifinality and implausible estimates (Shiklomanov et al., 2020).

In this work, we calibrate CLM-FATES parameters for a single site. However, there is evidence for substantial within-PFT parameter variability (Lin et al., 2015) and single-site calibration commonly leads to overfitting (Raoult et al., 2016; Mäkelä et al., 2019). To make the CES parameter estimation more robust, future efforts should therefore extend the setup to include additional sites, and then investigate the implications for large-scale simulations. For example, it would be possible to embed this CES method within a wider spatio-temporal framework allowing parameter information to be transferred across sites in particular and grid cells more generally (Alonso-González et al., 2023; Dokoohaki et al., 2022; Fer, Shiklomanov, et al., 2021). Closely related to this challenge is that LSMs provide a multitude of outputs, and prediction improvements for the assimilated variables through calibration can introduce biases for others (McCabe et al., 2005). Even if model predictions for variables not included in the assimilation do not degrade after calibration, the parameter estimation can still vary depending on the variables included (Bacour et al., 2015). To improve the robustness of the parameter estimation, it could thus be helpful to integrate additional types of observations into the workflow. Recent remote sensing

products appear particularly promising to help constrain LSM vegetation parameters (Qiu et al., 2018; Radeloff et al., 2024; Shiklomanov et al., 2021).

LSMs need to balance the necessary minimum complexity with computationally efficient algorithms that are flexible enough for global simulations—a challenge that requires the integration of ample amounts of data and knowledge across many disciplines (Argles et al., 2022; Bonan et al., 2021; Fer, Gardella, et al., 2021; Harrison et al., 2021; Keetz et al., 2023). Here, DA approaches are integral for building trust in model results, which means that they should be tightly knit into the existing open community-based LSM development efforts. Echoing widely held opinions (e.g., Collier et al., 2018; Dietze et al., 2013; Watson-Parris et al., 2021), we need model-data fusion systems that are technically accessible, well-documented, easy to use and interpret, and conveniently scalable to new modeling scenarios and approaches. Highly commendable DA systems pursuing these goals already exist (e.g., LeBauer et al., 2013; Watson-Parris et al., 2021). We thus hope that we can eventually incorporate the presented implementation (Keetz & Aalstad, 2024) into a more accessible technical infrastructure for wider applications (cf. Dunbar et al., 2024).

5. Conclusions

This work presents an adaptation of Cleary et al. (2021)'s approximate Bayesian DA framework, “calibrate, emulate, sample,” to estimate photosynthetic parameters with systematic uncertainty quantification in the land surface model CLM-FATES. We tested our implementation by assimilating GPP and ET fluxes in a single-site model setup representing a boreal forest site in southern Finland. We first demonstrate that the iterative ensemble Kalman smoother used within the “calibrate” step successfully approximated the constrained parameter posterior in a twin experiment. Here, the number of Kalman iterations was an important hyperparameter: Using more than one iteration provided considerably better posterior approximations. We further show that a GPR model successfully extrapolated model-data mismatches (the negative log-likelihood surface) in the typical posterior parameter region of interest after being trained on samples from the IEnKS calibration. This suggests that, in this study setup, GPR constituted an adequate surrogate model to replace the full forward CLM-FATES simulations in an adaptive MCMC algorithm to sample the posterior. However, it was crucial to penalize uncertain emulator predictions as unfettered extrapolations of parameter regions poorly represented by training data led to inadequate posterior approximations. We then tested different sampling point design strategies for the emulator training procedure and concluded that CES resulted in the most accurate overall representation of the synthetic truth parameters compared to simpler space-filling designs. Ensembles sampled from the other approximate posteriors also reduced the model-observation errors to similar margins but often converged to parameter values substantially diverging from the “synthetic truth.” This can probably be explained by expected equifinality and a lack of model output sensitivity to the parameters associated with less abundant plant functional types. Finally, assimilating real EC flux observations revealed that although the goal of reducing prediction errors was achieved, the chosen experimental setup was insufficient to draw generalizable conclusions for the model parametrization due to the more complex underlying dynamics and unexplored uncertainty sources. We could therefore not conclusively evaluate the potential benefit of CES over the other sampling design approaches in the real experiment. To do so more comprehensively, future work should include additional adjustable parameters, which CES in theory should be better suited for. It should also put more emphasis on jointly constraining LSM forcing data uncertainties and more rigorously incorporating available parameter prior knowledge. We conclude that CES is a promising new addition to the toolbox tackling the highly challenging task of uncertainty-aware yet computationally feasible LSM calibration.

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Data Availability Statement

The FI-Hyy eddy-covariance data were retrieved from Mammarella et al. (2019) and auxiliary meteorological site observations from Pastorello et al. (2020). Code for the CLM and FATES models can be found at <https://github.com/ESCOMP/CTSM> and <https://github.com/NGEET/fates>, respectively. The code for the DA framework, data

analysis, and figures presented in this work are available in a GitHub repository under MIT license (Keetz & Aalstad, 2024). Relevant analysis results are archived on Zenodo (Keetz, 2024).

Acknowledgments

This work contributes to the LATICE strategic research initiative funded by the Faculty of Mathematics and Natural Sciences at the University of Oslo (Grant UiO/GEO103920). The CLM is primarily supported by the U.S. National Science Foundation, and FATES by the U.S. Dept. of Energy, Office of Science, and Office of Biological and Environmental Research (Next Generation Ecosystem Experiments-Tropics project). KA and NP acknowledge funding from Grant 301552 (Spot-On) and YAY acknowledges funding from 294948 (EMERALD) from the Research Council of Norway. RF acknowledges funding by the European Union's Horizon 2020 (H2020) research and innovation program under Grant agreement nos. 101003536 (ESM2025—Earth System Models for the Future) and 821003 (4C, Climate—Carbon Interactions in the Coming Century). Thanks are due to Matvey Debolskiy for crucial technical support; and to Samuli Launila for kindly sharing auxiliary observation data and background information. We also want to express gratitude for the data provided by FLUXNET, the SMEAR II research infrastructure, and GSWP3v1. The CLM-FATES simulations were performed on resources provided by Sigma2—the National Infrastructure for High Performance Computing and Data Storage in Norway (project no. nn2806k). We sincerely thank the three anonymous reviewers for their valuable feedback and constructive suggestions, which have significantly improved the quality of this work.

References

- Aalstad, K., Westermann, S., Schuler, T. V., Boike, J., & Bertino, L. (2018). Ensemble-based assimilation of fractional snow-covered area satellite retrievals to estimate the snow distribution at Arctic sites. *The Cryosphere*, 12(1), 247–270. <https://doi.org/10.5194/tc-12-247-2018>
- Alonso-González, E., Aalstad, K., Baba, M. W., Revuelto, J., López-Moreno, J. I., Fiddes, J., et al. (2022). The Multiple Snow data Assimilation System (MuSA v1.0). *Geoscientific Model Development*. <https://doi.org/10.5194/gmd-15-9127-2022>
- Alonso-González, E., Aalstad, K., Pirk, N., Mazzolini, M., Treichler, D., Leclercq, P., et al. (2023). Spatio-temporal information propagation using sparse observations in hyper-resolution ensemble-based snow data assimilation. *Hydrology and Earth System Sciences*, 27(24), 4637–4659. <https://doi.org/10.5194/hess-27-4637-2023>
- Alsing, J., Charnock, T., Feeney, S., & Wandelt, B. (2019). Fast likelihood-free cosmology with neural density estimators and active learning. *Monthly Notices of the Royal Astronomical Society*, 488, 4440–4458. <https://doi.org/10.1093/mnras/stz1960>
- Argles, A. P. K., Moore, J. R., & Cox, P. M. (2022). Dynamic global vegetation models: Searching for the balance between demographic process representation and computational tractability. *PLOS Climate*, 1(9), e0000068. <https://doi.org/10.1371/journal.pclm.0000068>
- Arnold, C. P., & Dey, C. H. (1986). Observing-systems simulation experiments: Past, present, and future. *Bulletin of the American Meteorological Society*, 67(6), 687–695. [https://doi.org/10.1175/1520-0477\(1986\)067<0687:OSSEPP>2.0.CO;2](https://doi.org/10.1175/1520-0477(1986)067<0687:OSSEPP>2.0.CO;2)
- Aubinet, M. (2008). Eddy covariance CO₂ flux measurements in nocturnal conditions: An analysis of the problem. *Ecological Applications*, 18(6), 1368–1378. <https://doi.org/10.1890/06-1336.1>
- Bacour, C., Peylin, P., MacBean, N., Rayner, P. J., Delage, F., Chevallier, F., et al. (2015). Joint assimilation of eddy covariance flux measurements and fapar products over temperate forests within a process-oriented biosphere model. *Journal of Geophysical Research: Biogeosciences*, 120(9), 1839–1857. <https://doi.org/10.1002/2015JG002966>
- Baldocchi, D. D. (2003). Assessing the eddy covariance technique for evaluating carbon dioxide exchange rates of ecosystems: Past, present and future. *Global Change Biology*, 9(4), 479–492. <https://doi.org/10.1046/j.1365-2486.2003.00629.x>
- Ball, J. T., Woodrow, I. E., & Berry, J. A. (1987). A model predicting stomatal conductance and its contribution to the control of photosynthesis under different environmental conditions. In J. Biggins (Ed.), *Progress in photosynthesis research* (pp. 221–224). Springer Netherlands. https://doi.org/10.1007/978-94-017-0519-6_48
- Banner, K. M., Irvine, K. M., & Rodhouse, T. J. (2020). The use of Bayesian priors in Ecology: The good, the bad and the not great. *Methods in Ecology and Evolution*, 11(8), 882–889. <https://doi.org/10.1111/2041-210X.13407>
- Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, 54(3), 1937–1967. <https://doi.org/10.1007/s10462-020-09896-5>
- Beven, K. (2006). A manifesto for the equifinality thesis. *Journal of Hydrology*, 320(1), 18–36. <https://doi.org/10.1016/j.jhydrol.2005.07.007>
- Bonan, G. B., & Doney, S. C. (2018). Climate, ecosystems, and planetary futures: The challenge to predict life in Earth system models. *Science*, 359(6375), eaam8328. <https://doi.org/10.1126/science.aam8328>
- Bonan, G. B., Patton, E. G., Finnigan, J. J., Baldocchi, D. D., & Harman, I. N. (2021). Moving beyond the incorrect but useful paradigm: Reevaluating big-leaf and multilayer plant canopies to model biosphere-atmosphere fluxes—A review. *Agricultural and Forest Meteorology*, 306, 108435. <https://doi.org/10.1016/j.agrformet.2021.108435>
- Cameron, D., Hartig, F., Minunno, F., Oberpriller, J., Reineking, B., Van Oijen, M., & Dietze, M. (2022). Issues in calibrating models with multiple unbalanced constraints: The significance of systematic model and data errors. *Methods in Ecology and Evolution*, 13(12), 2757–2770. <https://doi.org/10.1111/2041-210X.14002>
- Carrassi, A., Bocquet, M., Bertino, L., & Evensen, G. (2018). Data assimilation in the geosciences: An overview of methods, issues, and perspectives. *WIREs Climate Change*, 9(5). <https://doi.org/10.1002/wcc.535>
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 785–794). ACM. <https://doi.org/10.1145/2939672.2939785>
- Chopin, N., & Papaspiliopoulos, O. (2020). *An introduction to sequential Monte Carlo*. Springer International Publishing. <https://doi.org/10.1007/978-3-030-47845-2>
- Cleary, E., Garbuno-Inigo, A., Lan, S., Schneider, T., & Stuart, A. M. (2021). Calibrate, emulate, sample. *Journal of Computational Physics*, 424, 109716. <https://doi.org/10.1016/j.jcp.2020.109716>
- Collier, N., Hoffman, F. M., Lawrence, D. M., Keppel-Aleks, G., Koven, C. D., Riley, W. J., et al. (2018). The International Land Model Benchmarking (ILAMB) System: Design, theory, and implementation. *Journal of Advances in Modeling Earth Systems*, 10(11), 2731–2754. <https://doi.org/10.1029/2018MS001354>
- Dagon, K. (2020). CLM5 perturbed parameter ensembles [Dataset]. UCAR/NCAR - GDEX. <https://doi.org/10.5065/9BCC-4A87>
- Dagon, K., Sanderson, B. M., Fisher, R. A., & Lawrence, D. M. (2020). A machine learning approach to emulation and biophysical parameter estimation with the community land model, version 5. *Advances in Statistical Climatology, Meteorology and Oceanography*, 6(2), 223–244. <https://doi.org/10.5194/ascmo-6-223-2020>
- Dee, D. P. (2005). Bias and data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 131(613), 3323–3343. <https://doi.org/10.1256/qj.05.137>
- Dietze, M. (2017). *Ecological forecasting*. Princeton University Press.
- Dietze, M. C., Lebauer, D. S., & Kooper, R. (2013). On improving the communication between models and data. *Plant, Cell and Environment*, 36(9), 1575–1585. <https://doi.org/10.1111/pce.12043>
- Dokoohaki, H., Morrison, B. D., Raiho, A., Serbin, S. P., Zarada, K., Dramko, L., & Dietze, M. (2022). Development of an open-source regional data assimilation system in pecan v. 1.7.2: Application to carbon cycle reanalysis across the contiguous us using Sipnet. *Geoscientific Model Development*, 15(8), 3233–3252. <https://doi.org/10.5194/gmd-15-3233-2022>
- Dunbar, O. R., Bieli, M., Garbuno-Inigo, A., Howland, M., de Souza, A. N., Mansfield, L. A., et al. (2024). CalibrateEmulateSample.jl: Accelerated parametric uncertainty quantification. *Journal of Open Source Software*, 9(97), 6372. <https://doi.org/10.21105/joss.06372>
- Dunbar, O. R., Garbuno-Inigo, A., Schneider, T., & Stuart, A. (2021). Calibration and uncertainty quantification of convective parameters in an idealized GCM. *Journal of Advances in Modeling Earth Systems*, 13(9). <https://doi.org/10.1029/2020MS002454>
- Duvenaud, D., Lloyd, J., Grosse, R., Tenenbaum, J., & Zoubin, G. (2013). Structure discovery in nonparametric regression through compositional kernel search. In S. Dasgupta & D. McAllester (Eds.), *Proceedings of the 30th international conference on machine learning* (Vol. 28, pp. 1166–1174). PMLR.

- Emerick, A. A., & Reynolds, A. C. (2013). Ensemble smoother with multiple data assimilation. *Computers and Geosciences*, 55, 3–15. <https://doi.org/10.1016/j.cageo.2012.03.011>
- Evensen, G. (2003). The ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean Dynamics*, 53(4), 343–367. <https://doi.org/10.1007/s10236-003-0036-9>
- Evensen, G., Vossepoel, F. C., & van Leeuwen, P. J. (2022). *Data assimilation fundamentals: A unified formulation of the state and parameter estimation problem*. Springer International Publishing. <https://doi.org/10.1007/978-3-030-96709-3>
- Fer, I., Gardella, A. K., Shiklomanov, A. N., Campbell, E. E., Cowdery, E. M., De Kauwe, M. G., et al. (2021). Beyond ecosystem modeling: A roadmap to community cyberinfrastructure for ecological data-model integration. *Global Change Biology*, 27(1), 13–26. <https://doi.org/10.1111/gcb.15409>
- Fer, I., Kelly, R., Moorcroft, P. R., Richardson, A. D., Cowdery, E. M., & Dietze, M. C. (2018). Linking big models to big data: Efficient ecosystem model calibration through Bayesian model emulation. *Biogeosciences*, 15(19), 5801–5830. <https://doi.org/10.5194/bg-15-5801-2018>
- Fer, I., Shiklomanov, A., Novick, K. A., Gough, C. M., Altaf Arain, M., Chen, J., et al. (2021). Capturing site-to-site variability through Hierarchical Bayesian calibration of a process-based dynamic vegetation model (preprint). *Ecology*. <https://doi.org/10.1101/2021.04.28.441243>
- Fernández-Martínez, M., Peñuelas, J., Chevallier, F., Ciais, P., Obersteiner, M., Rödenbeck, C., et al. (2023). Diagnosing destabilization risk in global land carbon sinks. *Nature*, 615(7954), 848–853. <https://doi.org/10.1038/s41586-023-05725-1>
- Fisher, R. A., Knox, R. G., Koven, C. D., Lemieux, G., Xu, C., Christoffersen, B. O., et al. (2023). *FATES technical documentation*. Zenodo. <https://doi.org/10.5281/ZENODO.3517271>
- Fisher, R. A., & Koven, C. D. (2020). Perspectives on the future of land surface models and the challenges of representing complex terrestrial systems. *Journal of Advances in Modeling Earth Systems*, 12(4). <https://doi.org/10.1029/2018MS001453>
- Fisher, R. A., Muszala, S., Versteinsten, M., Lawrence, P., Xu, C., McDowell, N. G., et al. (2015). Taking off the training wheels: The properties of a dynamic vegetation model without climate envelopes, CLM4.5. *Geoscientific Model Development*, 8(11), 3593–3619. <https://doi.org/10.5194/gmd-8-3593-2015>
- Friedlingstein, P., O'Sullivan, M., Jones, M. W., Andrew, R. M., Gregor, L., Hauck, J., et al. (2022). Global carbon budget 2022. *Earth System Science Data*, 14(11), 4811–4900. <https://doi.org/10.5194/essd-14-4811-2022>
- Garbuno-Inigo, A., Hoffmann, F., Li, W., & Stuart, A. (2020). Interacting Langevin diffusions: Gradient structure and ensemble Kalman sampler. *SIAM Journal on Applied Dynamical Systems*, 19(1), 412–441. <https://doi.org/10.1137/19M1251655>
- Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., & Rubin, D. (2013). *Bayesian data analysis* (3rd ed.). CRC Press.
- Gramacy, R. B. (2020). *Surrogates: Gaussian process modeling, design and optimization for the applied Sciences*. Chapman Hall/CRC.
- Groenke, B., Langer, M., Nitzbon, J., Westermann, S., Gallego, G., & Boike, J. (2023). Investigating the thermal state of permafrost with Bayesian inverse modeling of heat transfer. *The Cryosphere*, 17(8), 3505–3533. <https://doi.org/10.5194/tc-17-3505-2023>
- Guidicelli, M., Aalstad, K., Treichler, D., & Salzmänn, N. (2023). A combined data assimilation and deep learning approach for continuous spatio-temporal SWE reconstruction from sparse ground tracks. *SSRN. (Preprint)*. <https://doi.org/10.2139/ssrn.4489553>
- Haario, H., Saksman, E., & Tamminen, J. (2001). An adaptive Metropolis algorithm. *Bernoulli*, 7(2), 223. <https://doi.org/10.2307/3318737>
- Hari, P., & Kulmala, M. (2005). Station for Measuring Ecosystem-Atmosphere Relations (SMEAR II). *Boreal Environment Research*, 10(5), 315–322.
- Harrison, S. P., Cramer, W., Franklin, O., Prentice, I. C., Wang, H., Brännström, A., et al. (2021). Eco-evolutionary optimality as a means to improve vegetation and land-surface models. *New Phytologist*, 231(6), 2125–2141. <https://doi.org/10.1111/nph.17558>
- Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1), 97. <https://doi.org/10.2307/2334940>
- Helgeson, C., Srikrishnan, V., Keller, K., & Tuana, N. (2021). Why simpler computer simulation models can be epistemically better for informing decisions. *Philosophy in Science*, 88(2), 213–233. <https://doi.org/10.1086/711501>
- Hennig, P., Osborne, M., & Kersting, H. (2022). *Probabilistic numerics*. Cambridge University Press. <https://doi.org/10.1017/9781316681411>
- Horvath, P., Tang, H., Halvorsen, R., Stordal, F., Tallaksen, L. M., Berntsen, T. K., & Bryn, A. (2021). Improving the representation of high-latitude vegetation distribution in dynamic global vegetation models. *Biogeosciences*, 18(1), 95–112. <https://doi.org/10.5194/bg-18-95-2021>
- Huang, M., Ray, J., Hou, Z., Ren, H., Liu, Y., & Swiler, L. (2016). On the applicability of surrogate-based Markov chain Monte Carlo-Bayesian inversion to the Community Land Model: Case studies at flux tower sites. *Journal of Geophysical Research: Atmospheres*, 121(13), 7548–7563. <https://doi.org/10.1002/2015JD024339>
- IPCC. (2023). *Climate change 2021 – the physical science basis: Working group I contribution to the sixth assessment report of the intergovernmental panel on climate change* (1st ed.). Cambridge University Press. <https://doi.org/10.1017/9781009157896>
- Järvenpää, M., & Corander, J. (2024). Approximate Bayesian inference from noisy likelihoods with Gaussian process emulated MCMC. *Journal of Machine Learning Research*, 25(366), 1–55. Retrieved from <http://jmlr.org/papers/v25/21-0421.html>
- Kattge, J., Bönsch, G., Díaz, S., Lavorel, S., Prentice, I. C., Leadley, P., et al. (2020). TRY plant trait database—Enhanced coverage and open access. *Global Change Biology*, 26(1), 119–188. <https://doi.org/10.1111/gcb.14904>
- Katzfuss, M., Stroud, J. R., & Wikle, C. K. (2016). Understanding the ensemble Kalman filter. *The American Statistician*, 70(4), 350–357. <https://doi.org/10.1080/00031305.2016.1141709>
- Katzfuss, M., Stroud, J. R., & Wikle, C. K. (2020). Ensemble kalman methods for high-dimensional hierarchical dynamic space-time models. *Journal of the American Statistical Association*, 115(530), 866–885. <https://doi.org/10.1080/01621459.2019.1592753>
- Keetz, L. T. (2024). CLM-FATES parameter estimation using the “calibrate, emulate, sample” approach—Experimental results [Dataset]. Zenodo. <https://doi.org/10.5281/zenodo.11938487>
- Keetz, L. T., & Aalstad, K. (2024). CLM-FATES parameter estimation using the “calibrate, emulate, sample” approach [Software]. <https://doi.org/10.5281/zenodo.11840624>
- Keetz, L. T., Lieungh, E., Karimi-Asli, K., Geange, S. R., Gelati, E., Tang, H., et al. (2023). Climate–ecosystem modelling made easy: The Land Sites Platform. *Global Change Biology*, 29(15), 4440–4452. <https://doi.org/10.1111/gcb.16808>
- Kennedy, D., Dagon, K., Lawrence, D. M., Fisher, R. A., Sanderson, B. M., Collier, N., et al. (2024). One-at-a-time parameter perturbation ensemble of the community land model, version 5.1. *ESS Open Archive*. <https://doi.org/10.22541/essoar.172745082.24089296/v1>
- Kennedy, M. C., & O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society—Series B: Statistical Methodology*, 63(3), 425–464. <https://doi.org/10.1111/1467-9868.00294>
- Koven, C. D., Arora, V. K., Cadule, P., Fisher, R. A., Jones, C. D., Lawrence, D. M., et al. (2022). Multi-century dynamics of the climate and carbon cycle under both high and net negative emissions scenarios. *Earth System Dynamics*, 13(2), 885–909. <https://doi.org/10.5194/esd-13-885-2022>

- Koven, C. D., Knox, R. G., Fisher, R. A., Chambers, J. Q., Christoffersen, B. O., Davies, S. J., et al. (2020). Benchmarking and parameter sensitivity of physiological and vegetation dynamics using the Functionally Assembled Terrestrial Ecosystem Simulator (FATES) at Barro Colorado Island, Panama. *Biogeosciences*, 17(11), 3017–3044. <https://doi.org/10.5194/bg-17-3017-2020>
- Kulmala, L., Pumpanen, J., Kolari, P., Dengel, S., Berninger, F., Köster, K., et al. (2019). Inter- and intra-annual dynamics of photosynthesis differ between forest floor vegetation and tree canopy in a subarctic Scots pine stand. *Agricultural and Forest Meteorology*, 271, 1–11. <https://doi.org/10.1016/j.agrformet.2019.02.029>
- Kumar, S., Kolassa, J., Reichle, R., Crow, W., de Lannoy, G., de Rosnay, P., et al. (2022). An agenda for land data assimilation priorities: Realizing the promise of terrestrial water, energy, and vegetation observations from space. *Journal of Advances in Modeling Earth Systems*, 14(11), e2022MS003259. <https://doi.org/10.1029/2022MS003259>
- Launiainen, S., Guan, M., Salmivaara, A., & Kieloaho, A.-J. (2019). Modeling boreal forest evapotranspiration and water balance at stand and catchment scales: A spatial approach. *Hydrology and Earth System Sciences*, 23(8), 3457–3480. <https://doi.org/10.5194/hess-23-3457-2019>
- Launiainen, S., Katul, G. G., Leppä, K., Kolari, P., Aslan, T., Grönholm, T., et al. (2022). Does growing atmospheric CO₂ explain increasing carbon sink in a boreal coniferous forest? *Global Change Biology*, 28(9), 2910–2929. <https://doi.org/10.1111/gcb.16117>
- Lawrence, D. M., Fisher, R. A., Koven, C. D., Oleson, K. W., Swenson, S. C., Bonan, G., et al. (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>
- LeBauer, D. S., Wang, D., Richter, K. T., Davidson, C. C., & Dietze, M. C. (2013). Facilitating feedbacks between field measurements and ecosystem models. *Ecological Monographs*, 83(2), 133–154. <https://doi.org/10.1890/12-0137.1>
- Leonardi, S., Magnani, F., Nolè, A., Van Noije, T., & Borghetti, M. (2015). A global assessment of forest surface albedo and its relationships with climate and atmospheric nitrogen deposition. *Global Change Biology*, 21(1), 287–298. <https://doi.org/10.1111/gcb.12681>
- Li, L., Fang, Y., Zheng, Z., Shi, M., Longo, M., Koven, C. D., et al. (2023). A machine learning approach targeting parameter estimation for plant functional type coexistence modeling using ELM-FATES (v2.0). *Geoscientific Model Development*, 16(14), 4017–4040. <https://doi.org/10.5194/gmd-16-4017-2023>
- Liao, C., Lu, X., Huang, Y., Tao, F., Lawrence, D. M., Koven, C. D., et al. (2023). Matrix approach to accelerate spin-up of CLM5. *Journal of Advances in Modeling Earth Systems*, 15(8), e2023MS003625. <https://doi.org/10.1029/2023MS003625>
- Lin, Y.-S., Medlyn, B. E., Duursma, R. A., Prentice, I. C., Wang, H., Baig, S., et al. (2015). Optimal stomatal behaviour around the world. *Nature Climate Change*, 5(5), 459–464. <https://doi.org/10.1038/nclimate2550>
- Lorentzen, R. J., & Naevdal, G. (2011). An iterative ensemble kalman filter. *IEEE Transactions on Automatic Control*, 56(8), 1990–1995. <https://doi.org/10.1109/TAC.2011.2154430>
- Lu, D., Ricciuto, D., Walker, A., Safta, C., & Munger, W. (2017). Bayesian calibration of terrestrial ecosystem models: A study of advanced Markov chain Monte Carlo methods. *Biogeosciences*, 14(18), 4295–4314. <https://doi.org/10.5194/bg-14-4295-2017>
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon (Ed.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc.
- MacKay, D. (2003). *Information theory, inference, and learning algorithms*. Cambridge University Press.
- Mäkelä, J., Knauer, J., Aurela, M., Black, A., Heimann, M., Kobayashi, H., et al. (2019). Parameter calibration and stomatal conductance formulation comparison for boreal forests with adaptive population importance sampler in the land surface model JSBACH. *Geoscientific Model Development*, 12(9), 4075–4098. <https://doi.org/10.5194/gmd-12-4075-2019>
- Mäkelä, J., Susiluoto, J., Markkanen, T., Aurela, M., Järvinen, H., Mammarella, I., et al. (2016). Constraining ecosystem model with adaptive Metropolis algorithm using boreal forest site eddy covariance measurements. *Nonlinear Processes in Geophysics*, 23(6), 447–465. <https://doi.org/10.5194/npg-23-447-2016>
- Mametjanov, A., Norris, B., Zeng, X., Drewniak, B., Utke, J., Anitescu, M., & Hovland, P. (2012). Applying automatic differentiation to the community land model. In S. Forth, P. Hovland, E. Phipps, J. Utke, & A. Walther (Eds.), *Recent advances in algorithmic differentiation* (Vol. 87, pp. 47–57). Springer Berlin Heidelberg. (Series Title: Lecture Notes in Computational Science and Engineering). https://doi.org/10.1007/978-3-642-30023-3_5
- Mammarella, I., Peltola, O., Nordbo, A., Järvi, L., & Rannik, U. (2016). Quantifying the uncertainty of eddy covariance fluxes due to the use of different software packages and combinations of processing steps in two contrasting ecosystems. *Atmospheric Measurement Techniques*, 9(10), 4915–4933. <https://doi.org/10.5194/amt-9-4915-2016>
- Mammarella, I., Rannik, U., Launiainen, S., Alekseychik, P., Peltola, O., & Keronen, P. (2019). SMEAR II Hyytiälä forest eddy covariance [Dataset]. <https://doi.org/10.23729/40F64739-11D1-4E5F-8DC2-DA931512C91C>
- Massoud, E. C. (2019). Emulation of environmental models using polynomial chaos expansion. *Environmental Modelling and Software*, 111, 421–431. <https://doi.org/10.1016/j.envsoft.2018.10.008>
- Massoud, E. C., Xu, C., Fisher, R. A., Knox, R. G., Walker, A. P., Serbin, S. P., et al. (2019). Identification of key parameters controlling demographically structured vegetation dynamics in a land surface model: CLM4.5(FATES). *Geoscientific Model Development*, 12(9), 4133–4164. <https://doi.org/10.5194/gmd-12-4133-2019>
- McCabe, M., Franks, S., & Kalma, J. (2005). Calibration of a land surface model using multiple data sets. *Journal of Hydrology*, 302(1–4), 209–222. <https://doi.org/10.1016/j.jhydrol.2004.07.002>
- McElreath, R. (2020). *Statistical rethinking* (2nd ed.). CRC Press.
- McKay, M. D., Beckman, R. J., & Conover, W. J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21(2), 239. <https://doi.org/10.2307/1268522>
- McNeill, D., Robertson, E., & Wiltshire, A. (2024). Constraining the carbon cycle in JULES-ES-1.0. *Geoscientific Model Development*, 17(3), 1059–1089. <https://doi.org/10.5194/gmd-17-1059-2024>
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6), 1087–1092. <https://doi.org/10.1063/1.1699114>
- Miner, G. L., & Bauerle, W. L. (2017). Seasonal variability of the parameters of the Ball-Berry model of stomatal conductance in maize (*Zea mays* L.) and sunflower (*Helianthus annuus* L.) under well-watered and water-stressed conditions: Ball-Berry model parameters. *Plant, Cell and Environment*, 40(9), 1874–1886. <https://doi.org/10.1111/pce.12990>
- Murphy, K. (2023). *Probabilistic machine learning: Advanced topics*. MIT Press.
- Neal, R. M. (1993). *Probabilistic inference using Markov chain Monte Carlo methods. Technical report CRG-TR-93-1, department of computer science*. University of Toronto.
- O'Hagan, A. (2006). Bayesian analysis of computer code outputs: A tutorial. *Reliability Engineering and System Safety*, 91(10–11), 1290–1300. <https://doi.org/10.1016/j.ress.2005.11.025>

- Pastorello, G., Trotta, C., Canfora, E., Chu, H., Christianson, D., Cheah, Y.-W., et al. (2020). The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data. *Scientific Data*, 7(1), 225. <https://doi.org/10.1038/s41597-020-0534-3>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Pérez-Ruiz, E. R., Vivoni, E. R., Yépez, E. A., Rodríguez, J. C., Gochis, D. J., Robles-Morua, A., et al. (2021). Landscape controls on water-energy-carbon fluxes across different ecosystems during the North American monsoon. *Journal of Geophysical Research: Biogeosciences*, 126(5), e2020JG005809. <https://doi.org/10.1029/2020JG005809>
- Pirk, N., Aalstad, K., Westermann, S., Vatne, A., Van Hove, A., Tallaksen, L. M., et al. (2022). Inferring surface energy fluxes using drone data assimilation in large eddy simulations. *Atmospheric Measurement Techniques*, 15(24), 7293–7314. <https://doi.org/10.5194/amt-15-7293-2022>
- Pulido, M., Tandeo, P., Bocquet, M., Carrassi, A., & Lucini, M. (2018). Stochastic parameterization identification using ensemble kalman filtering combined with maximum likelihood methods. *Tellus A: Dynamic Meteorology and Oceanography*, 70(1), 1–17. <https://doi.org/10.1080/16000870.2018.1442099>
- Qiu, B., Xue, Y., Fisher, J. B., Guo, W., Berry, J. A., & Zhang, Y. (2018). Satellite chlorophyll fluorescence and soil moisture observations lead to advances in the predictive understanding of global terrestrial coupled carbon-water cycles. *Global Biogeochemical Cycles*, 32(3), 360–375. <https://doi.org/10.1002/2017GB005744>
- Radeloff, V. C., Roy, D. P., Wulder, M. A., Anderson, M., Cook, B., Crawford, C. J., et al. (2024). Need and vision for global medium-resolution Landsat and Sentinel-2 data products. *Remote Sensing of Environment*, 300, 113918. <https://doi.org/10.1016/j.rse.2023.113918>
- Rannik, U., Peltola, O., & Mammarella, I. (2016). Random uncertainties of flux measurements by the eddy covariance technique. *Atmospheric Measurement Techniques*, 9(10), 5163–5181. <https://doi.org/10.5194/amt-9-5163-2016>
- Raoult, N., Beylat, S., Salter, J. M., Hourdin, F., Bastrikov, V., Ottlé, C., & Peylin, P. (2024). Exploring the potential of history matching for land surface model calibration. *Geoscientific Model Development*, 17(15), 5779–5801. <https://doi.org/10.5194/gmd-17-5779-2024>
- Raoult, N., et al. (2024). *Parameter estimation in land surface models: Challenges and opportunities with data assimilation and machine learning*. ESS Open Archive. <https://doi.org/10.22541/essoar.172838640.01153603/v1>
- Raoult, N. M., Jupp, T. E., Cox, P. M., & Luke, C. M. (2016). Land-surface parameter optimisation using data assimilation techniques: The adjULES system V1.0. *Geoscientific Model Development*, 9(8), 2833–2852. <https://doi.org/10.5194/gmd-9-2833-2016>
- Rasmussen, C., & Williams, C. (2005). *Gaussian processes for machine learning*. The MIT Press. <https://doi.org/10.7551/mitpress/3206.001.0001>
- Richardson, A. D., Hollinger, D. Y., Burba, G. G., Davis, K. J., Flanagan, L. B., Katul, G. G., et al. (2006). A multi-site analysis of random error in tower-based measurements of carbon and energy fluxes. *Agricultural and Forest Meteorology*, 136(1–2), 1–18. <https://doi.org/10.1016/j.agrformet.2006.01.007>
- Roberts, G., Gelman, A., & Gilks, W. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Annals of Applied Probability*, 7(1), 110–120. <https://doi.org/10.1214/aop/1034625254>
- Rogers, A. (2014). The use and misuse of $V_{c,max}$ in Earth system models. *Photosynthesis Research*, 119(1–2), 15–29. <https://doi.org/10.1007/s11120-013-9818-1>
- Santner, T., Williams, B., & Notz, W. (2018). *The design and analysis of computer experiments*. Springer. <https://doi.org/10.1007/978-1-4939-8847-1>
- Särkkä, S., Solin, A., & Hartikainen, J. (2013). Spatiotemporal learning via infinite-dimensional bayesian filtering and smoothing: A look at Gaussian process regression through Kalman filtering. *IEEE Signal Processing Magazine*, 30(4), 51–61. <https://doi.org/10.1109/MSP.2013.2246292>
- Sawada, Y. (2020). Machine learning accelerates parameter optimization and uncertainty assessment of a land surface model. *Journal of Geophysical Research: Atmospheres*, 125(20). <https://doi.org/10.1029/2020JD032688>
- Schoups, G., & Vrugt, J. A. (2010). A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors. *Water Resources Research*, 46(10), W10531. <https://doi.org/10.1029/2009WR008933>
- Seiler, C., Melton, J. R., Arora, V. K., Sitch, S., Friedlingstein, P., Anthoni, P., et al. (2022). Are Terrestrial Biosphere Models fit for simulating the global land carbon sink? *Journal of Advances in Modeling Earth Systems*, 14(5). <https://doi.org/10.1029/2021MS002946>
- Shiklomanov, A. N., Cowdery, E. M., Bahn, M., Byun, C., Jansen, S., Kramer, K., et al. (2020). Does the leaf economic spectrum hold within plant functional types? A bayesian multivariate trait meta-analysis. *Ecological Applications*, 30(3), e02064. <https://doi.org/10.1002/eap.2064>
- Shiklomanov, A. N., Dietze, M. C., Fer, I., Viskari, T., & Serbin, S. P. (2021). Cutting out the middleman: Calibrating and validating a dynamic vegetation model (ED2-PROSPECT5) using remotely sensed surface reflectance. *Geoscientific Model Development*, 14(5), 2603–2633. <https://doi.org/10.5194/gmd-14-2603-2021>
- Shu, Z., Zhang, B., Tian, L., & Zhao, X. (2022). Improving dynamic vegetation modeling in Noah-MP by parameter optimization and data assimilation over China's loess plateau. *Journal of Geophysical Research: Atmospheres*, 127(19), e2022JD036703. <https://doi.org/10.1029/2022JD036703>
- Stordal, A., Lorentzen, R., & Fossum, K. (2023). Marginalized iterative ensemble smoothers for data assimilation. *Computational Geosciences*, 27(6), 975–986. <https://doi.org/10.1007/s10596-023-10242-1>
- Stordal, A. S., & Elsheikh, A. H. (2015). Iterative ensemble smoothers in the annealed importance sampling framework. *Advances in Water Resources*, 86, 231–239. <https://doi.org/10.1016/j.advwatres.2015.09.030>
- Strebel, L., Bogen, H., Vereecken, H., Andreassen, M., Aranda-Barranco, S., & Hendricks Franssen, H.-J. (2024). Evapotranspiration prediction for European forest sites does not improve with assimilation of in situ soil water content data. *Hydrology and Earth System Sciences*, 28(4), 1001–1026. <https://doi.org/10.5194/hess-28-1001-2024>
- Strebel, L., Bogen, H. R., Vereecken, H., & Hendricks Franssen, H.-J. (2022). Coupling the community land model version 5.0 to the parallel data assimilation framework PDAF: Description and applications. *Geoscientific Model Development*, 15(2), 395–411. <https://doi.org/10.5194/gmd-15-395-2022>
- Stuart, A. (2010). Inverse problems: A bayesian perspective. *Acta Numerica*, 19, 451–559. <https://doi.org/10.1017/S0962492910000061>
- Sun, T., Rinne, J., Reissell, A., Altimir, N., Keronen, P., Rannik, U., et al. (2003). Long-term measurements of surface fluxes above a Scots pine forest in Hyytiälä, Southern Finland, 1996–2001. *BOREAL ENVIRONMENT RESEARCH*, 8, 287–301.
- Terrer, C., Jackson, R. B., Prentice, I. C., Keenan, T. F., Kaiser, C., Vicca, S., et al. (2019). Nitrogen and phosphorus constrain the CO₂ fertilization of global plant biomass. *Nature Climate Change*, 9(9), 684–689. <https://doi.org/10.1038/s41558-019-0545-2>
- Tsai, W.-P., Feng, D., Pan, M., Beck, H., Lawson, K., Yang, Y., et al. (2021). From calibration to parameter learning: Harnessing the scaling effects of big data in geoscientific modeling. *Nature Communications*, 12(1), 5988. <https://doi.org/10.1038/s41467-021-26107-z>
- Van Oijen, M. (2017). Bayesian methods for quantifying and reducing uncertainty and error in forest models. *Current Forestry Reports*, 3(4), 269–280. <https://doi.org/10.1007/s40725-017-0069-9>

- Van Ravenzwaaij, D., Cassey, P., & Brown, S. D. (2018). A simple introduction to Markov chain Monte–Carlo sampling. *Psychonomic Bulletin and Review*, 25(1), 143–154. <https://doi.org/10.3758/s13423-016-1015-8>
- Vekuri, H., Tuovinen, J.-P., Kulmala, L., Papale, D., Kolari, P., Aurela, M., et al. (2023). A widely-used eddy covariance gap-filling method creates systematic bias in carbon balance estimates. *Scientific Reports*, 13(1), 1720. <https://doi.org/10.1038/s41598-023-28827-2>
- Vihola, M. (2012). Robust adaptive Metropolis algorithm with coerced acceptance rate. *Statistics and Computing*, 22(5), 997–1008. <https://doi.org/10.1007/s11222-011-9269-5>
- Walker, A. P., Beckerman, A. P., Gu, L., Kattge, J., Cernusak, L. A., Domingues, T. F., et al. (2014). The relationship of leaf photosynthetic traits - V_{cmax} and J_{max} —To leaf nitrogen, leaf phosphorus, and specific leaf area: A meta-analysis and modeling study. *Ecology and Evolution*, 4(16), 3218–3235. <https://doi.org/10.1002/ece3.1173>
- Wang, C., Duan, Q., Gong, W., Ye, A., Di, Z., & Miao, C. (2014). An evaluation of adaptive surrogate modeling based optimization with two benchmark problems. *Environmental Modelling and Software*, 60, 167–179. <https://doi.org/10.1016/j.envsoft.2014.05.026>
- Wang, Y., Xie, Z., Jia, B., Wang, L., Li, R., Liu, B., et al. (2020). Sensitivity of snow simulations to different atmospheric forcing data sets in the land surface model CAS-LSM. *Journal of Geophysical Research: Atmospheres*, 125(16), e2019JD032001. <https://doi.org/10.1029/2019JD032001>
- Watson-Parris, D., Williams, A., Deaconu, L., & Stier, P. (2021). Model calibration using ESEm v1.1.0—An open, scalable Earth system emulator. *Geoscientific Model Development*, 14(12), 7659–7672. <https://doi.org/10.5194/gmd-14-7659-2021>
- Wikle, C. K., & Berliner, L. M. (2007). A Bayesian tutorial for data assimilation. *Physica D: Nonlinear Phenomena*, 230(1–2), 1–16. <https://doi.org/10.1016/j.physd.2006.09.017>
- Yang, D., Yang, Y., & Xia, J. (2021). Hydrological cycle and water resources in a changing world: A review. *Geography and Sustainability*, 2(2), 115–122. <https://doi.org/10.1016/j.geosus.2021.05.003>
- Zaehle, S., & Friend, A. D. (2010). Carbon and nitrogen cycle dynamics in the O-CN land surface model: 1. Model description, site-scale evaluation, and sensitivity to parameter estimates. *Global Biogeochemical Cycles*, 24(1). <https://doi.org/10.1029/2009GB003521>