



A machine learning approach to emulation and biophysical parameter estimation with the Community Land Model, version 5

Katherine Dagon¹, Benjamin M. Sanderson^{1,2}, Rosie A. Fisher^{1,2}, and David M. Lawrence¹

¹National Center for Atmospheric Research, Boulder, CO, USA

²CERFACS, Toulouse, France

Correspondence: Katherine Dagon (kdagon@ucar.edu)

Received: 1 March 2020 – Revised: 19 October 2020 – Accepted: 27 October 2020 – Published: 22 December 2020

Abstract. Land models are essential tools for understanding and predicting terrestrial processes and climate–carbon feedbacks in the Earth system, but uncertainties in their future projections are poorly understood. Improvements in physical process realism and the representation of human influence arguably make models more comparable to reality but also increase the degrees of freedom in model configuration, leading to increased parametric uncertainty in projections. In this work we design and implement a machine learning approach to globally calibrate a subset of the parameters of the Community Land Model, version 5 (CLM5) to observations of carbon and water fluxes. We focus on parameters controlling biophysical features such as surface energy balance, hydrology, and carbon uptake. We first use parameter sensitivity simulations and a combination of objective metrics including ranked global mean sensitivity to multiple output variables and non-overlapping spatial pattern responses between parameters to narrow the parameter space and determine a subset of important CLM5 biophysical parameters for further analysis. Using a perturbed parameter ensemble, we then train a series of artificial feed-forward neural networks to emulate CLM5 output given parameter values as input. We use annual mean globally aggregated spatial variability in carbon and water fluxes as our emulation and calibration targets. Validation and out-of-sample tests are used to assess the predictive skill of the networks, and we utilize permutation feature importance and partial dependence methods to better interpret the results. The trained networks are then used to estimate global optimal parameter values with greater computational efficiency than achieved by hand tuning efforts and increased spatial scale relative to previous studies optimizing at a single site. By developing this methodology, our framework can help quantify the contribution of parameter uncertainty to overall uncertainty in land model projections.

1 Introduction

Land models were originally developed to provide lower boundary conditions for atmospheric general circulation models but have evolved to simulate important processes such as carbon cycling, ecosystem dynamics, terrestrial hydrology, and agriculture. Including these societally relevant processes helps provide insight into potential impacts on humans and ecosystems but also introduces additional sources of uncertainty in model predictions. This uncertainty is largely driven by a combination of insufficient observations and incomplete knowledge regarding mathematical represen-

tations of these processes. At the same time, adequately predicting terrestrial processes and climate–carbon feedbacks relies on improving these models and their predictive capabilities while minimizing sources of model error (Bonan and Doney, 2018).

For example, much of the uncertainty in projections of the terrestrial carbon cycle comes from differences in model parameterizations (Friedlingstein et al., 2006; Booth et al., 2012). Looking across 11 carbon-cycle climate models all forced by the same future climate scenario, Friedlingstein et al. (2014) found that not only was the magnitude of the land–atmosphere carbon flux different across models, but

there was also disagreement in the sign, implying that the land could be either a net carbon sink or source by the end of the century. The model spread in predictions of land–atmosphere carbon fluxes was found to be strongly tied to process representation, such as model treatment of carbon dioxide fertilization and the nitrogen cycle. This situation has remained unchanged in the latest iteration of the Climate Model Intercomparison Project (CMIP6) (Arora et al., 2020).

There are several different sources of uncertainty in Earth system models. Variations in initial conditions produce internal variability uncertainty, mimicking internal climate processes including stochastic weather noise. Differences in imposed forcing through future scenarios or boundary conditions produce forcing uncertainty, representing the uncertainty in predicting the pathway of future carbon emissions. Inherent model uncertainty encompasses uncertainty in model structure and parameters and represents the various choices in how models use mathematical abstractions of physics, chemistry, and biology to represent processes within the Earth system. When contrasting land versus ocean carbon cycle uncertainty over time, Lovenduski and Bonan (2017) found that while uncertainty in the forcing dominates the ocean carbon cycle, the land carbon cycle is dominated by uncertainty in model structure and parameters.

Parametric uncertainty in land models has been traditionally explored through experimentation with different parameter values to test how variations impact resulting model predictions (e.g., Bauerle et al., 2014; Fischer et al., 2011; Göhler et al., 2013; Hawkins et al., 2019; Huo et al., 2019; Ricciuto et al., 2018; White et al., 2000; Zaehle et al., 2005). Hand tuning parameter values can be computationally inefficient, requiring many model simulations and large amounts of computer time, especially when the spatial domain is large or global. Single point simulations provide a lower computational cost and the ability to more easily run large ensembles; however, results from parameter estimation at a single site may not be transferable to other regions due to differences in climate, soil, and vegetation types (Post et al., 2017; Cai et al., 2019; Huang et al., 2016; Lu et al., 2018; Ray et al., 2015). Undirected model calibration can be limited in its ability to objectively assess whether the optimal parameter configuration has been found due to the ambiguous and subjective nature of the hand tuning process (Mendoza et al., 2015; Prihodko et al., 2008). A machine learning approach can help streamline this process by providing increased computational efficiency and reduced analysis time as well as objective methods to assess calibration results (Reichstein et al., 2019). Artificial feed-forward neural networks are a type of machine learning algorithm that can learn relationships between input and output data (Russell and Norvig, 1995; Hagan et al., 1996; Mitchell, 1997). By using neural networks to build a surrogate model that emulates the behavior of a land model, additional parameter values can be tested and optimized quickly and efficiently without running the full model,

thus more effectively exploring the parameter space (Sanderson et al., 2008; Knutti et al., 2003).

In this work we seek to better understand land model uncertainty through variations in parameter choices. By exploring the CLM5 biophysical parameter space and determining sensitive parameters, we investigate the role that parameter choices play in overall model uncertainty. We first narrow the parameter space following a series of one-at-a-time sensitivity simulations assessed with objective metrics by ranking global mean sensitivity to multiple output variables and searching for low spatial pattern correlations between parameters. With a candidate list of important CLM5 biophysical parameters, we use model results from a perturbed parameter ensemble (PPE) to train a series of artificial feed-forward neural networks to emulate a subset of the outputs of CLM5. Here we utilize supervised learning, where the neural network is trained on known model parameter values and simulation output from the PPE. The networks are trained to predict annual mean spatial variability in carbon and water fluxes, given biophysical parameter values as input. The trained networks are then applied to globally estimate optimal parameter values with respect to observations, and these optimal values are tested with CLM5 to investigate changes in model predictive skill.

2 Model simulations and parameter selection

2.1 Community Land Model, version 5

For this work we use the latest version of the Community Land Model, CLM5 (Lawrence et al., 2019). We use the “satellite phenology” version of the model where vegetation distributions and leaf and stem area indices (LAI) are prescribed by remote sensing data (Bonan et al., 2002), thus reducing the degrees of freedom by turning off processes that in fully prognostic mode lead to the prediction of LAI. Vegetated land units in CLM are partitioned into up to 15 plant functional types (PFTs) plus bare ground. Some model parameter values (those associated with plant physiology) vary with PFT, an important consideration for this work. We select parameters associated with biogeophysical processes in CLM, such as surface energy fluxes, hydrology, and photosynthesis. Parametric uncertainty of biogeochemical processes in CLM5 (vegetation carbon and nitrogen cycling) under carbon dioxide and nitrogen fertilization scenarios were explored by Fisher et al. (2019), though that study did not include an optimization component.

We run the model globally with a horizontal resolution of 4° in latitude and 5° in longitude. This coarse resolution is intended to maximize computational efficiency and allow for a greater number of ensemble members in both the sensitivity simulations and the perturbed parameter ensemble. CLM is forced by the Global Soil Wetness Project, Phase 3 (GSWP3) meteorological forcing data, repeatedly sampled over the years 2000–2004. These years are cho-

sen to provide a consistent 5-year period of present-day climate forcing for each simulation. GSWP3 is a reanalysis-based product providing 3-hourly precipitation, solar radiation, wind speed, humidity, downward longwave radiation, and bottom atmospheric layer temperature and pressure (<http://hydro.iis.u-tokyo.ac.jp/GSWP3/>, last access: October 2020). Each CLM simulation is run for 20 years, with the last 5 years used in the analysis. The first 15 years are used as spin-up to equilibrate the soil moisture and soil temperature (Fig. S1 in the Supplement). All simulations reach equilibrium after 15 years.

2.2 Parameter sensitivity simulations

We begin with a set of parameter sensitivity simulations to explore the effect of one-at-a-time parameter perturbations for 34 CLM5 biophysical parameters, the results of which will inform the selection of a narrowed list of key parameters to utilize in the emulation and optimization steps. Initial parameter selection of the 34 parameters is based on identifying tunable quantities in the model which are important for the calculation of biophysical processes, including evapotranspiration, photosynthesis, and soil hydrology. These 34 parameters include a mix of empirically derived parameters (10) and parameters that describe biophysical properties (24). Parameter ranges are based on an extensive literature review and expert judgment, utilizing observational evidence whenever possible (Tables S1 and S2 in the Supplement, and references therein). We run two model simulations for each parameter using the setup described above, one simulation using the minimum value of the parameter sensitivity range and one using the maximum value. Ten of the parameters vary with PFT and thus have PFT-specific sensitivity ranges that are applied in these simulations. In most cases, these PFT-varying parameters are uniformly perturbed across PFTs by the same perturbation amount relative to their default value (e.g., $\pm 20\%$). This approach accounts for the fact that the default values for a given parameter may vary across PFTs. For some parameters where the default values do not vary much across PFTs, all PFTs are perturbed to the same minimum or maximum value. Table S2 shows the details of the minimum and maximum values used for each parameter in the one-at-a-time sensitivity simulations, including the perturbation amounts for PFT-varying parameters.

We use seven different model outputs to assess the sensitivity of the 34 parameters. These outputs are gross primary production (GPP), evapotranspiration (ET), transpiration fraction (TF = transpiration / ET), sensible heat flux, soil moisture of the top 10 cm, total column soil moisture, and water table depth. These output fields are selected to span the relevant biogeophysical processes in the model. To narrow the parameter space and identify candidate parameters for further uncertainty quantification, we use multiple parameter selection criteria. First we assess the sensitivity of each parameter to the seven different outputs using a simple sensi-

tivity metric. We call this the parameter effect (PE), detailed in Eq. (1), where $X_{p, \max}$ and $X_{p, \min}$ represent the model output of a quantity X from the simulations using the maximum and minimum values of parameter p , respectively. We calculate the 5-year annual mean PE at each grid point and then take the global mean of the resulting quantity.

$$\text{PE}(p) = |X_{p, \max} - X_{p, \min}| \quad (1)$$

In order to account for differences in units between variables, the global mean PE values for each parameter are ranked across each output variable from highest to lowest, with the average of the individual ranks across all seven outputs used to assess overall sensitivity rank. The PE ranks for each parameter and output variable are shown in Table S3, along with the average and overall ranks. The annual mean PE for GPP for the final six parameters is shown in Fig. 1.

The second selection criterion uses spatial pattern correlations to maximize the sampled range of model physics and responses in the final parameter set. Filtering parameters by spatial pattern correlations helps avoid the situation where all the selected parameters control behavior in a certain region (e.g., the tropics but not at the mid or high latitudes). We compute this metric by calculating the spatial pattern correlation of the PE between all pairwise parameter combinations for each output. This method reveals parameters with similar spatial pattern responses and provides an additional means by which to narrow the parameter space. While one-at-a-time sensitivity studies do not explicitly account for parameter interactions (Rosolem et al., 2012), by calculating the spatial correlation between parameters and selecting those with low correlations, we indirectly account for some of these potential interactions.

Equation (2) details the pattern correlation calculation, where $\text{PC}(p, q)_v$ is the spatial pattern correlation of the PE between parameters p and q for a given output variable v . The pairwise pattern correlations are then summed across the seven output variables ($V = 7$). A visual representation of the sum of pattern correlations across outputs for each parameter combination is shown in Fig. S2. Then the average pattern correlation is calculated across parameters ($Q = 34$), resulting in a single PC value for each parameter p . (This is equivalent to taking the average across rows or columns in Fig. S2.) The average pattern correlations are also ranked across parameters, though this time the ranks are computed from lowest to highest as we are looking for parameters which exhibit low spatial correlation between each other to sample the broadest possible space of model processes. The average pattern correlations and pattern correlation ranks for each parameter are shown in Table S4.

$$\text{PC}(p) = \frac{1}{Q} \sum_{q=1}^Q \sum_{v=1}^V \text{PC}(p, q)_v \quad (2)$$

The pattern correlation ranks are averaged with the PE ranks to help inform the final list of candidate parameters

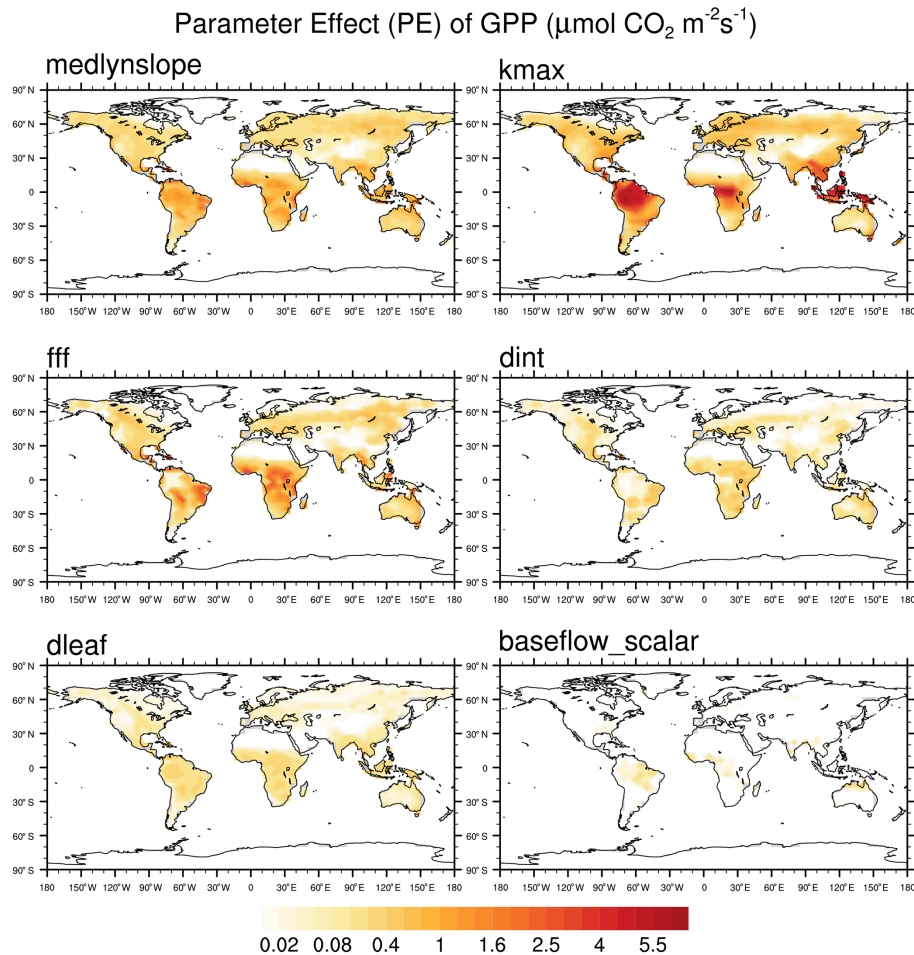


Figure 1. Annual mean parameter effect (PE) for GPP ($\mu\text{mol CO}_2 \text{ m}^{-2} \text{ s}^{-1}$) for the final six parameters.

(Table S5). We aim for a balance between parameters with high PE values across outputs and low spatial pattern correlations with other parameters. In this way we are selecting sensitive parameters in the model that do not all control the same behavior and looking across a range of outputs to inform this selection. We also prioritize parameters for which we are able to use observational data to better inform their uncertainty ranges, in particular for those parameters that vary with PFT. These results are detailed in Sect. 2.3. As a result of these parameter selection criteria, a list of 6 candidate parameters is chosen from the original set of 34. These six parameters, along with a short description of each, are shown in Table 1. A six-dimensional space was found during this study to be the practical limiting size for optimization, as we are motivated to include the most important parameters while also considering computational constraints. We first select the top four parameters by PE rank (medlynslope, kmax, fff, and dint). We then select dleaf because it has a relatively high parameter effect (PE rank = 10) and because we are able to use observations to generate new PFT-specific uncertainty ranges for this parameter. All five of these param-

eters are among the top 50% of parameters when looking by average rank across PE and PC ranks (Table S5). Finally, we select the parameter baseflow_scalar because it has the highest overall average rank, largely due to its low spatial pattern correlation with other parameters (PC rank = 2). Additional information for these parameters, including their default CLM5 values, their uncertainty ranges, and references used to determine their minimum and maximum bounds, can be found in the tables in the Supplement.

2.3 Using observations to inform PFT-specific parameter ranges

For some parameters that vary with PFT, we can improve estimates of plausible ranges by incorporating additional observational data. We can then define uncertainty ranges for a single parameter that vary in their bounds and widths across PFTs. The first parameter we constrain is dleaf, the characteristic dimension of leaves in the direction of wind flow. This parameter is relevant for boundary layer dynamics and relates to the structural form of the equation for leaf boundary layer

Table 1. CLM5 candidate parameters selected based on sensitivity tests.

Parameter name	Description	Varies with PFT?
medlynslope	Slope of stomatal conductance–photosynthesis relationship	Yes
dleaf	Characteristic dimension of leaves in the direction of wind flow	Yes
kmax	Plant segment maximum conductance	Yes
fff	Decay factor for fractional saturated area	No
dint	Fraction of saturated soil for moisture value at which dry surface layer initiates	No
baseflow_scalar	Scalar multiplier for base flow rate	No

resistance. The dleaf parameter has a constant fixed value across PFTs in the default CLM5 configuration but in the real world would likely vary by PFT. We use the TRY plant trait database (Kattge et al., 2011) to identify a dataset with concurrent measurements of leaf width and PFT-relevant information (e.g., leaf type, phenology, growth form). The TRY database included only one dataset with adequate leaf width measurements and enough information to associate the measurements with CLM PFTs. The particular dataset we use is from Northeast China (Prentice et al., 2011), which includes a total of 409 usable measurements of leaf width, each of which we assign to a CLM PFT. Some measurements are applied across multiple CLM PFTs based on the lack of biome variation in the dataset. Because we are utilizing measurements from one geographic location, we may not be adequately capturing biome variation (e.g., tropical versus temperate versus boreal PFTs), and this may impact the results. However, it is likely that this variation is small relative to other types of PFT variation (e.g., phenology) which we are able to capture using this dataset. Using the equations of Campbell and Norman (1998), which relate leaf width to dleaf by a leaf shape-dependent factor, we produce PFT-dependent uncertainty bounds for dleaf (Table S6). These minimum and maximum values are then applied consistently across all PFTs for the minimum and maximum dleaf perturbation simulations, respectively.

The second parameter we constrain using observational data is medlynslope, the slope of the stomatal conductance–photosynthesis relationship defined by Medlyn et al. (2011). This relationship is important for determining stomatal responses to environmental changes and calculating photosynthesis and transpiration. We use data from Lin et al. (2015) to perform a combination of genus- and species-based linear regressions of photosynthesis versus stomatal conductance to obtain slope values (Table S7). The minimum and maximum parameter values are taken from the set of slopes for each PFT. The only PFT we cannot constrain using this method is C3 Arctic grasses, due to a lack of available data. In this case we use a $\pm 10\%$ perturbation from the default medlynslope value for C3 Arctic grasses. As with dleaf, the minimum and maximum values are then applied simultaneously across all PFTs for the minimum and maximum medlynslope perturbation simulations.

2.4 Perturbed parameter ensemble

To generate the parameter values for a perturbed parameter ensemble, we use Latin hypercube (LHC) (Mckay et al., 2000) sampling to generate 100 unique parameter sets for the six CLM5 parameters identified in Table 1. The parameter uncertainty ranges are normalized to uniform linear parameter scalings such that the sampling generates random numbers between 0 (minimum of uncertainty range) and 1 (maximum of uncertainty range). This sampling allows us to include unique uncertainty ranges for each PFT for the PFT-specific parameters without increasing the dimensionality of the problem, albeit while assuming no cross-PFT interactions (Fer et al., 2018). We then run 100 simulations of CLM5 using the LHC-generated parameter sets and the same setup outlined at the beginning of this section. We use the resulting model output, along with the parameter scaling values, to build and train a set of neural networks to emulate the land model.

3 Land model emulation

In order to adequately explore the parameter space of CLM, we seek to emulate the output of the land model using a machine learning-based emulator. This approach will allow us to perform global model calibration in an objective way while reducing computational demands.

3.1 Neural network training and validation

We train feed-forward artificial neural networks to emulate the output of the perturbed parameter ensemble described above (hereafter referred to as the PPE simulations). These networks consist of a series of fully interconnected layers, each of which is made up of a number of nodes, or connected units (Fig. 2). The first layer is the input layer, and here the input values to the networks are the six normalized parameter scaling values (p_i) used in generating the CLM PPE. The last layer is the output layer or the CLM output we would like to predict. Our output layer consists of three outputs (z_m), which will be described in detail below. In between the input and output layers there are a variable number of “hidden layers” which can have different numbers of nodes within them (Mitchell, 1997). Specified activation functions transform the

input values into the values at each node in the hidden layers (n_j^1 and n_k^2) and the output layer (z_m). Weights and biases associated with the nodes and the interconnections are calculated during the training process based on chosen activation functions. An optimization algorithm minimizes the error between actual model output and network predictions and updates the weights and biases using backpropagation (Russell and Norvig, 1995).

The output values for training the neural network are derived from the global 5-year mean maps from the PPE simulations. To reduce the dimensionality of the output variables while preserving information on the spatial patterns observed globally, we calculate empirical orthogonal functions (EOFs) (Lorenz, 1956) and principal components (PCs) (Jolliffe, 2002) of area-weighted GPP and latent heat flux (LHF) anomalies for each ensemble member and use these as the targets for emulation. In this way we can emulate spatial variability of global fields without having to explicitly represent model output at each grid point, which would require a more complex network design. GPP and LHF are chosen based on the availability of globally gridded observations to calibrate the emulator predictions. We use singular value decomposition (SVD) to generate the EOFs and PCs for each variable (Hannachi et al., 2007). The SVD calculation is shown in Eq. (3), where \mathbf{X} represents the area-weighted anomalies where we have reshaped the output into two dimensions (number of ensemble members by number of grid points), \mathbf{U} represents the variability across space (number of ensemble members by number of modes), \mathbf{s} is a diagonal matrix where the diagonal elements are the singular values for each mode, and \mathbf{V} represents the variability across ensemble members (number of modes by number of grid points).

$$\mathbf{X} = \mathbf{U} \times \mathbf{s} \times \mathbf{V} \quad (3)$$

The first three modes of variability are used for each output variable, representing over 95 % of the spatial variance in each case. We choose to truncate at three modes because the higher modes are noisy and we are unable to sufficiently emulate them. The resulting ensemble PC distributions (column vectors of \mathbf{U} , shown in Fig. 3) represent the output data we are training the emulators to predict (z_m in Fig. 2), with separate networks constructed for GPP and LHF. We choose to construct separate neural network emulators for GPP and LHF based on performance and ease of interpretation. To further investigate the relationships between the inputs and outputs of the emulator, we plot the PCs versus parameter scaling values from the PPE for GPP and LHF in Figs. S3 and S4, respectively. Linear and nonlinear relationships are evident in both plots for certain parameters, demonstrating the importance of including different activation functions within the neural network architecture to account for these diverse relationships. The neural network architecture will be discussed in more detail below.

The associated EOF spatial patterns (reshaped row vectors of \mathbf{V}) are shown in Fig. 4. The maps show that the first

mode is primarily a tropical signal for both GPP and LHF, with different spatial patterns evident in the second and third modes. The second mode of GPP and third mode of LHF show similar spatial patterns and likely reflect arid locations. The second mode of LHF is similar to the first mode of LHF and could be a variation on a tropical signal, with some additional mid-latitude influences reminiscent of an east–west dipole pattern. The third mode of GPP is unstructured, noisy, and relatively unimportant as this mode is responsible for only about 1 % of the total GPP spatial variance. Despite using EOFs to preserve some spatial information, there will still be regional biases present in the calibration results due to the outsized influence of the tropics on global carbon and water cycles. In addition, we have spatially masked the low horizontal resolution model output to match the gridded observations, in anticipation of the global calibration procedure. (The observations will be discussed in more detail in Sect. 4.1.) This leads to the absence of EOF signals in certain locations in Fig. 4 (e.g., Sahara, Madagascar, Papua New Guinea). These details will also impact the calibration procedure in terms of which geographic areas the model is tuned to.

We utilize 60 % of the ensemble members as training, 20 % for validation during the training process, and keep the remaining 20 % completely separate for out-of-sample testing. We use the mean squared error (MSE) between the network predictions and the actual model output as our skill metric for training the emulator. Our calculation of MSE is detailed in Eq. (4), where n represents the number of ensemble members, \mathbf{U}_i represents the actual model output for ensemble member i (calculated in Eq. 3), and $\hat{\mathbf{U}}_i$ represents the associated neural network predictions. We calculate the MSE across ensemble members and modes for each subset of the data (training, test, and validation) and compare emulator skill during training with test data and out-of-sample prediction using the validation set. We also consider linear regressions between \mathbf{U} and $\hat{\mathbf{U}}$ for each mode in the validation set and aim to maximize the r^2 values for all three modes.

$$\text{MSE}(\mathbf{U}, \hat{\mathbf{U}}) = \frac{1}{n} \sum_{i=1}^n (\mathbf{U}_i - \hat{\mathbf{U}}_i)^2 \quad (4)$$

During the training process, we iterate on different network configurations to determine the values of some of the important neural network “hyperparameters”, such as number of hidden layers, number of nodes in each layer, activation functions between each layer, optimization algorithm and associated learning rate, batch size, and number of training epochs (Breuel, 2015; Kurth et al., 2018). We find that a network with two hidden layers improves the performance over a single hidden layer (not shown). We also find that a combination of linear and nonlinear activation functions provides the best performance, relative to a single type of activation function. We utilize rectified linear activation for the first hidden layer and hyperbolic tangent for the activation

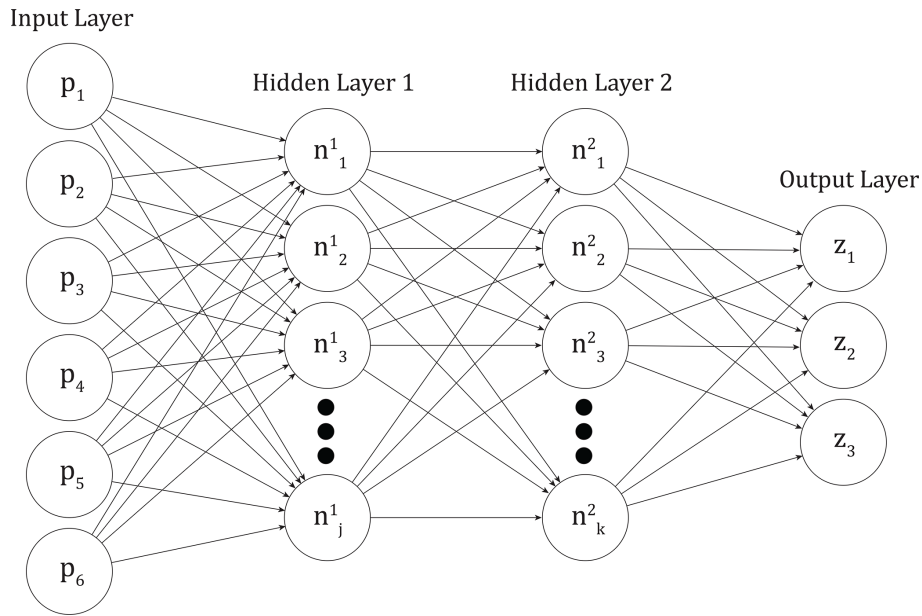


Figure 2. Representative neural network architecture, with input values as land model parameters p_i , three output values z_m , and two hidden layers with j and k nodes, respectively.

into the second hidden layer (which improved over a logistic sigmoidal function) (Krizhevsky et al., 2017; Breuel, 2015). Finally, we use a linear activation to transition from the second hidden layer into the output layer. We use some regularization in the form of the L2 norm to prevent overfitting and improve generalization (Belkin et al., 2019; Bengio, 2012). We find that the optimization algorithm RMSprop (Tieleman et al., 2012), which utilizes an adaptive learning rate, greatly improves the predictive skill of the network relative to optimization using stochastic gradient descent.

Once the basic structure of each network is defined (two hidden layers with linear and nonlinear activation functions), we test initial learning rates for the RMSprop optimizer and find that a value of 0.005 provides the best tradeoff between computational learning efficiency and accuracy of predictions (Bengio, 2012; Smith, 2017). We also test different batch sizes, defined as the number of samples used to estimate the error gradient at each epoch. Here we also utilize early stopping, where the number of epochs is limited based on the stability of the errors, another method commonly applied to combat overfitting (Belkin et al., 2019; Zhang et al., 2017). A batch size of 30 helps minimize errors during training and improves the ability of the emulator to generalize (Keskar et al., 2017; Bengio, 2012).

With the majority of our hyperparameters set, we then iteratively test the performance of the emulator using between 5 and 15 nodes in each layer and select the best performing configurations based on the mean squared error and predictive skill. For the 5–10 best configurations, we randomly resample the training data 100 times to test the stability of these configurations. This resampling also helps avoid overtraining

the network, an important factor considering our small sample size. Our final network configuration is selected based on a combination of high predictive skill and low variability of errors and correlation coefficients. Scatterplots comparing CLM model output with emulator predictions for the first three modes of variance for GPP and LHF are shown in Fig. 5. We further assess emulator performance by analyzing the combined mode predictions to show the error introduced by the EOF approximation and the combined error introduced by the EOF approximation and the emulation. We calculate the root mean squared error (RMSE) across spatial grid points for GPP and LHF relative to FLUXNET-MTE observations of GPP and LHF (Jung et al., 2011). The observations are sampled over the same years as the model simulations (2000–2004), and we calculate the area-weighted anomalies of the observations relative to the PPE ensemble mean. The RMSE is calculated for the original CLM PPE and the reconstructed climatology from the neural network (NN) predictions. We further decompose this error by also calculating the RMSE relative to observations for the reconstructed CLM PPE based on truncating at the first three modes of variability. Figure 6a, b show the error introduced by the EOF approximation, and Fig. 6c, d show the total error from the EOF approximation and the emulation combined. Comparing these top and bottom panels, most of the error comes from the emulation rather than the EOF approximation for both GPP and LHF. However, the emulation performance when viewing the combined mode predictions is still comparable to the performance across individual PC modes as shown in Fig. 5. Figure 6e shows the combined GPP and LHF normalized error performance relative to ob-

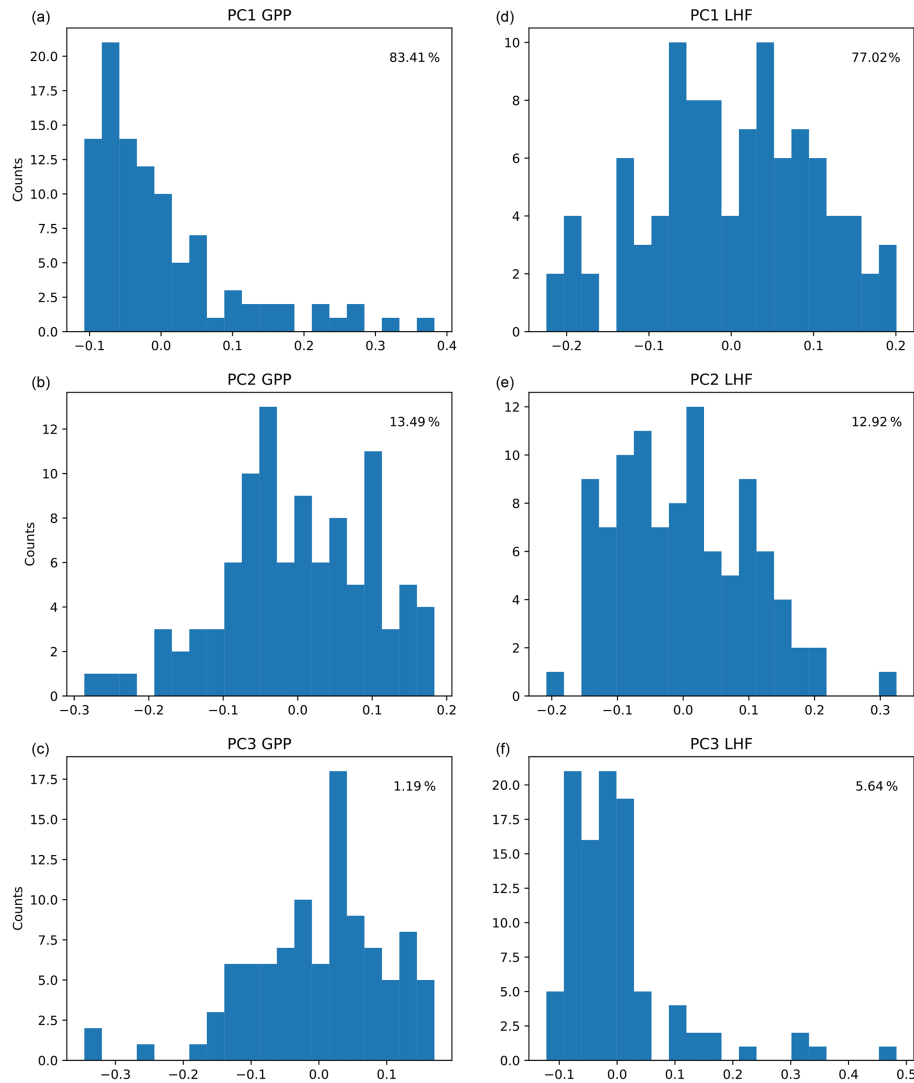


Figure 3. Distributions of the first three principal components (PCs) across CLM PPE ensemble members for GPP (a, b, c) and LHF (d, e, f). The percent variance explained by each mode is shown in each panel in the upper right corner.

servations, comparing the NN predictions with the original CLM PPE. The normalized error is calculated in the form of a weighted cost function, which will be discussed further in Sect. 4.2. We also perform a history matching type experiment (Williamson et al., 2015) in this section to study optimal regions of the parameter space, and for reference the error threshold for this experiment is shown as a vertical magenta line in this panel. All panels of Fig. 6 also show the error resulting from the model with default parameter values for comparison.

To provide another check on the performance of the emulator, we produce a second PPE using 100 different random combinations of parameter values for the same set of six parameters, also generated using Latin hypercube sampling but with a different LHC than the first PPE. We then use the emulator trained on the first PPE to test the predictive skill using

the information from the second PPE. The predictive skill is comparable, especially for the first two modes, which provides more confidence in the trained emulator and helps support the notion that the network is not overtrained (Fig. S5). Following Fig. 6, we also plot the RMSE comparisons and error breakdown for the second CLM PPE relative to observations (Fig. S6). Again, most of the error comes from the emulation rather than the EOF approximation, and the predictive skill is comparable to the first PPE.

3.2 Interpretation of emulator performance and skill

We use multiple interpretation methods to better understand how the neural network emulator makes its predictions and what it has learned (McGovern et al., 2019). The first method we use is called permutation feature importance, which is an

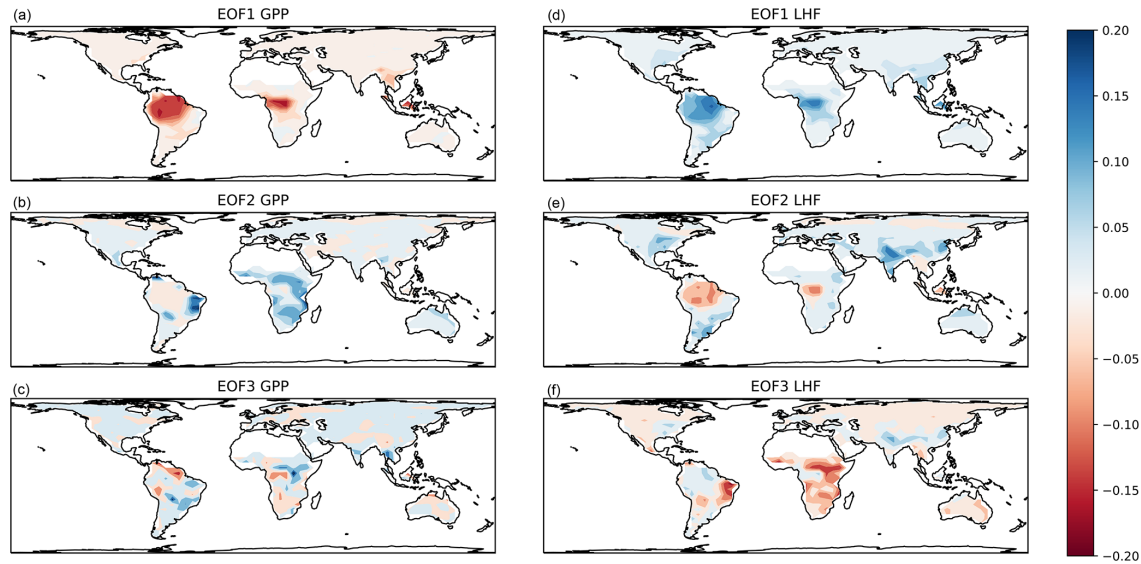


Figure 4. Associated spatial patterns of the first three EOFs for GPP (a, b, c) and LHF (d, e, f).

approach to ranking the importance of various model predictors. Permutation feature importance tests the importance of different inputs (in this case, land model parameters) in the predictive skill of the neural network (Gagne et al., 2019; Molnar, 2019). Feature importance is calculated by randomly shuffling the values of one parameter input while preserving others and testing the resulting performance of the emulator. The goal of this method is to determine the impact on emulator skill when the statistical link between a predictor input and the output is severed. The skill metric we use for these tests is the mean squared error between the predictions and the actual values (Eq. 4). A larger value implies that the parameter is more important to the predictive skill of the emulator, because when the link between this particular parameter and the emulator output is broken, the performance degrades and the prediction error increases. We plot the results of the permutation feature importance tests for PC1 of GPP and LHF as bar charts in Fig. 7, with larger bars reflecting greater prediction error and thus implying important information is stored in that parameter. We also plot the original emulator skill (i.e., the mean squared error without any permutations) to better compare the permutation results relative to the control with no permutations. For parameters with error very close to the original emulator skill, this implies that either these parameters are not very important to the predictive skill of the emulator or that the information in these parameters is present in a different predictor (McGovern et al., 2019). We find the permutation results are different for the first modes of GPP and LHF, where the skill of the GPP emulator is dominated by one parameter in particular, k_{max} , and none of the other parameters is particularly important. However, for LHF, there are several important parameters, including $medlynslope$, k_{max} , and $dint$. These results change

when you look at the higher modes of variability, demonstrating that there are different parameters important to predicting different modes of GPP and LHF (Fig. S7).

The second method we use is partial dependence plots, a technique to visualize the partial dependence of the predictions on an individual input variable (Friedman, 2001). Partial dependence plots further the analysis of permutation feature importance tests by helping to illuminate why a certain parameter is important (McGovern et al., 2019; Molnar, 2019). These plots show where in the uncertainty range a given parameter is most important to the skill of the emulator. To visualize these results, we first generate a set of 10 fixed values for each parameter by evenly sampling its uncertainty range. Taking one parameter and one fixed value at a time, we then replace all the values for that parameter in the original LHC-generated parameter sets with the fixed value. In this way we are removing any skill from that parameter across the entire ensemble. We then generate predictions using the trained emulator where we have artificially fixed one parameter to the same value across all ensemble members. We repeat for each fixed value and each parameter and calculate separately for GPP and LHF. We then average the predictions across the emulator output to average out the effects of the other predictors. We plot the PC1 results for each parameter in Fig. 8, where each line represents the average prediction across emulator output for the 10 fixed values (shown as points on the bold lines). In this way we can see how the predictions vary across the uncertainty ranges for each parameter. Regions of non-zero slope in the partial dependence plots indicate where in the parameter range the emulator is most sensitive. For example, we see that it is the low end of k_{max} values that has the greatest impact on the skill of the emulator for PC1 GPP. For PC1 LHF, the parameter $med-$

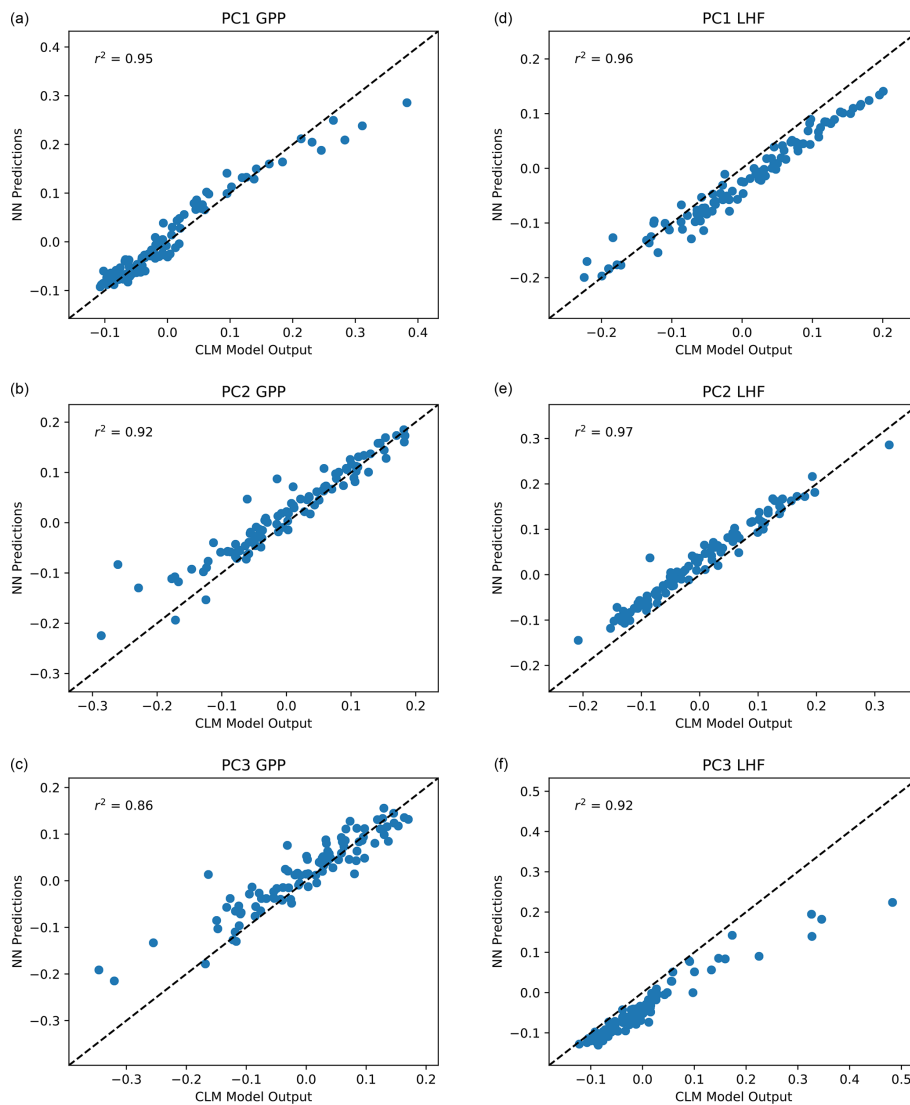


Figure 5. Scatterplot of predicted PC1, PC2, and PC3 for GPP (a, b, c) and LHF (d, e, f) from the neural network emulators versus actual values from CLM output. The one-to-one line is shown in each panel as a dashed black line, and r^2 values from a linear regression fit are included in each panel in the upper left corner.

lynslope is important fairly consistently across its range of values. As with feature importance, these results change for the higher modes and highlight the importance of different parameters (Fig. S8 and S9).

3.3 Ensemble inflation

As an additional test of the emulation performance, we utilize the trained neural networks to artificially inflate the PPE ensemble size. Here we are exploiting the full power of emulation, in that we can quickly predict CLM output given new and unseen parameter values. To test the emulation, we generate 1000 random parameter sets using Latin hypercube sampling to more densely span the uncertainty ranges for the six parameters of interest. These parameter sets are fed into

the trained emulators to produce predictions for the first three modes of variability for GPP and LHF. The emulation process takes seconds, whereas running CLM globally with this setup to test a given parameter set would take several hours per simulation. The distributions of predictions from the inflated ensemble relative to the original PPE for GPP and LHF are shown in Fig. S10. We see that the predictions with the inflated ensemble replicate both the overall range and the characteristic shape of the original CLM PPE distributions (see also Fig. 3 for the original PPE distributions).

4 Parameter estimation

We next apply our trained emulator of CLM to globally optimize parameter values based on best fit to observations. The

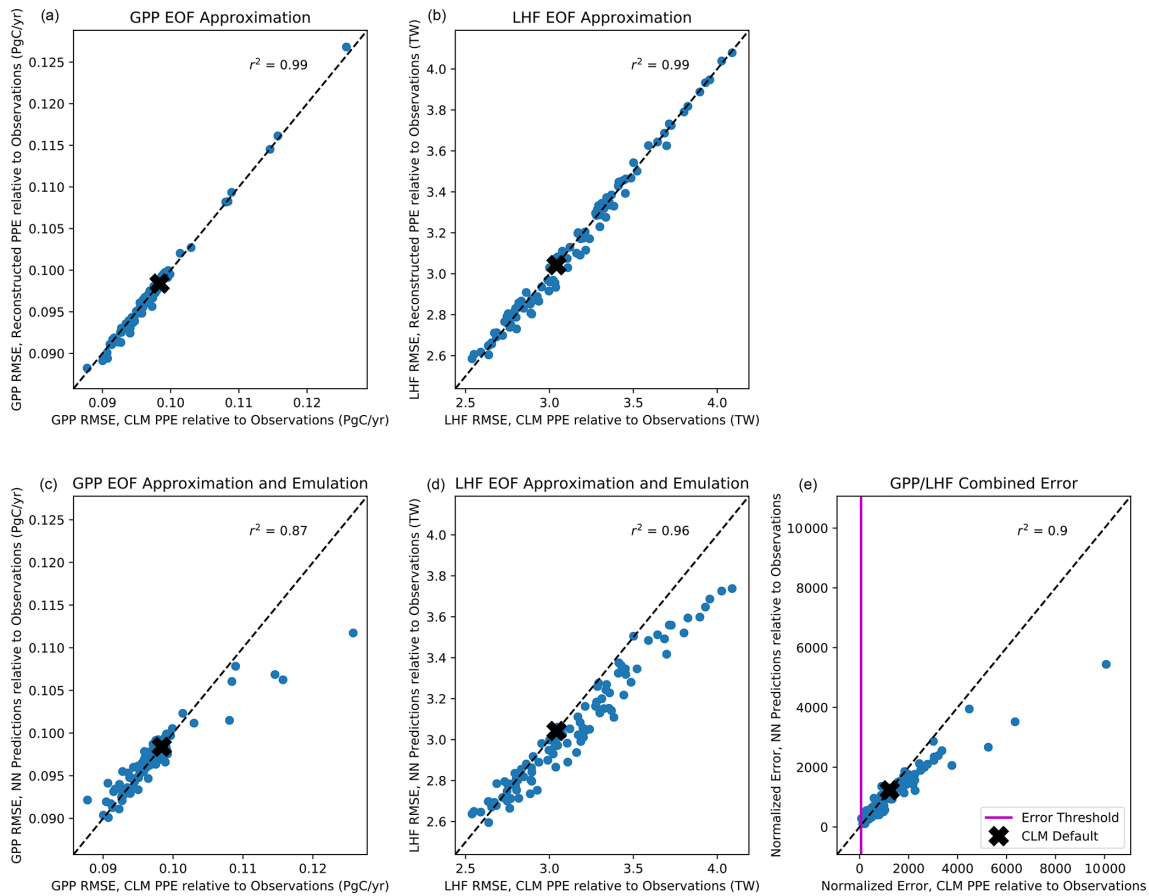


Figure 6. Scatterplot of root mean squared error (RMSE) across spatial grid points for GPP (a, c) and LHF (b, d) from the reconstructed EOF approximation relative to observations (a, b) and the reconstructed neural network (NN) predictions relative to observations (c, d) versus RMSE from CLM PPE relative to observations. Panel (e) shows the combined GPP/LHF normalized error relative to observations for the NN predictions versus the CLM PPE. The error threshold for the history matching experiment described in Sect. 4.2 is shown as a vertical magenta line. The errors resulting from the model with default parameters are shown as a black x in each panel. The one-to-one line is shown in each panel as a dashed black line, and r^2 values from a linear regression fit are included at the top of each panel.

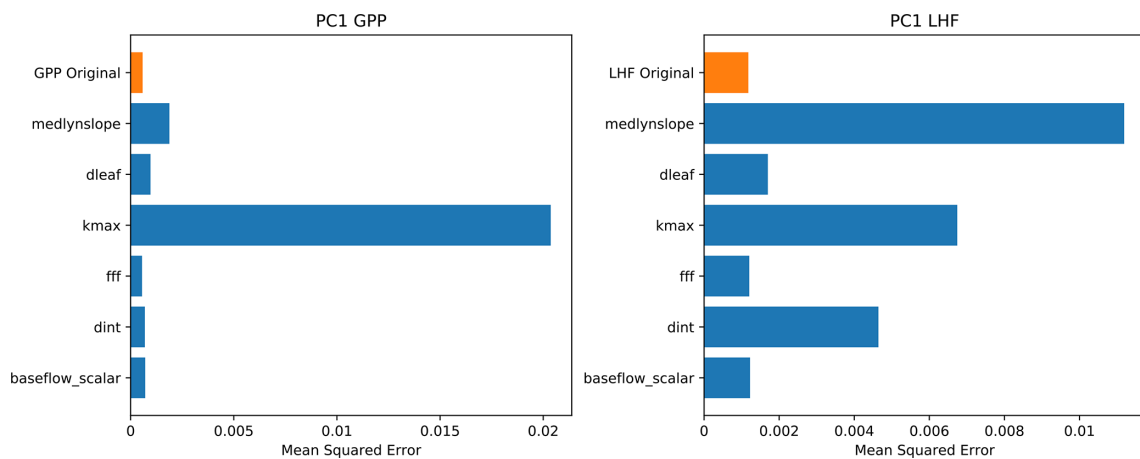


Figure 7. Permutation feature importance test of PC1 for GPP and LHF from the neural network emulators. The top bar (in orange) shows the mean squared error without any permutation (original model skill), and each row (in blue) shows the emulator performance when the information from each parameter is removed, one at a time.

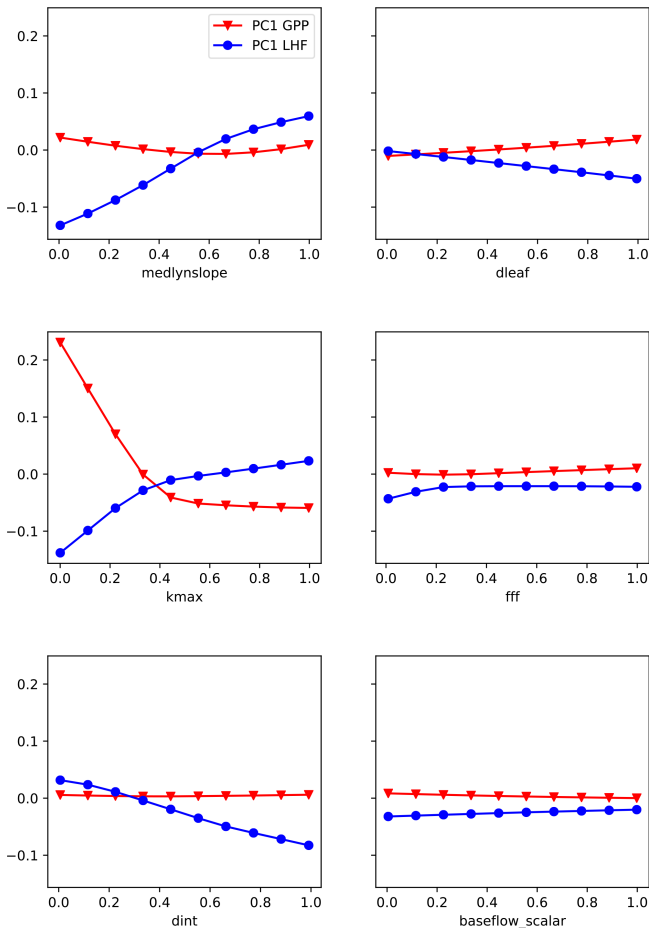


Figure 8. Partial dependence plots of PC1 for GPP (red) and LHF (blue) from the neural network emulators. Circles and triangles denote the fixed values used for predictions with each parameter. The x axis shows the parameter scaling values, and the y axis shows the average prediction of PC1.

neural network emulator can quickly produce predictions of CLM output for a given parameter set, and we aim to optimize the emulation process with the goal of reducing model error. In this section we outline the procedure for producing optimized parameter sets, including details of the observational datasets, construction of the cost function to minimize model error, and optimization results including testing parameter estimates with CLM.

4.1 Observational data

Here we use the FLUXNET-MTE product (Jung et al., 2011) to provide observational targets for our parameter estimation. This dataset includes globally gridded GPP and LHF estimates at 0.5° resolution from 1982 to 2008. We begin by sampling FLUXNET monthly GPP and LHF values from 2000 to 2004 to match the years of the CLM simulations. We regrid the observations to match the CLM output resolution

of 4° latitude by 5° longitude using bilinear interpolation. We then calculate area-weighted anomalies for the observations, where the anomalies are calculated relative to the CLM PPE mean. The observed anomalies are then projected into the same EOF space as the PPE to produce observational estimates consistent with the output metrics used to train the neural network. This calculation is detailed in Eq. (5), where X_{obs} represents the observed anomalies, \mathbf{s} and \mathbf{V} are taken from the SVD calculation for the PPE (Eq. 3), and U_{obs} represents the observational targets projected into the same SVD space. We also calculate, as a reference, the default CLM state without any variations in parameter values, where the default values are calculated in a similar manner to the observations. We can then optimize for parameter values that minimize the error with respect to the observational estimates, looking across output variables and modes of variability.

$$U_{\text{obs}} = X_{\text{obs}} \times (\mathbf{s} \times \mathbf{V})^{-1} \quad (5)$$

4.2 Cost function

With two output variables (GPP and LHF) and three modes of variability for each, we have six targets for parameter calibration. A key question is how to combine them into a single cost function representing model predictive skill relative to observations. Here we calculate the normalized mean squared error of the model predictions relative to observations, with a weighted sum over modes and a separate term for each output variable. The cost function $J(p)$ is detailed in Eq. (6), where $\hat{U}_{v,m}(p)$ represents the predictions from the neural network emulator for output variable v and mode m as a function of parameter values p . $U_{\text{obs},v,m}$ are the observational targets calculated in Eq. (5), and we normalize by the standard deviation across all observational years (1982–2008, represented by $\sigma(U_{\text{obs}^*,v,m})$) in order to account for natural variability using as many years as possible in this dataset. The sum for each variable is weighted by the percent variance explained by each mode ($\lambda_{v,m}$).

$$J(p) = \sum_{v=1}^2 \left[\sum_{m=1}^3 \lambda_{v,m} \left(\frac{\hat{U}_{v,m}(p) - U_{\text{obs},v,m}}{\sigma(U_{\text{obs}^*,v,m})} \right)^2 \right] \quad (6)$$

To explore parameter response surfaces of this cost function, we perform a history matching type experiment (Williamson et al., 2015). We generate an additional large Latin hypercube parameter set with 10 million members and predict the PCs for each member using the trained emulators. We then compute the cost function $J(p)$ following Eq. (6) for each member using the emulated PCs. We subset the results, selecting the 1000 members with the smallest normalized error as computed by the cost function. (For reference, the error threshold for this subset of parameter solutions is shown as a vertical magenta line in Fig. 6e.) We then take each parameter pair and plot the distribution of the parameter scaling values. Fig. 9 shows the resulting parameter space, highlighting the regions where the optimal solution would apply. The

diagonal panels show the distributions of optimal parameter scaling values for each parameter. For certain parameters (notably `dleaf` and `baseflow_scalar`), the range of values is not well constrained by selecting parameter sets with small normalized errors. Other parameters such as `fff` and `dint` favor the edges of their uncertainty bounds and also may not be as well constrained by this exercise. However, for the `med_lyslope` and `kmax` parameters, these plots show where the optimal solutions sit relative to their uncertainty ranges, as well as their relationships with other parameters.

4.3 Nonlinear optimization

We use the SciPy `optimize` function in Python to minimize the cost function $J(p)$ and find optimal parameter values (<https://docs.scipy.org/doc/scipy/reference/optimize.html>, last access: October 2020). There are many different nonlinear optimization and root finding methods available through this package, and we test several of these algorithms to explore their effectiveness at finding optimal parameter values. In particular, a bounded global nonlinear optimization approach using differential evolution produces the best results by quickly and efficiently searching the solution space (Storn and Price, 1997). We use random initial conditions to initialize the algorithm and impose bounds of $[0, 1]$ for each parameter scaling factor, representing the minimum and maximum of the uncertainty ranges.

We also utilize other global methods such as dual annealing (Xiang et al., 1997) and simplicial homology global optimization (SHGO) (Endres et al., 2018) to verify the results. SHGO fails to converge on an optimal solution, while dual annealing produces a similar result to differential evolution but takes several orders of magnitude more iterations and function evaluations. Local optimization methods such as sequential least squares programming (SLSQP), limited-memory bounded Broyden–Fletcher–Goldfarb–Shanno (LBFGS-B), truncated Newton (TNC), and trust-region constrained tend to get stuck at a local minimum and do not sufficiently explore the parameter space. All of the above methods are referenced in the SciPy package.

To test the sensitivity of our cost function formulation, we repeat the optimization process using an alternate cost function that removes the weighting by modes of variability ($\lambda_{v,m}$ in Eq. 6). The resulting parameter values are very similar to the approach using a cost function with mode weighting, demonstrating that this aspect of the cost function is not a significant factor in finding an optimal parameter set in this context.

4.4 CLM test case

Using the optimization results, we translate the parameter scaling factors back to parameter values and run a test simulation with CLM. We use the same model setup as before, spinning up the model for 15 years and using the subsequent

5 years for analysis. The results for the first three modes of GPP and LHF are summarized in Fig. 10. The results for PC1 show that the emulator predictions (blue bars) are very close to the results of the test case (green bars), particularly for PC1 GPP. This feature is less apparent for PC2–3 GPP and PC3 LHF, where the CLM test case is not able to capture the optimized predictions as accurately, possibly due to the noisy nature of the higher modes. Despite this mismatch, these results demonstrate another out-of-sample test of the emulator in predicting CLM output. From the first mode of GPP and LHF we also see that the optimized predictions do indeed match the observations (red bars), suggesting that the emulator has been successfully optimized to match observations for this mode. Furthermore, by using the optimal parameter values in CLM we are able to capture the predictive skill of the emulator and minimize model error. The results for the higher modes vary, but in general the CLM test case with optimal parameters moves the model closer to observations, especially when compared with the default model performance (black bars).

To further investigate the spatial response, we compare the mean bias maps for the optimized CLM test case and the default CLM configuration relative to observations (Fig. 11). We also calculate the global and regional mean biases for GPP and LHF and summarize these in Table 2. We see that the optimized model has some areas of improvement and some areas of degradation when comparing these results with the default model bias. For GPP we are able to decrease the positive biases in the Sahel and other parts of Africa, but in turn the negative bias in the Amazon increases. Positive biases in LHF become negative in the calibrated model, though some areas do show decreases in the bias magnitude. We also calculate the differences for sensible heat flux (FSH), an output variable that we did not calibrate to using the CLM emulator. Encouragingly, we see that areas of negative FSH bias in India and Africa decrease in magnitude with the optimized model, though some areas of positive bias increase. This result demonstrates we are able to achieve some cross-variable benefits to parameter calibration beyond the objectives that were selected for emulation. We also include the mean bias maps for the optimized CLM test case using the alternate cost function formulation that removes weighting by modes of variability (Fig. S11). As stated above, the results are nearly indistinguishable due to the close similarity in the optimal parameter values selected with this alternate cost function.

To highlight the responses in different regions, we also calculate the monthly climatology of GPP and LHF for four different regions representing different biomes (Figs. S12 and S13). The results for the Amazon show that we are unable to capture the phase of the seasonal cycle for GPP with either the default or optimized model, and in fact the seasonal cycle in the optimized model shifts further away from the observations. The seasonality of LHF in the Amazon is better captured, but the overall magnitude in the optimized model is

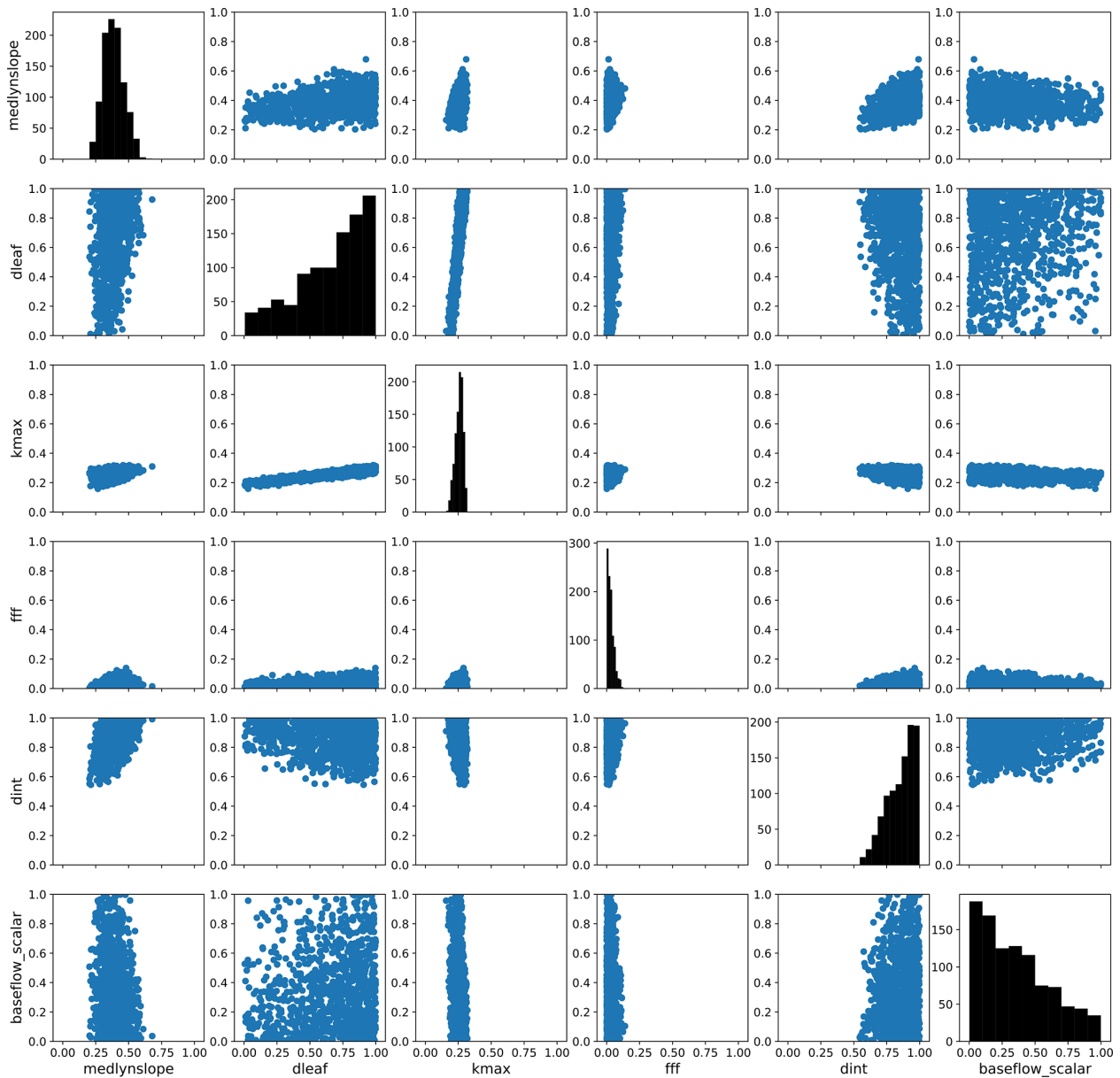


Figure 9. The parameter space of optimal solutions from an additional large emulated ensemble. Here we show a subset of 1000 parameter sets with the smallest predicted normalized error from that ensemble. The axes of these figures show the parameter scaling values, except for the y axes of the diagonals which show the distributions of optimal scaling values for each individual parameter.

too low. The other regions show some marginal improvement with the optimized model relative to the default model, particularly for the spring and fall in the Sahel and the summer in the eastern US for LHF. LHF in Siberia does not change much at all with the optimized model, though summer GPP does marginally improve there. Winter GPP in the Sahel improves, and while summer GPP gets closer to observations, it is still biased high.

5 Discussion and conclusions

In this study we utilize machine learning to explore parameter uncertainty in a global land model (CLM5). We begin by narrowing the parameter space within our domain of interest, using one-at-a-time sensitivity simulations to identify important biogeophysical parameters in CLM. We utilize multiple methods to assess importance, including global mean sensitivity to multiple model outputs and spatial pattern correlations between parameters across outputs. Our goal is to select overall important parameters that exert control on outputs

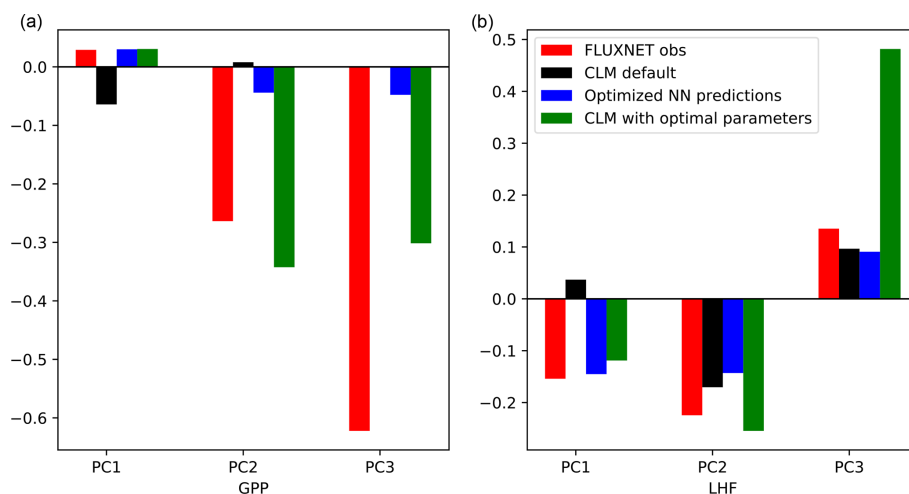


Figure 10. PC1, PC2, and PC3 for GPP (a) and LHF (b) comparing observations, model simulations, and emulator predictions. Observational estimates from FLUXNET are shown as red bars, CLM default values are shown as black bars, the optimized NN predictions are shown as blue bars, and the results of the CLM test case with optimal parameters are shown as green bars.

Table 2. Global and regional mean biases for GPP and LHF comparing CLM with optimized parameters and the CLM default configuration with observations.

Spatial region	Comparison	GPP	LHF
		($\text{gC m}^{-2} \text{d}^{-1}$)	(W m^{-2})
Global mean	Optimized CLM – observations	−0.042	−2.304
	Default CLM – observations	0.359	3.905
Amazon (0–10° S, 70–50° W)	Optimized CLM – observations	−1.425	−7.069
	Default CLM – observations	−0.605	8.164
Sahel (5–15° N, 0–20° E)	Optimized CLM – observations	0.614	−4.228
	Default CLM – observations	1.448	6.691
Eastern US (35–45° N, 100–80° W)	Optimized CLM – observations	−0.264	−3.386
	Default CLM – observations	0.008	1.011
Siberia (60–70° N, 90–110° E)	Optimized CLM – observations	0.132	−2.022
	Default CLM – observations	0.220	−2.251

that cover as much of the planet as possible. By perturbing parameters one at a time we do not include potential parameter interactions, which are a known factor in complex land models (Rosolem et al., 2012). However, by utilizing spatial pattern correlations as a metric for parameter selection, we weight our selection towards parameters without such strong correlations. We also incorporate updated observational data to better constrain uncertainty ranges of sensitive parameters that vary with plant functional type. While it is difficult, in such a process-rich model, to completely avoid relying on expert judgement to identify parameters and their uncertainty ranges, we use this multi-step assessment process in order to be as objective as possible in parameter selection. By identifying important biophysical parameters in CLM5, our approach helps identify which processes are important for accu-

rate land modeling and can help inform observational studies that aim to better quantify these parameters.

With our narrowed set of six parameters, we generate a perturbed parameter ensemble with CLM to span the uncertainty ranges of all parameters and explore parameter interactions. The results of this ensemble show that there are a variety of linear and nonlinear relationships evident between parameters and output fields, highlighting the importance of capturing these diverse relationships. This ensemble is then used to train a set of feed-forward artificial neural networks to act as emulators of CLM. Hyperparameter tuning along with techniques like regularization, early stopping, and resampling of the training data help instill confidence that the neural network is generalizable and not overfitted. While there is still much debate about establishing a procedure for this kind of hyperparameter optimization, we again

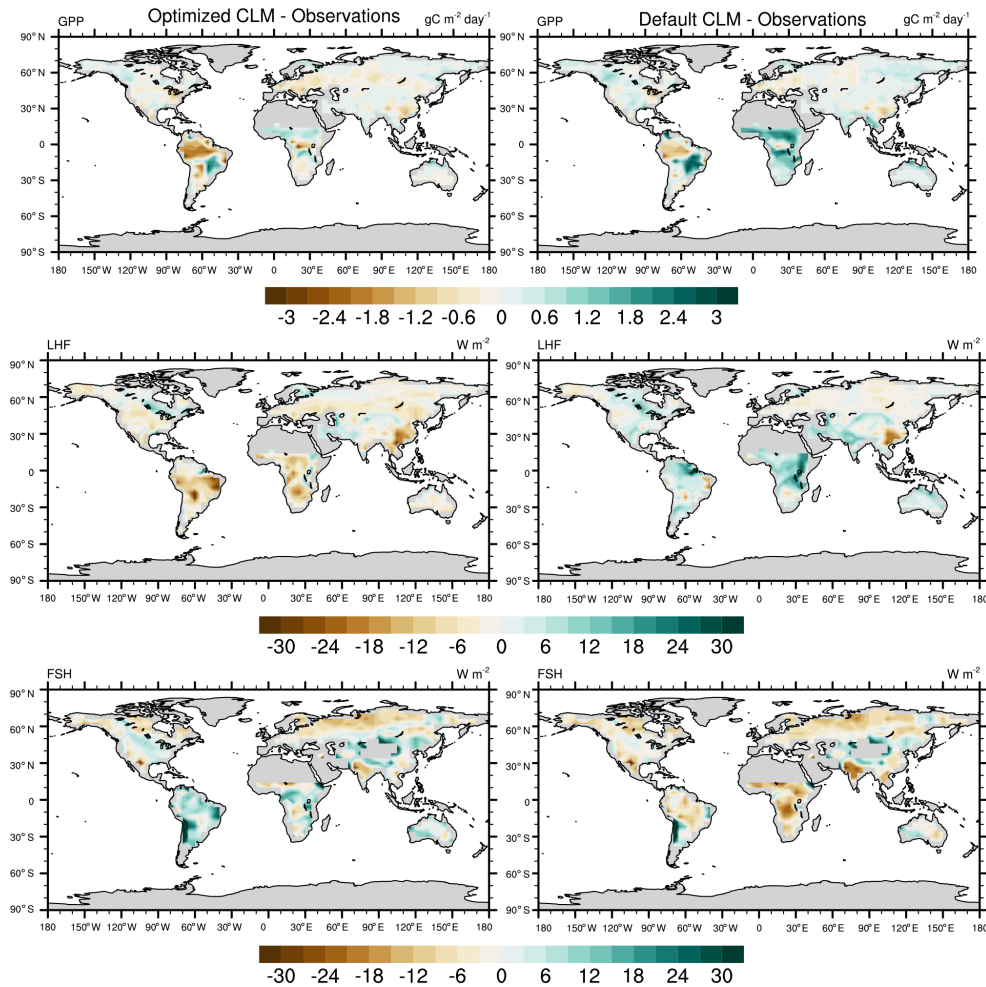


Figure 11. Annual mean bias maps comparing the CLM with optimized parameters (left column) and the CLM default configuration (right column) with observations. The differences are calculated for the calibrated output variables (GPP and LHF), along with sensible heat flux (FSH), an additional model output that we did not calibrate to.

strive to be as objective and transparent as possible in the construction, training, and validation of the neural network emulators. It is possible our results would change if we had utilized a single neural network emulator to simultaneously predict GPP and LHF, and a single network may be better at preserving correlations between output fields. However, we would be less able to clearly identify which parameters are important to the skill of predicting GPP and LHF individually, as we have shown here.

Our final network configurations provide good predictive skill in emulating the first three modes of annual mean spatial variability for GPP and LHF. We show that the EOF approximation of truncation at the first three modes does not introduce much error, and the minimal error that is present comes from the emulation of those modes by the neural networks. We see comparable performance using a second CLM PPE with the trained neural networks, which gives us additional confidence in the predictive skill of the emulator.

While our training dataset is small due to the limited ensemble size, it is encouraging to see that the networks are able to generalize to the second PPE with similar skill. We also explore different interpretation methods such as feature importance and partial dependence plots to better understand the inner workings of the neural networks as emulators for CLM. These methods help illuminate which parameters are most important to the predictive skill of the emulator and why they are important. In particular, we find that the parameter k_{max} , related to plant hydraulic conductivity (Kennedy et al., 2019), is most important in the predictive skill for the first mode of GPP. Furthermore, k_{max} shows a strong nonlinear response in predictive skill where the emulator is most sensitive to values at the low end of its uncertainty range. The parameter $\text{med}_{\text{lynslope}}$, or the slope of the stomatal conductance–photosynthesis relationship (Medlyn et al., 2011), is important for the first mode of LHF, roughly uniformly across its full uncertainty range. The $\text{med}_{\text{lynslope}}$

lynslope parameter is also important for the second and third modes of GPP, again somewhat consistently across its uncertainty range. The *kmax* and *medlynslope* parameters both relate to plant carbon and water use. Due to the strong influence of the tropics on the first modes of GPP and LHF it is perhaps not surprising that these parameters are important for their predictions. The nonlinear response of *kmax* indicates the parameter range may be too wide and could be narrowed by decreasing the maximum values or by developing PFT-specific values. Plant hydraulic stress models are difficult to constrain due to a sparsity of observations, but work is currently ongoing to address these parameter uncertainties (Kennedy et al., 2019). The parameter *dint* is shown to be important for the second mode of LHF and the parameter *fff* for the third mode of LHF. These parameters relate to soil hydrology, and thus they likely influence LHF through soil evaporation and possibly soil water available for transpiration (Swenson and Lawrence, 2014; Niu et al., 2005). The parameter *dint* is important for PC2 LHF across its full uncertainty range, but *fff* also shows some nonlinear behavior for PC3 LHF, where the emulator is most sensitive to low values. Again, this could indicate that the ranges of values for *fff* could be narrowed by decreasing the maximum value, though additional study is needed to better inform the values of these soil hydrology parameters.

We next use the trained emulators to produce globally optimized parameter values for our six parameters based on best fit to observations of GPP and LHF. There are many ways to formulate a cost function to assess model error relative to observational targets (Trudinger et al., 2007). Our approach uses a weighted sum over the first three modes of the squared differences between emulator predictions and observations, normalized by variability in the observations and with separate error terms for GPP and LHF. Our sensitivity test with an alternate cost function shows that the mode weighting is not important to determining the optimization results. We represent observational uncertainty through the normalization term as a substitute for measurement uncertainty or other sources of uncertainty that might arise in the creation of the globally gridded observational products, which are not quantified (Jung et al., 2011; Collier et al., 2018). We do not explicitly include emulator uncertainty in our cost function, or the idea that the emulator is not a perfect representation of CLM (McNeall et al., 2016; Williamson et al., 2015). While we do not expect this uncertainty to be zero, we can begin to quantify it by comparing the emulator predictions for optimized parameters and the results from the model tested with those estimated parameter values. Our test simulation with CLM shows that we can capture the behavior of the optimized predictions for PC1 by using the optimized parameter values in the land model, though the higher modes are not as well captured. This implies there is some uncertainty present in the emulation of PC2 and PC3, perhaps due to the sample size in conjunction with the lower variance associated with these modes. Despite this uncertainty, the CLM test simula-

tion moves the model closer to observations relative to the default for all modes (with the exception of PC3 LHF).

We also explore parameter relationships by visualizing the optimal parameter space for a subset of parameter values where the predicted normalized errors are small. Some parameters are not well constrained by this exercise, and these parameters (*dleaf* and *baseflow_scalar*) are also not shown to be important to the predictive skill of the emulators. Other parameters (*fff* and *dint*) are shown to sample the edges of their uncertainty bounds, indicating that the range of values chosen for these parameters may not be sufficient to capture optimal solutions. This result echoes what was found utilizing the partial dependence plots for the higher modes of GPP and LHF. Two parameters (*medlynslope* and *kmax*) appear to be more constrained by the optimization process and are consistent with parameters found to be important through the permutation feature important tests for different output modes. These parameters will be the focus of further study to better define their range of values and quantify uncertainty. In particular, the relationship between *medlynslope* and *kmax* is well constrained by the optimization and relates to the coupling of plant water stress and stomatal conductance. The new plant hydraulic stress formulation in CLM5 implements this coupling through a water stress factor to better incorporate theory of plant hydraulics (Kennedy et al., 2019). Future process-based modeling could further this approach by directly modeling hydraulic limitation as part of stomatal conductance (Anderegg et al., 2018).

For the optimized model relative to the default configuration, we find that mean spatial biases persist and in some cases worsen. For example, we cannot simultaneously correct mean GPP biases in the Amazon and Congo, similar to results shown in McNeall et al. (2016). This is suggestive of additional sources of uncertainty and structural biases present in the default model configuration (e.g., forcing uncertainty, epistemic uncertainty, or additional parametric uncertainty), which cannot be tuned with the limited set of parameters we have selected for this analysis. It has also been shown previously that it is difficult to find global optimal parameter values which consistently improve skill in a climate model (Jackson et al., 2008; Williamson et al., 2015; Li et al., 2018), and we find a similar outcome when studying biophysical processes with a land model. Cai et al. (2019) calibrated evapotranspiration (ET) across multiple paired FLUXNET sites in the Energy Exascale Earth System Model Land Model and found a reduction in the mean annual bias of ET when the global model was tested with the optimized parameters. This study demonstrates the potential for site-level calibration to improve a global model simulation. They also highlighted the importance of stomatal conductance and soil moisture related parameters, similar to the parameters we found to be most sensitive in CLM5. Li et al. (2018) found a reduction in GPP and LHF error using optimized parameter values in the CABLE and JULES land models; however, this reduction was not consistent across PFTs

nor was it consistent across all land area. Furthermore, the differences between the land models increased when both were calibrated to the same observations. Both studies reiterate the difficulty of globally calibrating a limited set of land model parameters and discuss additional sources of uncertainty contributing to model biases and inter-model differences.

Parameter estimation results will always be sensitive to the choice of cost function and how additional sources of uncertainty are accounted for (Fer et al., 2018; Jackson et al., 2008; Li et al., 2019; Ray et al., 2015). Though we account for multiple CLM outputs in the form of GPP and LHF, we combine them into a single objective cost function whereas a multi-objective approach could help avoid introducing additional biases into the calibration process (Rosolem et al., 2013). In the future we plan to explore different versions of cost functions to account for sources of uncertainty that we do not sample with this study, including structural uncertainty. Structural uncertainty relates to uncertainty in process representation, including how model equations are formulated. While parametric uncertainty can be explored through scrutinizing choices of parameter values, quantifying structural uncertainty is less clearly defined. The United Kingdom probabilistic climate projections (Murphy et al., 2009, 2007) define structural uncertainty using a multi-model ensemble (MME). The reasoning is that the MME represents a variety of structural choices and can represent this source of uncertainty outside of what a single model can achieve, assuming the MME is made up of “state-of-the-art” climate models that have been validated against historical observations (Sexton et al., 2012; Sexton and Murphy, 2012). However, this approach relies on the assumption that the MME is composed of climate models which are all unique representations of the true climate, while there is evidence that state-of-the-art climate models such as those participating in the Coupled Model Intercomparison Project (CMIP) share components and model development processes, which implies that they may also have common limitations (Sanderson and Knutti, 2012; Eyring et al., 2019).

There are a variety of limitations to our model setup. For the purposes of fully exploring the dynamics of the biogeophysics domain of CLM, we use offline CLM simulations without active biogeochemistry, land use change, or vegetation dynamics and thus are limited to a specific set of processes. We also do not account for cross-PFT interactions within parameters which vary across PFTs. While our approach using consistent scaling factors still allows for different uncertainty ranges for PFTs, our calibration results do not reflect possible variations in the sampling of PFT-specific values for a given parameter. This is primarily due to computational challenges in properly sampling PFT-varying parameters as demonstrated in prior work, though future study will consider possible approaches to better account for interactions between different PFTs. Our offline CLM simulations use a specific 5-year period of atmospheric forcing data, and

our results could be dependent on the choice of time period. This choice is somewhat constrained by available observational years, though we could also test additional 5-year or longer periods where we have observations to calibrate the model. The choice of atmospheric forcing data is based on best performance with CLM5 (Lawrence et al., 2019). While our model resolution is low and the simulation length is relatively short, these tradeoffs allow us to perform a global sensitivity analysis and parameter estimation with greater computational efficiency.

Our methodology also uses a smaller subset of output variables (GPP and LHF) for parameter calibration than the larger set of seven biophysical output variables used during the parameter selection stage. For the calibration we are somewhat limited by the availability of globally gridded observational products. However, we expect the results to depend on exactly which output variables are used in the calibration (Keenan et al., 2011). We are able to capture some additional benefits to calibration by reducing regional biases in sensible heat flux, despite not explicitly tuning to this output variable. It is an open question as to how many and which calibration fields should be utilized in a model tuning exercise, along with what metrics should be used to assess model biases and changes in model skill (Mendoza et al., 2015; Fer et al., 2018). Here we use annual mean spatial variability as determined by EOF analysis and thus do not consider seasonal cycle changes or interannual variability. As our monthly climatology results show, we gain some marginal improvement in certain regions and seasons, but we would not expect to see significant improvements due to not calibrating specifically to the seasonal cycle in these locations. In the future we plan to explore additional observational datasets and assessment metrics using the International Land Model Benchmarking (ILAMB) system (Collier et al., 2018).

As model sophistication and process representation increase, it is important to acknowledge there is a tradeoff between model complexity and model error (Saltelli, 2019). Our study provides an example of a mechanistic framework for evaluating parameter uncertainty in a complex global land model. By utilizing machine learning to build a fast emulator of the land model, we can more efficiently optimize parameter values with respect to observations and understand the different sources of uncertainty contributing to predictions in terrestrial processes.

Code and data availability. CLM5 is a publicly released version of the Community Land Model available from <https://github.com/ESCOMP/CTSM> (last access: October 2020; access to the latest version: <https://doi.org/10.5281/zenodo.3779821>, CTSM Development Team, 2020). All simulation output used in the analysis is available through the NCAR Digital Asset Services Hub (DASH) Repository and published under the following DOI: <https://doi.org/10.5065/9bcc-4a87> (Dagon,

2020a). All code used in the analysis is available through <https://doi.org/10.5281/zenodo.4302690> (Dagon, 2020b).

Supplement. The supplement related to this article is available online at: <https://doi.org/10.5194/ascmo-6-223-2020-supplement>.

Author contributions. All the authors contributed to the development of the methodology. KD performed the model simulations and analysis and prepared the manuscript with input from all authors.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. The CESM project is supported primarily by the National Science Foundation (NSF). This material is based upon work supported by the National Center for Atmospheric Research, which is a major facility sponsored by the National Science Foundation under cooperative agreement no. 1852977. Portions of this study were supported by the Regional and Global Model Analysis (RGMA) component of the Earth and Environmental System Modeling Program of the U.S. Department of Energy's Office of Biological and Environmental Research (BER) via National Science Foundation IA 1947282. BMS is supported by the French National Research Agency, ANR-17-MPGA-0016. Computing and data storage resources, including the Cheyenne supercomputer (<https://doi.org/10.5065/D6RX99HX>), were provided by the Computational and Information Systems Laboratory (CISL) at NCAR. We thank all the scientists, software engineers, and administrators who contributed to the development of CESM2. We thank David John Gagne for providing Python code examples for neural network configuration and interpretation and Jeremy Brown for designing the neural network diagram. We thank William Hsieh, Rafael Rosolem, Charles Jackson, and one anonymous reviewer for insightful comments that improved the manuscript.

Financial support. This research has been supported by the National Science Foundation, Division of Atmospheric and Geospace Sciences (cooperative agreement no. 1852977), the French National Research Agency (grant no. ANR-17-MPGA-0016), and the U.S. Department of Energy, Biological and Environmental Research (grant no. IA 1947282).

Review statement. This paper was edited by William Hsieh and reviewed by Charles Jackson, Rafael Rosolem, and one anonymous referee.

References

Anderegg, W. R. L., Wolf, A., Arango-Velez, A., Choat, B., Chmura, D. J., Jansen, S., Kolb, T., Li, S., Meinzer, F. C., Pita, P., Resco de Dios, V., Sperry, J. S., Wolfe, B. T., and Pacala, S.:

Woody plants optimise stomatal behaviour relative to hydraulic risk, *Ecol. Lett.*, 21, 968–977, <https://doi.org/10.1111/ele.12962>, 2018.

- Arora, V. K., Katavouta, A., Williams, R. G., Jones, C. D., Brovkin, V., Friedlingstein, P., Schwinger, J., Bopp, L., Boucher, O., Cadule, P., Chamberlain, M. A., Christian, J. R., Delire, C., Fisher, R. A., Hajima, T., Ilyina, T., Joetzjer, E., Kawamiya, M., Koven, C. D., Krasting, J. P., Law, R. M., Lawrence, D. M., Lenton, A., Lindsay, K., Pongratz, J., Raddatz, T., Séférian, R., Tachiiri, K., Tjiputra, J. F., Wiltshire, A., Wu, T., and Ziehn, T.: Carbon-concentration and carbon-climate feedbacks in CMIP6 models and their comparison to CMIP5 models, *Biogeosciences*, 17, 4173–4222, <https://doi.org/10.5194/bg-17-4173-2020>, 2020.
- Bauerle, W. L., Daniels, A. B., and Barnard, D. M.: Carbon and water flux responses to physiology by environment interactions: a sensitivity analysis of variation in climate on photosynthetic and stomatal parameters, *Clim. Dynam.*, 42, 2539–2554, <https://doi.org/10.1007/s00382-013-1894-6>, 2014.
- Belkin, M., Hsu, D., Ma, S., and Mandal, S.: Reconciling modern machine learning practice and the bias-variance trade-off, *arXiv [preprint]*, [arXiv:1812.11118](https://arxiv.org/abs/1812.11118), 10 September 2019.
- Bengio, Y.: Practical recommendations for gradient-based training of deep architectures, *arXiv [preprint]*, [arXiv:1206.5533](https://arxiv.org/abs/1206.5533), 16 September 2012.
- Bonan, G. B. and Doney, S. C.: Climate, ecosystems, and planetary futures: The challenge to predict life in Earth system models, *Science*, 359, eaam8328, <https://doi.org/10.1126/science.aam8328>, 2018.
- Bonan, G. B., Levis, S., Kergoat, L., and Oleson, K. W.: Landscapes as patches of plant functional types: An integrating concept for climate and ecosystem models, *Glob. Biogeochem. Cy.*, 16, 5–1–5–23, <https://doi.org/10.1029/2000GB001360>, 2002.
- Booth, B. B. B., Jones, C. D., Collins, M., Totterdell, I. J., Cox, P. M., Sitch, S., Huntingford, C., Betts, R. A., Harris, G. R., and Lloyd, J.: High sensitivity of future global warming to land carbon cycle processes, *Environ. Res. Lett.*, 7, 024002, <https://doi.org/10.1088/1748-9326/7/2/024002>, 2012.
- Breuel, T. M.: The Effects of Hyperparameters on SGD Training of Neural Networks, *arXiv [preprint]*, <https://arxiv.org/abs/1508.02788>, 12 August 2015.
- Cai, X., Riley, W. J., Zhu, Q., Tang, J., Zeng, Z., Bisht, G., and Randerson, J. T.: Improving Representation of Deforestation Effects on Evapotranspiration in the E3SM Land Model, *J. Adv. Model. Earth Sy.*, 11, 2412–2427, <https://doi.org/10.1029/2018MS001551>, 2019.
- Campbell, G. S. and Norman, J. M.: *Conductances for Heat and Mass Transfer*, Springer New York, New York, NY, 87–111, https://doi.org/10.1007/978-1-4612-1626-1_7, 1998.
- Collier, N., Hoffman, F. M., Lawrence, D. M., Keppel-Aleks, G., Koven, C. D., Riley, W. J., Mu, M., and Randerson, J. T.: The International Land Model Benchmarking (ILAMB) System: Design, Theory, and Implementation, *J. Adv. Model. Earth Sy.*, 10, 2731–2754, <https://doi.org/10.1029/2018MS001354>, 2018.
- CTSM Development Team: ESCOMP/CTSM: Update documentation for release-clm5.0 branch, and fix issues with no-anthro surface dataset creation, Version release-clm5.0.34, Zenodo, <https://doi.org/10.5281/zenodo.3779821>, 2020.
- Dagon, K.: CLM5 Perturbed Parameter Ensembles, UCAR/NCAR – DASH Repository, <https://doi.org/10.5065/9bcc-4a87>, 2020a.

- Dagon, K.: katiedagon/CLM5_ParameterUncertainty: Publication release, Version v1.0, Zenodo, <https://doi.org/10.5281/zenodo.4302690>, 2020b.
- Endres, S. C., Sandrock, C., and Focke, W. W.: A simplicial homology algorithm for Lipschitz optimisation, *J. Global Optim.*, 72, 181–217, <https://doi.org/10.1007/s10898-018-0645-y>, 2018.
- Eyring, V., Cox, P. M., Flato, G. M., Gleckler, P. J., Abramowitz, G., Caldwell, P., Collins, W. D., Gier, B. K., Hall, A. D., Hoffman, F. M., Hurtt, G. C., Jahn, A., Jones, C. D., Klein, S. A., Krasting, J. P., Kwiatkowski, L., Lorenz, R., Maloney, E., Meehl, G. A., Pendergrass, A. G., Pincus, R., Ruane, A. C., Russell, J. L., Sanderson, B. M., Santer, B. D., Sherwood, S. C., Simpson, I. R., Stouffer, R. J., and Williamson, M. S.: Taking climate model evaluation to the next level, *Nat. Clim. Change*, 9, 102–110, <https://doi.org/10.1038/s41558-018-0355-y>, 2019.
- Fer, I., Kelly, R., Moorcroft, P. R., Richardson, A. D., Cowdery, E. M., and Dietze, M. C.: Linking big models to big data: efficient ecosystem model calibration through Bayesian model emulation, *Biogeosciences*, 15, 5801–5830, <https://doi.org/10.5194/bg-15-5801-2018>, 2018.
- Fischer, E. M., Lawrence, D. M., and Sanderson, B. M.: Quantifying uncertainties in projections of extremes—a perturbed land surface parameter experiment, *Clim. Dynam.*, 37, 1381–1398, <https://doi.org/10.1007/s00382-010-0915-y>, 2011.
- Fisher, R. A., Wieder, W. R., Sanderson, B. M., Koven, C. D., Oleson, K. W., Xu, C., Fisher, J. B., Shi, M., Walker, A. P., and Lawrence, D. M.: Parametric Controls on Vegetation Responses to Biogeochemical Forcing in the CLM5, *J. Adv. Model. Earth Sy.*, 11, 2879–2895, <https://doi.org/10.1029/2019MS001609>, 2019.
- Friedlingstein, P., Cox, P., Betts, R., Bopp, L., von Bloh, W., Brovkin, V., Cadule, P., Doney, S., Eby, M., Fung, I., Bala, G., John, J., Jones, C., Joos, F., Kato, T., Kawamiya, M., Knorr, W., Lindsay, K., Matthews, H. D., Raddatz, T., Rayner, P., Reick, C., Roeckner, E., Schnitzler, K.-G., Schnur, R., Strassmann, K., Weaver, A. J., Yoshikawa, C., and Zeng, N.: Climate–Carbon Cycle Feedback Analysis: Results from the C4MIP Model Intercomparison, *J. Climate*, 19, 3337–3353, <https://doi.org/10.1175/JCLI3800.1>, 2006.
- Friedlingstein, P., Meinshausen, M., Arora, V. K., Jones, C. D., Anav, A., Liddicoat, S. K., and Knutti, R.: Uncertainties in CMIP5 Climate Projections due to Carbon Cycle Feedbacks, *J. Climate*, 27, 511–526, <https://doi.org/10.1175/JCLI-D-12-00579.1>, 2014.
- Friedman, J. H.: Greedy function approximation: A gradient boosting machine, *Ann. Stat.*, 29, 1189–1232, <https://doi.org/10.1214/aos/1013203451>, 2001.
- Gagne II, D. J., Haupt, S. E., Nychka, D. W., and Thompson, G.: Interpretable Deep Learning for Spatial Analysis of Severe Hailstorms, *Mon. Weather Rev.*, 147, 2827–2845, <https://doi.org/10.1175/MWR-D-18-0316.1>, 2019.
- Göhler, M., Mai, J., and Cuntz, M.: Use of eigendecomposition in a parameter sensitivity analysis of the Community Land Model, *J. Geophys. Res.-Biogeo.*, 118, 904–921, <https://doi.org/10.1002/jgrg.20072>, 2013.
- Hagan, M. T., Demuth, H. B., Beale, M. H., and Jesús, O. D.: Neural Network Design, available at: <http://hagan.okstate.edu/nnd.html> (last access: October 2020), 1996.
- Hannachi, A., Jolliffe, I. T., and Stephenson, D. B.: Empirical orthogonal functions and related techniques in atmospheric science: A review, *Int. J. Climatol.*, 27, 1119–1152, <https://doi.org/10.1002/joc.1499>, 2007.
- Hawkins, L. R., Rupp, D. E., McNeall, D. J., Li, S., Betts, R. A., Mote, P. W., Sparrow, S. N., and Wallom, D. C. H.: Parametric Sensitivity of Vegetation Dynamics in the TRIFFID Model and the Associated Uncertainty in Projected Climate Change Impacts on Western U.S. Forests, *J. Adv. Model. Earth Sy.*, 11, 2787–2813, <https://doi.org/10.1029/2018MS001577>, 2019.
- Huang, M., Ray, J., Hou, Z., Ren, H., Liu, Y., and Swiler, L.: On the applicability of surrogate-based Markov chain Monte Carlo-Bayesian inversion to the Community Land Model: Case studies at flux tower sites, *J. Geophys. Res.-Atmos.*, 121, 7548–7563, <https://doi.org/10.1002/2015JD024339>, 2016.
- Huo, X., Gupta, H., Niu, G.-Y., Gong, W., and Duan, Q.: Parameter Sensitivity Analysis for Computationally Intensive Spatially Distributed Dynamical Environmental Systems Models, *J. Adv. Model. Earth Sy.*, 11, 2896–2909, <https://doi.org/10.1029/2018MS001573>, 2019.
- Jackson, C. S., Sen, M. K., Huerta, G., Deng, Y., and Bowman, K. P.: Error Reduction and Convergence in Climate Prediction, *J. Climate*, 21, 6698–6709, <https://doi.org/10.1175/2008JCLI2112.1>, 2008.
- Jolliffe, I. T.: Principal Component Analysis, 2nd edn., Springer, New York, ISBN: 0-387-95442-2, 2002.
- Jung, M., Reichstein, M., Margolis, H. A., Cescatti, A., Richardson, A. D., Arain, M. A., Arneth, A., Bernhofer, C., Bonal, D., Chen, J., Gianelle, D., Gobron, N., Kiely, G., Kutsch, W., Lasslop, G., Law, B. E., Lindroth, A., Merbold, L., Montagnani, L., Moors, E. J., Papale, D., Sottocornola, M., Vaccari, F., and Williams, C.: Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations, *J. Geophys. Res.*, 116, G00J07, <https://doi.org/10.1029/2010JG001566>, 2011.
- Kattge, J., Díaz, S., Lavorel, S., Prentice, I. C., Leadley, P., Bönsch, G., Garnier, E., Westoby, M., Reich, P. B., Wright, I. J., Cornelissen, J. H. C., Violle, C., Harrison, S. P., Van Bodegom, P. M., Reichstein, M., Enquist, B. J., Soudzilovskaia, N. A., Ackerly, D. D., Anand, M., Atkin, O., Bahn, M., Baker, T. R., Baldocchi, D., Bekker, R., Blanco, C., Blonder, B., Bond, W. J., Bradstock, R., Bunker, D. E., Casanoves, F., Cavender-Bares, J., Chambers, J. Q., Chapin III, F. S., Chave, J., Coomes, D., Cornwell, W. K., Craine, J. M., Dobrin, B. H., Duarte, L., Durka, W., Elser, J., Esser, G., Estiarte, M., Fagan, W. F., Fang, J., Fernández-Méndez, F., Fidelis, A., Finegan, B., Flores, O., Ford, H., Frank, D., Freschet, G. T., Fyllas, N. M., Gallagher, R. V., Green, W. A., Gutierrez, A. G., Hickler, T., Higgins, S., Hodgson, J. G., Jalili, A., Jansen, S., Joly, C., Kerkhoff, A. J., Kirkup, D., Kitajima, K., Kleyer, M., Klotz, S., Knops, J. M. H., Kramer, K., Kühn, I., Kurokawa, H., Laughlin, D., Lee, T. D., Leishman, M., Lens, F., Lenz, T., Lewis, S. L., Lloyd, J., J. Llusà, F. L., Ma, S., Mahecha, M. D., Manning, P., Massad, T., Medlyn, B., Messier, J., Moles, A. T., Müller, S. C., Nadrowski, K., Naeem, S., Niinemets, U., Nöllert, S., Nüske, A., Ogaya, R., Oleksyn, J., Onipchenko, V. G., Onoda, Y., nez, J. O., Overbeck, G., Ozinga, W. A., no, S. P., Paula, S., Pausas, J. G., nuelas, J. P., Phillips, O. L., Pillar, V., Poorter, H., Poorter, L., Poschlod, P., Prinzing, A., Proulx, R., Rammig, A., Reinsch, S., Reu, B., Sack, L., Salgado-Negret, B.,

- Sardans, J., Shiodera, S., Shipley, B., Siefert, A., Sosinski, E., Soussana, J.-F., Swaine, E., Swenson, N., Thompson, K., Thornton, P., Waldram, M., Weiher, E., White, M., White, S., Wright, S. J., Yguel, B., Zaehle, S., Zanne, A. E., and Wirth, C.: TRY – a global database of plant traits, *Glob. Change Biol.*, 17, 2905–2935, <https://doi.org/10.1111/j.1365-2486.2011.02451.x>, 2011.
- Keenan, T. F., Carbone, M. S., Reichstein, M., and Richardson, A. D.: The model-data fusion pitfall: assuming certainty in an uncertain world, *Oecologia*, 167, 587, <https://doi.org/10.1007/s00442-011-2106-x>, 2011.
- Kennedy, D., Swenson, S., Oleson, K. W., Lawrence, D. M., Fisher, R., Lola da Costa, A. C., and Gentine, P.: Implementing Plant Hydraulics in the Community Land Model, Version 5, *J. Adv. Model. Earth Sy.*, 11, 485–513, <https://doi.org/10.1029/2018MS001500>, 2019.
- Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P.: On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima, *arXiv [preprint]*, arXiv:1609.04836, 9 February 2017.
- Knutti, R., Stocker, T. F., Joos, F., and Plattner, G.-K.: Probabilistic climate change projections using neural networks, *Clim. Dynam.*, 21, 257–272, <https://doi.org/10.1007/s00382-003-0345-1>, 2003.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E.: ImageNet Classification with Deep Convolutional Neural Networks, *Communications of the ACM*, 60, 84–90, <https://doi.org/10.1145/3065386>, 2017.
- Kurth, T., Treichler, S., Romero, J., Mudigonda, M., Luehr, N., Phillips, E., Mahesh, A., Matheson, M., Deslippe, J., Fatica, M., Prabhat, and Houston, M.: Exascale Deep Learning for Climate Analytics, *arXiv [preprint]*, arXiv:1810.01993, 3 October 2018.
- Lawrence, D. M., Fisher, R. A., Koven, C. D., Oleson, K. W., Swenson, S. C., Bonan, G., Collier, N., Ghimire, B., van Kampenhout, L., Kennedy, D., Kluzek, E., Lawrence, P. J., Li, F., Li, H., Lombardozi, D., Riley, W. J., Sacks, W. J., Shi, M., Vertenstein, M., Wieder, W. R., Xu, C., Ali, A. A., Badger, A. M., Bisht, G., van den Broeke, M., Brunke, M. A., Burns, S. P., Buzan, J., Clark, M., Craig, A., Dahlin, K., Drewniak, B., Fisher, J. B., Flanner, M., Fox, A. M., Gentine, P., Hoffman, F., Keppel-Aleks, G., Knox, R., Kumar, S., Lenaerts, J., Leung, L. R., Lipscomb, W. H., Lu, Y., Pandey, A., Pelletier, J. D., Perket, J., Randerson, J. T., Ricciuto, D. M., Sanderson, B. M., Slater, A., Subin, Z. M., Tang, J., Thomas, R. Q., Val Martin, M., and Zeng, X.: The Community Land Model Version 5: Description of New Features, Benchmarking, and Impact of Forcing Uncertainty, *J. Adv. Model. Earth Sy.*, 11, 4245–4287, <https://doi.org/10.1029/2018MS001583>, 2019.
- Li, J., Duan, Q., Wang, Y.-P., Gong, W., Gan, Y., and Wang, C.: Parameter optimization for carbon and water fluxes in two global land surface models based on surrogate modelling, *Int. J. Climatol.*, 38, e1016–e1031, <https://doi.org/10.1002/joc.5428>, 2018.
- Li, S., Rupp, D. E., Hawkins, L., Mote, P. W., McNeall, D., Sparrow, S. N., Wallom, D. C. H., Betts, R. A., and Wettstein, J. J.: Reducing climate model biases by exploring parameter space with large ensembles of climate model simulations and statistical emulation, *Geosci. Model Dev.*, 12, 3017–3043, <https://doi.org/10.5194/gmd-12-3017-2019>, 2019.
- Lin, Y.-S., Medlyn, B. E., Duursma, R. A., Prentice, I. C., Wang, H., Baig, S., Eamus, D., de Dios, V. R., Mitchell, P., Ellsworth, D. S., de Beeck, M. O., Wallin, G., Uddling, J., Tarvainen, L., Linderson, M.-L., Cernusak, L. A., Nippert, J. B., Ocheltree, T. W., Tissue, D. T., Martin-StPaul, N. K., Rogers, A., Warren, J. M., De Angelis, P., Hikosaka, K., Han, Q., Onoda, Y., Gimeno, T. E., Barton, C. V. M., Bennie, J., Bonal, D., Bosc, A., Löw, M., Macinins-Ng, C., Rey, A., Rowland, L., Setterfield, S. A., Tausz-Posch, S., Zaragoza-Castells, J., Broadmeadow, M. S. J., Drake, J. E., Freeman, M., Ghannoum, O., Hutley, L. B., Kelly, J. W., Kikuzawa, K., Kolari, P., Koyama, K., Limousin, J.-M., Meir, P., Lola da Costa, A. C., Mikkelsen, T. N., Salinas, N., Sun, W., and Wingate, L.: Optimal stomatal behaviour around the world, *Nat. Clim. Change*, 5, 459–464, <https://doi.org/10.1038/nclimate2550>, 2015.
- Lorenz, E. N.: Empirical orthogonal functions and statistical weather prediction, Massachusetts Institute of Technology, Department of Meteorology Cambridge, USA, Scientific Report for the MIT Statistical Forecasting Project, available at: https://eapsweb.mit.edu/sites/default/files/Empirical_Orthogonal_Functions_1956.pdf (last access: October 2020), 1956.
- Lovenduski, N. S. and Bonan, G. B.: Reducing uncertainty in projections of terrestrial carbon uptake, *Environ. Res. Lett.*, 12, 044020, <https://doi.org/10.1088/1748-9326/aa66b8>, 2017.
- Lu, D., Ricciuto, D., Stoyanov, M., and Gu, L.: Calibration of the E3SM Land Model Using Surrogate-Based Global Optimization, *J. Adv. Model. Earth Sy.*, 10, 1337–1356, <https://doi.org/10.1002/2017MS001134>, 2018.
- McGovern, A., Lagerquist, R., John Gagne, D., Jergensen, G. E., Elmore, K. L., Homeyer, C. R., and Smith, T.: Making the Black Box More Transparent: Understanding the Physical Implications of Machine Learning, *B. Am. Meteorol. Soc.*, 100, 2175–2199, <https://doi.org/10.1175/BAMS-D-18-0195.1>, 2019.
- Mckay, M. D., Beckman, R. J., and Conover, W. J.: A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output From a Computer Code, *Technometrics*, 42, 55–61, <https://doi.org/10.1080/00401706.2000.10485979>, 2000.
- McNeall, D., Williams, J., Booth, B., Betts, R., Challenor, P., Wiltshire, A., and Sexton, D.: The impact of structural error on parameter constraint in a climate model, *Earth Syst. Dynam.*, 7, 917–935, <https://doi.org/10.5194/esd-7-917-2016>, 2016.
- Medlyn, B. E., Duursma, R. A., Eamus, D., Ellsworth, D. S., Prentice, I. C., Barton, C. V., Crous, K. Y., De Angelis, P., Freeman, M., and Wingate, L.: Reconciling the optimal and empirical approaches to modelling stomatal conductance, *Glob. Change Biol.*, 17, 2134–2144, <https://doi.org/10.1111/j.1365-2486.2010.02375.x>, 2011.
- Mendoza, P. A., Clark, M. P., Barlage, M., Rajagopalan, B., Samaniego, L., Abramowitz, G., and Gupta, H.: Are we unnecessarily constraining the agility of complex process-based models?, *Water Resour. Res.*, 51, 716–728, <https://doi.org/10.1002/2014WR015820>, 2015.
- Mitchell, T. M.: *Machine Learning*, McGraw Hill, New York, ISBN: 0070428077, 1997.
- Molnar, C.: *Interpretable Machine Learning*, available at: <https://christophm.github.io/interpretable-ml-book/> (last access: October 2020), 2019.
- Murphy, J., Booth, B., Collins, M., Harris, G., Sexton, D., and Webb, M.: A methodology for probabilistic predictions of regional climate change from perturbed physics

- ensembles, *Philos. T. Roy. Soc. A*, 365, 1993–2028, <https://doi.org/10.1098/rsta.2007.2077>, 2007.
- Murphy, J., Sexton, D., Jenkins, G., Boorman, P., Booth, B., Brown, C., Clark, R., Collins, M., Harris, G., Kendon, E., Betts, R., Brown, S., Howard, T. P., Humphrey, K. A., McCarthy, M. P., McDonald, R. E., Stephens, A., Wallace, C., Warren, R., Wilby, R., and Wood, R. A.: UK Climate Projections Science Report: Climate change projections, Tech. rep., Met Office Hadley Centre, Exeter, ISBN: 978-1-906360-02-3, 2009.
- Niu, G.-Y., Yang, Z.-L., Dickinson, R. E., and Gulden, L. E.: A simple TOPMODEL-based runoff parameterization (SIMTOP) for use in global climate models, *J. Geophys. Res.*, 110, D21106, <https://doi.org/10.1029/2005JD006111>, 2005.
- Post, H., Vrugt, J. A., Fox, A., Vereecken, H., and Hendricks Franssen, H.-J.: Estimation of Community Land Model parameters for an improved assessment of net carbon fluxes at European sites, *J. Geophys. Res.-Biogeo.*, 122, 661–689, <https://doi.org/10.1002/2015JG003297>, 2017.
- Prentice, I. C., Meng, T., Wang, H., Harrison, S. P., Ni, J., and Wang, G.: Evidence of a universal scaling relationship for leaf CO₂ drawdown along an aridity gradient, *New Phytol.*, 190, 169–180, <https://doi.org/10.1111/j.1469-8137.2010.03579.x>, 2011.
- Prihodko, L., Denning, A., Hanan, N., Baker, I., and Davis, K.: Sensitivity, uncertainty and time dependence of parameters in a complex land surface model, *Agr. Forest Meteorol.*, 148, 268–287, <https://doi.org/10.1016/j.agrformet.2007.08.006>, 2008.
- Ray, J., Hou, Z., Huang, M., Sargsyan, K., and Swiler, L.: Bayesian Calibration of the Community Land Model Using Surrogates, *SIAM/ASA Journal on Uncertainty Quantification*, 3, 199–233, <https://doi.org/10.1137/140957998>, 2015.
- Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., and Prabhat: Deep learning and process understanding for data-driven Earth system science, *Nature*, 566, 195–204, <https://doi.org/10.1038/s41586-019-0912-1>, 2019.
- Ricciuto, D., Sargsyan, K., and Thornton, P.: The Impact of Parametric Uncertainties on Biogeochemistry in the E3SM Land Model, *J. Adv. Model. Earth Sy.*, 10, 297–319, <https://doi.org/10.1002/2017MS000962>, 2018.
- Rosolem, R., Gupta, H. V., Shuttleworth, W. J., Zeng, X., and de Gonçalves, L. G. G.: A fully multiple-criteria implementation of the Sobol' method for parameter sensitivity analysis, *J. Geophys. Res.*, 117, D07103, <https://doi.org/10.1029/2011JD016355>, 2012.
- Rosolem, R., Gupta, H. V., Shuttleworth, W. J., de Gonçalves, L. G. G., and Zeng, X.: Towards a comprehensive approach to parameter estimation in land surface parameterization schemes, *Hydrol. Process.*, 27, 2075–2097, <https://doi.org/10.1002/hyp.9362>, 2013.
- Russell, S. J. and Norvig, P.: *Artificial Intelligence: A Modern Approach*, chap.: Learning in Neural and Belief Networks, Prentice Hall, New Jersey, ISBN: 0-13-103805-2, 1995.
- Saltelli, A.: A short comment on statistical versus mathematical modelling, *Nat. Commun.*, 10, 3870, <https://doi.org/10.1038/s41467-019-11865-8>, 2019.
- Sanderson, B. M. and Knutti, R.: On the interpretation of constrained climate model ensembles, *Geophys. Res. Lett.*, 39, L16708, <https://doi.org/10.1029/2012GL052665>, 2012.
- Sanderson, B. M., Knutti, R., Aina, T., Christensen, C., Faull, N., Frame, D. J., Ingram, W. J., Piani, C., Stainforth, D. A., Stone, D. A., and Allen, M. R.: Constraints on Model Response to Greenhouse Gas Forcing and the Role of Subgrid-Scale Processes, *J. Climate*, 21, 2384–2400, <https://doi.org/10.1175/2008JCLI1869.1>, 2008.
- Sexton, D. M. H. and Murphy, J. M.: Multivariate probabilistic projections using imperfect climate models. Part II: robustness of methodological choices and consequences for climate sensitivity, *Clim. Dynam.*, 38, 2543–2558, <https://doi.org/10.1007/s00382-011-1209-8>, 2012.
- Sexton, D. M. H., Murphy, J. M., Collins, M., and Webb, M. J.: Multivariate probabilistic projections using imperfect climate models. Part I: outline of methodology, *Clim. Dynam.*, 38, 2513–2542, <https://doi.org/10.1007/s00382-011-1208-9>, 2012.
- Smith, L. N.: Cyclical Learning Rates for Training Neural Networks, arXiv [preprint], arXiv:1506.01186, 4 April 2017.
- Storn, R. and Price, K.: Differential Evolution – A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces, *J. Global Optim.*, 11, 341–359, <https://doi.org/10.1023/A:1008202821328>, 1997.
- Swenson, S. C. and Lawrence, D. M.: Assessing a dry surface layer-based soil resistance parameterization for the Community Land Model using GRACE and FLUXNET-MTE data, *J. Geophys. Res.-Atmos.*, 119, 10299–10312, 2014.
- Tieleman, T., Hinton, G. E., Srivastava, N., and Swersky, K.: Lecture 6e rmsprop: Divide the gradient by a running average of its recent magnitude, COURSERA: Neural Networks for Machine Learning, 26–31, available at: https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf (last access: October 2020), 2012.
- Trudinger, C. M., Raupach, M. R., Rayner, P. J., Kattge, J., Liu, Q., Pak, B., Reichstein, M., Renzullo, L., Richardson, A. D., Roxburgh, S. H., Styles, J., Wang, Y. P., Briggs, P., Barrett, D., and Nikolova, S.: OptIC project: An intercomparison of optimization techniques for parameter estimation in terrestrial biogeochemical models, *J. Geophys. Res.*, 112, G02027, <https://doi.org/10.1029/2006JG000367>, 2007.
- White, M. A., Thornton, P. E., Running, S. W., and Nemani, R. R.: Parameterization and Sensitivity Analysis of the BIOME-BGC Terrestrial Ecosystem Model: Net Primary Production Controls, *Earth Interact.*, 4, 1–85, [https://doi.org/10.1175/1087-3562\(2000\)004<0003:PASAOT>2.0.CO;2](https://doi.org/10.1175/1087-3562(2000)004<0003:PASAOT>2.0.CO;2), 2000.
- Williamson, D., Blaker, A. T., Hampton, C., and Salter, J.: Identifying and removing structural biases in climate models with history matching, *Clim. Dynam.*, 45, 1299–1324, <https://doi.org/10.1007/s00382-014-2378-z>, 2015.
- Xiang, Y., Sun, D., Fan, W., and Gong, X.: Generalized simulated annealing algorithm and its application to the Thomson model, *Phys. Lett. A*, 233, 216–220, [https://doi.org/10.1016/S0375-9601\(97\)00474-X](https://doi.org/10.1016/S0375-9601(97)00474-X), 1997.
- Zachle, S., Sitch, S., Smith, B., and Hatterman, F.: Effects of parameter uncertainties on the modeling of terrestrial biosphere dynamics, *Global Biogeochem. Cy.*, 19, GB3020, <https://doi.org/10.1029/2004GB002395>, 2005.
- Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O.: Understanding deep learning requires rethinking generalization, arXiv [preprint], arXiv:1611.03530, 26 February 2017.