https://doi.org/10.5194/egusphere-2026-4009
Preprint. Discussion started: 24 August 2026

Enhancing Parameter Calibration in Land Surface Models Using a Multi-Task Surrogate Model within a Differentiable Parameter Learning Framework

Wenpeng Xie[1], Hongmei Li[2], Kei Yoshimura[1,3]

[1]Institute of Industrial Sciences, The University of Tokyo, Kashiwa, Japan.
[2]School of Civil and Environmental Engineering, Cornell University, USA
[3]Earth Observation Research Center, Japan Aerospace Exploration Agency, Tsukuba, Japan.

Correspondence to: Wenpeng Xie (wenpeng@iis.u-tokyo.ac.jp)

Abstract. Land surface models (LSMs) are essential for simulating terrestrial processes and their interactions with the atmosphere, but parameter calibration remains challenging because of complex process coupling, parameter uncertainty, and limited observational constraints. In this study, we introduce multi-task differentiable parameter learning (MdPL), a deep learning framework that combines a multi-task neural surrogate with a differentiable parameter generator for efficient and diagnostically transparent LSM parameter calibration. The framework was evaluated at 20 PLUMBER2 sites spanning four plant functional types (PFTs). Because MdPL optimizes multiple parameters simultaneously through a neural surrogate, we first assessed surrogate–ILS consistency under joint parameter perturbations. Eighteen of the 20 sites satisfied the consistency criterion, defined as median Pearson's R > 0.9 for both sensible and latent heat fluxes, and were retained for the main calibration analysis. Across these consistency-screened sites, the MdPL-calibrated Integrated Land Simulator (ILS) reduced RMSE by 14.1% and 13.9% at the PFT-mean level for sensible and latent heat flux simulations, respectively, relative to the default parameter set. Benchmarking with the PLUMBER2 dataset showed that the MdPL-calibrated ILS achieved competitive performance relative to standard LSMs, including CLM5, JULES, Noah, and GFDL, and performance comparable to an LSTM benchmark, although the relative advantage varied across variables, sites, and temporal resolutions. A comparison with a traditional calibration baseline based on Sobol sequence sampling at four representative sites further showed that MdPL identified competitive parameter solutions without exhaustive full-model evaluation, but its advantage remained site- and metric-dependent. An out-of-target evaluation using gross primary productivity (GPP) suggested that calibration based on sensible and latent heat fluxes did not lead to systematic degradation of a non-target carbon-related flux. Parameter plausibility analysis showed that some calibrated solutions approached predefined parameter bounds, indicating limited parameter identifiability under flux-only observational constraints. Therefore, the calibrated parameters should be interpreted as effective parameters under the given model structure and observational constraints, rather than as uniquely identifiable physical quantities. Overall, MdPL provides a scalable and diagnostically transparent approach for improving parameter calibration in complex LSMs.

https://doi.org/10.5194/egusphere-2026-4009
Preprint. Discussion started: 24 August 2026

1 Introduction

To address the unprecedented challenges posed by climate change, extreme weather, and biodiversity loss, it is critical to accurately forecast terrestrial ecosystem dynamics under these global change scenarios (Cardinale et al., 2012; Raoult et al., 2024; Rivera et al., 2017). Land surface models (LSMs) are key components of Earth System Models that provide mechanistic simulations of land-atmosphere interactions (Blyth et al., 2021). They also capture biogeochemical cycles, water fluxes, and energy exchanges. Thus, they offer valuable insights into land management, water management, and agricultural production at the global scale for ecosystems (Nitta et al., 2020). However, ecosystem complexity, driven by spatial heterogeneity and nonlinear process couplings, introduce significant uncertainties in LSM parameterization (Fisher & Koven, 2020; Li et al., 2024).

Further, LSM parameters often lack direct observational constraints (Famiglietti et al., 2021), exhibit multiscale dependencies, and interact nonlinearly within coupled systems. For example, soil hydraulic conductivity and minimum stomatal conductance are often estimated empirically or using sparse field measurements, which introduce biases and inconsistencies (Buotte et al., 2021; Exbrayat et al., 2014; Oberpriller et al., 2022). Moreover, simplifying model processes, such as radiation transfer or turbulent exchange schemes for wind speed, can introduce systematic biases, which can further complicate parameter calibration (Sawada, 2020).

For a long time, manual parameter tuning was the only strategy for calibrating LSM parameters, and owing to the limited availability of computational resources and observational data, module parameters were primarily derived from existing knowledge (Blyth et al., 2021). However, over the past two decades, advances in computing power have enhanced parameter data assimilation (PDA), which combines observational data with model states to update parameters iteratively (Medvigy et al., 2009; Rayner et al., 2005). PDA also synergistically fuses observations and model forecasts within a probabilistic framework and thus, simultaneously quantifies data and models formulation uncertainties (Rayner et al., 2019). This approach has yielded substantial progress in the calibration of parameters related to crop, carbon, and hydrological cycles (Kaminski et al., 2002; Keenan et al., 2013; Peylin et al., 2016; Weng et al., 2011). However, it is associated with several limitations, particularly in the context of LSMs, including high computational cost, sensitivity to nonlinear processes, and dependence on dense observational data (Bacour et al., 2023; Raiho et al., 2021; Schürmann et al., 2016). Its stability and accuracy in converging optimal parameter sets are also affected by nonlinear processes, such as soil moisture feedback and energy exchange (Huang et al., 2018; Massoud et al., 2019). Moreover, its effectiveness is limited by its dependence on dense and high-quality observations, meanwhile existing observations are often sparse and noisy across space and time (Cameron et al., 2022; MacBean et al., 2016).

Machine learning (ML) offers novel possibilities for addressing PDA-associated limitations. Notably, ML methods can be employed to accurately capture complex and nonlinear relationships and optimize high-dimensional parameter spaces, thereby offering new avenues for augmenting PDA workflow (Chen et al., 2022; Kolassa et al., 2017; Kwon et al., 2019; Rodríguez-Fernández et al., 2019). To date however, most ML efforts have served as adjuncts within PDA frameworks,

https://doi.org/10.5194/egusphere-2026-4009
Preprint. Discussion started: 24 August 2026

rather than being the primary learning objective responsible for parameter estimation (McNeall et al., 2024). Tsai et al. (2021) introduced differentiable parameter learning (dPL), an end-to-end deep learning scheme that directly embeds parameter estimation within model training to enhance calibration efficiency. Subsequent studies have further explored differentiable modelling frameworks for hydrological systems (e.g., Bao et al., 2023). For efficient parameter inference, dPL, e.g., the variable infiltration capacity (VIC) model, leverages neural-network differentiability to couple process-based simulators (Liang et al., 1994) with gradient-based training. In this end-to-end workflow, the neural network, g_z (·) uses forcings, historical site data (forcings + flux observations), and static land surface properties to estimate the parameters of the VIC model. The predicted parameters were used to drive a neural surrogate of the VIC and the resulting outputs were then compared with observation data, enabling backpropagation-based calibration. Thus, dPL is highly advantageous across different datasets, such as soil moisture and runoff. Similarly, Feng et al. (2023) applied dPL to calibrate the parameters of the Hydrologiska Byråns Vattenbalansavdelning hydrological model (Tibangayuka et al., 2022). Their results showed notable improvements in the accuracy of soil and groundwater storage, snow accumulation, evapotranspiration, and baseflow simulations across different basins in the United States. Therefore, relative to PDA models, dPL is advantageous in that it offers the possibility to transform the parameter calibration problem into a multi-model problem and uses deep learning frameworks to optimize parameters efficiently via backpropagation and gradient descent algorithms (Shen et al., 2023). Consequently, the parameters for every site are learned in one pass, minimizing manual tuning and improving model transferability. Second, by exploiting model differentiability, dPL sidesteps costly repeated simulator runs and reduces calibration time by several orders of magnitude relative to the cost and time requirements associated with PDA (Tsai et al., 2021). dPL's gradient-based approach also adeptly learns nonlinear couplings, which are critical for processes such as snow-soil interactions in hydrology (Feng et al., 2022).

However, applying dPL in LSMs is associated with several limitations. First, surrogate fidelity sets the ceiling on parameter accuracy, and any mismatch between the neural surrogate and the true LSM increases calibration errors. Therefore, surrogate-based calibration requires explicit evaluation of whether the surrogate can reproduce the original LSM response in the parameter space relevant to calibration. This is particularly important when the surrogate is pretrained using structured parameter perturbations, whereas the subsequent calibration procedure updates multiple parameters simultaneously. Second, LSMs, unlike standalone hydrological simulators, integrate energy fluxes, carbon cycling, and vegetation physiology, resulting in a significant increase in emulation difficulty (MATSIRO6 Document Writing Team, 2021). This multidomain coupling and strong nonlinearity complicate surrogate architecture and training. Third, simultaneously calibrating multiple fluxes and states in LSMs requires the careful mapping of parameters to each target variable. Further, different parameters exhibit varying levels of sensitivity and impact on each output. The effects of coupling and interactions on different processes also amplify uncertainties in parameter calibration (Hou et al., 2015). Finally, even when calibrated parameters improve target flux simulations, they may not be uniquely identifiable from flux observations alone. This identifiability issue is particularly important for flux-only calibration, because multiple parameter combinations can produce similar sensible and latent heat fluxes through compensatory adjustments. Therefore, the key challenge associated with applying dPL in LSMs is

https://doi.org/10.5194/egusphere-2026-4009
Preprint. Discussion started: 24 August 2026

the design of surrogate models, management of model complexity, and diagnostic evaluation of surrogate reliability and parameter plausibility.

To address the multi-process and multi-output complexity of LSMs with respect to dPL, in this study, we propose a multi-task differentiable parameter learning (MdPL) framework as a unified calibration framework for application in LSMs. The MdPL framework comprises two modules, namely, a multi-task surrogate and a differentiable parameter generator, (g_z(\cdot)). Based on multitask learning (Ruder, 2017), the multi-task surrogate uses shared bases to capture common dynamics, and thereafter employs separate task-specific branches to fine-tune the parameters of each target flux. We also integrated gate layers that filter and route parameter signals and automatically emphasize those relevant to each output, while suppressing irrelevant ones, thereby reducing overfitting. Further, the (g_z(\cdot)) parameter generator was employed to generate parameters that drive the multi-task surrogate, and its predictions were compared against observed fluxes for end-to-end optimization. Before calibration, we designed the framework to incorporate pretraining using data synthesized by perturbing the key parameters, ensuring that it learned their influence on flux outputs. Because the MdPL optimization updates multiple parameters simultaneously, we further introduced a surrogate–ILS consistency screening under joint parameter perturbations before interpreting MdPL-derived parameter estimates. We also assessed whether calibrated parameters approached their predefined bounds, as such boundary-proximal solutions may indicate limited parameter identifiability under flux-only constraints.

Compared with single-task dPL, our multitask design is intended to reduce cross-output interference and handle integrated process complexity, providing a more efficient approach for calibrating parameters in complex land surface models. Further, to validate the performance of the proposed framework, we used it to calibrate the Integrated Land Simulator (ILS) (Nitta et al., 2020) across 20 sites from the PLUMBER2 dataset (PALS Land Surface Model Benchmarking Evaluation Project Phase 2), representing four plant functional types. Because the reliability of MdPL depends on the surrogate approximation, the main calibration analysis was conducted for sites that satisfied the surrogate–ILS consistency criterion. We thereafter benchmarked the MdPL-calibrated ILS against four leading LSMs, Community Land Model version 5 (CLM5; NCAR, 2020), Joint UK Land Environment Simulator (JULES; McNeall et al., 2024), Noah Land Surface Model (Noah; Ek et al., 2003), and Geophysical Fluid Dynamics Laboratory Land Model (GFDL; Shevliakova et al., 2024), and a Long Short-Term Memory (LSTM) model using half-hourly sensible and latent heat observations. We further compared MdPL with a traditional calibration baseline based on Sobol-based joint parameter sampling at four representative sites, and conducted an out-of-target evaluation using gross primary productivity (GPP) to assess whether calibration on energy fluxes affects non-target carbon-related variables. Additional transferability analysis using leave-one-out cross-validation is provided in the Supplementary Material.

Thus, we present MdPL as an efficient calibration framework that addresses process complexity, multi-output coupling, and parameter uncertainty, which limit LSM accuracy. Further, using the multitask learning-inspired surrogate model, we aim to capture nonlinear and coupled processes, reduce computational costs through gradient-based optimization, and potentially improve robustness under sparse and noisy observational conditions. At the same time, we emphasize that surrogate-based

https://doi.org/10.5194/egusphere-2026-4009
Preprint. Discussion started: 24 August 2026

calibration should be accompanied by diagnostic checks of surrogate consistency and parameter plausibility before calibrated parameters are physically interpreted. Together, these improvements provide a scalable and diagnostically transparent strategy for LSM parameter calibration, supporting ecosystem simulations under climate change. The rest of the manuscript is structured as follows. Section 2 provides details regarding the MdPL framework, evaluation metrics, data preprocessing, surrogate consistency screening, parameter plausibility assessment, and the experiments performed. Section 3 provides results and discussion, and Section 4 provides a conclusion of the study, summarizing the key findings of the study and their implications.

2 Materials and Methods

2.1 Framework outline

The MdPL framework is illustrated in **Fig. 1(a)**, from which it is evident that the calibration of LSM parameters was performed in two main steps, the first being to train the multi-task surrogate model for the LSM, as shown in **Fig. 1(a1)**, ensuring that the surrogate model replicated the LSM outputs as closely as possible by minimizing the following loss function:

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(\hat{f}(X_i, P_i, Encoding(DOY); \theta), f(X_i, P_i)) \tag{1}$$

In Eq. (1), X_i represents the meteorological forcing of the i-th sample, P_i represents the calibration parameters of the i-th sample, $Encoding(DOY)$ represents the time positional encoding of the day of the year, $f(X_i, P_i)$ denotes the LSM output, $\hat{f}(X_i, P_i; \theta)$ denotes the predicted outputs of the surrogate model, with θ referring to the trainable parameters (e.g., weights and biases) of the neural network. Further, $\mathcal{L}$ represents the loss function, typically expressed as the mean squared error (MSE), and N represents the total number of training samples.

The surrogate model, which was trained using the same inputs as those in the LSM, i.e., meteorological forcing and calibration parameters, with LSM outputs as the target data, learned the LSM input-output mapping pattern by minimizing the error between predictions and targets. Further, surrogate models established based on differentiable function, were employed to approximate LSM behavior and support automatic differentiation. Thus, they played critical roles in gradient-based optimization in MdPL.

The second step involved learning the optimal LSM parameters using a parameter generation model, g_z, as shown in **Fig. 1(a2)**. Specifically, g_z is a deep learning framework with meteorological forcing data, X and auxiliary information, A(archived site-level data and land surface properties) as inputs. Here, A represents static site-level attributes, including geographic information (e.g., longitude and latitude), surface properties (e.g., albedo, slope, and elevation variability), hydrological characteristics (e.g., lake depth and lake fraction), and land surface classification variables (e.g., soil type and surface type fractions). The parameters of g_z are optimized by minimizing the following objective function:

https://doi.org/10.5194/egusphere-2026-4009
Preprint. Discussion started: 24 August 2026

$$\min_{\emptyset} \frac{1}{N} \sum_{j=1}^{N} \mathcal{L}\left(\hat{f}\left(X_j, g_z\left(X_j, A_j; \emptyset\right), Encoding(DOY); \bar{\theta}\right), Y_{obs,j}\right) \quad (2)$$

In Eq. (2), A_j represents archived site-level data and land surface properties for the j-th sample, $g_z(X_j, A_j; \emptyset)$ denotes calibration parameters generated by the parameter generation model, $\emptyset$ represents the trainable parameters in the parameter generation model, $\bar{\theta}$ represents the frozen parameters of the surrogate model, and $Y_{obs,j}$ represents the observed data for the j-th sample. This learning process leverages the parameter generation model, g_z to map the observational data and dynamic forcing inputs to the LSM parameter space, thereby enabling automated parameter calibration. Since both the surrogate model and the parameter generator g_z are implemented as differentiable neural networks, gradients can be propagated from the loss function to the generated parameters and further back to the weights of g_z through backpropagation, enabling end-to-end optimization within the MdPL framework.

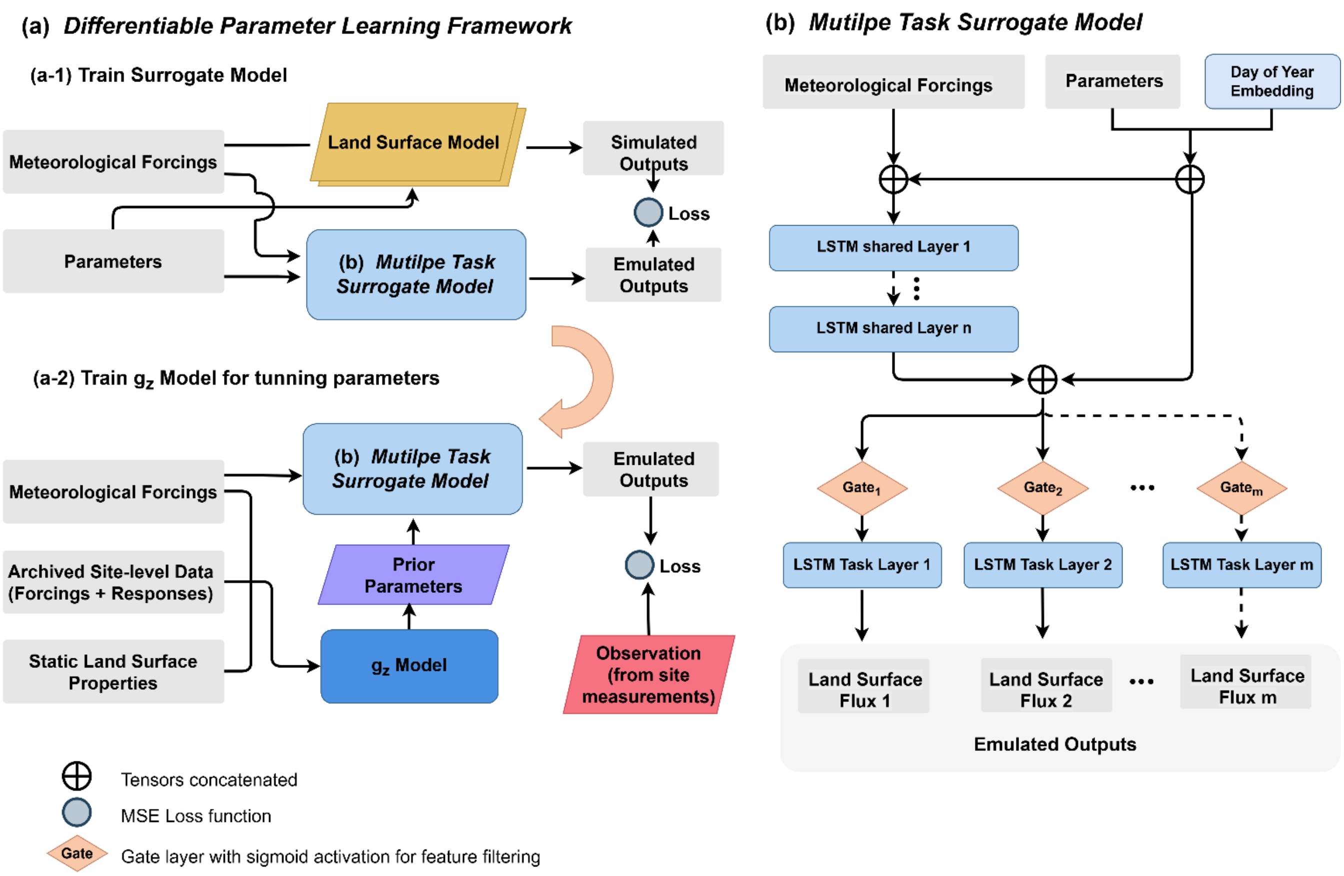


Figure 1. Overview of the multi-task differentiable parameter learning (MdPL) framework for land surface models (LSMs). The surrogate model emulated LSM outputs based on meteorological forcings and parameters. A secondary model, g_z inferred site-specific parameter priors using static land surface properties and historical site data (forcings + flux observations). Loss was computed by comparing surrogate predictions and observed fluxes.

https://doi.org/10.5194/egusphere-2026-4009
Preprint. Discussion started: 24 August 2026

2.2 Multi-task surrogate model

Details regarding the multi-task surrogate model are shown in **Fig. 1(b)**. Unlike conventional single-task dPL surrogates that rely on a single shared predictor after concatenating forcing data and calibration parameters, the proposed multi-task surrogate augments the combined inputs with temporal positional encoding and routes the shared representations through task-specific gated branches. This architecture is intended to better distinguish shared temporal dynamics from task-specific parameter effects. Differences in scale and variation frequency between static and dynamic features can otherwise interfere with the extraction of critical parameter information and limit surrogate performance (Leontjeva & Kuzovkin, 2016; Han et al., 2022).

To capture seasonal variability in LSM responses, we introduced time positional encoding (TPE) based on the day of year (DOY) (Foumani et al., 2024). This encoding method transformed temporal information into high-dimensional feature representations, and thus, enhanced the ability of the model to perceive seasonal variability. The TPE was computed according to Eqs. (3) and (4)

$$DOY_{rad} = DOY \cdot \frac{2\pi}{365} \tag{3}$$

$$Encoding(DOY) = [sin(f_l \cdot DOY_{rad}), cos(f_l \cdot DOY_{rad})]_{l=0}^{L-1} \tag{4}$$

where $f_l = 2^l$ represents the frequency of each encoding dimension, with $l \in \{0,1,\dots,L-1\}$, where L is the number of encoding frequencies. TPE also captured multiscale temporal periodicity and provided a robust representation of seasonal effects on model responses.

In this study, to extract global features from dynamic inputs, we employed a LSTM network as a shared layer (Hochreiter & Schmidhuber, 1997). The output features of the LSTM, denoted as H_{shared}, were expressed as shown in Eq. (5).

$$H_{\text{shared}} = \text{LSTM}_{\text{shared}}(X_{\text{combined}}) \tag{5}$$

The dynamic forcing data, X_{forcing}, time positional encoding, $Encoding(DOY)$, and calibration parameters, P were concatenated to form the combined feature representation, denoted X_{combined} and expressed according to Eq. (6).

$$X_{\text{combined}} = [X_{\text{forcing}}, \text{Encoding(DOY)}, P] \tag{6}$$

Further, to address task competition in multitask learning and reduce the influence of parameters that are irrelevant to specific outputs, a gate layer was introduced before the task-specific layers. This gate layer dynamically adjusted the flow of features based on the input as defined in Eqs. (7), (8), and (9).

$$g_i = \sigma\left(W_{\text{gate},i} \cdot X_{\text{combined}} + b_{\text{gate},i}\right) \tag{7}$$

$$X_{\text{gated},i} = g_i \odot H_{\text{shared}} \tag{8}$$

$$Y_{\text{pre},i} = LSTM_{\text{Task},i}(X_{\text{gated},i}) \tag{9}$$

where $W_{\text{gate},i}$ and $b_{\text{gate},i}$ represent the weights and biases of the gate layer for the i-th task, respectively, σ denotes the sigmoid activation function, and $\odot$ represents element-wise multiplication. The gated features, X_{gated} were then fed into

https://doi.org/10.5194/egusphere-2026-4009
Preprint. Discussion started: 24 August 2026

each task-specific layer to learn the features relevant to different output objectives. $Y_{\mathrm{pre},i}$ represents the output of model task $LSTM_{\mathrm{Task},i}$ for the i-th task.

The LSTM shared layer extracted global features from the time series and thus, provided a common representation for multitask learning. The task-specific layers, combined with the gate layer, effectively mitigated competition among tasks and enhanced the ability of the model to fit individual task objectives. Further, by dynamically controlling feature flow, the gate layer limited the impact of irrelevant information on the training process, thereby improving the stability and generalizability of the model.

2.3 Data and ILS

2.3.1 Data sources

The PLUMBER2 dataset (PALS Land Surface Model Benchmarking Evaluation Project), a public framework for the standardized inter-comparison of LSMs and data-driven models, was used to simulate energy, water, and carbon fluxes across global ecosystems (Abramowitz et al., 2024). The dataset comprises: (1) meteorological forcings, (2) site attributes, (3) flux observations, and (4) model outputs spanning LSMs and data-driven methods (e.g., linear regression, ML, Deep Learning). Thus, it provides a comprehensive foundation for benchmarking.

To ensure geographic and ecosystem diversity, we selected 20 globally distributed sites representing four PFTs: broadleaf evergreen forests (hereafter evergreen forests), mixed coniferous and broadleaf deciduous forests and woodlands (hereafter woodland), wooded and grassland sites (hereafter grassland), and cultivation sites. Details of these PFTs are provided in Table 1. Available half-hourly flux measurements spanning approximately 1–3 years, depending on the site, were used. Site attributes included soil texture, vegetation classification, and geographic coordinates. To unify the influences of various variables on the neural network, the forcing and outputs were standardized using means and STDs.

Table 1. Summary of the characteristics of the observation sites for different plant functional types.

Plant function type	Site name	Location (latitude, longitude)	Observation period	Climate	Country and region
Broadleaf evergreen forest	CN-Din	23.17, 112.54	2003 - 2005	Humid Subtropical	China
	ID-Pag	-2.32, 113.90	2002 - 2003	Tropical rainforest	Indonesia
	PT-Esp	38.64, -8.60	2002 - 2004	Subtropical	Portugal
	PT-Mi1	38.54, -8.00	2005 - 2005	Subtropical	Portugal
Mixed	AR-SLu	-33.46, -66.46	2010 -2010	Subtropical	Argentina

coniferous & broadleaf deciduous forest & woodland	CN-Cha	42.40, 128.10	2003 - 2005	Temperate	China
	DE-Meh	51.28, 10.66	2004 - 2006	Temperate	Germany
	JP-SMF	35.26, 137.08	2003 - 2006	Subtropical	Japan
	US-Bar	44.06, -71.29	2005 - 2005	Temperate	USA
Wooded & grassland	AU-Emr	-23.86, 148.47	2012 - 2013		Australia
	CN-Dan	30.85, 91.08	2004 - 2005	Arctic	China
	CN-Du2	42.05, 116.28	2007 - 2008	Temperate	China
	DK-Lva	55.68, 12.08	2005 - 2006	Temperate	Denmark
	IE-Dri	51.99, -8.75	2003 - 2005	Temperate	Ireland
	PL-wet	52.76, 16.31	2004 - 2005	Temperate	Poland
Cultivation land	DE-Seh	50.87, 6.45	2008 - 2010	Temperate	Germany
	DK-Fou	56.48, 9.59	2005 - 2005	Temperate	Denmark
	IE-Ca1	52.86, -6.92	2004 - 2006	Temperate	Ireland
	IT-BCi	40.52, 14.96	2005 - 2010	Subtropical	Italy
	IT-CA2	42.38, 12.03	2012 - 2013	Subtropical	Italy

Note: The data shown are sourced from FLUXNET2015 (FLUXNET 2015 dataset for micrometeorological measurements), LaThuile (FLUXNET LaThuile 2007 synthesis dataset), and OzFlux (Australian and New Zealand flux research and monitoring networks) (Pastorello et al., 2020). Site names follow the FLUXNET/ICOS convention, using two-letter country codes and three-letter site abbreviations (e.g., CN-Din = Dinghushan, China). The climate classifications are based on the Köppen climate classification system: humid subtropical, Cfa/Cwa; tropical rainforest, Af; mediterranean climate, Csa/Csb; temperate, Cfb/Cfa; and arctic, ET. Coordinates are presented as (latitudes and longitudes) in decimal degrees.

Although the full PLUMBER2 archive contains a larger number of sites, we selected 20 sites to balance ecosystem diversity, data completeness, and computational feasibility within the differentiable calibration framework. The selected subset spans four representative PFT categories and provides continuous observations suitable for both surrogate training and parameter evaluation. Therefore, this study is intended primarily as a controlled multi-site proof-of-concept evaluation of the MdPL framework, rather than an exhaustive benchmark across the full PLUMBER2 dataset.

2.3.2 ILS

The land surface model used in this study was an ILS specifically designed to simulate the interactions between the terrestrial surface and the atmosphere (Nitta et al., 2020). This ILS offered the possibility to evaluate the performance of the calibrated parameters and was used to conduct sensitivity analysis experiments.

https://doi.org/10.5194/egusphere-2026-4009
Preprint. Discussion started: 24 August 2026

2.4 Parameter perturbation and surrogate pretraining

We performed one-factor-at-a-time sensitivity analysis to identify the model parameters with the most significant effects on heat fluxes. The candidate PFT-related parameters considered in this study were selected based on previously reported parameter ranges (Poulter et al., 2011, 2015; Li et al., 2025). Each parameter was sampled at four uniformly spaced levels across its plausible range (as listed in **Table 2**), resulting in four simulations per parameter. With all other parameters held constant, each parameter was individually perturbed to drive 1-year ILS simulations of sensible and latent heat fluxes. This one-factor-at-a-time design focuses on first-order sensitivity and was used only to rank the relative influence of individual parameters on the target heat fluxes.

Further, we computed the RMSE between default and perturbed flux simulations to rank parameter sensitivities. The sensitivity-ranking results are provided in **Fig. S1**. Via this analysis, we identified the 12 parameters with the most significant impact on the outputs selected for calibration.

Table 2. Key plant functional type parameters and their definitions, default values, and ranges.

Parameters	Explanation	Default Value (Evergreen Forest)	Default Value (Woodland)	Default Value (Grassland)	Default Value (Cultivation)	Unit	Range
vegh	Vegetation height	35	20	1	1	m	[0.5, 40]
rlfv	Leaf albedo (visible)	0.1	0.07	0.11	0.11	-	[0.05, 0.3]
rlfn	Leaf albedo (near-infrared)	0.45	0.4	0.58	0.58	-	[0.3, 0.7]
tlfn	Leaf transmittance (near-infrared)	0.25	0.15	0.25	0.25	-	[0.1, 0.5]
vgcov	Vegetation coverage	1	1	1	1	-	[0.1, 1]
cdl	Leaf exchange coefficient (vapor)	0.11	0.111	9.82d-2	0.098	s m-2	[0.01, 0.5]
chl	Leaf exchange coefficient (thermal)	0.0274	0.0277	2.46d-2	0.0246	s m-2	[0.01, 0.1]
vmax0	Maximum Rubisco capacity	6.00E-05	6.00E-05	6.00E-05	6.00E-05	mol m-2 s-1	[1E-6, 1E-4]
gradm	Stomatal conductance slope	9	7.5	4	9	-	[1, 20]
binter	Minimum stomatal conductance	0.01	0.01	0.04	0.01	mol m-2 s-1	[0.001, 0.1]
effcon	Intrinsic quantum efficiency	0.08	0.08	0.05	0.08	mol mol-1	[0.01, 0.2]
psicr	Plant critical water potential	-200	-200	-200	-200	kPa	[-500, -50]

https://doi.org/10.5194/egusphere-2026-4009
Preprint. Discussion started: 24 August 2026

The parameter psicr represents a plant-related water stress threshold associated with each PFT. It controls the sensitivity of vegetation water uptake to soil moisture conditions. In the ILS model, soil water potential is compared with psicr to determine the degree of water stress.

A pre-training dataset was generated using the same perturbation strategy applied in the sensitivity analysis to enable the surrogate model to better capture parameter information. Specifically, for each site, the 12 calibration parameters listed in Table 2 were individually perturbed at four uniformly spaced levels within their predefined ranges, resulting in 48 simulations per site (12 parameters × 4 levels). This design provides structured coverage of first-order parameter sensitivities. The simulation results from this perturbed dataset were then concatenated with those obtained using the default parameter set to form the final pre-training dataset.

Joint parameter perturbations were not used for parameter ranking or surrogate pretraining in this section; instead, they were used later as a diagnostic test of surrogate reliability under simultaneous parameter variations, as described in Section 2.5.

MdPL training was performed using the PyTorch framework on an NVIDIA A100 GPU. The LSM simulations were supported by an Intel Xeon Gold 6126 CPU utilizing six nodes for parallel processing.

2.5 Surrogate reliability and parameter plausibility diagnostics

We first evaluated whether the trained surrogate could reliably reproduce the ILS response under joint parameter perturbations. This diagnostic was designed to assess surrogate reliability in the parameter space relevant to calibration, where multiple parameters may vary simultaneously rather than one at a time.

For each site, we generated 100 joint parameter perturbation samples by varying the selected calibration parameters within their predefined plausible ranges. These parameter sets were then used to drive both the original ILS and the trained surrogate model under identical meteorological forcing conditions. Surrogate reliability was evaluated by comparing surrogate-predicted and ILS-simulated sensible and latent heat fluxes. For each joint parameter perturbation sample, Pearson's correlation coefficient (R) was calculated across the time series between the surrogate prediction and the corresponding ILS simulation. This yielded 100 R values for each site and target variable. The site-wise distributions of surrogate error to perfect correlation, defined as (1 - R), are summarized in **Fig. 2**. A site was considered to satisfy the surrogate–ILS consistency criterion only when the median R across the 100 joint parameter perturbation samples exceeded 0.9 for both sensible and latent heat fluxes. Sites that did not satisfy this criterion were excluded from the main MdPL calibration analysis to avoid interpreting parameter estimates derived from insufficiently reliable surrogate behavior.

In addition to surrogate reliability, we assessed the plausibility of the calibrated parameters. Because flux-only calibration may lead to equifinal parameter combinations, the calibrated parameters were not interpreted as uniquely identifiable physical quantities. Instead, we diagnosed whether the optimized parameters remained within the predefined physically

https://doi.org/10.5194/egusphere-2026-4009
Preprint. Discussion started: 24 August 2026

plausible ranges and whether they systematically approached the lower or upper bounds of these ranges. For each calibrated parameter, the optimized value was normalized by its prescribed range as

$$p^*_{norm} = \frac{p^* - p_{min}}{p_{max} - p_{min}} \tag{10}$$

where p^* is the calibrated parameter value, and p_{min} and p_{max} denote the lower and upper bounds of the corresponding parameter range. Values close to 0 or 1 indicate boundary-proximal solutions. We therefore identified parameters as boundary-proximal when $p^*_{norm} \leq 0.05$ or $p^*_{norm} \geq 0.95$. The frequency of boundary-proximal solutions was then summarized across sites and PFTs.

This diagnostic was used to evaluate whether MdPL calibration produced broadly plausible effective parameter sets or whether improvements in heat flux simulations were associated with parameters being pushed toward artificial range limits. Boundary-proximal solutions were interpreted as a potential indication of limited parameter identifiability or compensatory parameter adjustment under the given model structure and observational constraints.

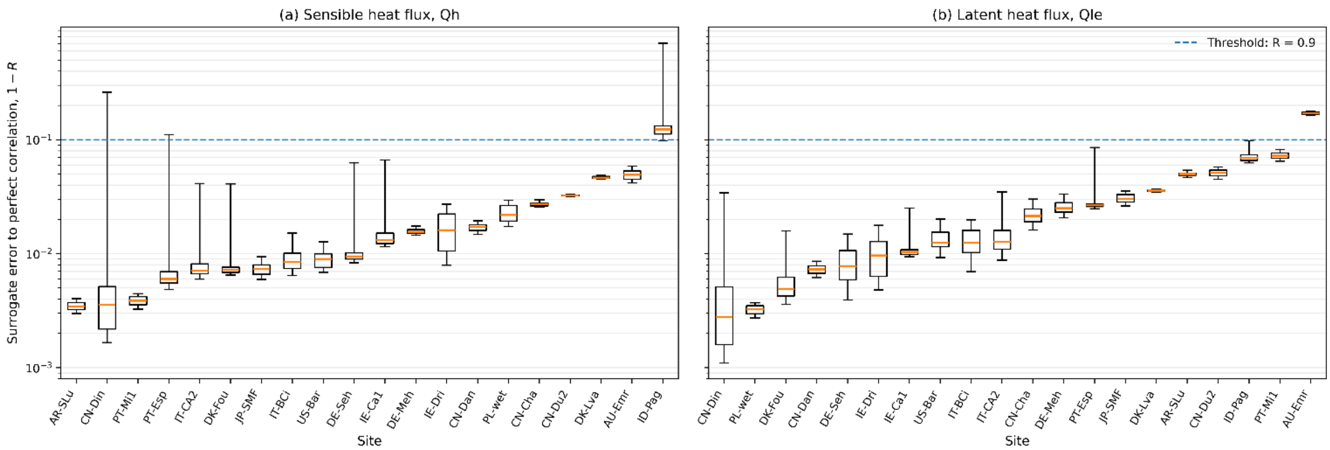


Figure 2. Site-wise surrogate–ILS consistency under joint parameter perturbations. Distributions of surrogate error to perfect correlation, defined as (1 - R), are shown for (a) sensible heat flux (Qh) and (b) latent heat flux (Qle). Each boxplot summarizes the results from 100 joint parameter perturbation samples at one site. Sites are ranked by median R in descending order. The y-axis is shown on a logarithmic scale to highlight differences among high-correlation sites. The dashed horizontal line corresponds to the consistency threshold of R = 0.9, equivalent to 1 - R = 0.1.

2.6 Experimental design and evaluation metrics

2.6.1 Experiment 1: Site-specific parameter calibration

Experiment 1 was designed to evaluate the performance of the MdPL framework for site-specific parameter calibration. This experiment involved 20 PLUMBER2 sites spanning four PFTs, including evergreen forest, woodland, grassland, and cultivation. The 20 sites were first used for the surrogate–ILS consistency screening described in Section 2.5. The main site-specific calibration analysis was then conducted for the sites that satisfied the consistency criterion for both sensible and latent heat fluxes. The final parameter vector used in ILS simulations was obtained by averaging the time-dependent parameter estimates produced by the MdPL model over the full training period. To further assess the impact of model architecture, a subset of four evergreen forest sites was selected to compare the multi-task surrogate with a single-task dPL surrogate.

The multi-task surrogate model consisted of a shared 4-layer LSTM with a hidden size of 128, followed by task-specific 2-layer LSTM branches with the same hidden size. In contrast, the single-task surrogate model adopted a 6-layer LSTM with a hidden size of 190. Both models used a time sequence length of 48 time steps, corresponding to 2 days of data. The total number of trainable parameters was 1,096,962 for the multi-task model and 980,400 for the single-task model. The parameter generator g_z was implemented as a 4-layer LSTM with a hidden size of 128 and the same sequence length.

The surrogate training dataset was constructed by combining default and perturbed-parameter ILS simulations. Both surrogate models were trained for 200 epochs using the Adam optimizer with a learning rate of 0.005. For the multi-task surrogate, a rolling training strategy was applied, in which the learning rate of the shared layer was reduced to one-third of the base learning rate. The parameter generator g_z was trained using forcing data, flux observations, and site attributes as inputs, with the same optimizer settings but for 2000 epochs.

For each site, one MdPL model was trained to generate a site-specific parameter set, which was subsequently used to drive the ILS model. The resulting simulations of sensible and latent heat fluxes were compared with observations using RMSE and Pearson's R to quantify calibration performance. In total, MdPL models were trained for the 20 candidate sites, whereas the main performance analysis focused on the consistency-screened subset. Four additional single-task dPL models were trained for the evergreen forest sites to isolate the effect of the multi-task surrogate architecture. This experiment demonstrates the ability of the MdPL framework to improve parameter calibration across different PFTs while maintaining task-specific accuracy.

2.6.2 Experiment 2: Benchmarking against LSMs and LSTM

Experiment 2 was designed to benchmark the MdPL-calibrated ILS against standard LSMs and data-driven approaches. The objectives were twofold: (1) to evaluate whether a deep-learning-calibrated physical model can outperform purely data-driven methods, and (2) to assess the performance of the MdPL-calibrated ILS relative to widely used LSMs.

This experiment was conducted using PLUMBER2 outputs from 14 sites. Starting from the 20 candidate sites, four sites (AR-SLu, CN-Din, PT-Mi1, and JP-SMF) were excluded because corresponding LSTM benchmarks were unavailable, and two additional sites (ID-Pag and AU-Emr) were excluded because they did not satisfy the surrogate–ILS consistency criterion. The evaluated models included the MdPL-calibrated ILS (ILS_MdPL), the default ILS (ILS_ORI), four widely used LSMs (CLM5, JULES, GFDL, and Noah), and a Long Short-Term Memory (LSTM) model.

https://doi.org/10.5194/egusphere-2026-4009
Preprint. Discussion started: 24 August 2026

In addition, the Surface Flux Equilibrium (SFE) method (McColl et al., 2019) was evaluated at 10 sites (DE-Meh, DK-Lva, IE-Dri, DK-Fou, DE-Seh, IE-Ca1, AU-Emr, PL-wet, IT-BCi, and IT-CA2), where the required ground heat flux observations were available .

No additional model training was performed in this experiment. Instead, the calibrated parameter sets obtained from Experiment 1 were used to drive the ILS simulations. All models were evaluated using consistent inputs and compared against observations of sensible and latent heat fluxes to ensure a fair comparison.

Furthermore, by comparing ILS_MdPL with LSTM and other LSMs, this experiment highlights the potential of the calibrated physical model to achieve performance comparable to or exceeding that of purely data-driven approaches.

2.6.3 Experiment 3: Comparison with traditional calibration baseline

To further evaluate the effectiveness of the proposed MdPL framework, we conducted an additional experiment comparing MdPL with a traditional calibration baseline based on Sobol-based random sampling.

Due to the high computational cost of running the ILS model for multiple parameter combinations, this experiment was performed at four representative sites selected from the main PFT categories. A total of 256 parameter sets were generated using Sobol sequence sampling with joint perturbations across all calibration parameters. Each parameter was sampled within predefined bounded ranges around its default value and constrained within physically meaningful limits.

For each sampled parameter set, the ILS model was executed, and model performance was evaluated using a unified objective function defined as RMSE(Qh) + RMSE(Qle). The results were then compared with those obtained from ILS_ORI, and ILS_MdPL to assess the relative performance and efficiency of the proposed framework.

2.6.4 Experiment 4: Out-of-target evaluation (GPP)

Experiment 4 was conducted to assess whether calibration based only on sensible and latent heat fluxes affects the simulation of another variable that was not directly included in the calibration targets. Gross Primary Productivity (GPP), which is available from both model outputs and FLUXNET observations, was used as an out-of-target carbon-related flux for additional evaluation. Specifically, we compared GPP simulated by ILS_ORI and ILS_MdPL against observed GPP at selected representative sites with available records.

2.6.5 Evaluation Metrics

To comprehensively assess the performance of the MdPL framework, multiple evaluation metrics were employed to characterize model accuracy, variability, and overall performance. The Root Mean Square Error (RMSE) was used to quantify the average magnitude of prediction errors, defined as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{11}$$

https://doi.org/10.5194/egusphere-2026-4009
Preprint. Discussion started: 24 August 2026

where y_i and $\hat{y}_i$ represent observed and predicted values, respectively, and n is the number of samples. The standard deviation (STD) of the predicted values was used to evaluate the variability of model outputs:

$$STD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}}_i)^2} \tag{12}$$

where $\bar{\hat{y}}_i$ denotes the mean of the predicted values. This metric is also used in the Taylor diagram to assess how well the model reproduces the variability of the observations. The Pearson correlation coefficient (R) was used to measure the linear relationship between predictions and observations:

$$R = \frac{\sum_{i=1}^{n}(y_i - \bar{y}_i)(\hat{y}_i - \bar{\hat{y}}_i)}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}}_i)^2}} \tag{13}$$

Kling–Gupta efficiency (KGE), an advanced performance metric that combines correlation, bias, and variability into a single score (Gupta et al., 2009), was also employed in this study. It was determined according to Eq. (14).

$$KGE = 1 - \sqrt{(r-1)^2 + (\beta-1)^2 + (\gamma-1)^2} \tag{14}$$

where r represents correlation coefficient, β represents the ratio of the mean predicted and observed values, and γ represents the ratio of the predicted and observed standard deviations. Nash–Sutcliffe efficiency (NSE), which provides a measure of how well predicted values match observed data relative to the mean of the observations, was also employed in this study. It was determined according to Eq. (15).

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{15}$$

NSE values close to 1 were indicative of a better fit, whereas negative values indicated that the mean of the observed data performs better than that of the model. In addition, relative performance improvements were quantified using ΔRMSE and ΔR. Specifically, ΔRMSE was defined as:

$$\Delta RMSE = (RMSE\_ORI - RMSE\_MdPL) / RMSE\_ORI \times 100\% \tag{16}$$

such that positive values indicate a reduction in error (i.e., improved performance). Similarly, ΔR was defined as:

$$\Delta R = (R\_MdPL - R\_ORI) / R\_ORI \times 100\% \tag{17}$$

where R represents the Pearson correlation coefficient. Positive values of ΔR indicate improved correlation between model simulations and observations.

3 Results and discussion

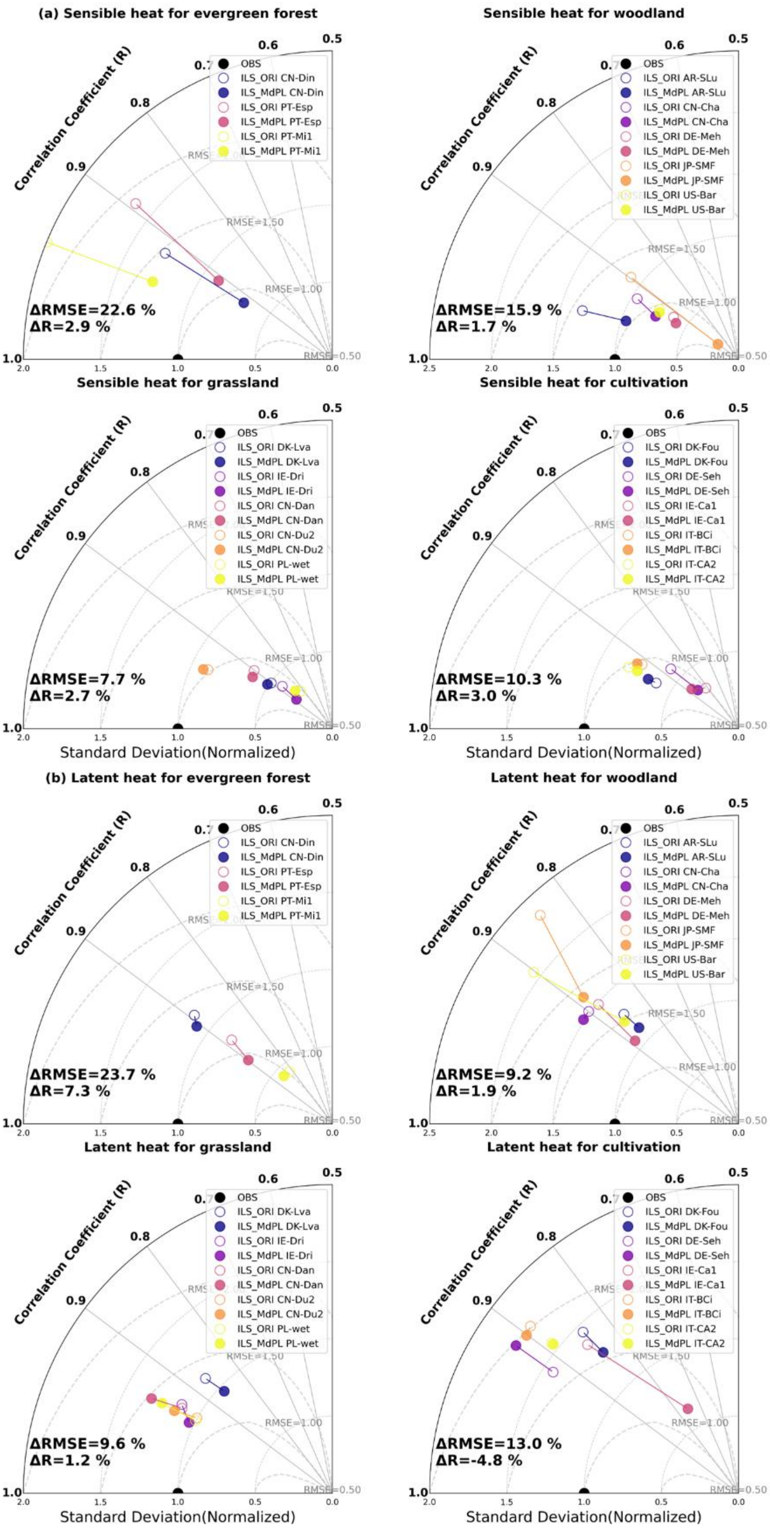
3.1 Site-specific parameter calibration and parameter plausibility

This section presents the results of **Experiment 1.** After the surrogate–ILS consistency screening described in Section 2.5, 18 of the 20 candidate sites were retained for the main site-specific calibration analysis. The calibrated parameter sets obtained from MdPL were then used to drive the original ILS, and the resulting sensible and latent heat flux simulations

https://doi.org/10.5194/egusphere-2026-4009
Preprint. Discussion started: 24 August 2026

were compared with observations and with simulations using the default parameter set. **Fig. 3** shows Taylor diagrams for sensible and latent heat flux simulations across the 18 consistency-screened sites. Detailed site-level calibration metrics and calibrated parameter values are provided in **Tables S1–S4**.

https://doi.org/10.5194/egusphere-2026-4009
Preprint. Discussion started: 24 August 2026



https://doi.org/10.5194/egusphere-2026-4009
Preprint. Discussion started: 24 August 2026

Figure 3. Taylor diagrams showing correlation coefficient R, normalized standard deviation (STD), and centered root mean square error (RMSE) for sensible and latent heat simulations across the 18 consistency-screened sites. (a) Sensible heat flux (Qh) and (b) latent heat flux (Qle). OBS denotes observations; ILS_ORI denotes simulations using default PFT parameters; ILS_MdPL denotes simulations using parameters calibrated with the MdPL framework. The symbol Δ indicates the relative improvement of ILS_MdPL compared with ILS_ORI, such as ΔRMSE and ΔR, expressed as percentages.

Overall, MdPL calibration improves sensible heat simulations across most PFTs, with the largest gains observed in evergreen forest sites. **Fig. 3(a)** shows that MdPL calibration significantly improved model performance for sensible heat flux. The most significant decrease in RMSE was observed for evergreen forest sites (~22.6%), whereas smaller improvements were observed for grassland and other PFTs (~7.7%). Further, cultivation sites exhibited the highest improvement in correlation, whereas woodland sites showed relatively smaller changes. In addition, the normalized standard deviations for sensible heat at most sites were closer to the observed values, indicating that the MdPL framework improved the representation of variability.

Fig. 3(b) shows latent heat simulation results. From this figure, it is evident that RMSEs decreased across all sites, with evergreen forest sites showing the most significant reduction (23.7%), whereas woodland sites showed a reduction (9.2%). In terms of correlation, most PFTs showed positive changes (ΔR = 7.3%, 1.9%, and 1.2% for evergreen forest, woodland, and grassland sites, respectively), whereas cultivation sites showed a decrease (ΔR = −4.8%). This indicates that reducing the magnitude of flux errors does not necessarily guarantee improved temporal correlation. One possible reason is that the calibration objective was based on flux error minimization rather than directly optimizing correlation. In addition, cultivation sites are strongly affected by crop phenology, management practices, and seasonal changes in vegetation structure. Because ILS uses fixed PFT-level parameters in this experiment, it may not fully capture time-varying controls on latent heat flux at cultivated sites.These results indicate that MdPL calibration generally improved site-specific ILS simulations for both sensible and latent heat fluxes, although the magnitude of improvement varied across PFTs and variables. Because the main analysis was restricted to the 18 sites that satisfied the surrogate–ILS consistency criterion, these improvements should be interpreted as calibration results obtained under verified surrogate reliability rather than as results for all initially selected PLUMBER2 sites.

In addition to model performance, we examined the plausibility of the calibrated parameter values. **Fig. 4** summarizes the normalized calibrated parameter values across the 18 consistency-screened sites. The normalized value indicates the position of each calibrated parameter within its predefined range, with 0 and 1 corresponding to the lower and upper bounds, respectively. Values close to 0 or 1 therefore indicate boundary-proximal solutions.

Several parameters showed clear boundary-proximal behavior. Boundary-proximal solutions were particularly frequent for vegh, rlfv, chl, and tlfn, indicating that vegetation structural, radiative, and leaf-exchange-related parameters were often pushed toward the limits of their predefined ranges during flux-only calibration. In contrast, parameters such as cdl and psicr showed broader distributions within their prescribed ranges, suggesting weaker systematic tendency toward the bounds (Gupta et al., 2008; Vrugt et al., 2005).

The presence of boundary-proximal solutions suggests that some calibrated parameters should not be interpreted as uniquely identifiable physical traits. In this study, parameter calibration was constrained only by sensible and latent heat flux observations. Under such flux-only constraints, multiple parameter combinations may produce similar heat flux simulations through compensatory adjustments, leading to equifinality and limited parameter identifiability. Therefore, the calibrated parameters should be interpreted as effective parameters under the given model structure and observational constraints, rather than as uniquely identifiable physical quantities (Beven and Freer, 2001; Beven, 2006; Kuppel et al., 2014; Rosolem et al., 2012).

This interpretation is particularly important for parameters related to vegetation structure, radiative transfer, and leaf exchange processes. Boundary-proximal behavior indicates where additional observational constraints may be needed. For example, vegetation structural information from site metadata or remote sensing, soil moisture observations, and carbon flux constraints could help reduce equifinality and improve parameter identifiability in future applications.

As an additional diagnostic, the comparison between the multi-task MdPL surrogate and the single-task dPL surrogate was conducted for four evergreen forest sites and is reported in Supplementary **Section S1**, **Table S5**, and **Fig. S2**. These supplementary results show that the multi-task surrogate achieved competitive calibration performance while jointly learning sensible and latent heat fluxes within a unified framework.

In summary, MdPL improved site-specific ILS simulations across the consistency-screened PLUMBER2 sites, particularly for evergreen forest sites. The parameter plausibility diagnostics further showed that some calibrated parameters approached predefined bounds, indicating limited parameter identifiability under flux-only calibration. Therefore, the MdPL-derived parameters are best interpreted as effective parameter sets that improve ILS behavior under the current observational constraints, rather than as uniquely identifiable physical parameter values.

https://doi.org/10.5194/egusphere-2026-4009
Preprint. Discussion started: 24 August 2026

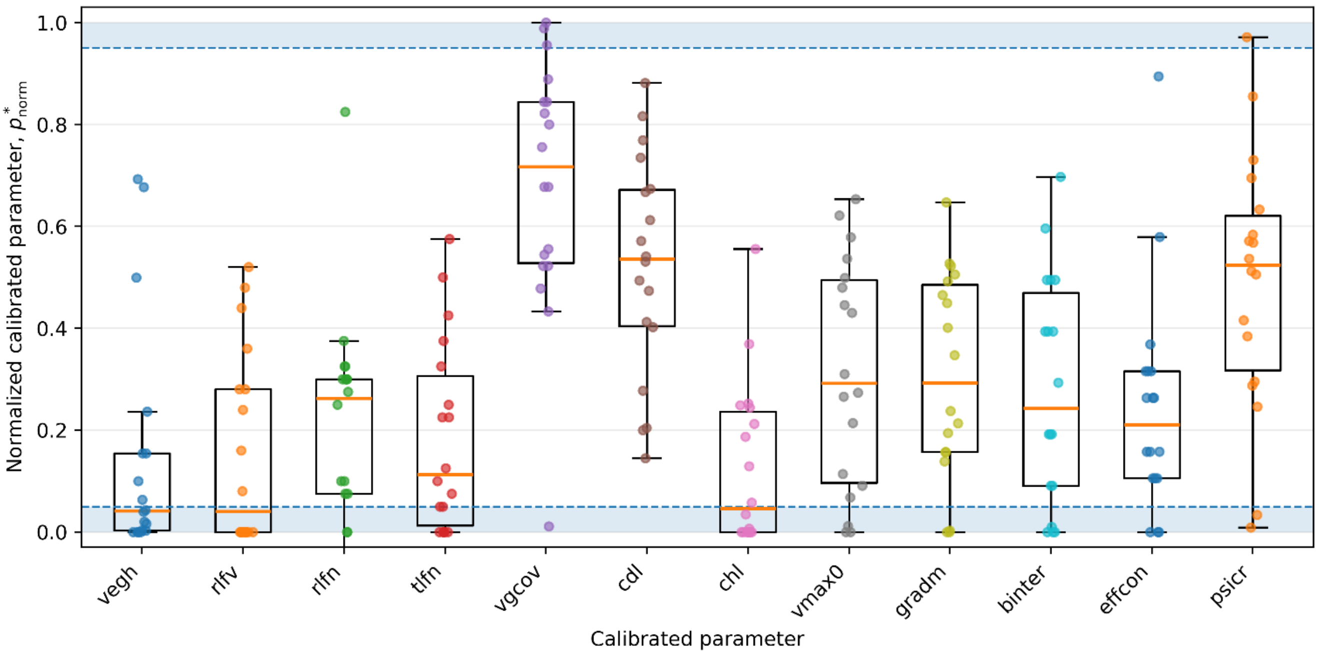


Figure 4. Parameter plausibility diagnostics for MdPL-calibrated parameters across the 18 consistency-screened sites.
470 Distributions of normalized calibrated parameter values are shown for each calibrated parameter, where 0 and 1 indicate the lower and upper bounds of the predefined parameter range, respectively. Each point represents the calibrated value for one site, and boxplots summarize the site-level distributions. Shaded regions indicate boundary-proximal zones, defined as normalized values lower than or equal to 0.05 or higher than or equal to 0.95.

https://doi.org/10.5194/egusphere-2026-4009
Preprint. Discussion started: 24 August 2026

3.2 Comparative Evaluation of Calibrated ILS and Widely Used LSMs Against LSTM

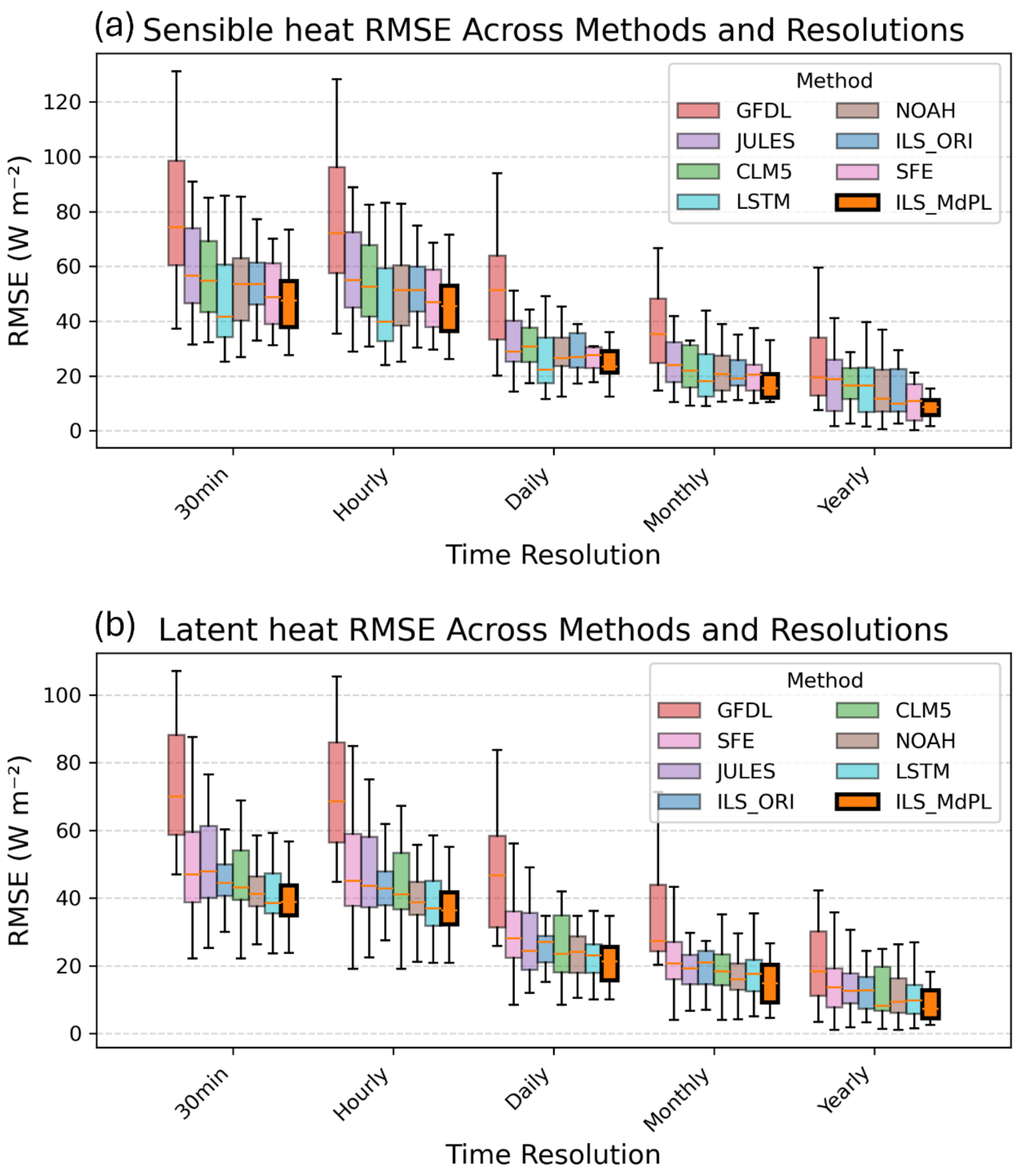


Figure 5. Comparison of RMSEs for sensible and latent heat fluxes across methods and resolutions. (a) Box plots of RMSE distributions for sensible heat flux. The numbers of sites included in the LSM/LSTM and SFE comparisons were 14 and 10, respectively. (b) Box plots showing the distribution of RMSEs for latent heat flux. The yellow line represents the median. The upper and lower box boundaries indicate 75% and 25% interquartile ranges, respectively.

https://doi.org/10.5194/egusphere-2026-4009
Preprint. Discussion started: 24 August 2026

This section presents the results of **Experiment 2**. **Fig. 5** shows the distribution of RMSEs across different time resolutions. In this study, we primarily focused on comparing the calibrated land surface model (ILS_MdPL) with the deep-learning-based data-driven model (LSTM), as this comparison provided more insightful implications than those obtained by merely comparing calibrated and uncalibrated LSMs.

At the finest time resolution (30 min), ILS_MdPL and LSTM exhibit comparable RMSEs. As the temporal resolution becomes coarser, ILS_MdPL generally shows slightly lower RMSEs than LSTM for both sensible and latent heat fluxes, although the magnitude of the difference varies across resolutions. This indicates that the physically constrained calibration approach of ILS_MdPL provides stable performance across different temporal scales. For latent heat, the advantage of ILS_MdPL is more evident in some resolutions, but does not consistently increase with coarser aggregation. For example, at the 30 min resolution, ILS_MdPL achieves a slightly lower RMSE (41.45 W m^{-2}) than LSTM (44.02 W m^{-2}). Overall, these results suggest that ILS_MdPL provides competitive performance, particularly for latent heat flux simulations, while avoiding strong dependence on temporal resolution.

Regarding sensible heat, the accuracy of SFE was similar to that of ILS_MdPL. However, for latent, it performed poorly at high-frequency resolutions (e.g., 30 min), but shows considerable improvement as the resolution increased. Specifically, at the monthly scale, its performance in predicting sensible heat ranked second only to that of ILS_MdPL, implying that it outperformed all the other traditional LSMs. This observation suggested that even simplified diagnostic methods such as SFE can be competitive at aggregated scales, possibly owing to their ability to capture large-scale patterns with substantially lower model complexity.

In summary, the evaluation across different models and temporal resolutions demonstrates the effectiveness and robustness of the proposed MdPL calibration method. Relative to the default parameter set and several widely used LSMs, the ILS_MdPL framework generally improves model performance in reproducing both sensible and latent heat fluxes. Compared with LSTM, the performance is comparable at fine temporal resolutions, and in some cases slightly better at coarser scales, although the advantage varies across resolutions and variables. Even though empirical methods such as SFE show limited performance at high-frequency resolutions, their competitiveness at aggregated time scales suggests their potential for simplified large-scale assessments. Overall, the proposed approach balances retention of the process-based model structure, multi-output calibration capability, and computational efficiency.

3.3 Comparison with traditional calibration baseline

To further evaluate the effectiveness of the proposed MdPL framework, we compared it with a traditional calibration baseline based on Sobol-based joint parameter sampling.

Fig. 6 shows the distribution of model performance for 256 Sobol sequence sampled parameter sets at four representative sites, along with the results obtained from ILS_ORI and ILS_MdPL. Each grey point represents one sampled parameter set generated through joint perturbations across all calibrated parameters within predefined bounded ranges. The x-axis shows

https://doi.org/10.5194/egusphere-2026-4009
Preprint. Discussion started: 24 August 2026

total RMSE, defined as $RMSE_{Qh} + RMSE_{Qle}$, whereas the y-axis shows total correlation, defined as $R_{Qh} + R_{Qle}$. Therefore, better performance corresponds to lower total RMSE and higher total correlation.

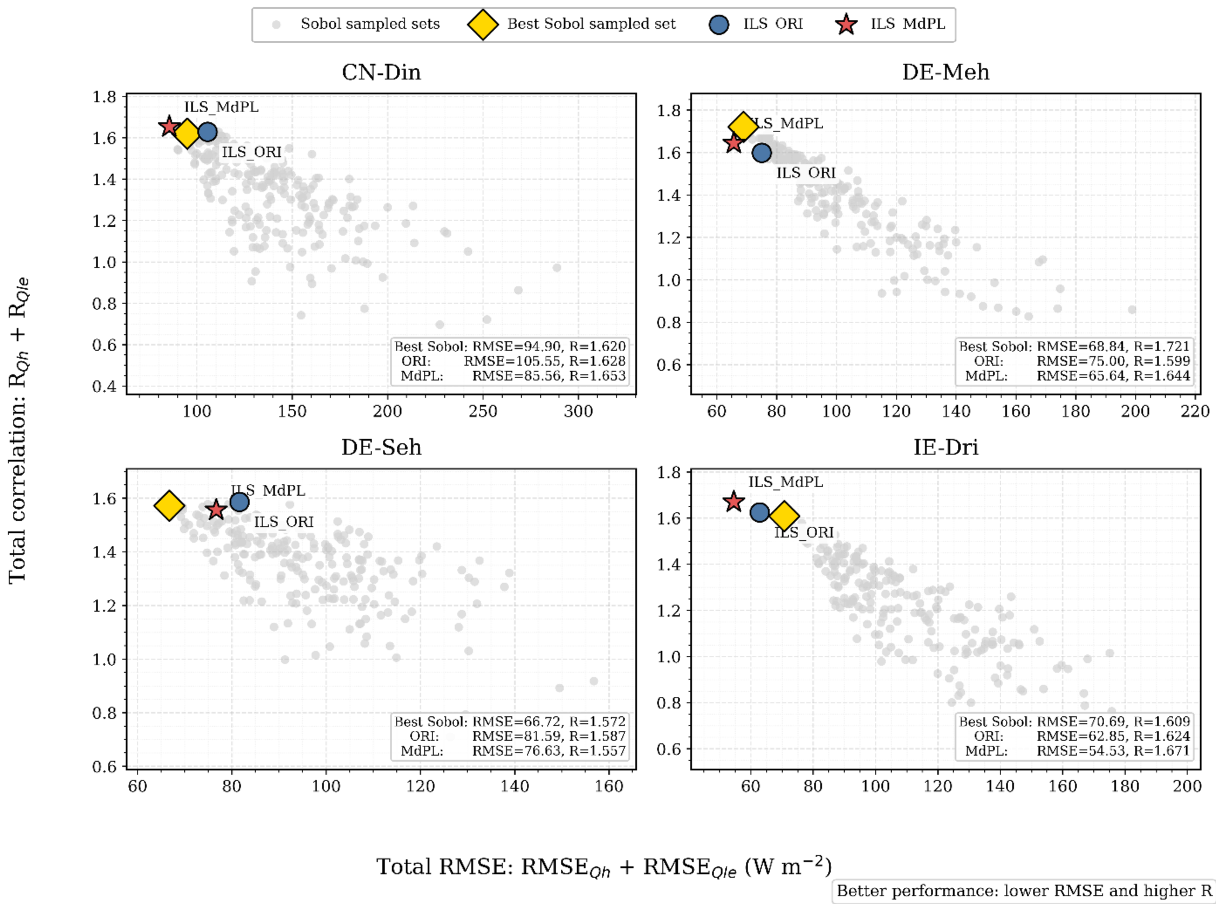


Figure 6. Comparison between MdPL and a traditional calibration baseline based on Sobol sampling at four representative sites. Each grey point represents one of the 256 parameter sets generated through joint perturbations across all calibration parameters. The x-axis shows total RMSE, defined as $RMSE_{Qh} + RMSE_{Qle}$, and the y-axis shows the total correlation, defined as $R_{Qh} + R_{Qle}$. The best-performing Sobol sequence sampled parameter set, selected based on the lowest total RMSE is highlighted by a yellow diamond.

The Sobol-sampled parameter sets showed substantial variability in both total RMSE and total correlation across the four sites, indicating that joint perturbations of the calibrated parameters can lead to a wide range of ILS responses. In most cases, parameter sets with lower total RMSE also tended to show higher total correlation, although this relationship was not strictly

https://doi.org/10.5194/egusphere-2026-4009
Preprint. Discussion started: 24 August 2026

monotonic. This result highlights the difficulty of identifying high-quality parameter combinations through direct sampling alone.

Compared with the Sobol baseline, ILS_MdPL achieved competitive but site-dependent performance. At CN-Din and IE-Dri, ILS_MdPL achieved both lower total RMSE and higher total correlation than the best Sobol sequence sampled parameter set and ILS_ORI. At DE-Meh, ILS_MdPL achieved the lowest total RMSE, although the best Sobol sequence sampled parameter set showed a higher total correlation. At DE-Seh, ILS_MdPL reduced total RMSE relative to ILS_ORI, but it did not outperform the best Sobol-sampled parameter set and showed slightly lower total correlation. These results indicate that MdPL can identify high-quality parameter solutions, but its relative advantage over direct Sobol sequence sampling depends on site characteristics and on whether performance is assessed primarily by RMSE or correlation.In terms of computational efficiency, the Sobol sequence sampling required 256 full ILS simulations, whereas MdPL used the trained differentiable surrogate to optimize parameters through gradient-based learning, thereby avoiding exhaustive full-model evaluation of all candidate parameter combinations during calibration.

Although MdPL requires surrogate pretraining, this cost can be amortized when calibration is applied across multiple sites or parameter sets. Therefore, the main advantage of MdPL is not that it always outperforms Sobol sequence sampling at every site, but that it provides a scalable and differentiable calibration pathway that can identify competitive parameter solutions without repeated full ILS simulations for every candidate parameter set.

Overall, the four-site Sobol comparison provides a direct baseline for assessing MdPL against traditional sampling-based calibration. The results show that MdPL generally produced competitive parameter solutions and often achieved lower total RMSE than the best Sobol sequence sampled solution, while also revealing that calibration performance remains site- and metric-dependent.

3.4 Out-of-target evaluation using GPP

To further assess whether calibration based on sensible and latent heat fluxes affects other simulated variables, we evaluated GPP, which was not included in the calibration targets. While net ecosystem exchange (NEE) would be an important variable for such an analysis, it is not available in the current ILS configuration used in this study. Therefore, GPP was used as an alternative carbon-related flux for out-of-target evaluation.

For this analysis, one representative site was selected for each PFT category. The calibrated parameter sets obtained from MdPL using Qh and Qle were applied to simulate GPP at these sites. The resulting GPP simulations were then compared with observations from the PLUMBER2 dataset, alongside simulations using the default parameter configuration (ILS_ORI). **Fig. 7** shows the comparison between observed GPP and simulations from ILS_ORI and ILS_MdPL at four representative sites. The results indicate that the impact of MdPL calibration on GPP is site-dependent. At CN-Din, the MdPL-calibrated model shows a clear improvement compared to the original configuration, with reduced RMSE and improved correlation. At DE-Meh, both configurations perform well, with MdPL showing slight improvements. At DE-Seh, the performance remains

https://doi.org/10.5194/egusphere-2026-4009
Preprint. Discussion started: 24 August 2026

nearly unchanged between the two models. In contrast, at IE-Dri, a degradation in GPP performance is observed after calibration.

Overall, these results suggest that calibration against sensible and latent heat fluxes does not lead to systematic degradation in GPP. Although improvements are not consistently observed across all sites, the absence of widespread deterioration provides additional confidence that the MdPL framework maintains reasonable performance for a non-target carbon-related flux.

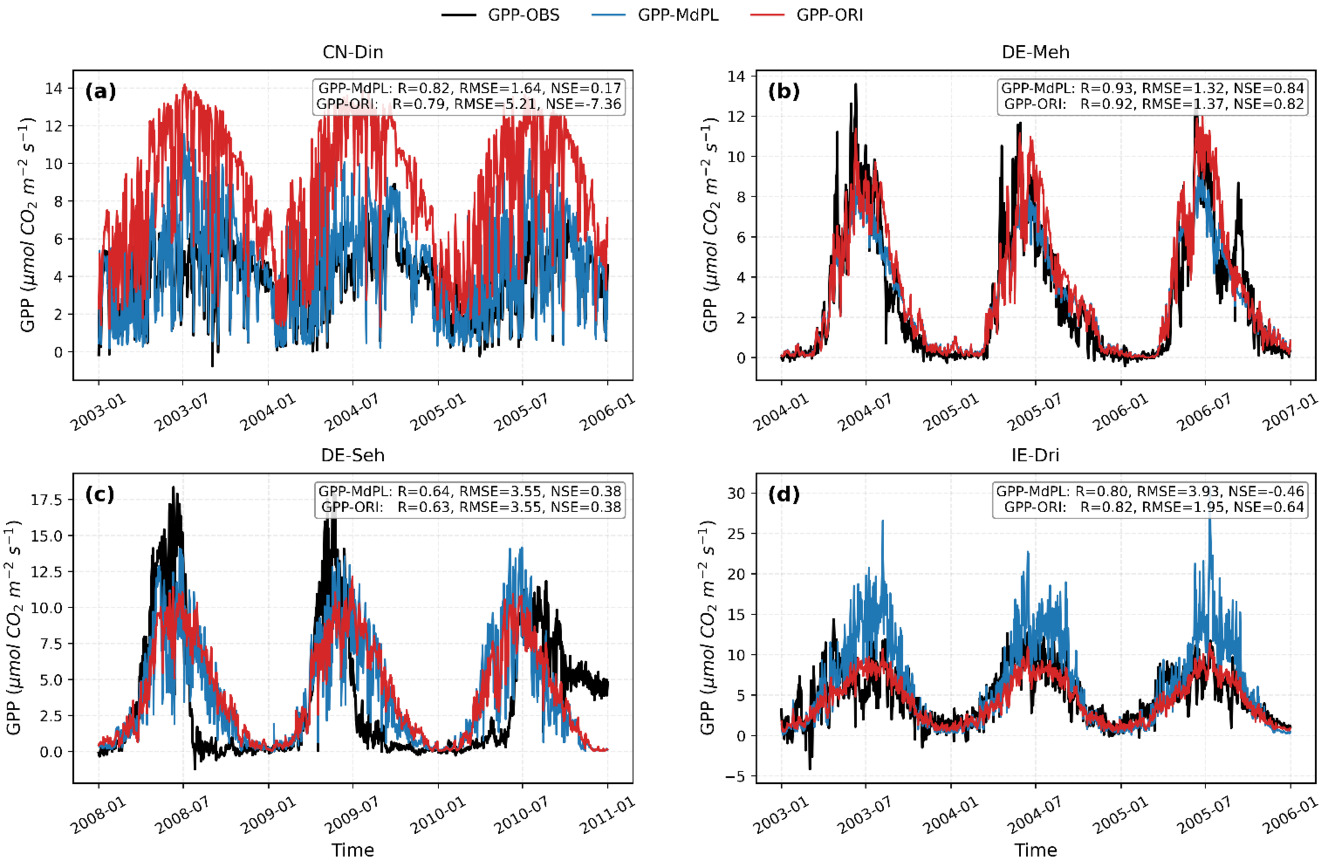


Figure 7. Out-of-target evaluation of GPP at four representative sites using MdPL-calibrated and default parameter sets. GPP-OBS represents observed GPP from FLUXNET2015. GPP-MdPL represents simulations using parameters calibrated by the MdPL framework based on sensible and latent heat fluxes. GPP-ORI represents simulations using the default parameter configuration of the model.

4. Conclusion

In this study, we introduced MdPL, a deep-learning-based calibration framework that combines a multi-task neural surrogate with a differentiable parameter generator to improve parameter calibration in land surface models. The framework was

https://doi.org/10.5194/egusphere-2026-4009
Preprint. Discussion started: 24 August 2026

evaluated using 20 PLUMBER2 sites spanning four PFTs. After surrogate–ILS consistency screening under joint parameter perturbations, 18 sites were retained for the main site-specific calibration analysis. Across these consistency-screened sites, the MdPL-calibrated ILS reduced RMSE relative to the default parameter set by 14.1% and 13.9% at the PFT-mean level for sensible and latent heat fluxes, respectively.

Benchmarking against PLUMBER2 model outputs showed that the MdPL-calibrated ILS achieved competitive performance relative to widely used LSMs, including CLM5, JULES, Noah, and GFDL, as well as an LSTM benchmark. However, the relative advantage of ILS_MdPL varied across variables, sites, temporal resolutions, and evaluation metrics, indicating that MdPL did not uniformly outperform all alternative models. Instead, the results suggest that MdPL provides a competitive calibration-based pathway for improving a process-based LSM while retaining its physical model structure.

The comparison with a traditional calibration baseline based on Sobol sequence sampling at four representative sites further showed that MdPL can identify high-quality parameter solutions without exhaustive full-model evaluation of all candidate parameter combinations. However, this comparison also revealed site- and metric-dependent behavior: MdPL achieved clear advantages at some sites, whereas the best Sobol sequence sampled parameter set remained competitive at others. Thus, the main advantage of MdPL is not that it always outperforms direct sampling, but that it provides a scalable and differentiable calibration pathway for high-dimensional parameter optimization.

The out-of-target evaluation using GPP suggested that calibration against sensible and latent heat fluxes did not lead to systematic degradation of a non-target carbon-related flux, although the impact of calibration remained site dependent. This result provides additional evidence that MdPL calibration did not broadly degrade carbon-related model behavior in the selected representative sites, but broader evaluation using additional carbon-cycle variables and more sites would be needed before drawing general conclusions about carbon-cycle performance.

The parameter plausibility diagnostics showed that some calibrated parameters approached predefined bounds, especially for vegetation structural, radiative, and leaf-exchange-related parameters. This boundary-proximal behavior indicates limited parameter identifiability under flux-only observational constraints and suggests that some improvements in heat flux simulations may arise from compensatory parameter adjustments. Therefore, the MdPL-derived parameters should be interpreted as effective parameter sets under the given model structure and observational constraints, rather than as uniquely identifiable physical quantities.

Overall, these results show that MdPL provides a scalable, efficient, and diagnostically transparent framework for improving parameter calibration in complex LSMs. Future work should extend the calibration targets to additional fluxes and state variables, incorporate independent observational constraints such as soil moisture, vegetation structure, and carbon fluxes, and evaluate the framework across a broader range of sites in the full PLUMBER2 archive.

https://doi.org/10.5194/egusphere-2026-4009
Preprint. Discussion started: 24 August 2026

Data Availability

The PLUMBER2 benchmarking dataset used for model evaluation is publicly available at the Australian Research Data Commons: https://researchdata.edu.au/plumber2-forcing-evaluation-surface-models/1656048. The underlying FLUXNET2015, LaThuile and OzFlux tower data are provided at https://fluxnet.org/. All data generated and analyzed during this study are publicly available in the Zenodo repository at https://doi.org/10.5281/zenodo.15753067. Detailed instruction are provided in the repository README. These resources are released under a CC-BY 4.0 license.

Code Availability

All custom code for data preprocessing, model training, ILS simulation, calibration experiments, and figure generation is publicly available at Zenodo: https://doi.org/10.5281/zenodo.15748737. Detailed instructions, environment specification, and scripts are provided in the repository README.

Author Contributions

WX conceptualized the study, developed the MdPL framework, ran the ILS simulations and analyses, and drafted the manuscript. HL carried out data preprocessing (PLUMBER2), conceptualized the study, and contributed to manuscript revision. KY supervised the research, provided scientific guidance, and contributed to manuscript revision. All authors discussed the results and approved the final version.

Financial support

This study was supported by the water environment and resource research project at the Earth Observation Research Center, Japan Aerospace Exploration Agency (JAXA EORC). This work was supported by the Japan Society for the Promotion of Science (JSPS) via Grants-in-Aid 22H04938 and 21H05002; the Japan Science and Technology Agency (JST) via the "New Social Challenges" project (Grant JPMJMI24I1).

Competing Interests

The authors declare that they have no conflict of interest.

https://doi.org/10.5194/egusphere-2026-4009
Preprint. Discussion started: 24 August 2026

References

Abramowitz, G., Ukkola, A., Hobeichi, S., Cranko Page, J., Lipson, M., De Kauwe, et al. (2024). On the predictability of turbulent fluxes from land: PLUMBER2 MIP experimental description and preliminary results. EGUsphere [preprint]. https://doi.org/10.5194/egusphere-2023-3084.

Bacour, C., Macbean, N., Chevallier, F., Léonard, S., Koffi, E. N., & Peylin, P. (2023). Assimilation of multiple datasets results in large differences in regional-to global-scale NEE and GPP budgets simulated by a terrestrial biosphere model. Biogeosciences, 20(6), 1089–1111. https://doi.org/10.5194/bg-20-1089-2023.

Bao, S., Alonso, L., Wang, S., Gensheimer, J., De, R., & Carvalhais, N. (2023). Toward robust parameterizations in ecosystem-level photosynthesis models. Journal of Advances in Modeling Earth Systems, 15, e2022MS003464. https://doi.org/10.1029/2022MS003464

Beven, K., & Freer, J. (2001). Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems. *Journal of Hydrology*, 249(1–4), 11–29. https://doi.org/10.1016/S0022-1694(01)00421-8.

Beven, K. (2006). A manifesto for the equifinality thesis. *Journal of Hydrology*, 320(1–2), 18–36. https://doi.org/10.1016/j.jhydrol.2005.07.007.

Blyth, E. M., Arora, V. K., Clark, D. B., Dadson, S. J., De Kauwe, M. G., Lawrence, D. M., et al. (2021). Advances in Land Surface Modelling. Current Climate Change Reports, 7(2), 45–71. https://doi.org/10.1007/s40641-021-00171-5.

Buotte, P. C., Koven, C. D., Xu, C., Shuman, J. K., Goulden, M. L., Levis, et al. (2021). Capturing functional strategies and compositional dynamics in vegetation demographic models. Biogeosciences, 18(14), 4473–4490. https://doi.org/10.5194/bg-18-4473-2021.

Cameron, D., Hartig, F., Minnuno, F., Oberpriller, J., Reineking, B., Van Oijen, M., et al. (2022). Issues in calibrating models with multiple unbalanced constraints: the significance of systematic model and data errors. Methods in Ecology and Evolution, 13(12), 2757–2770. https://doi.org/10.1111/2041-210X.14002.

Cardinale, B. J., Duffy, J. E., Gonzalez, A., Hooper, D. U., Perrings, C., Venail, P., et al. (2012). Biodiversity loss and its impact on humanity. Nature, 486(7401), 59–67. https://doi.org/10.1038/nature11148.

Chen, X., Huang, Y., Nie, C., Zhang, S., Wang, G., Chen, S., et al. (2022). A long-term reconstructed TROPOMI solar-induced fluorescence dataset using machine learning algorithms. Scientific Data, 9(1), 1–11. https://doi.org/10.1038/s41597-022-01520-1.

Ek, M. B., Mitchell, K. E., Lin, Y., Rogers, E., Grunmann, P., Koren, V., et al. (2003). Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model. Journal of Geophysical Research: Atmospheres, 108(22), 1–16. https://doi.org/10.1029/2002jd003296.

https://doi.org/10.5194/egusphere-2026-4009
Preprint. Discussion started: 24 August 2026

Exbrayat, J. F., Pitman, A. J., & Abramowitz, G. (2014). Response of microbial decomposition to spin-up explains CMIP5 soil carbon range until 2100. Geoscientific Model Development, 7(6), 2683–2692. https://doi.org/10.5194/gmd-7-2683-2014.

Famiglietti, C. A., Smallman, T. L., Levine, P. A., Flack-Prain, S., Quetin, G. R., Meyer, V., et al. (2021). Optimal model complexity for terrestrial carbon cycle prediction. Biogeosciences, 18(8), 2727–2754. https://doi.org/10.5194/bg-18-2727-2021.

Feng, D., Liu, J., Lawson, K., & Shen, C. (2022). Differentiable, learnable, regionalized process-based models with multiphysical outputs can approach state-of-the-art hydrologic prediction accuracy. Water Resources Research, 58(10), e2022WR032404. https://doi.org/10.1029/2022WR032404.

Feng, D., Beck, H., Lawson, K., & Shen, C. (2023). The suitability of differentiable, physics-informed machine learning hydrologic models for ungauged regions and climate change impact assessment. Hydrology and Earth System Sciences, 27(12), 2357–2373. https://doi.org/10.5194/hess-27-2357-2023.

Fisher, R. A., & Koven, C. D. (2020). Perspectives on the Future of Land Surface Models and the Challenges of Representing Complex Terrestrial Systems. Journal of Advances in Modeling Earth Systems, 12(4), e2018MS001453. https://doi.org/10.1029/2018MS001453.

Foumani, N. M., Tan, C. W., Webb, G. I., & Salehi, M. (2024). Improving position encoding of transformers for multivariate time series classification. Data Mining and Knowledge Discovery, 38(1), 22–48. https://doi.org/10.1007/s10618-023-00948-2.

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. Journal of Hydrology, 377(1–2), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003.

Gupta, H. V., Wagener, T., & Liu, Y. (2008). Toward improved identification of hydrological models: A diagnostic evaluation framework. Water Resources Research, 44(1), W00B03. https://doi.org/10.1029/2007WR006716.

MATSIRO6 Document Writing Team (2021). Description of MATSIRO6, CCSR Report No. 66, Division of Climate System Research, Atmosphere and Ocean Research Institute, The University of Tokyo. https://doi.org/10.15083/0002000181.

Han, Y., Huang, G., Song, S., Yang, L., Wang, H., & Wang, Y. (2022). Dynamic Neural Networks: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(11), 7436–7456. https://doi.org/10.1109/TPAMI.2021.3117837.

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

Hoppe, H., Dietrich, P., Marzahn, P., Weiß, T., Nitzsche, C., Freiherr von Lukas, U., et al. (2024). Transferability of Machine Learning Models for Crop Classification in Remote Sensing Imagery Using a New Test Methodology: A Study on Phenological, Temporal, and Spatial Influences. Remote Sensing, 16(9), 1–19. https://doi.org/10.3390/rs16091493.

Hou, T., Zhu, Y., Lü, H., Sudicky, E., Yu, Z., & Ouyang, F. (2015). Parameter sensitivity analysis and optimization of Noah land surface model with field measurements from Huaihe River Basin, China. Stochastic Environmental Research and Risk Assessment, 29(5), 1383–1401. https://doi.org/10.1007/s00477-015-1033-5.

Huang, Y., Zhu, D., Ciais, P., Guenet, B., Huang, Y., Goll, D. S., et al. (2018). Matrix-based sensitivity assessment of soil organic carbon storage: A case study from the ORCHIDEE-MICT model. Journal of Advances in Modeling Earth Systems, 10(8), 1790–1808. https://doi.org/10.1029/2017MS001237.

Kaminski, T., Knorr, W., Rayner, P. J., & Heimann, M. (2002). Assimilating atmospheric data into a terrestrial biosphere model: A case study of the seasonal cycle. Global Biogeochemical Cycles, 16(4), 14-1–14-16. https://doi.org/10.1029/2001gb001463.

Keenan, T. F., Davidson, E. A., Munger, J. W., & Richardson, A. D. (2013). Rate my data: Quantifying the value of ecological data for the development of models of the terrestrial carbon cycle. Ecological Applications, 23(1), 273–286. https://doi.org/10.1890/12-0747.1.

Kolassa, J., Reichle, R. H., Liu, Q., Cosh, M., Bosch, D. D., Caldwell, T. G., et al. (2017). Data assimilation to extract soil moisture information from SMAP observations. Remote Sensing, 9(11), 1–24. https://doi.org/10.3390/rs9111179.

Kwon, Y., Forman, B. A., Ahmad, J. A., Kumar, S. V., & Yoon, Y. (2019). Exploring the utility of machine learning-based passive microwave brightness temperature data assimilation over terrestrial snow in high Mountain Asia. Remote Sensing, 11(19). https://doi.org/10.3390/rs11192265.

Kuppel, S., Peylin, P., Chevallier, F., Bacour, C., Maignan, F., & Richardson, A. D. (2014). Model–data fusion across ecosystems: from multisite optimizations to global simulations. *Geoscientific Model Development*, 7(6), 2581–2597. https://doi.org/10.5194/gmd-7-2581-2014.

Leontjeva, A., & Kuzovkin, I. (2016). Combining static and dynamic features for multivariate sequence classification. Proceedings - 3rd IEEE International Conference on Data Science and Advanced Analytics, DSAA 2016, 21–30. https://doi.org/10.1109/DSAA.2016.10.

Li, H., Xie, W., Li, X., & Yoshimura, K. (2025). Improvement of simulating gross primary production and evapotranspiration in the land surface model based on parameter calibration. *Hydrological Research Letters*, 19(4), 252–259. https://doi.org/10.3178/hrl.25-0003.

Li, L., Bisht, G., Hao, D., & Leung, L. R. (2024). Global 1 km land surface parameters for kilometer-scale Earth system modeling. Earth System Science Data, 16(4), 2007–2032. https://doi.org/10.5194/essd-16-2007-2024.

Liang, X., Lettenmaier, D. P., Wood, E. F., & Burges, S. J. (1994). A simple hydrologically based model of land surface water and energy fluxes for general circulation models. Journal of Geophysical Research, 99(D7), 14415–14428. https://doi.org/10.1029/94jd00483.

Lumumba, V. W., Kiprotich, D., Makena, N. G., Kavita, M. D., & Mpaine, M. L. (2024). Comparative Analysis of Cross-Validation Techniques: LOOCV, K-folds Cross-Validation, and Repeated K-folds Cross-Validation in Machine Learning

https://doi.org/10.5194/egusphere-2026-4009
Preprint. Discussion started: 24 August 2026

Models. American Journal of Theoretical and Applied Statistics, 13(5), 127–137. https://doi.org/10.11648/j.ajtas.20241305.13.

MacBean, N., Peylin, P., Chevallier, F., Scholze, M., & Schürmann, G. (2016). Consistent assimilation of multiple data streams in a carbon cycle data assimilation system. Geoscientific Model Development, 9(10), 3569–3588. https://doi.org/10.5194/gmd-9-3569-2016.

Martin, A. R., & Isaac, M. E. (2018). Functional traits in agroecology: Advancing description and prediction in agroecosystems. Journal of Applied Ecology, 55(1), 5–11. https://doi.org/10.1111/1365-2664.13039.

Massoud, E. C., Xu, C., Fisher, R. A., Knox, R. G., Walker, A. P., Serbin, S. P., et al. (2019). Identification of key parameters controlling demographically structured vegetation dynamics in a land surface model: CLM4.5(FATES). Geoscientific Model Development, 12(9), 4133–4164. https://doi.org/10.5194/gmd-12-4133-2019.

McColl, K. A., Salvucci, G. D., & Gentine, P. (2019). Surface flux equilibrium theory explains an empirical estimate of water-limited daily evapotranspiration. Journal of Advances in Modeling Earth Systems, 11, 2036–2049.

McNeall, D., Robertson, E., & Wiltshire, A. (2024). Constraining the carbon cycle in JULES-ES-1.0. Geoscientific Model Development, 17(3), 1059–1089. https://doi.org/10.5194/gmd-17-1059-2024.

Medvigy, D., Wofsy, S. C., Munger, J. W., Hollinger, D. Y., & Moorcroft, P. R. (2009). Mechanistic scaling of ecosystem function and dynamics in space and time: Ecosystem Demography model version 2. Journal of Geophysical Research: Biogeosciences, 114(1), 1–21. https://doi.org/10.1029/2008JG000812.

NCAR. (2020). CLM5 Documentation. March 23, 337. https://www.cesm.ucar.edu/.

Nitta, T., Arakawa, T., Hatono, M., Takeshima, A., & Yoshimura, K. (2020). Development of Integrated Land Simulator. Progress in Earth and Planetary Science, 7(1), 68. https://doi.org/10.1186/s40645-020-00383-7.

Pastorello, G., Trotta, C., Canfora, E., Chu, H., Christianson, D., Cheah, Y. W., et al. (2020). The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data. Scientific Data 7(1), 225. https://doi.org/10.1038/s41597-020-0534-3.

Oberpriller, J., Herschlein, C., Anthoni, P., Arneth, A., Krause, A., Rammig, et al. (2022). Climate and parameter sensitivity and induced uncertainties in carbon stock projections for European forests (using LPJ-GUESS 4.0). Geoscientific Model Development, 15(16), 6495–6519. https://doi.org/10.5194/gmd-15-6495-2022.

Peylin, P., Bacour, C., MacBean, N., Leonard, S., Rayner, P., Kuppel, S., et al. (2016). A new stepwise carbon cycle data assimilation system using multiple data streams to constrain the simulated land surface carbon cycle. Geoscientific Model Development, 9(9), 3321–3346. https://doi.org/10.5194/gmd-9-3321-2016.

Poulter, B., Ciais, P., Hodson, E., Lischke, H., Maignan, F., Plummer, S., et al. (2011). Plant functional type mapping for earth system models. Geoscientific Model Development, 4(4), 993–1010. https://doi.org/10.5194/gmd-4-993-2011.

Poulter, B., MacBean, N., Hartley, A., Khlystova, I., Arino, O., Betts, R., et al. (2015). Plant functional type classification for earth system models: Results from the European Space Agency's Land Cover Climate Change Initiative. Geoscientific Model Development, 8(7), 2315–2328. https://doi.org/10.5194/gmd-8-2315-2015.

Raiho, A. M., Nicklen, E. F., Foster, A. C., Roland, C. A., & Hooten, M. B. (2021). Bridging implementation gaps to connect large ecological datasets and complex models. Ecology and Evolution, 11(24), 18271–18287. https://doi.org/10.1002/ece3.8420.

Raoult, N., Douglas, N., Macbean, N., Kolassa, J., Quaife, T., Roberts, A. G., et al. (2024). Parameter Estimation in Land Surface Models: Challenges and Opportunities with Data Assimilation and Machine Learning. ESS Open Archive. https://doi.org/10.22541/essoar.172838640.01153603/v1.

Rayner, P. J., Michalak, A. M., & Chevallier, F. (2019). Fundamentals of data assimilation applied to biogeochemistry. Atmospheric Chemistry and Physics, 19(22), 13911–13932. https://doi.org/10.5194/acp-19-13911-2019.

Rayner, P. J., Scholze, M., Knorr, W., Kaminski, T., Giering, R., & Widmann, H. (2005). Two decades of terrestrial carbon fluxes from a carbon cycle data assimilation system (CCDAS). Global Biogeochemical Cycles, 19(2), 1–20. https://doi.org/10.1029/2004GB002254.

Rivera, A., Bravo, C., & Buob, G. (2017). Climate Change and Land Ice. International Encyclopedia of Geography, 1–15. https://doi.org/10.1002/9781118786352.wbieg0538.

Rodríguez-Fernández, N., de Rosnay, P., Albergel, C., Richaume, P., Aires, F., Prigent, C., et al. (2019). SMOS neural network soil moisture data assimilation in a land surface model and atmospheric impact. Remote Sensing, 11(11). https://doi.org/10.3390/rs11111334.

Rosolem, R., Gupta, H. V., Shuttleworth, W. J., Zeng, X., & Gonçalves, L. G. G. (2012). Towards a consistent soil moisture estimation framework: integrating remote sensing, modeling and data assimilation. Hydrology and Earth System Sciences, 16(10), 3877–3890. https://doi.org/10.5194/hess-16-3877-2012.

Ruder, S. (2017). An overview of multi-task learning in deep neural networks. arXiv,arXiv:1706.05098. https://doi.org/10.48550/arXiv.1706.05098.

Sawada, Y. (2020). Machine Learning Accelerates Parameter Optimization and Uncertainty Assessment of a Land Surface Model. Journal of Geophysical Research: Atmospheres, 125(20), e2020JD032688. https://doi.org/10.1029/2020JD032688.

Schürmann, G. J., Kaminski, T., Köstler, C., Carvalhais, N., Voßbeck, M., Kattge, J., et al. (2016). Constraining a land-surface model with multiple observations by application of the MPI-Carbon Cycle Data Assimilation System V1.0. Geoscientific Model Development, 9(9), 2999–3026. https://doi.org/10.5194/gmd-9-2999-2016.

Shen, C., Appling, A. P., Gentine, P., Bandai, T., Gupta, H., Tartakovsky, A., et al. (2023). Differentiable modelling to unify machine learning and physical models for geosciences. Nature Reviews Earth and Environment, 4(8), 552–567. https://doi.org/10.1038/s43017-023-00450-9.

Shevliakova, E., Malyshev, S., Martinez-Cano, I., Milly, P. C. D., Pacala, S. W., Ginoux, P., et al. (2024). The Land Component LM4.1 of the GFDL Earth System Model ESM4.1: Model Description and Characteristics of Land Surface Climate and Carbon Cycling in the Historical Simulation. Journal of Advances in Modeling Earth Systems, 16(5), 1–47. https://doi.org/10.1029/2023MS003922.

Tibangayuka, N., Mulungu, D. M. M., & Izdori, F. (2022). Evaluating the performance of HBV, HEC-HMS and ANN models in simulating streamflow for a data scarce high-humid tropical catchment in Tanzania. Hydrological Sciences Journal, 67(14), 2191–2204. https://doi.org/10.1080/02626667.2022.2137417.

Tsai, W. P., Feng, D., Pan, M., Beck, H., Lawson, K., Yang, Y., et al. 2021. From calibration to parameter learning: Harnessing the scaling effects of big data in geoscientific modeling. Nature Communications, 12(1), 5988. https://doi.org/10.1038/s41467-021-26107-z.

Vrugt, J. A., Gupta, H. V., Bouten, W., & Sorooshian, S. (2005). Improved treatment of uncertainty in hydrologic modeling: Combining the strengths of global optimization and data assimilation. Water Resources Research, 41(1), W01017. https://doi.org/10.1029/2004WR003059.

Weng, E., Luo, Y., Gao, C., & Oren, R. (2011). Uncertainty analysis of forest carbon sink forecast with varying measurement errors: A data assimilation approach. Journal of Plant Ecology, 4(3), 178–191. https://doi.org/10.1093/jpe/rtr018.