Reviewer #3

Reviewer Comment:

This reviewer finds the manuscript interesting and could be a contribution to the community; however, it is not in its current form. Below are some key comments this reviewer found it important to address:

- Please clarify and correct the parameter-bounding implementation and reporting. If values are clipped, the reported values should reflect the clipping; if not, the claim is wrong, and the optimisation may be physically invalid.

For example, in Appendix A1, vmax0 (or vamx0 as written in the manuscript) is -9.35E-06 for ID-Pag, and in Appendix A5, vmax0 = -8.14E-07 for PT-Mi1. Negative vmax0 is not physically possible.

Authors' Response:

We thank the reviewer for the careful evaluation of our manuscript and for the constructive and insightful comments. We appreciate the recognition of the potential contribution of this work to the community. The comments have been very helpful in improving the clarity, robustness, and physical interpretation of our study.

We also thank the reviewer for pointing out this important issue regarding parameter bounding and reporting, as well as for the careful examination of the Appendix tables.

In the current MdPL framework, two design choices contribute to the appearance of out-of-range values in intermediate outputs. First, parameter bounds are enforced using a soft constraint rather than hard clipping. This design preserves differentiability during training, as direct clipping would break gradient propagation and hinder optimization. Under this mechanism, when predicted parameters exceed the predefined bounds, a penalty term is added to the loss function to encourage them to remain within the valid range. Second, all parameters are normalized during training to reduce the influence of differences in parameter magnitudes. The combination of normalization and inverse transformation may lead to slight overshooting beyond the predefined bounds in the raw outputs.

As a result, a small number of intermediate parameter values may appear outside the predefined ranges during optimization. However, these values are not used in the ILS simulations. Before being passed to the ILS model, all parameters are mapped back to their valid physical ranges, and the simulations are conducted using these bounded values (in Table A1-A5, bold font means out of range during training).

The issue identified by the reviewer arose because the Appendix tables in the previous version did not fully reflect the bounded values actually used in ILS. In other words, although the simulations were conducted with constrained parameter values, a few raw outputs that had not been updated to their bounded counterparts were inadvertently retained in the reported tables. This was an oversight in reporting on our part, and we thank the reviewer for highlighting this inconsistency.

In the revised manuscript, we have carefully checked and corrected all Appendix parameter tables so that the reported values now correspond to the bounded parameter values actually used in the ILS simulations. We have also added an explicit note in the table description to clarify that the values reported in the Appendix are the constrained values used in the ILS runs.

Reviewer Comment:

Several “calibrated parameters” appear to be compensatory knobs rather than plausible PFT traits. Please discuss equifinality/identifiability: many optimised parameters hit bounds or take implausible values (e.g., forest height at 0.5 m), suggesting compensation for structural model–data mismatch. Consider constraining structural traits with independent information (site metadata/remote sensing) or excluding them from calibration when they are not identifiable from H/LE alone.

For example, Evergreen forest PT-Mi1: vegh = 0.5 m (lower bound) — very unlikely for a forest canopy. Evergreen forest PT-Esp: vgcov = 0.11 — implies ~11% vegetation cover for a forest PFT; again suspicious. Multiple sites show parameters sticking to bounds (e.g., chl = 0.01; rlfv = 0.05; gradm = 1).

When structural parameters (height, cover, optical traits) are used to match only H and LE, they can compensate for missing/biased processes (roughness length schemes, LAI/phenology, soil moisture dynamics, energy closure issues), producing “effective parameters” that are not transferable.

Authors’ Response:

We agree that some calibrated parameters, particularly structural parameters such as vegh and vgcov, may take values that are not directly interpretable as realistic PFT traits under the current calibration setup. This behavior is consistent with the issue of equifinality in environmental and land surface modeling, where multiple parameter combinations can produce similar model outputs under limited observational constraints (Beven and Freer, 2001; Beven, 2006).

In this study, parameter calibration is performed using only sensible and latent heat fluxes as target variables. Under such limited observational constraints, the identifiability of certain parameters—especially structural and optical parameters—is inherently limited. Similar identifiability challenges have been reported in land surface modeling studies (e.g., Kuppel et al., 2014; Rosolem et al., 2012).

As a result, some parameters may approach their predefined bounds or take values that are not directly interpretable as realistic physical traits. This reflects compensatory behavior, where parameters adjust to account for structural model–data mismatches or missing processes (e.g., roughness parameterization, phenology, or soil moisture dynamics). Such behavior is commonly observed in environmental model calibration and indicates that parameters may act as effective representations of unresolved processes (Gupta et al., 2008; Vrugt et al., 2005).

Therefore, the calibrated parameters in this study should be interpreted as *effective parameters* under the given model structure and observational constraints, rather than uniquely identifiable physical quantities. We have clarified this point in the revised manuscript and explicitly discussed equifinality and identifiability. See Section 3.1.

We also agree with the reviewer that incorporating independent constraints (e.g., vegetation structural information from site metadata or remote sensing) would improve parameter identifiability and physical interpretability. This represents an important direction for future work. In the current study, however, our primary objective is to evaluate the effectiveness of the differentiable calibration framework in improving model performance, rather than retrieving physically unique parameter values.

Reviewer Comment:

The “critical water potential” parameter range/meaning is unclear and may be physically inconsistent. You define psicr as “Critical water potential” with range [−500, −50] kPa, but is this soil matric potential threshold? leaf water potential? root-zone? The magnitude (−0.05 to −0.5 MPa) may be too low for many stomatal regulation

thresholds (depending on the definition), and the paper doesn't provide the governing equation in ILS for how psicr controls stress.

Please define psicr precisely (soil vs plant potential, sign convention, and where it enters the stomatal/water-stress formulation) and justify the chosen range for the modelled processes. Without this, it's hard to assess whether optimised values are physiologically meaningful.

Authors' Response:

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

In the model implementation, psicr is expressed in units of kPa, following the conventional sign convention in soil physics, where negative values represent suction. Within the code, the parameter is internally scaled (×100) and enters a logistic-type function that regulates root water uptake, resulting in a smooth transition between unstressed and stressed conditions.

Therefore, psicr should be interpreted as an effective plant-level threshold parameter governing the onset of water stress, rather than a direct representation of a specific physiological variable such as leaf water potential or soil matric potential alone. It links plant response to soil moisture conditions through the model formulation.

Regarding the selected range (−500 to −50 kPa), we acknowledge that this range may not fully represent the diversity of physiological thresholds across vegetation types. However, the chosen bounds are consistent with typical ranges of soil water potentials influencing plant water availability in land surface models. In addition, this range is consistent with that adopted in our previous study on parameter calibration in land surface modeling (Li et al., 2025), where similar parameter bounds were used and demonstrated to provide stable and physically reasonable simulations of evapotranspiration and gross primary productivity.

Under the current calibration setup, psicr functions as an effective parameter controlling water stress response. We have clarified the definition of psicr, its role in the model formulation, its unit and sign convention, and its implementation in the water-stress function in the revised manuscript to improve transparency.

Reviewer Comment:

The surrogate training design likely under-samples parameter interactions, risking incorrect gradients. The surrogate's training data should include joint sampling of parameter space (e.g., Latin hypercube / Sobol / random combinations), and the paper should quantify surrogate fidelity on held-out parameter combinations. Otherwise, the backpropagated gradients used for "calibration" may be inaccurate.

Authors' Response:

We agree that sufficient coverage of parameter interactions is important for ensuring reliable surrogate-based gradients.

To address this concern, we have explicitly designed a surrogate–ILS consistency evaluation based on joint parameter perturbations (Appendix A, Figure A-6). In this experiment, one representative site was selected from each PFT (four sites in total), and for each site, 100 parameter sets were generated by simultaneously perturbing all calibration parameters.

The perturbation strategy was centered around the default parameter values of each PFT. Each parameter was sampled within a predefined bounded range using relative perturbations. Importantly, all parameters were perturbed jointly rather than individually, enabling the exploration of parameter interactions across the multi-dimensional parameter space.

For each sampled parameter set, simulations were performed using both the original ILS and the trained surrogate model. The results (Figure A-6) show strong agreement between surrogate and ILS outputs across all sites, with high correlation coefficients (generally $R > 0.97$) and low RMSE values.

This consistency under joint perturbations indicates that the surrogate model provides a reliable approximation of the ILS behavior within the explored parameter space, including regions where multiple parameters interact. Therefore, the gradients propagated through the surrogate are expected to be meaningful for calibration within this parameter domain.

Reviewer Comment:

Clarify whether g_z outputs a single constant parameter vector per site or time-varying parameters per window, and how the final parameter vector used in ILS runs is derived; If parameters vary in time, please rename them as "effective parameters" and discuss the implications for physical interpretability and transferability.

Authors' Response:

In the MdPL framework, g_z is implemented using a recurrent neural network (RNN), which produces a sequence of parameter estimates over time (i.e., one parameter vector per time window). These intermediate outputs can be interpreted as time-dependent effective parameters reflecting temporal variability in the input data.

However, the final parameter set used for ILS simulations is not time-varying. Instead, a single parameter vector is derived for each site by averaging the predicted parameter sequences over the full time period. This averaged parameter vector is then used in all ILS simulations.

Therefore, while the intermediate outputs of g_z are time-dependent, the final calibrated parameters applied in ILS are time-invariant and site-specific. We have clarified this procedure in the revised manuscript to avoid ambiguity.

We also acknowledge that the temporal variability in the intermediate parameter estimates may have implications for physical interpretability and transferability. In this sense, these parameters can be viewed as effective parameters during training, while the final averaged parameters represent a practical compromise between flexibility and interpretability. A more explicit treatment of time-varying parameters is an interesting direction for future work.

Reviewer Comment:

Calibration target is only (H, LE); energy-balance realism is not enforced. The authors acknowledge calibration "primarily targeted energy fluxes" and did not evaluate surface temperature or soil moisture. Without checking (or penalising) the full surface energy balance and related states, tuned parameters can improve H/LE while worsening Rn partitioning, canopy/soil temperature dynamics, or water balance.

Authors' Response:

In this study, the calibration objective focuses on H and LE, which are the most widely available and reliable flux observations in the FLUXNET2015 dataset. We acknowledge that other components of the surface energy and water balance (e.g., net radiation partitioning, surface temperature, and soil moisture) are not explicitly constrained in the loss function due to the lack of consistent observational data across sites.

We agree that calibrating only H and LE may potentially introduce compensatory effects in other model states. To address this concern, we conducted an additional out-of-target evaluation using gross primary productivity (GPP), which is not included in the calibration targets but is closely linked to surface energy and carbon processes.

The results (Section 3.5) show that calibration based on H and LE does not lead to systematic degradation in GPP simulations. While improvements are site-dependent, the absence of widespread deterioration suggests that the MdPL framework maintains reasonable consistency for non-target variables.

We acknowledge that a more comprehensive evaluation including surface temperature, soil moisture, and full energy balance closure would further strengthen the physical consistency of the calibration. However, such analysis is currently limited by data availability. Incorporating multi-variable constraints represents an important direction for future work.

- ***End of response to Reviewer #3***

References

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.

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.

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.

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.

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.

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.