Reviewer #1

Reviewer Comment:

The paper develops a multi-task differentiable parameter learning (MdPL) framework as a unified calibration approach for land surface models (LSMs). It builds on a previous deep learning scheme, dPL, that directly embeds parameter estimation within model training, and improves its design while also enabling multi-output calibration, with latent heat flux and sensible heat flux as the outputs.

MdPL consists of a multi-task surrogate model of LSM and a parameter generator. It was used in this study to calibrate the Integrated Land Simulator (ILS) using site observations representing four functional types. Different experiments were used to assess the performance of the calibrated MdPL-calibrated ILS and compare with the original out-of-sample observation, dPL simulations, ILS simulations, and multiple LSMs.

Authors' Response:

We appreciate the reviewer's clear summary of the manuscript and the key contributions of this work. In the revised manuscript, we have further clarified the structure of the MdPL framework and its experimental design to improve readability.

Reviewer Comment:

I support this work as an important step in exploring how AI can improve LSMs. However, I have three major comments that may enhance the reliability and clarity of the performance conclusions.

- First, the study uses only a limited number of sites spanning a few years, which weakens the strength of the conclusions. The PLUMBER2 dataset includes 170 sites, yet only 20 were used here across four PFTs, with each PFT represented by only a handful of sites, and in some cases results are shown for a single site per PFT.
- Second, the model was calibrated and evaluated using latent and sensible heat flux, but the authors do not discuss other fluxes simulated by LSMs that were not included in the calibration, such as NEE. For example, does the MdPL-ILS improve the simulation of NEE? Showing an example of how the model performs on a new flux like NEE would also be interesting.
- Third, while the authors provide estimated parameters for sites across different PFTs in the appendix, there is no interpretation of whether the new values make physical sense more than the original values. Even if a full explanation is not possible, some discussion would help demonstrate that MdPL captures site relationships in a meaningful way.

Authors' Response:

We thank the reviewer for these important comments. We address each point below.

(1) Site number

We acknowledge that the use of 20 sites limits the scope of the conclusions and that the revised manuscript should more clearly distinguish between proof-of-concept evaluation and broad generalization across the full PLUMBER2 archive.

In this study, the 20 sites were selected to balance ecosystem diversity, data completeness, and computational feasibility within the MdPL. Our intention was to establish a controlled multi-site proof-

of-concept for applying MdPL to a complex land surface model, rather than to provide an exhaustive benchmark across all PLUMBER2 sites.

To address this concern, we have revised the manuscript in three ways. First, we added a clearer explanation of the site selection rationale in Section 2.4.1. Second, we explicitly stated the limited scope of the current dataset and framed the study as a proof-of-concept evaluation. Third, we moderated several statements regarding robustness and generalizability in the Results (Section 3.1) and Conclusion sections. We also note in the revised manuscript that broader validation across more PLUMBER2 sites remains an important direction for future work.

(2) NEE / other flux

We agree that evaluating only sensible and latent heat fluxes is not sufficient to assess whether the MdPL-calibrated model preserves performance for other simulated variables.

In principle, net ecosystem exchange (NEE) would be an important variable for such an out-of-target evaluation. However, NEE is not directly available in the current ILS configuration used in this study. To address this concern, we have added an additional evaluation using gross primary productivity (GPP), which is available from both model outputs and PLUMBER2 observations.

Specifically, we introduced a new experiment (Section 2.4.6) and a corresponding Results subsection (Section 3.5), in which one representative site was selected for each PFT category. The parameter sets calibrated using Qh and Qle were applied to simulate GPP, and the results were compared against observations and simulations using the default parameter configuration (ILS_ORI).

The results show that the impact of MdPL calibration on GPP is site-dependent: improvements are observed at some sites, while performance remains similar or slightly degraded at others. Importantly, no systematic degradation in GPP is observed across all sites. This suggests that calibration based on energy fluxes does not compromise the simulation of a non-target carbon-related variable.

We also clarify in the revised manuscript that this GPP-based evaluation does not fully replace a direct assessment using NEE, but provides an additional and independent check on model robustness beyond the calibration targets.

(3) Parameter interpretability

In the revised manuscript, we have added a short discussion at the end of Section 3.1 to clarify that the calibrated parameters should not be interpreted strictly as physically optimal values, but may partly act as effective parameters due to model complexity and interactions among processes. We also note that some consistency can be observed within PFT groups, suggesting that the MdPL framework captures certain site-level characteristics.

In addition, we included a brief remark in Section 3.5 to relate parameter behavior to transferability results, further supporting that the calibrated parameters reflect site differences to some extent. A more detailed physical interpretation of individual parameters is beyond the scope of this study and is left for future work.

Reviewer Comment:

"Exceeded those of LSTM-based approaches": it is actually a single LSTM approach, not multiple.

Authors' Response:

We have corrected the wording in the revised manuscript. Please check Line 19.

Reviewer Comment:

Line 25: The statement “close-to-optimal” overestimates the performance, since the KGE scores were far from (the optimal value).

Authors’ Response:

The term “close-to-optimal” has been revised to avoid overstating the model performance. We have revised the wording to “reasonable transferability” to provide a more appropriate description. Please check Line 23.

Reviewer Comment:

Line 27: Replace “Despite its” with “Despite the”.

Authors’ Response:

We have revised the wording accordingly. Please check Line 22-24.

Reviewer Comment:

Line 51: “However, it is associated with several limitations, particularly in the context of LSMs”: the statement neither has a reference nor is it explained.

Authors’ Response:

We have revised the sentence to explicitly state the key limitations and included appropriate references for clarification. Please see Line 52-54.

Reviewer Comment:

Line 103: PLUMBER2 should be introduced when first mentioned.

Authors’ Response:

We have introduced PLUMBER2 with its full name at first mention in the revised manuscript. Please check Line 105.

Reviewer Comment:

Line 103: A reference for ILS is needed.

Authors’ Response:

A reference for the ILS has been added in the revised manuscript. Please check in Line 104.

Reviewer Comment:

Line 104: A reference is needed for each of these LSMs, and the acronyms should be expanded on first mention.

Authors’ Response:

We have expanded all model acronyms at first mention and added appropriate references for each LSM in the revised manuscript. Please check Line 107-109.

Reviewer Comment:

Lines 91–112: The introduction section includes methodological details and results that are better suited to the Methods or Results sections. For example, you mention the role of gate layers (a methodological detail) and how you pretrained the surrogate, and then describe results such as "accurately captured nonlinear and coupled processes". These belong later in the paper. You could reframe these points as objectives instead. For instance: state that the aim is to develop a scalable framework, and that you intend to validate it against observational data.

Authors' Response:

The Introduction has been revised to remove methodological details and result-oriented statements, focusing instead on the study objectives.

In the revised manuscript, we have rephrased these parts to focus on the objectives and design of the proposed framework, rather than implementation details or results.

Reviewer Comment:

The way the experiments are explained is a bit confusing. It is worth using subheadings for Experiment 1, Experiment 2, and Experiment 3 and revising the text. Also, some helpful details are missing. For each experiment, indicate how many models you will train and how many simulations each model will generate. Later, in the results, state explicitly that section 3.1 is the results of Experiment 1 and section 3.2 is the results of Experiment 2…

Authors' Response:

The experiment section has been reorganized to improve clarity and structure. We reorganized the experiment section by separating the original Section 2.4.3 into multiple subsections (Sections 2.4.3–2.4.7), each corresponding to a specific experiment. We also explicitly state the number of trained models and clarify the simulation procedures for each experiment.

In addition, we revised the Results section by streamlining the descriptions and indicating, at the beginning of each subsection, the corresponding experiment (e.g., Section 3.1 for Experiment 1 and Section 3.2 for Experiment 2).

Reviewer Comment:

It would be useful to clarify the differentiable aspect of the parameter generator g_z

Authors' Response:

We have added a clarification on the differentiable aspect of the parameter generator g_z in Section 2.1 (Lines 150–153).

Reviewer Comment:

In Figure 1 (b), is the arrow connecting Parameters and Day of Year Embedding correct? Shouldn't the parameters go directly to the tensor concatenation symbol?

Authors' Response:

Thank you for pointing this out. In the revised version of Figure 1(b), we have clarified the data flow. Specifically, the calibration parameters and DOY embedding are first concatenated to form a combined

representation. This combined feature is then concatenated with the meteorological forcing inputs before being fed into the shared layers.

Reviewer Comment:

Line 125: The subject of Eq.1 (left-hand side) is missing. Is this the loss function of the surrogate model? This needs to be mentioned before displaying the equation.

Authors' Response:

We have clarified that Eq. (1) defines the loss function of the surrogate model and added a corresponding description before the equation.

Reviewer Comment:

Line 139: Again, Eq. 2 is missing the first part. Is this the cost function?

Authors' Response:

We have clarified that Eq. (2) defines the objective function for training the parameter generator and added a corresponding description before the equation.

Reviewer Comment:

In this part, you describe the multi-task surrogate model in 1b and wrote: "Further, differences in scale and variation frequency between static and dynamic features…" Do the Parameters include static ones?

Authors' Response:

We apologize for the ambiguity. No, the "Parameters" do not include static features. In this study, "Parameters" refer to the calibration parameters of the LSM, whereas the static features correspond to A, i.e., site-level attributes used as inputs to the parameter generator. We have clarified this in Lines 141–144.

Reviewer Comment:

Line 193: Better to omit natural and keep only variability.

Authors' Response:

We have removed the word "natural". Please check Line 205.

Reviewer Comment:

Line 202-203: What is the baseline model? Is it the ILS or the dPL-ILS?

Authors' Response:

This has now been clarified in the revised manuscript. The baseline in NSE refers to the mean of the observed data. Please check 211.

Reviewer Comment:

Line 216: Is averaging correct here, or do you mean including?

Authors' Response:

It means including, we have revised "averaging" to "with approximately 3 years of continuous data" in Line 234.

Reviewer Comment:

An evaluation of the surrogate model against ILS (as described in Figure A-1) is important to demonstrate that the surrogate closely approximates the ILS model itself.

Authors' Response:

A surrogate–ILS consistency evaluation has been added in Appendix A to demonstrate that the surrogate closely approximates the original ILS. In this experiment, one representative site from each PFT was selected, and 100 jointly perturbed parameter sets were generated per site. The results show strong agreement between the surrogate and ILS outputs across both sensible and latent heat fluxes, demonstrating that the surrogate closely approximates the ILS behavior.

Reviewer Comment:

Figure 3: Could you explain how ΔR^2 and ΔRMSE are calculated? For instance, here ΔRMSE = 24%. This should indicate that IL_MdPL is worse, since RMSE (error) has increased. This is different from ΔR^2, where larger values indicate improvement in skill. Is R^2 the proportion of variance explained, or is it the relative difference in standard deviation (the x-axis)? The Taylor diagram shows results for RMSE, correlation, and standard deviation. In the results you discuss R and RMSE, but not standard deviation.

Authors' Response:

The definitions of the relative performance metrics have been clarified in the revised manuscript. We have clarified the definitions of ΔRMSE and ΔR in Section 2.3. ΔRMSE is defined as (RMSE_ORI − RMSE_MdPL) / RMSE_ORI, such that positive values indicate a reduction in error. Similarly, ΔR is defined as the relative change in the Pearson correlation coefficient, such that positive values indicate improved correlation.

We corrected the notation in Figure 3 from ΔR^2 to ΔR. The values were in fact calculated using R, but were mistakenly labeled as R^2. The caption and related text have been updated for consistency with the Taylor diagram (RMSE, R, and standard deviation).

Reviewer Comment:

Line 317: Explicitly state that these results are for sensible heat flux.

Authors' Response:

We have revised the text to explicitly state that the results shown in Fig. 3(a) correspond to sensible heat flux. Please check Line 358.

Reviewer Comment:

Figure 3: There is a typo in the title of some subplots (Latnet).

Authors' Response:

We thank the reviewer for pointing this out. The typo has been corrected.

Reviewer Comment:

Line 324-353: The paragraph mentions R, but it seems you are referring to the normalised standard deviation. In the first point explaining the reduction in performance related to R, could you clarify what you mean?

Authors' Response:

We apologize for the confusion caused by the original wording. The percentage values refer to ΔR (i.e., the relative change in the correlation coefficient), rather than the correlation coefficient itself or the normalized standard deviation.

We have revised the text to explicitly refer to ΔR and clarified the descriptions to ensure consistency between the correlation metric and the Taylor diagram components.

Reviewer Comment:

Throughout the results, the use of significantly is inconsistent. For example, in line 328 you mention decreased significantly (-4.77%), whereas in line 326 you say only showed minor reduction (8.74%), despite |8.74| being larger than |4.77|.

Authors' Response:

Qualitative terms such as "significantly" and "minor" have been removed to ensure consistency with the reported numerical values. Check Line 366-368.

Reviewer Comment:

Figure 4: This shows that ILS_MdPL is better for sensible heat but worse than ILS_dPL for latent heat. The conclusion that "it exhibited good scalability and could adapt to multiple variable requirements… high potential for application in parameter calibration in complex LSMs" is not strongly demonstrated.

Authors' Response:

The original summary overstated the advantages of the multiple-task model. This has been revised to clarify that its performance relative to ILS_dPL depends on the evaluation metric and is not consistently superior across all cases. Please 419-428.

Reviewer Comment:

Line 391: "In summary, all the evaluation metrics and time resolutions consistently demonstrated the effectiveness and robustness of the proposed MdPL." You state all the evaluation metrics, but you only show one. To claim robustness, you need to demonstrate this across multiple metrics. A supplementary plot for other metrics (e.g. R, NSE, or NGE) would be needed to support this claim.

Authors' Response:

We agree that the statement referring to "all evaluation metrics" was not sufficiently supported. In the revised manuscript, we have clarified that multiple metrics were considered in the analysis and have revised the summary to avoid overgeneralization. Please check Line 443-450.

Reviewer Comment:

Tables 3 and 4: Please add vertical lines to improve readability. For example, in Table 4, you could place a line separating latent heat and sensible heat.

Authors' Response:

Thank you for the suggestion. Vertical lines have been added to Tables 3 and 4 to improve readability.

Reviewer Comment:

Figure 6: In the caption, it is clearer to state that these are monthly time-series of sensible heat (top) and latent heat (bottom) simulated by the models, with KGE provided. Why are evaluation results shown for a very short record (one year or less) for the first two sites? Why not use sites with longer coverage?

Authors' Response:

We have revised the caption and text of Figure 6 to clarify that it shows monthly time series of KGE for sensible heat and latent heat.

Regarding the length of the evaluation period, in the LOOCV experiment, sites with relatively longer and more continuous records were used for training to ensure sufficient information for parameter learning, while the remaining sites were used for validation.

This design allows us to evaluate the transferability of calibrated parameters under realistic data constraints, rather than selecting sites based on performance.

Reviewer Comment:

Line 424–426: "with LOOCV showing a narrow gap between observations" does not appear correct for Evergreen Forest or Woodland, for either sensible heat flux or latent heat flux.

Authors' Response:

We agree that the original description was not sufficiently accurate. The statement has been revised to avoid implying a close agreement with observations and now reflects that LOOCV shows intermediate performance between ILS_ORI and ILS_MdPL. Please see Line 500-507.

- ***End of response to Reviewer #1***