Reviewer #2

Reviewer Comment:

The manuscript entitled *"Enhancing Parameter Calibration in Land Surface Models Using a Multi-Task Surrogate Model within a Differentiable Parameter Learning Framework"* aimed to calibrate parameters of the ILS model by leveraging the differentiability of neural networks. The study demonstrates that the proposed multi-task differentiable parameter learning (MdPL) framework achieves better sensible and latent heat simulation performance than the default parameter set and outperforms the single-task version.

Authors' Response:

We thank the reviewer for the clear summary of our work and for recognizing the contributions of the proposed MdPL framework.

Reviewer Comment:

While the topic is relevant and the approach technically interesting, I have serious concerns about the validity and scientific contribution of the study in its current form. Specifically, there are two major issues that critically affect the plausibility and impact of the work:

1. **Lack of direct physical connection between calibration and model parameters:**
   The ILS model parameters were not directly optimized through differentiable learning. Instead, the learning process was conducted through a surrogate model. This inevitably introduces fitting errors and weakens the physical interpretability of the results. The surrogate model may not faithfully represent the true relationships between model parameters and physical processes within the ILS framework.

2. **Insufficient evaluation and benchmarking:**
   The study assesses calibrated model outputs only against simulations using the default parameter set. Without comparison against other calibration methods (e.g., PFT-specific parameter optimization and etc), it is difficult to judge the value or robustness of the proposed approach. Given the many existing parameter calibration techniques, it is essential to demonstrate that MdPL performs competitively or superiorly to conventional parameter calibration methods.

These two issues fundamentally limit the scientific credibility and general applicability of the work. Nonetheless, the authors' efforts are appreciated, and I encourage them to consider these points in future work to strengthen the study's methodological and physical rigor.

Authors' Response:

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

(1) Lack of direct physical connection between calibration and model parameters

We acknowledge that the use of a surrogate model introduces an additional approximation layer, which may affect the direct physical interpretability of the calibration process.

To address this concern, we have added a surrogate–ILS consistency evaluation in Appendix A (Figure A-6), demonstrating that the surrogate model closely approximates the behavior of the original ILS across different PFTs and under jointly perturbed parameter settings. The results show strong agreement between the surrogate and ILS outputs, with high correlation coefficients (generally $R > 0.97$) and low

RMSE values, indicating that the surrogate preserves the key input–output relationships of the physical model.

In addition, it is important to note that the calibrated parameters are ultimately evaluated using the original ILS, rather than the surrogate model itself. Therefore, the final performance improvements are achieved within the physical model framework.

Regarding parameter interpretability, we have added a discussion in Section 3.1 to clarify that the calibrated parameters should not be interpreted strictly as physically optimal values, but may act as effective parameters due to model complexity and interactions among processes. Nevertheless, some consistency is observed within PFT groups, suggesting that the MdPL framework captures meaningful site-level characteristics.

(2) Insufficient evaluation and benchmarking

We acknowledge that the original manuscript did not include sufficient benchmarking against conventional calibration approaches.

To address this, we have added an additional experiment based on a traditional calibration baseline using Sobol-based random sampling, as described in Section 2.4.5. The corresponding results are presented in Section 3.3. While this comparison is conducted for a representative site due to computational constraints, it serves as a proof-of-concept demonstration of the relative performance and efficiency of MdPL.

In this experiment, 256 parameter sets were generated through joint perturbations across all calibration parameters for a representative site (CN-Din), and model performance was evaluated using a unified objective function (RMSE(Qh) + RMSE(Qle)).

The results show that the MdPL-calibrated parameters achieve better performance than the best parameter set identified through random sampling. This indicates that the proposed framework can effectively explore the parameter space and identify high-quality solutions without exhaustive sampling.

These additions provide a direct comparison with a conventional calibration approach and strengthen the evaluation of the proposed method.

Reviewer Comment:

Introduction: The literature review on differentiable parameter learning is incomplete. Please include more relevant studies (e.g., Bao et al., JAMES, 2023) to better contextualize the contribution.

Authors' Response:

We have expanded the Introduction to include additional relevant studies on differentiable parameter learning, including Bao et al. (JAMES, 2023), to better contextualize the contribution of this work. (in Line 66)

Reviewer Comment:

Line 110: The manuscript claims to mitigate the impact of sparse and noisy observational data, but this is not clearly demonstrated. Please elaborate or revise this claim.

Authors' Response:

The statement has been revised to avoid overclaiming, and we now describe the framework as potentially improving robustness under sparse and noisy observational conditions. (Line 116)

Reviewer Comment:

Lines 112–114: The notation for section references (“Section 2,” “Sect. 3,” “Sect. 4”) should be unified.

Authors’ Response:

The notation has been unified to “Section” throughout the manuscript. Please check Line 120.

Reviewer Comment:

Figure 1: This figure should appear after the corresponding description. As currently presented, it is difficult to interpret without prior explanation of the framework.

Authors’ Response:

Figure 1 has been moved to appear after the corresponding description of the framework.

Reviewer Comment:

Line 158: Please define the symbol ‘L’

Authors’ Response:

The symbol *L* has now been defined in the revised manuscript as the number of encoding frequencies used in the temporal positional encoding. See Line 172.

Reviewer Comment:

Section 2.3: Consider summarizing the evaluation metrics in a concise table instead of repeating similar textual descriptions.

Authors’ Response:

Section 2.3 has been revised to improve conciseness by reducing repetitive descriptions of the evaluation metrics, while retaining their definitions and formulations for clarity.

Reviewer Comment:

Table 1: Add horizontal lines to clearly separate plant functional types.

Authors’ Response:

We thank the reviewer for this suggestion. Horizontal lines have been added to Table 1.

Reviewer Comment:

Line 225: The term “Mediterranean climate” is more appropriate than “subtropical climate” for ‘Csa’ and ‘Csb’.

Authors’ Response:

The term “subtropical climate” has been corrected to “Mediterranean climate”. Please check Line 243.

Reviewer Comment:

Line 230: Provide details on parameter sampling: Was it random? How many samples were drawn per range? Did the authors account for potential nonlinear parameter–output relationships (e.g., exponential)?

Authors' Response:

We thank the reviewer for this comment. In this study, parameters were sampled deterministically using four uniformly spaced values within the predefined ranges listed in Table 2, rather than random sampling. This results in four simulations per parameter in the one-factor-at-a-time sensitivity analysis.

We note that this design focuses on first-order sensitivity and does not explicitly capture nonlinear parameter interactions. To address this, we refer to the joint parameter perturbation experiments presented in Appendix A, where multiple parameters are varied simultaneously, providing additional insight into nonlinear and coupled effects.

Reviewer Comment:

Table 2: Explain how the key parameters influence model outputs, and consider providing response curves to illustrate these relationships.

Authors' Response:

Thank you for this helpful suggestion. The influence of key parameters on model outputs is already reflected in the one-factor-at-a-time sensitivity analysis presented in Figure 2, where the relative changes in RMSE of sensible and latent heat fluxes are shown for each parameter.

This analysis provides a quantitative indication of parameter importance and their effects on model outputs. Therefore, we have clarified this point in the manuscript, rather than introducing additional response curves or modifying Table 2.

Reviewer Comment:

Line 241: Justify the use of only one pre-training dataset and provide its distribution in a figure.

Authors' Response:

We thank the reviewer for this important comment.

In this study, the pre-training dataset was constructed using a one-factor-at-a-time perturbation strategy. Specifically, for each site, 12 calibration parameters were individually perturbed at four uniformly spaced levels within their predefined ranges (Table 2), resulting in 48 simulations per site. Across all 20 sites, this yields a dataset consisting of nearly 1000 simulations.

Therefore, the dataset does not correspond to a single configuration, but rather a structured sampling of the parameter space capturing first-order parameter effects.

We acknowledge that this design does not fully represent all joint parameter combinations. However, this dataset is intended to provide sufficient coverage of parameter sensitivities for training the surrogate model. The subsequent optimization in the MdPL framework is not restricted to these sampled points, as parameters are updated continuously through gradient-based learning.

Importantly, the calibrated parameters are ultimately evaluated using the original ILS model rather than the surrogate model. Therefore, the final results are not limited by the distribution of the pre-training dataset.

In addition, the MdPL framework performs joint parameter calibration, allowing interactions among parameters to be implicitly captured during optimization. We have clarified these points in the revised manuscript.

Reviewer Comment:

Line 257: Correct grammatical errors.

Authors' Response:

The grammatical errors have been corrected. Please check Line 283.

Reviewer Comment:

Line 266: Explain the rationale for using different hidden layer sizes in single-task and multi-task surrogate models, or directly use the same size.

Authors' Response:

The hidden sizes were not set identically because the two models have different architectures (standard LSTM vs. shared–branch multi-task structure). Using the same hidden size would lead to substantially different numbers of trainable parameters. Therefore, the hidden sizes were adjusted to ensure comparable model capacity, enabling a fair comparison between the two models (trainable parameters: 1,096,962 vs 980,400).

Reviewer Comment:

Section 2.4.3: The three experiments appear sequential rather than parallel. Consider dividing them into three sub-sections—e.g., (2.4.3) *Comparison between Multi-Task and Single-Task Models*, (2.4.4) *Benchmarking*, and (2.4.5) *Transferability Evaluation*.

Authors' Response:

The experimental design has been reorganized into separate subsections (Sections 2.4.3–2.4.7) to improve clarity and structure.

Reviewer Comment:

Line 300: Clarify why only three plant functional types were evaluated.

Authors' Response:

We thank the reviewer for this comment. Although grassland includes a relatively large number of sites, it was not included in the transferability experiment because its default parameter values are highly similar to those of the cultivation PFT (Table 2).

As a result, the parameter space represented by grassland substantially overlaps with that of cultivation, and including both categories would introduce redundancy rather than additional insight into transferability behavior.

Therefore, we selected three representative PFT categories (evergreen forest, woodland, and cultivation) to cover distinct vegetation types while avoiding redundant parameter configurations.

We have clarified this point in the revised manuscript. Please check Line 331-334.

Reviewer Comment:

Results and Discussion: This section should be streamlined to highlight key findings rather than listing results exhaustively. Focus on the major insights and implications.

Authors' Response:

We thank the reviewer for this helpful suggestion. The Results and Discussion section has been revised to improve clarity and readability by emphasizing key findings and reducing excessive numerical descriptions.

Reviewer Comment:

Line 371: 'LSM'?

Authors' Response:

The "LSM" has already been defined as "land surface model" in the Abstract and Introduction. For clarity, we have ensured that its usage is consistent throughout the manuscript. Please check Line 11 and Line 33.

Reviewer Comment:

Lines 376–379: The text mentions identical RMSEs between LSTM and ILS_MdPL and an increasing difference at larger timescales, but this is not supported by Figure 5. Please check for consistency.

Authors' Response:

We thank the reviewer for this comment. The description has been revised to avoid overstating the trend and to better reflect the results shown in Figure 5. We now describe the performance differences between ILS_MdPL and LSTM as generally small and dependent on temporal resolution, rather than implying a systematic increase.

Reviewer Comment:

Table 4: The PFT-calibrated parameters perform worse than the mean of site-specific parameters, which is counterintuitive (LSTM should be more flexible than mean and should have better performance). Please investigate potential causes (e.g., missing features or model structure issues).

Authors' Response:

The comparison in Table 4 is intended to evaluate parameter transferability rather than model flexibility. The ILS_MdPL_LOOCV represents PFT-level averaged parameters, while ILS_MdPL corresponds to site-specific calibrated parameters.

The observed performance difference is therefore expected, as site-specific calibration can better capture local characteristics, whereas the averaged parameters may lose site-specific information.

The purpose of this experiment is to assess how well calibrated parameters can be transferred across sites within the same PFT. The reduced performance of the LOOCV configuration highlights the limitations of parameter transferability, particularly in heterogeneous environments such as cultivation sites.

Reviewer Comment:

Figure 6: The figure caption does not match the content—it does not directly compare KGE but instead presents temporal LE and H. Please revise accordingly.

Authors' Response:

We agree that the previous caption was misleading. The figure presents time series of sensible and latent heat fluxes, while KGE values are provided within each panel as performance indicators. The caption has been revised accordingly.

- ***End of response to Reviewer #2***