Reviewer #1

Reviewer's general assessment:

This study presents a multi-task differentiable parameter learning (MdPL) framework for calibrating parameters in the Integrated Land Simulator (ILS). The approach first trains a differentiable multi-task surrogate to emulate ILS sensible and latent heat fluxes under different parameter settings, and then uses a parameter-generation network to infer parameter sets constrained by flux observations. The calibrated parameters are subsequently evaluated in the original ILS and compared with the default ILS, five land surface models, an LSTM benchmark, SFE, and a Sobol-based calibration baseline.

Overall, I found the manuscript well written, clearly explained, and well organised. The methodological workflow is generally easy to follow. I have a few recommendations and requests for clarification that I believe would strengthen the manuscript; none of these should be considered major.

A few comments suggest additional parameter diagnostics. These are intended as recommendations, as they could provide useful insight into parameter behaviour and identifiability where the relevant outputs or analyses are readily available.

Authors' Response:

We thank the reviewer for the positive assessment of our manuscript and for the thoughtful comments and constructive suggestions. Below, we respond to each comment and explain the corresponding clarifications, additional analyses, and planned revisions.

Reviewer Comment 1:

Lines 226. The wording that the 20 sites were selected "to ensure geographic and ecosystem diversity" seems somewhat strong given the relatively small number of sites, four broad PFT groups, and uneven geographic representation. I suggest softening this statement.

Authors' Response:

Thank you for this suggestion. We will replace "to ensure geographic and ecosystem diversity" with the following wording:

"We selected 20 flux-tower sites covering four broad PFT groups and a range of geographic locations."

Reviewer Comment 2:

It would be useful to provide a clear list of the meteorological forcing variables used by the surrogate and g_z, together with their units. I also noticed that the flux observations are described as half-hourly, whereas 48 time steps are said to correspond to two days in line 326. At half-hourly resolution, 48 steps correspond to one day.

Authors' Response:

Thank you for pointing out the missing input. **Table R1-1** lists the 11 time-varying inputs used by the surrogate and g_z, together with their field names and physical units before standardization.

The parameter-generation network g_z additionally receives site attributes and observed Qh and Qle. We will include the input table in the revised manuscript.

At half-hourly resolution, 48 time steps correspond to one day, whereas 96 time steps correspond to two days. We will correct the statement in line 326 and specify the sequence lengths according to the actual model configurations.

Table R1-1 Time-varying inputs used by the surrogate and g_z.

Input variable	Field name	Unit
Cloud cover	ERA5_ccover	1 (fraction, 0–1)
Downward longwave radiation	LWdown	W m^{-2}
Surface pressure	Psurf	Pa
Specific humidity	Qair	kg kg^{-1}
Precipitation rate	Precip	kg m^{-2} s^{-1}
Downward shortwave radiation	SWdown	W m^{-2}
Air temperature	Tair	K
Wind speed	Wind	m s^{-1}
Leaf area index	LAI	m^2 m^{-2}
Black carbon deposition flux	MIROC6_tdftbc	kg m^{-2} s^{-1}
Dust deposition flux	MIROC6_tdftdu	kg m^{-2} s^{-1}

Reviewer Comment 3:

Lines 188–195. The motivation for including day-of-year (DOY) to represent seasonal variability is clear. It would nevertheless be helpful to clarify why this particular multi-frequency representation was chosen, particularly the choice f_l = 2^l, and whether its benefit over a simpler annual or seasonal representation was assessed.

Authors' Response:

Thank you for this helpful comment. Although the formula allows multiple frequencies through f_l=2^l, we used (L=1), giving (f_0=1).

We will revise the references to "high-dimensional" encoding and "multiscale temporal periodicity" accordingly.

Reviewer Comment 4:

(1)–(2), around lines 145–173. For completeness in Section 2.2, please clarify how the errors associated with Q_h and Q_le are combined during surrogate training and parameter calibration, including whether separate task losses are calculated and whether any relative weighting is applied. The current equations use a generic loss L, but the balance between the two target fluxes is not explicit.

Authors' Response:

Thank you for pointing out that the treatment of the two flux losses was insufficiently explicit. Separate MSEs are calculated for Qh and Qle, using standardized flux values.

During surrogate training, the two task losses are used sequentially within each batch to update the shared layers and the corresponding task branches. No additional task-specific weighting coefficients are applied.

During parameter calibration, the loss for Qh and Qle is the sum of the two standardized MSEs. Thus, the two standardized MSE terms are added directly, without assigning different coefficients to Qh and Qle. We will make these definitions explicit in Section 2.2.

Reviewer Comment 5:

Lines 320–335. Please clarify the temporal partitioning used for surrogate training, g_z training, validation or model selection, and final evaluation. Additionally, it would be useful to state whether the means and standard deviations used for standardisation were estimated using training data only.

Authors' Response:

Thank you for this comment. In the original experiments, surrogate training, parameter calibration, and evaluation used the same analysis period. We will explicitly describe the original results as ILS performance during the training period.

We have completed an additional experiment at CN-Din, DE-Meh, IE-Dri and JP-SMF, using the first 80% of each time series for parameter-generator training and the final 20% for testing. These four-site experiments provide an initial out-of-sample evaluation. We plan to extend the same temporal hold-out evaluation to all 20 sites in the revised manuscript. Table R1-2 shows the training and test periods.

Table R1-2. Training and test periods for the four-site experiment with a chronological 80/20 split.

Site	**Training period**	**Test period**
CN-Din	2003-01-01–2005-05-26	2005-05-26–2005-12-31
DE-Meh	2004-01-01–2006-05-26	2006-05-26–2006-12-31
IE-Dri	2003-01-01–2005-05-26	2005-05-26–2005-12-31
JP-SMF	2003-01-01–2006-03-14	2006-03-14–2006-12-31

The means and standard deviations used for standardisation were calculated using the training set only. The results are presented in **Table R2-1** and **Figure R2-3**; the training runs and parameter-set selection are described in our response to **Reviewer 2, Comment 16.**

Reviewer Comment 6:

Lines 240–245 and 334–335. The practical pathway from site-level calibration to gridded LSM application could be discussed more explicitly. Operational LSMs require parameters at every grid cell, including locations without flux observations. The leave-one-site-out analysis is a useful initial test of transferability, but the main analysis trains a site-specific MdPL model and subsequently excludes sites where the surrogate is insufficiently reliable. Please clarify how the authors envision g_z being applied to unseen grid cells with different meteorological and land-surface characteristics, and what additional spatial validation would be required before regional or global deployment. This would help distinguish generalisation within the calibrated parameter space from spatial generalisation to new environments.

Authors' Response:

Thank you for this suggestion. We will move the leave-one-site-out results into the main text.

In this analysis, the calibrated parameters from other sites are averaged and used to run ILS at the target site, providing an initial evaluation of how these parameters perform at another site.

For future application to grid cells without flux observations, a shared parameter-generation network could be trained jointly across multiple sites. Meteorological forcing and land-surface attributes would serve as its inputs, while flux observations at the training sites would constrain parameter learning. The trained network could then generate parameters for new grid cells from their meteorological and land-surface characteristics. We will explain this future development pathway in the Discussion.

Reviewer Comment 7:

Lines 320–321 and Fig. 4. The parameter generator produces time-dependent estimates, which are then averaged over the entire training period before being used in ILS. It would be informative to provide some insight into the temporal distributions before averaging. This does not need to be done for every site; a few representative sites would be sufficient to show whether the estimates are relatively stable or exhibit systematic seasonal variation, and whether some parameters repeatedly approach their prescribed bounds at particular times. Such an analysis could help establish whether averaging mainly removes noise or suppresses meaningful temporal structure.

Authors' Response:

Thank you for this helpful suggestion. We have examined the saved outputs of the parameter-generation network before temporal averaging. **Figure R1-1** presents examples from DE-Seh, IT-CA2, and JP-SMF. The figure

shows the transformed parameter values used by the network, together with the parameter bounds after the same transformation.

The distributions show that the amount and pattern of variation differ among parameters and sites. For example, *rlfv* at DE-Seh is relatively stable and remains within its prescribed range, whereas *vgcov* and *vmax0* vary substantially over time. The JP-SMF outputs also show recurring patterns within each year. Some estimates repeatedly cross the prescribed bounds: 51.6% of the *chl* estimates before averaging at DE-Seh are below the lower bound, while *rlfv* at JP-SMF remains below that bound throughout the saved record.

These results indicate that averaging reduces both systematic changes over time and short-term fluctuations. In this study, averaging is used to obtain one effective parameter set per site, which is held fixed during the ILS simulation. We will clarify this purpose and discuss the temporal structure revealed by the diagnostic. Developing a dynamic parameterization, and assessing its effects in ILS, remains a direction for future work.

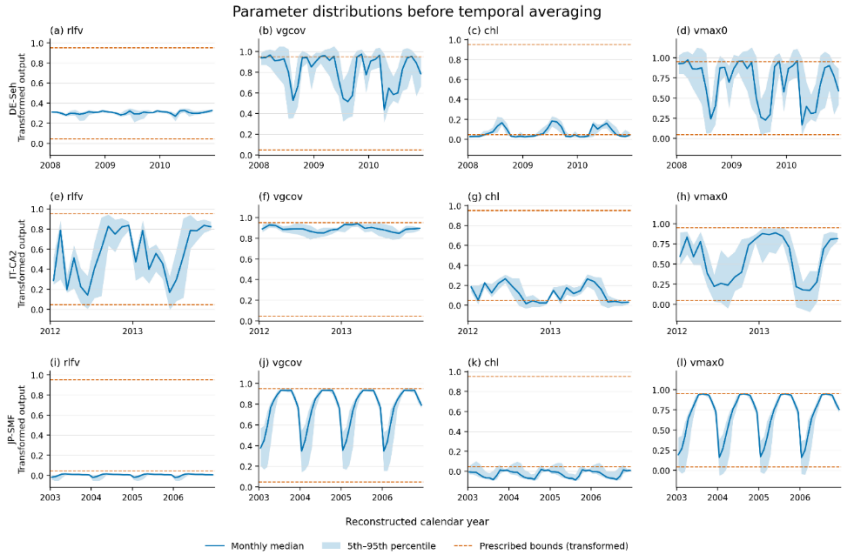


Figure R1-1. Temporal variation in parameter-generator outputs before averaging at three representative sites. Solid lines represent monthly medians, shaded areas indicate the 5th–95th percentile ranges within each month, and dashed lines mark the prescribed parameter bounds. The figure shows the network's transformed parameter values and the corresponding transformed bounds.

Reviewer Comment 8:

Lines 447–461. Given the discussion of equifinality and parameter identifiability, it would be interesting to know whether calibration against Q_h alone or Q_le alone was explored and how the resulting parameter sets differ. Such an analysis could help identify which parameters are primarily constrained by each flux and illustrate the additional information gained by joint calibration. I view this as an additional diagnostic rather than a necessary new experiment. If these results already exist, they would be valuable to report; otherwise, the question could simply be noted as an interesting direction for future work.

Authors' Response:

Thank you for this suggestion. Separate Qh-only and Qle-only parameter-calibration experiments were not conducted in this study. We will identify a controlled comparison of calibration using Qh alone, Qle alone, and both fluxes together as a future direction for determining which parameters are constrained by each flux and how those constraints complement one another.

Reviewer Comment 9:

Lines 370–373 and 545–560. Evaluating GPP outside the calibration is useful because it tests whether improving Q_h and Q_le inadvertently degrades another process. It would nevertheless be helpful to clarify the rationale for keeping GPP entirely outside the calibration when GPP observations are available at some sites. In practical applications, jointly constraining energy and carbon fluxes could potentially provide additional information about the parameters and reduce compensatory parameter combinations. I am not suggesting that an additional multi-variable calibration experiment is required for this study; a brief discussion of this possibility and its implications for parameter identifiability would be sufficient.

Authors' Response:

Thank you for this comment. GPP was excluded from calibration to evaluate how parameters calibrated against Q_h and Q_le affect a carbon flux that was not included in the calibration objective. Retaining GPP for this evaluation allows us to examine whether improvements in the target energy fluxes are accompanied by improvements or deterioration in another process, even where GPP observations are available.

We agree that joint energy–carbon calibration is a valuable extension. We will explain why GPP was excluded from calibration and discuss this possibility and its implications for parameter identifiability.

Reviewer Comment 10:

Please carefully check a small number of references where there appear to be substantive bibliographic or citation issues. The Gupta et al. (2008) entry appears to combine bibliographic details from different papers and should be corrected against the intended publication. The author lists for Kuppel et al. (2014) and Vrugt et al. (2005) also appear to require correction against the publisher records, and the Rosolem et al. (2012) entry should be verified because the stated bibliographic details are inconsistent with the journal record. In addition, Exbrayat et al. (2014) does not appear to support the statement concerning empirical or sparse estimation of soil hydraulic conductivity and minimum stomatal conductance, and the description of Feng et al. (2023) appears broader than the quantities evaluated in that study. Please revise these citations or the associated statements as appropriate.

Authors' Response:

Thank you for identifying these bibliographic and citation issues. We will correct the author lists for Kuppel et al. (2014) and Vrugt et al. (2005) against the publisher records.

For Gupta et al. (2008), we will correct the bibliographic details. We will remove the Rosolem et al. (2012) entry and its associated citation, as stated in our response to **Reviewer 2, Comment 40**.

We agree that Exbrayat et al. (2014), which examines soil carbon responses to model spin-up. We will remove this citation from that statement and revise the wording.

For Feng et al. (2023), we will restrict our description to the daily streamflow performance, decadal streamflow trends, and evapotranspiration evaluation reported in that study.

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

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

Gupta, H. V., Wagener, T., & Liu, Y. (2008). Reconciling theory with observations: elements of a diagnostic approach to model evaluation. *Hydrological Processes, 22*, 3802–3813. https://doi.org/10.1002/hyp.6989.

Reviewer Comment 11:
Line 17. ILS is used before it is defined. Please introduce Integrated Land Simulator (ILS) at its first occurrence in the Abstract and use the abbreviation thereafter.
Authors' Response:
Thank you for pointing this out. We will introduce "Integrated Land Simulator (ILS)" at its first occurrence in the Abstract and use "ILS" thereafter.

Reviewer Comment 12:
Lines 105 and 108. Please revise the notation around g_z(.)
Authors' Response:
Thank you for identifying this typesetting issue. We will correct both occurrences and consistently denote the parameter-generation network as $g_z(\cdot)$.

Reviewer Comment 13:
Line 128. Please replace 'leave-one-out cross-validation' with 'leave-one-site-out cross-validation'.
Authors' Response:
Thank you for the suggestion. We will replace "leave-one-out cross-validation" with "leave-one-site-out cross-validation".

Reviewer Comment 14:
Line 350. The SFE results are presented and discussed later in the manuscript, but the method itself is introduced only by name and citation. For completeness, please provide a brief description of SFE.
Authors' Response:
Thank you for this suggestion. We will add a brief description of the SFE method to **Section 2.6.2**. SFE relates the partitioning of surface energy to an approximate balance between surface heating and moistening in the near-surface relative-humidity budget (McColl et al., 2019).
In our implementation, air temperature (T_a) and specific humidity (q_a) are used to estimate the Bowen ratio:

$$B = \frac{R_\mathrm{v} c_\mathrm{p} T_\mathrm{a}^2}{L_\mathrm{v}^2 q_\mathrm{a}}$$

Observed net radiation (R_n) and ground heat flux (G) then determine the available energy, which is partitioned as

$$Q_\mathrm{le} = \frac{R_\mathrm{n} - G}{1 + B}, \quad Q_\mathrm{h} = R_\mathrm{n} - G - Q_\mathrm{le}$$

Here, R_v, c_p, and L_v are the gas constant for water vapour, the specific heat capacity of air, and the latent heat of vaporization, respectively. The Bowen ratio varies with temperature and humidity.
SFE is included as a simplified physical reference to assess how well this energy-partitioning relationship reproduces the fluxes compared with the more complex land surface models.

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*(7), 2036–2049. https://doi.org/10.1029/2019MS001685.

Reviewer Comment 15:
(16)–(17), Lines 395–400. For clarity, please define ORI when it is first introduced in these equations, as it could initially be interpreted as referring to observations. A short statement that ORI represents the original or default ILS parameter configuration would make the equations immediately clear.
Authors' Response:
Thank you for pointing out this ambiguity. We will add the following definition when ORI first appears in Eqs. (16)–(17):
The subscript ORI refers to the ILS simulation using the default parameters and the metrics calculated for that simulation.