

Comments on egusphere-2026-4009

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.

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.
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.
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.
4. **Eqs. (1)-(2), around lines 145-173.** For completeness in Section 2.2, please clarify how the errors associated with Q_h and Q_l 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.
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.
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.

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.
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.
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.
10. **References.** 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.
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.
12. **Lines 105 and 108.** Please revise the notation around $g_z(.)$

13. **Line 128.** Please replace 'leave-one-out cross-validation' with 'leave-one-site-out cross-validation'.
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.
15. **Eqs. (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.