Reviewer #3

Reviewer Comment 1：

This study presents a multi-task differentiable parameter learning (MdPL) framework for calibrating parameters in the Integrated Land Simulator (ILS). The framework combines a multi-task neural surrogate with a parameter-generation network, and the resulting parameter sets are evaluated using the original ILS.

The study includes surrogate-ILS consistency screening under joint parameter perturbations, benchmarking against other land surface models and an LSTM, comparison with a Sobol-based calibration baseline, and an out-of-target evaluation using GPP.

Overall, I found the manuscript well written and well organised, and the methodological workflow is generally easy to follow. The consideration of surrogate reliability and non-target flux behaviour is valuable. However, several questions concerning independent evaluation, the construction of the final parameter vector, and the interpretation of the methodological and generalisation claims need to be addressed before the conclusions can be fully supported. I therefore recommend major revision.

Some comments below request clarification of the existing experiments, whereas others suggest additional diagnostics. These diagnostics are intended to strengthen particular claims rather than require an extensive new experimental programme. Where additional analyses are not feasible, clearer limitations and more cautious interpretation would be appropriate.

Authors' Response:

We thank the reviewer for the careful assessment of our manuscript and for recognising the value of evaluating surrogate reliability and the behaviour of fluxes not used for parameter calibration. The comments have helped us clarify the experimental design and strengthen the evidence supporting the framework.

Reviewer Comment 2:

(1) Lines 320-335, Eq. (2), and Fig. 1. Please clarify the temporal partitioning used for surrogate training, parameter-generator training, validation or model selection, and final evaluation. The current description does not establish whether the main performance estimates are independent of calibration.

For the fixed-parameter application evaluated here, an independent temporal test should use parameters estimated from the calibration period and held fixed during evaluation, with standardisation statistics calculated using training data only. Fig. 1 indicates that historical flux observations are supplied to the generator; please explain their temporal relationship to the evaluation targets and reconcile this description with the definition of the auxiliary inputs in the text.

Authors' Response:

Thank you for this comment. In the original experiments, surrogate training, parameter-generator training, and evaluation used the same analysis period, without a separate temporal test set. Therefore, the reported results represent training-period performance.

To address this limitation, we conducted additional experiments at four sites (CN-Din, DE-Meh, IE-Dri, and JP-SMF), using the first 80% of each record for calibration and the final 20% for independent testing. We plan to extend this evaluation to all 20 sites in the revised manuscript.

The additional ILS evaluations showed lower test RMSE for both Qh and Qle at all four sites, with mean reductions ranging from 11.2% to 36.2%. The training and test periods and corresponding RMSE results are reported in **Table R2-1 in our response to Reviewer 2, Comment 1**. The results from the three random seeds and the variability of the calibrated parameters are presented in **Figure R2-3 and our response to Reviewer 2, Comment 16**.

Based on this revised experimental procedure, we have updated Fig. 1 to clearly distinguish surrogate training, parameter-generator training, and independent ILS evaluation. We will also clarify the temporal use of historical flux observations and the auxiliary inputs in Eq. (2).

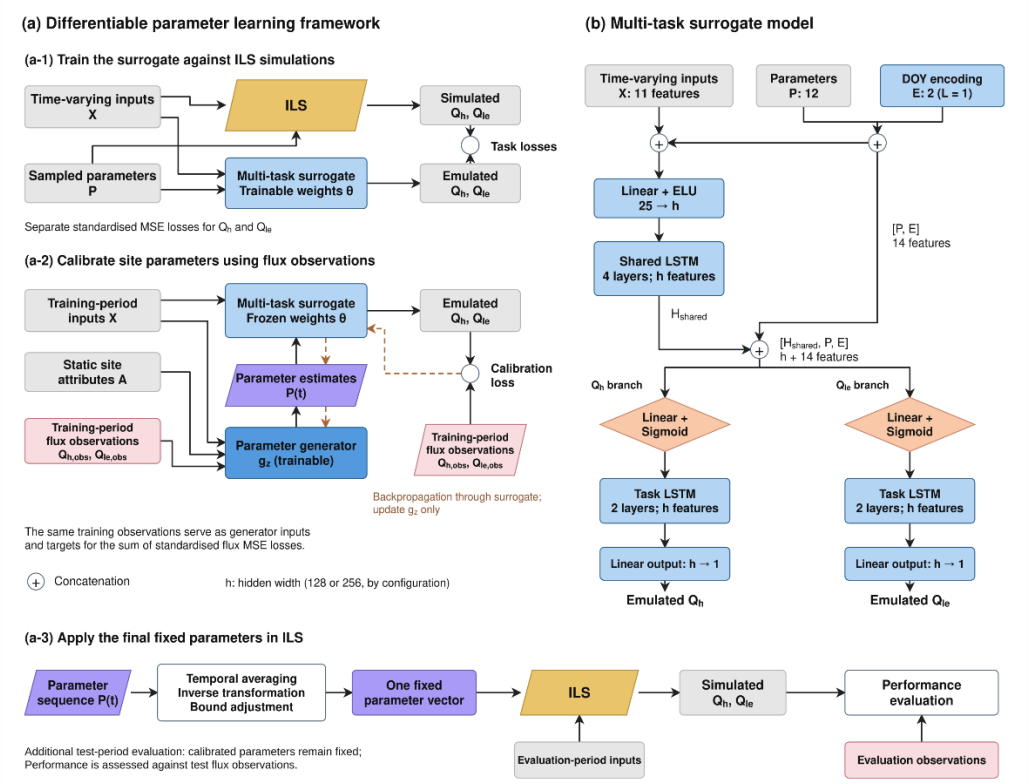


Figure 1. Overview of the multi-task differentiable parameter learning (MdPL) framework for land surface models (LSMs). (a) The workflow comprises surrogate training using Integrated Land Simulator (ILS) simulations (a-1), site-specific parameter calibration using flux observations (a-2), and evaluation of the final fixed parameters in ILS (a-3). The parameter generator, g_z, receives time-varying inputs, static site attributes, and flux observations from the training period. Its weights are optimized by propagating the flux prediction errors through the surrogate, whose weights remain frozen during calibration. The transformed parameter values generated by the network are averaged, converted back to physical values, and adjusted to the prescribed bounds before application in ILS. The resulting ILS simulations are evaluated against flux observations. (b) Architecture of the multi-task surrogate, combining time-varying inputs, ILS parameter values, and day-of-year encoding through a shared long short-term memory (LSTM) network and separate task-specific branches for sensible (Qh) and latent (Qle) heat fluxes. The "+" symbols denote concatenation.

Reviewer Comment 3:

(2) Lines 280-292 and Fig. 2. The joint-perturbation screening is useful, but correlation between complete flux time series does not adequately characterise bias or errors in parameter-dependent responses. Shared meteorological and diurnal variation could produce high correlations even when parameter effects are imperfectly represented.

Please supplement R with RMSE or bias, preferably including comparisons of changes relative to the default simulation and surrogate errors at the final calibrated parameters. It would also be helpful to describe how the 100 joint samples were generated and justify the median R > 0.9 threshold.

Authors' Response:

Thank you for this comment. We agree that a high correlation between full flux time series does not necessarily indicate accurate surrogate responses to parameter changes. The original median R > 0.9 threshold was an empirical criterion.

The 100 joint parameter combinations used for screening were randomly generated within the prescribed parameter ranges (**Table 2**).

Following the reviewer's suggestion, we conducted additional comparisons between the surrogate and ILS at four sites. We compared their flux changes relative to the default parameter configuration and calculated the corresponding response RMSE, as reported in **Table R2-2** in our response to **Reviewer 2, Comment 9**. We also evaluated surrogate errors at the final calibrated parameters using RMSE, bias, and R, as reported in **Table R2-3**.

These results show that surrogate accuracy in reproducing parameter-induced flux changes varies among sites, confirming the limitations of correlation-based screening. However, all 12 final parameter sets were directly evaluated using ILS against independent test-period observations, with improvements in both fluxes at all four sites (**Table R2-1, Reviewer 2, Comment 1**).

We will clarify the screening criterion and its limitations in the revised manuscript and distinguish the original screening results from the additional surrogate diagnostics.

Reviewer Comment 4:

(3) Eqs. (1)-(2), around lines 145-173. Please explicitly define how the Qh and Qle losses are combined during surrogate training and parameter calibration, including whether standardised values are used, whether relative weights are applied, and how missing observations are handled. The Sobol baseline selects parameter sets using RMSE(Qh) + RMSE(Qle), which is not generally equivalent to minimising a sum of MSE losses.

Although common performance metrics are subsequently reported, please explain whether differences in the optimisation or selection objectives affect the interpretation of the comparison.

Authors' Response:

Thank you for this comment. Separate MSE terms are calculated for Qh and Qle. During surrogate training, the two task losses are applied sequentially within each mini-batch. During parameter calibration, the flux loss is the sum of these two MSE terms.

Each term is the mean squared error over the observations. No additional coefficients assign different weights to the two standardised losses. The original calibration implementation uses separate masks for Qh and Qle, retaining standardised target values. We will explicitly describe the standardisation, task updates and masking rule in the Methods.

We agree that selecting parameters by the sum of RMSEs is not equivalent to selecting them by the sum of MSEs. We have therefore re-ranked the archived Sobol simulations using $\mathrm{MSE}_{Q_\mathrm{h}} + \mathrm{MSE}_{Q_\mathrm{le}}$ and use this same metric to compare the resulting ILS simulations. Sobol candidates are selected using the sum of MSE. The comparison therefore assesses the resulting ILS performance using a common evaluation metric while retaining each method's calibration procedure.

Under this criterion, MdPL has a lower combined MSE than the selected Sobol parameter set at CN-Din, DE-Meh and IE-Dri by 10.38%, 35.28% and 47.02%, respectively. At DE-Seh, the MdPL combined MSE is 5.52% higher. The updated comparison is presented in our response to **Reviewer 2, Comment 14**.

Reviewer Comment 5:

(4) Lines 320-321, Table 2, and Tables S1-S4. The supplementary captions describe clipping unconstrained parameter estimates, but Table S2 reports rlfn values of 0.23 and 0.22 for DK-Fou and IE-Ca1, below the lower bound of 0.30 in Table 2. Please resolve this discrepancy and explain the order of optimisation, clipping, and temporal averaging, identifying the final parameter vectors used in ILS.

Because the surrogate is nonlinear, optimisation using time-dependent parameters is not equivalent to optimisation using their mean. For a few representative sites, it would therefore be useful to compare surrogate performance using the generated parameter sequence and the final constant vector, together with a brief description of parameter variability before averaging.

Authors' Response:

Thank you for requesting clarification of how the final parameter set was obtained. After optimisation, the transformed parameter values generated by the network were averaged. The averages were then converted back to physical units, and values outside the prescribed ranges were manually adjusted to the corresponding bounds before being supplied to ILS. The *rlfn* values of 0.23 and 0.22 reported for DK-Fou and IE-Ca1 represent estimates before this final adjustment to the parameter bounds; applying the lower bound gives 0.30 at both sites. We will revise **Table S2** and explicitly describe the sequence of averaging, inverse transformation and adjustment to the parameter bounds.

We have also examined parameter variability before averaging in **Reviewer 1, Comment 7**. They show that some parameters are relatively stable, while others vary substantially over time and repeatedly approach or exceed the specified bounds before the final adjustment.

In the additional four-site experiments, we compared predictions from the same frozen surrogate using the generated parameter set. The effect varied among sites. Differences were small at CN-Din and IE-Dri. At DE-Meh, the mean training-period latent-heat RMSE across the three runs increased from 26.49 to 35.85 W m^{-2}. At JP-SMF, the corresponding changes were 43.57 to 52.83 W m^{-2} for sensible heat and 43.73 to 46.07 W m^{-2} for latent heat. These results quantify the site-dependent effect of replacing the generated parameter sequence with the final fixed parameter set used in ILS.

We also evaluated the final fixed parameter vectors directly in ILS against independent test-period observations, as reported in **Table R2-1** in our response to **Reviewer 2, Comment 1**. We will distinguish these ILS results from the surrogate diagnostic. Developing and evaluating a physically constrained dynamic parameterisation in ILS remains a direction for future work.

Reviewer Comment 6:

(5) Table 2, Fig. 4, and lines 591-595. The methods appropriately describe boundary proximity as a potential indication of limited identifiability, but the conclusions use stronger wording. Please retain the same caution throughout: boundary-proximal solutions may reflect parameter compensation, unsuitable bounds, or a

constrained optimum and do not alone establish non-identifiability. The interpretation of vgcov should also account for its default value already lying at the upper bound.

In addition, please clarify whether relevant joint physical constraints are enforced. For example, if rlfn and tlfn represent reflectance and transmittance in the same spectral band, their sum should not exceed one, whereas the listed independent ranges permit combinations above this value.

Authors' Response:

Thank you for this clarification. We agree that proximity to a prescribed bound is a diagnostic feature of the calibration solution. We will use consistent wording throughout the manuscript and explain that parameter estimates close to their bounds may reflect parameter compensation.

We will also explicitly account for the default *vgcov* value of 1, which already lies at its upper bound. Consequently, proximity to this boundary cannot automatically be attributed to calibration, and reductions from the default value should be interpreted together with their magnitude and the associated flux responses.

The original optimisation did not explicitly impose the joint constraint. We therefore checked the site-level parameter combinations reported for all 18 sites that passed the surrogate screening. All satisfy this condition, with a maximum sum of 0.88. Adjusting the two *rlfn* entries discussed in **Comment 5** to the lower bound of 0.30 gives sums of 0.49 and 0.40 at DK-Fou and IE-Ca1, respectively, and does not alter this conclusion.

We will report this check and clarify that the calibrated parameter combinations satisfy the optical constraint, while the independent parameter ranges alone do not enforce it. This distinction will be included alongside the parameter-range description.

Reviewer Comment 7:

(6) Lines 320-338 and 458-461. The comparison between multi-task and single-task configurations is useful, but differences in parameter count and learning-rate strategy limit attribution of the results specifically to task sharing or gating. Please distinguish evidence supporting the complete framework from evidence supporting its individual components. If improved generalisation or reduced overfitting is attributed specifically to the gates, a focused ablation would be informative. Otherwise, these statements should be presented as design motivations rather than demonstrated findings. I do not consider a broader architecture search necessary.

Authors' Response:

Thank you for this comment. We agree that differences in parameter count and learning-rate strategy prevent us from attributing the observed performance differences specifically to task sharing or gating.

We will clarify that the comparison evaluates the complete model. Claims regarding reduced overfitting, improved generalisation, and strong collaborative effects will be removed or revised as architectural design motivations rather than demonstrated findings.

Reviewer Comment 8:

(7) Lines 240-245 and 334-335, Supplementary Section S2, and the conclusions. The leave-one-site-out analysis provides useful evidence of transferability, including its limitations, and its principal findings should be

summarised in the main text. Please clarify how parameters are generated for the held-out site and how the framework would operate without local flux observations. Since the principal implementation trains a separate model at each site, computational scalability should be distinguished from demonstrated spatial generalisation. The abstract and conclusions should reflect the scope of the available spatial validation.

Authors' Response:

Thank you for this suggestion. We will move the main leave-one-site-out results from Supplementary **Section S2** into the main text.

For each target site, the transferred parameter set was obtained by averaging the calibrated parameters from the other sites and then applied as a fixed parameter set in ILS. The results show that transfer performance depends on the site group and target flux.

For grid cells without local flux observations, a future extension could train a shared parameter-generation network across multiple sites using meteorological inputs and land-surface attributes available at the target cells. This would require evaluation at sites excluded from training, covering different climates and vegetation conditions, together with verification of surrogate reliability in those environments.

We will revise the Abstract and Conclusions to distinguish the potential for computational scaling from the spatial transferability demonstrated here. Further details are provided in our responses to **Reviewer 1, Comment 6**, and **Reviewer 2, Comment 21**.

Reviewer Comment 9:

(8) Lines 345-355, Fig. 5, and Section 3.2. Please describe the training and calibration information available to the LSTM and LSM benchmarks, because common meteorological forcing alone does not establish equivalent evaluation conditions. SFE and the other methods also use different site sets; comparisons over common sites and valid timestamps would be preferable for ranking their performance. Please briefly describe SFE and its implementation, together with temporal aggregation and missing-data rules. When interpreting performance across resolutions, distinguish error cancellation caused by aggregation from a sustained advantage relative to the benchmarks, and qualify the temporal-robustness claim accordingly.

Authors' Response:

Thank you for this comment. The benchmarking was intended to compare the default and calibrated ILS with established modelling approaches. The other LSMs use the default PLUMBER2 configurations and were not recalibrated to the site flux observations in this study. This comparison shows how calibration changes ILS performance relative to these benchmark models.

The LSTM predictions were supplied by PLUMBER2. Its published protocol trains the model using 167 sites and predicts fluxes at the three sites excluded from training, repeating the procedure across the dataset. It uses meteorological variables, atmospheric CO2, LAI and static site attributes to predict Qh, Qle and NEE jointly. The training data used by this LSTM benchmark differ from those used for site-specific MdPL calibration. We will state this distinction explicitly, with reference to Abramowitz et al. (2024, Sections 2.2–2.3).

Our SFE implementation partitions observed available energy, $Rn - G$, between Qh and Qle using a Bowen ratio calculated from air temperature and specific humidity, without fitting parameters to the target fluxes. Its use of

observed net radiation and ground heat flux will be stated explicitly. Further methodological details are provided in our response to **Reviewer 1, Comment 14**.

We have examined the results over shared sites, and relative performance remains dependent on the flux and site set. The existing calculations use only time steps with valid values in both the simulations and observations. For longer averaging intervals, the time series are averaged before calculating the metrics. We will restrict direct rankings to common sites and the same time steps with valid data, and document the aggregation and missing-data rules.

Abramowitz, G., Ukkola, A., Hobeichi, S., Cranko Page, J., Lipson, M., De Kauwe, M. G., Green, S., Brenner, C., Frame, J., Nearing, G., Clark, M., Best, M., Anthoni, P., Arduini, G., Boussetta, S., Caldararu, S., Cho, K., Cuntz, M., Fairbairn, D., Ferguson, C. R., Kim, H., Kim, Y., Knauer, J., Lawrence, D., Luo, X., Malyshev, S., Nitta, T., Ogee, J., Oleson, K., Ottlé, C., Peylin, P., de Rosnay, P., Rumbold, H., Su, B., Vuichard, N., Walker, A. P., Wang-Faivre, X., Wang, Y., & Zeng, Y. (2024). On the predictability of turbulent fluxes from land: PLUMBER2 MIP experimental description and preliminary results. *Biogeosciences, 21*(23), 5517–5538. https://doi.org/10.5194/bg-21-5517-2024. bg.copernicus.org

Reviewer Comment 10:

(9) Lines 545-562 and Fig. 7. The manuscript already acknowledges site-dependent GPP responses and the need for broader evaluation. Nevertheless, the magnitude of the IE-Dri deterioration deserves explicit discussion: RMSE increases from 1.95 to 3.93 micromoles of CO2 per square metre per second, and NSE decreases from 0.64 to -0.46. Please discuss the associated seasonal response or parameter changes where possible and distinguish relative improvement from absolute performance. The abstract should reflect the mixed outcomes and limited site coverage as carefully as the conclusions. Please also identify the GPP reference product, explain its relationship to PLUMBER2 and FLUXNET2015, and specify the temporal resolution used for evaluation.

Authors' Response:

Thank you for pointing out this inconsistency. We agree that four sites, with one site per PFT, do not support a general statement that GPP is not degraded. We will replace the statement in the Abstract with: "An evaluation of GPP at four sites revealed site-dependent responses to energy-flux calibration, including degradation at IE-Dri."

At IE-Dri, the deterioration is associated with pronounced summer GPP overestimation after calibration. For example, the July mean GPP is approximately 8.21 μmol CO2 m−2 s−1 in the reference data, 8.26 in the default simulation and 14.49 in the calibrated simulation. The daily-mean RMSE increases from 1.95 to 3.93 μmol CO2 m−2 s−1, while NSE decreases from 0.64 to −0.46. The negative NSE indicates poorer performance than using the observed mean as a constant prediction during this evaluation period. Several canopy and stomatal parameters changed simultaneously, so the net GPP response cannot be inferred from the direction of *vmax0* alone. The broader analysis of how parameter changes affect the simulated fluxes is provided in our response to **Reviewer 2, Comment 19**.

The GPP reference series were obtained from the PLUMBER2 flux files. CN-Din and DE-Seh use the FLUXNET2015 GPP_NT_VUT_REF product, whereas DE-Meh and IE-Dri use the GPP_f series from La Thuile. Thus, PLUMBER2 is the distribution source, while FLUXNET2015 and La Thuile identify the original

data sources. The metrics in **Fig. 7** were calculated from daily mean GPP derived from the half-hourly series. We will specify these products and the evaluation resolution consistently in the text and figure caption.

Reviewer Comment 11:

(10) Lines 226-245 and Table 1. The manuscript already explains the broad considerations behind site selection and appropriately identifies the study as a proof of concept. Please supplement this with more specific inclusion criteria and soften the statement that the subset "ensures" geographic and ecosystem diversity. It would also be useful to distinguish available observation records from the periods actually analysed. In particular, the description of approximately 1-3 years does not clearly correspond to all listed periods, including JP-SMF and IT-BCi. Please report the effective data coverage where it differs from the nominal record span.

Authors' Response:

Thank you for this suggestion. Site selection began with the vegetation classifications reported in the flux-site datasets. We first identified the four broad PFT groups and then selected sites within these groups. The subset was limited to 20 sites to keep the site-specific ILS simulations, surrogate training, and parameter calibration computationally manageable. This selection supports a multi-site proof-of-concept evaluation. We will therefore replace "to ensure geographic and ecosystem diversity" with "to include sites from four broad PFT groups across several geographic regions."

We have checked the temporal coverage of the datasets for all 20 sites. All have a regular half-hourly time axis, and the analysed calendar periods span 1–6 years. We will correct the "approximately 1–3 years" description and rename the "Observation period" column in **Table 1** as "Analysed period." To distinguish record length from the availability of usable flux values, **Table R3-1** below reports the number and percentage of time steps with valid observations of both Qh and Qle. We will include this information in a supplementary table accompanying **Table 1**.

Table R3-1. Analysed periods and availability of records with valid values for both sensible and latent heat fluxes at the 20 candidate sites. All records have a half-hourly temporal resolution.

Site	Analysed period	Total time steps	Uncorrected fluxes: valid pairs (%)	Corrected fluxes: valid pairs (%)
CN-Din	2003–2005	52,608	52,608 (100.0%)	—
ID-Pag	2002–2003	35,040	35,040 (100.0%)	—
PT-Esp	2002–2004	52,608	52,608 (100.0%)	—
PT-Mi1	2005	17,520	17,520 (100.0%)	—
AR-SLu	2010	17,520	17,520 (100.0%)	17,520 (100.0%)
CN-Cha	2003–2005	52,608	52,608 (100.0%)	—
DE-Meh	2004–2006	52,608	52,608 (100.0%)	52,608 (100.0%)
JP-SMF	2003–2006	70,128	70,128 (100.0%)	—
US-Bar	2005	17,520	17,520 (100.0%)	—
AU-Emr	2012–2013	35,088	35,088 (100.0%)	35,088 (100.0%)

CN-Dan	2004–2005	35,088	35,088 (100.0%)	—
CN-Du2	2007–2008	35,088	35,088 (100.0%)	28,084 (80.0%)
DK-Lva	2005–2006	35,040	35,040 (100.0%)	9,264 (26.4%)
IE-Dri	2003–2005	52,608	52,608 (100.0%)	47,664 (90.6%)
PL-wet	2004–2005	35,088	35,088 (100.0%)	31,679 (90.3%)
DE-Seh	2008–2010	52,608	52,608 (100.0%)	52,608 (100.0%)
DK-Fou	2005	17,520	17,520 (100.0%)	14,648 (83.6%)
IE-Ca1	2004–2006	52,608	52,608 (100%)	51,159 (97.2%)
IT-BCi	2005–2010	105,168	105,168 (100.0%)	51,062 (48.6%)
IT-CA2	2012–2013	35,088	35,088 (100.0%)	35,088 (100.0%)

Notes: "Uncorrected" and "corrected" refer to the absence or presence of energy-balance-closure correction, respectively.

For the two sites specifically identified by the reviewer, JP-SMF covers four calendar years, with 70,128 valid paired flux records. IT-BCi covers six calendar years, with 105,168 half-hourly time steps, but its energy-balance-corrected fluxes used in the evaluation contain 51,062 valid pairs, or 48.6% coverage. These corrected fluxes are unavailable in 2005–2006 and only partially available in 2007. The six-year simulation period therefore does not imply valid flux observations throughout the six years. This difference concerns the availability of corrected flux values, not their temporal resolution. We will clarify this distinction in the data description and remove wording that implies uninterrupted valid flux observations at every site.

Reviewer Comment 12:

(11) Sections 2.3 and 2.6.1. Please list the meteorological forcing variables and their units, distinguish the inputs supplied to ILS and the two networks, and specify the ILS version and configuration. At line 326, 48 half-hourly time steps correspond to one day rather than two days, unless resampling was applied. Please clarify the actual input frequency and sequence duration. It would also be helpful to explain sequence construction, recurrent-state handling across separate simulations, the final-model or checkpoint-selection rule, and whether the rolling training strategy involves anything beyond the reported learning-rate adjustment.

Authors' Response:

Thank you for requesting these implementation details. ILS, the surrogate, and the parameter generator use the same prescribed time-varying forcing variables. Their names and physical units are listed in the table accompanying our response to **Reviewer 1, Comment 2.** The surrogate additionally receives the 12 ILS parameter values and the DOY encoding. The parameter generator additionally receives site attributes and the Qh and Qle observations from the training period. In ILS, the prescribed forcing is combined with the model's boundary conditions, initial states, and parameter configuration. The revised **Figure 1** distinguishes these inputs and their roles.

The simulations use MATSIRO within ILS in an offline, site-scale configuration, with a 1,800-s model time step and half-hourly flux output. The initialization procedure and additional initialization-sensitivity tests are described in our response to **Reviewer 2, Comment 6**.

The input interval is 30 minutes: 48 steps correspond to one day and 96 steps to two days. Both sequence lengths occur in the site-specific configurations. We will correct line 326 and report the configurations explicitly. Training examples are overlapping windows with a stride of one time step, constructed separately within each annual simulation sequence. Recurrent hidden and cell states are initialized independently for each window, and the training loss is evaluated at the final time step of the window.

In the original experiments, surrogate training stopped when the training-loss stopping criterion was met or the maximum number of epochs was reached. The parameter generator was trained for a fixed number of epochs specified for each site. The surrogate configurations are described in our response to **Reviewer 1, Comment 5**.

Finally, "rolling training" refers to alternating task-wise updates within each mini-batch. Qh and Qle are processed sequentially, with each update adjusting the shared network and the corresponding task-specific branch. The shared-network learning rate is one-third of the task-branch learning rate. We will replace "rolling training" with this explicit description.

Reviewer Comment 13:

(12) Lines 188-195 and Eqs. (3)-(4). The motivation for representing seasonality is clear, but please report the number of frequencies, L, and explain the choice f_l = 2^l. Was this representation compared with a simpler annual sine-cosine encoding? If no comparison was conducted, a design rationale and appropriately cautious wording would be sufficient; its superiority or robustness should not be implied as an established result.

Authors' Response:

Thank you for this clarification. The actual implementation used L = 1. Thus, the configuration used in this study already corresponds to the simpler annual representation.

The purpose of this encoding was to provide two features representing the annual seasonal cycle alongside the time-varying forcing. We did not conduct an ablation comparing alternative seasonal encodings. We will explicitly report L = 1, simplify the equations and description to match the implemented annual encoding, and remove wording suggesting that multi-frequency benefits or improved robustness were demonstrated.

Reviewer Comment 14:

(13) Fig. 1(b) and Eqs. (7)-(8). The diagram appears to concatenate the shared LSTM representation with a pathway carrying parameter and DOY information before gating, whereas the equations generate gates from X_combined and apply them element-wise to H_shared. Please make these descriptions consistent with the implementation and provide the relevant tensor dimensions. Since gating operates on learned features, statements about selecting individual physical parameters should also be qualified.

Authors' Response:

Thank you for identifying the ambiguity between the diagram and the equations. We have revised **Figure 1(b)** to show the implemented sequence of operations. The 11 time-varying inputs, 12 parameter values, and two DOY features are first concatenated, projected to the hidden dimension through a linear layer with ELU activation, and passed to the shared LSTM. The shared representation is then concatenated with the parameter and DOY features before entering each task-specific linear–sigmoid transformation and LSTM branch.

The corresponding computation is:

$$C = [X, P, E], \quad H_{\text{shared}} = \text{LSTM}_{\text{shared}}(\text{ELU}(W_0 C + b_0)),$$

$$Z_k = \sigma(W_k[H_{\text{shared}}, P, E] + b_k), \quad \hat{Q}_k = W_{\text{o},k}\text{LSTM}_k(Z_k) + b_{\text{o},k}.$$

where k denotes the Qh or Qle task. The sigmoid-transformed features Z_k are supplied directly to the corresponding task LSTM. We will update Eqs. (7)–(8) and the surrounding description accordingly.

For sequence length T, batch size B, and hidden width h, the relevant dimensions are:

Table R3-2. Tensor dimensions at each stage of the multi-task surrogate model.

Representation	**Tensor dimensions**
Concatenated inputs $[X, P, E]$	T × B × 25
Shared LSTM representation	T × B × h
Input to each task branch $[H_{\text{shared}}, P, E]$	T × B × (h + 14)
Task-specific sigmoid output	T × B × h
Output of each flux branch	T × B × 1

Here, h is 128 or 256 according to the site-specific configuration.

We also agree that these transformations act on learned features. We will describe their role as constructing features for each task using the shared representation, parameter values and time-of-year information.

Reviewer Comment 15:

(14) Fig. 3, Section 2.6.5, and line 419. The manuscript defines ordinary RMSE in Eq. (11) and identifies the Taylor-diagram contours as centred RMSE. For clarity, please state explicitly in the Fig. 3 caption that the relative RMSE improvement annotations use the ordinary RMSE defined in Eqs. (11) and (16). Reporting mean bias would provide a useful complementary diagnostic. For key architecture or performance comparisons, variability across repeated training runs would help assess stability, since cross-site variability does not establish robustness to network initialisation. Please replace "significantly improved" with descriptive wording unless statistical significance is demonstrated, and qualify robustness claims that were not directly tested.

Authors' Response:

Thank you for these suggestions. We have redrawn **Figure 3**, to distinguish ordinary RMSE from the centred RMSE represented in the Taylor diagrams. Each site's simulated standard deviation is normalized by its own observed standard deviation, and the dashed contours represent centred RMSE normalized by that same observed standard deviation.

Within each PFT group, we first calculate the arithmetic mean of the site-level RMSE values for each model and then calculate their relative change. Under the revised Eq. (16), negative ΔRMSE indicates lower error. The relative-correlation change retains the definition in Eq. (17).

The revised figure shows lower PFT-mean RMSE for both fluxes in all four groups. The reductions range from 5.7% to 15.1% for Qh and from 8.7% to 13.5% for Qle. Correlation changes remain mixed: the mean Qle

correlation decreases by 4.8% in the cultivation group, despite its lower mean RMSE. We have additionally calculated mean bias for the default and calibrated ILS simulations and will include these values alongside RMSE and correlation in the site-level supplementary table.

To assess sensitivity to parameter-generator initialization, we conducted three training runs with different random seeds at each of four sites. Figure R2-3 shows three parameter sets per site and their ILS performance. All eight site–flux combinations improved relative to the default ILS for these reported sets; the differences between the highest and lowest RMSE were 0.04–6.22 W m^{-2} across the eight combinations. The selection of runs is described in our response to **Reviewer 2, Comment 16.**

We will replace “significantly improved” with descriptive wording wherever statistical significance was not tested. Statements about initialization robustness will refer to the repeated-run evidence, and claims concerning individual architectural components will be qualified as explained in our response to **Comment 7**.

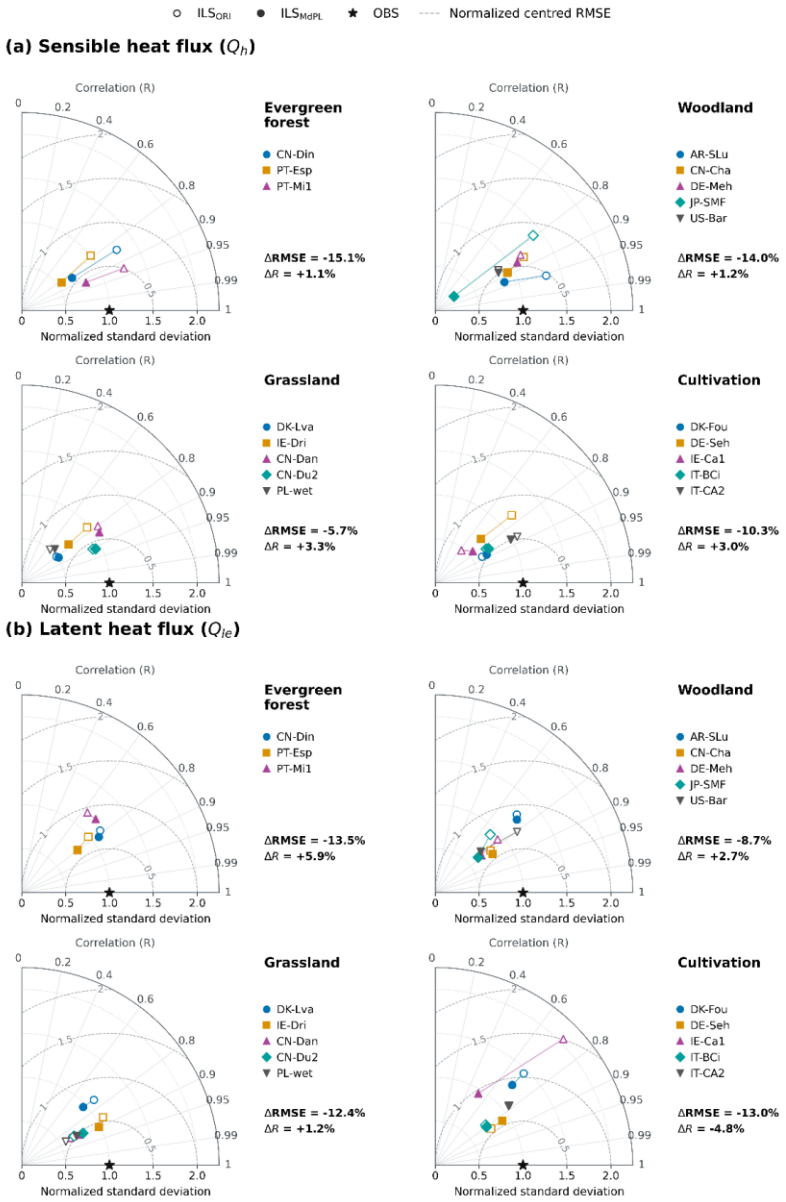


Figure 3. Normalized Taylor diagrams comparing the default and MdPL-calibrated ILS at 18 sites that passed the surrogate–ILS consistency screening, grouped by plant functional type (PFT), for (a) sensible heat flux (Qh) and (b) latent heat flux (Qle). Colours and marker shapes identify sites. Open and filled markers denote the default (ILS_ORI) and calibrated (ILS_MdPL) configurations, respectively, with lines connecting the two configurations at each site. The black star represents the observations. Angular position indicates Pearson's correlation coefficient (R), radial distance indicates the simulated standard deviation normalized by the observed standard deviation, and grey dashed contours indicate centred RMSE normalized by the same observed standard deviation. The ΔRMSE annotations use ordinary RMSE, including mean bias. Both ΔRMSE and ΔR are calculated from the arithmetic means of the site-level metrics within each PFT group. Negative ΔRMSE indicates lower error, and positive ΔR indicates higher correlation. Results are shown for the training periods used in the original experiments.

Reviewer Comment 16:

(15) Please introduce Integrated Land Simulator (ILS) at its first occurrence in the abstract, correct the rendering of the parameter-generator notation around lines 105-108, and use "leave-one-site-out" where sites are the validation units. Although ILS_ORI is defined earlier, briefly restating the meaning of ORI alongside Eqs. (16)-(17) would make the equations easier to read. In Fig. 1, please replace "parameter priors" with "parameter estimates" if the outputs are point estimates rather than probabilistic priors.

Authors' Response:

Thank you for these editorial suggestions. We will introduce "Integrated Land Simulator (ILS)" at its first occurrence in the Abstract, standardize the parameter-generator notation as g_z(.), and use "leave-one-site-out cross-validation" when one site is excluded at a time.

Alongside Eqs. (16)–(17), we will explicitly state that ORI denotes ILS simulations using the default parameter configuration, while MdPL denotes ILS simulations using the calibrated parameters. This will distinguish the model configurations from the observations.

In the revised **Figure 1**, "parameter priors" has already been replaced with "parameter estimates." This terminology reflects the actual outputs of the parameter-generation network: parameter point estimates, rather than a probability distribution representing prior uncertainty.