# Reviewer #2

**Reviewer Comment 1：**

This study presents MdPL, a calibration framework pairing a multi-task neural surrogate of the ILS land surface model with a differentiable parameter generator. Twelve PFT-level parameters are calibrated against sensible and latent heat fluxes at 20 PLUMBER2 sites covering four plant functional types, with a surrogate reliability screening, a benchmark against four LSMs and an LSTM, a comparison with Sobol sampling, and an evaluation against GPP as a non-target flux.

It is an interesting and timely study, relevant to HESS. The manuscript is clearly written, and I was glad to see a surrogate reliability check and a parameter plausibility diagnostic included at all, since both are often missing from this kind of study. Nevertheless, I have a few comments that must be addressed before the manuscript can be published.

Most importantly, I could not find any out-of-sample evaluation. The generator is trained on the flux observations and the parameters are averaged over the full training period (L320–321), so the 14% RMSE reductions and the comparisons against the other LSMs appear to be calculated over the same period used for the calibration. As it stands, these numbers show how well the model can be fitted rather than how well it predicts. Could the authors state which periods were used for training and evaluation, and add a hold-out test? Related to this, the only test of generalisation is the leave-one-site-out analysis in the supplementary material, which shows that parameters transfer poorly, worse than the default parameters for woodland latent heat and for both fluxes at cultivation sites. Since scalability is claimed in the abstract and conclusions, I think this analysis belongs in the main text.

I am also not convinced that the surrogate screening tests the property the calibration relies on, and no results are reported with uncertainty. Finally, I feel there's a missed opportunity to discuss what the calibrated parameters themselves mean for the model. I believe one of the main strengths of parameter calibration is what it teaches us about the base model and how we can use that information to guide future model development and understanding. For example, several parameters that shift in the same direction at nearly every site can tell us something about the model rather than just about the fit.

I hope my comments will help clarify and strengthen the manuscript. Below are my specific comments:

**Authors' Response：**

We thank the reviewer for the careful assessment of our manuscript and the constructive suggestions. We appreciate the positive evaluation of the framework and the emphasis on independent evaluation, surrogate reliability, and the interpretation of calibrated parameters.

We agree that a separate evaluation is needed to establish how the calibrated parameters perform outside that period. We therefore conducted additional experiments at CN-Din, DE-Meh, IE-Dri and JP-SMF, using the first 80% of each record for parameter-generator training and the final 20% for testing. Fixed parameters were obtained from the training-period outputs and applied to ILS. 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.

The results are summarized in **Table R2-1** below. All eight site–flux combinations showed lower test RMSE than the default ILS. The reductions in mean RMSE range from 11.2% to 36.2%. We will report these results separately from the original training-period results and include the results for each random seed.

We will also move the leave-one-site-out analysis into the main text and clarify how parameters were transferred to the held-out sites, as detailed in our response to **Comment 21**. The scalability statements and limitations of application to unobserved grid cells are addressed in our response to **Comment 22**.

Table R2-1. Training and test periods and test RMSE of the default and MdPL-calibrated ILS at four sites. MdPL values are means of the three random seeds parameter set RMSE, in W $m^{-2}$.

| **Site** | **Training period** | **Test period** | **Qh RMSE: Default → MdPL** | **Qle RMSE: Default → MdPL** |
|---|---|---|---|---|
| CN-Din | 2003-01-01–2005-05-26 | 2005-05-26–2005-12-31 | 45.19 → 40.12 | 65.08 → 53.25 |
| DE-Meh | 2004-01-01–2006-05-26 | 2006-05-26–2006-12-31 | 39.35 → 26.49 | 42.21 → 34.55 |
| IE-Dri | 2003-01-01–2005-05-26 | 2005-05-26–2005-12-31 | 22.36 → 18.30 | 25.97 → 18.96 |
| JP-SMF | 2003-01-01–2006-03-14 | 2006-03-14–2006-12-31 | 72.46 → 46.21 | 65.96 → 54.19 |

**Reviewer Comment 2：**

L37: "did not lead to systematic degradation of a non-target carbon-related flux" rests on one site per PFT (L550), and GPP degrades at IE-Dri (L557). The conclusions (L586–590) are more careful, and the abstract should be brought in line with them.

**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."

**Reviewer Comment 3:**

L41: "scalable", see my comment on the transferability analysis below.

**Authors' Response:**

Thank you for this comment. We will move the leave-one-site-out analysis into the main text and remove the unqualified description of the framework as "scalable" from the Abstract.

We will discuss regional and global application as a future development requiring multi-site parameter learning and independent spatial evaluation.

**Reviewer Comment 4:**

L108, L133, L195 and L213–217: the claims that the gates reduce overfitting and improve stability and generalisability, that the day-of-year encoding gives a robust representation, and that the framework is robust under sparse and noisy observations, are not tested anywhere I could see. These should be phrased as motivation, unless the authors can add an ablation with and without the gates

**Authors' Response:**

Thank you for this comment. We agree that these claims are not directly supported by the current experiments. We will revise the relevant statements to present the gates and DOY encoding as architectural design choices. Claims regarding reduced overfitting, improved stability and generalisability, and robustness under sparse and noisy observations will be identified as requiring further evaluation.

**Reviewer Comment 5:**

L229 and L240–245: the site selection needs more justification. The records are approximately 1–3 years, and four sites (AR-SLu, DK-Fou, PT-Mi1, US-Bar) have only a single year, which is short for training the surrogate and for calibration, and leaves no room for a hold-out period. Twenty sites is also a small subset of what is now available, both within PLUMBER2 (42 sites) and in the larger flux collections that have appeared since (FluxnetShuttle). The authors state that the subset balances ecosystem diversity, data completeness and computational feasibility, but it would help to say which of these is the binding constraint, presumably the roughly 150 ILS simulations needed per site for pretraining and screening. How many sites in the archive would have met the data requirements? This matters for the transferability analysis in Sect. S2, where only 3–5 sites per PFT are available, and where more sites per PFT would be the obvious way to make the test meaningful.

**Authors' Response:**

Thank you for this comment. We will clarify the site selection and study limitations as follows.
First, the study selected 20 sites covering four PFT groups. Computational cost was the main constraint, as each site required multiple ILS simulations to generate surrogate training data and evaluate surrogate performance, followed by neural-network training and parameter calibration. We did not systematically screen the full archive and therefore cannot provide an exact number of eligible sites.
Second, the 1–3 year periods reported in the manuscript refer to the periods used in our experiments, not necessarily the full observational records available at those sites. The selection of these periods was also influenced by computational cost. We will clarify that results from the one-year experiments were evaluated during the training period and do not demonstrate temporal generalisability.
Third, we agree that the limited number of sites within each PFT (3–5 sites) restricts the spatial transferability analysis in **Sect. S2**. We will explicitly acknowledge this limitation and avoid drawing broad conclusions about spatial transferability.

**Reviewer Comment 6:**

L247–249: ILS is described in two sentences. Which version and configuration were used, what are the forcing variables and units, and what is the time step?

I could not find any mention of spin-up. Parameters such as psicr, vgcov and vmax0 change how quickly soil water is drawn down, and therefore the equilibrium soil moisture. Was each perturbed simulation spun up separately, for the 48 pretraining runs, the 100 screening runs, the 256 Sobol runs and the final calibrated run? If not, differences between runs will partly be drift from inconsistent initial states, which over 1–3 years could be a large part of the signal, and it would also add to the cost of the Sobol baseline.

**Authors' Response:**

Thank you for this comment. We will clarify the model configuration, initialization procedure, and its influence on simulation results.

The simulations used ILS with MATSIRO6 in an offline, site-scale configuration. The land-model time step was 1,800 s, and Qh and Qle were output as half-hourly averages. ILS, the surrogate, and the parameter generator used the same prescribed forcing variables, whose names and units are provided in our response to **Reviewer 1, Comment 2**.

For initialization, each site was first spun up for 100 years using the default parameters. The resulting restart state was then reused for parameter-perturbation, screening, Sobol, and calibrated simulations, without repeating the 100-year spin-up for each parameter set.

To assess the influence of this initialization procedure, we conducted additional experiments at four sites (CN-Din, DE-Meh, IE-Dri, and JP-SMF). Run A started from the shared default-parameter restart, while Run B used 20 additional annual spin-up cycles with the calibrated parameters. Both runs used identical forcing, parameters, and executable.

The results show that additional spin-up did not systematically improve agreement with observations. As shown in Figure R2-2b, the changes in RMSE for Qh and Qle were no greater than 0.37 W $m^{-2}$ at CN-Din, IE-Dri, and JP-SMF. At DE-Meh, additional spin-up increased the mean RMSE by 2.22 W $m^{-2}$ for Qh and 1.45 W $m^{-2}$ for Qle.

Figure R2-1 further illustrates the model adjustment and convergence during the additional spin-up. All parameter sets met the convergence criterion within 15 annual cycles, and the largest half-hourly RMS difference

between cycles 19 and 20 was only 0.0052 W m$^{-2}$. Figure R2-2a shows the monthly RMS differences between the two initialization methods during the first simulation year, with relatively larger differences at DE-Meh.

We will include the initialization procedure and these results in the revised manuscript. The additional ILS evaluations on independent test periods used 20 annual spin-up cycles.

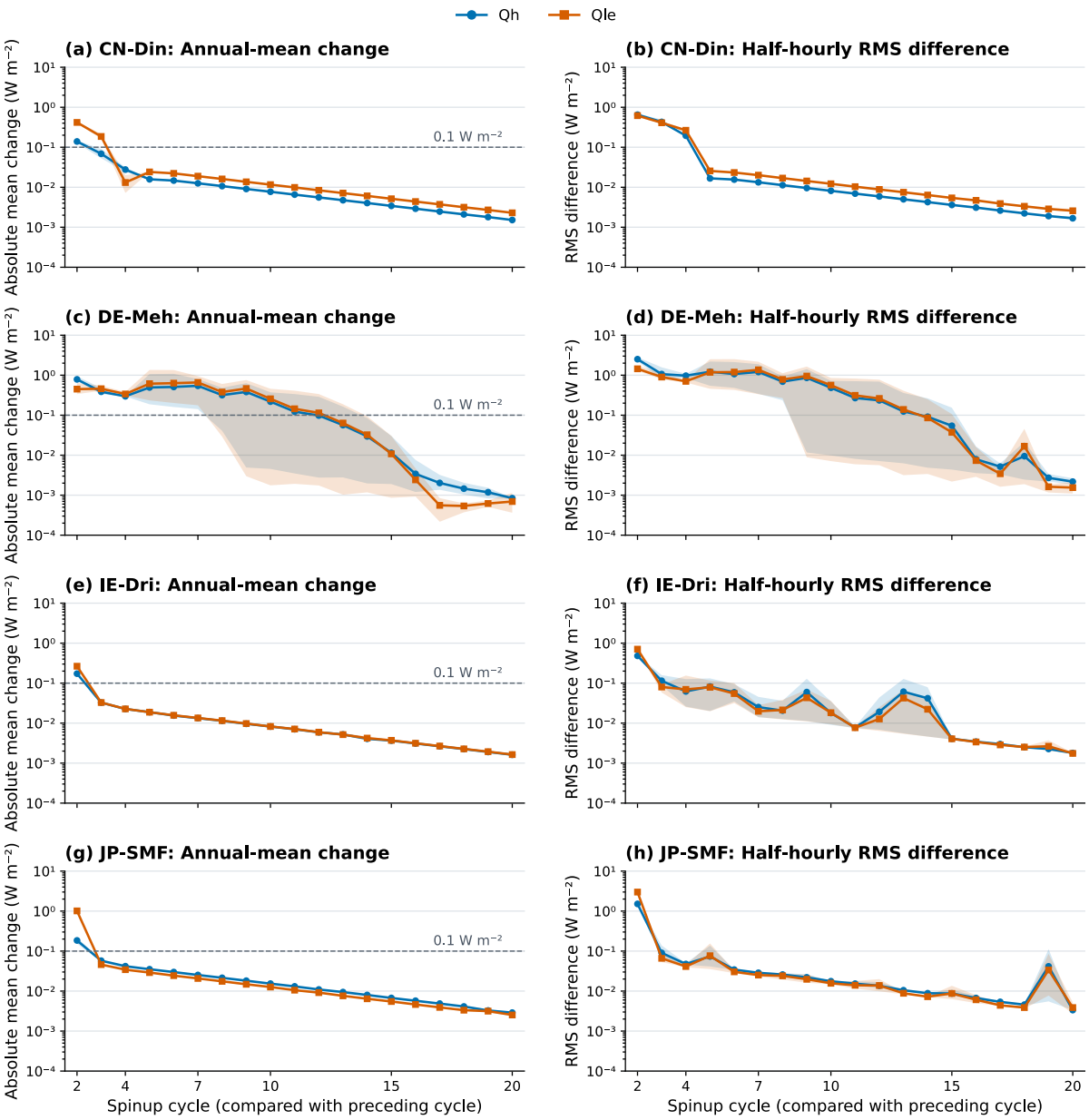



Figure R2-1. Changes in simulated Qh and Qle over 20 annual spin-up cycles at four sites. Left panels show absolute changes in annual-mean fluxes between successive cycles; right panels show half-hourly RMS differences. Lines show the means across three parameter sets and shading shows their ranges. The dashed line marks 0.1 W m$^{-2}$.

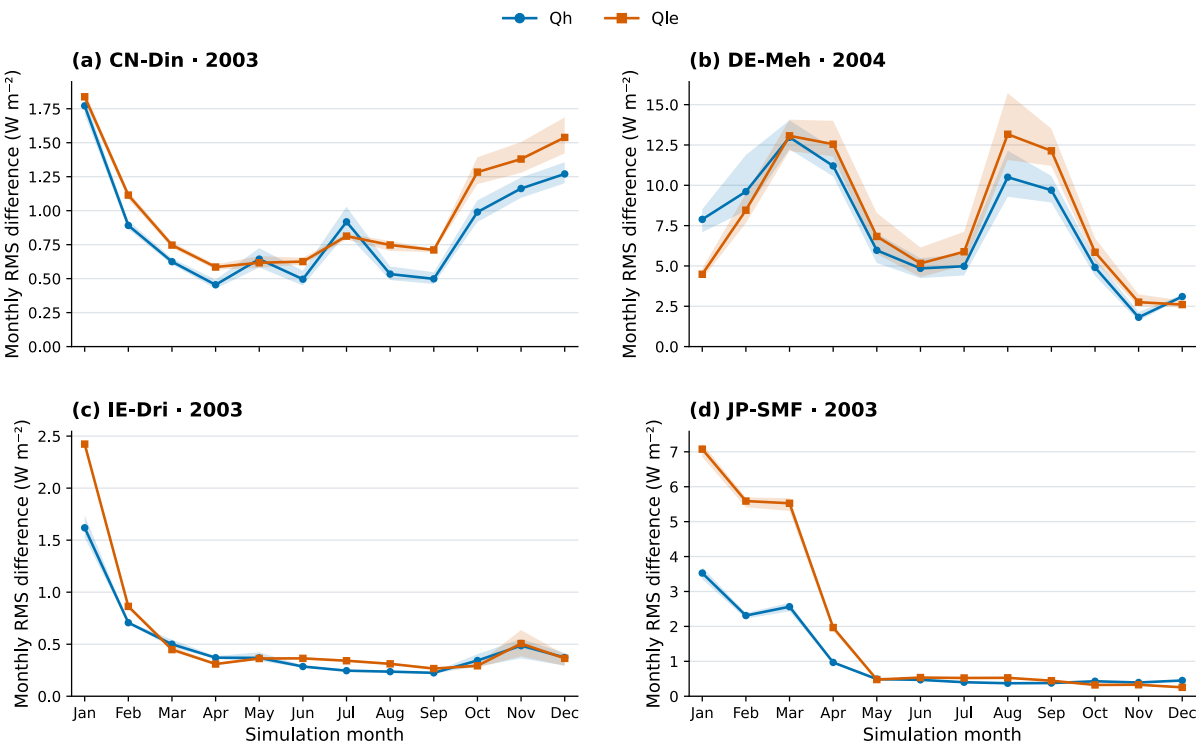



Figure R2-2a. Monthly RMS differences between first-year ILS fluxes initialized from the shared default restart and after 20 additional annual spin-up cycles with the same calibrated parameters. Lines show means across three parameter sets and shading shows their ranges.

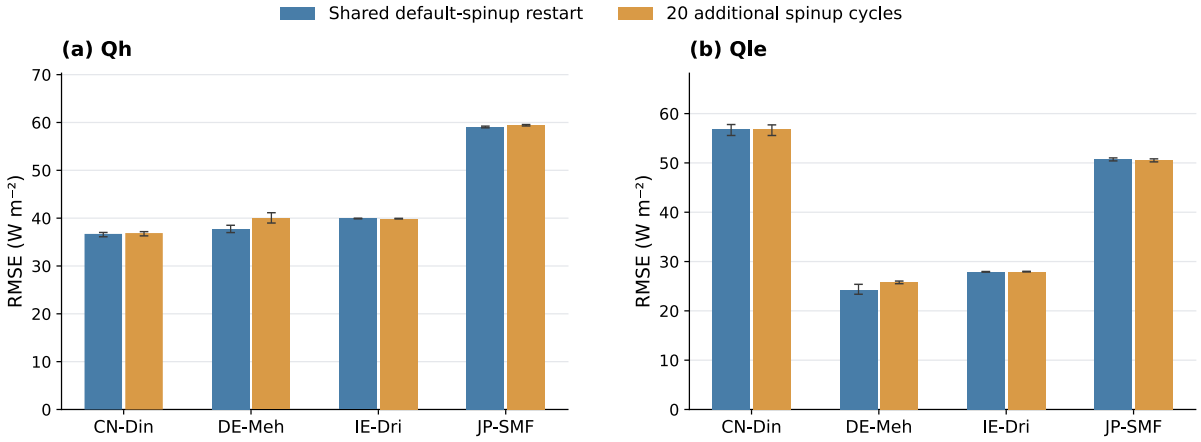



Figure R2-2b. First-year RMSE against observations for ILS initialized from the shared default restart and after 20 additional annual spin-up cycles using the same calibrated parameters. Bars show means across three parameter sets and whiskers show their ranges.

MATSIRO6 Document Writing Team (2021). *Description of MATSIRO6*. CCSR Report No. 66, Division of Climate System Research, Atmosphere and Ocean Research Institute, The University of Tokyo. https://doi.org/10.15083/0002000181

**Reviewer Comment 7:**

Fig. S1: the sensitivity analysis appears to have been run at a single site, ES-LgS, which is not one of the 20 study sites. Could the authors explain this, and justify using one site to select parameters for all four PFTs? The size of the candidate pool would also be useful.

**Authors' Response:**

Thank you for requesting this clarification. ES-LgS was used as an external site for preliminary parameter screening, so that this step was conducted separately from calibration at the 20 target sites. The candidate pool comprised all 21 PFT-dependent parameters defined in ILS. We ranked their sensitivity using the changes in simulated Qh and Qle under parameter perturbations and selected the 12 highest-ranked parameters for subsequent calibration.

We agree that a sensitivity ranking obtained at one site should not be interpreted as a universal ranking for all PFTs. We will clarify this distinction in the Methods and the caption of **Fig. S1**, and provide the complete candidate list and sensitivity scores in the Supplement.

**Reviewer Comment 8:**

L269–275: the surrogate is pretrained on one-at-a-time perturbations only, then used to optimise 12 parameters jointly. Can the authors comment on the consequences, and ideally show whether adding some joint samples to the pretraining changes the calibrated parameters?

**Authors' Response:**

Thank you for raising this point. The main limitation of one-at-a-time sampling is that parameter interactions are not systematically represented. Therefore, accurate surrogate predictions for individual parameter perturbations do not necessarily guarantee accuracy when multiple parameters change simultaneously.

To assess this limitation, we compared the surrogate and ILS responses at the calibrated parameter combinations across four sites. The Qle response correlations ranged from 0.568 to 0.623 at CN-Din, 0.055 to 0.904 at DE-Meh, 0.799 to 0.821 at IE-Dri, and 0.879 to 0.897 at JP-SMF. These results indicate varying surrogate accuracy under joint parameter changes, as discussed in our response to **Comment 9**.

Regarding the addition of joint parameter samples, we have not conducted a controlled comparison of surrogate training with and without these samples. Therefore, we cannot determine whether adding joint samples would change the calibrated parameter values. However, all 12 calibrated parameter sets were evaluated directly using ILS against independent test-period observations, as reported in our general response. We will explicitly acknowledge this limitation and identify joint-parameter sampling as a direction for future improvement.

**Reviewer Comment 9:**

L283–291: I am not convinced this screening tests what the calibration depends on. Pearson's R between two half-hourly flux series is dominated by the diurnal cycle, so a surrogate responding only weakly to the parameters could still pass a threshold of 0.9, and R ignores bias, which is what the MSE-based calibration acts on. Could

the anomalies relative to the default run be compared instead, (ILS(p) − ILS($p_0$)) against the same for the surrogate, with RMSE or bias reported alongside R? It would also help to justify the 0.9 threshold, to say how the 100 joint samples were drawn, and to report the surrogate error at the calibrated parameters themselves, since that is where it matters most.

**Authors' Response:**

Thank you for this suggestion. We agree that the correlation between full flux time series can be influenced by the diurnal cycle, and therefore a high R value does not necessarily indicate that the surrogate accurately reproduces the flux responses to parameter changes.

We would like to clarify that the surrogate was trained to minimise the prediction errors of Qh and Qle. The original median R > 0.9 threshold was an empirical criterion for preliminary screening of agreement between simulated flux time series. The 100 joint parameter samples used for screening were randomly generated within the prescribed parameter ranges (**Table 1**).

Following the reviewer's suggestion, we further compared the flux changes relative to the default parameter configuration between the surrogate and ILS at the trained calibrated parameters. The corresponding correlations and RMSE values are reported in **Table R2-2**. We also directly evaluated the surrogate predictions at the final calibrated parameters using RMSE, bias, and R, as shown in **Table R2-3**.

These results show that the surrogate accuracy in reproducing parameter-induced flux changes varies among sites. This confirms that the original correlation-based screening alone is insufficient to fully assess the surrogate's ability to reproduce parameter responses. We will explicitly acknowledge this limitation in the revised manuscript.

However, the primary objective of this study is to use the surrogate to support parameter calibration. Therefore, we directly evaluated all 12 calibrated parameter sets using ILS against independent test-period observations. The results show that both Qh and Qle had lower RMSE than the default ILS simulations at all four sites, as reported in **Table R2-1**.

Thus, despite errors in the surrogate's representation of parameter responses, the independent ILS evaluations demonstrate that the calibrated parameters improved simulation performance at these sites. We will clarify the distinction between surrogate prediction accuracy and the performance of the final calibrated ILS simulations, without interpreting a high correlation coefficient as evidence of accurate parameter-response predictions.

Table R2-2. RMSE between the flux changes simulated by the surrogate and ILS during the test period.

| **Site** | **Response RMSE: Qh (W m$^{-2}$)** | **Response RMSE: Qle (W m$^{-2}$)** |
|---|---|---|
| CN-Din | 24.52–26.44 | 34.12–36.87 |
| DE-Meh | 22.14–35.86 | 19.01–35.65 |
| IE-Dri | 5.08–5.14 | 6.95–7.43 |
| JP-SMF | 37.20–39.90 | 30.90–34.53 |

Table R2-3. Agreement between surrogate and ILS flux predictions at the final calibrated parameters during the test period.

| Site | Flux | RMSE (W m$^{-2}$) | Bias (W m$^{-2}$) | R |
|---|---|---|---|---|
| CN-Din | Qh | 19.84 | +8.39 | 0.970 |
| CN-Din | Qle | 34.00 | −17.30 | 0.977 |
| DE-Meh | Qh | 23.96 | +4.59 | 0.949 |
| DE-Meh | Qle | 26.44 | −9.04 | 0.960 |
| IE-Dri | Qh | 4.01 | −1.08 | 0.994 |
| IE-Dri | Qle | 7.16 | −2.22 | 0.996 |
| JP-SMF | Qh | 34.27 | +9.72 | 0.883 |
| JP-SMF | Qle | 29.34 | −5.87 | 0.977 |

**Reviewer Comment 10:**

On a related point, the surrogate sees only forcing, parameters and day of year over 48-step sequences, with no soil moisture or other ILS state. How can it then represent parameters that act through soil water storage over weeks to months, as psicr does (L264–266)? Is the LSTM hidden state carried over between sequences or reset?

**Authors' Response:**

Thank you for requesting clarification of how input sequences and LSTM states are handled. The LSTM hidden and cell states are reset for each input window and are not carried between windows. The implemented sequence lengths are 48 or 96 half-hourly steps, corresponding to one or two days, respectively. We will correct the description of these durations and specify the configurations used.

The surrogate is trained against fluxes from continuous ILS simulations and learns their relationship with the forcing, parameter values and seasonal information available within each window. However, soil water storage is not explicitly supplied or propagated as a surrogate state. The short input windows do not explicitly account for how earlier soil conditions affect the simulated fluxes, including their response to *psicr*.

In the final evaluation, the calibrated parameters are applied to ILS, which continuously evolves its soil water and other physical states throughout the simulation. The additional ILS evaluation during the test period therefore tests the practical performance of the resulting fixed parameters with soil water and other physical states evolving continuously.

**Reviewer Comment 11:**

L320–321: the generator produces time-varying parameters but ILS is run with their average, so the vector actually used was never itself optimised. It would help to see p(t) at a few sites, and to compare surrogate performance under p(t) with that under the average.

**Authors' Response:**

Thank you for this suggestion. We have examined both the temporal variation of the generator outputs and the effect of replacing them with the final fixed parameter set.

First, the figure in our response to **Reviewer 1, Comment 7** shows the generator outputs before averaging at representative sites. The figure shows that their temporal behaviour differs among sites and parameters, including relatively stable outputs and systematic changes over time.

Second, in the additional experiment at four sites, we compared two predictions from the same trained surrogate: one using the parameters generated by g_z for each input window, and the other using the final fixed parameter set used in ILS. The forcing data, training period and valid observation times were identical between the two predictions. **Table R2-4** reports the resulting RMSE against observations, averaged across the three random seeds.

Replacing the window-specific parameters with the final fixed parameter set had little effect on training-period surrogate RMSE at CN-Din and IE-Dri. At DE-Meh, Qle RMSE increased from 26.49 to 35.85 W $m^{-2}$, while Qh RMSE changed little. At JP-SMF, Qh RMSE increased from 43.57 to 52.83 W $m^{-2}$ and Qle RMSE from 43.73 to 46.07 W $m^{-2}$.

We will clarify that the transformed parameter values are averaged, followed by conversion back to physical parameter values, and that the resulting fixed parameter set is not separately re-optimised. Our objective is to obtain a single fixed parameter set for each site. Although the averaged parameters were not separately optimised, all 12 final parameter sets were directly evaluated in ILS against independent test-period observations, as reported in **Table R2-1**.

Table R2-4. Training-period RMSE of surrogate predictions using parameters generated for each input window and the final fixed parameter set at four sites. Values are reported in W $m^{-2}$ and averaged across the three random seeds.

| Site | Qh: Time-varying | Qh: Fixed | Qle: Time-varying | Qle: Fixed |
|---|---|---|---|---|
| CN-Din | 34.55 | 34.57 | 40.87 | 40.87 |
| DE-Meh | 42.57 | 42.48 | 26.49 | 35.85 |
| IE-Dri | 31.41 | 31.53 | 23.14 | 23.79 |
| JP-SMF | 43.57 | 52.83 | 43.73 | 46.07 |

**Reviewer Comment 12:**

Eqs. 1–2: how are the two losses weighted, and are both computed on standardised fluxes?

**Authors' Response:**

Thank you for requesting this clarification. Separate MSE losses are calculated for Q_h and Q_le after each flux has been standardized using its corresponding mean and standard deviation.

During surrogate training, the two task losses are applied sequentially to update the shared network and the corresponding task-specific branch. During parameter calibration, the flux loss is their sum.

**Reviewer Comment 13:**

L345–348: excluding the two sites that fail the screening will bias the reported means upwards so could the results for all 20 sites also be given? ID-Pag failed the screening but is still used in the multi-task comparison (Table S5), where it shows the largest gain of the four sites, and in the leave-one-site-out analysis. Why?

**Authors' Response:**

Thank you for raising this concern. We agree that reporting the full site set alongside the screened subset is necessary to show how screening affects the summarized performance.

To examine this effect, we recalculated the results for all 20 sites and for the 18-site subset excluding ID-Pag and AU-Emr.

Table R2-5. Mean relative RMSE changes from default to MdPL-calibrated ILS for all 20 sites and the 18 sites passing the surrogate–ILS consistency screening. Negative values indicate reductions in RMSE.

| Site set | Mean Qh RMSE change (%) | Mean Qle RMSE change (%) |
|---|---|---|
| All 20 sites | −14.01 | −12.86 |
| 18 sites passing the consistency screening | −11.27 | −11.92 |

In this comparison, excluding the two sites reduces the estimated average improvement rather than increasing it. The screening criterion concerns agreement between the surrogate and ILS, rather than the magnitude of the calibration gain. We will report both site sets and provide the individual default and calibrated ILS metrics, together with each site's screening status, in the supplementary site-level table described in our response to **Comment 16**.

Regarding ID-Pag in **Table S5**, the single-task versus multi-task comparison evaluates the relative effect of surrogate architecture on calibration performance. A substantial improvement with the multi-task architecture can coexist with insufficient surrogate–ILS consistency. ID-Pag is therefore informative for this architectural comparison, but its improvement does not establish that it passes the consistency criterion. We will explicitly identify its screening status in **Table S5** and distinguish the architectural comparison from the consistency-screened main analysis.

The leave-one-site-out experiment serves a different, specific purpose. When ID-Pag is the held-out target, the transferred parameter vector is obtained by averaging calibrated parameters from other sites and is then applied directly to ILS at ID-Pag. Its local surrogate is not used to generate this transferred vector. Failure of the local surrogate consistency screening therefore does not, by itself, preclude evaluating parameter transfer at ID-Pag. We will clarify this procedure and explicitly document the donor and target site sets.

**Reviewer Comment 14:**

L364, L366 and L534–540: are the Sobol ranges the same as the Table 2 ranges available to MdPL? The Sobol sets are also ranked on RMSE(Qh) + RMSE(Qle) in physical units, while MdPL minimises a standardised loss. The efficiency argument would be much more convincing with wall-clock or CPU/GPU costs for both approaches, and I am not sure pretraining "can be amortized… across multiple sites" when it uses 48 runs per site and one model is trained per site.

**Authors' Response:**

Thank you for identifying these distinctions. The original Sobol sampling ranges were not identical to the MdPL ranges for every PFT, particularly for vegetation height at the grassland and cropland sites. We have therefore restricted selection of the best Sobol candidate to parameter vectors within the MdPL bounds and satisfying the joint physical constraint.

We have also recalculated the Sobol ranking and revised the comparison figure using the sum of the two MSEs in physical units. The best Sobol vector is selected solely by minimizing this quantity. For the original-paper MdPL and default ILS results, each flux MSE is obtained by squaring its corresponding reported RMSE.

Table R2-6. Comparison of summed MSE and summed Pearson correlation coefficients for Qh and Qle between MdPL and the best Sobol-sampled parameter sets at four sites. The best Sobol set is selected by minimizing the summed MSE. MSE sums are reported in $W^2\ m^{-4}$.

| Site | MdPL MSE sum | Best Sobol MSE sum | MdPL R sum | Best Sobol R sum |
|---|---|---|---|---|
| CN-Din | 3721.55 | 4152.68 | 1.653 | 1.544 |
| DE-Meh | 2154.64 | 3329.09 | 1.644 | 1.675 |
| DE-Seh | 2964.99 | 2809.82 | 1.557 | 1.577 |
| IE-Dri | 1511.99 | 2853.91 | 1.671 | 1.608 |

MdPL has a lower MSE sum at three sites, while Sobol performs better at DE-Seh. The correlation comparison also shows site-dependent differences. We will describe these results accordingly.

The trained surrogate can be reused for repeated calibrations at the same site, as demonstrated by the repeated parameter-generation network trainings in our additional experiment. We will describe this site-specific reuse explicitly.

During calibration, parameter updates use gradients through the trained surrogate, avoiding a full ILS simulation at every update. A systematic computational-cost comparison will be addressed in future work, including ILS training-data generation, surrogate training, parameter calibration and final ILS evaluation.

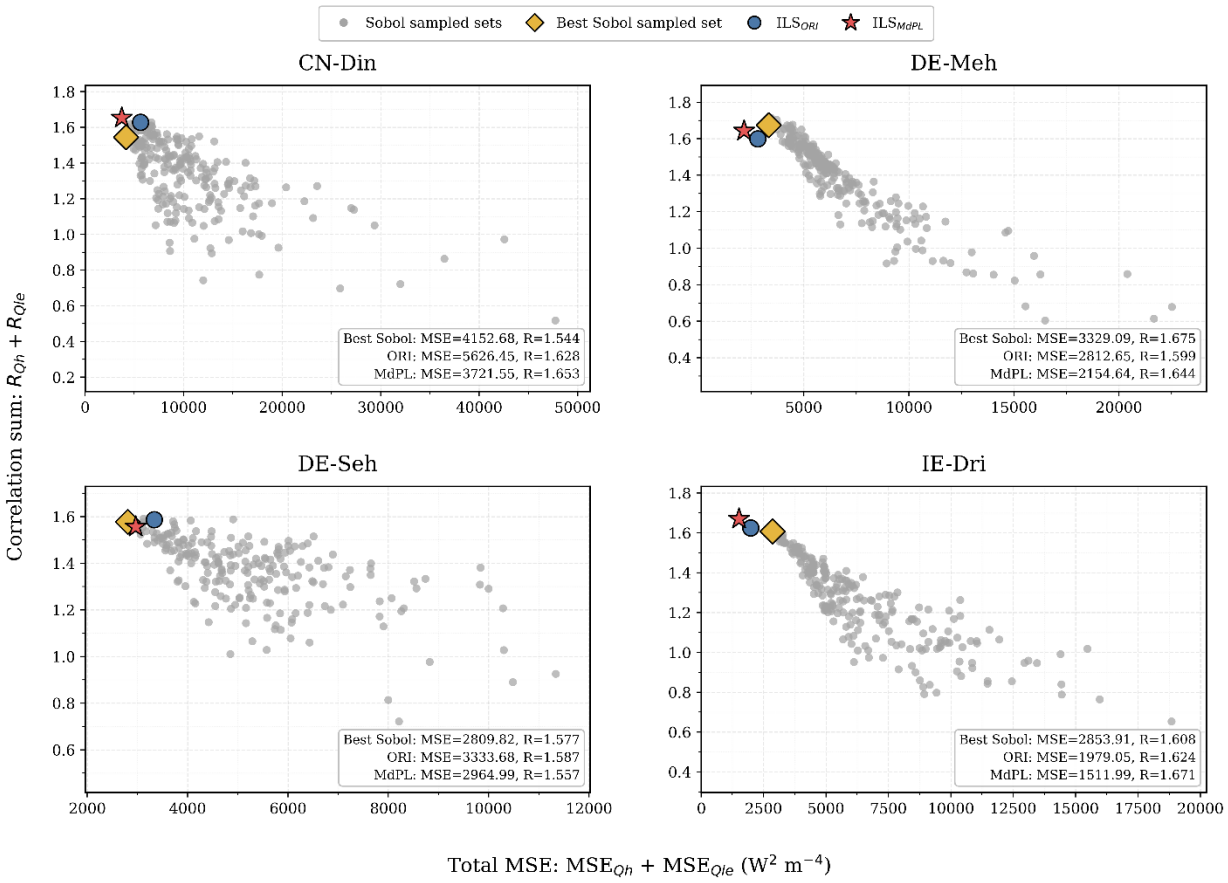



Figure 6. Comparison of ILS performance using default, MdPL-calibrated, and Sobol-sampled parameter sets at four sites. The horizontal axis shows the sum of the Qh and Qle MSE values, and the vertical axis shows the sum of their Pearson correlation coefficients. The best Sobol parameter set at each site is selected by the lowest summed MSE.

**Reviewer Comment 15:**

L419: "significantly improved" reads as a statistical claim, but no test is reported.

**Authors' Response:**

Thank you for pointing this out. We will replace "significantly improved" with "improved" and report the corresponding numerical changes.

**Reviewer Comment 16:**

No results are given with any uncertainty: no repeated seeds, generator ensembles or confidence intervals. Given the equifinality the authors rightly discuss, the spread of p* across random initialisations would be interesting in itself and would show whether the gains are robust. Site-level metrics would also be easier to read as a table than from the Taylor diagrams.

**Authors' Response:**

Thank you for this suggestion. We trained the parameter-generation network with three random seeds at each of four sites. Figure R2-3 shows three parameter sets per site, alongside the default ILS and their mean test RMSE. We will include all evaluated candidates in the Supplement.

Both fluxes had lower test RMSE than the default ILS for all three parameter sets at each site. Across the eight site–flux combinations, the difference between the highest and lowest RMSE among these sets ranged from 0.04 to 6.22 W m$^{-2}$.

**Figure R2-3c** shows the spread of all 12 final parameters across the three seeds. To compare parameters with different units, the maximum-minus-minimum value is expressed as a percentage of each parameter's prescribed interval. The largest spans at CN-Din, DE-Meh, IE-Dri and JP-SMF were 11.4%, 43.9%, 5.5% and 3.8%, respectively, corresponding to *effcon*, *vmax0*, rlfn and *rlfv.* Together, these results allow parameter variability and predictive performance to be assessed separately.

We will include the complete parameter values and performance metrics for each random seed in the Supplement. We will also add a site-level table of default and calibrated ILS RMSE and correlation coefficients for the original 20-site analysis, with the screening status indicated, so that individual-site results can be read directly alongside the Taylor diagrams.

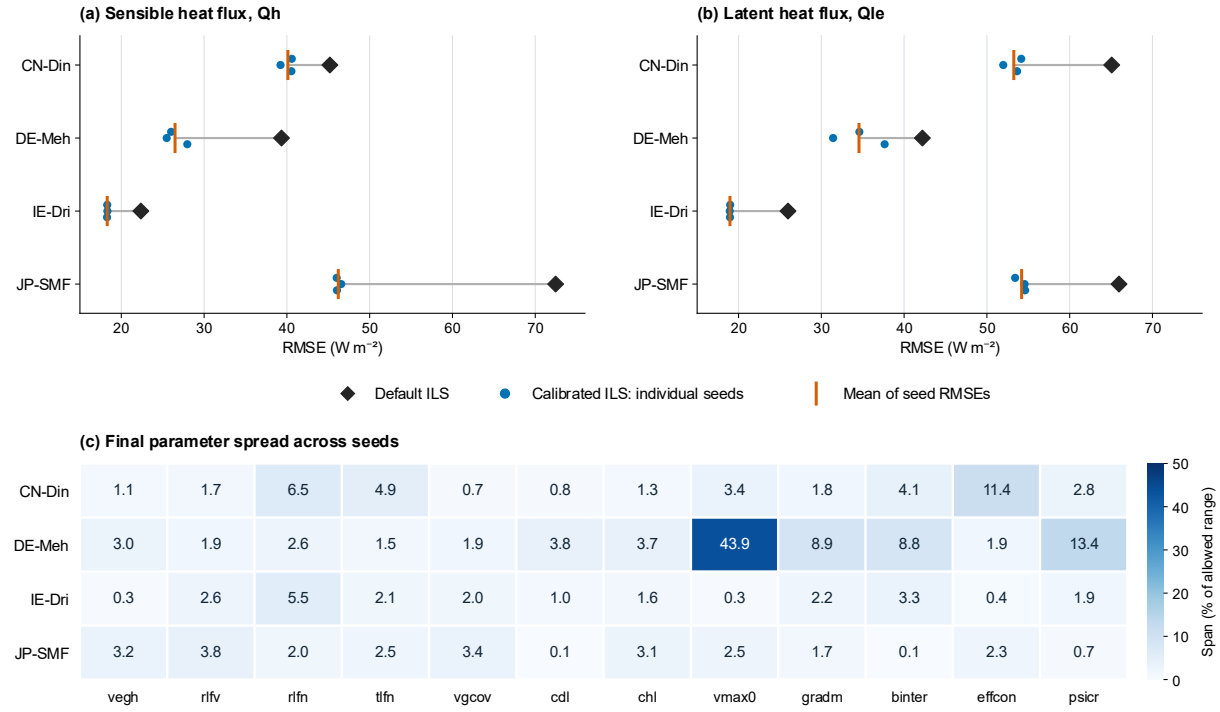



Figure R2-3. ILS test performance and final parameter variability for three parameter sets at four sites. (a–b) Test RMSE for Qh and Qle, showing parameter set, their mean RMSE and the default ILS. (c) Parameter ranges across three sets, expressed as percentages of their prescribed intervals.

**Reviewer Comment 17:**

L420–428: the PFT-mean weights each PFT equally despite 3–6 sites per PFT; could the site-mean also be given? The decrease in latent heat correlation at cultivation sites (−4.8%) is interesting and could be explored a little more.

**Authors' Response:**

Thank you for this suggestion. The original PFT-level mean assigns equal weight to each PFT to avoid dominance by groups with more sites. We have additionally calculated the site-level mean, giving equal weight to each site. The relative RMSE changes are calculated from the mean RMSE values under each weighting method.

Table R2-7. Comparison of relative RMSE changes under equal PFT weighting and equal site weighting for all 20 sites and the 18 sites passing the surrogate–ILS consistency screening.

| Site set | Aggregation | Qh RMSE change | Qle RMSE change |
|---|---|---|---|
| All 20 sites | Equal weighting of PFTs | −14.01% | −12.86% |
| All 20 sites | Equal weighting of sites | −13.26% | −12.45% |
| Screened 18 sites | Equal weighting of PFTs | −11.27% | −11.92% |
| Screened 18 sites | Equal weighting of sites | −10.99% | −11.79% |

The differences between the two aggregation methods are less than one percentage point for both fluxes, indicating that the overall RMSE improvements are not strongly affected by the weighting method.

Regarding the 4.8% decrease in Qle correlation at cultivation sites, the calibration minimises squared errors rather than maximising correlation. Therefore, a reduction in RMSE does not necessarily imply an improvement in correlation. We will clarify this distinction in the revised manuscript.

**Reviewer Comment 18:**

Fig. 4: the diagnostic is confounded for vgcov, whose default of 1 is also its upper bound, so the default itself counts as boundary-proximal. Counts per parameter would help.

**Authors' Response:**

Thank you for identifying this ambiguity. The default vgcov value is 1 at all sites, which is also its prescribed upper bound. Its closeness to the upper bound therefore cannot, by itself, be attributed to calibration.

Using the reported values in **Tables S1–S4** and the existing threshold of 5% of the prescribed parameter interval, all 18 default *vgcov* values are at the upper boundary. After calibration, the values remain close to the upper bound at three sites: CN-Din (0.96), CN-Cha (0.99), and DE-Meh (1.00). At PT-Esp, the value of 0.11 is close to the lower bound. Overall, *vgcov* decreases from its default value at 17 of the 18 sites.

We will revise the interpretation of **Fig. 4** accordingly and provide the numbers of values close to the lower and upper bounds for each parameter, together with the corresponding default positions. This will distinguish parameters that were already near a boundary in the default configuration from those that approached a boundary during calibration.

**Reviewer Comment 19:**

My main comment on this section is that the calibrated parameters are called "effective parameters" and then left there. In Tables S1–S4 several move the same way almost everywhere: vgcov is reduced at 17 of 18 sites, cdl increased at 16, vmax0 reduced at 15. That consistency suggests a systematic bias rather than site-specific noise. Lower cover and lower photosynthetic capacity would be consistent with the model over-estimating transpiration, or mis-partitioning energy between canopy and soil. Alternatively, since tower Qh + Qle usually falls short of the available energy, calibrating to uncorrected fluxes would push the parameters to suppress turbulent fluxes: which flux product was used here, corrected for energy balance closure or not? Could these shifts be reported and related to the biases of the default run in Qh, Qle and their partitioning? That would make the calibration genuinely useful for model development. And if vmax0 falls nearly everywhere, would we not expect GPP to fall too?

**Authors' Response:**

Thank you for highlighting the value of these systematic parameter changes for understanding model behaviour. We have examined the parameter changes in **Tables S1–S4** together with the default and calibrated ILS fluxes at all 18 retained sites.

Across these sites, *vgcov* decreases at 17 sites, *cdl* increases at 16 sites, and *vmax0* decreases at 14 sites. For *vgcov*, however, the default value is 1 at every site, already at the upper bound of its prescribed range. The predominance of downward adjustments therefore needs to be interpreted alongside their magnitude and the resulting flux responses, rather than taken alone as evidence of a common model deficiency.

Using the corresponding calibration observation product at each site, the default ILS overestimates mean Q_le at 11 of the 18 sites. Following calibration, mean Q_h increases while mean Q_le decreases at 10 sites; both fluxes decrease at only four sites. For example, at IE-Ca1, the Q_le mean bias decreases from +23.81 to +2.39 W $m^{-2}$, while the Q_h mean bias changes from −26.47 to −6.10 W $m^{-2}$. Thus, the calibrated parameter combination reduces opposing biases in the two fluxes. These reductions in opposing biases are consistent with adjustments to canopy exchange and energy partitioning.

Regarding energy-balance closure, calibration used the PLUMBER2 Q_h and Q_le observations corrected for energy-balance closure at 11 sites and uncorrected observations at seven sites. We have re-evaluated the existing ILS simulations using the corresponding observation product consistently at each site. Q_h RMSE improves at 16 of the 18 sites and Q_le RMSE at 17 sites. Moreover, seven of the 11 sites calibrated against corrected observations also show increased Q_h and decreased Q_le. Total Q_h + Q_le nevertheless decreases at 11 sites, so observational uncertainty associated with energy-balance closure remains relevant. We will explicitly report the observation products and these site-level diagnostics.

The GPP results partly follow the reviewer's expectation: mean GPP decreases at three of the four evaluated sites. At CN-Din, mean GPP decreases from 8.66 to 4.56 μmol $CO_2$ $m^{-2}$ $s^{-1}$, approaching the observed mean of 3.96. At IE-Dri, GPP instead increases from 4.35 to 6.38 μmol $CO_2$ $m^{-2}$ $s^{-1}$ against an observed mean of 4.91, and performance deteriorates despite the reduction in vmax0. Since the parameters were calibrated jointly, the GPP changes cannot be attributed to vmax0 alone. At DE-Meh and DE-Seh, mean GPP bias decreases, although half-hourly RMSE increases.

These results relate the systematic parameter changes to changes in Qh, Qle and GPP, suggesting that canopy exchange and carbon–water coupling warrant further investigation.

**Reviewer Comment 20:**

Sect. S1: the multi-task and single-task results are close and go in both directions (mean Qh RMSE 42.82 vs 41.43 W $m^{-2}$, Qle 37.8 vs 39.5). With four sites and no uncertainty, I am not sure this supports "strong collaborative effects between key variables" (S1, L51–52).

**Authors' Response:**

Thank you for pointing this out. We agree that the four-site comparison without uncertainty analysis does not support the claim of strong collaborative effects. The multi-task configuration reduced mean Qle RMSE from 39.50 to 37.80 W $m^{-2}$, while mean Qh RMSE increased from 41.43 to 42.82 W $m^{-2}$.

We will remove the claim of strong collaborative effects and revise the statement to: "Across the four evaluated sites, the multi-task configuration achieved lower mean Qle RMSE and higher mean correlations for both fluxes, while mean Qh RMSE was higher than in the single-task configuration."

**Reviewer Comment 21:**

Sect. S2: as above, this analysis belongs in the main text. A few questions on it:

•Why is grassland excluded, and why are only 3 of the 5 cultivation sites used?
• How are the held-out site's parameters produced: as one set per PFT, or by applying g_z to that site's attributes?
• How well does the multi-site generator fit the sites it was trained on? Only the held-out performance is given.
• Is the surrogate trained per site, or pooled across sites?

This is also the only place where the site attributes can matter. In the single-site experiments A is constant during training (L160–163), so it cannot inform the parameters.

**Authors' Response:**

Thank you for these comments. We will move the leave-one-site-out analysis into the main text and clarify the experimental procedure.

The transfer analysis included evergreen forest, woodland, and three cultivation sites, but did not cover grassland or the remaining two cultivation sites. We acknowledge that this limited site coverage was not adequately explained. We will clarify the scope of this exploratory analysis and avoid drawing general conclusions about transferability across all PFTs.

For each held-out site, the calibrated parameter vectors from the other sites within the same PFT were averaged and directly applied to ILS at the target site. The surrogate and parameter generator were trained separately at each site. No multi-site generator was trained, and therefore there is no multi-site training performance to report. The performance of the individual site-specific calibrations is presented in the main text.

We also agree that the static site attributes remain constant within each site-specific training experiment and therefore cannot provide information about parameter relationships across sites. Learning such relationships would require training a shared generator across multiple sites.

We will explicitly identify the donor and target sites and present the transfer results alongside the site-specific calibration results. The limited transfer performance, particularly at cultivation sites, will also be discussed.

**Reviewer Comment 22:**

The framework is described as scalable (L584–585, L597), but as implemented it is trained per site, needs flux observations and depends on passing the surrogate screening. How would it be applied to grid cells where no flux observations exist? I would also welcome a paragraph on the limitations of the study; those given in Supplement Sect. S2 would sit more naturally here.

**Authors' Response:**

Thank you for this suggestion. We will move the leave-one-site-out analysis into the main text and describe the framework's demonstrated application as site-level parameter calibration.

For application to grid cells without local flux observations, a future extension would train a shared parameter-generation network across multiple sites. Meteorological forcing and land-surface attributes would provide its inputs, while flux observations at the training sites would constrain parameter learning. The trained network could then generate parameters for new grid cells using their available forcing and land-surface characteristics. We will describe this development pathway in the Discussion and revise the scalability statements accordingly.

**Reviewer Comment 23:**

L45: I do not think Exbrayat et al. (2014) supports this statement, that paper is about microbial decomposition and spin-up in CMIP5 models.

**Authors' Response:**

Thank you for pointing this out. We agree that Exbrayat et al. (2014) does not support the specific statement about soil hydraulic conductivity and minimum stomatal conductance. We will remove this citation from the sentence and replace the specific examples with a more focused statement:

"Parameter calibration in LSMs is challenging because model parameters may be weakly constrained by the observations available for calibration."

**Reviewer Comment 24:**

L69–71: something has gone wrong with this sentence: "dPL, e.g., the variable infiltration capacity (VIC) model, leverages…".

**Authors' Response:**

Thank you for identifying this unclear sentence. We will revise it to distinguish the parameter-learning method from the process-based model to which it is applied:

"For efficient parameter inference, dPL uses neural networks to estimate parameters of process-based models, such as the Variable Infiltration Capacity (VIC) model (Liang et al., 1994), and optimises these estimates through a differentiable model or surrogate."

Liang, X., Lettenmaier, D. P., Wood, E. F., & Burges, S. J. (1994). A simple hydrologically based model of land surface water and energy fluxes for general circulation models. *Journal of Geophysical Research: Atmospheres, 99*(D7), 14415–14428. https://doi.org/10.1029/94JD00483.

**Reviewer Comment 25:**

L76: HBV is attributed to Tibangayuka et al. (2022); the original HBV reference should be cited.

**Authors' Response:**

Thank you for this suggestion. We will replace Tibangayuka et al. (2022) with the original model reference, Bergström and Forsman (1973).

**Reviewer Comment 26:**

L122–123: CLM5 is cited to "NCAR, 2020" and JULES to McNeall et al. (2024); the model description papers would be more appropriate.

**Authors' Response:**

Thank you for this recommendation. We will replace "NCAR, 2020" with Lawrence et al. (2019), the CLM5 model description paper, and replace McNeall et al. (2024) with Best et al. (2011), which describes the energy and water components of JULES.

Lawrence, D. M., et al. (2019). The Community Land Model Version 5: Description of New Features, Benchmarking, and Impact of Forcing Uncertainty. *Journal of Advances in Modeling Earth Systems*, 11, 4245–4287. DOI: 10.1029/2018MS001583。Wiley Online Library

Best, M. J., et al. (2011). The Joint UK Land Environment Simulator (JULES), model description – Part 1: Energy and water fluxes. *Geoscientific Model Development*, 4, 677–699. DOI: 10.5194/gmd-4-677-2011

**Reviewer Comment 27:**

L226: 20 sites concentrated in Europe and East Asia do not really "ensure geographic and ecosystem diversity", I would soften this.

**Authors' Response:**

Thank you for this suggestion. We will replace "To ensure geographic and ecosystem diversity" with "For this proof-of-concept evaluation, we selected 20 sites spanning four broad PFT groups."

**Reviewer Comment 28:**

L17: ILS is used before it is defined (L19).

**Authors' Response:**

Thank you for noting this. We will introduce "Integrated Land Simulator (ILS)" at its first occurrence in the Abstract and use the abbreviation thereafter.

**Reviewer Comment 29:**

L105 and L108: the g_z notation has not rendered properly – "$g_z(\cdot)$".

**Authors' Response:**

Thank you for pointing this out. We will correct the notation to $g_z(\cdot)$, and use it consistently for the parameter-generation network throughout the text, equations and figures.

**Reviewer Comment 30:**

L128: "leave-one-out" should be "leave-one-site-out".

**Authors' Response:**

Thank you for this correction. We will replace "leave-one-out cross-validation" with "leave-one-site-out cross-validation" throughout the manuscript to make clear that sites are the units being held out.

**Reviewer Comment 31:**

L188–195: L, the number of encoding frequencies, is never given.

**Authors' Response:**

Thank you for noting this omission. We used (L=1), corresponding to a single annual sine–cosine pair with (f_0=1). We will state this value explicitly in Section 2.2 and clarify that the experiments used annual sine–cosine encoding.

**Reviewer Comment 32:**

L177 and Fig. 1: the generator's outputs are called "parameter priors" but are used as point estimates, which suggests a Bayesian treatment that is not carried out.

**Authors' Response:**

Thank you for this clarification. We will replace "parameter priors" with "parameter estimates" in the text.

**Reviewer Comment 33:**

L230: the forcing variables and their units are not listed.

**Authors' Response:**

Thank you for this suggestion. **Table R1-1** in our response to **Reviewer 1, Comment 2** lists the time-varying input variables, their field names and their physical units before standardisation. These include the meteorological variables, leaf area index, and black carbon and dust deposition fluxes. We will include this table in the revised manuscript and clarify the additional inputs supplied to each network.

**Reviewer Comment 34:**

L326: 48 half-hourly time steps is 1 day, not 2 days.

**Authors' Response:**

Thank you for pointing this out. The training configurations include both sequence lengths (48 and 96). We will correct the statement at L326 and report the sequence lengths and their corresponding durations consistently in the Methods.

**Reviewer Comment 35:**

L350: SFE is used without being described – a sentence would help.

**Authors' Response:**

Thank you for this suggestion. Please refer to our response to **Reviewer 1, Comment 14**, for a description of the SFE method, its inputs and its implementation in this study. We will include this description in Section 2.6.2 of the revised manuscript.

**Reviewer Comment 36:**

Eqs. 16–17: ORI is not defined at first use.

**Authors' Response:**

Thank you for pointing this out. We will define ORI as the ILS simulation using the original default parameter configuration at its first occurrence in the revised manuscript.

**Reviewer Comment 37:**

L552 and Fig. 7 caption: observed GPP is attributed to PLUMBER2 in the text and to FLUXNET2015 in the caption.

**Authors' Response:**

Thank you for pointing out the inconsistent attribution. The GPP reference data were obtained from the PLUMBER2 observation files. Their underlying sources are FLUXNET2015 for CN-Din and DE-Seh, and La Thuile for DE-Meh and IE-Dri. We will revise the text and Figure 7 caption to consistently attribute the GPP observations to PLUMBER2.

**Reviewer Comment 38:**

Table 1: AU-Emr has no climate class; PT-Esp, PT-Mi1 and AR-SLu are listed as "Subtropical", which is not one of the classes defined in the note; and PL-wet, a wetland site, is grouped with grassland.

**Authors' Response:**

Thank you for pointing out. We will use consistent Köppen climate labels and matching definitions throughout the table.

Regarding PL-wet, its PFT assignment was taken directly from the vegetation classification prescribed in the ILS boundary-condition dataset. Although PL-wet is ecologically a wetland, the boundary data assign it to vegetation category 7, which corresponds to the grassland parameter category used in our simulations.

**Reviewer Comment 39:**

Table 2: the units of cdl and chl are given as "s m-2", which does not look right for exchange coefficients. Fortran notation ("9.82d-2") is also mixed with plain decimals.

**Authors' Response:**

Thank you for pointing this out. We will correct their units in **Table 2**. We will also use consistent decimal or scientific notation throughout the table, replacing "9.82d-2" and "2.46d-2" with "0.0982" and "0.0246", respectively.

**Reviewer Comment 40:**

References: please check the entries for Gupta (2008), Kuppel (2014), Vrugt (2005) and Rosolem (2012) and Raoult (2024), which is no longer a preprint.

**Authors' Response:**

Thank you for identifying these references. We have checked the entries against the publication records. We will revise the Gupta et al. (2008) entry to cite "Reconciling theory with observations: elements of a diagnostic approach to model evaluation" and correct the author lists for Kuppel et al. (2014) and Vrugt et al. (2005).

We will remove the Rosolem et al. (2012) entry and its associated citation, while retaining the relevant references already cited in the discussion of equifinality and parameter interpretation. We will also replace the Raoult et al. (2024) preprint with the published article, Raoult et al. (2025), in Journal of Advances in Modeling Earth Systems (doi:10.1029/2024MS004733), and update the corresponding in-text citation.