This is a preliminary point-by-point response prepared for the discussion stage to summarize our current position and planned revisions. A complete response, including the corresponding manuscript changes and finalized additional analyses, will be provided together with the revised manuscript.

In this paper, the authors present a unified platform for easily running many experiments to determine important parameters in the earth system and optimize such parameters in an easily configurable manner. Given the necessity of better constraining parameters in land modeling, such a platform is a genuinely important scientific advance that will allow the science of parameter optimization to continue to scale. This will allow differentiable parameter optimization to become more easily usable "out-of-the-box" by other groups and allow for quick and reproducible testing of different parameterization ideas.

The authors present a well-written paper, with great storytelling of the relevant literature in the introduction. My comments primarily involve further explaining their reasoning for certain high-level decisions, but this work will be a valuable addition to the literature in this area.

Response: Thank you very much for your encouraging assessment of ADELM. We also appreciate your positive comments on the manuscript and your constructive suggestions on clarifying the rationale behind our modeling choices, which will help us strengthen the presentation and interpretation of the study.

Major comments

- Section 2 / 2.1: The individual modules of your model are very nicely documented and the reasoning for each well-cited. However, justification of the overall modules included and citations that you have included the modules necessary to accurate simulate ecohydrological processes are currently missing. For example, without justifying the level of complexity used across modules (e.g. by citing land surface models that were the inspiration for your scheme) one could easily ask why a more advanced plant hydraulics scheme was not incorporated. More discussion of the inspiration for your overall model theory (whether that be an LSM directly influencing your model development plans, or other inspiration for your choices) before diving into the details of each module would help convince potential users of your model that the overall theory is just.

Response: Thank you for this important suggestion. We agree that the rationale for the overall process representation should be explained before introducing the individual modules. ADELM v1.0 was designed to retain explicit biophysical pathways linking learned parameters to simulated fluxes and states, while keeping the process complexity manageable for observation-constrained learning. Our intention is not to suggest that the selected processes constitute the only appropriate or most comprehensive ecohydrological representation, but to provide a physically structured and configurable foundation for investigating parameterization choices. In the revision, we will expand the beginning of Sect. 2 to explain this rationale, relate ADELM to established land surface models and distinguish its approach from differentiable frameworks that learn more aggregated process relationships. We will also expand the discussion of how the chosen process representation affects

parameter identifiability. We will discuss how more mechanistic formulations could be incorporated and compared within the same differentiable framework.

- Lines 325-329: Please further justify your inclusion of the Lv,site term in your loss function, as I would consider this a non-standard term in such a loss function. Did using only Lv,sample lead to poor performance? I agree that getting yearly mean fluxes correct is important, but are you not risking over-smoothing your learned parameters with this term, therefore not capturing parameters that represent extreme days, which may be the most important for various applications? Additional justification of the necessity and impacts of this additional loss term would be appreciated.

Response: Thank you for raising this concern. We introduced the site-level loss to constrain differences in multi-year mean fluxes across locations, which are important for learning spatial attribute-to-parameter mappings. To evaluate the consequences of including the site-level loss, we conducted an ablation experiment for Exp. 3 (comb), comparing training with the sample-level loss alone against training with both terms. The results suggest that the additional loss improves cross-site agreement without systematically degrading predictions of high-flux conditions, although some trade-offs occur at other flux levels. We will include this ablation analysis in the Supplement and clarify the complementary roles of the two loss terms in the Methods.

- Lines 392-423: This part of the results deals primarily in qualitative differences between the experiments, which is important, but neglects to discuss in detail the performance differences between the experiments. Across all of Europe, did you parameterization schemes improve model performance? In any individual region, did you parameterization schemes increase performance? Furthermore, as mentioned below, the y-axis range on Figures 6c, 6d, 7c, and 7d make it very hard to glean this information from the figure. A deeper discussion of quantitative performance differences—both overall and in individual regions—across your experiments, along with changes to figures that makes experiment-to-experiment difference more easy to decipher, would substantially improve my understanding of the impact of your parameterization schemes on model performance.

Response: We agree that the original presentation placed too much emphasis on qualitative spatial comparisons. We have conducted additional quantitative evaluations of GPP and ET against gridded reference products. At the European scale, these include spatial Pearson correlation, Lin's concordance correlation coefficient (CCC), root mean squared difference (RMSD), and mean difference. We have also examined how these metrics vary across the eight biogeographical regions. The results reveal contrasting responses between GPP and ET. Learning vegetation-related parameters substantially improves the agreement of simulated ET with the gridded reference products, particularly in flux magnitude. For GPP, however, parameter learning does not consistently improve spatial agreement, and some regions exhibit greater discrepancies than under the prescribed baseline. We will present the European-scale statistics in a new Table, provide the regional metrics in supplementary figures, and expand Sect. 4.2 to discuss these results quantitatively. We will also revise Figs. 6 and 7 to make differences among experiments easier to distinguish.

Discussion section: I appreciate your in-depth discussion of the various ways that your model—both the structure and the way that parameters are retrieved—is not perfect. However, I think that your discussion still could use more specific connection to your results. For example, is there a reason why any of the limitations you mention would mean that your parameter retrieval process really helps ET prediction while being negligible for GPP prediction at the site scale, which I found surprising? Or could any of these limitations specifically explain why there are many regions where experiment 2 harms model performance compared to the baseline for predicting regional GPP? While I appreciate the in-depth analysis of the limitations, there is a need to directly connect these limitations to some of the more surprising results that were found in your experiments.

Response: Thank you for this suggestion. We agree that discussing model limitations is more informative when those limitations are connected directly to the experimental findings. The stronger improvement in ET can be understood partly through the role of the learned vegetation parameters in regulating stomatal resistance and transpiration. In contrast, GPP depends on additional controls, such as light availability, prescribed LAI, and co-limitation between light-limited and CO2-diffusion-limited photosynthesis. The prescribed configuration also already achieves relatively good GPP performance at held-out sites, leaving less room for improvement through the selected parameters. We will revise Sect. 5.3 to connect these process pathways explicitly to the observed ET improvements and relatively small GPP changes.

Minor comments

Introduction, lines 66-77: This is key in introducing what is different about your model compared to other differentiable frameworks, but this is still a bit opaque / confusing here. The key sentence seems to be lines 69-71, “Errors in simulated fluxes or states 70 (compared with observations) therefore constrain selected parameters through the same coupled pathways that generate the model outputs.” Could you be more specific here to help readers? Does the model get to choose which parameters to learn, as opposed to the user? Or is the advance that parameters are easily turned “on / off” to be learned? I assume that this will become clear later in the paper, but being more specific here I believe will clarify the novelty of your advance.

Response: Thank you for pointing out this ambiguity. ADELM does not automatically decide which parameters should be learned. Instead, users can configure each parameter as prescribed, globally learned, or learned as a spatial function of environmental attributes. The central contribution is the ability to change and systematically compare these parameterization choices without modifying the underlying process equations. We will make this point more explicit in the Introduction and Sect. 2.

Line 191-194: Just to clarify, my read here is that all learned spatially variable parameters a) are predicted only via MLPs, not more advanced architectures, and b) must share MLPs of the same dimensions. Is this correct? The writing in these lines could be clarified and made more concise to make this point (or the contrary point, if I am incorrect) more clearly. Furthermore, is this true across

any setup, or just for the setup you chose for your experiments? I think the answer is in this text, but it's a bit convoluted and confusing as written.

Response: Thank you for this question. In the current implementation, each spatially varying learned parameter is represented by a separate MLP with independently trained weights. The MLPs use shared input attributes and network architecture specified through the configuration. The network dimensions can be changed through the configuration. We will clarify this distinction in Sect. 2.2 and describe the network architecture used in the European experiments separately.

Lines 217-218: Here you describe what you chose to do in Section 3, but it would be helpful to know (and this applies to similar descriptions throughout this section) the degree to which this can be configured for a potential user. That is, is your adaptive framework for these learning hyperparameters easily flexible beyond what was chosen for your European experiment?

Response: Training options, including the optimizer, learning rate, number of epochs, weight decay, gradient clipping, early stopping, learning-rate scheduling, and cross-validation protocol, can be specified through the YAML configuration. We will clarify this flexibility in Sect. 2.3, and retain the numerical settings of the European experiment in Sect. 3.2.

Line 254-256: Just want to call out that it is awesome that such thought has been given to defining different cross-validation techniques by default! Bravo!

Response: Thank you very much for this encouraging comment.

Line 275: Cycled how many times? Justification and/or evidence that this was sufficient to reach some sort of equilibrium?

Response: Thank you for this question. We repeated the meteorological forcing for 2000 three times, carrying the model states forward between cycles. To assess the adequacy of the spin-up, we compared water storage at the end of the second and third cycles. The area-weighted mean absolute change in soil water storage was 0.78–1.10% across the four configurations, and changes in snow storage were negligible. These relatively small residual changes suggest that three cycles provide a reasonable spin-up for the European simulations. We will clarify the initialization procedure and this diagnostic in Sect. 3.1.

Line 285: What does "when the quality-control fraction is at least 0.7" mean intuitively, for those not familiar with the specific FluxnetLSM package?

Response: Thank you for pointing out this ambiguity. The quality control fraction refers to the proportion of valid subdaily observations within each day. Daily fluxes were calculated from valid subdaily records and retained only when these accounted for at least 70% of the expected observations. We will explain this criterion more explicitly in Sect. 3.1.

Line 286: Line 275 says you use daily timestep; was ERA-5 Land therefore averaged to the daily timestep? Please describe the processing done here.

Response: Yes, ADELM uses daily meteorological forcing derived from hourly ERA5-Land data. Daily temperature statistics, wind speed, and vapour pressure deficit are calculated from hourly values, and precipitation and radiation are processed from accumulated fields. The data are then prepared on the common 0.1° model grid. We will add these processing details to Sect. 3.1.

Line 289: Can you be more specific about what exactly a "mean-standard-deviation bias correction towards site observations" means? What changes to the data did you perform, exactly, for your site simulations?

Response: Thank you for this suggestion. For site simulations, we use a site- and variable-specific linear transformation that adjusts the mean and standard deviation of ERA5-Land forcing to match the corresponding available tower observations over their overlapping valid days. The resulting transformation is applied to the site's meteorological time series. GIMMS LAI is adjusted similarly at flux sites, and the European gridded simulations use the unadjusted forcing and LAI fields. We will specify the transformation and its application in Sect. 3.1.

Line 343: What is the "site-month metric"? The loss function defines a site-year metric, if I am not mistaken, but I don't believe this site-month metric has been defined at this point? If it has been, can you cite where it has been defined here?

Response: Thank you for identifying this potential confusion. The site-month metric is an evaluation metric and is distinct from the site-level term in the training loss. For evaluation, daily simulated and observed fluxes are aggregated separately for each site, month, and year. NSE is then calculated across the resulting site-month pairs from held-out sites. We will clarify the evaluation procedure in Sect. 3.3.

Figure 5a and related description in the text: statistical comparisons of these distributions (e.g. significance tests) could be helpful in the discussion of how model performance changes across the experiments.

Response: We agree. We have conducted paired Wilcoxon signed-rank tests on the held-out site-level daily NSE values. The results indicate no statistically significant differences among experiments for GPP, whereas the vegetation-focused and combined experiments show significantly better ET performance than the prescribed baseline and hydro-focused experiment. We will add the statistical groupings to Fig. 5a and explain the procedure in its caption.

Figure 6c, 6d, 7c, 7d and related discussion: Again could be helpful to have some statistical significance comparisons between the distributions represented by the box plots. Also, please justify why correlation (r) is a better metric in this figure than coefficient of determination, since this is testing the ability of a model to reproduce truth, not the association between two random variables. Finally, the axes on these sub-panels make it extremely difficult to tell differences between your experiments in locations where the model generally performs well, which is a very interesting part of this figure.

Response: Thank you for these suggestions. To make regional performance differences easier to assess, we will supplement Figs. 6 and 7 with larger panels showing the distributions of grid-wise Pearson correlation and CCC for GPP and ET. The significance groups will be included in the supplementary figures. Regarding the choice of evaluation metrics, we agree that correlation alone does not quantify agreement in flux magnitude. We therefore complement Pearson correlation with CCC, which accounts for differences in means and variability. This is particularly useful when comparing ADELM with gridded reference products that themselves contain uncertainty.

I really appreciate Section 4.3!

Response: We thank the referee for this encouraging feedback.