RC1: 'Comment on egusphere-2025-3517', Anonymous Referee #1, 05 Nov 2025

This manuscript extends the widely-used benchmarking tool ILAMB to include soil moisture and uses it to perform a comprehensive evaluation of soil moisture Earth System Models from CMIP6, both near the surface and to a depth of 1m. It represents a substantial contribution to the field. The evaluation is extensive and reproducible; of particular note is that a variety of different metrics are used, including the relationship between soil moisture and other variables. The presentation quality is high throughout.

- Author Response: We thank the referee for the positive assessment of our manuscript and for the constructive and insightful comments provided below.

Specific comments

Add some sentences mentioning that soil moisture below 1m can be important too (particularly in areas with deep roots), even though evaluating it is beyond the scope of this analysis.

- Author Response: We thank the referee for this suggestion. We added a sentence in the Introduction acknowledging that soil moisture below 1 m can be important for ecosystem functioning, particularly in regions with deep rooting systems, while clarifying that our analysis focuses on surface and rootzone soil moisture. This addition is supported by recent literature (e.g., Stocker et al., 2023; Kühnhammer et al., 2023). This can be found on Lines 107-110 of the revised manuscript.

Line 160: this is an important point you make here – refer back to it in section 3.2

- Author Response: We thank the referee for this suggestion. We added a sentence in Section 3.2 explicitly referring back to the discussion in Section 2.2, noting that some models have first soil layers deeper than 5 cm and that integrating `mrsol` to 5 cm can introduce additional uncertainty in comparisons with ESA-CCI, which likely contributes to part of the spread in surface soil moisture performance. This can be found on Lines 301-305 of the revised manuscript.

Line 212-3 This statement needs citations. Even better would be to add a brief justification of this statement and mention the limitations of these datasets, so that the reader can bear these in mind when interpreting your results.

- Author Response: We thank the referee for this comment. We revised the paragraph in Section 2.3 to add explicit citations for the WECANN GPP, AVH15C1 LAI, and GLEAM ET datasets, and to briefly justify their use as benchmarks. We also expanded the text to clarify that these products are observationally informed estimates derived from models or statistical algorithms, and to note their key limitations and associated uncertainties, so that readers can better interpret the results. This can be found on Lines 200-232 of the revised manuscript.

Line 249 Add a description of how Overall Score is calculated from Bias Score, RMSE Score, Seasonal Cycle Score, and Spatial Distribution Score. Add few words to clarify the “Seasonal Cycle Score” and “Spatial Distribution Score”.

- Author Response: We thank the referee for this suggestion. We added a brief explanation in Section 3.1 describing how the ILAMB Overall Score is constructed as a composite of the Bias, RMSE, Seasonal Cycle, and Spatial Distribution scores following Collier et al. (2018), and we clarified that the Seasonal Cycle Score reflects agreement in the timing and amplitude of the annual cycle, while the Spatial Distribution Score reflects agreement in the spatial patterns of the fields. This can be found on Lines 250-256 of the revised manuscript.

Line 272: I couldn’t see anywhere what time resolution was used to calculate the Taylor plots (e.g. annual, monthly or daily?). Same for fig. 6 and 7.

- Author Response: We thank the referee for pointing this out. Although the monthly temporal resolution is stated in Section 2.1, we added explicit clarification in Sections 3.2 (Line 306) and 3.4 (Line 365) indicating that the Taylor diagrams (Figure 4) and the SM–ecohydrology relationship plots (Figures 6 and 7) are based on monthly mean fields, to make the temporal resolution clear to the reader.

Line 299: what do you mean by “parameters that mask deficiencies in SM representation”? Maybe add a clarifying phrase, or an example.

- Author Response: We thank the referee for this suggestion. We clarified this statement in Section 3.3 by adding a brief explanation and example, noting that compensating errors in parameters such as soil hydraulic properties or effective soil water holding capacity can partially compensate for missing or simplified processes (e.g., groundwater or root water uptake), leading to apparently realistic ET despite underlying deficiencies in soil moisture representation. This can be found on Lines 348-351 of the revised manuscript.

Line 300: I disagree with this statement because I would characterize transpiration as a vegetation-related process. The representation of water flux through the canopy and carbon flux through the canopy typically both rely structurally on very similar parts of the code.

- Author Response: We thank the referee for this comment and agree with the point raised. We revised the text in Section 3.3 to clarify that transpiration, carbon uptake, and phenology are all vegetation-controlled processes and often rely on similar canopy and stomatal formulations in land models. We now emphasize that the differing apparent performance across ET, GPP, and LAI more likely reflects differences in observational constraints and sources of uncertainty, rather than a fundamental separation between vegetation and hydrologic processes. This can be found on Lines 341-346 of the revised manuscript.

Line 319: “However, it is important to note that the ILAMB spatial climatology used in Figures 6 and 7 may be affected by ESA-CCI’s inconsistent spatiotemporal coverage” I don’t understand this sentence – what does the phrase “ILAMB spatial climatology” mean here?

- Author Response: We thank the referee for pointing out this ambiguity. We revised the sentence in Section 3.4 to clarify that the climatologies shown in Figures 6 and 7 are computed from the available gridded data within ILAMB, and that gaps in ESA-CCI’s spatial and temporal coverage can bias the inferred climatological means and relationships, potentially affecting the model–data comparison. This can be found on Lines 380-384 of the revised manuscript.

Technical corrections

Line 43: consider replacing “, most commonly relying” with “. The majority rely” (the existing sentence has an ambiguity about whether “most” refers to “ESMs” or “commonly” which hampers the sentence flow)

- Author Response: We thank the referee for this suggestion. We revised the sentence (Line 40) in the Introduction to remove the ambiguity by splitting it into two sentences and clarifying that the majority of ESMs rely on bucket-type soil moisture schemes.

Line 45: consider replacing “by incorporating” with “using” or “incorporating it into”

- Author Response: We thank the referee for this suggestion. We revised the wording in the Introduction (Line 44) by replacing “by incorporating” with “using” to improve clarity and sentence flow.

Line 121: the minus sign in both units needs to be in the superscript

- Author Response: We thank the referee for noting this. We corrected the unit formatting so that the minus signs in the exponents are properly shown as superscripts (Line 131).

Consider putting mrsol and mrsos in monospace font, given that it is the name of a variable

- Author Response: We thank the referee for this suggestion. We revised the manuscript to format the variable names `mrsol` and `mrsos` in a monospace font to clearly distinguish them as model variable names. We also did this for `gpp`, `lai`, and `evspsbl` whenever these processes are actually the variables from the model outputs.

Equation 1: It is more standard to not use italics if the variable contains more than one letter i.e. consider making all letters non-italic apart from n, l, z, w, ρ. Consider renaming mrsol in equation 1 with something briefer e.g. mSM, θm

- Author Response: We thank the referee for this suggestion. We revised Equation 1 to avoid italicizing multi-letter variable names, while retaining italic formatting for single-letter variables

and indices (Line 139). We kept mrsol(i) to avoid confusion that this variable is indeed the same one from the model output.

Line 137: put the numbers in the volumetric SM units into superscript

- Author Response: We thank the referee for noting this. We corrected the formatting of the volumetric soil moisture units so that the exponents are properly shown as superscripts (Line 148).

RC2: 'Comment on egusphere-2025-3517', Anonymous Referee #2, 19 Dec 2025

The paper extends ILAMB to benchmark soil moisture (SM) in CMIP6 Earth System Models, evaluating both surface and root-zone depths and relating SM skill to ecohydrologic variables (ET, GPP, LAI), including climate-region analysis. It is a great effort to conduct this type of benchmarking, but I hope to see a more thorough description and interpretations of the differences in performance among these ESM models to help guide model improvement.

- Author Response: We thank the referee for this constructive comment. We agree that the initial version of the manuscript placed too much emphasis on describing benchmarking results and not enough on interpreting the differences in performance among models. In the revised manuscript, we have substantially expanded the interpretation of the results and strengthened the links between model performance and underlying land surface model structure and process representations.

Specifically, we (i) added a new summary table (Table A1) documenting key structural features of the land surface models (e.g., hydrology scheme, rooting representation, lower boundary condition, and treatment of soil–plant coupling) for those models where this information is clearly documented and readily available, (ii) revised Sections 3.1–3.5 to emphasize cross-model patterns, state-versus-coupling behavior, and regime-dependent performance rather than figure-by-figure description, and (iii) substantially expanded Sections 4.1 and 4.2 to explicitly relate observed performance differences to model process representations and land model families (e.g., CLM5, JSBACH, and JULES). We note in the manuscript that comparable process-level documentation is not always easily accessible or consistently reported for all CMIP6 models, which limits the completeness of such a synthesis.

In addition to these structural revisions, we now explicitly summarize the key diagnostic interpretations emerging from the analysis. In brief, models sharing the same land surface framework exhibit coherent performance signatures across both global and regional benchmarks. For example, CLM5-based models (CESM2 and NorESM2-LM), which include explicit soil–plant hydraulic stress and multi-layer Richards-based hydrology, tend to show strong ecohydrological coherence and realistic SM–vegetation coupling in Tropical and Temperate regions, but exhibit systematic high-latitude SM biases likely linked to organic soil and cold-region process representations. In contrast, JSBACH-based models (MPI-ESM1-2-LR and AWI-ESM-1-1-LR) often reproduce surface SM states reasonably well but show weaker SM–vegetation coupling, suggesting limitations in propagating soil moisture anomalies into vegetation responses under more implicit soil–plant coupling schemes. JULES-based models (UKESM1-0-LL) display depth-dependent behavior, with comparatively better rootzone performance than surface SM in some regions, consistent with prescribed rooting structures and implicit stress formulations. These patterns indicate that structural choices in hydrology, rooting depth, and soil–plant coupling exert a first-order control on both state accuracy and cross-variable coherence.

We now also summarize these process-level interpretations explicitly in the revised manuscript, including in the concluding synthesis (Lines 609-648), where we identify structural differences in

vertical soil discretization, rooting representation, and soil–plant hydraulic coupling as likely priority areas for improvement. A forward-looking summary of these implications is also included in the final sentence of the Abstract to ensure that the diagnostic insights and development priorities are clearly communicated.

Together, these revisions shift the paper from a primarily descriptive benchmarking study to a more diagnostic analysis that connects performance differences to model structure and process choices, with clearer implications for future land model development.

Then, the motivation to relate soil moisture (SM) skill to ecosystem functioning is well taken, but the choice of GPP and LAI as primary ecohydrological variables raises concerns regarding interpretability. GPP and LAI are highly integrated, bulk ecosystem variables that reflect multiple interacting processes beyond soil moisture. As a result, discrepancies in SM–GPP or SM–LAI relationships cannot be straightforwardly attributed to deficiencies in soil hydrology, right? In contrast, hydrological variables more directly linked to soil moisture dynamics—such as runoff, drainage, evaporation and transpiration partitioning, and canopy interception—would provide a clearer and more mechanistic pathway for diagnosing soil-moisture–related model behaviour. Including such variables would strengthen the causal interpretability of the proposed benchmarking framework and its relevance for guiding land-model development.

- Author Response: We thank the referee for this thoughtful and important comment. We fully agree that GPP and LAI are integrated ecosystem variables influenced by many interacting processes beyond soil moisture, and that discrepancies in SM–GPP or SM–LAI relationships cannot be interpreted mechanistically as arising solely from deficiencies in soil hydrology. In the revised manuscript, we now clarify this limitation explicitly and refine the interpretation of these coupling metrics.

Our intent in including GPP, LAI, and ET is not to provide a mechanistic diagnosis of individual soil hydrological processes, but rather to assess whether models reproduce the emergent ecosystem-scale coupling between soil moisture variability and vegetation functioning. Accordingly, we have revised Section 4.3 (Lines 566-607) to state explicitly that these variables are used as integrated diagnostics of model behavior and not as process-isolating benchmarks.

We agree that more process-level hydrological variables (e.g., runoff, drainage, evaporation–transpiration partitioning, canopy interception) would provide a more direct pathway for mechanistic attribution. However, we also note that such quantities are not yet consistently observed at the global scale and therefore cannot currently serve as robust, uniform benchmarking targets across CMIP6 models. We have added text to the Discussion (Section 4.3, Lines 580-583) acknowledging this limitation and highlighting these variables as a priority for future benchmarking efforts as observational capabilities improve.

We believe these revisions strengthen the interpretability of the results by more clearly delineating what can (and cannot) be inferred from the SM–ecohydrology coupling diagnostics used in this study.

A further limitation of the current benchmarking framework is the lack of systematic documentation of key land-surface input differences among the evaluated models. In particular, it is unclear whether—and how—soil texture, soil hydraulic parameters, rooting depth distributions, or lower boundary conditions differ across models. Because soil moisture dynamics are highly sensitive to these prescribed or semi-prescribed inputs, benchmarking model outputs without explicitly accounting for such differences makes it difficult to attribute performance differences to model structure rather than to input choices. Providing a concise overview of relevant soil and hydrological inputs for each model, or at least discussing their expected variability and implications, would improve the interpretability of the results and strengthen the paper's ability to inform model development.

- Author Response: We thank the referee for this important and constructive comment. We agree that differences in prescribed land-surface inputs (e.g., soil texture, hydraulic parameters, rooting depth distributions, and lower boundary conditions) can strongly influence soil moisture dynamics and complicate attribution of performance differences to model structure alone.

In response, we have taken several steps in the revised manuscript. First, we added a new summary table (Table A1) that compiles key structural and process-level characteristics of the land surface models (e.g., hydrology scheme, rooting representation, lower boundary condition, and soil–plant coupling) for those CMIP6 models where this information is clearly documented and readily available. This provides a concise overview of major similarities and differences among model frameworks and helps contextualize the benchmarking results.

Second, we substantially expanded the Discussion (Sections 4.1 and 4.2) to explicitly address how differences in soil properties, hydrologic parameterizations, vertical structure, and boundary conditions may contribute to the observed patterns in both global and regional performance. We also discuss the limitations of attributing performance differences uniquely to model structure versus prescribed inputs, and highlight this as a key source of uncertainty.

Finally, we note in the manuscript that a comprehensive, uniform documentation of land-surface inputs is not always available across all CMIP6 models, which limits the completeness of such a synthesis. Where detailed information is lacking, we explicitly acknowledge this and discuss the expected implications qualitatively. We believe these additions improve the interpretability of the results while remaining consistent with the scope of the present study, and they help clarify how both model structure and prescribed inputs likely shape the benchmarking outcomes.

Some Detailed comments:

P2 L45-47: It needs more details about the ILAMB framework and why it suits evaluating ESMs.

- Author Response: We thank the referee for this suggestion. We have revised the Introduction (Section 1, Lines 44-52) to provide a clearer and more detailed description of the ILAMB framework and why it is well suited for evaluating ESMs. Specifically, we added text explaining that ILAMB provides a standardized, reproducible benchmarking system, uses a suite of complementary performance metrics (e.g., bias, variability, seasonal cycle, and spatial pattern

agreement), and enables consistent multi-model intercomparison and diagnostic evaluation across variables. We also clarified how ILAMB has been applied to a range of ecohydrological variables in previous studies and how this work extends its application to soil moisture. We believe these additions better motivate the choice of ILAMB and clarify its relevance for this study.

P3, l92-94: The authors first augue the drawback for the Qiao et al., 2022 data is that they used reanalysis data as reference to evaluating the SM, but instead here, they said they used the Wang et al., 2021 data which is averaged across both observational and also model simulation outputs, why the Wang et al., 2021 data is better that the reanalysis data?

- Author Response: We thank the referee for raising this important point. We agree that the Wang et al. (2021) product also incorporates model-based information and should not be interpreted as a definitive or purely observational "truth." In the revised manuscript (Lines 94-111), we have clarified this distinction and removed any implication that Wang2021OLC is inherently superior to reanalysis-based products.

We now emphasize that the key difference relative to Qiao et al. (2022) is not the absence of model influence, but rather the use of multiple, complementary reference datasets with different methodological assumptions and sources of uncertainty. Specifically, we combine a satellite-based observational product (ESA-CCI) with a blended product that integrates offline land surface models, reanalysis, and satellite constraints (Wang2021OLC). This multi-product approach is intended to reduce reliance on any single reference dataset and to better characterize observational uncertainty, rather than to privilege one product as a "truth."

We revised the Introduction and the Discussion (Section 4.3, Lines 567-575) to make this rationale explicit and to state clearly that none of the benchmark datasets used here represents a definitive reference, and that differences among them are themselves part of the uncertainty being explored.

PA6, section 2.3, please describe why you chose these different GPP, LAI and ET products

- Author Response: We thank the referee for this suggestion. We revised Section 2.3 (Lines 201-224) to more clearly justify the choice of the GPP, LAI, and ET benchmark products by briefly describing their construction, strengths, and typical use in large-scale model evaluation, and by adding appropriate citations (including WECANN GPP, AVHRR/AVH15C1 LAI and the NOAA CDR, and GLEAM v3 ET). The revised text now clarifies why these products were selected as suitable and widely used references for global-scale benchmarking, while also acknowledging their underlying assumptions and limitations.

Move Figure 3 to the Results section.

- Author Response: We thank the referee for this suggestion. We moved Figure 3 to the Results section, where it is now discussed in the context of the benchmarking results.

L250-254, you don’t need to describe what each figure is about, but focus on what you found from different figures.

- Author Response: We thank the referee for this suggestion and agree that the Results section should emphasize interpretation rather than figure-by-figure description. In the revised manuscript, we have substantially rewritten the relevant parts of Section 3 (including the text around Lines 250–254 from the original manuscript) to focus on the key findings and patterns emerging from the figures, rather than describing what each figure shows. The Results section now highlights cross-model contrasts, state-versus-coupling behavior, and regime-dependent performance, and explicitly links these patterns to model structure where appropriate. We believe this revision makes the Results more concise, more interpretive, and more informative for guiding model evaluation and development.

Section 3.1 did not provide a clear summative description of the overall benchmark score. The description of what each figure is about is not what you have have in the Results sections

- Author Response: We thank the referee for this comment and agree that the original Section 3.1 did not sufficiently emphasize a summative interpretation of the overall benchmark scores. In the revised manuscript, we have substantially rewritten Section 3.1 (Lines 250-280) to provide a clear synthesis of the overall ILAMB results, focusing on cross-model patterns, contrasts between state and coupling skill, and similarities among models sharing the same land surface framework, rather than describing individual figures. We now explicitly explain what the Overall Score represents, how it relates to its component metrics, and what the main performance patterns imply for model behavior. We believe these changes address the referee’s concern and make Section 3.1 a more effective and informative summary of the benchmarking results.

L299-301: unclear to me why it is important to benchmark GPP and LAI in this study. Yes, they are linked to the SM, but their variations are also affected by many other factors.

- Author Response: We thank the referee for this comment. We agree that GPP and LAI are influenced by many factors beyond soil moisture and that their variability cannot be attributed solely to soil hydrologic processes. We revised the manuscript to clarify this point, (including edits around Lines 299–301 from the original manuscript) and the addition of a new paragraph in Discussion Section 4.3 (Lines 576-584), where we now explicitly state that GPP and LAI are used here as integrative ecosystem-scale diagnostics and that SM–GPP and SM–LAI relationships are interpreted as emergent coupling rather than direct mechanistic attribution. The revised text more clearly explains why benchmarking GPP and LAI is informative in this context while acknowledging the associated interpretive limitations.