

Benchmarking soil moisture and its relationship to ecohydrologic variables in Earth System Models

Elias C. Massoud¹, Nathan Collier¹, Yaoping Wang¹, Jiafu Mao¹, Adrian Harpold², Steven A. Kannenberg³, Gerbrand Koren⁴, Mukesh Kumar⁵, Pushpendra Raghav⁵, Pallav Ray⁶, Mingjie Shi⁷, Jing Tao⁸, Sreedevi P. Vasu⁶, Huiqi Wang⁹, Qing Zhu⁸, Forrest M. Hoffman¹

¹ Oak Ridge National Laboratory, Oak Ridge, TN 37830, USA

² University of Nevada, Reno, Reno, NV 89557, USA

³ Department of Biology, West Virginia University, Morgantown, WV 26506, USA

⁴ Copernicus Institute of Sustainable Development, Utrecht University, Princetonlaan 8a, 3584 CB Utrecht, Netherlands

⁵ Department of Civil, Construction, and Environmental Engineering, University of Alabama, Tuscaloosa, AL 35487, USA

⁶ Florida Institute of Technology, Melbourne, FL 32901, USA

⁷ Pacific Northwest National Laboratory, Richland, WA 99352, USA

⁸ Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

⁹ University of California, Berkeley, Berkeley, CA 94720, USA

Correspondence to: Elias C. Massoud (massoudec@ornl.gov)

Abstract. Soil moisture (SM) is a key regulator of ecosystem biogeophysics, influencing plant water relations and land-atmosphere energy exchanges. [We evaluate the representation of SM in 16 CMIP6 Earth System Models using the International Land Model Benchmarking \(ILAMB\) framework, focusing on surface \(0–5 cm, 0–10 cm\) and rootzone \(0–100 cm\) depths, key ecohydrological variables \(GPP, LAI, ET\), and their coupling. Models are benchmarked against multiple observational and assimilated datasets to assess both state variables and cross-variable relationships. Surface SM is generally well represented \(\$r > 0.87\$ \), while rootzone SM variability is systematically overestimated \(normalized standard deviation \$> 1\$ \). ET shows strong agreement with observations \(\$r > 0.9\$ \), whereas GPP and LAI exhibit larger inter-model spread. Skill in individual variables does not guarantee realistic SM–ecohydrology coupling, which varies strongly across models and depends on the reference dataset. Köppen-based regional analyses reveal strong regime dependence, with several models performing well in Tropical and Temperate regions but degrading in Continental \(high-latitude\) zones. Across both global and regional benchmarks, models cluster by land surface framework, indicating that structural choices in soil hydrology and soil–plant coupling exert a first-order control on performance. These results provide process-relevant benchmarks and suggest that improving the representation of vertical soil structure, rooting depth distributions, and soil–plant hydraulic coupling will be central to advancing soil moisture realism in next-generation Earth system models.](#)

~~This study evaluates the representation of SM in Earth System Models (ESMs) using the International Land Model Benchmarking (ILAMB) framework, focusing on both surface (0–5 cm, 0–10 cm) and rootzone (0–100 cm) depths. We~~

35 ~~benchmark Coupled Model Intercomparison Project Phase 6 (CMIP6) models against multiple observational and assimilated~~
~~datasets to evaluate their performance in simulating SM, as well as their relationships with ecohydrological processes and~~
~~vegetation traits such as gross primary productivity (GPP), leaf area index (LAI), and evapotranspiration (ET). Results show~~
~~that while surface SM is generally well represented ($r > 0.87$), rootzone SM variability is overestimated (normalized standard~~
~~deviation > 1). Simulated ET agrees strongly with observations ($r > 0.9$; normalized standard deviation 0.8–1.2), whereas~~
40 ~~ecoecology relationships varies with model structure and observational dataset, with better consistency observed when~~
~~assimilated SM products are used. Regional analyses using Köppen classifications reveal distinct model behaviors, with~~
~~stronger performance in tropical zones and reduced skill in high-latitude regions, likely due to challenges in simulating~~
~~freeze-thaw and permafrost dynamics. These findings offer quantitative benchmarks of model performance, highlighting~~
~~specific areas for improving SM representation and its coupling with vegetation and hydrological processes in future ESM~~
45 ~~development.~~

1 Introduction

1 Introduction

50 Soil moisture (SM) plays a central role in regulating Earth system processes by controlling the storage and exchange of
water, carbon, and energy between the land surface and atmosphere (Clark et al., 2015; Trugman et al., 2018; Green et al.,
2019; Massoud et al., 2020). Accurately representing SM in Earth System Models (ESMs) is essential for improving
predictions of the Earth system (Seneviratne et al., 2010; Hauser et al., 2016; Humphrey et al., 2021). ESMs simulate these
processes through coupled biogeochemical and hydrological cycles, both of which are strongly influenced by SM. However,
55 accurately modeling SM at ESM grid scales remains challenging due to heterogeneity in soil properties, scale mismatches
between physical processes and model resolution, and limited knowledge of subsurface boundary conditions (e.g.,
groundwater depth). As a result, ESMs adopt a range of approaches to simulate SM, ~~most commonly~~ The majority relying on
"bucket-type" models that route water through discrete soil layers using threshold-based parameters (e.g., field capacity) and
rate-dependent functions (e.g., hydraulic conductivity).

60 This work seeks to evaluate the representation of SM in ESMs by ~~incorporating using~~ the International Land Model
Benchmarking (ILAMB) framework (Collier et al., 2018), a tool extensively used to assess the performance of land models.
ILAMB provides a standardized, reproducible evaluation framework that compares model output against observational
benchmarks using a suite of complementary metrics, including measures of bias, variability, seasonal cycle, and spatial
pattern agreement. Its modular design and consistent scoring methodology make it well suited for systematic, multi-model
intercomparison and for diagnosing strengths and weaknesses across different components of the land surface system.
65 Although ILAMB has been applied to various ecohydrologic processes and properties such as gross primary production
(GPP) (Caen et al., 2022), evapotranspiration (ET) (Wu et al., 2020), and leaf area index (LAI) (Yang et al., 2023), SM has
remained mostly underrepresented in ILAMB-based model evaluations until now.

70 Despite advances in ESMs, the accurate simulation of SM in Coupled Model Intercomparison Project Phase 6 (CMIP6)
remains a persistent challenge due to structural biases and uncertainties in land surface processes. Several studies have
highlighted both improvements and limitations in how CMIP6 models represent SM. For example, Yuan et al. (2021)
showed that CMIP6 models better capture historical surface SM trends over the contiguous United States (CONUS)

75 compared to CMIP5, particularly in regions like the Northwest and Midwest. However, considerable inter-model variability remains, suggesting a need for further refinement in future model generations. Similarly, Wang et al. (2022) conducted a comprehensive evaluation of CMIP6 SM simulations over China and found that while the multi-model mean (MME) generally captured observed spatial patterns and seasonal cycles of both near-surface and rootzone SM, substantial inter-model spread persisted, particularly in trends and interannual variability. Their findings also emphasized the dominant role of land surface schemes in driving model behavior, as models developed by the same institution often exhibited similar performance. Moreover, Purdy et al. (2018) illustrated the potential for significant improvement in model performance through better SM representation, showing that integrating SM information from satellite data into model simulations reduced global ET errors by up to 23% in dry regions. These studies point to the critical need to identify and correct structural biases that limit current model skill in simulating SM.

85 At the same time, the importance of accurately simulating SM extends beyond model performance metrics. Zuo et al. (2024) highlighted that maintaining current global SM levels could reduce nearly a third of projected land warming under low-emission scenarios. Their results, based on outputs from historical CMIP6 experiments and other model intercomparison projects, highlight the central role SM plays in climate feedbacks and the reliance on models like CMIP6 to inform future projections. Given both the challenges and stakes involved, this study benchmarks SM and its coupling with ecohydrologic processes such as GPP, LAI, and ET, with the goal of identifying model limitations and guiding improvements as we look toward CMIP7 and beyond.

90 The evaluation of SM in global models has a long history, with early efforts dating back to Robock et al., (1998) as part of the Atmospheric Model Intercomparison Project (AMIP). In their analysis, they found significant discrepancies in how SM was represented and simulated across different models, an issue that was prominent 25 years ago and remains a challenge today. More recently, Qiao et al., (2022) conducted a detailed evaluation of SM using a suite of CMIP6 models, examining both surface and deeper SM (up to 2 meters) across various subregions around the globe. They found that the multimodel ensemble mean generally produces reasonable representations for overall climatology. However, their study relied on reanalysis products and data assimilation systems as the reference for benchmarking (Qiao et al., 2022), which, while widely used, are not true SM observational products but rather model-based estimates. Furthermore, Qiao et al. (2022) assumed that the model outputs `mrsos` (surface SM) and `mrsol` (layered SM) represent the same variable, despite their structural differences (Massoud et al., 2025, in review). In reality, `mrsol` provides moisture values at multiple soil layers, from which `mrsos`, representing moisture in the top 10 cm of soil, is derived. Other studies used a combination of observation and model-based benchmarks, but are limited to regional domains and similarly do not distinguish the different nature of `mrsos` and `mrsol` (Yuan et al. 2021; Wang et al. 2022). In this current study, we distinguish between these two variables and apply depth-specific derivations to more accurately evaluate SM representation across different layers in the models. Here, we not only benchmark SM at various depths, but also evaluate key ecohydrologic variables such as GPP, LAI, and ET, both individually and in relation to SM, to provide a more comprehensive assessment of land surface processes in CMIP6 models.

110 One of the persistent challenges in benchmarking SM is the limited availability of high-quality datasets. To address this, we utilize two key datasets for global-scale benchmarking in this study. The first is the Wang et al. (2021) product, which is a weighted average of multiple sources, including offline land surface model simulations, remote sensing data, and reanalysis products, that was found to outperform the original sources in that study. This dataset provides estimates for both surface SM (top 10 cm) and rootzone SM (up to 1 m), offering a view of SM at different depths. The second dataset is the European Space Agency Climate Change Initiative SM (ESA-CCI SM) product (Dorigo et al., 2017; Gruber et al., 2019; Preimesberger et al., 2021), which is derived from a blend of passive and active satellite sensors. ESA-CCI SM represents surface SM down to 5 cm, and its combination of different satellite platforms helps mitigate the limitations of individual sensors, providing an observational estimate for surface SM that is robust. Together, these datasets allow for a more thorough evaluation of SM in CMIP6 models. We emphasize that neither of these products represents a definitive “truth,” as both incorporate model-based information and are subject to their own structural assumptions and uncertainties. In contrast to studies that rely on a single reanalysis-based reference (e.g., Qiao et al., 2022), our approach intentionally combines complementary datasets with differing methodologies, including satellite-based observations (ESA-CCI) and a blended product integrating offline models, reanalysis, and satellite constraints (Wang et al., 2021). By using multiple reference

products rather than a single benchmark, our goal is to provide a more robust and balanced assessment of model performance and to reduce reliance on any one dataset. We note that soil moisture below 1 m can also be important for ecosystem functioning (Stocker et al., 2023), particularly in regions with deep rooting systems (Kühnhammer et al., 2023), but evaluating deeper soil layers is beyond the scope of this study and we therefore focus on surface and rootzone soil moisture.

While benchmarking SM alone can offer insights into model performance, understanding how SM interacts with other ecosystem processes such as GPP, ET, and LAI can yield additional clues into the strengths and weaknesses of these models (Guswa et al., 2002; Wang et al., 2019). By examining how models' skills are related to certain model processes, we aim to uncover patterns that point to specific model limitations, whether they stem from structural design, input data, or parameterizations. Ultimately, this work seeks to guide improvements in model development and reduce uncertainties in global SM simulations. The goals of this study are threefold. First, we benchmark CMIP6 models in their simulation of SM at multiple depths, specifically at 5 cm, 10 cm, and 100 cm, using various datasets. Second, we assess model performance in simulating key ecohydrologic variables, including GPP, LAI, ET, as well as the relationships between SM and each of these variables. Third, we aim to identify specific areas for improvement, whether in individual models or as systematic issues across the CMIP6 ensemble. To support these objectives, we also implement a Köppen climate region analysis (Geiger 1954) within the ILAMB framework to evaluate model performance across distinct climate zones.

2 Materials and Methods

2.1 Models and Variables in CMIP6 Simulations

The model data used in this study come from CMIP6 (Eyring et al., 2016), an international effort to standardize ESM simulations, enabling direct comparison across models. We use a suite of CMIP6 models (detailed in Table 1) that differ in their spatial resolutions and land surface model components. Table 1 lists each model's horizontal grid spacing and the land surface scheme used to simulate SM, GPP, LAI, and ET.

The SM variables analyzed are `mrsol` and `mrsos`, which represent layered and surface soil moisture, respectively. Both variables are provided in units of mass per unit area [kg m^{-2}]. Other simulated variables include `gpp` (units $\text{g m}^{-2} \text{d}^{-1}$), `lai` (unitless), and `evspsbl` (units mm d^{-1}) corresponding to GPP, LAI, and ET. All variables are analyzed at monthly temporal resolution.

To derive SM at specific depths (e.g., 0–5 cm, 0–10 cm, or 0–100 cm), we calculate depth-integrated soil moisture estimates from the `mrsol` variable, which provides total SM contained in each discrete model soil layer. Since the vertical layering varies across models, we apply a depth-weighted integration approach that also converts units from mass to volumetric SM (i.e., from [kg m^{-2}] to [m^3/m^3]). This conversion and integration is expressed as:

$$SM_{\text{integrated}} = \left\{ \left[\sum_{\text{layer } i=1}^{\text{layer } n-1} \frac{mrsol(i)}{\rho_w * dz(i)} * dz(i) \right] + \left[\frac{mrsol(n)}{\rho_w * dz(i)} * z_{\text{remaining}} \right] \right\} / z_{\text{total}} \quad (1)$$

In this equation, $mrsol(i)$ represents the mass of SM in the i -th model-defined soil layer, reported in units of [kg m^{-2}]. The variable ρ_w is the density of liquid water, assumed to be a constant value of 1000 kg/m^3 . The quantity $dz(i)$ refers to the thickness of the i -th soil layer in meters [m], which is used to compute the volumetric contribution of each layer. The term $z_{\text{remaining}}$ represents the portion in [m] of the final layer that partially overlaps with the target integration depth (e.g., if the target is 10 cm and the final layer spans from 8–15 cm, then $z_{\text{remaining}} = 2 \text{ cm} = 0.02 \text{ m}$). The total integration depth, z_{total} , is the sum of all full-layer thicknesses plus $z_{\text{remaining}}$, defining the vertical extent over which SM is integrated.

This formula first converts each layer's SM from mass per unit area [kg m^{-2}] to volumetric SM [m^3/m^3] by dividing by the product of water density and layer thickness. Then, it multiplies by the layer thickness to compute the volume per unit

165 area. Summing over all layers and dividing by the total soil depth yields the average volumetric SM over the target depth.
 This method standardizes SM across models with differing vertical discretizations and unit conventions, enabling accurate
 and consistent comparisons with benchmarking datasets, which report SM as a volume fraction. For surface SM ($mrsos$),
 which represents a shallow fixed-depth layer (i.e., 0.1 m), we apply the same conversion logic by assuming that fixed depth
 during volumetric transformation. This approach is consistent with prior studies (e.g., Qiao et al., 2022; Wang et al., 2022;
 170 Massoud et al., 2025).

CMIP6 Model Name	Spatial Resolution ($\sim lon^\circ \times lat^\circ$)	Land Model	Total Soil Depth (m)	# of Soil Layers
ACCESS-ESM1-5	1.875 × 1.25	CABLE	2.87	6
AWI-ESM-1-1-LR	1.875 × 1.875	JSBACH	6.98	5
BCC-ESM1	2.8 × 2.8	BCC-AVIM2	2.86	10
CanESM5-1	2.8 × 2.8	CLASS	4.10	3
CESM2	1.25 × 0.94	CLM5	8.03	20
CMCC-ESM2	1.25 × 0.94	CLM4.5	35.18	15
CNRM-ESM2-1	1.41 × 1.41	ISBA	≤12 m for soil temp; rooting ≤8 m	14
EC-EARTH3-CC	0.70 × 0.70	HTESEL and LPJ-GUESSv4	1.945	4
GFDL-ESM4	1.0 × 1.0	LM4.1-PPA	8.75	20
GISS-E3-G	2.5 × 2.0	GISS LSM	3.50	6
MPI-ESM1-2-LR	1.875 × 1.875	JSBACH	6.98	5
MRI-ESM2-0	1.125 × 1.125	AGCM	10.0	14
NorESM2-LM	2.5 × 1.875	CLM5	8.03	20
SAM0-UNICON	1.25 × 0.94	CLM4.0	8.03	20
TaiESM1	1.25 × 0.94	CLM4.0	8.03	20
UKESM1-0-LL	1.875 × 1.25	JULES-ES-1.0	3.0	4

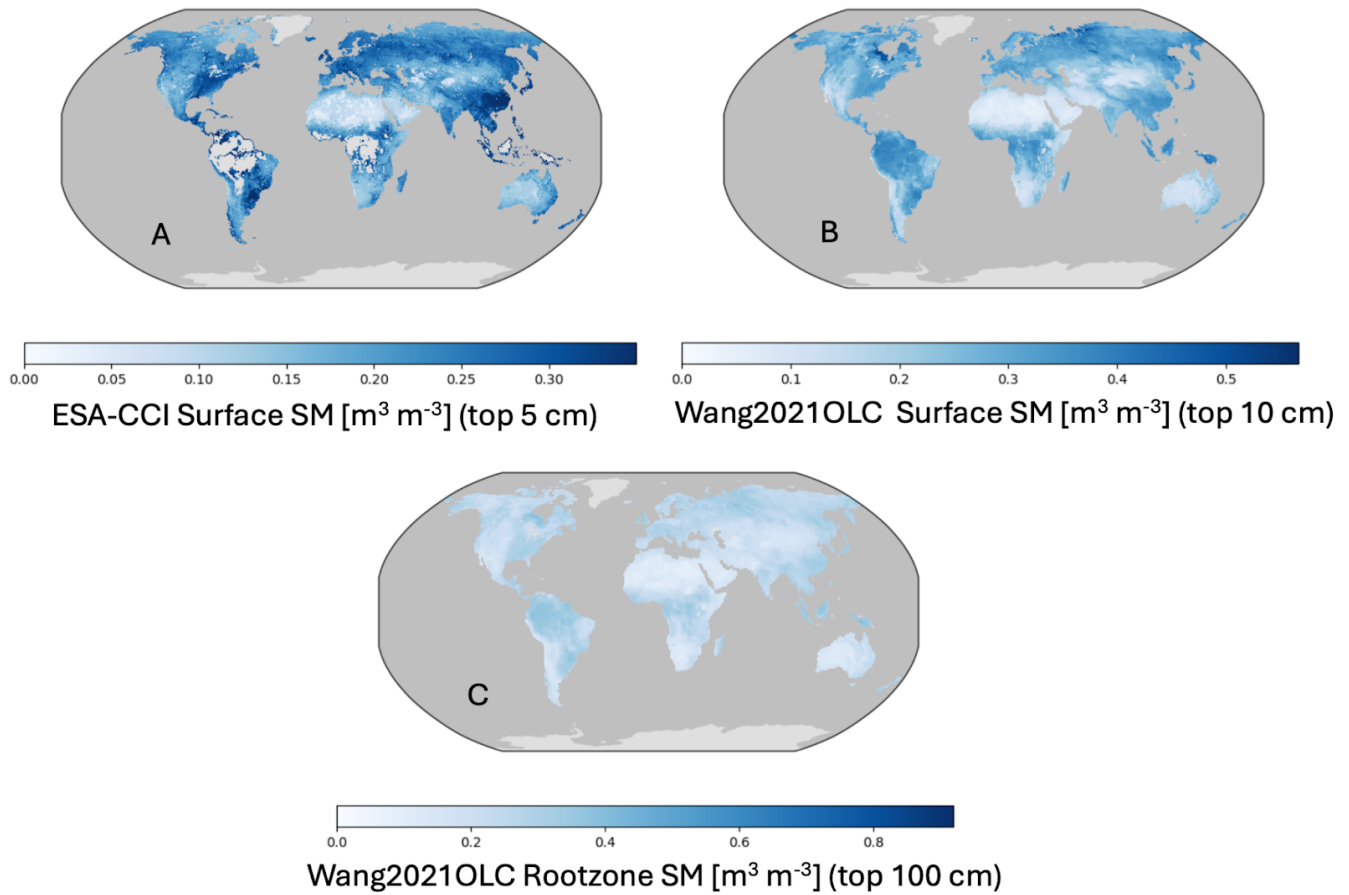
Table 1: CMIP6 models used in this study, along with their latitude and longitude grid sizes and the land models with their soil depths and number of soil layers used in each model.

2.2 Soil Moisture Datasets Used for Benchmarking

175 In this study, we evaluate the SM performance of CMIP6 models using two primary benchmark datasets. The first is the
180 ESA-CCI SM product (Dorigo et al., 2017; Gruber et al., 2019; Preimesberger et al., 2021), which provides a global, long-
term record of surface SM (~2-5 cm) spanning over 40 years (1978-2023) at a daily temporal resolution that is aggregate to
reflect monthly values and spatial resolution of approximately 25 km (Figure 1A). The ESA-CCI SM product is updated
annually through an algorithmic process that incorporates data from both passive and active satellite sensors, producing a
185 blended dataset that integrates multiple sensor sources while extending the time series with each update. This dataset has
been extensively used in hydrological and climatological research (e.g., An et al., 2016; McNally et al., 2016; Ciabatta et al.,
2018; Massoud et al., 2023; Li et al., 2025a), including the Bulletin of the American Meteorological Society's annual "State
of the Climate" reports. Its long-term, global coverage makes it a valuable resource for evaluating surface SM in ESMs. To
facilitate direct comparison with the ESA-CCI SM product that typically represents SM at a depth of 5 cm, the `mrsol`
variable in each CMIP6 model is integrated to the 5 cm layer using Eq. 1. Given that some models have first layers that are
deeper than 5 cm, this integration may introduce additional uncertainty in the estimated SM.

The second SM dataset used in this study is the Wang et al. (2021) product, which provides a global, gap-free, long-term
190 record of SM across four depths (0–10, 10–30, 30–50, and 50–100 cm) from 1970 to 2016, with a monthly temporal
resolution and spatial resolution of 0.5 degrees. This dataset synthesizes SM information from diverse sources, including in
situ observations, satellite data, reanalysis products, and offline land surface model simulations. It employs three statistical
approaches, unweighted averaging, optimal linear combination (OLC), and emergent constraint (EC), to produce a merged
product that outperforms individual source datasets in terms of bias, root mean square error (RMSE), and correlation when
195 compared to in situ observations. For this study, we utilize the OLC version of the Wang et al. (2021) product, which we
hereafter refer to as Wang2021OLC, because it is constrained by in situ observational values and performs among the best of
the paper's reported method-data source combinations. This hybrid dataset offers harmonized spatial, temporal, and vertical
coverage, making it highly suitable for large-scale ESM benchmarking of both surface (Figure 1B) and rootzone (Figure 1C)
SM. Since this dataset provides SM estimates at both 10 cm and 100 cm depths, we integrate the `mrsol` variable from each
200 CMIP6 model to these depths using Eq. 1 to enable direct comparison. In addition, because the `mrsos` variable in CMIP6
models also represents surface SM at approximately 10 cm, it is separately benchmarked against the 10 cm layer from the
Wang2021OLC product. This dual use of `mrsol` and `mrsos` allows us to assess the internal consistency and depth
representation of SM across different model variables and observational references.

205 While these datasets are widely used and provide valuable long-term, global-scale estimates, each has inherent limitations,
including retrieval uncertainties, differences in spatial and temporal resolution, and dependence on model-based or
algorithmic assumptions. Because no single dataset can fully capture the complexity of SM dynamics, the use of multiple,
complementary observational and assimilated products helps quantify uncertainty in the benchmark results presented here.



210 **Figure 1: Long-term mean soil moisture (SM) from observational datasets. (A) ESA-CCI Surface SM [$\text{m}^3 \text{m}^{-3}$] (top 5 cm) (1978-2023), providing global estimates based on satellite data. (B) Wang et al. (2021) OLC Surface SM [$\text{m}^3 \text{m}^{-3}$] (top 10 cm) (1970-2016), derived from a merged dataset of satellite and in-situ observations. (C) Wang et al. (2021) OLC Rootzone SM [$\text{m}^3 \text{m}^{-3}$] (top 100 cm), representing long-term mean soil moisture across the rootzone. Each subplot shows the global distribution of soil moisture at the depths represented in each product.**

215 2.3 Other Benchmark Datasets

For the ecohydrologic relationship variables, we incorporate observational datasets specific to each variable that exist in the ILAMB data library. GPP observations are derived from the Water, Energy, and Carbon with Artificial Neural Networks (WECANN) dataset, which provides globally gridded estimates based on advanced machine learning approaches that integrate remote sensing and meteorological inputs (Alemohammad et al., 2017) from 2007 to 2016. LAI data is sourced from the NOAA Climate Data Record (CDR) of AVHRR Leaf Area Index (AVH15C1) from 1981 to 2019 (Claverie et al., 2016). ET data comes from the Global Land Evaporation Amsterdam Model (GLEAM) v3.3a, which provides daily estimates (aggregated here to reflect monthly values) from 1980 to 2018 (Miralles et al., 2011; Martens et al., 2017). The spatial resolution of the WECANN GPP product (Figure 2A) is 0.5° , that of the AVH15C1 LAI data (Figure 2B) is 0.05° , and that of the GLEAMv3.3a ET data (Figure 2C) is 0.25° .

225

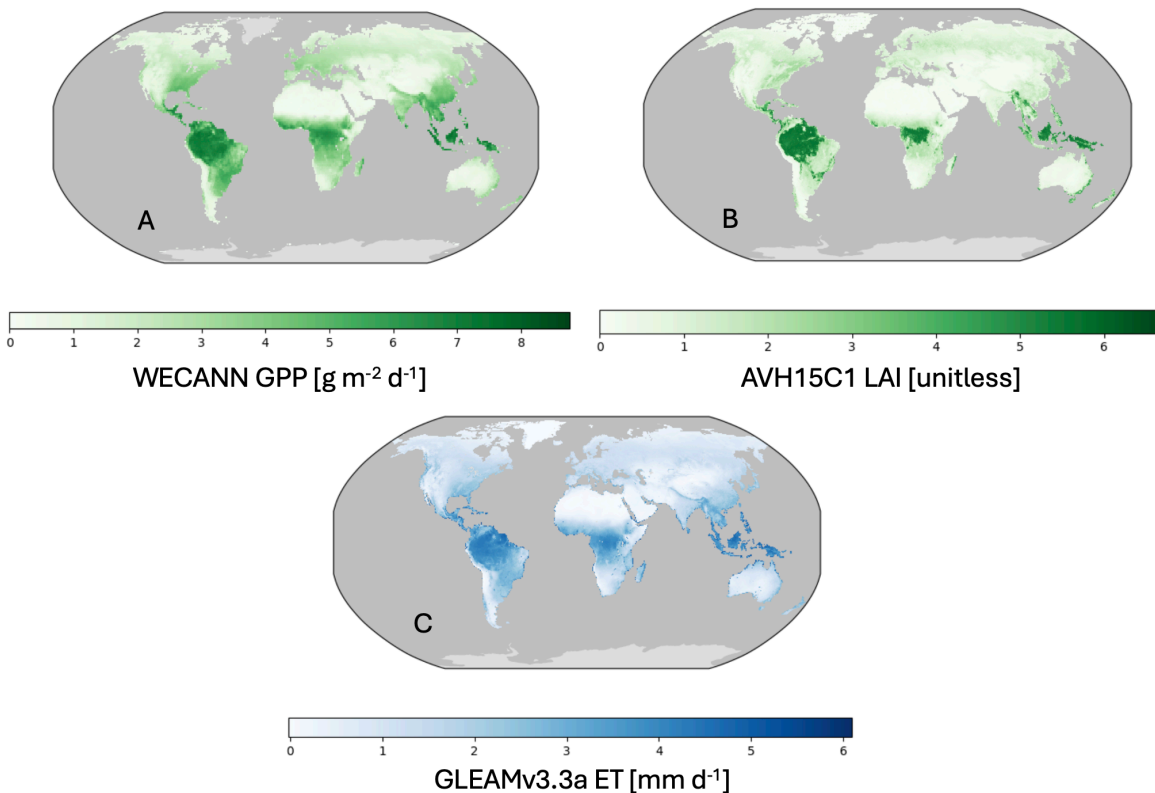


Figure 2: Long-term mean ecohydrological variables from observational datasets. (A) WECANN GPP [$\text{g m}^{-2} \text{d}^{-1}$] (2007–2016), providing global estimates of gross primary productivity derived from machine learning techniques. (B) AVH15C1 LAI [unitless] (1981–2019), representing global leaf area index values from the NOAA Climate Data Record. (C) GLEAMv3.3a ET [mm d^{-1}] (1980–2018), showing global evapotranspiration estimates derived from satellite observations and meteorological data. Each subplot displays the global distribution of these ecohydrological variables.

While ILAMB supports multiple observational datasets for each variable, we utilize a single benchmark dataset per variable in this study to maintain consistency and simplicity in our analysis. [GPP observations are taken from the WECANN product \(Alemohammad et al., 2017\), which provides globally gridded estimates derived from machine-learning approaches that integrate satellite and meteorological data.](#) LAI is evaluated using the NOAA Climate Data Record AVH15C1 product [\(Claverie et al., 2016; NOAA CDR AVH15C1\), which offers a long-term, internally consistent record based on AVHRR observations.](#) ET is benchmarked against GLEAM v3.3a [\(Miralles et al., 2011; Martens et al., 2017\), a widely used dataset that combines satellite observations with a physically based modeling framework.](#)

It is worth noting that although these products are widely used as observational references, they are themselves derived from models or statistical algorithms informed by observational inputs, [and therefore should be interpreted as observationally informed estimates rather than direct measurements.](#) Each dataset also carries known limitations, including [uncertainties related to retrieval algorithms, input data quality, and assumptions embedded in the underlying models.](#) [Nonetheless, these datasets have been extensively evaluated and applied in previous studies and provide reliable.](#) ~~As such, they should be interpreted as observationally informed estimates rather than direct measurements. Nonetheless, these datasets provide~~

reliable and widely accepted benchmarks for evaluating modeled GPP, LAI, and ET (represented by the `gpp`, `lai`, and `evspsbl` variables), supporting a comprehensive evaluation of ecohydrologic processes and their relationships with SM.

2.4 ILAMB Framework

- 250 The ILAMB framework is an open-source software package used to assess the performance of ESMs by comparing model outputs to a suite of observational datasets. ILAMB's scoring system integrates several key performance metrics: bias, RMSE, seasonal cycle representation, and spatial distribution (Figure A1). These metrics are synthesized into an overall score using the method detailed in Collier et al. (2018), offering a quantitative view of model fidelity. Bias measures the average deviation from observational data, RMSE quantifies error magnitude, and the seasonal cycle and spatial metrics
- 255 assess the temporal and geographic accuracy of model outputs. By combining these metrics, ILAMB generates diagnostic graphics and scores to help identify strengths and weaknesses in model simulations. The ILAMB framework has been widely adopted for model evaluation and intercomparison, aiding in the continuous development of more accurate land model components (Collier et al., 2018, 2023).
- 260 In this study, we extend the ILAMB analysis by incorporating Köppen climate classifications, which allows for a more detailed evaluation of model performance across diverse climate zones. These regions, which include tropical, desert and semi-arid, temperate, and continental climates, reflect varying environmental conditions that significantly influence SM and vegetation dynamics. By examining model skill within these distinct climate zones, we gain deeper insights into region-specific strengths and weaknesses, allowing for targeted improvements in land models across climate types.
- 265

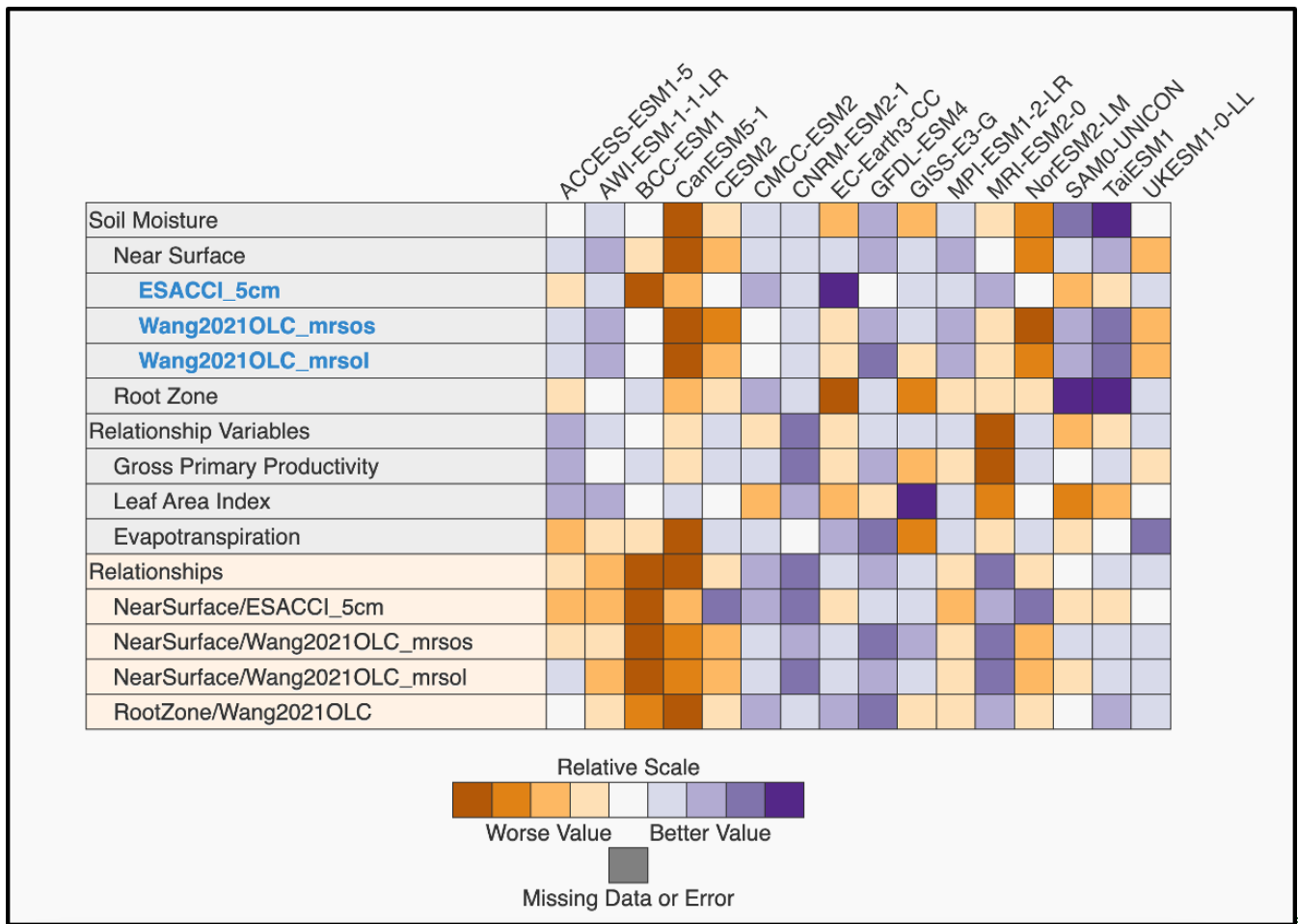


Figure 3: Benchmarking results from the ILAMB global run. Each model is compared against observational datasets: ESA-CCI Surface SM [$\text{m}^3 \text{m}^{-3}$] (top 5 cm), Wang2021OLC Surface SM [$\text{m}^3 \text{m}^{-3}$] (top 10 cm), Wang2021OLC Rootzone SM [$\text{m}^3 \text{m}^{-3}$] (top 100 cm), WECANN GPP [$\text{g m}^{-2} \text{d}^{-1}$], AVH15C1 LAI [unitless], and GLEAMv3.3a ET [mm d^{-1}]. Additionally, the relationships between soil moisture and ecohydrological variables (GPP, LAI, and ET) are also benchmarked. Purple represents better relative skill, while orange indicates relatively worse performance.

3 Results from ILAMB

3.1 Overall benchmark scores in ILAMB

In this section, we summarize the global performance of CMIP6 models using the ILAMB framework across surface and rootzone soil moisture (SM) and key ecohydrological variables. The combined results are shown in the ILAMB portrait plot (Figure 3), where colors indicate relative performance across models and metrics, with purple denoting higher skill and orange lower skill. The Overall Score is a composite metric that combines normalized measures of bias, error magnitude (RMSE), seasonal cycle representation, and spatial pattern agreement following the ILAMB methodology (Collier et al., 2018). The Seasonal Cycle Score evaluates how well models capture the timing and amplitude of the observed annual cycle, while the Spatial Distribution Score measures agreement in spatial patterns across grid cells.

285 Rather than showing uniform behavior across variables, the models exhibit distinct and structured performance patterns. Some models display relatively strong skill for ecohydrological variables (GPP, LAI, and ET) but weaker performance for soil moisture itself, while others perform well for soil moisture states but less well for vegetation and flux variables. As a result, the coupling metrics between SM and ecohydrological variables often differ markedly from the skill in the individual variables alone. This indicates that reproducing realistic mean states or variability does not necessarily imply that models capture the correct sensitivity of ecosystem processes to soil moisture.

290 Notably, models based on the same land surface framework tend to exhibit similar performance signatures (e.g., CESM2 and NorESM2-LM using CLM5, and MPI-ESM1-2-LR and AWI-ESM-1-1-LR using JSBACH; see Table A1). This structural coherence suggests that differences in land model formulation play an important role in shaping benchmarking outcomes, motivating a more detailed examination of soil moisture, ecohydrological variables, and their coupling in the following subsections. We note that comparable process-level documentation is not always consistently available across all CMIP6 models; therefore, Table A1 summarizes key characteristics only for those models for which this information is clearly documented in the literature, and the absence of entries for some models reflects limitations in publicly available documentation rather than an exhaustive comparison.

300 Figures A1–A3 provide additional context for these patterns. Figure A1 shows the individual ILAMB component scores for global surface SM ($mrsol$ integrated to 10 cm) against Wang2021OLC, Figure A2 presents the quantitative Overall Scores underlying Figure 3, and Figure A3 shows global maps of model bias. Together, these diagnostics illustrate that the overall rankings emerge from a combination of biases, variability errors, and spatial pattern mismatches, rather than from any single performance metric alone.

305 In this section, we present the evaluation of CMIP6 model performance using the ILAMB framework, focusing on key land surface variables. Specifically, we benchmark model outputs against observational datasets for surface and rootzone SM, as well as ecohydrological variables. Surface SM is evaluated using the ESA CCI product (top 5 cm, Figure 1A) and the Wang2021OLC dataset (top 10 cm, Figure 1B), while rootzone SM (up to 100 cm) is assessed using the Wang2021OLC product (Figure 1C). For ecohydrological benchmarking, we use GPP from WECANN (Figure 2A), LAI from AVH15C1 (Figure 2B), and ET from GLEAMv3.3a (Figure 2C).

310 Overall model performance across these variables is summarized in the ILAMB portrait plot (Figure 3), where models are color-coded from orange (lower performance) to purple (higher performance). These benchmarks are derived from quantitative ILAMB evaluation metrics, including the Bias Score, RMSE Score, Seasonal Cycle Score, and Spatial Distribution Score. These scores contribute to the final benchmark, represented as the Overall Score, and are the results that are presented in Figure 3. Figure A1 shows these metrics based on global SM simulations using the $mrsol$ variable integrated to 10 cm, compared against the Wang2021OLC dataset at the same depth. Figure A2 presents the quantitative overall ILAMB scores used to generate the colored portrait plot in Figure 3. Figure A3 shows global maps of model bias ($m^3 m^{-3}$), calculated as the difference between simulated and observed surface SM ($mrsol$ to 10 cm vs. Wang2021OLC top 10 cm), providing spatial insight into model performance. While the results in Figure 3 and Figures A1–A3 show the overall quantitative benchmarking scores generated with ILAMB, the following subsections provide a more detailed analysis of model performance across SM, ecohydrological variables, and their interrelationships.

315

320

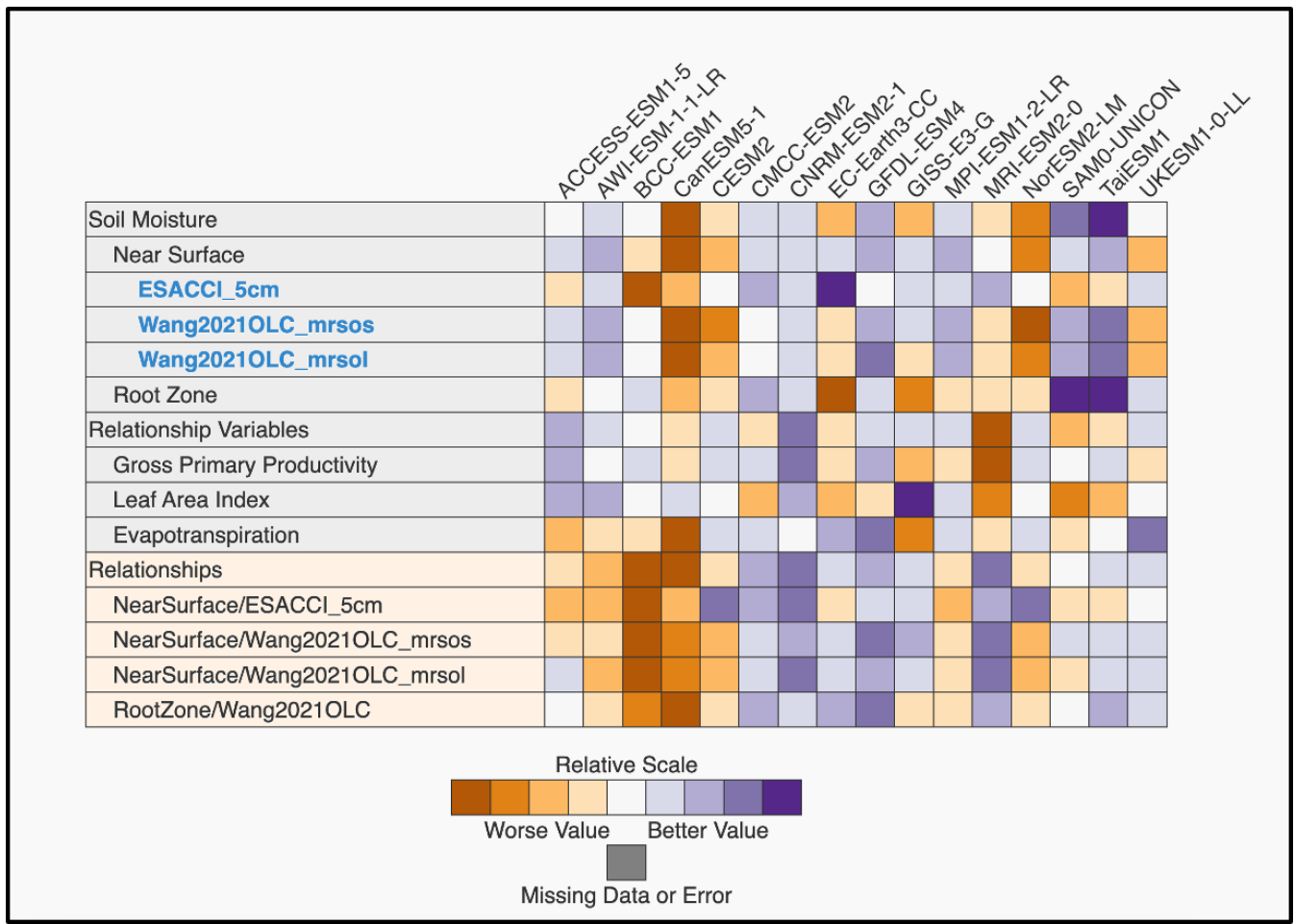


Figure 3: Benchmarking results from the ILAMB global run. Each model is compared against observational datasets: ESA-CCI Surface SM [$\text{m}^3 \text{m}^{-3}$] (top 5 cm), Wang2021OLC Surface SM [$\text{m}^3 \text{m}^{-3}$] (top 10 cm), Wang2021OLC Rootzone SM [$\text{m}^3 \text{m}^{-3}$] (top 100 cm), WECANN GPP [$\text{g m}^{-2} \text{d}^{-1}$], AVH15C1 LAI [unitless], and GLEAMv3.3a ET [mm d^{-1}]. Additionally, the relationships between soil moisture and ecohydrological variables (GPP, LAI, and ET) are also benchmarked. Purple represents better relative skill, while orange indicates relatively worse performance.

3.2 Surface and Rootzone Soil Moisture

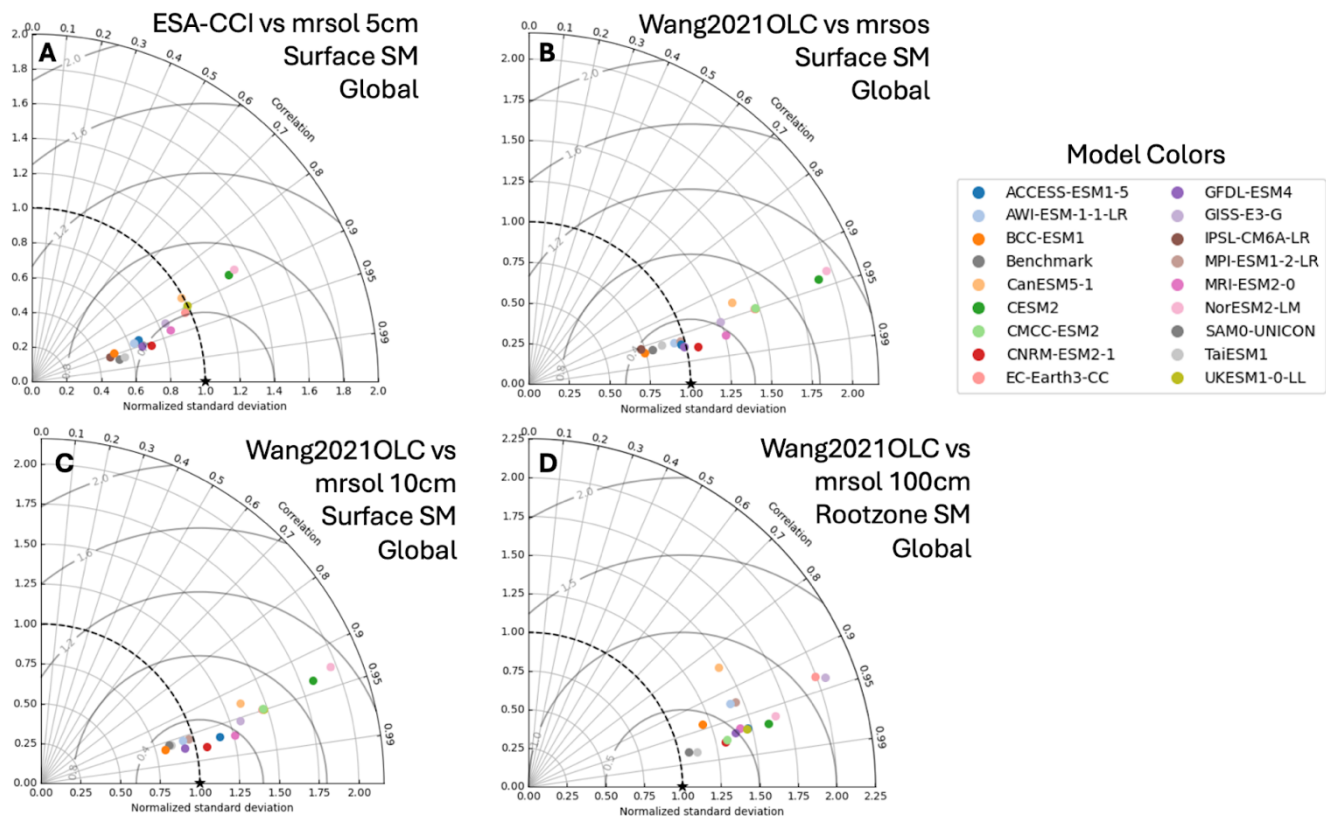
The evaluation of surface and rootzone SM highlights sensitivity both to the observational reference and to the soil depth considered. When surface SM is evaluated using ESA-CCI (top 5 cm) and Wang2021OLC (top 10 cm), roughly half of the models change their relative skill ranking while others remain similar (Figure 3), indicating that benchmarking outcomes at shallow depths depend in part on the choice of observational dataset. In contrast, when different model representations of shallow SM (mrsos and mrsol integrated to 10 cm) are evaluated against Wang2021OLC, most models retain similar rankings, suggesting relatively consistent performance across comparable near-surface depths within a common benchmark. Rootzone SM (0–100 cm), also evaluated against Wang2021OLC, shows a mixture of stable and shifting rankings relative to the surface evaluation: several models change skill category or show modest shifts, while many remain broadly similar.

340 Together, these patterns indicate that differences among observational products influence surface benchmarking outcomes, while depth-dependent differences in model performance emerge more clearly when extending the evaluation to the rootzone.

345 The Taylor diagram analysis (Figure 4), based on monthly mean fields from both models and observational datasets, further illustrates systematic differences between surface and rootzone performance. Across models, correlations with observed surface SM are generally higher than agreement in variability, indicating that models tend to better capture relative wetness and dryness patterns than absolute soil moisture magnitudes. In contrast, rootzone SM exhibits a systematic overestimation of variability across nearly all models. This suggests that ESMs tend to simulate larger fluctuations in deeper soil moisture than indicated by available reference data, pointing to a persistent challenge in representing subsurface water storage and buffering processes. At the same time, the limited availability and greater uncertainty of deep soil moisture observations likely contribute to the reduced observed variability at depth, implying that part of the model–data discrepancy may reflect observational limitations as well as model deficiencies.

355 Together, these results indicate that while many models achieve reasonable performance for surface SM patterns, substantial challenges remain in representing the magnitude and variability of rootzone SM. The contrast between relatively robust surface correlations and systematically inflated deep-soil variability highlights the importance of improving subsurface hydrological processes and their vertical coupling in land surface models

360 The evaluation of surface SM using both ESA CCI (top 5 cm) and Wang2021OLC (top 10 cm) datasets shows broad agreement in overall model rankings (Figure 3), indicating consistent model performance across shallow soil depths. This consistency extends to rootzone SM (0–100 cm) when assessed using the Wang2021OLC dataset, suggesting that several models are generally stable in their representation of SM across soil layers. However, notable discrepancies emerge depending on the benchmark dataset used. Some models perform well relative to ESA CCI but show significantly lower skill against Wang2021OLC, particularly for rootzone SM. Conversely, other models show improved performance when benchmarked against Wang2021OLC compared to ESA CCI. This variation underscores the sensitivity of model evaluation to the choice of reference dataset and highlights the need for multi-dataset benchmarking.



370 **Figure 4: Taylor diagrams evaluating the performance of CMIP6 model SM simulations compared to different observational datasets: A) Surface SM from ESA-CCI (top 5 cm), B) Wang2021OLC Surface SM using *mrsos* from the models (top 10 cm), C) Wang2021OLC Surface SM using *mrsol* from the models (top 10 cm), and D) Wang2021OLC Rootzone SM (top 100 cm).**

375 [Figure 4 presents Taylor diagrams comparing model performance against the four SM benchmark datasets used in this study. These diagrams show that models generally capture the correlation and spatial patterns of surface SM well, though with varying degrees of bias and spread. The results show that models exhibit greater skill in capturing correlations than standard deviations, suggesting they better represent relative wetness and dryness patterns than absolute soil moisture levels. Furthermore, the evaluation of the *mrsos* and *mrsol* to 10 cm SM exhibits similar but non-identical model performance, verifying the difference between the two variables \(Figure 4BC\). In contrast to surface SM, rootzone SM exhibits systematic overestimation of variability across all models. This tendency suggests that ESMs simulate larger fluctuations in deeper SM than are observed, pointing to a key area for improvement in land surface hydrology representations. However, the scarcity of observations for deeper SM likely contributes to the smaller observed variability at this depth, suggesting that part of the model-data discrepancy may stem from limited observational coverage.](#)

380

3.3 Ecohydrological Variables

[The evaluation of ecohydrological variables \(GPP, LAI, and ET\) reveals substantial diversity in model performance across variables and across models \(Figure 3\). Unlike soil moisture, where some coherence in relative ranking is evident, models that perform well for one ecohydrological variable often perform poorly for others, resulting in a wide spread of rankings](#)

385 and few consistently high-performing models across all three benchmarks. This lack of coherence highlights the difficulty of achieving balanced realism in vegetation productivity, canopy structure, and surface fluxes within a single model framework.

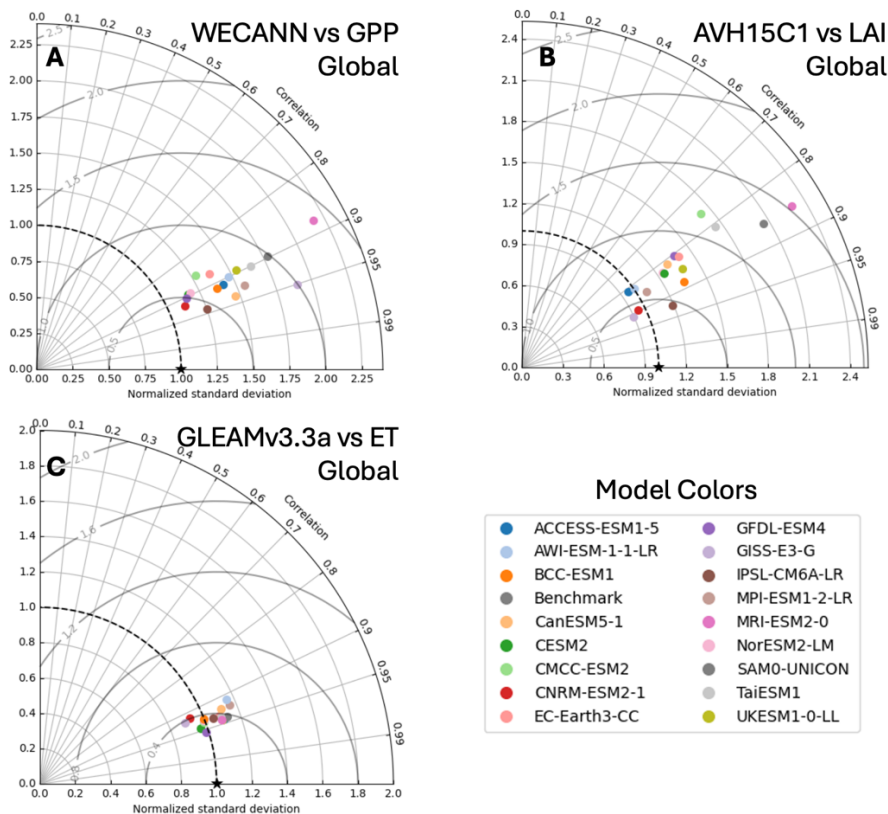
390 The Taylor diagram analysis (Figure 5), based on monthly mean fields from models and observational datasets, further illustrates systematic differences in how these variables are represented. Across most models, both GPP and LAI exhibit a pronounced high bias in variability, indicating that models tend to overestimate fluctuations in vegetation productivity and canopy structure relative to observations. In contrast, ET generally shows stronger agreement with observations, both in terms of spatial correlation and variability. This pattern suggests that while models capture the large-scale controls on surface energy and water fluxes relatively well, they tend to exaggerate the amplitude of vegetation responses.

395 The comparatively stronger performance in ET likely reflects, at least in part, tighter physical constraints through the surface energy balance and the availability of multiple independent observational products used to evaluate and calibrate this variable. Although transpiration is fundamentally a vegetation-controlled process and is represented in land models through the same canopy and stomatal conductance formulations that regulate carbon uptake, ET can nonetheless appear better constrained because it is subject to additional physical and observational constraints.

400 In addition, this relatively strong ET skill may also arise from compensating parameter choices that mask deficiencies in soil moisture representation. For example, biases in soil hydraulic properties or effective soil water holding capacity may compensate for missing or simplified processes such as groundwater interactions or root water uptake, allowing models to reproduce realistic surface fluxes despite underlying errors in subsurface moisture dynamics. By comparison, GPP and LAI are influenced by a broader set of interacting biological and phenological processes, which are represented differently across models and remain more weakly constrained by observations (Massoud et al., 2019; Li et al., 2025b). As a result, larger inter-model spread and systematic variability biases in GPP and LAI are not unexpected, even in cases where ET appears relatively well simulated.

410 Together, these results indicate that skill in simulating surface fluxes does not necessarily imply comparable skill in representing vegetation structure or productivity, and that improvements in ecohydrological realism will likely require more consistent treatment of vegetation dynamics in addition to improvements in hydrological processes.

415 The evaluation of ecohydrological variables (GPP, LAI, and ET) reveals limited consistency in model performance across these variables (Figure 3). That is, models that perform well in simulating one variable often perform poorly in others. This results in a wide spread of rankings, with few models consistently performing well across all three benchmarks. For instance, while some models demonstrate relatively strong performance across GPP, LAI, and ET, others excel in only one variable or underperform across all variables, highlighting the challenge of achieving balanced ecohydrological realism in ESMs.



420

Figure 5: Taylor diagrams evaluating the performance of CMIP6 models in simulating ecohydrologic variables: A) Gross Primary Productivity (GPP) from WECANN, B) Leaf Area Index (LAI) from AVH15C1, and C) Evapotranspiration (ET) from GLEAMv3.3a.

425

430

Taylor diagrams in Figure 5 provide a more detailed comparison of model performance against the ecohydrological observational datasets. Across most models, a systematic high bias in variability is evident for both GPP and LAI, suggesting that models tend to overestimate fluctuations in vegetation productivity and canopy structure. In contrast, model simulations of ET show stronger agreement with observations, both in terms of variability and spatial correlation. This comparatively better performance in ET may reflect stronger observational constraints and more developed parameterizations in hydrologic and surface energy fluxes. However, it may also indicate compensating errors within model processes or parameters that mask deficiencies in SM representation. By comparison, vegetation-related processes such as carbon uptake and phenology likely carry greater structural uncertainty (Massoud et al., 2019; Li et al., 2025b), contributing to more pronounced biases in GPP and LAI.

3.4 Relationship of Soil Moisture to Ecohydrological Variables

435

The relationship between soil moisture and ecohydrological variables provides a more stringent test of model behavior than evaluation of the individual variables alone. Figures 6 and 7 illustrate these relationships for the ACCESS-ESM1-5 model using two different observational surface SM products, based on monthly mean fields consistent with the temporal resolution of the analysis. The comparisons reveal that apparent agreement in individual variables does not necessarily translate into agreement in their coupled behavior, with notable differences emerging depending on the benchmark dataset and the depth of soil moisture considered.

440

When evaluated against the ESA-CCI surface SM product (0–5 cm; Figure 6), models exhibit substantial discrepancies in the diagnosed relationships between SM and GPP, LAI, and ET, indicating that vegetation sensitivity to near-surface moisture is not consistently represented. In contrast, when the Wang2021OLC product is used with SM integrated to 10 cm (Figure 7), agreement between models and observations improves for most variables, suggesting that part of the mismatch in surface-based comparisons arises from differences in observational products and soil depth representation. This highlights the strong sensitivity of coupling diagnostics to both the choice of reference dataset and the effective depth over which soil moisture is evaluated. We note that the climatological fields shown in Figures 6 and 7 are computed within ILAMB from the available gridded data over the evaluation period. Because the ESA-CCI product contains spatial and temporal gaps in coverage, the resulting inferred climatological means and relationships may be influenced by incomplete sampling, potentially affecting the model–data comparison.

445

450

455

Extending this analysis across the full CMIP6 ensemble (Figure 8) confirms that coupling skill varies widely among models and does not simply mirror performance in the individual SM or ecohydrological variables. Some models that perform well for SM or for GPP, LAI, and ET individually nevertheless show weak performance in the coupling metrics, indicating that reproducing realistic mean states or variability is insufficient to guarantee realistic dynamic sensitivity between soil moisture and ecosystem processes. Conversely, other models exhibit relatively stronger coupling skill despite more modest performance in the individual variables, underscoring that different aspects of model behavior are being tested by these diagnostics.

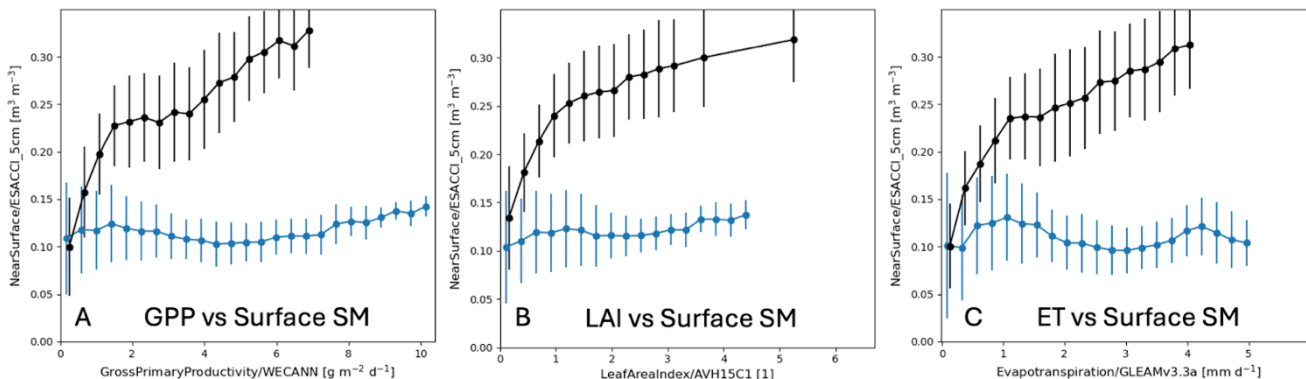
460

Together, these results demonstrate that benchmarking SM–ecosystem relationships provides complementary information to traditional state-based evaluations. The systematic differences in coupling performance across models and datasets point to persistent challenges in representing how soil moisture anomalies propagate into vegetation and surface flux responses, and motivate further examination of the underlying structural differences among land models

465

Figures 6 and 7 illustrate the relationships between SM and key ecohydrological variables (GPP, LAI, and ET) for the ACCESS-ESM1-5 model, comparing model outputs to two different observational SM products. Figure 6 uses the ESA-CCI surface SM product (0–5 cm) alongside model SM integrated to the same depth (*mrsol* to 5 cm). The dot and whisker plots reveal notable discrepancies between model (blue) and observations (black), indicating areas where the model may misrepresent vegetation sensitivity to near-surface SM.

Relationship - *mrsol* to 5 cm (ESA-CCI) and Ecohydrological Variables



470

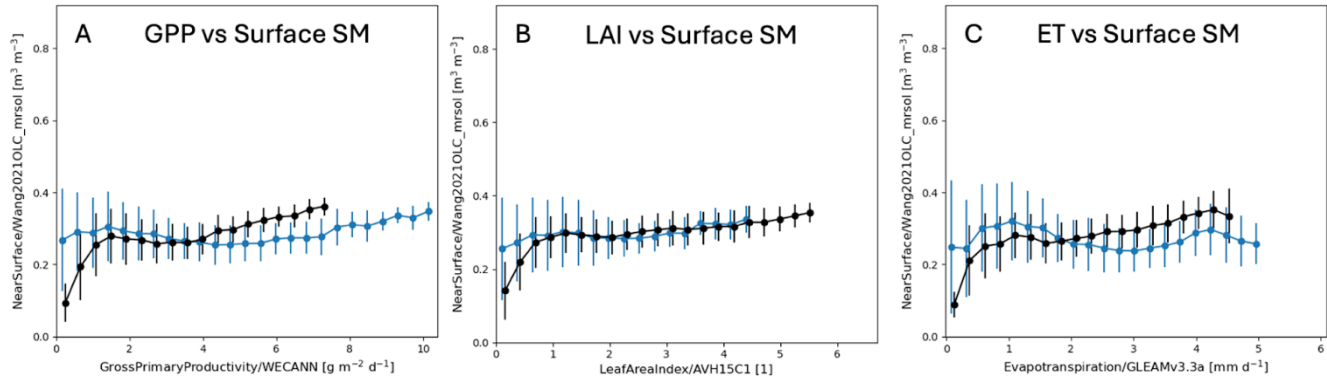
Figure 6: Relationship of SM (y-axis) versus ecohydrological variables (x-axis) shown through dot-and-whisker plots for the ACCESS-ESM1-5 model (blue) and observations (black). The ESA-CCI SM product (typically to 5 cm) is used here for SM, compared to the ACCESS-ESM1-5 model’s SM integrated to the same depth (*mrsol* to 5 cm). The relationship of these SM estimates are shown for (a) WECANN GPP, (b) AVH15C1 LAI, and (c) GLEAMv3.3a ET at

475 the global scale. Whiskers indicate interquartile ranges across land grid cells. These plots reveal a significant
discrepancy between the model and observations when using ESA-CCI as the SM benchmark.

480 Figure 7 presents an analogous analysis using the Wang2021OLC SM product (top 10 cm) and model SM integrated to the
same depth ($mrsol$ to 10 cm). Here, models show improved agreement with observations across all ecohydrological
variables, suggesting that part of the mismatch seen with ESA-CCI may stem from differences in observational datasets or
soil depth representation. 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 (e.f., Preimesberger et al., 2025), potentially biasing its
evaluation. Conversely, Wang2021OLC relies heavily on reanalysis data, which incorporates model structures and may thus
reduce apparent model biases by design. Despite this improvement, persistent biases highlight ongoing challenges in
accurately simulating SM-vegetation coupling within ESMs.

485 Extending this comparison across all models, Figure 8 shows SM-GPP relationships for the full CMIP6 ensemble using
 $mrsol$ to 10 cm and the Wang2021OLC and WECANN datasets. Consistent with ACCESS-ESM1-5 results, models
generally exhibit better agreement when benchmarked against Wang2021OLC compared to ESA-CCI (as in Figure 6);
490 further validating the utility of Wang2021OLC for evaluating SM-vegetation coupling. These findings highlight both the
strengths and limitations of current Earth System Models in representing key ecohydrological interactions, while
demonstrating the value of ILAMB for detailed inter-model diagnostics.

Relationship - $mrsol$ to 10 cm (Wang2021OLC) and Ecohydrological Variables



495 Figure 7: Similar to Figure 6, but using the Wang2021OLC SM product (integrated to 10 cm) and the corresponding
ACCESS-ESM1-5 model SM at the same depth ($mrsol$ to 10 cm). Dot-and-whisker plots show the relationship of soil
moisture (y-axis) versus (a) WECANN GPP, (b) AVH15C1 LAI, and (c) GLEAMv3.3a ET for both model (blue) and
observations (black) at the global scale. Whiskers represent interquartile ranges. Compared to Figure 6, these plots
demonstrate a marked reduction in discrepancies between modeled and observed relationships.

3.5 Other ILAMB capabilities: Köppen Classification

500 While the preceding analyses focus on global-scale benchmarking, one of ILAMB's key strengths is its ability to evaluate
model performance within specific biogeographic regimes. Using Köppen classifications, we assess model behavior across
Tropical, Desert and Semi-arid, Temperate, and Continental climate zones (Figure A4). The regional results from ILAMB
(Figures A5-A9) reveal that model performance is strongly dependent on the climate zone and that global scores can obscure
substantial regional contrasts.

505 In the Tropical region (Figures A5-A6), CLM5-based models (CESM2 and NorESM2-LM) exhibit consistently high skill
across soil moisture, ecohydrological variables, and their relationships, indicating coherent performance in energy- and

510 water-rich environments. In contrast, JSBACH-based models (MPI-ESM1-2-LR and AWI-ESM-1-1-LR) show systematically weaker performance across many of these diagnostics, while UKESM1-0-LL (JULES) exhibits mixed behavior, with relatively stronger performance for rootzone SM and surface fluxes but weaker performance for surface SM and GPP. These patterns suggest that model differences that are muted in global averages can become more pronounced when evaluated in climate regimes dominated by deep rooting, strong transpiration control, and tight soil–vegetation coupling.

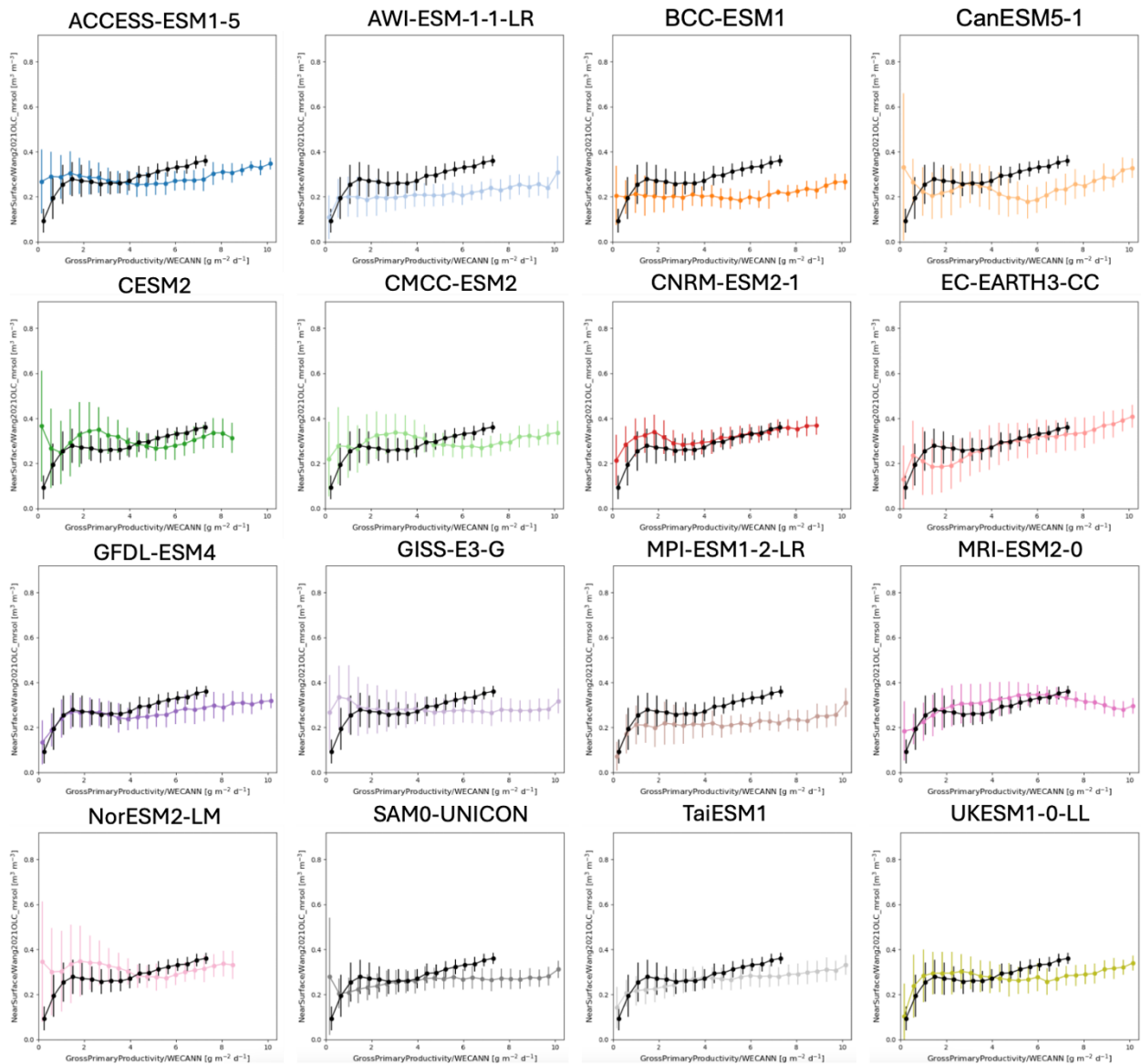
515 In Desert and Semi-arid regions (Figure A7), CLM5-based models again show generally strong performance across most variables, although skill in rootzone SM is reduced, leading to weaker coupling metrics involving deep soil moisture. JSBACH-based models tend to perform relatively well for soil moisture states but less well for ecohydrological variables, resulting in degraded coupling diagnostics. UKESM1-0-LL again shows mixed behavior, with relatively strong performance for rootzone SM and ET but weaker performance for surface SM and GPP, and correspondingly weaker coupling metrics. These results highlight that in water-limited environments, differences in how soil moisture anomalies propagate into vegetation and flux responses become particularly important.

520 In Temperate regions (Figure A8), CLM5-based models maintain generally high performance across most benchmarks, although some degradation is evident for LAI. JSBACH-based models show relatively good performance for surface SM and LAI but weaker performance for rootzone SM, GPP, and ET, leading to poor coupling metrics. UKESM1-0-LL again exhibits stronger performance for rootzone SM and ecohydrological variables than for surface SM, with coupling metrics remaining comparatively weak. This intermediate regime underscores the growing importance of vertical soil moisture structure and rooting depth in shaping both state and coupling performance.

525 The Continental (high-latitude) region (Figure A9) shows the most pronounced divergence among models. Here, CLM5-based models exhibit markedly degraded performance across soil moisture, ecohydrological variables, and their relationships, likely reflecting the strong influence of cold-region processes such as organic soils, freeze–thaw dynamics, and high soil porosity, which lead to substantially higher simulated soil moisture than indicated by current observational products. In contrast, JSBACH-based models show relatively good performance for surface SM and ecohydrological variables but weaker performance for rootzone SM, resulting in depth-dependent differences in coupling skill. These results demonstrate that high-latitude regions pose distinct challenges for both observations and models, and that process representations that improve performance in one climate regime may degrade it in another.

530 Taken together, the regional diagnostics show that no single model performs uniformly well across all climate regimes and benchmarks. The strong regime dependence of model skill highlights the importance of regional evaluation for identifying both strengths and weaknesses that are not evident from global metrics alone, and reinforces the need for process-based interpretation of benchmarking results.

535 While the preceding analyses focus on global-scale benchmarking (which is the default in ILAMB) one of ILAMB’s key strengths is its ability to evaluate model performance across specific biogeographic zones. Using Köppen classifications, ILAMB enables targeted regional analysis across zones such as Tropical, Desert and Semi-arid, Temperate, and Continental climates. Figure A4 illustrates these regions and shows regional mean values of surface SM from the ESA-CCI product to provide geographic context.



550 **Figure 8: Similar to the bias plots shown in Figure 7A, but for all models. These plots highlight the observed relationships (black curves) between the surface SM (from the Wang2021OLC top 10 cm SM product) and GPP (from WECANN), as well as the relationship shown in each model (colored curves) using the simulated m_{rsol} to 10 cm and gpp variables.**

555 [To demonstrate the utility of regional benchmarking, Figure A5 presents ILAMB portrait style evaluations \(using the same methodology as Figure 3\) and Figure A6 shows the evaluated relationships for the Tropical region \(using the same methodology as Figures 6–7\). The ILAMB portrait style evaluations for the remaining Köppen regions are shown in Figure A7 \(Desert and semi arid\), A8 \(Temperate\), and A9 \(Continental\), again using the same methodology as Figure 3.](#)

560 These localized portraits highlight how model performance can vary significantly by region. For example, some models like CESM2 and NorESM2-LM exhibit a high bias in high northern latitudes when evaluated globally (as depicted in Figure A2), yet perform relatively well in the Tropical region (Figure A5). Conversely, their performance declines in the Continental zone (Figure A8). EC-Earth3-CC scores highly in the Tropical region when evaluated against ESA-CCI but ranks among the lowest when evaluated against the Wang2021OLC dataset for the same region, highlighting uncertainty in the benchmarking datasets in the region. In contrast, it shows more consistent agreement across both observational products in Continental areas.

570 ILAMB's regional diagnostics also extend to ecohydrological relationships. Figure A6 mirrors the SM-ecohydrology analysis shown in Figure 7 but focuses specifically on the Tropical region and surface SM. In this example, the ACCESS-ESM1-5 model shows reasonably good agreement with observed relationships in the Tropical region when evaluated against the Wang2021OLC dataset. This supports earlier findings that surface SM comparisons with Wang2021OLC tend to yield consistent model-observation alignment. These regional results further emphasize the value of localized analyses for identifying where models are performing well and where region-specific improvements are needed.

4 Discussion

4.1 Potential Drivers for Divergences in Model Simulations

575 The substantial spread in SM estimates among CMIP6 models reflects differences in how land surface models (LSMs) represent key processes and input data, as evidenced by both the global and Köppen-region ILAMB diagnostics (Sections 3.1 and 3.5). Although precipitation is a major driver of SM, it does not explain much of the mean inter-model variability in this study. Most CMIP6 models simulate similar global mean precipitation rates of approximately 2.5 mm day⁻¹ (Tapiador et al., 2018), suggesting that differences in total precipitation alone are insufficient to account for the spread in simulated SM (see Figure A1, column "Model Period Mean"). However, the timing and intensity of precipitation events, rather than total amounts, may contribute significantly to model divergence, especially in regions with highly seasonal precipitation regimes where the timing of precipitation governs the length of dry-down periods. This effect is particularly pronounced in seasonally snowmelt-dominated systems (Harpold et al., 2015), where variations in precipitation and snowmelt seasonality strongly influence the partitioning between infiltration, runoff, and ET. Differences in how models represent storm frequency, sub-daily rainfall variability, and snowmelt dynamics can therefore propagate into systematic differences in SM variability and ecohydrological coupling, as seen in the regional results.

580 Similarly, the treatment of soil ice is unlikely to contribute significantly to SM variability in a direct sense, as the CMIP6 definition of SM includes both liquid water and ice, which should be handled consistently across models and is only relevant in cold regions. This implies that other factors, such as soil properties, hydrologic parameterizations, and land-atmosphere feedbacks, are likely more important contributors to the inter-model spread highlighted in Sections 3.2–3.5.

595 One important driver is the treatment of soil properties, particularly in high-latitude regions. Models that include representations of organic soils, such as CESM2 and NorESM2-LM, tend to simulate higher SM levels, likely due to the greater water-holding capacity of organic matter. Both models use the CLM5 land surface scheme, which explicitly accounts for organic soil processes in northern latitudes (Lu et al., 2020). This feature provides a plausible explanation for the large positive SM biases seen for these models in the Continental region (Figure A9 and Figure A3), while also being consistent with their comparatively strong performance in Tropical and Temperate regions (Figures A5 and A8), where organic-rich soils and strong vegetation-soil coupling are widespread. It remains unclear whether all CMIP6 models have updated their treatment of organic soils to a similar degree, which could contribute to persistent inter-model differences in both SM states and SM-vegetation relationships.

600

605 Soil porosity is another critical input affecting SM simulations. Previous work (Dai et al., 2019) has shown that many CMIP5 models used outdated soil maps (e.g., FAO-UNESCO, 1981), and some CMIP6 models may still rely on these legacy datasets. More modern global soil datasets, such as the Harmonized World Soil Database v2.0 (Nachtergaele et al., 2023), the Global Soil Dataset for Earth System Modeling (Shangguan et al., 2014), and SoilGrids (Hengl et al., 2014), offer improved estimates of soil properties, including porosity. Differences in the underlying soil maps and parameter choices may therefore help explain why some models perform well for surface SM but poorly for rootzone SM, or vice versa, as seen in several Köppen regions (Section 3.5).

610 Model-specific representations of hydrologic processes further influence SM estimates and coupling behavior. For example, EC-Earth3-CC consistently simulates high SM values, potentially due to its inclusion of groundwater–soil interactions not present in many other models. With a few exceptions (Oleson et al., 2013; Lawrence et al., 2019), ESMs generally rely on “bucket-type” soil moisture representations that drain water between layers based on thresholds (e.g., field capacity) and rates (e.g., hydraulic conductivity), rather than fully pressure-driven flow as in Richards’ equation. Across the CMIP6 ensemble, differences in the number and spacing of soil layers, rooting profiles, ET parameterizations, infiltration schemes, soil hydraulic properties, total soil depth, and lower boundary conditions (e.g., free drainage versus groundwater influence; see Table A1) likely contribute to the contrasting behaviors observed among model families, such as the different surface-versus-rootzone and state-versus-coupling skill patterns seen for CLM5-, JSBACH-, and JULES-based models.

620 Finally, land–atmosphere feedbacks may also contribute to the inter-model spread. CMIP models have been shown to overestimate the strength of feedbacks between SM and atmospheric variables, including temperature, ET, and surface fluxes (Levine et al., 2016). These exaggerated feedbacks may amplify SM variability and introduce additional divergence across models. For instance, models that strongly couple soil moisture to surface heat fluxes may simulate more aggressive drying or wetting cycles, depending on local conditions (Schumacher et al., 2022), potentially contributing to the widespread pattern of relatively good ET skill but weaker SM–ET coupling seen in several models (Section 3.4). The degree of inter-model variability in feedback strength is probably quite high, but further studies are needed to quantify how this affects inter-model spread in SM (Vogel et al., 2016; Talib et al., 2023).

630 In summary, the observed variability in SM simulations across CMIP6 models likely stems from a combination of factors, including differences in soil properties, legacy input datasets, precipitation event timing, hydrologic parameterizations, vertical structure and boundary conditions, and land–atmosphere feedbacks. The regime-dependent patterns revealed by the Köppen analysis suggest that these factors operate differently across climate zones, reinforcing the need for both global and regional diagnostics when evaluating and improving SM representation in ESMs.

635 The substantial spread in SM estimates among CMIP6 models reflects differences in how land surface models (LSMs) represent key processes and input data. Although precipitation is a major driver of SM, it does not explain much of the mean inter-model variability in this study. Most CMIP6 models simulate similar global mean precipitation rates of approximately 2.5 mm/day (Tapiador et al., 2018), suggesting that differences in total precipitation alone are insufficient to account for the spread in simulated SM (see Figure A1, column ‘Model Period Mean’). However, the timing and intensity of precipitation events, rather than total amounts, may contribute significantly to model divergence, especially in regions with highly seasonal precipitation regimes where the timing of precipitation governs the length of dry down periods. This effect is particularly pronounced in seasonally snowmelt-dominated systems (Harpoel et al., 2015), where variations in precipitation and snowmelt seasonality strongly influence the partitioning between infiltration, runoff, and ET. Differences in how models represent storm frequency, sub-daily rainfall variability, and snowmelt dynamics can further impact SM through their interactions with soil hydrology and vegetation water uptake. Future work should investigate the role of precipitation and snowmelt timing, along with event scale dynamics, in driving model spread.

650 Similarly, the treatment of soil ice is unlikely to contribute significantly to SM variability, as the CMIP6 definition of SM includes both liquid water and ice, which should be handled consistently across models and only present in cold areas with minimum snow cover. This implies that other factors, such as soil properties, hydrologic parameterizations, and land–atmosphere feedbacks, are likely more important contributors.

655 One important driver is the treatment of soil properties, particularly in high latitude regions. Models that include
representations of organic soils, such as CESM2 and NorESM2 LM, tend to simulate higher SM levels, likely due to the
greater water holding capacity of organic matter. Both models use the CLM5 land surface scheme, which explicitly accounts
for organic soil processes in northern latitudes (Lu et al., 2020). This feature may explain the higher SM values observed in
660 these models' bias maps (e.g., Figure A3). However, the inclusion of organic soils in CESM2 and NorESM2 LM models
may have led to better SM representation in the tropics (Figure A5), where organic soil is prevalent in the tropical forests. It
remains unclear whether all CMIP6 models have updated their treatment of organic soils, which could contribute to
persistent inter-model differences.

665 Soil porosity is another critical input affecting SM simulations. Previous work (Dai et al., 2019) has shown that many
CMIP5 models used outdated soil maps (e.g., FAO UNESCO, 1981), and some CMIP6 models may still rely on these
legacy datasets. More modern and accurate global soil datasets, such as the Harmonized World Soil Database v2.0
(Nachtergaele et al., 2023), the Global Soil Dataset for Earth System Modeling (Shangguan et al., 2014), and SoilGrids
(Hengl et al., 2014), offer improved estimates of soil properties, including porosity. Comparing SM outputs against these
updated datasets could help reveal systematic biases tied to inaccurate soil input data.

670 Model specific representation of hydrologic processes can also influence SM estimates. For example, the EC Earth3 CC
model consistently simulates high SM values, potentially due to its inclusion of groundwater soil interactions not present in
other models. With a few exceptions (Oleson et al., 2013; Lawrence et al., 2019), ESMs generally use 'bucket type' soil
moisture representations that drain water between different soil layers based on predefined thresholds (e.g., field capacity)
and rates (e.g., hydraulic conductivity) rather than the pressure driven movement represented by more physically based
675 models like the Richard's Equation. Although ESMs differ, several key processes and parameters are common and should be
considered in model performance differences: number and spacing of soil layers, rooting profile and ET process
representation, infiltration process representation, soil properties (i.e. porosity, wilting point and field capacity water content
at each layer, and hydraulic conductivity), and soil depth and the bottom boundary condition (e.g., groundwater versus freely
draining). Such process level differences highlight the role of internal model design in shaping SM outputs that remain
680 challenging to disentangle.

685 Finally, land atmosphere feedbacks may also contribute to the inter model spread. CMIP models have been shown to
overestimate the strength of feedbacks between SM and atmospheric variables, including temperature, ET, and surface fluxes
(Levine et al., 2016). These exaggerated feedbacks may amplify SM variability and introduce additional divergence across
models. For instance, models that strongly couple soil moisture to surface heat fluxes may simulate more aggressive drying
or wetting cycles, depending on local conditions (Schumacher et al., 2022). The degree of inter model variability in feedback
strength is probably quite high, but more studies will be needed to understand how this affects inter model spread in SM
(Vogel et al., 2016; Talib et al., 2023). Further research is needed to examine how different models simulate the sensitivity
of ET and energy partitioning to meteorological conditions such as vapor pressure deficit, solar radiation, or wind speed,
690 factors that could also shape SM variability via land atmosphere interactions.

695 In summary, the observed variability in SM simulations across CMIP6 models likely stems from a combination of factors,
including differences in soil properties, legacy input datasets, precipitation event timing, hydrologic parameterizations and
process representations, and feedback mechanisms. Future research should aim to disentangle the relative contributions of
these drivers to better constrain SM representation in ESMs.

4.2 Shared characteristics of models with similar performance

The global and regional ILAMB diagnostics reveal clear clustering of model behavior by land surface framework (Sections 3.1 and 3.5; Table A1), indicating that structural choices in land model formulation exert a first-order control on both state and coupling performance. Models that share a common land surface model tend to exhibit similar strengths and weaknesses

700 across soil moisture (SM), ecohydrological variables, and their interrelationships, suggesting that performance differences are not purely idiosyncratic but are linked to shared process representations.

705 CESM2 and NorESM2-LM, which both use CLM5, provide a clear example of this structural coherence. These models show relatively strong and internally consistent performance for ecohydrological variables (GPP, LAI, and ET) and their coupling to SM across several regions, particularly in the Tropical and Temperate zones (Section 3.5, Figures A5 and A8). This behavior is consistent with CLM5’s more detailed treatment of soil hydrology and explicit soil–plant hydraulic stress, which promotes coherent interactions between soil water availability, transpiration, and photosynthesis (Table A1). At the same time, both models exhibit degraded performance in the Continental (high-latitude) region, where their inclusion of organic soils and cold-region processes leads to systematically higher simulated SM than indicated by current observational products (Figure A9). This combination of strong ecohydrological coupling but biased SM states highlights how a given structural choice can improve process coherence while simultaneously introducing regional biases.

710 In contrast, the JSBACH-based models (MPI-ESM1-2-LR and AWI-ESM-1-1-LR) tend to show relatively good performance for some SM state variables, particularly near the surface in several regions, but weaker performance for ecohydrological variables and, notably, for SM–vegetation coupling metrics. This pattern is especially evident in arid and temperate regimes (Section 3.5), where these models often reproduce mean SM reasonably well but fail to translate SM variability into realistic vegetation or flux responses. A plausible explanation is that JSBACH relies on more implicit vegetation water-stress formulations and simplified soil–plant coupling (Table A1), which may limit the model’s ability to propagate soil moisture anomalies consistently into photosynthesis and transpiration, even when the soil moisture state itself is reasonably represented.

715 UKESM1-0-LL (JULES) exhibits a different, mixed behavior. This model often shows weaker performance for surface SM but comparatively better performance for rootzone SM and surface fluxes, along with moderate skill in some coupling metrics. This depth-dependent pattern is consistent with JULES’s prescribed root distributions and implicit soil moisture stress functions, which can yield reasonable integrated flux responses while still allowing biases in shallow soil layers (Table A1). The recurring finding that UKESM1-0-LL captures some aspects of ecohydrological behavior despite weaker SM states reinforces the idea that realistic fluxes do not necessarily guarantee realistic subsurface moisture dynamics.

720 Taken together, these results demonstrate that similarities in land surface model architecture translate into recognizable performance “signatures” across both global and regional benchmarks. They also show that models can achieve good performance in individual variables while still failing to capture the correct coupling between soil moisture and ecosystem processes. This underscores the importance of evaluating not only state variables but also cross-variable coherence, and suggests that targeted improvements to soil hydrology, rooting depth distributions, and soil–plant coupling schemes could yield broad benefits across multiple aspects of model performance.

725 Models that exhibit similar performance across multiple variables often share underlying structural and process-based characteristics. For example, CESM2 and NorESM2-LM, while performing relatively poorly in simulating SM (e.g., Figure 3), show stronger agreement with observations for ecohydrological variables such as GPP, LAI, ET. A shared feature of these models is their use of the CLM5, which likely contributes to their consistent performance in simulating vegetation dynamics and surface fluxes. However, despite these strengths, both models apparently exhibit a persistent high-latitude bias in SM (Figure A3), likely related to CLM5’s representation of soil properties and its inclusion of organic matter in northern soils (Lu et al., 2020). However, considering the bias of satellite observations under snow- and ice-covered conditions (Dorigo et al., 2015) and the relatively high influence of reanalysis and land surface models on the Wang2021OLC product (Wang et al. 2021), this bias in high latitudes warrants further investigation with more observational datasets.

730 In addition to similarities in their overall performance scores, CESM2 and NorESM2-LM also exhibit relatively consistent relationships between SM and ecohydrological variables. These relationships, such as SM-GPP and SM-ET coupling, reflect each model’s ability to capture key land-atmosphere feedbacks, including how soil water availability influences stomatal conductance, photosynthetic rates, and vegetation phenology. Notably, this suggests that even when absolute SM values are

735

740

745

750 biased, some models may still effectively represent the underlying functional dynamics between water and vegetation. In contrast, models with weaker SM performance tend to display greater mismatches in these coupling relationships, indicating broader limitations in their ability to simulate ecohydrological processes.

755 Taken together, these patterns highlight that model performance is influenced not only by how individual processes are parameterized, but also by how different components of the land system interact. Structural choices, such as vegetation and soil representations, hydrologic schemes, and coupling strategies, shape a model's capacity to simulate both state variables and their interrelationships. As such, evaluating models across multiple variables and focusing on cross-variable coherence can provide deeper diagnostic insight than isolated benchmark scores. These shared behaviors among models with similar architectures may point to areas where coordinated model development and targeted improvements, such as refining soil parameterizations or vegetation hydrology interactions, could yield broad performance gains.

4.3 Consideration of uncertainty and limitations in the benchmark results

765 Evaluating model performance inherently involves uncertainty, both from observational datasets and from model representations. To better characterize this uncertainty, we employed multiple observational products for surface and rootzone SM, including ESA-CCI and Wang2021OLC surface SM, as well as Wang2021OLC rootzone SM. By comparing models against more than one observational reference, we aimed to capture a broader envelope of observational variability, rather than relying on any single product that may contain its own structural or systematic biases. We reiterate that none of the benchmark datasets used here represents a definitive truth, and that our multi-product approach is intended to reduce dependence on any single reference and to better characterize observational uncertainty. Similarly, we assessed different model representations of SM, including the `mrsos` and `mrsol` variables integrated to 5 cm, 10 cm, and 100 cm depths, to
770 examine how model configuration affects evaluation outcomes.

775 We note that GPP and LAI are integrative ecosystem variables influenced by multiple interacting processes beyond soil moisture alone, and therefore discrepancies in SM–GPP or SM–LAI relationships cannot be interpreted mechanistically as arising solely from deficiencies in soil hydrology. In this study, these relationships are instead used to assess whether models reproduce the emergent ecosystem-scale coupling between soil moisture variability and vegetation functioning. More process-level hydrological quantities, such as runoff, drainage, or evaporation–transpiration partitioning, would be valuable for mechanistic diagnosis of soil moisture controls, but their use as global benchmarking targets is currently limited by observational availability and consistency. We therefore focus on GPP, LAI, and ET as comparatively well-constrained, widely evaluated ecosystem-scale diagnostics, while acknowledging the associated limitations in process-level attribution.

780 Soil moisture presents unique challenges in benchmarking due to the diversity and limitations of observational data sources. Satellite-derived products offer broad spatial coverage but are limited to shallow depths and may have coarser resolution for passive microwave technology, whereas active microwave retrievals can achieve higher resolution but are subject to greater uncertainty from surface roughness and vegetation structure (Zeng et al., 2023). In contrast, in-situ observations provide
785 deeper and more accurate point measurements but are geographically sparse and face substantial scaling challenges when compared against ESM grid cells. While our study did not include in-situ datasets, their inclusion could strengthen future benchmarks, particularly for deeper soil layers and for regions where satellite coverage is limited. The variation in spatial resolution, vertical depth coverage, and methodological differences between observational datasets contributes to uncertainty in benchmarking, and any conclusions drawn from SM comparisons, especially in data-sparse regions such as high latitudes and arid zones, should be interpreted within this context.

790 To address these complexities, we adopted a multi-product comparison approach. However, additional strategies may further enhance benchmarking efforts. In atmospheric science, for example, it is common to evaluate models not only against individual datasets but also against an ensemble of observations (Yamaguchi et al., 2015), which helps define a consensus observational baseline. While this strategy is less frequently used in the SM modeling community, it may offer a valuable pathway forward, particularly when observational datasets diverge or when no single product can be considered definitively
795 superior.

800 Beyond observational uncertainty, ensemble modeling (using a collection of model runs or parameter sets) can also help characterize structural and parametric uncertainty within the models themselves. While not implemented in this study, future benchmarking efforts could benefit from integrating probabilistic or ensemble-based approaches to better understand model sensitivity (cf. Massoud et al., 2019) and confidence in simulated soil moisture.

805 Evaluating model performance inherently involves uncertainty, both from observational datasets and model representations. To better characterize this uncertainty, we employed multiple observational products for surface and rootzone SM, including ESA-CCI and Wang2021OLC surface SM, as well as Wang2021OLC rootzone SM. By comparing models against more than one observational reference, we aimed to capture a broader envelope of observational variability, rather than relying on any single product that may contain its own structural or systematic biases. Similarly, we assessed different model representations of SM, including the `mrsos` and `mrsol` variables integrated to 5 cm, 10 cm, and 100 cm depths, to examine how model configuration affects evaluation outcomes.

815 Soil moisture presents unique challenges in benchmarking due to the diversity and limitations of observational data sources. Satellite-derived products, for example, offer broad spatial coverage but are limited to shallow depths and may have coarser resolution for passive microwave technology, whereas active microwave technology can achieve higher resolution but is subject to greater uncertainty from surface roughness and vegetation structure variations (Zeng et al., 2023). In contrast, in-situ observations provide deeper and more accurate point measurements but are geographically sparse and face the scaling-up problem when compared against ESMs. While our study did not include in-situ datasets, their inclusion could strengthen future benchmarks by adding another layer of observational comparison, particularly for deeper soil layers. The variation in spatial resolution, vertical depth coverage, and methodological differences between observational datasets contributes to uncertainty in benchmarking, and any conclusions drawn from SM comparisons should be interpreted within this context.

825 To address these complexities, we adopted a multi-product comparison approach. However, additional strategies exist and may further enhance benchmarking efforts. In atmospheric science, for example, it is common to evaluate models not only against individual datasets but also against an ensemble of observations (Yamaguchi et al., 2015), which helps define a consensus observational baseline. While this strategy is less frequently used in the SM modeling community, it may offer a valuable pathway forward, particularly when observational datasets diverge or when no single product can be considered definitively superior.

830 Beyond observational uncertainty, ensemble modeling (using a collection of model runs or parameter sets) can also help characterize structural and parametric uncertainty within the models themselves. While not implemented in this study, future benchmarking efforts could benefit from integrating probabilistic or ensemble-based approaches to better understand model sensitivity (e.f., Massoud et al., 2019) and confidence in simulated soil moisture.

5 Conclusion

835 This study evaluated the performance of 16 CMIP6 models in simulating key land surface and ecohydrological variables, including surface and rootzone soil moisture (SM), gross primary productivity (GPP), leaf area index (LAI), and evapotranspiration (ET), using the ILAMB framework. By combining state-variable benchmarks with diagnostics of SM–ecohydrology coupling and regional (Köppen) analyses, we provide a multi-faceted assessment of land model behavior. The main findings are summarized as follows:

840 (i) Model performance in simulating soil moisture (SM) varies substantially depending on the observational benchmark used. For example, some models (e.g., EC-Earth3-CC) perform well against ESA-CCI surface SM but perform substantially worse relative to the Wang2021OLC product. This sensitivity highlights the importance of using multiple observational references to better characterize uncertainty and avoid conclusions based on a single dataset.

845 (ii) Model skill differs markedly across variables and reflects structural similarities among land surface model frameworks. For instance, CESM2 and NorESM2-LM exhibit relatively strong performance for ecohydrological variables (GPP, LAI, and ET) but show persistent SM biases in high-latitude regions. These shared behaviors reflect their common use of the CLM5 land surface model and illustrate how structural choices shape both strengths and weaknesses in simulated land surface states and fluxes. In particular, models with more explicit soil–plant hydraulic coupling, such as CLM5-based models, tend to
850 reproduce SM–vegetation relationships more consistently across regions, whereas models relying on more implicit vegetation water-stress formulations (e.g., JSBACH- and JULES-based models) more often reproduce soil moisture states without fully capturing the associated vegetation or flux responses.

855 (iii) Evaluating SM–ecohydrology relationships further shows that models more consistently reproduce observed coupling patterns when assessed against the Wang2021OLC dataset than against ESA-CCI. This suggests that deeper soil moisture estimates may provide a more stable reference for diagnosing vegetation–water interactions, while also reinforcing that good performance in individual variables does not necessarily imply realistic cross-variable sensitivity.

860 (iv) Model skill also varies strongly by climate regime. For example, CESM2 and NorESM2-LM perform relatively well in Tropical and Temperate regions but degrade substantially in Continental (high-latitude) zones, where cold-region processes and organic soils play a larger role. These results highlight the limitations of relying on global mean scores alone and demonstrate the value of regional benchmarking for diagnosing where and why models succeed or fail.

865 Taken together, these results indicate that no single CMIP6 model performs best across all variables, coupling metrics, or regions. Instead, performance patterns often cluster by land surface model family, suggesting that structural choices in soil hydrology, rooting representation, and soil–plant coupling influence both soil moisture states and ecohydrological interactions. The ILAMB framework therefore provides a useful approach not only for benchmarking individual variables but also for diagnosing cross-variable coherence and regional dependencies in model behavior.

870 Future work should expand these benchmarking efforts by incorporating in-situ observations, applying ensemble-based and probabilistic approaches to both models and observations, and further refining regional analyses. Such developments will be important for improving constraints on soil moisture dynamics, strengthening understanding of land–atmosphere coupling, and guiding the development of next-generation Earth system models, including those participating in CMIP7.

875 This study evaluated the performance of 16 CMIP6 models in simulating key land surface and ecohydrological variables, including surface and rootzone SM, GPP, LAI, and ET, using the ILAMB framework. The main findings are summarized as follows:

880 (i) Our analysis showed substantial variability in model performance in simulating SM, particularly when compared against more than one observational dataset. For instance, models such as EC Earth3-CC performed well against ESA-CCI surface SM observations but underperformed when benchmarked against the Wang2021OLC dataset. This stresses the importance of using multiple observational products to better capture uncertainty and avoid drawing conclusions based on a single reference dataset.

885 (ii) Models often showed differing levels of skill across variables. For example, CESM2 and NorESM2-LM exhibited strong performance in ecohydrological variables like GPP, LAI, and ET, but showed consistent SM biases in high-latitude regions. These shared behaviors likely reflect common model structures, as both rely on the CLM5 land surface model. This highlights how model architecture and parameterizations influence simulation outcomes.

890 (iii) Beyond individual variables, we examined the relationships between soil moisture and ecohydrological processes. This analysis revealed that models tended to capture these relationships more effectively when compared to the Wang2021OLC dataset than to ESA-CCI, indicating that deeper or better-integrated soil moisture estimates may provide a more consistent benchmark for evaluating vegetation–water coupling.

895 (iv) The regional analysis using Köppen climate zones demonstrated ILAMB's capacity to localize model evaluations. Results showed that model skill can vary significantly by region. For instance, CESM2 and NorESM2-LM performed relatively well in tropical zones but struggled in continental regions. This highlights the need to move beyond global averages to uncover spatially varying model behavior. Regional benchmarking enables more precise identification of where and why individual ESMs succeed or fall short.

900 Overall, our findings showcase the strengths and limitations of current CMIP6 models in simulating land surface and ecohydrological processes. While some models excel in specific areas, no single model performs best across all metrics or regions. The ILAMB framework provides a powerful, systematic approach for benchmarking not only model outputs but also functional relationships among variables. This study also reinforces the value of considering multiple observational products and model configurations to account for uncertainty in benchmarking efforts.

905 Future work should continue to expand benchmarking approaches by incorporating in-situ observations, applying ensemble-based and probabilistic methods for both models and observational datasets, and further refining regional analyses. Ensemble approaches, in particular, can help quantify structural uncertainty and improve robustness in performance assessments by leveraging the diversity across models and observational products. Such efforts will be critical for improving our understanding of SM and land-atmosphere interactions and enhancing the development of next-generation ESMs.

910

Appendix

<u>CMIP6 Model</u>	<u>Land Model</u>	<u>Hydrology Scheme</u>	<u>Root Distribution</u>	<u>Lower Boundary Condition</u>	<u>Plant Hydraulic s</u>	<u>Notes</u>	<u>Source</u>
<u>CESM2 / NorESM2 (CLM5)</u>							
<u>CESM2</u>	<u>CLM5</u>	<u>Richards-based (multi-layer soil)</u>	<u>PFT-dependent, depth-varying</u>	<u>Free drainage + runoff scheme</u>	<u>Explicit soil-plant hydraulics</u>	<u>Includes plant hydraulic stress and multi-layer soil water flow</u>	<u>Lawrence et al., 2019, JAMES (CLM5)</u>
<u>NorESM2-LM</u>	<u>CLM5</u>	<u>Richards-based</u>	<u>PFT-dependent, depth-varying</u>	<u>Free drainage + runoff</u>	<u>Explicit</u>	<u>Same land model as CESM2</u>	<u>Lawrence et al., 2019</u>
<u>UKESM1 (JULES)</u>							

UKESM1-0-LL	JULES-ES	Richards-based	Prescribed, PFT-dependent	Free drainage	Implicit (soil moisture stress functions)	No explicit plant hydraulic continuum in standard JULES	Best et al., 2011; Clark et al., 2011 (JULES)
MPI-ESM / AWI-ESM (JSBACH)							
MPI-ESM1-2-LR	JSBACH	Richards-based (multi-layer)	Prescribed, PFT-dependent	Free drainage	Implicit	Soil hydrology with layered diffusion and runoff parameterization	Reick et al., 2021 (JSBACH)
AWI-ESM1-1-LR	JSBACH	Richards-based	Prescribed, PFT-dependent	Free drainage	Implicit	Same JSBACH core	Reick et al., 2021

915 **Table A1: Overview of key land surface process representations for some of the selected CMIP6 models used in this study, including hydrology scheme, root distribution, lower boundary condition, and treatment of plant hydraulics. This table provides structural context for interpreting differences in soil moisture, ecohydrological variables, and their coupling across models. Sources for each entry are listed in the table. Because process-level documentation is not consistently available for all CMIP6 models, this table summarizes key characteristics only for models with clearly documented information in the literature.**

920

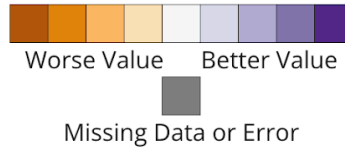
Benchmark	Download Data Period Mean	Model Period Mean (original grids) [m3 m-3]	Benchmark Period Mean (intersection) [m3 m-3]	Model Period Mean (intersection) [m3 m-3]	Benchmark Period Mean (complement) [m3 m-3]	Model Period Mean (complement) [m3 m-3]	Bias [m3 m-3]	RMSE [m3 m-3]	Phase Shift [months]	Bias Score [1]	RMSE Score [1]	Seasonal Cycle Score [1]	Spatial Distribution Score [1]	Overall Score [1]
Benchmark	0.240													
ACCESS-ESM1-5	0.251	0.246	0.240	0.307	0.227		0.00644	0.0557	1.72	0.379	0.506	0.752	0.962	0.621
AWI-ESM1-1-LR	0.180	0.182	0.239	0.149	0.244		-0.0577	0.0699	1.28	0.278	0.567	0.824	0.974	0.642
BCC-ESM1	0.187	0.182	0.240	0.241	0.158		-0.0576	0.0723	1.36	0.265	0.525	0.811	0.939	0.613
CanESM5-1	0.217	0.203	0.240	0.370			-0.0364	0.0828	1.49	0.267	0.352	0.793	0.882	0.529
CESM2	0.282	0.278	0.240	0.333	0.188		0.0390	0.0726	1.21	0.377	0.498	0.831	0.637	0.578
CMCC-ESM2	0.251	0.249	0.240	0.285	0.199		0.00983	0.0625	1.22	0.378	0.493	0.830	0.841	0.607
CNRM-ESM2-1	0.271	0.267	0.240	0.316	0.180		0.0287	0.0573	1.21	0.365	0.482	0.842	0.984	0.631
EC-Earth3-CC	0.194	0.221	0.240	0.0717			-0.0201	0.0689	1.05	0.322	0.475	0.862	0.844	0.596
GFDL-ESM4	0.219	0.221	0.240	0.198			-0.0177	0.0516	1.37	0.376	0.562	0.807	0.981	0.658
GISS-E3-G	0.260	0.257	0.240	0.305	0.166		0.0198	0.0643	1.27	0.356	0.466	0.823	0.907	0.604
MPI-ESM1-2-LR	0.181	0.183	0.239	0.156	0.245		-0.0562	0.0679	1.35	0.277	0.576	0.811	0.978	0.644
MRI-ESM2-0	0.243	0.240	0.240	0.279	0.220		0.000579	0.0567	1.89	0.356	0.496	0.711	0.936	0.599
NorESM2-LM	0.289	0.287	0.240	0.301			0.0483	0.0829	1.26	0.345	0.476	0.825	0.631	0.550
SAM0-UNICON	0.229	0.215	0.240	0.405	0.181		-0.0236	0.0541	1.23	0.375	0.554	0.825	0.949	0.651
TaiESM1	0.218	0.218	0.240	0.212			-0.0216	0.0534	1.20	0.383	0.562	0.833	0.957	0.659
UKESM1-0-LL	0.207	0.207	0.240	0.200	0.212		-0.0314	0.0745	1.46	0.275	0.466	0.795	0.841	0.569

Figure A1: Evaluation of global simulations of SM using the mrs_{01} variable to 10 cm, compared to the Wang et al. (2021) dataset at the same depth. The various metrics shown are the different benchmark scores used in the ILAMB evaluation, including Bias Score, RMSE Score, Seasonal Cycle Score, and Spatial Distribution Score. These scores contribute to the final benchmark, represented here as the Overall Score, and are the results that are presented in Figure 3 (and Figures S5 and S7-9).

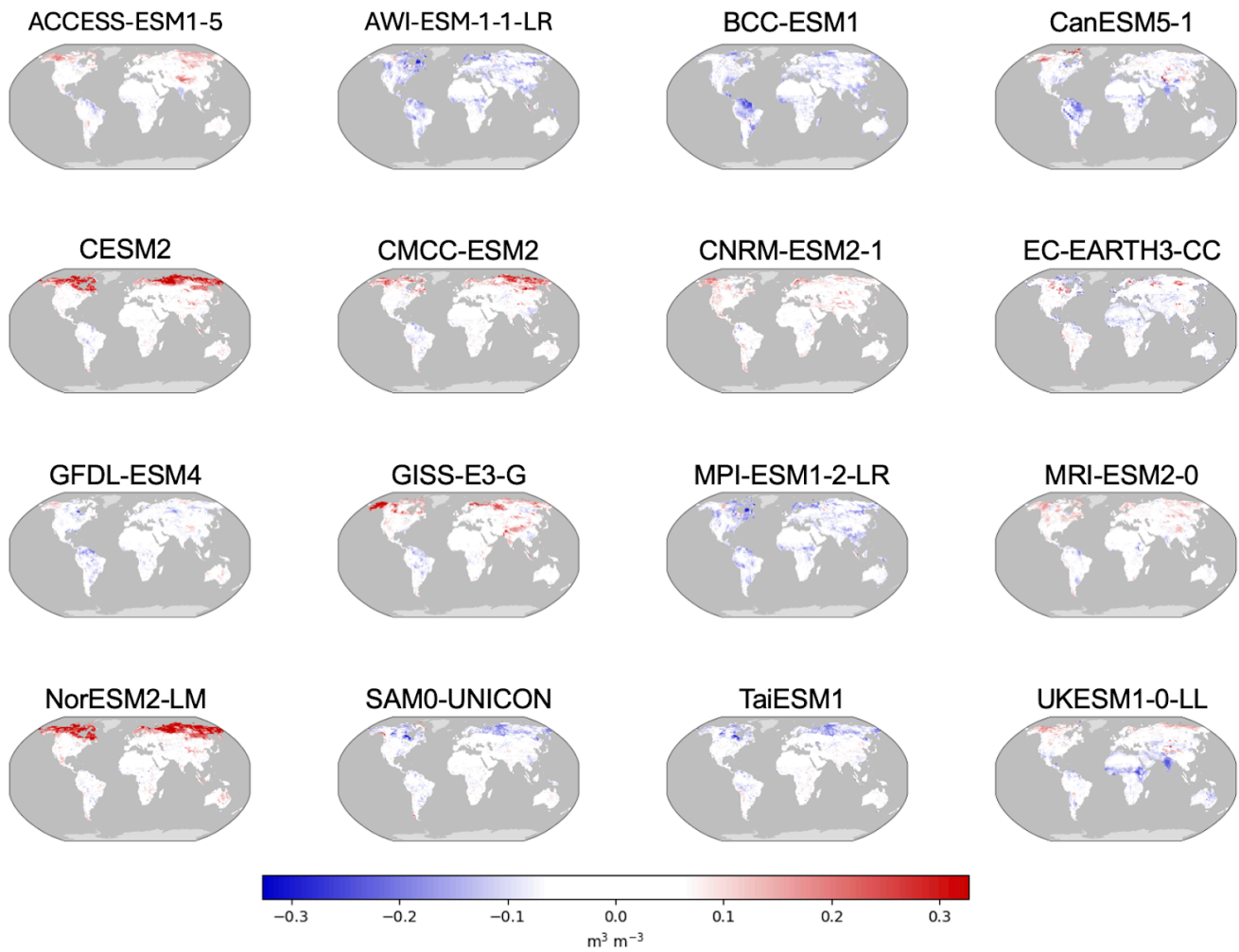
925

	ACCESS-ESM1-5	AWI-ESM-1-1-LR	BCC-ESM1	CanESM5-1	CESM2	CMCC-ESM2	CNRM-ESM2-1	EC-Earth3-CC	GFDL-ESM4	GISS-E3-G	MPI-ESM1-2-LR	MRI-ESM2-0	NotESM2-LM	SAM0-UNICON	TaiESM1	UKESM1-0-LL
[-] Soil Moisture	-0.25	0.54	-0.33	-2.31	-0.52	0.79	0.40	-0.78	0.84	-0.76	0.52	-0.22	-1.10	1.54	1.83	-0.20
[+] Near Surface	-0.18	0.93	-1.35	-2.52	-0.54	0.48	0.58	0.82	1.11	0.13	1.04	0.12	-1.15	0.41	0.95	-0.84
[+] Root Zone	-0.22	-0.03	0.74	-1.16	-0.28	0.74	0.07	-1.93	0.24	-1.25	-0.17	-0.44	-0.60	1.93	1.88	0.47
[-] Relationship Variables	0.94	0.39	0.07	-0.62	0.57	-0.41	1.47	-0.37	0.65	0.68	0.21	-3.04	0.46	-1.02	-0.39	0.43
[+] Gross Primary Productivity	1.01	-0.02	0.54	-0.32	0.55	0.55	1.27	-0.28	0.92	-0.87	-0.58	-3.16	0.50	-0.08	0.22	-0.26
[+] Leaf Area Index	0.94	0.79	-0.11	0.14	0.05	-1.24	1.08	-0.78	-0.41	2.34	0.71	-1.66	0.09	-1.25	-0.81	0.10
[+] Evapotranspiration	-0.99	-0.47	-0.61	-1.99	0.80	0.36	0.05	1.16	1.44	-1.42	0.23	-0.43	0.42	-0.49	0.08	1.89
[-] Relationships	-0.25	-0.83	-2.02	-2.07	-0.26	0.85	1.16	0.50	1.20	0.15	-0.66	1.32	-0.23	-0.02	0.82	0.35
[+] Near Surface/ESACCI_5cm	-0.81	-1.20	-1.85	-0.95	1.58	0.79	1.39	-0.36	0.20	0.37	-1.05	0.94	1.45	-0.30	-0.27	0.04
[+] Near Surface/Wang2021OLC_mrsos	-0.47	-0.68	-2.17	-1.28	-1.07	0.63	1.21	0.36	1.36	0.79	-0.47	1.27	-0.89	0.57	0.57	0.26
[+] Near Surface/Wang2021OLC_mrsol	0.65	-0.93	-1.90	-1.61	-0.78	0.71	1.42	0.38	0.92	0.62	-0.69	1.49	-0.86	-0.31	0.64	0.26
[+] Root Zone/Wang2021OLC	-0.11	-0.53	-1.53	-2.54	-0.37	0.83	0.69	0.81	1.34	-0.38	-0.46	1.09	-0.29	-0.01	1.22	0.22

Relative Scale



930 Figure A2: Overall scores that provide the ILAMB results in Figure 3 are depicted numerically here.



935 | **Figure A3: Bias plots (in $\text{m}^3 \text{m}^{-3}$) comparing Wang2021OLC top 10 cm SM product to the *mrso1* variable up to 10 cm for all models. The bias is calculated as the difference between model simulations and observations in each grid cell, providing insight into simulated SM performance in various regions around the globe.**

Köppen Climate Regions

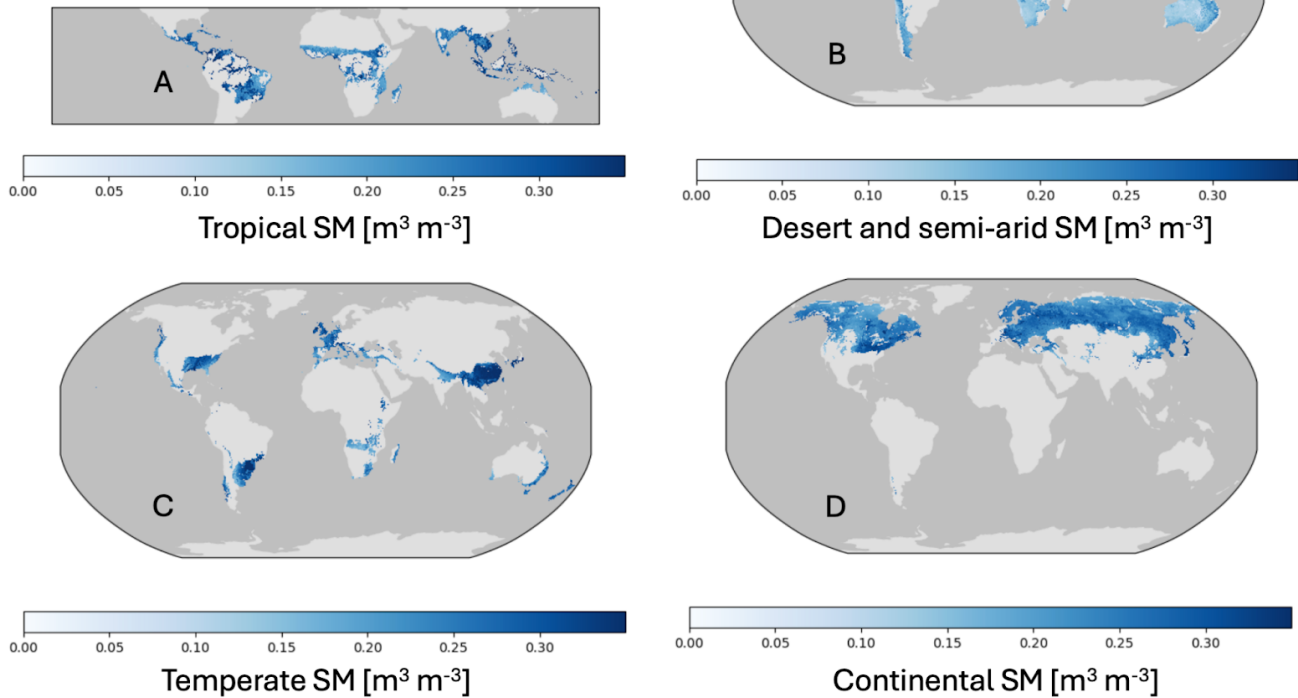


Figure A4: Köppen climate regions evaluated in this study through ILAMB. The default region is "global," while other regions include A) Tropical, B) Desert and Semi-arid, C) Temperate, and D) Continental. Regional mean values of surface SM derived from the ESA-CCI product are shown. The ESA-CCI dataset includes some data gaps, particularly in densely forested regions (e.g., the Amazon and Congo), ice-covered areas, and urban zones, due to limitations in microwave satellite observations. These gaps are more prevalent in earlier years and become less frequent over time. However, the primary purpose of this figure is to broadly illustrate the Köppen climate regions rather than serve as a detailed analysis of SM.

940

945

Tropical climates

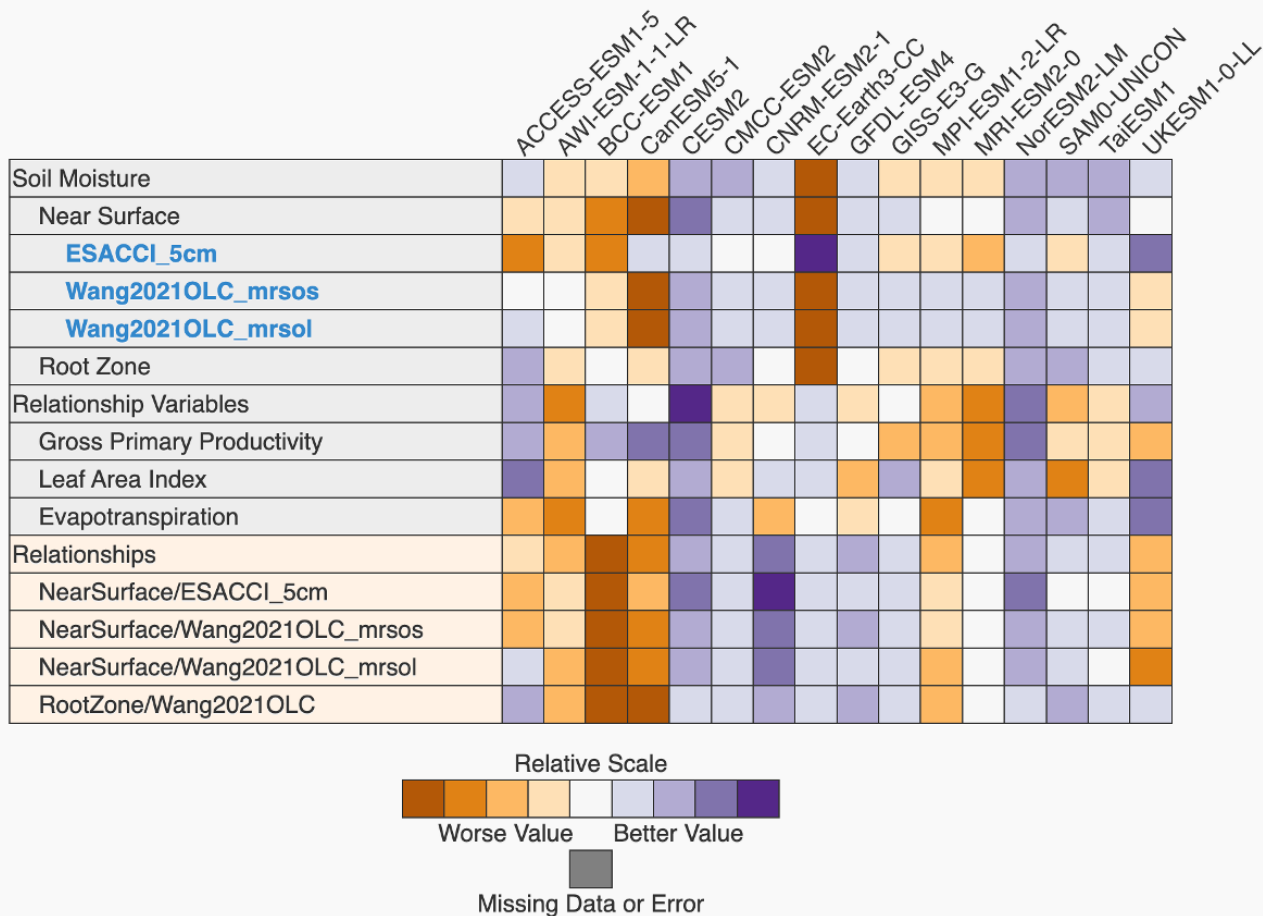
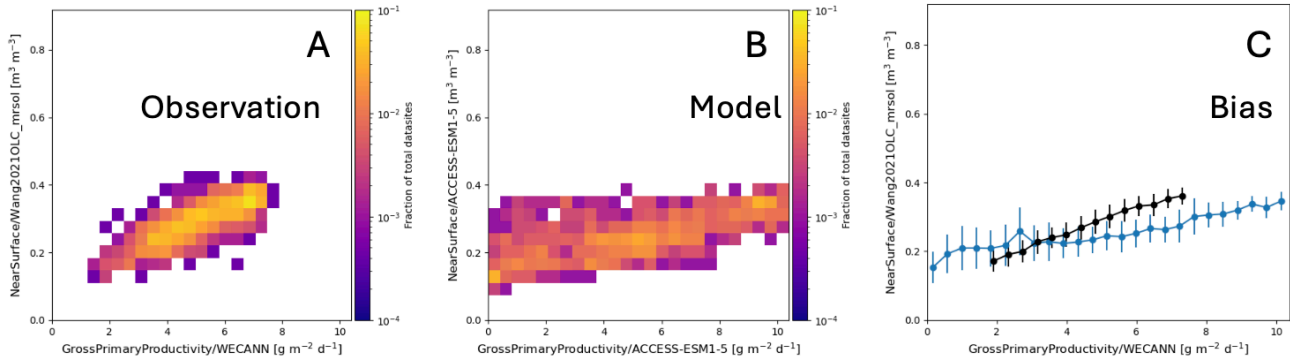
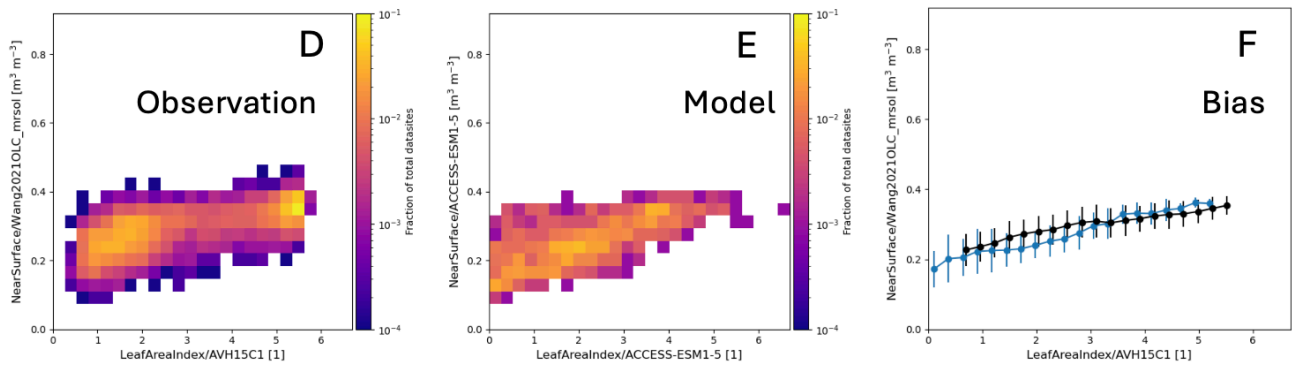


Figure A5: ILAMB results similar to Figure 3, but for the “Tropical” Köppen region.

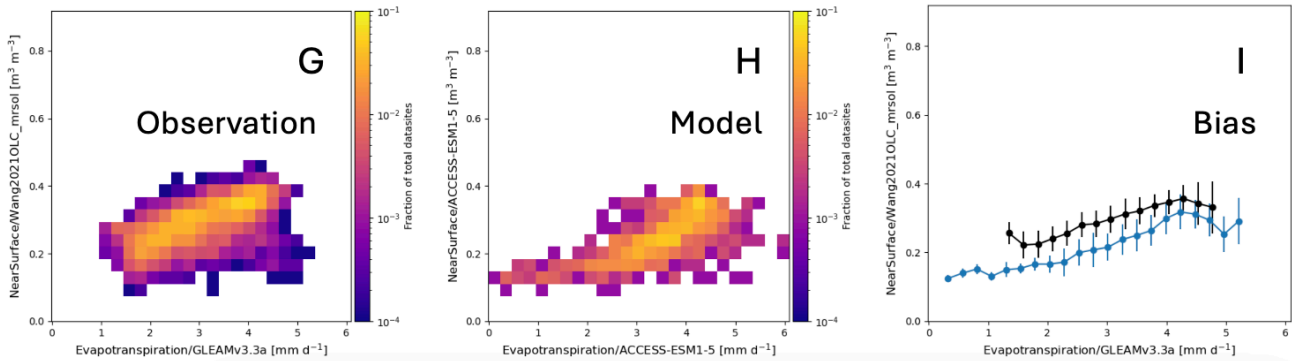
GPP vs Surface SM – Tropical Climate



LAI vs Surface SM – Tropical Climate



ET vs Surface SM – Tropical Climate



950 **Figure A6: Relationships between surface SM and ecohydrologic variables, similar to Figure 7, but for the “Tropical” Köppen region. The model results shown here are for the ACCESS-ESM1-5 model. Panels A, D, and G show the heat maps of the observed relationships, panels B, E, and H show the heat maps of simulated relationships in the ACCESS-ESM1-5 model, and panels C, F, and I portray both the observed and simulated relationships in the form of dot and whisker plots for easy comparison.**

955

Desert and semi-arid climates

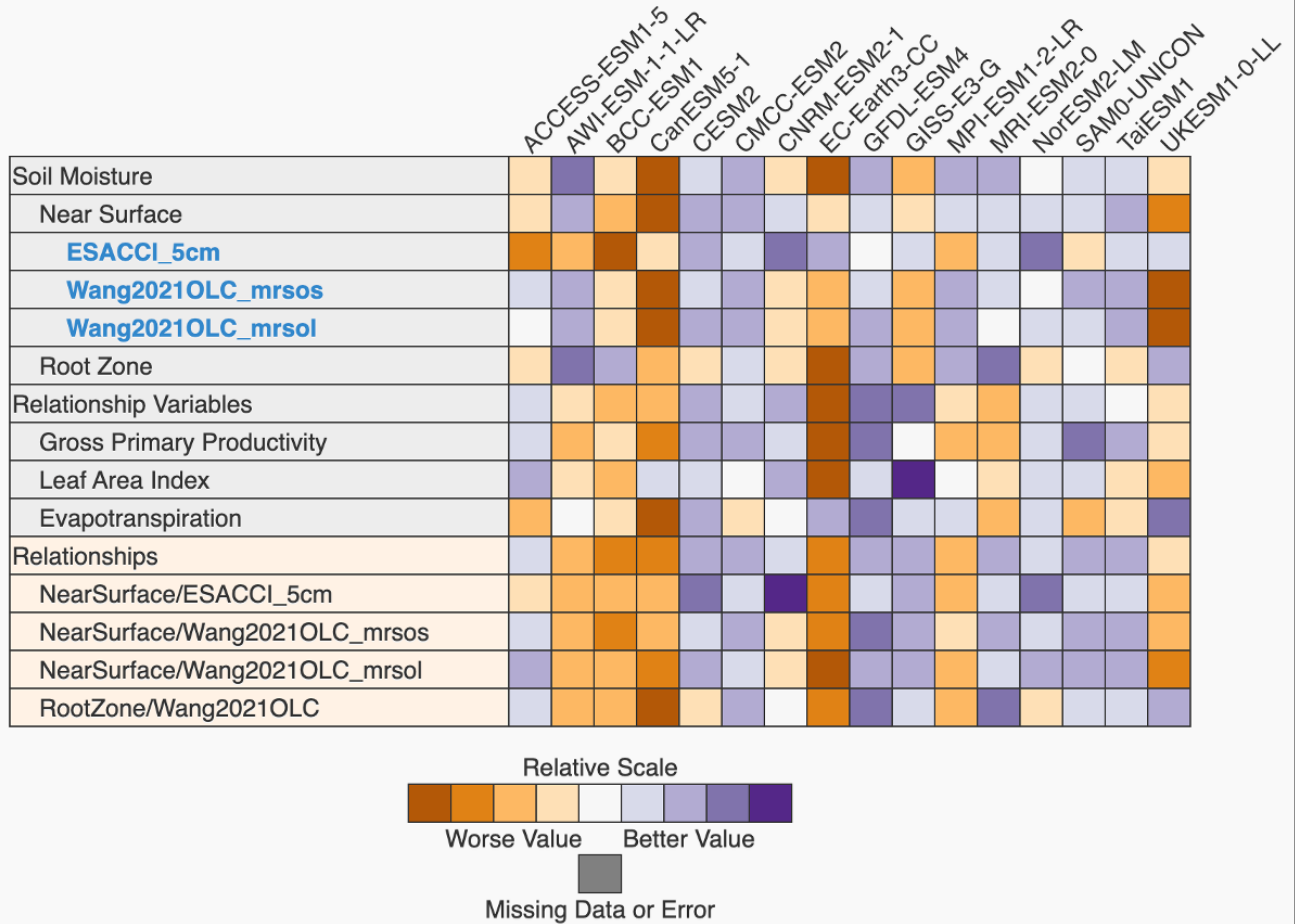
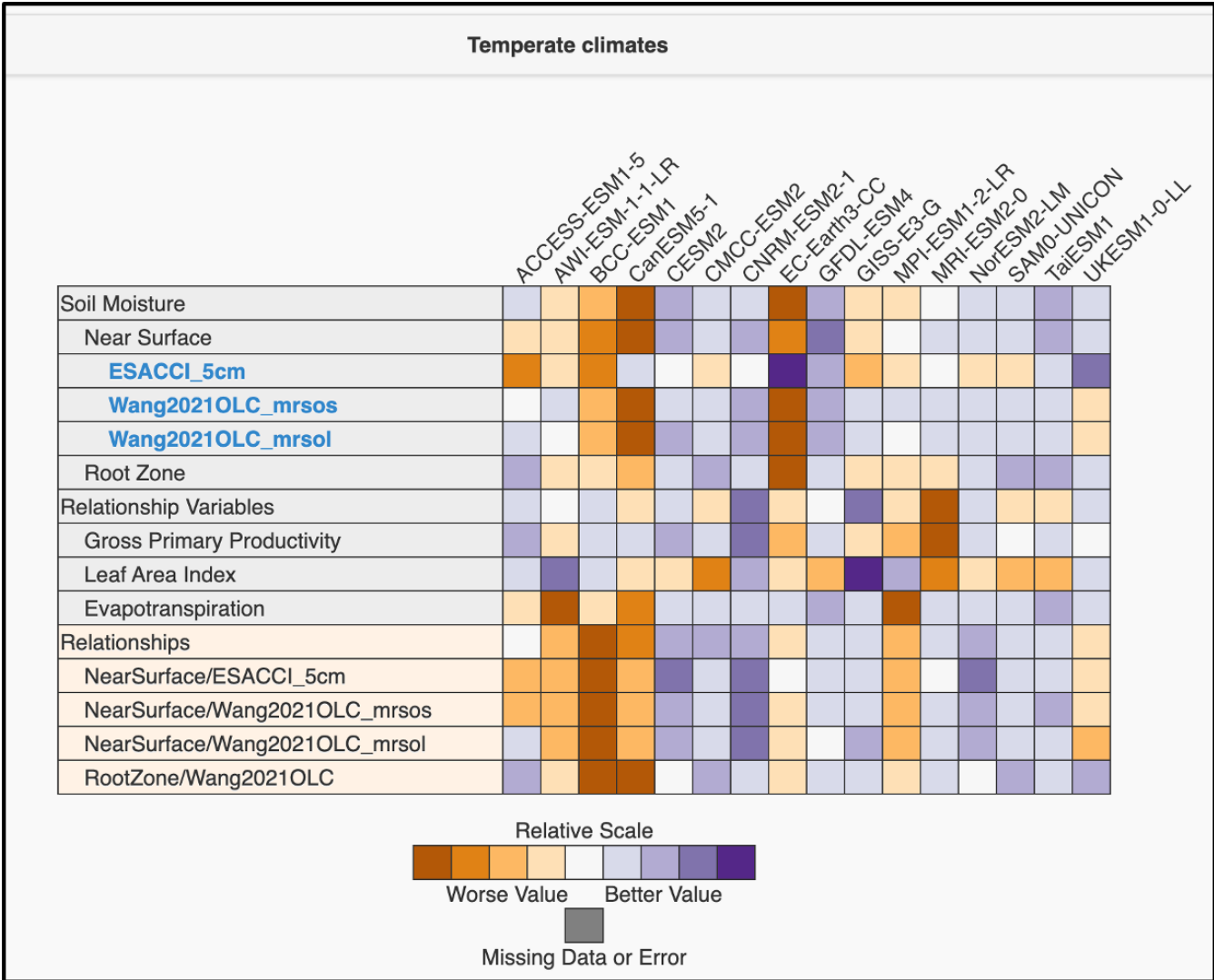


Figure A7: ILAMB results similar to Figure 3 and Figure A5, but for the “Desert and semi-arid” Köppen region.



960 Figure A8: ILAMB results similar to Figure 3 and Figure A5, but for the “Temperate” Köppen region.

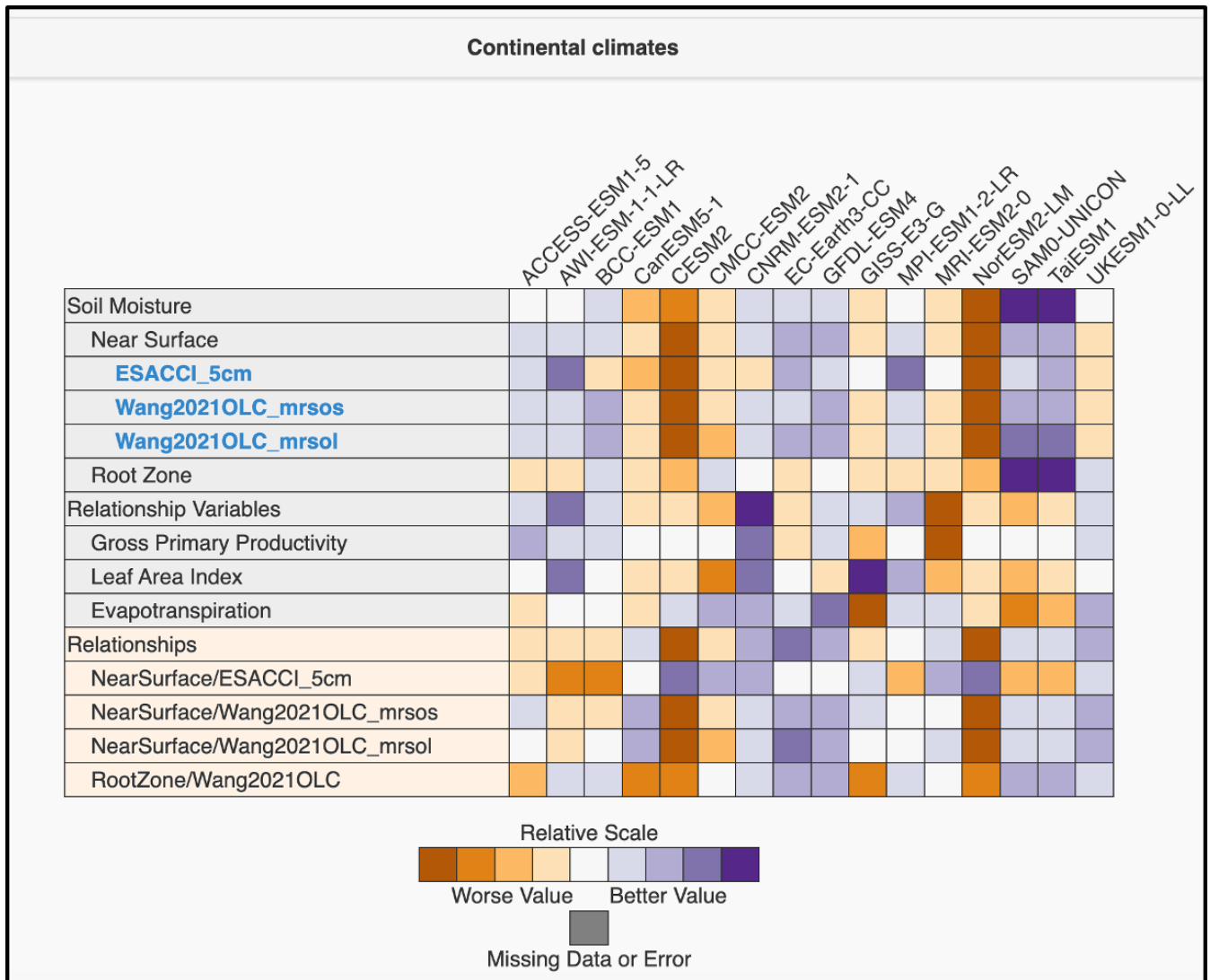


Figure A9: ILAMB results similar to Figure 3 and Figure A5, but for the “Continental” Köppen region.

Acknowledgements

The authors declare no conflicts of interest. This research was partially supported by the RUBISCO Science Focus Area (RUBISCO SFA KP1703), which is sponsored by the Regional and Global Model Analysis (RGMA) activity of the Earth and Environmental Systems Modeling (EESM) Program in the Earth and Environmental Systems Sciences Division (EESDD) of the Office of Biological and Environmental Research (BER) in the U.S. Department of Energy Office of Science. This work is the result of the RUBISCO project Soil Moisture Working Group effort on benchmarking soil moisture in Earth System Models, supported by the RUBISCO SFA. This paper has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the U.S. Department of Energy (DOE). P.R. acknowledges a grant from the Climate Program Office (CPO) of the National Oceanic and Atmospheric Administration (NOAA, NA22OAR4310612). The U.S. government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this paper, or allow others to do so, for U.S. government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan.

Author Contributions

This work arises from collaborative efforts and discussions within the RUBISCO Soil Moisture Working Group, which inspired the ideas behind this paper. E.C.M. performed the analysis, wrote the original draft, and led the editing process. N.C., Y.W., and J.M. helped with the analysis, contributed to discussions that shaped the paper, and assisted with editing. A.H., S.A.K., G.K., M.K., Pu.R., Pa.R., M.S., J.T., S.P.V., H.W., Q.Z., and F.M.H. contributed to the development of ideas through discussions and assisted with manuscript editing.

Code/Data Availability

The ILAMB software used in this study is publicly available at <https://github.com/rubisco-sfa/ILAMB> and archived with DOI: 10.18139/ILAMB.v002.00/1251621 (Collier et al., 2023). The specific assets related to the RUBISCO Soil Moisture Working Group benchmarking effort are available at <https://github.com/rubisco-sfa/soil-moisture>.

References

Alemohammad, S. H., Fang, B., Konings, A. G., Aires, F., Green, J. K., Kolassa, J., Miralles, D., Prigent, C., and Gentine, P. (2017). Water, Energy, and Carbon with Artificial Neural Networks (WECANN): a statistically based estimate of global surface turbulent fluxes and gross primary productivity using solar-induced fluorescence, *Biogeosciences*, 14, 4101–4124, <https://doi.org/10.5194/bg-14-4101-2017>.

- 1005 An, R., Zhang, L., Wang, Z., Quaye-Ballard, J.A., You, J., Shen, X., Gao, W., Huang, L., Zhao, Y. and Ke, Z., 2016. Validation of the ESA CCI soil moisture product in China. *International Journal of Applied Earth Observation and Geoinformation*, 48, pp.28-36. <https://doi.org/10.1016/j.jag.2015.09.009>
- 1010 [Best, M. J., Pryor, M., Clark, D. B., Rooney, G. G., Essery, R. L. H., Ménard, C. B., Edwards, J. M., Hendry, M. A., Porson, A., Gedney, N., Mercado, L. M., Sitch, S., Blyth, E., Boucher, O., Cox, P. M., Grimmond, C. S. B., and Harding, R. J.: The Joint UK Land Environment Simulator \(JULES\), model description – Part 1: Energy and water fluxes, *Geosci. Model Dev.*, 4, 677–699, <https://doi.org/10.5194/gmd-4-677-2011>, 2011.](https://doi.org/10.5194/gmd-4-677-2011)
- 1015 [Clark, D. B., Mercado, L. M., Sitch, S., Jones, C. D., Gedney, N., Best, M. J., Pryor, M., Rooney, G. G., Essery, R. L. H., Blyth, E., Boucher, O., Harding, R. J., Huntingford, C., and Cox, P. M.: The Joint UK Land Environment Simulator \(JULES\), model description – Part 2: Carbon fluxes and vegetation dynamics, *Geosci. Model Dev.*, 4, 701–722, <https://doi.org/10.5194/gmd-4-701-2011>, 2011.](https://doi.org/10.5194/gmd-4-701-2011)
- 1020 Clark, M. P., Fan, Y., Lawrence, D.M., Adam, J.C., Bolster, D., Gochis, D.J., Hooper, R.P. et al. (2015). Improving the representation of hydrologic processes in Earth System Models. *Water Resources Research*, 51, no. 8, 5929-5956. <https://doi.org/10.1002/2015WR017096>
- 1025 Caen, A., Smallman, T.L., de Castro, A.A., Robertson, E., von Randow, C., Cardoso, M. and Williams, M., 2022. Evaluating two land surface models for Brazil using a full carbon cycle benchmark with uncertainties. *Climate Resilience and Sustainability*, 1(1), p.e10. <https://doi.org/10.1002/cli.2.10>
- 1030 Claverie, M., Matthews, J., Vermote, E., and Justice, C. (2016), A 30+ Year AVHRR LAI and FAPAR Climate Data Record: Algorithm Description and Validation, *Remote Sensing*, 8, 263, <https://doi.org/10.3390/rs8030263>
- 1035 Ciabatta, L., Massari, C., Brocca, L., Gruber, A., Reimer, C., Hahn, S., Paulik, C., Dorigo, W., Kidd, R., and Wagner, W., 2018: SM2RAIN-CCI: a new global long-term rainfall data set derived from ESA CCI soil moisture, *Earth Syst. Sci. Data*, 10, 267–280, <https://doi.org/10.5194/essd-10-267-2018>.
- 1040 Collier, N., Hoffman, F.M., Lawrence, D.M., Keppel-Aleks, G., Koven, C.D., Riley, W.J., Mu, M. and Randerson, J.T., 2018. The International Land Model Benchmarking (ILAMB) system: design, theory, and implementation. *Journal of Advances in Modeling Earth Systems*, 10(11), pp.2731-2754. <https://doi.org/10.1029/2018MS001354>
- 1045 Collier, N., Hoffman, F. M., Mu, M., Randerson, J. T., Riley, W. J., Koven, C. D., Keppel-Aleks, G., & Lawrence, D. M. (2023). International Land Model Benchmarking (ILAMB) Package (v002.00) [Computer software]. RUBISCO Science Focus Area. <https://doi.org/10.18139/ILAMB.v002.00/1251621>.

- Dai, Y., Shangguan, W., Wei, N., Xin, Q., Yuan, H., Zhang, S., Liu, S., Lu, X., Wang, D., and Yan, F.: A review of the global soil property maps for Earth system models, *SOIL*, 5, 137–158, <https://doi.org/10.5194/soil-5-137-2019>, 2019.
- 1050 Dorigo, W. A., Gruber, A., De Jeu, R. A. M., Wagner, W., Stacke, T., Loew, A., ... & Kidd, R. (2015). Evaluation of the ESA CCI soil moisture product using ground-based observations. *Remote Sensing of Environment*, 162, 380-395. <https://doi.org/10.1016/j.rse.2014.07.023>
- 1055 Dorigo, W., Wagner, W., Albergel, C., Albrecht, F., Balsamo, G., Brocca, L., Chung, D., Ertl, M., Forkel, M., Gruber, A. and Haas, E., 2017. ESA CCI Soil Moisture for improved Earth system understanding: State-of-the art and future directions. *Remote Sensing of Environment*, 203, pp.185-215. <https://doi.org/10.1016/j.rse.2017.07.001>
- 1060 Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E.: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, *Geosci. Model Dev.*, 9, 1937–1958, <https://doi.org/10.5194/gmd-9-1937-2016>, 2016.
- 1065 Geiger, Rudolf (1954). "Klassifikation der Klimate nach W. Köppen" [Classification of climates after W. Köppen]. *Landolt-Börnstein – Zahlenwerte und Funktionen aus Physik, Chemie, Astronomie, Geophysik und Technik, alte Serie*. Vol. 3. Berlin: Springer. pp. 603–607.
- Green, J.K., Seneviratne, S.I., Berg, A.M. et al. Large influence of soil moisture on long-term terrestrial carbon uptake. *Nature* 565, 476–479 (2019). <https://doi.org/10.1038/s41586-018-0848-x>
- 1070 Gruber, A., Scanlon, T., van der Scalie, R., Wagner, W., Dorigo, W. (2019), Evolution of the ESA CCI Soil Moisture climate data records and their underlying merging methodology, *Earth System Science Data*, 11, 717-739, <https://doi.org/10.5194/essd-11-717-2019>.
- 1075 Guswa, A.J., Celia, M.A. and Rodriguez-Iturbe, I., 2002. Models of soil moisture dynamics in ecohydrology: A comparative study. *Water Resources Research*, 38(9), pp.5-1. <https://doi.org/10.1029/2001WR000826>
- 1080 Harpold, A. A., & Molotch, N. P. (2015). Sensitivity of soil water availability to changing snowmelt timing in the western US. *Geophysical Research Letters*, 42(19), 8011-8020. <https://doi.org/10.1002/2015GL065855>
- 1085 Hauser, M., Orth, R., and Seneviratne, S. I. Role of soil moisture versus recent climate change for the 2010 heat wave in western Russia. *Geophysical Research Letters* 43, no. 6 (2016): 2819-2826. <https://doi.org/10.1002/2016GL068036>
- Hengl, T., De Jesus, J.M., MacMillan, R.A., Batjes, N.H., Heuvelink, G.B., Ribeiro, E., Samuel-Rosa, A., Kempen, B., Leenaars, J.G., Walsh, M.G. and Gonzalez, M.R., 2014. SoilGrids1km—global soil

information based on automated mapping. *PloS one*, 9(8), p.e105992.
<https://doi.org/10.1371/journal.pone.0105992>

1090

Humphrey, V., Berg, A., Ciais, P. et al. Soil moisture–atmosphere feedback dominates land carbon uptake variability. *Nature*, 592, 65–69 (2021). <https://doi.org/10.1038/s41586-021-03325-5>

1095

[Kühnhammer, K., van Haren, J., Kübert, A., Bailey, K., Dubbert, M., Hu, J., ... & Beyer, M. \(2023\). Deep roots mitigate drought impacts on tropical trees despite limited quantitative contribution to transpiration. *Science of the Total Environment*, 893, 164763. <https://doi.org/10.1016/j.scitotenv.2023.164763>](https://doi.org/10.1016/j.scitotenv.2023.164763)

1100

Lawrence, D. M., Fisher, R. A., Koven, C. D., Oleson, K. W., Swenson, S. C., Bonan, G., ... & Zeng, X. (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(12), 4245–4287. <https://doi.org/10.1029/2018MS001583>

1105

Levine, P. A., Randerson, J. T., Swenson, S. C., and Lawrence, D. M., 2016. Evaluating the strength of the land–atmosphere moisture feedback in Earth system models using satellite observations, *Hydrol. Earth Syst. Sci.*, 20, 4837–4856, <https://doi.org/10.5194/hess-20-4837-2016>

1110

Li, L., Lin, X., Fang, Y. et al. A unified ensemble soil moisture dataset across the continental United States. *Sci Data*, 12, 546 (2025a). <https://doi.org/10.1038/s41597-025-04657-x>

Li, W., Migliavacca, M., Miralles, D.G., Reichstein, M., Anderegg, W.R., Yang, H. and Orth, R. Disentangling Effects of Vegetation Structure and Physiology on Land–Atmosphere Coupling. *Global Change Biology*, 31(1), p.e70035. (2025b). <https://doi.org/10.1111/gcb.70035>

1115

Lu, X., Du, Z., Huang, Y., Lawrence, D., Kluzek, E., Collier, N., Lombardozzi, D., Sobhani, N., Schuur, E.A. and Luo, Y., 2020. Full implementation of matrix approach to biogeochemistry module of CLM5. *Journal of Advances in Modeling Earth Systems*, 12(11), p.e2020MS002105. <https://doi.org/10.1029/2020MS002105>

1120

Martens, B., Miralles, D. G., Lievens, H., van der Schalie, R., de Jeu, R. A. M., Fernández-Prieto, D., Beck, H. E., Dorigo, W. A., and Verhoest, N. E. C., 2017: GLEAM v3: satellite-based land evaporation and root-zone soil moisture, *Geosci. Model Dev.*, 10, 1903–1925, <https://doi.org/10.5194/gmd-10-1903-2017>

1125

Massoud, E. C., Xu, C., Fisher, R. A., Knox, R. G., Walker, A. P., Serbin, S. P., Christoffersen, B. O., Holm, J. A., Kueppers, L. M., Ricciuto, D. M., Wei, L., Johnson, D. J., Chambers, J. Q., Koven, C. D., McDowell, N. G., and Vrugt, J. A.: Identification of key parameters controlling demographically structured vegetation dynamics in a land surface model: CLM4.5(FATES), *Geosci. Model Dev.*, 12, 4133–4164, <https://doi.org/10.5194/gmd-12-4133-2019>, 2019.

1130

Massoud, E., Turmon, M., Reager, J., Hobbs, J., Liu, Z. and David, C.H., 2020. Cascading dynamics of the hydrologic cycle in California explored through observations and model simulations. *Geosciences*, 10(2), p.71. <https://doi.org/10.3390/geosciences10020071>

1135

Massoud, E. C., Andrews, L., Reichle, R., Molod, A., Park, J., Ruehr, S., and Giroto, M.: Seasonal forecasting skill for the High Mountain Asia region in the Goddard Earth Observing System, *Earth Syst. Dynam.*, 14, 147–171, <https://doi.org/10.5194/esd-14-147-2023>, 2023.

1140

Massoud, E.C., Collier, N.O., M. Xu, M. Shi, Hoffman, F.M. 2025. Discrepancies in the representation of surface and layered soil moisture in Earth System Models. *Geophysical Research Letters*. Submitted. 2025.

1145

McNally, A., Shukla, S., Arsenault, K.R., Wang, S., Peters-Lidard, C.D. and Verdin, J.P., 2016. Evaluating ESA CCI soil moisture in East Africa. *International Journal of Applied Earth Observation and Geoinformation*, 48, pp.96-109. <https://doi.org/10.1016/j.jag.2016.01.001>

1150

Miralles, D. G., Holmes, T. R. H., De Jeu, R. A. M., Gash, J. H., Meesters, A. G. C. A., and Dolman, A. J.: Global land-surface evaporation estimated from satellite-based observations, *Hydrol. Earth Syst. Sci.*, 15, 453–469, <https://doi.org/10.5194/hess-15-453-2011>, 2011.

Nachtergaele, F., van Velthuisen, H., Verelst, L., Wiberg, D., Henry, M., Chiozza, F., Yigini, Y., Aksoy, E., Batjes, N., Boateng, E. and Fisher, G., 2023. Harmonized world soil database version 2.0. FAO.

1155

Oleson, K. W., Lawrence, D. M., Bonan, G. B., Fisher, R. A., Lawrence, P. J., & Muszala, S. P. (2013). Technical description of version 4.5 of the Community Land Model (CLM). Technical description of version 4.5 of the Community Land Model (CLM)(2013) NCAR/TN-503+ STR, 503. DOI: 10.5065/D6RR1W7M

1160

Preimesberger, W., Scanlon, T., Su, C. -H., Gruber, A. and Dorigo, W. (2021). Homogenization of Structural Breaks in the Global ESA CCI Soil Moisture Multisatellite Climate Data Record, in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 4, pp. 2845-2862, April 2021, doi: 10.1109/TGRS.2020.3012896.

1165

Preimesberger, W., Stradiotti, P., and Dorigo, W.: ESA CCI Soil Moisture GAPPILLED: An independent global gap-free satellite climate data record with uncertainty estimates, *Earth Syst. Sci. Data Discuss.* [preprint], <https://doi.org/10.5194/essd-2024-610>, in review, 2025.

1170

Purdy, A. J., Fisher, J. B., Goulden, M. L., Colliander, A., Halverson, G., Tu, K., & Famiglietti, J. S. (2018). SMAP soil moisture improves global evapotranspiration. *Remote Sensing of Environment*, 219, 1-14. <https://doi.org/10.1016/j.rse.2018.09.023>

- 1175 Qiao, L., Zuo, Z., and Xiao, D. "Evaluation of soil moisture in CMIP6 simulations." *Journal of Climate* 35, no. 2 (2022): 779-800. <https://doi.org/10.1175/JCLI-D-20-0827.1>
- 1180 [Reick, C. H., Gayler, V., Goll, D., Hagemann, S., Heidkamp, M., Nabel, J. E. M. S., Raddatz, T., Roeckner, E., Schnur, R., & Wilkenskield, S. \(2021\). JSBACH 3 – The land component of the MPI Earth System Model: Documentation of version 3.2 \(Berichte zur Erdsystemforschung, 240\). Max-Planck-Institut für Meteorologie. <https://doi.org/10.17617/2.3279802>](https://doi.org/10.17617/2.3279802)
- 1185 Robock, A., Schlosser, C.A., Vinnikov, K.Y., Speranskaya, N.A., Entin, J.K. and Qiu, S., 1998. Evaluation of the AMIP soil moisture simulations. *Global and Planetary Change*, 19(1-4), pp.181-208. [https://doi.org/10.1016/S0921-8181\(98\)00047-2](https://doi.org/10.1016/S0921-8181(98)00047-2)
- 1190 Seneviratne, S. I., Corti, T., Davin, E. L., Hirschi, M., Jaeger, E. B., Lehner, I., Orlowsky, B., and Teuling, A. J.. "Investigating soil moisture–climate interactions in a changing climate: A review." *Earth-Science Reviews* 99, no. 3-4 (2010): 125-161. <https://doi.org/10.1016/j.earscirev.2010.02.004>
- 1195 Shangguan, W., Dai, Y., Duan, Q., Liu, B. and Yuan, H., 2014. A global soil data set for earth system modeling. *Journal of Advances in Modeling Earth Systems*, 6(1), pp.249-263. <https://doi.org/10.1002/2013MS000293>
- 1200 Schumacher, D.L., Keune, J., Dirmeyer, P. et al. Drought self-propagation in drylands due to land–atmosphere feedbacks. *Nat. Geosci.* 15, 262–268 (2022). <https://doi.org/10.1038/s41561-022-00912-7>
- 1205 [Stocker, B. D., Tumber-Dávila, S. J., Konings, A. G., Anderson, M. C., Hain, C., & Jackson, R. B. \(2023\). Global patterns of water storage in the rooting zones of vegetation. *Nature Geoscience*, 16, 250–256. <https://doi.org/10.1038/s41561-023-01125-2>](https://doi.org/10.1038/s41561-023-01125-2)
- 1210 Talib, J., Müller, O.V., Barton, E.J. et al. The Representation of Soil Moisture-Atmosphere Feedbacks across the Tibetan Plateau in CMIP6. *Adv. Atmos. Sci.* 40, 2063–2081 (2023). <https://doi.org/10.1007/s00376-023-2296-2>
- 1215 Tapiador, F.J., Navarro, A., Jiménez, A., Moreno, R. and García-Ortega, E., 2018. Discrepancies with satellite observations in the spatial structure of global precipitation as derived from global climate models. *Quarterly Journal of the Royal Meteorological Society*, 144, pp.419-435. <https://doi.org/10.1002/qj.3289>
- 1220 Trugman, A. T., D. Medvigy, J. S. Mankin, and W. R. L. Anderegg. "Soil moisture stress as a major driver of carbon cycle uncertainty." *Geophysical Research Letters* 45, no. 13 (2018): 6495-6503. <https://doi.org/10.1029/2018GL078131>

- 1215 Vogel, M. M., Zscheischler, J., and Seneviratne, S. I.: Varying soil moisture–atmosphere feedbacks explain divergent temperature extremes and precipitation projections in central Europe, *Earth Syst. Dynam.*, 9, 1107–1125, <https://doi.org/10.5194/esd-9-1107-2018>, 2018.
- Wang, C., Fu, B., Zhang, L. and Xu, Z., 2019. Soil moisture–plant interactions: an ecohydrological review. *Journal of Soils and Sediments*, 19, pp.1-9. <https://doi.org/10.1007/s11368-018-2167-0>
- 1220 Wang, Y., Mao, J., Jin, M., Hoffman, F. M., Shi, X., Wullschleger, S. D., and Dai, Y.: Development of observation-based global multilayer soil moisture products for 1970 to 2016, *Earth Syst. Sci. Data*, 13, 4385–4405, <https://doi.org/10.5194/essd-13-4385-2021>, 2021.
- 1225 Wang, A., Kong, X., Chen, Y., & Ma, X. (2022). Evaluation of soil moisture in CMIP6 multi model simulations over conterminous China. *Journal of Geophysical Research: Atmospheres*, 127, e2022JD037072. <https://doi.org/10.1029/2022JD037072>
- 1230 Wu, G., Cai, X., Keenan, T.F., Li, S., Luo, X., Fisher, J.B., Cao, R., Li, F., Purdy, A.J., Zhao, W. and Sun, X., 2020. Evaluating three evapotranspiration estimates from model of different complexity over China using the ILAMB benchmarking system. *Journal of Hydrology*, 590, p.125553. <https://doi.org/10.1016/j.jhydrol.2020.125553>
- 1235 Yamaguchi, M., Lang, S.T., Leutbecher, M., Rodwell, M.J., Radnoti, G. and Bormann, N., 2016. Observation-based evaluation of ensemble reliability. *Quarterly Journal of the Royal Meteorological Society*, 142(694), pp.506-514. <https://doi.org/10.1002/qj.2675>
- 1240 Yang, X., Thornton, P., Ricciuto, D., Wang, Y., and Hoffman, F.: Global evaluation of terrestrial biogeochemistry in the Energy Exascale Earth System Model (E3SM) and the role of the phosphorus cycle in the historical terrestrial carbon balance, *Biogeosciences*, 20, 2813–2836, <https://doi.org/10.5194/bg-20-2813-2023>, 2023.
- 1245 Yuan, S., Quiring, S. M., & Leason, Z. T. (2021). Historical changes in surface soil moisture over the contiguous United States: An assessment of CMIP6. *Geophysical Research Letters*, 48(1), e2020GL089991. <https://doi.org/10.1029/2020GL089991>
- Zeng, J., Peng, J., Zhao, W., Ma, C., & Ma, H. (2023). Microwave remote sensing of soil moisture. *Remote Sensing*, 15(17), 4243. <https://doi.org/10.3390/rs15174243>
- 1250 Zuo, Z., Qiao, L., Zhang, R., Chen, D., Piao, S., Xiao, D., & Zhang, K. (2024). Importance of soil moisture conservation in mitigating climate change. *Science Bulletin*, 69(9), 1332-1341. <https://doi.org/10.1016/j.scib.2024.02.033>