Biogeosciences, 23, 263–282, 2026
https://doi.org/10.5194/bg-23-263-2026

Biogeosciences Open Access EGU

Land surface model underperformance tied to specific meteorological conditions

Jon Cranko Page[1,2,a], Martin G. De Kauwe[3], Andy J. Pitman[1,2], Isaac R. Towers[1,2,4], Gabriele Arduini[5], Martin J. Best[6], Craig R. Ferguson[7], Jürgen Knauer[8], Hyungjun Kim[9,10,11,12], David M. Lawrence[13], Tomoko Nitta[14], Keith W. Oleson[13], Catherine Ottlé[15], Anna Ukkola[1,2], Nicholas Vuichard[15], Xiaoni Wang-Faivre[15], and Gab Abramowitz[1,2]

[1]Climate Change Research Centre, University of New South Wales, Sydney, NSW 2052, Australia
[2]ARC Centre of Excellence in Climate Extremes, Sydney, NSW 2052, Australia
[3]School of Biological Sciences, University of Bristol, 24 Tyndall Avenue, Bristol, BS8 1TQ, UK
[4]Ecology and Evolution Research Centre, University of New South Wales, Sydney, NSW 2052, Australia
[5]European Centre for Medium-Range Weather Forecasts, Reading, UK
[6]Met Office, Exeter, UK
[7]Atmospheric Sciences Research Center, University at Albany, State University of New York, Albany, NY, USA
[8]School of Life Sciences, Faculty of Science, University of Technology Sydney, Ultimo, NSW 2007, Australia
[9]Moon Soul Graduate School of Future Strategy, Korea Advanced Institute of Science and Technology, Daejeon, Korea
[10]Department of Civil and Environmental Engineering, Korea Advanced Institute of Science and Technology, Daejeon, Korea
[11]Graduate School of Green Growth and Sustainability, Korea Advanced Institute of Science and Technology, Daejeon, Korea
[12]Graduate School of Data Science, Korea Advanced Institute of Science and Technology, Daejeon, Korea
[13]NSF National Center for Atmospheric Research, Boulder, CO, USA
[14]Faculty of Science and Engineering, Chuo University, Japan
[15]Laboratoire des Sciences du Climat et de l'Environnement, IPSL, CEA-CNRS-Université Paris-Saclay,
Orme des Merisiers, Gif-sur-Yvette, 91190, France
[a]current address: Water, Energy and Environmental Engineering Research Unit, Faculty of Technology,
University of Oulu, Oulu 90570, Finland

Correspondence: Jon Cranko Page (joncrankopage@gmail.com)

Received: 25 August 2025 – Discussion started: 2 September 2025
Revised: 9 December 2025 – Accepted: 9 December 2025 – Published: 9 January 2026

Abstract. The exchange of carbon, water, and energy fluxes between the land and the atmosphere plays a vital role in shaping global change and extreme events. Yet our understanding of the theory of this surface-atmosphere exchange, represented via land surface models (LSMs), continues to be limited, highlighted by marked biases in model-data benchmarking exercises. Here, we leveraged the PLUMBER2 dataset of observations and model simulations of terrestrial sensible heat, latent heat, and net ecosystem exchange fluxes from 153 international eddy-covariance sites to identify the meteorological conditions under which land surface models are performing worse than independent benchmark expectations. By defining performance relative to three sophisticated out-of-sample empirical models, we generated a lower bound of performance in turbulent flux prediction that can be achieved with the input information available to the land surface models during testing at flux tower sites. We found that land surface model performance relative to empirical models is worse at edge conditions – that is, LSMs underperform in timesteps where the meteorological conditions consist of coinciding relative extreme values. Conversely, LSMs perform much better under "typical" conditions within the centre of the meteorological variable distributions. Constraining analysis to exclude the edge conditions results in the LSMs outperforming strong empirical benchmarks. Encouragingly, we show that refinement of the performance of land surface

Published by Copernicus Publications on behalf of the European Geosciences Union.

models in these edge conditions, consisting of only 12 %–31 % of all site-timesteps, would see large improvements (22 %–114 %) in an aggregated performance metric. Better performance in the edge conditions could see mean relative improvements in the aggregated metric of 77 % for the latent heat flux, 48 % for the sensible heat flux, and 36 % for the net ecosystem exchange on average across all LSMs and sites. Precise targeting of model development towards these meteorological edge conditions offers a fruitful avenue to focus model development, ensuring future improvements have the greatest impact.

1 Introduction

Our ability to predict future climate and its impact on the places where we live largely rests on our ability to model the land surface (Charney, 1975; Friedlingstein et al., 2014, 2025; Arora et al., 2020; Canadell et al., 2021). Turbulent fluxes of carbon, water, and energy link terrestrial processes with atmospheric dynamics and oceanic freshwater input. As the climate changes, these land-atmosphere interactions will be altered as the behaviour of both the land surface and the atmosphere is impacted (Cao et al., 2021; Walker et al., 2021). As such, it is necessary that our knowledge of terrestrial processes is robust and well-developed.

Land Surface Models (LSMs) simulate the surface carbon, water, and energy cycles and their interaction with the boundary layer (the lowest part of the atmosphere directly influenced by the land surface) and are important components of Earth System Models (ESMs) that are used to create future climate projections. Since LSMs integrate our current knowledge of terrestrial processes, they are an ideal testbed for evaluating the extent and efficacy of our understanding. While all LSMs are based on fundamental theory and physical processes, implementations can vary substantially whether via equations, parametrisation, approach to approximations, or the number of processes represented, with multiple independent modelling teams producing their own unique LSMs (Fisher and Koven, 2020). In turn, the outputs from LSMs can also exhibit significant differences, resulting in a wide range of contemporary simulations and future projections of the terrestrial carbon, water, and energy cycles (Arora et al., 2020).

Analysing LSM performance can provide important feedback on our understanding of terrestrial processes. Analysis conducted against observations yield estimates of accuracy (Blyth et al., 2011). Comparing LSMs against each other, either in coupled simulations as part of ESMs as seen within progressive phases of the Coupled Model Intercomparison Projects (CMIP, Eyring et al., 2016) such as CMIP6 (Gier et al., 2024), stand-alone implementations such as the PILPS (Henderson-Sellers et al., 1996) and TRENDY projects (Sitch et al., 2015, 2024), or even the combination of coupled and offline simulations (e.g. LS3MIP, van den Hurk et al., 2016), provides a measure of model uncertainty that can advance our understanding of terrestrial processes. A relatively novel approach to LSM evaluation is that of benchmarking LSMs against *a priori* expectations of performance as used in the PLUMBER framework (Best et al., 2015; Haughton et al., 2018b). Model intercomparisons alone rarely guide model refinement as quantifying intermodel spread or demonstrating that one model is marginally superior does not provide insight into the processes or input conditions causing such performance discrepancies. Here we used data from the PLUMBER2 benchmarking experiment to answer three questions that might more constructively contribute to LSM development:

1. Under what specific meteorological input conditions does each land surface model consistently underperform in simulations of net ecosystem exchange, latent heat, and sensible heat fluxes?

2. Are the conditions where each LSM underperforms similar across different LSMs and fluxes?

3. How much do these conditions, and the subsequent underperformance when simulating fluxes, influence the overall performance of LSMs in benchmarking studies, such as PLUMBER2?

To answer the first question, it is necessary to define "underperformance". To do so, we turn to the PLUMBER benchmarking framework and its second iteration, PLUMBER2 (Best et al., 2015; Abramowitz et al., 2024). PLUMBER2 includes multiple LSM simulations using harmonised input data from a set of eddy-covariance sites (Ukkola et al., 2022). In addition, it includes a series of out-of-sample empirical flux models (EFMs) which also simulate the fluxes at the same sites to provide performance measures independent of the LSMs. By comparing LSM outputs to those of EFMs of increasing complexity, PLUMBER2 creates a range of potential performance expectations predicated on the information available to the LSMs – the EFMs provide lower bounds of flux prediction skill based on the input data (Haughton et al., 2018a) – and LSMs can then be assessed on their accuracy relative to these benchmarks (Abramowitz et al., 2024). As such, here we define "underperformance" as instances where the LSM does not achieve these empirically-derived lower bounds of predictive skill.

Next we must consider what is meant by "conditions". The PLUMBER2 dataset includes a wide variety of eddy-covariance sites – over 150 eddy-covariance sites situated in more than 20 countries and comprising of 11 vegetation classes (Ukkola et al., 2022). While this is a biased sampling of the global land surface (Alton, 2020; Chu et al., 2017, 2021; Griebel et al., 2020), the dataset does capture a broad range of ecosystems, climate regimes and weather conditions (van der Horst et al., 2019; Beringer et al., 2022). Therefore, the PLUMBER2 domain space is likely to be suitable for assessing LSM performance and identifying specific conditions where LSMs underperform. Identifying "poor performance" conditions based on plant functional type (PFT) classifications or individual site behaviour is unlikely to provide clear directions for improvement (Haughton et al., 2018a). PFTs are poor descriptors of flux-meteorology interactions, so grouping sites in this manner is likely to obscure the cause of performance issues (Cranko Page et al., 2024). Similarly, studies have repeatedly analysed LSM performance during meteorological conditions defined at longer timescales – such as the monthly scales used in the TRENDY project (Sitch et al., 2024) – with conditions such as droughts being poorly modelled by LSMs (Bastos et al., 2021; De Kauwe et al., 2015; Gu et al., 2016). However, most LSMs operate with approximately half-hourly timesteps meaning that any emergent biases ultimately originate from biases at this timescale, and so in this study, we consider meteorological conditions at this higher temporal resolution. Analysis at half-hourly resolution is unlikely to adversely penalise LSMs (Haughton et al., 2016) but could provide crucial information on conditions in which LSMs underperform and the processes responsible, including issues of process representation associated with diurnal cycles, fast processes such stomata functioning, or particular transient weather conditions (e.g. clouds).

To answer whether all LSMs reliably underperform in similar conditions we can again lean on the PLUMBER benchmarking framework. Multiple LSMs were run with the same information at the same eddy-covariance sites, facilitating comparison between LSMs in a robust manner (Abramowitz et al., 2024). If all LSMs routinely underperform in the same half-hourly meteorological conditions, this would imply there are processes at play that are either consistently missing or incorrectly represented across all models. However, if the conditions under which each LSM underperforms differ, then information on correct process representation and parametrisation might be derived from these differences.

When considering LSM performance, the initial PLUMBER2 analysis found that LSMs perform poorly against benchmarking EFMs (Abramowitz et al., 2024). In fact, LSM simulations of turbulent fluxes were often of worse quality than those from simple empirical models when the modelled fluxes were compared via an aggregated measure of seven independent performance metrics. This result may derive from a base level of LSM underperformance under all meteorological conditions – the LSMs may just consistently be slightly worse than the EFMs. However, potentially, this poor LSM performance against EFMs could be caused by the specific conditions in which LSMs underperform. If this is the case, we could identify the region of input space of meteorological conditions that is responsible for the poor LSM performance relative to the benchmarks in PLUMBER2 and help constrain areas for future model development.

We might hypothesise that the conditions of worst EFM performance will be in those areas where the EFMs may be lacking training data. Such conditions will be associated with meteorological extremes, which due to their nature are observed less frequently. Meanwhile, the EFMs should perform very well under "average" conditions – those meteorological conditions experienced the most across the available training data. As such, we hypothesise that the LSMs may underperform generally across the more temperate and hence "populated" meteorological conditions due to the better EFM performance under these conditions. However, since most LSM parametrisation is based on typical or good conditions conducive to normal ecosystem functioning, two possible contrasting hypotheses are posited regarding the extreme meteorological conditions. Firstly, that the lack of training data is a greater penalty to the EFMs and so LSMs have good performance at meteorological extremes relative to the EFMs. The alternative hypothesis is that the LSMs' lack of parametrisation at extremes is a bigger disadvantage than the lack of training data for EFMs, and hence the LSMs underperform in the extreme conditions.

A final hypothesis is that, under conditions where the observations used are erroneous but consistently biased (for example, conditions where the eddy-covariance method is biased due to violated assumptions such as during low wind speeds), the LSMs will underperform compared to the EFMs. This is because the eddy-covariance data is compromised and therefore the LSMs, being process-based, cannot model the biased observations, while the EFMs have no such constraints and can learn the biased behaviour under such conditions.

2 Methods

2.1 Data

2.1.1 Eddy-covariance Data

As part of the PLUMBER2 benchmarking framework, this study utilised the PLUMBER2 eddy-covariance dataset of observations (Ukkola et al., 2022). Of the 170 available sites, 153 were used in this study, as 17 sites were found to exhibit problematic or missing precipitation data (Abramowitz et al., 2024). For each site, observed fluxes of sensible heat (Qh), latent heat (Qle), and net ecosystem exchange (NEE) were available at a half-hourly timestep. Note that the PLUMBER2 dataset includes energy balance-corrected versions for each of these fluxes but these were not used in this study since Abramowitz et al. (2024) showed that using the energy balance-corrected data did not improve the overall performance of the LSMs. Locally-observed meteorology including downwelling surface shortwave radiation (SWdown), downwelling surface longwave radiation (LWdown), air temperature (Tair), vapour pressure deficit (VPD), specific humidity (Qair), relative humidity (RH), precipitation (Precip), surface air pressure (PSurf), CO_2 concentration (CO2air), and wind speed (Wind) was used as inputs to the models. The leaf area index (LAI) provided for each site in the PLUMBER2 dataset was also utilised. This LAI timeseries was derived from one of two remote sensing products, with this study using the preferred timeseries as detailed in Ukkola et al. (2022). In total, over 16 million individual site-timesteps were available in the dataset. Throughout this study, the three fluxes were analysed separately.

2.1.2 Models

The PLUMBER2 benchmarking framework involved 33 models, including seven empirical flux models and 26 process-based models. The details are specified in Abramowitz et al. (2024). To confine this study and ensure models were comparable, we here focussed on the 11 process-based models that are "Land Surface Models" or LSMs – that is, those models that are designed to be used in coupled climate modelling simulations. The models meeting this requirement are detailed in Table 1. Simulated NEE fluxes were not available for CLM5, JULES_GL9, JULES_GL9_LAI, and MATSIRO. In addition, ORCHIDEE2 and ORCHIDEE3 were lacking model outputs for all three fluxes at two and four sites, respectively. Additionally, this study used five of the PLUMBER2 EFMs in two different groups – the "benchmark" EFMs and the "best" EFMs. The "benchmark EFMs" were used to illustrate the full range of EFM performance when using models of increasing complexity and data needs. As such, they can be used to benchmark LSM performance aligning with their use in PLUMBER2. These were:

- the "1lin" model, a simple linear regression of flux against SWdown.
- the "3km27" model that used k-means clustering plus regression on three meteorological variables, namely SWdown, Tair and RH. 27 clusters were used to theoretically allow clusters where each meteorological variable is "low", "medium", and "high".
- the "LSTM" model, a long short-term memory model that was provided similar information to the LSMs, including static site parameters such as vegetation type and canopy height.

The "best EFMs" were, as the name suggests, the three EFMs that performed best according to the aggregated metrics for NEE, Qle, and Qh in the PLUMBER2 framework (Abramowitz et al., 2024). These EFMs were the most complex EFMs used in the framework and were:

- the "6km729lag" model, which used k-means clustering plus regression across 6 meteorological variables – SWdown, Tair, RH, Wind, Precip, and LWdown – as well as lagged Precip and Tair in the form of mean values over the prior 1–7, 8–30, and 31–90 d. Similar to the "3km27", the number of clusters was chosen such that each meteorological variable could be "low", "medium", or "high", resulting in 729 clusters.
- the "RF" model which used the Random Forest method with SWdown, LWdown, RH, Qair, surface air pressure (Psurf), Wind, CO_2 concentration (CO2air), VPD, and LAI as predictors.
- the "LSTM" model as above.

By using three EFMs to define the best EFM performance, we reduced the influence of individual model structure on our results while also limiting any instances of a single EFM performing anonymously well and affecting the assessment of LSM performance.

In all cases, the EFMs were run out-of-sample for each site. For all but the "LSTM" model, this involved training the models on the timesteps from all but a single site to then predict the remaining site out-of-sample. For the "LSTM" model, three randomly chosen sites were held back from each model and this was repeated until all sites had been simulated out-of-sample. As such, each EFM is the collective output from multiple models trained on all but one or three sites and then simulating the individual unseen sites. Similarly, for all but the "LSTM" model, each of the three fluxes used separate models. The "LSTM" model predicted NEE, Qle, and Qh in each model.

2.2 Analysis

2.2.1 Defining Poor LSM Performance

To analyse the conditions under which the LSMs could be expected to perform better, we compared LSM error to the errors of the best EFMs at each site-timestep. The absolute model error was calculated as:

$$\mathrm{Error}_{\mathrm{Mod}}(t) = |\mathrm{Flux}_{\mathrm{Mod}}(t) - \mathrm{Flux}_{\mathrm{Obs}}(t)| \tag{1}$$

for each LSM and the three best EFMs (step 2 in Fig. 1). We then defined an individual timestep as being a "LSM Loss" (LL) if and only if the absolute error of the LSM was greater than the absolute error of each of the three best EFMs:

$$\begin{aligned}\text{Timestep } t = \text{LSM Loss} \Longleftrightarrow &\mathrm{Error}_{\mathrm{LSM}}(t)\\ &> \max(\mathrm{Error}_{\mathrm{EFM}}(t))\end{aligned} \tag{2}$$

If a timestep was not an LSM Loss, then we defined it as an "LSM Win" such that each timestep had a binary classification of whether or not we could expect the LSM to perform better under the corresponding conditions (step 3, Fig. 1). This definition was chosen to favour the LSMs. If the absolute LSM error was smaller than the absolute error of any of

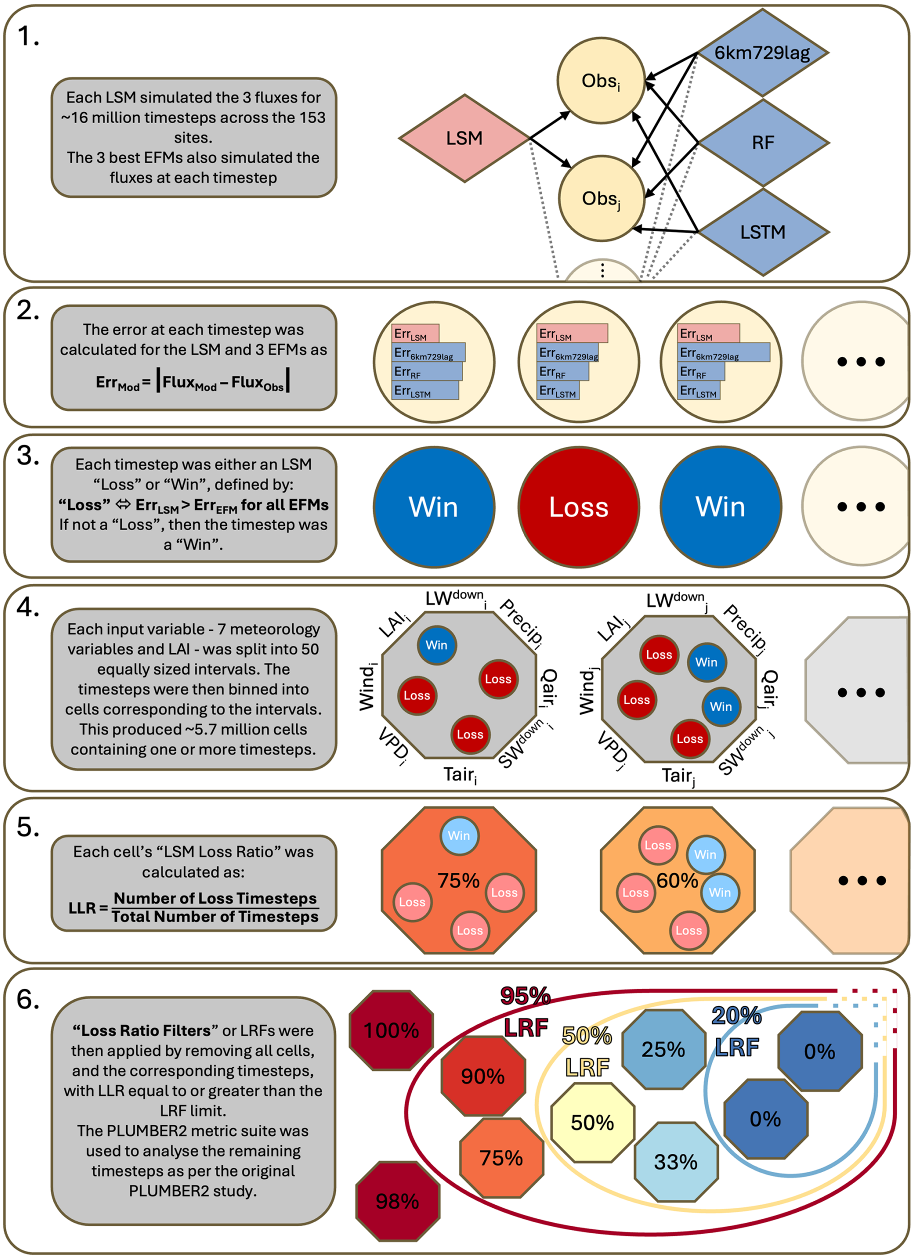
Table 1. Land Surface Models included in this study. "No NEE flux" here refers only to the necessary 30 min outputs required for this analysis – the model may produce NEE simulations at a larger timestep.

Model	LAI	Notes	References
CABLE	Prescribed	No C-N cycle	Kowalczyk et al. (2006), Wang et al. (2011)
CABLE-POP	Prescribed	Contains C-N cycle	Haverd et al. (2013, 2016, 2018)
CHTESSEL_1	Prescribed	–	Balsamo et al. (2009), Boussetta et al. (2013), Dutra et al. (2010), van den Hurk et al. (2000)
CLM5	Prescribed	No NEE flux	Lawrence et al. (2019)
GFDL	Computed	–	Dunne et al. (2020), Shevliakova et al. (2024)
JULES_GL9	Prescribed	No NEE flux, internal LAI values	Best et al. (2011), Clark et al. (2011)
JULES_GL9_LAI	Prescribed	No NEE flux, PLUMBER2 LAI values	Best et al. (2011), Clark et al. (2011)
MATSIRO	Prescribed	No NEE flux	MATSIRO6 Document Writing Team (2021)
NoahMP	Computed	–	Niu et al. (2011), He et al. (2023)
ORCHIDEE2	Computed	Two sites missing, no C-N cycle	Krinner et al. (2005)
ORCHIDEE3	Computed	Four sites missing, contains C-N cycle	Vuichard et al. (2019)

the best EFMs – that is, if the LSM outperformed just one of the three EFMs – we did not count this towards the definition of poor LSM performance. As previously mentioned, this reduced the influence of individual EFM model structure, meaning that poor LSM performance is more likely due to a general underutilisation of the information available in the input variables by the LSM. This in turn aimed to ensure that any areas of the input domain that exhibit poor performance under this definition are more likely to yield performance improvements if targeted for development. To illustrate the impact of such a definition of LSM Loss, consider the situation where an LSM and the best EFMs had equal performance across the entire domain of site-timesteps so that the differences between the four models were simply noise. We'd then expect a ratio of LSM Loss to LSM Win of 1 : 3. In other words, 25 % of timesteps would be an LSM Loss.

To explore the conditions under which the LSMs perform better or worse, we binned the input space based on eight input variables from the data measured at each site – SWdown, LWdown, Tair, Qair, VPD, Precip, Wind, and LAI. Each variable's range was split into fifty equally-spaced bins. With eight variables and fifty bins each, the input space was then split into 4×10^{13} unique cells across the 8-dimensional domain. Each cell was defined by the small sub-range of each of the eight input variables it contained. However, due to site distribution and the nature of meteorology, many of these cells were empty and some contained many more timesteps than others. For instance, cells corresponding to low SWdown values contained many timesteps due to roughly half of all the observational timesteps occurring at nighttime. Meanwhile, the PLUMBER2 dataset contains fewer sites that experience the extremes of temperature compared to temperate locations. Therefore, due to natural conditions and physically-determined interactions, the number of populated cells (i.e. cells with at least one timestep assigned to them) was far fewer than the total number of cells at slightly under six million (step 4, Fig. 1). The "LSM Loss Ratio" or LLR was then defined for each input cell as the percentage of timesteps in the cell that were an LSM Loss (step 5, Fig. 1). Hence, the LLR took a range of 0 % (the LSM outperformed at least one of the three best EFMs for every timestep in the cell) to 100 % (every timestep saw the LSM perform worse than all three of the best EFMs). The LLR provided a metric for LSM performance relative to the EFMs, and indicates the degree to which the LSMs are able to match the EFM's lower bound estimate of flux predictability for the small domain space contained within each cell. Cells with high LLR represent meteorological conditions under which the LSM(s) could be expected to perform better. In contrast, low LLR indicates meteorological conditions where our process understanding, LSM structure, and parametrisation are effective at utilising the available information to predict terrestrial fluxes. As noted above, if the LSM and best EFMs are equally capable, such that differences in their errors are random, then the LLR would be 25 %. Therefore, a LLR of 25 % or less indicates conditions where LSMs are reliably adding value.

We enabled visualisation of the results by collapsing the eight-dimensional input space into two-dimensional "fingerprints" via grouping the cells based on only two input variables at a time. We present plots of these fingerprints with Tair as one dimension and the other variables as the second dimension in turn. Tair was chosen as the constant x axis because it is intuitively easy to understand, has clear relationships with other meteorological variables, and its domain space is relatively uniformly represented by the site-timesteps, as opposed to the zero-dominated shortwave radiation for instance. Thus, we have two-dimensional fingerprints consisting of cells for which we calculated the LLR. We do not present the fingerprints for Qair because Qair and VPD are highly coupled.


1. Each LSM simulated the 3 fluxes for ~16 million timesteps across the 153 sites. The 3 best EFMs also simulated the fluxes at each timestep

2. The error at each timestep was calculated for the LSM and 3 EFMs as $\mathrm{Err}_{\mathrm{Mod}} = |\mathrm{Flux}_{\mathrm{Mod}} - \mathrm{Flux}_{\mathrm{Obs}}|$

3. Each timestep was either an LSM "Loss" or "Win", defined by: **"Loss" ⇔ $\mathrm{Err}_{\mathrm{LSM}} > \mathrm{Err}_{\mathrm{EFM}}$ for all EFMs** If not a "Loss", then the timestep was a "Win".

4. Each input variable - 7 meteorology variables and LAI - was split into 50 equally sized intervals. The timesteps were then binned into cells corresponding to the intervals. This produced ~5.7 million cells containing one or more timesteps.

5. Each cell's "LSM Loss Ratio" was calculated as: $\mathrm{LLR} = \frac{\text{Number of Loss Timesteps}}{\text{Total Number of Timesteps}}$

6. **"Loss Ratio Filters"** or LRFs were then applied by removing all cells, and the corresponding timesteps, with LLR equal to or greater than the LRF limit. The PLUMBER2 metric suite was used to analyse the remaining timesteps as per the original PLUMBER2 study.

Figure 1. Schematic of Analysis Process. Circles indicate a timestep and octagons indicate a cell containing all timesteps that fall within a specific interval for each meteorological variable. The ellipses indicate that many other timesteps and cells are included in the full analysis.

2.2.2 Quantifying Impacts of Poor LSM Performance

While high LLR implies that the LSM(s) are underperforming compared to the EFMs, it does not quantify the magnitude of this underperformance. To do this, we utilised the metric suite from PLUMBER2 (Abramowitz et al., 2024). This consists of seven independent metrics (here meaning that any timeseries can be modified such that any one metric changes while the others remain constant), namely Mean Bias Error, Standard Deviation Difference, Correlation Coefficient, Normalised Mean Error, 5th Percentile Difference, 95th Percentile Difference, and Density Overlap Percentage. These capture information about the averages, distributions, and extremes of the timeseries as well as temporal correlation. Note that these were calculated on a site-by-site basis. To summarise these seven metrics, we utilised the independent normalised metric value (iNMV), introduced by PLUMBER2 and again calculated on a site-by-site basis. For metrics where lower values are better, the metric $iNMV^m$ was defined as:

$$iNMV^m_{\mathrm{LSM}} = \frac{m_{\mathrm{LSM}} - \min(m_{\mathrm{benchmarkEFMs}})}{\max(m_{\mathrm{benchmarkEFMs}}) - \min(m_{\mathrm{benchmarkEFMs}})} \quad (3)$$

while for metrics where higher values are better, it was:

$$iNMV^m_{\text{LSM}} = \frac{\max(m_{\text{benchmarkEFMs}}) - m_{\text{LSM}}}{\max(m_{\text{benchmarkEFMs}}) - \min(m_{\text{benchmarkEFMs}})} \quad (4)$$

where m is the metric value, and the set of benchmark EFMs is as listed above. A lower $iNMV^m$ is better. The mean $iNMV^m$ across metrics and sites, iNMV, then provides a single number that synthesises LSM performance relative to EFMs. Since each of the seven metrics is weighted equally within this summary metric, the iNMV should adequately account for performance related to averages, extremes, and the distribution of the modelled timeseries as far as any aggregated metric can.

To assess the impact of poor LSM performance under the LSMs' worst meteorological conditions against other potential sources of poor LSM performance, we filtered the site-timesteps in five ways and compared the iNMV to the iNMV of the unfiltered data. Three of these filters were related to physical conditions, and two were "Loss Ratio Filters" (LRFs), where all timesteps that belong to input cells with a LLR above the threshold value were removed from the analysis, and performance impact of removing them was assessed (Fig. 1). Importantly, the LRFs were applied on the eight-dimensional input cells, not the two-dimensional fingerprints. The five filters were:

- A LRF of 95 %. All timesteps belonging to input cells with a LLR of 95 % or above were removed. This captured all conditions where the LSM was nearly always outperformed by the best EFMs.
- A LRF of 50 %. All timesteps belonging to input cells with a LLR of 50 % or above were removed. This provided an analysis that included only cells where the number of LSM wins outnumbered LSM losses. Due to the preference shown to the LSMs in the definition of an LSM Loss, this LRF significantly weighted the analysis towards conditions where the LSMs perform well compared to the best EFMs.
- A "Physical" consistency filter. This filter removed timesteps that consisted of physically impossible meteorological conditions. This involved calculating the saturated vapour pressure (SVP) at the observed air temperature and then removing any timesteps with a Qair or VPD value that violated this SVP. Such timesteps may occur due to observational error and could result in empirical models having an advantage where LSMs have strict constraints.
- A "Daytime" filter. All timesteps with SWdown less than or equal to $10\,\text{W}\,\text{m}^{-2}$ were removed. The assumptions underlying the eddy-covariance method can be violated at night-time (for instance, due to increased instances of insufficient turbulence), and hence night-time data can be heavily gap-filled (Aubinet et al., 2010, 2012; Pastorello et al., 2020).
- A "Windy" filter. All timesteps with Wind less than or equal to $2\,\text{m}\,\text{s}^{-1}$ were removed. This again was to account for potential measurements at times of insufficient turbulence. Wind speed was chosen over the friction velocity $u*$ to maintain consistency across the PLUMBER2 studies, with the wind speed threshold also being used in Abramowitz et al. (2024).

By applying these filters, we could gain an understanding of the areas of weak performance of LSMs, whether this was at night, being fed physically inconsistent inputs, or under particular meteorological conditions.

3 Results

Figure 2 shows the sensible heat flux LSM Loss Ratio (LLR) for the LSMs. Yellow indicates the null expectation of LSM performance (an LLR of around 25 %) where the LSM performed similarly to the best EFMs. This progresses to oranges and reds as LSM performance decreased and LLR increased. In contrast, LLRs of less than 20 % are coloured blue-green and indicate conditions where the LSMs were providing additional performance over the EFMs. The first column is an aggregated measure of LLR across all the LSMs using the root mean square of the individual LLRs. The Timesteps column shows the density distribution of the full dataset. Interestingly, across all potential input variable pairings with Tair, the largest LLRs for the root mean square (RMS) of all LSMs were seen in the edge cases – that is, located around the edge of the fingerprints. High LLRs occurred over the widest range of conditions when considering the LWdown-Tair interaction, with VPD-Tair and Wind-Tair also having more substantial areas of low LSM performance. There were no large discrepancies between the individual LSMs at LLRs above 80 % (light and dark red), with the "fingerprints" exhibiting similar patterns for poor LSM performance across models, and this is reflected too in the RMS fingerprints in the first column. At a LLR of 50 % or more (i.e. the LSM performing poorly at least half the time), differences between LSMs were clearer. For the VPD-Tair interaction, CABLE-POP had a high LLR across all conditions, while CABLE and both JULES implementations did worse at high VPD for the corresponding Tair (high LLR in the top right of the VPD-Tair domain). GFDL exhibited an inverse result with worse performance at low VPD-Tair values. The ORCHIDEE LSMs had a LLR of 50 % or more at VPD values with high but not extreme temperatures as seen from the dark orange streak to the right of the fingerprint, with surrounding lighter areas at higher temperatures. CHTESSEL_1, CLM5, and NoahMP had poor performance only at the boundaries of the Tair-VPD input space.

Figure 3 shows the Tair interaction LLR fingerprints for latent heat flux. While Qle had lower LLR values than Qh in general, it is still the case that the majority of high LLRs occurred in the edges of the fingerprints. However, in com-

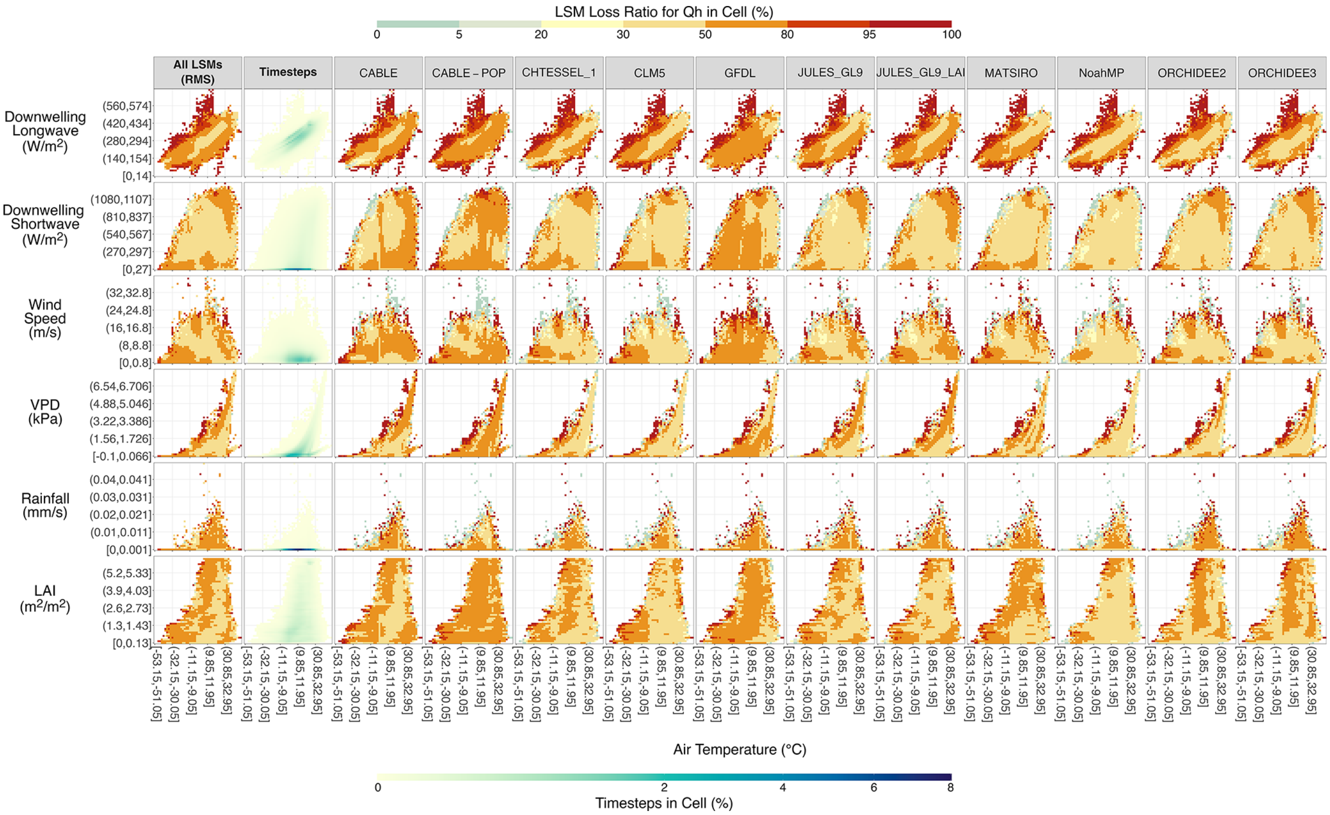


Figure 2. LSM Loss Ratio by Driving Variable and Model for Predicting Sensible Heat Flux. The x axis of each plot is air temperature and the y axis the other variables, each split into 50 equal-sized bins of which only every 10th bin is labeled. For the columns of individual models, each cell is coloured by the LSM Loss Ratio, the percentage of timesteps within the corresponding 2-D variable cell that are classified as an LSM Loss for the Qh flux. An LSM Loss of 0 occurs where, for every timestep within the cell, the LSM's absolute flux error is smaller than at least one of the absolute flux errors of the three best EFMs. Conversely, an LSM Loss of 100 occurs when the LSM's absolute flux error is greater than all three absolute flux errors of the best EFMs for all timesteps in the cell. For the timesteps column, the colour indicates the percentage of the total number of timesteps that fell within the 2-D variable cell. The "All LSMs" column is coloured by the root mean square of the LSM Loss across all models. Note that since all models simulated the same site-timesteps, the cells and timesteps within each cell are approximately the same but not equal for every model (some LSMs did not submit simulations for every site, missing four sites at most).

parison to Qh (Fig. 2), it is not the case that most edge cells had poor LLR. In fact, many edge condition cells had LLR as low as 20 % or less. This indicates that, while LSMs still had areas of poor performance under edge conditions, there were also edge conditions under which Qle was well simulated relative to the EFMs. Hence, there were edge conditions where the LSMs have poor performance for Qh but good performance for Qle (for instance, compare the LWdown – Tair fingerprints in Figs. 2 and 3 where dark red areas in Fig. 2 are blue in Fig. 3). In turn, there were also edge conditions where both fluxes showed poor performance compared to the EFMs, such as can be seen in the VPD – Tair fingerprints. Particularly high LLR and poor LSM performance for Qle was seen at high VPD values relative to the concurrent Tair observation. Generally, as seen in the RMS of LLR for all LSMs and across most models independently, poor LSM performance for Qle did not appear to be dependent on LAI values, with the edge LLR rarely exceeding even 80 %, and the rest of the fingerprints' areas at 50 % or below.

At low Tair (−30 °C or less), LLR was sub-20 % irrespective of input interactions across all LSMs, seen from the blue regions in columns 2 onwards in Fig. 3. For Qle, GFDL stands out as having the most potential for improvement in flux simulations. Extensive high LLR areas exist for all input condition interactions. GFDL has a LLR of around 20 %–50 % at high but not extreme temperatures indicating capacity for satisfactory simulation, yet performance began to decline around 10 °C. Other behaviours of note are the ORCHIDEE models having a similar drop in performance around a Tair value of 10 °C, although here the effect was moderated by interaction with other input conditions. The performance decline was seen only at the high/low LWdown boundaries, for SWdown values above ~ 300 W m^{-2}, and for high VPD values. MATSIRO had high LLR (and hence poor

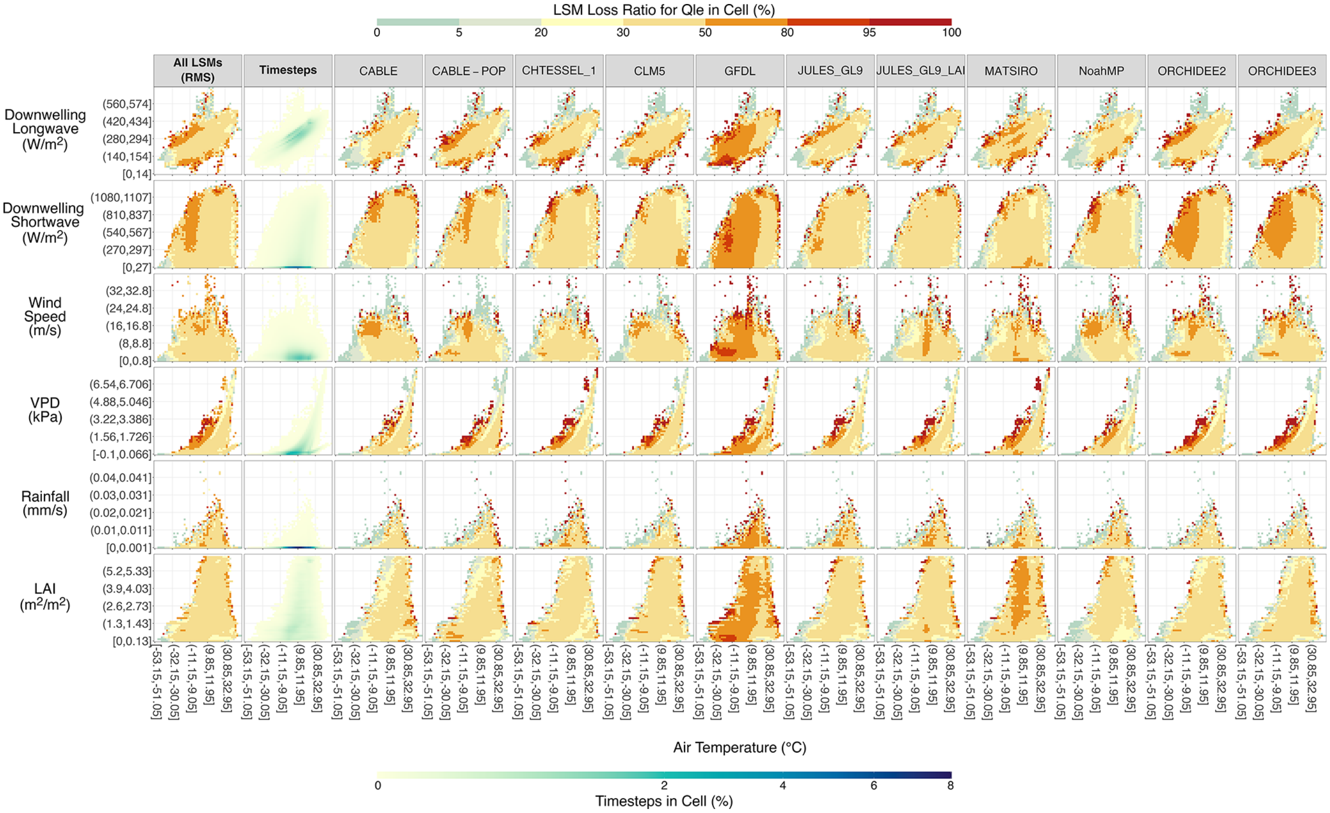


Figure 3. LSM Loss Ratio by Driving Variable and Model for Predicting Latent Heat Flux. The details are the same as Fig. 2 but for predictions of the Qle flux.

relative performance) at Tair values around 0 °C when LAI is above ∼ 1 $m^2\,m^{-2}$ but performed well at higher and lower temperatures.

Figure 4 completes the fingerprinting with the results for NEE. The LSM performance under edge conditions was much improved for NEE compared to the energy fluxes. In fact, there are substantial areas in Fig. 4 where the LSMs were providing additional value above the EFM simulations, often around the edge of the fingerprints. High LLRs still occurred at the boundaries of the Tair-VPD and Tair-Wind fingerprints yet were interspersed with cells of good LSM performance. However, inter-LSM differences were conspicuous for the carbon flux. NoahMP struggled against the EFMs relative to the rest of the LSMs: there was a pronounced temperature effect with a gradient of generally increasing LLR as Tair values increased. This dominant role of Tair appeared, but with a more subtle impact, for the other LSMs to varying degrees with greater regions of worse LLR (orange and red tones) towards higher temperatures and good performance (blue tones) towards lower temperatures. CABLE and CHTESSEL_1 were the other two LSMs that exhibited LLRs of 50 % or more at higher temperatures, while the two ORCHIDEE models in particular have only small regions of high LLR.

Figure 5 shows the independent normalised metric values (iNMV) for each LSM defined by the range of metric values of the three benchmark EFMs. This is a measure of LSM performance based on the benchmark EFM performance, directly comparable across LSMs (note the differing *y* axes). The grey lines indicate the 3 PLUMBER2 benchmark EFMs – 1lin, 3km27, and LSTM – with different symbols to differentiate between the EFMs and darker greys as the complexity increases. The black line is the base LSM performance. The blue lines are the LSM results when filters were applied, either using only daytime timesteps (dark blue), timesteps with sufficient wind (mid blue), or timesteps that did not violate physical humidity limits (light blue). The orange (95 %) and red lines (50 %) are the LSM results when timesteps are filtered based on the LSM Loss Ratio. In all cases, iNMV was calculated using the benchmark metrics as calculated on the original dataset i.e. no filters were ever applied to the benchmark models – their iNMV values are always calculated on the entire dataset. The black lines indicate the raw LSM performance as in the PLUMBER2 benchmarking framework and are therefore almost exactly the same result as Abramowitz et al. (2024) (there are slight differences due to using one less site here). As per Abramowitz et al. (2024), LSMs mostly performed worse than the benchmark EFMs, with all 11 LSMs performing worse than a simple lin-

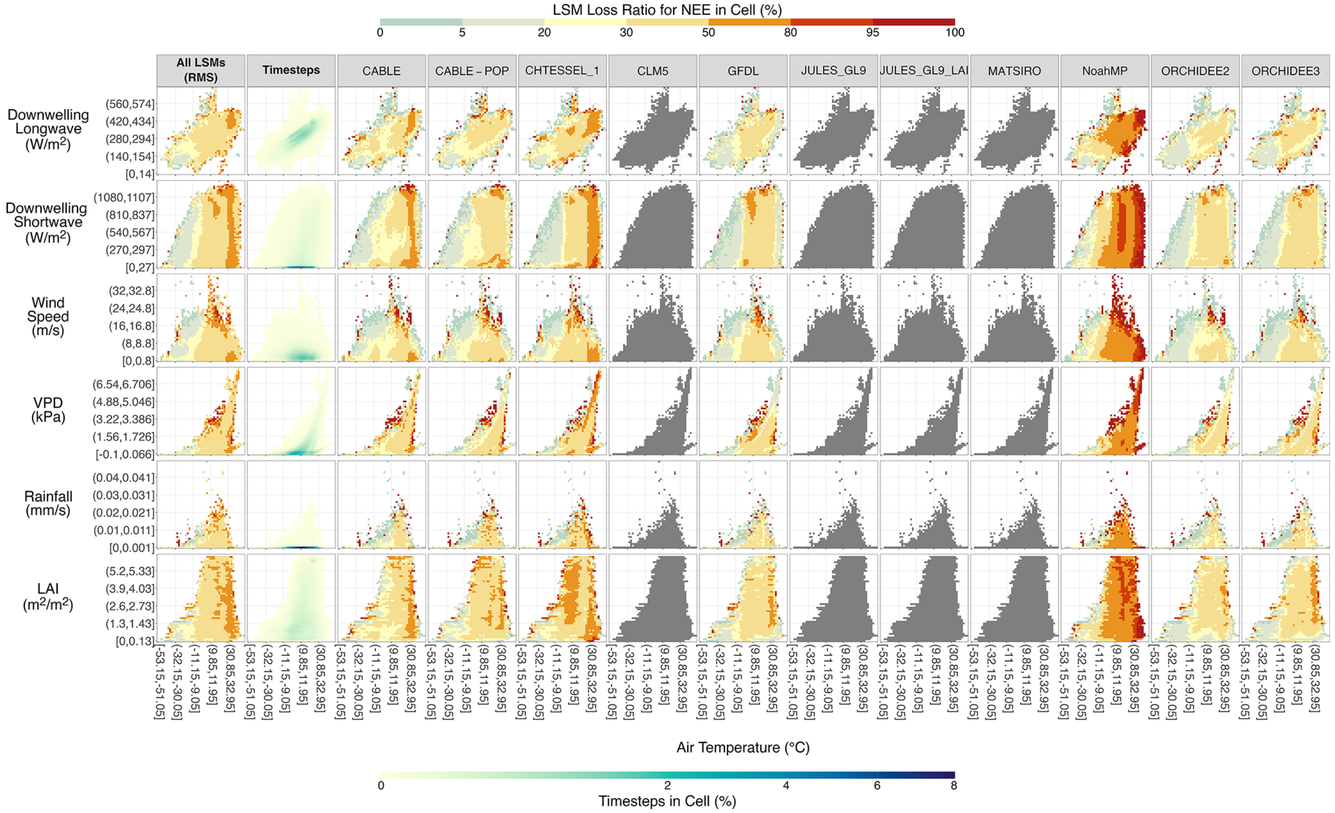


Figure 4. LSM Loss Ratio by Driving Variable and Model for Predicting Net Ecosystem Exchange. The details are the same as Fig. 2 but for predictions of the NEE flux. Note that greyed out LSMs did not provide half-hourly NEE outputs to the PLUMBER2 experiment.

ear model for simulating Qh. Better performance was seen in the other two fluxes, with most LSMs' performance falling between the linear model and the "3km27" k-means clustering model for Qle. For NEE, three of the seven LSMs in this analysis beat the linear model performance.

The iNMVs for two different LRFs are also shown in Fig. 5. The first, a LRF of 95 %, indicates the LSM performance when timesteps in poorly modelled input cells were removed from the analysis. These are the darker cells from Figs. 2–4. The percentage of timesteps removed by this filter varied from 12 % (CABLE and ORCHIDEE2 for NEE, CABLE-POP and NoahMP for Qle) to 31 % (also NoahMP but for NEE). The mean percentage of timesteps removed by the 95 % LRF was 17 % for Qh, 14 % for Qle, and 16 % for NEE. The 95 % LRF substantially improved iNMV for many of the models for all three fluxes (indicated by the difference between the black "None" and orange "95" points). The mean relative improvement in iNMV across all fluxes and models was 56 % (Table S1 in the Supplement). Noteworthy cases include many models (CABLE, CHTESSEL_1, CLM5, JULES_GL9, and the ORCHIDEE LSMs) improving from beating only the "1lin" model to being better than even the "LSTM" when simulating Qle (the mean relative improvement of iNMV for Qle under the 95 % LRF was 77 %). The Qh simulations from CABLE-POP, CLM5, MATSIRO, and NoahMP improved from a drastically worse performance than any of the EFM benchmarks to at least beating the "1lin" model (iNMV improvements of 69 %, 58 %, 68 %, and 57 % respectively). The impact of the 95 % LRF was less substantial for the NEE flux with smaller improvements across the LSM suite (mean iNMV relative improvement of 36 %).

The 50 % LRF was a drastic filter application. Any timesteps located in input cells with a LLR of 50 % or higher were removed from the metric calculations. This means that theoretically, within any single removed input cell, there were still up to 50 % of the removed timesteps that were instances where the LSMs actually beat (at least one of) the best EFMs. The largest improvement here was seen in the Qh flux simulations, where all LSMs but CHTESSEL_1 and GFDL beat the "LSTM" model under the 50 % LRF (mean relative iNMV improvement of 97 %, Table S2 in the Supplement). Differences between the 50 % and 95 % LRFs were minimal for Qle and NEE. The exception was Qle from GFDL where at the 95 % LRF, GFDL was still outperformed by the "1lin" EFM but at the 50 % LRF, GFDL can beat the "LSTM" EFM. Necessarily, the 50 % LRF resulted in the exclusion of a greater percentage of timesteps. Individual LSMs range from 24 % to 72 % (ORCHIDEE2 and NoahMP respectively, both for NEE) of timesteps removed. A mean of

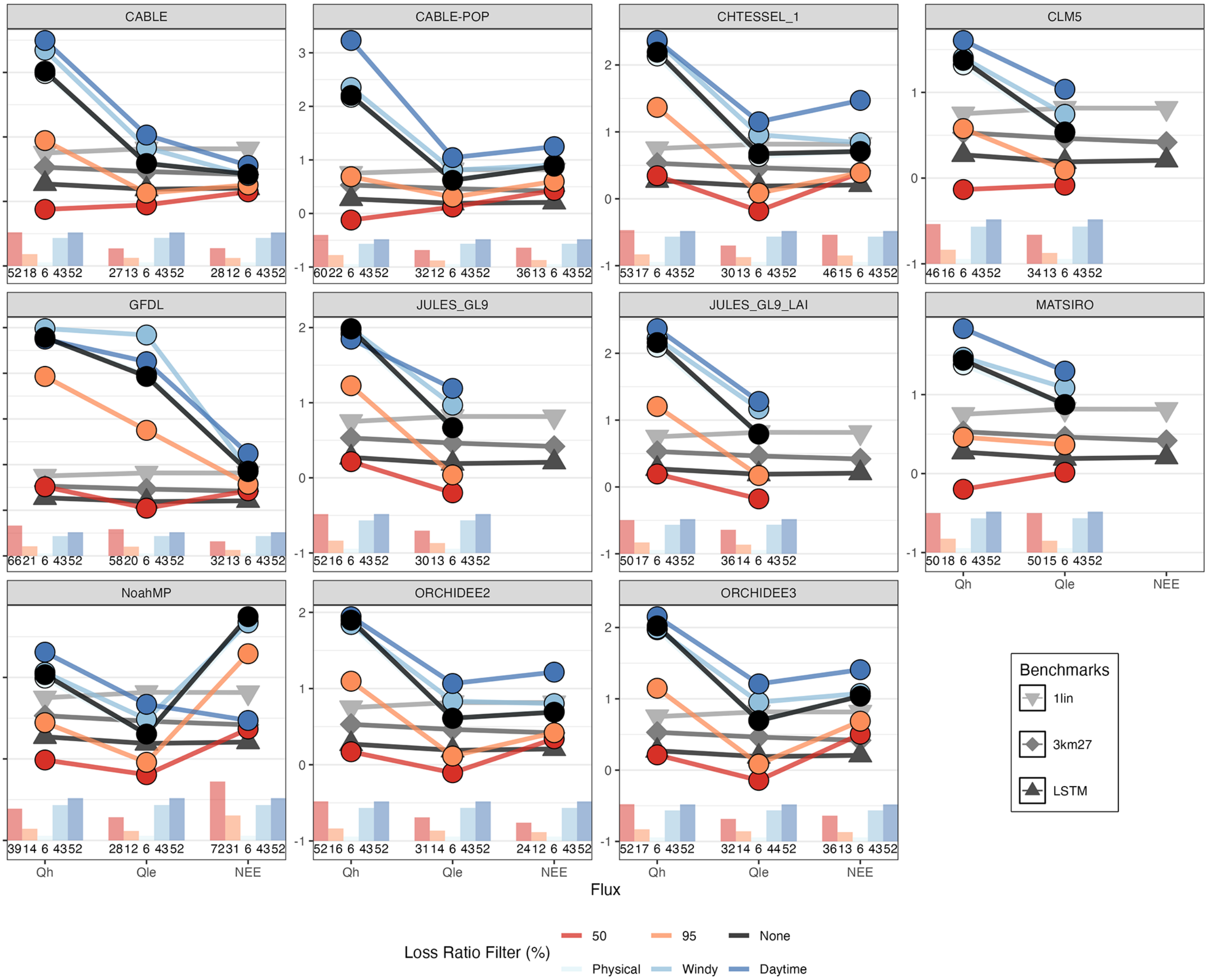


Figure 5. LSM Performance after Application of Different Timestep Filters. The three fluxes are on the x axis, and the y axis is the independent normalised metric value iNMV. Note that the y axis varies for each facet. Each facet is for an individual LSM. Lower iNMV is better. The grey lines are the benchmark EFMs, the black line is the original LSM, and the blue and red lines are the LSM under different filters including the 95 % and 50 % Loss Ratio Filters as well as Daytime, Windy, and Physical filters. The bar charts and labels indicate the percentage of timesteps removed by each of the five filters. Note the differing y axis scales and that the "Physical" filter is often obscured by the "None" LRF/original model performance.

52 % of timesteps were removed for the Qh flux, 35 % for Qle, and 39 % for NEE.

As clear in Figs. 2–4, the original PLUMBER2 dataset includes timesteps that represent physically-impossible conditions. For instance, there are timesteps that fall outside of the VPD-temperature curve which is well-defined by the laws of thermodynamics. Such instances could be due to observational error, poor gap-filling, or incorrectly applied quality control. These timesteps have been retained in the LRF filters in this analysis as they feature in the published PLUMBER2 dataset. However, we also plotted the iNMV of the LSMs when the physically inconsistent timesteps were removed (the "Physical" LRF in light blue). The performance for most models and fluxes was almost indistinguishable from the raw LSM performance (mean relative iNMV improvement of 1 % across all fluxes, Table S3 in the Supplement) and only 6 % of timesteps were removed.

Another non-LRF filter is the "Daytime" filter, implemented by removing all timesteps with a SWdown value of 10 W m^{-2} or less. This removed 52 % of the timesteps in the base analysis, a substantial amount. However, in nearly all cases, the LSM performance for Qh was degraded by applying this filter (Table S4 in the Supplement), with higher iNMV for all LSMs except GFDL and JULES_GL9 (relative iNMV improvement of 1 % and 7 % respectively). For Qle, the LSM performance was always worse for the Daytime filter. In fact, for many LSMs, the base performance beat the "1lin" benchmark while the Daytime filter resulted in even this simple EFM performing better than the LSM. The relative difference in iNMV varied from −11 % (GFDL) to −120 % for NoahMP. The Daytime filter also resulted in worse NEE performance for all LSMs except NoahMP which struggled with simulating NEE under the original conditions (Fig. 4). NoahMP saw a dramatic performance increase (73 % relative iNMV improvement) from failing to beat the "1lin" benchmark to nearly matching the "3km27" benchmark. Significant declines in performance (i.e. a negative change in the EFM benchmark beaten) were seen for CHTESSEL_1 (−108 % iNMV relative difference) and ORCHIDEE2 (−76 % iNMV change) simulations of NEE in terms of EFM benchmark thresholds (from beating "1lin" to not).

The final filter is a "Windy" filter where timesteps with a wind speed of 2 $\mathrm{m\,s^{-1}}$ or less were removed from the iNMV calculations for the LSMs. With the intention of removing timesteps of insuffïcent turbulence for effective eddy-covariance measurements, the iNMV under this filter was generally degraded (Table S5 in the Supplement), falling between the original LSM iNMV and the Daytime NMV. This filter removed 43 % of timesteps. Performance was degraded for Qle (mean relative iNMV difference of −39 %) while Qh and NEE saw little change (mean relative iNMV differences of −4 % and −5 % respectively).

4 Discussion

LSMs are under constant cycles of development with multiple modelling teams across the globe dedicated to improving their performance. By benchmarking LSMs against three different out-of-sample EFMs that performed well when predicting site fluxes, we provide insight for model development and assessment. This was not a "beauty contest" between LSMs – comparing LSMs directly against each other fails to account, for example, for instances where all models are underperforming. Similarly, direct comparison to observations fails to account for the level of inherent predictability in flux measurements (Haughton et al., 2018a, b; Di et al., 2023). Importantly, we also do not address here the absolute performance of the LSMs. Instances of poor LSM performance, as defined in this study, may in fact occur where both LSMs and EFMs are performing objectively well compared to the observations. On the converse, we may here report good LSM performance where the LSMs and EFMs have significant absolute errors compared to observations. Instead, by considering performance relative to out-of-sample EFMs, we can infer a lower bound estimate of predictability that LSMs theoretically should be capable of achieving (Nearing et al., 2018).

Importantly, all EFMs simulations were produced out-of-sample on a site-by-site basis, meaning that they were not exposed to any information from the site they were tasked with simulating. This increases the information provided by the benchmark of prediction skill. For instance, suppose there exists a process that LSMs are assumed to struggle to represent accurately, for example, the temperature acclimation of photosynthesis (Oliver et al., 2022; Ren et al., 2025). If the EFMs also struggle under these conditions, then LSM development may be ill-spent focused on improving the representation of this process because the poor EFM performance would indicate that we lack the necessary information (whether missing observations of related processes or even the processes themselves) to easily improve LSM performance. Instead, in areas where the EFMs outperform the LSMs by a substantial margin, we can hypothesise that we have the information required to model the process accurately. In such areas, targeted LSM development may see easier performance gains.

In line with this, our results show that model improvements targeted to behaviour in a small number of specific meteorological conditions could significantly improve LSM performance. For instance, consider the MATSIRO LSM simulations of sensible heat. As illustrated in Fig. 5, in the PLUMBER2 analysis across all site-timesteps, MATSIRO underperformed with simulations worse than a simple linear regression. However, when we applied the 95 % LRF, the performance of MATSIRO dramatically improved. Removing 17 % of timesteps resulted in MATSIRO leapfrogging the "1lin" model and even beating the "3km27" benchmark. Figure 2 shows where these 17 % of timesteps are located in meteorological space – conditions of high VPD and high wind speed relative to air temperature as well as high and low temperature values relative to incoming longwave radiation. That is, removing a small number of timesteps located in discrete meteorological conditions resulted in a substantial performance improvement. This was especially true for sensible heat flux where the message of the PLUMBER2 analysis (Abramowitz et al., 2024) – that for sensible heat simulations, LSMs were consistently beaten by a simple linear model – would significantly change. The LSM simulations of Qle improved dramatically under the 95 % LRF, frequently improving the EFM benchmark threshold they could beat to the best EFM in the analysis. This implies that, compared to the other two fluxes, the performance of the LSM in predicting Qle was dominated by a few poorly modelled instances. In other words, there were timesteps (less than 20 % for all models, see dark red cells in Fig. 3) where the LSM performance was so much worse than the average performance that these outliers significantly skewed the metrics towards worse performance.

An interesting result is that the only substantial improvement in model performance when applying the "Daytime" filter is for the NEE flux simulated by NoahMP. For all other models, the "Daytime" LRF reduces NEE performance. This would suggest that the poor performance of NoahMP for NEE is related to soil respiration which dominates nighttime fluxes. The lack of similar improvement in the "Windy" filter applied to NoahMP supports this. Similarly, the other LSMs having worse NEE performance when filtering out nighttime timesteps would indicate that these LSMs more accurately capture the soil respiration. This might in turn mean that the other LSMs are not as competent in simulating GPP as the original PLUMBER2 results for NEE suggest. Of course, such inferences might be tempered by the possible poor nighttime flux data meaning that the LSMs are attempting to simulate biased data at nighttime.

Figures 2–4 visualise the LSM input space through two-dimensional fingerprints in combinations of input variables interacting with Tair. There was a clear "edge effect" for all three fluxes, albeit of varying magnitude. This indicates that LSM performance was weak whenever an extreme value in the input space of a variable coincided with an extreme air temperature *relative to the variable value in question*. In other words, the edge effect was itself two-dimensional. For example, the LSM performance for sensible heat was notably weaker at the edge values of longwave radiation and air temperature. These edge conditions would imply extreme levels of humidity and/or cloudiness (the drivers of longwave radiation) compared to the temperature. The weaker LSM performance at these timesteps may relate to site water availability and the LSMs incorrectly partitioning energy between sensible and latent heat, poor representation of the effects of direct vs diffuse radiation, or perhaps even changes in the heat capacity of the air that are not considered within the LSMs. Another notable edge effect is seen in the VPD-Tair fingerprints where LSMs have high LSM Loss Ratios for both sensible and latent heat in the upper edge cells (Figs. 2 and 3). This implies that at times of relatively high VPD and low humidity for the given temperature, LSMs struggle to predict Qh and Qle well relative to the EFMs. This could potentially arise from incorrect energy partitioning in the LSMs (Yuan et al., 2022). Similar high LLRs are seen at high relative VPD for NEE (Fig. 4) but to a lesser extent, with cells at the lower and upper ends of the temperature range exhibiting good performance for the carbon flux. Interestingly, NEE performance also suffers under high relative temperatures for a given VPD. This may be further evidence for the energy partitioning explanation since the LSMs can perform well for NEE in some of the same regions that they struggle for Qh and Qle. This would align with Haughton et al. (2016) who found that, in the original PLUMBER analysis, LSM errors were largely attributable to energy partitioning within the LSMs. Yet another edge case is at high wind speeds, relative to Tair, where substantial LSM losses occur for all three fluxes. Future work should investigate which of these possible causes may be responsible for the specific edge effects.

Figure 4 indicates that a relationship between temperature and LSM performance appears to exist for NEE. For NoahMP in particular, and CABLE and CHTESSEL_1 to a lesser degree, there is a clear progression from low LLR to high LLR as air temperature increases. This relationship also exists for CABLE-POP, GFDL, ORCHIDEE2, and ORCHIDEE3 but with generally better LSM performance. One potential reason for better LSM performance at low temperatures is simply that photosynthesis and repsiration are significantly diminished or zero at cold temperatures which is easy to incorporate in LSMs through temperature thresholds. Another potential cause could be the temperature acclimation of photosynthesis at higher temperatures with LSMs struggling to model this process (Blyth et al., 2021; Mengoli et al., 2022). Interestingly, all LSMs except NoahMP still exhibit an edge effect at higher temperatures with the most extreme high temperatures also showing good LSM performance.

LSMs are often tested against certain types of climatic extreme events as part of their development. For instance, LSM performance under drought conditions (De Kauwe et al., 2015; Ukkola et al., 2016; Huang et al., 2016; Harper et al., 2021) and heatwaves (Mu et al., 2021a, b) is heavily studied. It is worth noting that while these extremes operate at longer timescales than the half-hourly data explored here, they necessarily need to result from processes simulated at the half-hourly timescale in these models. When data is temporally aggregated from this half-hourly timestep – even to the 6-hourly inputs of e.g. TRENDY (Sitch et al., 2024) – the extreme edge cases are averaged out. However, it is likely more difficult for LSMs to capture the high-resolution behaviour at the sub-daily scale than averages over longer periods. For example, it would be theoretically possible to simulate perfect monthly average fluxes and have poor, but compensating, process representations at the shorter timescales. In the extreme case, an LSM could even simulate the diurnal cycle out of phase by 12 h, and this may not be apparent when performance is analysed at monthly timescales. Interestingly, Haughton et al. (2016) showed that, for the original PLUMBER experiment, temporal aggregation made no significant difference to LSM performance relative to EFMs. This may indicate that LSMs submitted to other benchmarking and comparison studies are too heavily calibrated to the aggregated data on which they are applied. Our results indicate that LSM behaviour under all types of short-term climatic extremes is worthy of investigation, including during colder temperatures and varying humidity levels. This edge effect is then most likely explained by the parametrisation of LSMs not being tested on meteorological edge conditions at high temporal resolution to the degree necessary. A potential reason for this lack of parametrisation testing at the input boundaries is the complexity of the LSMs. The number of parameters in LSMs is increasing over time (Fisher and Koven, 2020) as development attempts to improve the representation of numerous model components, such as soil structure and hydrology (Fatichi et al., 2020; Xu et al., 2023). This increasing complexity may not be ideal given that most parameters are poorly constrained (Shiklomanov et al., 2020; Famiglietti et al., 2021).

Interestingly, by their nature and as illustrated in the fingerprints of Figs. 2–4, these edge cells contained fewer timesteps than the cells near the centre of the variable distribution. As such, there was much less training information available for these conditions for the EFMs – in fact, these edge cells were often populated with timesteps entirely from a single site (see Fig. S1 in the Supplement). Since the EFMs were trained out-of-sample, they potentially had zero training data from these extreme conditions. As such, it might be assumed that EFM performance would be at its worst at the edge. Yet it was within these data-sparse regions that the EFMs outperformed the LSMs. This would imply that the flux behaviour under the extreme conditions can be learned from other areas of the input space. Hence, it is unlikely to be novel processes or missing biophysical interactions that limit LSM performance in this space. In fact, Haughton et al. (2016) found in the first PLUMBER experiment that, on a site-by-site basis, LSMs could perform much better than EFMs at unique sites dominated by less-common behaviour or unusual processes. Another consideration is the temporal nature of these edge timesteps – did the extremes come from consecutive timesteps during an extreme event or do they occur only as random noise in the measurements? While we did not explicitly assess the temporal connections within our framework, if these coinciding extreme values did occur only due to random noise, then we might expect the EFMs to perform worse than the LSMs here. Since most of the extreme values are observed at single sites (Fig. S1), the EFMs have no visibility of data from the exact same conditions when simulating the timestep and should therefore equally struggle to accurately model random noise. A next step here may be exploring the results at site-level to identify more explicitly the processes in action within each cell with a high LLR.

Figures S2–S4 in the Supplement show the LSM fingerprints for the mean LSM error relative to the observations within each cell for Qh, Qle, and NEE respectively. While not considered in-depth in the same manner as the benchmarking, a few key findings stand out. Firstly, the magnitude of mean error exhibits a similar edge effect to the benchmarking losses with larger mean errors towards the edges of the two-dimensional. As such, targeting model developments at these cells is not only feasible as shown by the benchmarking results in Figs. 2–4 but also targets the objectively worst performing conditions when considered against observations. Evidence for the power of our benchmarking and LSM Loss framework is evinced by the LSMs' mean error for NEE, especially NoahMP (Fig. S4). While the mean error tends to have a greater magnitude at higher temperatures as most evident for NoahMP, the clear relationship between temperature and LSM performance (in terms of the LLR) with higher temperatures having higher LSM Loss Ratios (Fig. 4) is not discernible. In general, high LLR does not align with consistent under- or overestimation. A good example is GFDL's performance for latent heat (Figs. 3 and S3). A large area of high LLR at mid to low temperatures contains cells with positive and negative mean errors. Since no clear relationship exists between areas of high LLR and the LSM mean error, we recommend that model developers use LLR benchmarking to identify priority areas before assessing the model error and other metrics to pinpoint which processes or parameters may need additional consideration.

The importance of observations at particular meteorological extremes is clear from the edge effect observed here. Such analysis is only possible because the existing eddy-covariance networks have managed to capture these conditions. However, van der Horst et al. (2019) showed that the current network of flux towers undersamples high temperature conditions. It is likely also the case that the potential extremes of other variables are also underobserved. As such, there is a need for increased observations from ecosystems exposed to extreme meteorology (in this multi-dimensional sense). This is important both because of the performance edge effect we noted here, and the fact that these conditions will likely become more common than they have been historically. While additional flux towers are an easy and obvious request, two other actions could be taken. The first and simplest is ensuring that existing towers can operate during the extremes experienced at their locations by enhancing protection from the elements and disturbances. Secondly, the rapid deployment of portable flux measurements when meteorological extremes are forecast could help develop the necessary dataset of extreme observations. Such deployments could be eddy-covariance towers (Billesbach et al., 2004; Ocheltree and Loescher, 2007), leaf-canopy or footprint-scale (respiration chambers) measurement campaigns, that could complement ecosystem-scale understanding.

As mentioned for the Fig. 5 results, the PLUMBER2 analysis contains some timesteps that are physically inconsistent based on temperature and humidity relationships. It was possible that such timesteps may have negatively affected this analysis – the EFMs were not constrained by physics and so were more likely to be able to accurately model these timesteps than the LSMs. The manner in which individual LSMs handle physical inconsistency in inputs may also differ in methodology and therefore the impact on LSM performance. However, removing these timesteps from the analysis did not improve LSM performance, with any changes in iNMV minimal. Approximately 6 % of the timesteps were removed, representing a non-negligible portion of the data. Whether the ability of the LSMs to robustly deal with such physically-inconsistent inputs is reassuring (since model outputs did not massively degrade when forced by them) or of concern (why should we expect LSMs to be able to simulate fluxes under these unpredictable conditions?) is unclear and is an area for further research.

5 Conclusions

We have shown that the LSM underperformance relative to empirical benchmarks reported by the PLUMBER2 experiment can largely be attributed to LSM performance under specific conditions. In other words, LSMs perform significantly better and can beat better EFM benchmarks when certain meteorological conditions are excluded from the analysis. Notably, the conditions of markedly poor LSM performance occur in only a fraction of the total site-timesteps being considered with the majority occurring under coinciding extremes of two or more meteorological variables. In particular, the concurrent extremes, both low and high, of incoming longwave radiation and air temperature result in poor LSM performance for sensible heat. In addition, LSM performance under high values of VPD relative to air temperature show strong potential for improvement for all three fluxes of sensible heat, latent heat, and net ecosystem exchange. Hence, the timesteps of poor performance are mostly described by a edge effect visible in 2-dimensional fingerprints of meteorological input space. By focussing LSM development on the dominant processes at these half-hourly mutual extremes of multiple drivers, we have shown that substantial performance gains could be realised. This places clear value on observations of ecosystem fluxes at these meteorological edge conditions. Without observations of these conditions, the continued advancement of LSMs is constrained by a lack of calibration and testing data. Targeted field campaigns, using rapid deployments to sample responses, may be a necessary step to guide efficient future LSM development and process refinement.

Code and data availability. The flux tower data used here are available at https://doi.org/10.25914/5fdb0902607e1 as per Ukkola (2020) and use data acquired and shared by the FLUXNET community, including these networks: AmeriFlux, AfriFlux, AsiaFlux, CarboAfrica, CarboEuropeIP, CarboItaly, CarboMont, ChinaFlux, Fluxnet-Canada, GreenGrass, ICOS, KoFlux, LBA, NECC, OzFlux-TERN, TCOS-Siberia, and USCCC. The ERA-Interim reanalysis data are provided by the European Centre for Medium-Range Weather Forecasts (ECMWF) and are processed by LSCE. The FLUXNET eddy covariance data processing and harmonisation was carried out by the European Fluxes Database Cluster, AmeriFlux Management Project, and Fluxdata project of FLUXNET, with the support of CDIAC, the ICOS Ecosystem Thematic Center, and the OzFlux, ChinaFlux, and AsiaFlux offices. All the land model simulations in this experiment are hosted at https://modelevaluation.org (last access: 8 January 2024) and, to the extent that participants have no legal barriers to sharing them, are available after registration at https://modelevaluation.org (last access: 8 January 2024). All analyses were performed using the R software (R Core Team, 2020). Analysis code is available at https://doi.org/10.5281/zenodo.17962998.

Supplement. The supplement related to this article is available online at https://doi.org/10.5194/bg-23-263-2026-supplement.

Author contributions. JCP conceived the study with input from GAb, AJP, and IRT. GAr, MJB, CRF, JK, HK, DML, TN, KWO, CO, AU, NV, and XWF performed the land model runs for PLUMBER2. JCP conducted the analysis with continuous input from GAb, AJP, MGDK, and IRT. JCP wrote the first draft of the manuscript. All authors contributed to the final revision of the manuscript and participated in the peer review process.

Competing interests. The contact author has declared that none of the authors has any competing interests.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. The authors bear the ultimate responsibility for providing appropriate place names. Views expressed in the text are those of the authors and do not necessarily reflect the views of the publisher.

Acknowledgements. Jon Cranko Page, Gab Abramowitz, Martin G. De Kauwe, Andy J. Pitman, and Isaac R. Towers were supported by the Australian Research Council Centre of Excellence for Climate Extremes (CE170100023). Martin G. De Kauwe was additionally supported by the ARC Discovery Grants (Grant DP190102025 and DP190101823), the NSW Research Attraction and Acceleration Program, and acknowledges funding from the UK Natural Environment Research Council (NE/W010003/1). We thank the National Computational Infrastructure at the Australian National University, an initiative of the Australian Government for access to supercomputer resources. Hyungjun Kim was supported by Brain Pool Plus program funded by the Ministry of Science and ICT through the National Research Foundation of Korea (RS-2025-02312954) and the Korea Meteorological Administration Research and Development Program under Grant RS-2025-02222417. Keith W. Oleson would like to acknowledge high-performance computing support from the Derecho system (https://doi.org/10.5065/qx9a-pg09) provided by the NSF National Center for Atmospheric Research (NCAR), sponsored by the National Science Foundation.

Financial support. This research has been supported by the Australian Research Council Centre of Excellence for Climate Extremes (grant no. CE170100023).

Review statement. This paper was edited by Benjamin Stocker and reviewed by three anonymous referees.

References

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

Alton, P. B.: Representativeness of global climate and vegetation by carbon-monitoring networks; implications for estimates of gross and net primary productivity at biome and global levels, Agr. Forest Meteorol., 290, 108017, https://doi.org/10.1016/j.agrformet.2020.108017, 2020.

Arora, V. K., Katavouta, A., Williams, R. G., Jones, C. D., Brovkin, V., Friedlingstein, P., Schwinger, J., Bopp, L., Boucher, O., Cadule, P., Chamberlain, M. A., Christian, J. R., Delire, C., Fisher, R. A., Hajima, T., Ilyina, T., Joetzjer, E., Kawamiya, M., Koven, C. D., Krasting, J. P., Law, R. M., Lawrence, D. M., Lenton, A., Lindsay, K., Pongratz, J., Raddatz, T., Séférian, R., Tachiiri, K., Tjiputra, J. F., Wiltshire, A., Wu, T., and Ziehn, T.: Carbon–concentration and carbon–climate feedbacks in CMIP6 models and their comparison to CMIP5 models, Biogeosciences, 17, 4173–4222, https://doi.org/10.5194/bg-17-4173-2020, 2020.

Aubinet, M., Feigenwinter, C., Heinesch, B., Bernhofer, C., Canepa, E., Lindroth, A., Montagnani, L., Rebmann, C., Sedlak, P., and Van Gorsel, E.: Direct advection measurements do not help to solve the night-time CO_2 closure problem: evidence from three different forests, Agr. Forest Meteorol., 150, 655–664, https://doi.org/10.1016/j.agrformet.2010.01.016, 2010.

Aubinet, M., Vesala, T., and Papale, D.: Eddy Covariance: A Practical Guide to Measurement and Data Analysis, Springer Netherlands, Dordrecht, the Netherlands, https://doi.org/10.1007/978-94-007-2351-1, 2012.

Balsamo, G., Beljaars, A., Scipal, K., Viterbo, P., van den Hurk, B., Hirschi, M., and Betts, A. K.: A revised hydrology for the ECMWF model: verification from field site to terrestrial water storage and impact in the integrated forecast system, J. Hydrometeorol., 10, 623–643, https://doi.org/10.1175/2008JHM1068.1, 2009.

Bastos, A., Orth, R., Reichstein, M., Ciais, P., Viovy, N., Zaehle, S., Anthoni, P., Arneth, A., Gentine, P., Joetzjer, E., Lienert, S., Loughran, T., McGuire, P. C., O, S., Pongratz, J., and Sitch, S.: Vulnerability of European ecosystems to two compound dry and hot summers in 2018 and 2019, Earth Syst. Dynam., 12, 1015–1035, https://doi.org/10.5194/esd-12-1015-2021, 2021.

Beringer, J., Moore, C. E., Cleverly, J., Campbell, D. I., Cleugh, H., De Kauwe, M. G., Kirschbaum, M. U. F., Griebel, A., Grover, S., Huete, A., Hutley, L. B., Laubach, J., Van Niel, T., Arndt, S. K., Bennett, A. C., Cernusak, L. A., Eamus, D., Ewenz, C. M., Goodrich, J. P., Jiang, M., Hinko-Najera, N., Isaac, P., Hobeichi, S., Knauer, J., Koerber, G. R., Liddell, M., Ma, X., Macfarlane, C., McHugh, I. D., Medlyn, B. E., Meyer, W. S., Norton, A. J., Owens, J., Pitman, A., Pendall, E., Prober, S. M., Ray, R. L., Restrepo-Coupe, N., Rifai, S. W., Rowlings, D., Schipper, L., Silberstein, R. P., Teckentrup, L., Thompson, S. E., Ukkola, A. M., Wall, A., Wang, Y.-P., Wardlaw, T. J., and Woodgate, W.: Bridge to the future: important lessons from 20 years of ecosystem observations made by the OzFlux network, Glob. Change Biol., 28, 3489–3514, https://doi.org/10.1111/gcb.16141, 2022.

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.

Best, M. J., Abramowitz, G., Johnson, H. R., Pitman, A. J., Balsamo, G., Boone, A., Cuntz, M., Decharme, B., Dirmeyer, P. A., Dong, J., Ek, M., Guo, Z., Haverd, V., van den Hurk, B. J. J., Nearing, G. S., Pak, B., Peters-Lidard, C., Santanello, J. A., Stevens, L., and Vuichard, N.: The plumbing of land surface models: benchmarking model performance, J. Hydrometeorol., 16, 1425–1442, https://doi.org/10.1175/JHM-D-14-0158.1, 2015.

Billesbach, D. P., Fischer, M. L., Torn, M. S., and Berry, J. A.: A portable eddy covariance system for the measurement of ecosystem–atmosphere exchange of CO_2, water vapor, and energy, J. Atmos. Ocean. Tech., 21, 639–650, https://doi.org/10.1175/1520-0426(2004)021<0639:APECSF>2.0.CO;2, 2004.

Blyth, E., Clark, D. B., Ellis, R., Huntingford, C., Los, S., Pryor, M., Best, M., and Sitch, S.: A comprehensive set of benchmark tests for a land surface model of simultaneous fluxes of water and carbon at both the global and seasonal scale, Geosci. Model Dev., 4, 255–269, https://doi.org/10.5194/gmd-4-255-2011, 2011.

Blyth, E. M., Arora, V. K., Clark, D. B., Dadson, S. J., De Kauwe, M. G., Lawrence, D. M., Melton, J. R., Pongratz, J., Turton, R. H., Yoshimura, K., and Yuan, H.: Advances in land surface modelling, Current Climate Change Reports, 7, 45–71, https://doi.org/10.1007/s40641-021-00171-5, 2021.

Boussetta, S., Balsamo, G., Beljaars, A., Panareda, A.-A., Calvet, J.-C., Jacobs, C., van den Hurk, B., Viterbo, P., Lafont, S., Dutra, E., Jarlan, L., Balzarolo, M., Papale, D., and van der Werf, G.: Natural land carbon dioxide exchanges in the ECMWF integrated forecasting system: implementation and offline validation, J. Geophys. Res.-Atmos., 118, 5923–5946, https://doi.org/10.1002/jgrd.50488, 2013.

Canadell, J., Monteiro, P., Costa, M., Cotrim da Cunha, L., Cox, P., Eliseev, A., Henson, S., Ishii, M., Jaccard, S., Koven, C., Lohila, A., Patra, P., Piao, S., Rogelj, J., Syampungani, S., Zaehle, S., and Zickfeld, K.: Global carbon and other biogeochemical cycles and feedbacks, in: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, edited by: Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M., Huang, M., Leitzell, K., Lonnoy, E., Matthews, J., Maycock, T., Waterfield, T., Yelekçi, O., Yu, R., and Zhou, B., Cambridge University Press, Cambridge, UK and New York, NY, USA, https://doi.org/10.1017/9781009157896.007, 673–816, 2021.

Cao, D., Zhang, J., Xun, L., Yang, S., Wang, J., and Yao, F.: Spatiotemporal variations of global terrestrial vegetation climate potential productivity under climate change, Sci. Total Environ., 770, 145320, https://doi.org/10.1016/j.scitotenv.2021.145320, 2021.

Charney, J. G.: Dynamics of deserts and drought in the Sahel, Q. J. Roy. Meteor. Soc., 101, 193–202, https://doi.org/10.1002/qj.49710142802, 1975.

Chu, H., Baldocchi, D. D., John, R., Wolf, S., and Reichstein, M.: Fluxes all of the time? A primer on the temporal representativeness of FLUXNET, J. Geophys. Res.-Biogeo., 122, 289–307, https://doi.org/10.1002/2016JG003576, 2017.

Chu, H., Luo, X., Ouyang, Z., Chan, W. S., Dengel, S., Biraud, S. C., Torn, M. S., Metzger, S., Kumar, J., Arain, M. A., Arkebauer, T. J., Baldocchi, D., Bernacchi, C., Billesbach, D., Black, T. A., Blanken, P. D., Bohrer, G., Bracho, R., Brown, S., Brunsell, N. A., Chen, J., Chen, X., Clark, K., Desai, A. R., Duman, T., Durden, D., Fares, S., Forbrich, I., Gamon, J. A., Gough, C. M., Griffis, T., Helbig, M., Hollinger, D., Humphreys, E., Ikawa, H., Iwata, H., Ju, Y., Knowles, J. F., Knox, S. H., Kobayashi, H., Kolb, T., Law, B., Lee, X., Litvak, M., Liu, H., Munger, J. W., Noormets, A., Novick, K., Oberbauer, S. F., Oechel, W., Oikawa, P., Papuga, S. A., Pendall, E., Prajapati, P., Prueger, J., Quinton, W. L., Richardson, A. D., Russell, E. S., Scott, R. L., Starr, G., Staebler, R., Stoy, P. C., Stuart-Haëntjens, E., Sonnentag, O., Sullivan, R. C., Suyker, A., Ueyama, M., Vargas, R., Wood, J. D., and Zona, D.: Representativeness of eddy-covariance flux footprints for areas surrounding AmeriFlux sites, Agr. Forest Meteorol., 301–302, 108350, https://doi.org/10.1016/j.agrformet.2021.108350, 2021.

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.

Cranko Page, J., Abramowitz, G., De Kauwe, M. G., and Pitman, A. J.: Are plant functional types fit for purpose?, Geophys. Res. Lett., 51, e2023GL104962, https://doi.org/10.1029/2023GL104962, 2024.

De Kauwe, M. G., Zhou, S.-X., Medlyn, B. E., Pitman, A. J., Wang, Y.-P., Duursma, R. A., and Prentice, I. C.: Do land surface models need to include differential plant species responses to drought? Examining model predictions across a mesic-xeric gradient in Europe, Biogeosciences, 12, 7503–7518, https://doi.org/10.5194/bg-12-7503-2015, 2015.

Di, C., Wang, T., Han, Q., Mai, M., Wang, L., and Chen, X.: Complexity and predictability of daily actual evapotranspiration across climate regimes, Water Resour. Res., 59, e2022WR032811, https://doi.org/10.1029/2022WR032811, 2023.

Dunne, J. P., Horowitz, L. W., Adcroft, A. J., Ginoux, P., Held, I. M., John, J. G., Krasting, J. P., Malyshev, S., Naik, V., Paulot, F., Shevliakova, E., Stock, C. A., Zadeh, N., Balaji, V., Blanton, C., Dunne, K. A., Dupuis, C., Durachta, J., Dussin, R., Gauthier, P. P. G., Griffies, S. M., Guo, H., Hallberg, R. W., Harrison, M., He, J., Hurlin, W., McHugh, C., Menzel, R., Milly, P. C. D., Nikonov, S., Paynter, D. J., Ploshay, J., Radhakrishnan, A., Rand, K., Reichl, B. G., Robinson, T., Schwarzkopf, D. M., Sentman, L. T., Underwood, S., Vahlenkamp, H., Winton, M., Wittenberg, A. T., Wyman, B., Zeng, Y., and Zhao, M.: The GFDL Earth System Model Version 4.1 (GFDL-ESM 4.1): overall coupled model description and simulation characteristics, J. Adv. Model. Earth Sy., 12, e2019MS002015, https://doi.org/10.1029/2019MS002015, 2020.

Dutra, E., Balsamo, G., Viterbo, P., Miranda, P. M. A., Beljaars, A., Schär, C., and Elder, K.: An improved snow scheme for the ECMWF land surface model: description and offline validation, J. Hydrometeorol., 11, 899–916, https://doi.org/10.1175/2010JHM1249.1, 2010.

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.

Famiglietti, C. A., Smallman, T. L., Levine, P. A., Flack-Prain, S., Quetin, G. R., Meyer, V., Parazoo, N. C., Stettz, S. G., Yang, Y., Bonal, D., Bloom, A. A., Williams, M., and Konings, A. G.: Optimal model complexity for terrestrial carbon cycle prediction, Biogeosciences, 18, 2727–2754, https://doi.org/10.5194/bg-18-2727-2021, 2021.

Fatichi, S., Or, D., Walko, R., Vereecken, H., Young, M. H., Ghezzehei, T. A., Hengl, T., Kollet, S., Agam, N., and Avissar, R.: Soil structure is an important omission in Earth system models, Nat. Commun., 11, 522, https://doi.org/10.1038/s41467-020-14411-z, 2020.

Fisher, R. A. and Koven, C. D.: Perspectives on the future of land surface models and the challenges of representing complex terrestrial systems, J. Adv. Model. Earth Sy., 12, https://doi.org/10.1029/2018MS001453, 2020.

Friedlingstein, P., Meinshausen, M., Arora, V. K., Jones, C. D., Anav, A., Liddicoat, S. K., and Knutti, R.: Uncertainties in CMIP5 climate projections due to carbon cycle feedbacks, J. Climate, 27, 511–526, https://doi.org/10.1175/JCLI-D-12-00579.1, 2014.

Friedlingstein, P., O'Sullivan, M., Jones, M. W., Andrew, R. M., Hauck, J., Landschützer, P., Le Quéré, C., Li, H., Luijkx, I. T., Olsen, A., Peters, G. P., Peters, W., Pongratz, J., Schwingshackl, C., Sitch, S., Canadell, J. G., Ciais, P., Jackson, R. B., Alin, S. R., Arneth, A., Arora, V., Bates, N. R., Becker, M., Bellouin, N., Berghoff, C. F., Bittig, H. C., Bopp, L., Cadule, P., Campbell, K., Chamberlain, M. A., Chandra, N., Chevallier, F., Chini, L. P., Colligan, T., Decayeux, J., Djeutchouang, L. M., Dou, X., Duran Rojas, C., Enyo, K., Evans, W., Fay, A. R., Feely, R. A., Ford, D. J., Foster, A., Gasser, T., Gehlen, M., Gkritzalis, T., Grassi, G., Gregor, L., Gruber, N., Gürses, Ö., Harris, I., Hefner, M., Heinke, J., Hurtt, G. C., Iida, Y., Ilyina, T., Jacobson, A. R., Jain, A. K., Jarníková, T., Jersild, A., Jiang, F., Jin, Z., Kato, E., Keeling, R. F., Klein Goldewijk, K., Knauer, J., Korsbakken, J. I., Lan, X., Lauvset, S. K., Lefèvre, N., Liu, Z., Liu, J., Ma, L., Maksyutov, S., Marland, G., Mayot, N., McGuire, P. C., Metzl, N., Monacci, N. M., Morgan, E. J., Nakaoka, S.-I., Neill, C., Niwa, Y., Nützel, T., Olivier, L., Ono, T., Palmer, P. I., Pierrot, D., Qin, Z., Resplandy, L., Roobaert, A., Rosan, T. M., Rödenbeck, C., Schwinger, J., Smallman, T. L., Smith, S. M., Sospedra-Alfonso, R., Steinhoff, T., Sun, Q., Sutton, A. J., Séférian, R., Takao, S., Tatebe, H., Tian, H., Tilbrook, B., Torres, O., Tourigny, E., Tsujino, H., Tubiello, F., van der Werf, G., Wanninkhof, R., Wang, X., Yang, D., Yang, X., Yu, Z., Yuan, W., Yue, X., Zaehle, S., Zeng, N., and Zeng, J.: Global Carbon Budget 2024, Earth Syst. Sci. Data, 17, 965–1039, https://doi.org/10.5194/essd-17-965-2025, 2025.

Gier, B. K., Schlund, M., Friedlingstein, P., Jones, C. D., Jones, C., Zaehle, S., and Eyring, V.: Representation of the terrestrial carbon cycle in CMIP6, Biogeosciences, 21, 5321–5360, https://doi.org/10.5194/bg-21-5321-2024, 2024.

Griebel, A., Metzen, D., Pendall, E., Burba, G., and Metzger, S.: Generating spatially robust carbon budgets from flux tower observations, Geophys. Res. Lett., 47, e2019GL085942, https://doi.org/10.1029/2019GL085942, 2020.

Gu, L., Pallardy, S. G., Yang, B., Hosman, K. P., Mao, J., Ricciuto, D., Shi, X., and Sun, Y.: Testing a land model in ecosystem functional space via a comparison of observed and modeled ecosystem flux responses to precipitation regimes and associated stresses in a Central U.S. forest, J. Geophys. Res.-Biogeo., 121, 1884–1902, https://doi.org/10.1002/2015JG003302, 2016.

Harper, A. B., Williams, K. E., McGuire, P. C., Duran Rojas, M. C., Hemming, D., Verhoef, A., Huntingford, C., Rowland, L., Marthews, T., Breder Eller, C., Mathison, C., Nobrega, R. L. B., Gedney, N., Vidale, P. L., Otu-Larbi, F., Pandey, D., Garrigues, S., Wright, A., Slevin, D., De Kauwe, M. G., Blyth, E., Ardö, J., Black, A., Bonal, D., Buchmann, N., Burban, B., Fuchs, K., de Grandcourt, A., Mammarella, I., Merbold, L., Montagnani, L., Nouvellon, Y., Restrepo-Coupe, N., and Wohlfahrt, G.: Improvement of modeling plant responses to low soil moisture in JULESvn4.9 and evaluation against flux tower measurements, Geosci. Model Dev., 14, 3269–3294, https://doi.org/10.5194/gmd-14-3269-2021, 2021.

Haughton, N., Abramowitz, G., Pitman, A. J., Or, D., Best, M. J., Johnson, H. R., Balsamo, G., Boone, A., Cuntz, M., Decharme, B., Dirmeyer, P. A., Dong, J., Ek, M., Guo, Z., Haverd, V., van den Hurk, B. J. J., Nearing, G. S., Pak, B., Santanello, J. A., Stevens, L. E., and Vuichard, N.: The plumbing of land surface models: is poor performance a result of methodology or data quality?, J. Hydrometeorol., 17, 1705–1723, https://doi.org/10.1175/JHM-D-15-0171.1, 2016.

Haughton, N., Abramowitz, G., De Kauwe, M. G., and Pitman, A. J.: Does predictability of fluxes vary between FLUXNET sites?, Biogeosciences, 15, 4495–4513, https://doi.org/10.5194/bg-15-4495-2018, 2018a.

Haughton, N., Abramowitz, G., and Pitman, A. J.: On the predictability of land surface fluxes from meteorological variables, Geosci. Model Dev., 11, 195–212, https://doi.org/10.5194/gmd-11-195-2018, 2018b.

Haverd, V., Smith, B., Cook, G. D., Briggs, P. R., Nieradzik, L., Roxburgh, S. H., Liedloff, A., Meyer, C. P., and Canadell, J. G.: A stand-alone tree demography and landscape structure module for Earth system models, Geophys. Res. Lett., 40, 5234–5239, https://doi.org/10.1002/grl.50972, 2013.

Haverd, V., Cuntz, M., Nieradzik, L. P., and Harman, I. N.: Improved representations of coupled soil–canopy processes in the CABLE land surface model (Subversion revision 3432), Geosci. Model Dev., 9, 3111–3122, https://doi.org/10.5194/gmd-9-3111-2016, 2016.

Haverd, V., Smith, B., Nieradzik, L., Briggs, P. R., Woodgate, W., Trudinger, C. M., Canadell, J. G., and Cuntz, M.: A new version of the CABLE land surface model (Subversion revision r4601) incorporating land use and land cover change, woody vegetation demography, and a novel optimisation-based approach to plant coordination of photosynthesis, Geosci. Model Dev., 11, 2995–3026, https://doi.org/10.5194/gmd-11-2995-2018, 2018.

He, C., Valayamkunnath, P., Barlage, M., Chen, F., Gochis, D., Cabell, R., Schneider, T., Rasmussen, R., Niu, G.-Y., Yang, Z.-L., Niyogi, D., and Ek, M.: Modernizing the open-source community Noah with multi-parameterization options (Noah-MP) land surface model (version 5.0) with enhanced modularity, interoperability, and applicability, Geosci. Model Dev., 16, 5131–5151, https://doi.org/10.5194/gmd-16-5131-2023, 2023.

Henderson-Sellers, A., McGuffie, K., and Pitman, A. J.: The Project for Intercomparison of Land-surface Parametrization Schemes (PILPS): 1992 to 1995, Clim. Dynam., 12, 849–859, https://doi.org/10.1007/s003820050147, 1996.

Huang, Y., Gerber, S., Huang, T., and Lichstein, J. W.: Evaluating the drought response of CMIP5 models using global gross primary productivity, leaf area, precipitation, and soil moisture data, Global Biogeochem. Cy., 30, 1827–1846, https://doi.org/10.1002/2016GB005480, 2016.

Kowalczyk, E. A., Wang, Y. P., Law, R. M., Davies, H., Mcgregor, J., and Abramowitz, G.: The CSIRO Atmosphere Biosphere Land Exchange (CABLE) Model for Use in Climate Models and as an Offline Model, Tech. Rep. 43, CSIRO, Aspendale, Vic., https://doi.org/10.4225/08/58615c6a9a51d, 2006.

Krinner, G., Viovy, N., de Noblet-Ducoudré, N., Ogée, J., Polcher, J., Friedlingstein, P., Ciais, P., Sitch, S., and Prentice, I. C.: A dynamic global vegetation model for studies of the coupled atmosphere-biosphere system, Global Biogeochem. Cy., 19, https://doi.org/10.1029/2003GB002199, 2005.

Lawrence, D. M., Fisher, R. A., Koven, C. D., Oleson, K. W., Swenson, S. C., Bonan, G., Collier, N., Ghimire, B., van Kampenhout, L., Kennedy, D., Kluzek, E., Lawrence, P. J., Li, F., Li, H., Lombardozzi, D., Riley, W. J., Sacks, W. J., Shi, M., Vertenstein, M., Wieder, W. R., Xu, C., Ali, A. A., Badger, A. M., Bisht, G., van den Broeke, M., Brunke, M. A., Burns, S. P., Buzan, J., Clark, M., Craig, A., Dahlin, K., Drewniak, B., Fisher, J. B., Flanner, M., Fox, A. M., Gentine, P., Hoffman, F., Keppel-Aleks, G., Knox, R., Kumar, S., Lenaerts, J., Leung, L. R., Lipscomb, W. H., Lu, Y., Pandey, A., Pelletier, J. D., Perket, J., Randerson, J. T., Ricciuto, D. M., Sanderson, B. M., Slater, A., Subin, Z. M., Tang, J., Thomas, R. Q., Val Martin, M., and Zeng, X.: The Community Land Model Version 5: description of new features, benchmarking, and impact of forcing uncertainty, J. Adv. Model. Earth Sy., 11, 4245–4287, https://doi.org/10.1029/2018MS001583, 2019.

MATSIRO6 Document Writing Team: Description of MATSIRO6, Tech. Rep. 66, Division of Climate System Research, Atmosphere and Ocean Research Institute, The University of Tokyo, https://doi.org/10.15083/0002000181, 2021.

Mengoli, G., Agustí-Panareda, A., Boussetta, S., Harrison, S. P., Trotta, C., and Prentice, I. C.: Ecosystem photosynthesis in land-surface models: a first-principles approach incorporating acclimation, J. Adv. Model. Earth Sy., 14, e2021MS002767, https://doi.org/10.1029/2021MS002767, 2022.

Mu, M., De Kauwe, M. G., Ukkola, A. M., Pitman, A. J., Gimeno, T. E., Medlyn, B. E., Or, D., Yang, J., and Ellsworth, D. S.: Evaluating a land surface model at a water-limited site: implications for land surface contributions to droughts and heatwaves, Hydrol. Earth Syst. Sci., 25, 447–471, https://doi.org/10.5194/hess-25-447-2021, 2021a.

Mu, M., De Kauwe, M. G., Ukkola, A. M., Pitman, A. J., Guo, W., Hobeichi, S., and Briggs, P. R.: Exploring how groundwater buffers the influence of heatwaves on vegetation function during multi-year droughts, Earth Syst. Dynam., 12, 919–938, https://doi.org/10.5194/esd-12-919-2021, 2021b.

Nearing, G. S., Ruddell, B. L., Clark, M. P., Nijssen, B., and Peters-Lidard, C.: Benchmarking and process diagnostics of land models, J. Hydrometeorol., 19, 1835–1852, https://doi.org/10.1175/JHM-D-17-0209.1, 2018.

Niu, G.-Y., Yang, Z.-L., Mitchell, K. E., Chen, F., Ek, M. B., Barlage, M., Kumar, A., Manning, K., Niyogi, D., Rosero, E., Tewari, M., and Xia, Y.: The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements, J. Geophys. Res.-Atmos., 116, https://doi.org/10.1029/2010JD015139, 2011.

Ocheltree, T. W. and Loescher, H. W.: Design of the AmeriFlux portable eddy covariance system and uncertainty analysis of carbon measurements, J. Atmos. Ocean. Tech., 24, 1389–1406, https://doi.org/10.1175/JTECH2064.1, 2007.

Oliver, R. J., Mercado, L. M., Clark, D. B., Huntingford, C., Taylor, C. M., Vidale, P. L., McGuire, P. C., Todt, M., Folwell, S., Shamsudheen Semeena, V., and Medlyn, B. E.: Improved representation of plant physiology in the JULES-vn5.6 land surface model: photosynthesis, stomatal conductance and thermal acclimation, Geosci. Model Dev., 15, 5567–5592, https://doi.org/10.5194/gmd-15-5567-2022, 2022.

Pastorello, G., Trotta, C., Canfora, E., Chu, H., Christianson, D., Cheah, Y.-W., Poindexter, C., Chen, J., Elbashandy, A., Humphrey, M., Isaac, P., Polidori, D., Ribeca, A., van Ingen, C., Zhang, L., Amiro, B., Ammann, C., Arain, M. A., Ardö, J., Arkebauer, T., Arndt, S. K., Arriga, N., Aubinet, M., Aurela, M., Baldocchi, D., Barr, A., Beamesderfer, E., Marchesini, L. B., Bergeron, O., Beringer, J., Bernhofer, C., Berveiller, D., Billesbach, D., Black, T. A., Blanken, P. D., Bohrer, G., Boike, J., Bolstad, P. V., Bonal, D., Bonnefond, J.-M., Bowling, D. R., Bracho, R., Brodeur, J., Brümmer, C., Buchmann, N., Burban, B., Burns, S. P., Buysse, P., Cale, P., Cavagna, M., Cellier, P., Chen, S., Chini, I., Christensen, T. R., Cleverly, J., Collalti, A., Consalvo, C., Cook, B. D., Cook, D., Coursolle, C., Cremonese, E., Curtis, P. S., D'Andrea, E., da Rocha, H., Dai, X., Davis, K. J., De Cinti, B., de Grandcourt, A., De Ligne, A., De Oliveira, R. C., Delpierre, N., Desai, A. R., Di Bella, C. M., di Tommasi, P., Dolman, H., Domingo, F., Dong, G., Dore, S., Duce, P., Dufrêne, E., Dunn, A., Dušek, J., Eamus, D., Eichelmann, U., ElKhidir, H. A. M., Eugster, W., Ewenz, C. M., Ewers, B., Famulari, D., Fares, S., Feigenwinter, I., Feitz, A., Fensholt, R., Filippa, G., Fischer, M., Frank, J., Galvagno, M., Gharun, M., Gianelle, D., Gielen, B., Gioli, B., Gitelson, A., Goded, I., Goeckede, M., Goldstein, A. H., Gough, C. M., Goulden, M. L., Graf, A., Griebel, A., Gruening, C., Grünwald, T., Hammerle, A., Han, S., Han, X., Hansen, B. U., Hanson, C., Hatakka, J., He, Y., Hehn, M., Heinesch, B., Hinko-Najera, N., Hörtnagl, L., Hutley, L., Ibrom, A., Ikawa, H., Jackowicz-Korczynski, M., Janouš, D., Jans, W., Jassal, R., Jiang, S., Kato, T., Khomik, M., Klatt, J., Knohl, A., Knox, S., Kobayashi, H., Koerber, G., Kolle, O., Kosugi, Y., Kotani, A., Kowalski, A., Kruijt, B., Kurbatova, J., Kutsch, W. L., Kwon, H., Launiainen, S., Laurila, T., Law, B., Leuning, R., Li, Y., Liddell, M., Limousin, J.-M., Lion, M., Liska, A. J., Lohila, A., López-Ballesteros, A., López-Blanco, E., Loubet, B., Loustau, D., Lucas-Moffat, A., Lüers, J., Ma, S., Macfarlane, C., Magliulo, V., Maier, R., Mammarella, I., Manca, G., Marcolla, B., Margolis, H. A., Marras, S., Massman, W., Mastepanov, M., Matamala, R., Matthes, J. H., Mazzenga, F., McCaughey, H., McHugh, I., McMillan, A. M. S., Merbold, L., Meyer, W., Meyers, T., Miller, S. D., Minerbi, S., Moderow, U., Monson, R. K., Montagnani, L., Moore, C. E., Moors, E., Moreaux, V., Moureaux, C., Munger, J. W., Nakai, T., Neirynck, J., Nesic, Z., Nicolini, G., Noormets, A., Northwood, M., Nosetto, M., Nouvellon, Y., Novick, K., Oechel, W., Olesen, J. E., Ourcival, J.-M., Papuga, S. A., Parmentier, F.-J., Paul-Limoges, E., Pavelka, M., Peichl, M., Pendall, E., Phillips, R. P., Pilegaard, K., Pirk, N., Posse, G., Powell, T., Prasse, H., Prober, S. M., Rambal, S., Rannik, Ü., Raz-Yaseef, N., Reed, D., de Dios, V. R., Restrepo-Coupe, N., Reverter, B. R., Roland, M., Sabbatini, S., Sachs, T., Saleska, S. R., Sánchez-Cañete, E. P., Sanchez-Mejia, Z. M., Schmid, H. P., Schmidt, M., Schneider, K., Schrader, F., Schroder, I., Scott, R. L., Sedlák, P., Serrano-Ortíz, P., Shao, C., Shi, P., Shironya, I., Siebicke, L., Šigut, L., Silberstein, R., Sirca, C., Spano, D., Steinbrecher, R., Stevens, R. M., Sturtevant, C., Suyker, A., Tagesson, T., Takanashi, S., Tang, Y., Tapper, N., Thom, J., Tiedemann, F., Tomassucci, M., Tuovinen, J.-P., Urbanski, S., Valentini, R., van der Molen, M., van Gorsel, E., van Huissteden, K., Varlagin, A., Verfaillie, J., Vesala, T., Vincke, C., Vitale, D., Vygodskaya, N., Walker, J. P., Walter-Shea, E., Wang, H., Weber, R., Westermann, S., Wille, C., Wofsy, S., Wohlfahrt, G., Wolf, S., Woodgate, W., Li, Y., Zampedri, R., Zhang, J., Zhou, G., Zona, D., Agarwal, D., Biraud, S., Torn, M., and Papale, D.: The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data, Scientific Data, 7, 225, https://doi.org/10.1038/s41597-020-0534-3, 2020.

R Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, available at: , 2020.

Ren, Y., Wang, H., Harrison, S. P., Prentice, I. C., Mengoli, G., Zhao, L., Reich, P. B., and Yang, K.: Incorporating the acclimation of photosynthesis and leaf respiration in the Noah-MP land surface model: model development and evaluation, J. Adv. Model. Earth Sy., 17, e2024MS004599, https://doi.org/10.1029/2024MS004599, 2025.

Shevliakova, E., Malyshev, S., Martinez-Cano, I., Milly, P. C. D., Pacala, S. W., Ginoux, P., Dunne, K. A., Dunne, J. P., Dupuis, C., Findell, K. L., Ghannam, K., Horowitz, L. W., Knutson, T. R., Krasting, J. P., Naik, V., Phillipps, P., Zadeh, N., Yu, Y., Zeng, F., and Zeng, Y.: The land component LM4.1 of the GFDL Earth system model ESM4.1: model description and characteristics of land surface climate and carbon cycling in the historical simulation, J. Adv. Model. Earth Sy., 16, e2023MS003922, https://doi.org/10.1029/2023MS003922, 2024.

Shiklomanov, A. N., Bond-Lamberty, B., Atkins, J. W., and Gough, C. M.: Structure and parameter uncertainty in centennial projections of forest community structure and carbon cycling, Glob. Change Biol., 26, 6080–6096, https://doi.org/10.1111/gcb.15164, 2020.

Sitch, S., Friedlingstein, P., Gruber, N., Jones, S. D., Murray-Tortarolo, G., Ahlström, A., Doney, S. C., Graven, H., Heinze, C., Huntingford, C., Levis, S., Levy, P. E., Lomas, M., Poulter, B., Viovy, N., Zaehle, S., Zeng, N., Arneth, A., Bonan, G., Bopp, L., Canadell, J. G., Chevallier, F., Ciais, P., Ellis, R., Gloor, M., Peylin, P., Piao, S. L., Le Quéré, C., Smith, B., Zhu, Z., and Myneni, R.: Recent trends and drivers of regional sources and sinks of carbon dioxide, Biogeosciences, 12, 653–679, https://doi.org/10.5194/bg-12-653-2015, 2015.

Sitch, S., O'Sullivan, M., Robertson, E., Friedlingstein, P., Albergel, C., Anthoni, P., Arneth, A., Arora, V. K., Bastos, A., Bastrikov, V., Bellouin, N., Canadell, J. G., Chini, L., Ciais, P., Falk, S., Harris, I., Hurtt, G., Ito, A., Jain, A. K., Jones, M. W., Joos, F., Kato, E., Kennedy, D., Klein Goldewijk, K., Kluzek, E., Knauer, J., Lawrence, P. J., Lombardozzi, D., Melton, J. R., Nabel, J. E. M. S., Pan, N., Peylin, P., Pongratz, J., Poulter, B., Rosan, T. M., Sun, Q., Tian, H., Walker, A. P., Weber, U., Yuan, W., Yue, X., and Zaehle, S.: Trends and drivers of terrestrial sources and sinks of carbon dioxide: an overview of the TRENDY project, Global Biogeochem. Cy., 38, e2024GB008102, https://doi.org/10.1029/2024GB008102, 2024.

Ukkola, A.: PLUMBER2: Forcing and Evaluation Datasets for a Model Intercomparison Project for Land Surface Models v1.0, NCI Australia [data set], https://doi.org/10.25914/5FDB0902607E1, 2020.

Ukkola, A. M., De Kauwe, M. G., Pitman, A. J., Best, M. J., Abramowitz, G., Haverd, V., Decker, M., and Haughton, N.: Land surface models systematically overestimate the intensity, duration and magnitude of seasonal-scale evaporative droughts, Environ. Res. Lett., 11, 104012, https://doi.org/10.1088/1748-9326/11/10/104012, 2016.

Ukkola, A. M., Abramowitz, G., and De Kauwe, M. G.: A flux tower dataset tailored for land model evaluation, Earth Syst. Sci. Data, 14, 449–461, https://doi.org/10.5194/essd-14-449-2022, 2022.

van den Hurk, B., Viterbo, P., Beljaars, A., and Betts, A.: Offline Validation of the ERA40 Surface Scheme, Technical Memorandum 295, ECMWF, Shinfield Park, Reading, https://doi.org/10.21957/9aoaspz8, 2000.

van den Hurk, B., Kim, H., Krinner, G., Seneviratne, S. I., Derksen, C., Oki, T., Douville, H., Colin, J., Ducharne, A., Cheruy, F., Viovy, N., Puma, M. J., Wada, Y., Li, W., Jia, B., Alessandri, A., Lawrence, D. M., Weedon, G. P., Ellis, R., Hagemann, S., Mao, J., Flanner, M. G., Zampieri, M., Materia, S., Law, R. M., and Sheffield, J.: LS3MIP (v1.0) contribution to CMIP6: the Land Surface, Snow and Soil moisture Model Intercomparison Project – aims, setup and expected outcome, Geosci. Model Dev., 9, 2809–2832, https://doi.org/10.5194/gmd-9-2809-2016, 2016.

van der Horst, S. V. J., Pitman, A. J., De Kauwe, M. G., Ukkola, A., Abramowitz, G., and Isaac, P.: How representative are FLUXNET measurements of surface fluxes during temperature extremes?, Biogeosciences, 16, 1829–1844, https://doi.org/10.5194/bg-16-1829-2019, 2019.

Vuichard, N., Messina, P., Luyssaert, S., Guenet, B., Zaehle, S., Ghattas, J., Bastrikov, V., and Peylin, P.: Accounting for carbon and nitrogen interactions in the global terrestrial ecosystem model ORCHIDEE (trunk version, rev 4999): multi-scale evaluation of gross primary production, Geosci. Model Dev., 12, 4751–4779, https://doi.org/10.5194/gmd-12-4751-2019, 2019.

Walker, A. P., De Kauwe, M. G., Bastos, A., Belmecheri, S., Georgiou, K., Keeling, R. F., McMahon, S. M., Medlyn, B. E., Moore, D. J. P., Norby, R. J., Zaehle, S., Anderson-Teixeira, K. J., Battipaglia, G., Brienen, R. J. W., Cabugao, K. G., Cailleret, M., Campbell, E., Canadell, J. G., Ciais, P., Craig, M. E., Ellsworth, D. S., Farquhar, G. D., Fatichi, S., Fisher, J. B., Frank, D. C., Graven, H., Gu, L., Haverd, V., Heilman, K., Heimann, M., Hungate, B. A., Iversen, C. M., Joos, F., Jiang, M., Keenan, T. F., Knauer, J., Körner, C., Leshyk, V. O., Leuzinger, S., Liu, Y., MacBean, N., Malhi, Y., McVicar, T. R., Penuelas, J., Pongratz, J., Powell, A. S., Riutta, T., Sabot, M. E. B., Schleucher, J., Sitch, S., Smith, W. K., Sulman, B., Taylor, B., Terrer, C., Torn, M. S., Treseder, K. K., Trugman, A. T., Trumbore, S. E., Mantgem, P. J., Voelker, S. L., Whelan, M. E., and Zuidema, P. A.: Integrating the evidence for a terrestrial carbon sink caused by increasing atmospheric CO_2, New Phytol., 229, 2413–2445, https://doi.org/10.1111/nph.16866, 2021.

Wang, Y. P., Kowalczyk, E., Leuning, R., Abramowitz, G., Raupach, M. R., Pak, B., van Gorsel, E., and Luhar, A.: Diagnosing errors in a land surface model (CABLE) in the time and frequency domains, J. Geophys. Res.-Biogeo., 116, https://doi.org/10.1029/2010JG001385, 2011.

Xu, C., Torres-Rojas, L., Vergopolan, N., and Chaney, N. W.: The benefits of using state-of-the-art digital soil properties maps to improve the modeling of soil moisture in land surface models, Water Resour. Res., 59, e2022WR032336, https://doi.org/10.1029/2022WR032336, 2023.

Yuan, K., Zhu, Q., Riley, W. J., Li, F., and Wu, H.: Understanding and reducing the uncertainties of land surface energy flux partitioning within CMIP6 land models, Agr. Forest Meteorol., 319, 108920, https://doi.org/10.1016/j.agrformet.2022.108920, 2022.