On the predictability of turbulent fluxes from land: PLUMBER2 MIP experimental description and preliminary results

Gab Abramowitz[1,2], Anna Ukkola[1,2], Sanaa Hobeichi[1,2], Jon Cranko Page[1,2], Mathew Lipson[3], Martin G. De Kauwe[4], Samuel Green[1,2], Claire Brenner[5], Jonathan Frame[5], Grey Nearing[6], Martyn Clark[7], Martin Best[8], Peter Anthoni[9], Gabriele Arduini[10], Souhail Boussetta[10], Silvia Caldararu[11,12], Kyeungwoo Cho[13], Matthias Cuntz[14], David Fairbairn[10], Craig R. Ferguson[15], Hyungjun Kim[16], Yeonjoo Kim[13], Jürgen Knauer[17,18], David Lawrence[19], Xiangzhong Luo[20], Sergey Malyshev[21], Tomoko Nitta[22], Jerome Ogee[14], Keith Oleson[19], Catherine Ottlé[23], Phillipe Peylin[23], Patricia de Rosnay[10], Heather Rumbold[8], Bob Su[24], Nicolas Vuichard[23], Anthony P. Walker[25], Xiaoni Wang-Faivre[23], Yunfei Wang[24], Yijian Zeng[24]

[1] CLEX, UNSW Sydney, Australia
[2] CCRC, UNSW Sydney, Australia
[3] Bureau of Meteorology, Australia
[4] School of Biological Sciences, University of Bristol, Bristol, BS8 1TQ, UK
[5] University of Alabama, USA
[6] Google, USA
[7] University of Calgary, Canada
[8] UKMO, UK
[9] Karlsruhe Institute of Technology, Institute of Meteorology and Climate Research/Atmospheric Environmental Research, 82467 Garmisch-Partenkirchen, Germany
[10] European Centre for Medium-Range Weather Forecasts (ECMWF), UK
[11] Max Planck Institute for Biogeochemistry, Jena, Germany
[12] Discipline of Botany, School of Natural Sciences, trinity College Dublin, Dublin, Ireland
[13] Yonsei University, Seoul, Korea
[14] INRAE, France
[15] Atmospheric Sciences Research Center, University at Albany, State University of New York, Albany, NY, USA
[16] HydroKlima Lab, KAIST, Daejeon, Korea
[17] CSIRO Environment, Australia
[18] Western Sydney University, Australia
[19] NCAR, USA
[20] National University of Singapore, Singapore
[21] NOAA GFDL, USA
[22] The University of Tokyo, Japan
[23] LSCE, France
[24] University of Twente, Netherlands
[25] Oak Ridge National Laboratory, USA

Correspondence to: Gab Abramowitz (gabriel@unsw.edu.au)

Abstract. Accurate representation of the turbulent exchange of carbon, water, and heat between the land surface and the atmosphere is critical for modelling global energy, water, and carbon cycles, both in future climate projections and weather forecasts. Evaluation of models' ability to do this is performed in a wide range of simulation environments, often without explicit consideration of the degree of observational constraint or uncertainty, and typically without quantification of benchmark performance expectations. We describe a Model Intercomparison Project (MIP) that attempts to resolve these shortcomings, comparing the surface turbulent heat flux predictions of around 20 different land models provided with in-situ meteorological forcing, evaluated with measured surface fluxes using quality-controlled data from 170 eddy-covariance based flux tower sites.

Several out-of-sample empirical model predictions are used to quantify the information available to land models in their forcing data, and so the potential for land model performance improvement. Sites with unusual behaviour, complicated processes, poor data quality or uncommon flux magnitude are more difficult to predict for both mechanistic and empirical models, providing a means for fairer assessment of land model performance. When examining observational uncertainty, model performance does not appear to improve in low turbulence periods, or with energy-balance corrected flux tower data, and

Formatted: Width: 20.99 cm, Height: 29.7 cm
Deleted: es
Deleted: of site fluxes
Deleted: as benchmarks
Deleted: the degree to which
Deleted: could
Deleted: across a broad range of metrics. The performance discrepancy between empirical and mechanistic model predictions also provides a potential pathway to understand sources of model error…
Deleted: will be
Moved (insertion) [5]

indeed some results raise questions about whether the energy-balance correction process itself is appropriate. In all cases results are broadly consistent, with simple out-of-sample empirical models, including linear regression, comfortably outperforming mechanistic land models.

Generally, latent heat flux and net ecosystem exchange of CO_2 are better predicted by land models than sensible heat flux, despite seeming to have fewer physical controlling processes. Land models that are implemented in Earth System Models also appear to perform notably better than stand-alone ecosystem (including demographic) models, at least in terms of the fluxes examined here. The approach we outline enables isolation of the locations and conditions in which model developers can *know* that a land model can improve, allowing information pathways and discrete parametrisations in models to be identified and targeted for future model development.

1 Introduction

Land models (LMs) simulate terrestrial water, energy and biogeochemical cycles. They simulate the exchange of heat and moisture between the land and atmosphere inside weather forecast models (e.g. Bousetta et al., 2013; Bush et al., 2023), soil moisture and streamflow in hydrological and agricultural applications (e.g. Clark et al., 2015a, 2015b, Buechel, 2021), ecological dynamics and carbon exchange in ecosystem modelling (e.g. Knauer et al., 2023; Bennet et al., 2024), and most of these processes combined inside climate models (e.g. Lawrence et al., 2019; Vuichard et al., 2019; Bi et al., 2020). The fidelity of LM simulations is therefore consequential economically, socially and environmentally.

This paper focuses on a relatively simple question: how should we fairly assess the fidelity of land models? We aim to develop an evaluation framework that gives us confidence that LM evaluation is not partial - not dependent upon a particular metric, observational data choice, over-calibration or overfitting, a particular location or time, or subset of processes - that it is the closest we can reasonably expect to a summative understanding of the shortcomings or strengths of a particular model. This aim is the basis of a LM comparison experiment, PLUMBER2, and we use results from PLUMBER2 to illustrate the framework. It follows from the first Protocol for the Analysis of Land Surface Models (PALS) Land Surface Model Benchmarking Evaluation Project (PLUMBER; Best et al., 2015; Haughton et al., 2016), and addresses many of the shortcomings in its first iteration.

Our question is intentionally methodological, since the consequence of getting the answer wrong is very real – we rely on LMs for a great deal of scientific inference and societally relevant predictions. We consider our aim in two parts. First, what kind of simulation environment allows for the best observational constraint of LMs, so that poor model performance might fairly be attributed to a LM? Second, how do we best structure an evaluation framework to give us confidence in this kind of attribution? We discuss these two questions in turn and highlight how the experimental framework of PLUMBER2 addresses them in a way that the original PLUMBER experiment could not.

While most LMs are applied on regional or global grids, diagnostic evaluation of LMs (that is, understanding why they might be wrong) at these scales is difficult (Li et al., 2018; Wartenburger et al., 2018; Seiler et al., 2022). First, at this scale LMs need to be driven by reanalysis-based meteorology with unquantified uncertainty (Arora et al., 2023), making the attribution of model-observation mismatch inconclusive. Next, observationally based flux evaluation products at these scales, typically

Deleted: d

Deleted: wa

Deleted: both

Deleted: were

Deleted: PLUMBER2

Deleted: , and its openly available data,

Deleted:

Deleted: precise

Deleted: given

Deleted: Results suggest that latent heat flux and net ecosystem exchange of CO_2 are better predicted by land models than sensible heat flux, which at least conceptually would appear to have fewer physical processes controlling it. Land models that are implemented in Earth System Models also appear to perform notably better than stand-alone ecosystem (including demographic) models, at least in terms of the fluxes examined here. ¶

Moved up [5]: does not appear to improve with energy-balance corrected data, and indeed some results raised questions about whether the correction process itself was appropriate. In both cases results were broadly consistent, with simple out-of-sample empirical models, including linear regression, comfortably outperforming mechanistic land models. The PLUMBER2 approach, and its openly available data, enable precise isolation of the locations and conditions in which model developers can *know* that a given land model can improve, allowing information pathways and discrete parametrisations in models to be identified and targeted for model development.¶

Deleted: Flux tower data quality is also explored as an uncertainty source, with the difference between energy-balance corrected versus raw fluxes examined, as well as filtering for low wind speed periods. Land model performance does not appear to improve with energy-balance corrected data, and indeed some results raised questions about whether the correction process itself was appropriate. In both cases results were broadly consistent, with simple out-of-sample empirical models, including linear regression, comfortably outperforming mechanistic land models. The PLUMBER2 approach, and its openly available data, enable precise isolation of the locations and conditions in which model developers can *know* that a given land model can improve, allowing information pathways and discrete parametrisations in models to be identified and targeted for model development.¶

Deleted: with

also without quantified uncertainty, are usually at a low time resolution (e.g. monthly; Pan et al., 2020), so assessment of process representation in LMs can only be done using emergent outcomes, rather than directly.

Site-based LM simulations using observational data collected at flux towers offer a solution to some of these issues, but they come with their own challenges. On the positive side, meteorological variables that drive LMs are directly measured at tower sites, at a time resolution appropriate for LM simulation (typically 30 minutes), and uncertainties are relatively small and quantifiable (Schmidt et al., 2012). Vegetation properties are often documented at site locations, reducing parameter uncertainty in LM simulations (Falge et al., 2016). The fluxes to the atmosphere that LMs are evaluated against are also measured and aggregated to the same time step size as the meteorological driving data. For these reasons, both PLUMBER and PLUMBER2 involve the evaluation of LMs at flux tower sites, but PLUMBER2 examines a much broader range of environments (170 sites instead of 20).

There are however several complicating factors that also make LM constraint using tower data incomplete. The spatial scale represented by a flux tower's fetch – typically not larger than $1km^2$ – is at the very highest resolution of application scales for most LMs (Chu et al., 2021). However, all LMs are designed using leaf-scale or canopy-scale theories (Bonan et al., 2021), and do not contain an explicit length scale that modifies simulation characteristics for the size of the grid cell, so it is unclear whether this represents a problem. Next, not all LM parameters are measured at sites, and indeed at some sites, little information is available beyond a broad indication of vegetation type (Falge et al., 2016). The available information is, however, likely closer to being representative than parameter values prescribed at the grid scale that cannot be directly observed. Finally, and most importantly, the measurement of turbulent fluxes comes with significant uncertainty. Site measurements regularly do not close the energy balance (Stoy et al., 2013; Mauder et al., 2020), in a way that LM are structurally required to do, and measurements are likely to have much greater uncertainty in periods of low turbulence (Goulden et al., 1996; Aubinet et al., 2012). These issues were both ignored in the first PLUMBER experiment, but are directly addressed in PLUMBER2, with evaluation involving both raw and energy-balance corrected site data, as well as filtering to remove periods of low turbulence.

To address our second question, we outline three key aspects of an evaluation framework that will allow attribution of poor model-observation agreement to poor model performance. They are (a) an appropriate suite of metrics, (b) a mechanism to establish threshold values in these metrics that reasonably define "good" or "poor" performance, and (c) a summative indicator that can fairly synthesise information using (a) and (b) to provide a representative overall picture.

For a given variable time series, there are of course many metrics one might use to assess model performance, and it is well recognised that using a single metric will generally not allow for holistic assessment of model performance (Collier et al., 2018; Abramowitz et al., 2019). Both PLUMBER and PLUMBER2 focused on a broad collection of metrics that (i) assessed a wide range of aspects of model performance, and (ii) were independent, in the sense that a change to a model prediction might affect any one of these metrics without affecting others (see Gupta et al, 2009).

Next, establishing *a priori* performance expectations in the form of thresholds in these metrics is key to defining "good" model performance. Models will never agree with observations exactly, but if we could understand how well a perfect model *could* simulate a given environment, given the information provided in its driving variables and observational uncertainty, it would tell us exactly which aspects of observed site behaviour were predictable, and which were not. This idea can be approximated by using out-of-sample empirical models to predict site fluxes, using the same meteorological driving variables as the LMs as predictors (e.g. Abramowitz, 2005; Best et al. 2015; Whitley et al., 2017). While the PLUMBER experiment investigated using

simple out-of-sample empirical models to do this, here we offer a much more comprehensive range of empirical approaches. By varying the complexity of empirical models, and the number and type of predictors we provide it with, we create a hierarchy of benchmark levels of performance in any given metric that reflects different structural assumptions. For example, a LM should provide a more sophisticated prediction of evapotranspiration than a simple empirical model based on incoming shortwave radiation alone, since it contains information about soil moisture availability, soil temperature, vegetation and evaporative demand. By providing some empirical models with lagged variables, or using machine learning structures that allow internal states, we can begin to quantify how much predictive ability model states like soil moisture should provide.

Finally, with many variables, metrics, sites, empirical benchmarks and LMs, the importance of a summative indicator that appropriately synthesizes information and reduces the dimensionality of results should be clear. In PLUMBER, each LM was ranked against benchmark models from best to worst performing, and ranks were averaged over sites and metrics. However, it has since become clear that this can create misleading results. Consider the following example. A LM and three benchmarks have biases in latent heat flux at three sites of (32, 30, 31, 29), (48, 47, 45, 46) and (12, 52, 29, 85) Wm^{-2} respectively, translating to ranks of (4,2,3,1), (4,3,1,2) and (1,3,2,4) and an average rank of (3.3, 2.6, 2.3, 2.3). This summative indicator misleadingly suggests that the LM is comfortably the worst model of the four, when the actual site biases suggest that models were nearly indistinguishable in the first two sites, and the LM notably superior in the third. Using results from PLUMBER2, we examine two alternative summative indicators that resolve this issue. A summary of the main differences between PLUMBER and PLUMBER2 is shown in Table 1.

Before detailing our methodology below, we reinforce that this experimental description paper does not investigate process representations or flaws of any particular model – given the number of models, sites and benchmarks doing so would necessarily present an incomplete picture. We instead focus on developing a fair, holistic framework for LM evaluation, comparison and quantification of performance expectations and present a high-level overview of PLUMBER2 results that will serve as the basis for future detailed, process level analyses that are already underway.

Table 1: A summary of differences between the PLUMBER2 and PLUMBER experiments.

PLUMBER2	PLUMBER
170 (154) sites; 1040 site-years; 1-21 year record length	20 sites; 105 site-years; 1-10 year record length
Site data quality control detailed in Ukkola et al., (2022)	Ad-hoc site data quality control
Sensitivity to energy balance correction or uncorrected fluxes	Uncorrected fluxes only
Sensitivity to night / low turbulence	None
Land surface, ecological and hydrological models	Land surface models only
Linear regression, 3 and 6-variable cluster+regression, random forest (RF), long short-term memory (LSTM) models	Linear regression or 3-variable cluster+regression
Dependent and independent normalised metric value as summative indicators	Rank as only summative indicator

2 Methodology

2.1 Flux tower data

LMs completed simulations at 170 flux tower sites for PLUMBER2, forced with in-situ half-hourly or hourly meteorological variables. The aim was to maximise the number of sites that met variable availability and quality control requirements, as well

Deleted: ¶
Given these LM applications, we have reason to ask 'what makes a good land model?' Too often, our answer has been something akin to "the best performing model", or anything "better than the previous version" of a model. These are unsatisfactory answers for a number of reasons, not least that both discount the possibility that all the models we have are poor or unfit for a given application. The key question is "How can we do better?" ¶
¶
Model benchmarking, as opposed to model evaluation, requires setting expectations of model performance *a priori*, before we know how well our model performs. There are obviously many different ways we could choose to do this. We may require qualitative performance improvement relative to a simpler model, for example. Or, if a model's application is well known, and threshold levels of performance in particular metrics can decide what is sufficient, then the answer is relatively clear - a model is good enough when these thresholds are met. Yet this approach still only tells us about the suitability of the model for a given application, rather than understand whether we have a good model *per se*. What a 'good model' means, without a given application seems like an entirely subjective question, but we suggest this should relate to the fidelity of its representation, given its complexity.¶
¶
Perhaps the most universal approach to answering this broader question is to understand whether a model does a good job of utilising the information that is available to it for prediction (in its input time series, parameters and initial states). A 'perfect model' with all required driving variables would, for example, tell us exactly which aspects of observed site behaviour were predictable, and which were not - it would define predictability. But we might also imagine a 'perfect model' that is only provided with a subset of the required driving variables, and expect that it performs with less fidelity. It would still nevertheless define predictability, but conditional on only a subset of predictors being available. We would, for example, expect that a LM would provide a more sophisticated prediction of evapotranspiration than a response to incoming shortwave radiation alone, since it contains information about soil moisture availability, soil temperature, vegetation and evaporative demand. This is the approach to defining utilisation of information by land models that has been most commonly used - comparing land models to a range of empirical models of increasing complexity, using in-sample empirical approaches to compare to land models calibrated on evaluation data (e.g. eddy covariance fluxes) and out-of-sample empirical approaches to compare to land models without calibration on evaluation data (e.g. Abramowitz, 2005; Best et al. 2015; Whitley et al., 2017). In our case, the best possible utilisation of information towards flux prediction at a flux tower site defines site predictability.¶
¶
The Protocol for the Analysis of Land Surface Models (PALS) Land Surface Model Benchmarking Evaluation Project (PLUMBER; Best et al. 2015; Haughton et al., 2016) explored some of these questions in the form of a model comparison experiment, using 20 flux tower sites and simulations at these sites from nine land surface models (LSMs), driven by half hourly meteorology measured at the si ... [1]

Deleted: Since many of the LMs in this experiment are most often run at a global scale on coarse grids (typically 0.5-1°), it also seems reasonable to question whether an evaluation experiment using site-based data provides a fair test of models' ability. There are obvious differences in the nature of fluxes at the site scale of the flux tower data (typically < 1km^2) and the larger grid cell sizes used in a global simulation. Almost all LMs, however, are designed using leaf-scale or canopy-scale theories, and do not contain an explicit length scale that modifies simulation characteristics for the size of the grid cell. Indeed many of these LMs are being used within regional modelling frameworks at resolutions that approximate this flux tower fetch. There are also very real benefits at the site scale. When direct observations of meteorological forcing and fluxes are co-located, we have an ability to quantify uncertainties, and errors in measurements of meteorological forcing are relatively small. Flux tower sites allow us to ascribe simulation errors to LMs in a way that is just not possible at gridded scales, where meteorological inputs to LMs cannot be directly measured. Avoiding the 'garbage in, garbage out' problem, and ability to evaluate models at the time step size th ... [2]

Deleted: In contrast to the 20 sites used in the PLUMBER experiment (Best et al., 2015), models

Deleted: While detailed explanations of the motivation, processing steps and quality control of site data are given by Ukkola et al. (2022), a brief overview is given here. At the time of processing, t

as having open-access data. FLUXNET2015, FLUXNET La Thuile Free-Fair-Use, and OzFlux collections were used as the starting point, and after processing with the FluxnetLSM package (Ukkola et al., 2017), it was ensured that sites had: reference (measurement) height, canopy height and IGBP (International Geosphere–Biosphere Programme) vegetation type; whole years of data; and were not missing significant periods of key forcing variables (where gap-filling counted as missing), specifically incoming solar radiation (SWdown), air temperature (Tair), specific humidity (Qair) or precipitation (Rainf). Discerning thresholds in these variables was clearly subjective, but involved consideration of the proportion of time series with measured data, length of gaps, coincidence between variables, and ubiquity of site type - see Ukkola et al. (2022) for detail. Gap-filling (including allowing 100% synthesised data) of downwelling longwave radiation (LWdown) used the approach from Abramowitz et al. (2012). Surface air pressure (PSurf) was based on elevation and temperature, and ambient CO_2 was based on global values (Ukkola et al., 2022).

Since most sites had no publicly available leaf area index (LAI) data, and none had time evolving LAI data, we specified a remotely-sensed LAI time series for each site to try to minimise differences between LMs. LMs that predict LAI would clearly not utilise this (Table S3). The LAI time series were derived from either MODIS (8-daily MCD15A2H product, 2002-2019) or Copernicus Global Land Service (monthly, 1999-2017), with one of these chosen for each site based on a site-by-site analysis considering plausibility and some in-situ data, provided for each time step of meteorological forcing. Time-varying LAI was provided for the time period covered by the remotely-sensed products and otherwise a climatology was constructed from all available years. Some LMs utilised this LAI estimate for a single vegetation type simulation and others partitioned it in a mixed vegetation type representation. LAI estimates remain a key issue for observational constraint of LMs at the site and global scales.

Energy balance closure in flux tower data is particularly relevant in the context of this experiment. At a range of time scales, most sites do not obey the assumed equality of net radiation with the sum of latent heat flux, sensible heat flux and ground heat flux (see Wilson et al., 2002; Stoy et al., 2013; Mauder et al., 2020; Moderow et al., 2021). We therefore need to be careful attributing model-observation mismatch to model error, since LMs are fundamentally constrained to conserve energy. Energy-balance closure correction was part of the FLUXNET2015 release (the bulk of sites here) and we replicated this approach for sites from the other sources. Analyses below consider both uncorrected and corrected latent and sensible heat fluxes, were conducted only on flux time steps that were not gap-filled, and were also run separately filtered by time steps with wind speed above 2 ms^{-1} so that potential concerns about measurement fidelity in low turbulence periods (typically night time) could be investigated.

Forcing and evaluation files were produced in an updated version of ALMA NetCDF (Polcher et al., 1998; 2000), with CF-NetCDF standard name attributes and CMIP equivalent names included where possible. A complete list of these variables, as well as those requested in LM output, are shown in Table S1. Table S2 shows a complete site and site property list. Each site has a page on modelevaluation.org with more detail, including additional references, meta data, photographs and time series plots. Site locations are shown by Ukkola et al. (2022). Site vegetation types and distribution in mean precipitation-temperature space are shown in Fig. S1. Their location on a Budyko style dryness index versus (water) evaporative fraction plot (Budyko, 1974; Chen and Sivapalan, 2020) is shown in Fig. S2a. It is typically assumed that all sites will lie below 1 on the horizontal axis (i.e. evapotranspiration will be less than precipitation) and to the right of the 1-1 line (potential evapotranspiration > evapotranspiration), with drier, water limited sites close to 1 on the horizontal axis on the right hand side and wetter, energy limited sites towards the bottom left hand side close to the 1-1 line.

Deleted: ¶ ¶
Deleted: :
Deleted: information
Deleted: ¶ ¶
Deleted: l
Deleted: and
Deleted: as a single value
Deleted: We note that s
Deleted: While all observational data contain measurement uncertainty, the issue of
Deleted: e
Deleted: ¶ ¶ The simplest approach to dealing with this issue is to correct flux tower data for closure, where sufficient data exist.
Deleted: raw
Deleted: The final f
Deleted: for all sites
Deleted: ultimately
Deleted: , with each site's included years, IGBP vegetation type, mean annual temperature (MAT) and precipitation (MAP), canopy height, reference height, elevation, chosen LAI representation, and references…
Deleted: tower
Deleted: We refer readers to Ukkola et al. (2022) for a global map showing …s s
Deleted: s
Deleted: MAP-MAT

This is however clearly not true for these site data. To understand why, we first examined cumulative precipitation at each site, compared to an in-situ based gridded precipitation product - REGEN (Contractor et al, 2020) - and identified those sites that appeared anomalous. Clearly there are many good reasons why site-based precipitation might disagree with a gridded product, even if it were perfect. A subset of the sites were nevertheless identified as having precipitation data that were *a priori* not realistic, either because missing data had not been gap-filled (and was not flagged as missing, so precipitation flat lined), units had been reported incorrectly (e.g. US-SP1 appears to use inches rather than mm) or winter snowfall was apparently not included in precipitation totals (see Fig. S3). 16 sites were removed from the analysis as a result. These issues were unfortunately only identified after all modelling groups had completed their 170 site simulations, so the LM analyses below are conducted on the remaining 154 sites.

While removing these sites did lessen the extent of the problem, it did not by any means solve it (see Fig. S2b - the same as Fig. S2a but with 154 instead of 170 sites). Next, we examined if using the entire time series for each site, instead of filtering out gap-filled time steps (Fig. S2a has gap-filled data removed) resulted in any qualitative change - it did not (see Fig. S2c). Finally, we investigated whether using energy-balance corrected fluxes had an impact. Fig. S2d shows that it did indeed have a marked effect - but the proportion of sites where evapotranspiration exceeds precipitation *increased*.

Figures S2a-d reinforce how complicated the simulation task is for LMs, with around 30% of sites showing an average evapotranspiration exceeding average precipitation. Despite posing this as a data quality problem above, there are many sound, physically plausible reasons for this, such as hillslope or preferential flow, irrigation or groundwater access by vegetation. Needless to say, most LMs will simply be unable to reproduce this behaviour since these inputs and processes are usually not included. We discuss more about this issue, its influence on results and implications for LM evaluation in the Results and Discussion.

2.2 Land model simulations

Mechanistic LMs ran offline in single-site mode (as opposed to gridded simulations), forced by observed meteorology from the 170 sites. Simulations were requested as "out-of-the-box", using default (usually vegetation type based) parameters for each site, as if the LMs were running a global simulation. Models used the IGBP vegetation type prescribed in each forcing file where possible, mapped to the PFT schemes used by each model. In addition, site canopy height and reference height (measurement / lowest atmospheric model layer height) were provided. No additional parameter information for sites was prescribed.

The rationale for this setup was to understand the fidelity in flux prediction that LMs provide in a well-constructed global simulation noting that different LMs had to adapt their representation approaches in slightly different ways to achieve this (e.g. some use mixed vegetation types to describe a single location). While we might ideally additionally like to ensure that LMs used an appropriate soil type for each site, these are not universally measured or available for all sites, so LMs used their default global soil type grid.

Models were not allowed to calibrate to site fluxes, as we are primarily interested in the insight LMs provide about the system, rather than their fitting ability, which might leave little to distinguish them from machine learning approaches that we already know will perform better (Abramowitz 2012; Beaudry and Renner, 2012; Best et al., 2015; Nearing et al., 2018). Out-of-sample testing for any model, even if only partly empirical, is key to understanding its predictive ability (see Abramowitz et al., 2019), especially when it needs to be applied globally.

Deleted: this was the case

Deleted: as

Formatted: Font: Italic

Deleted: The model simulation protocol was broadly similar to that of PLUMBER. Groups ran

Deleted: m

Deleted: standardised, locally

Deleted: for

Deleted:

Deleted: in each forcing file

Deleted: try to

Deleted: degree of

Deleted: (i.e. where meteorological forcing is as close as possible to being error-free and vegetation properties are appropriately represented by vegetation type),

Deleted: types from their default

Different LMs require different periods of spin-up until model states reach an equilibrium, depending on whether they include a dynamic carbon (C) and/or nitrogen (N) / phosphorus (P) cycle(s), vegetation or stand dynamics. For models where soil temperature and moisture spin-up is sufficient (e.g. if LAI is prescribed rather than predicted), we suggested that model spin-up use the site forcing file and repeatedly simulate *the entire* period, for at least 10 years of simulation, before beginning a simulation on the first year of site data.

For LMs with prognostic LAI and/or soil C, N, and P pools, the process was more complicated. LM simulations were initialised with a spin-up routine resulting in equilibrium conditions of C stocks (and N and P if available) representing the year 1850. Climatic forcing for the spin-up came from the site eddy-covariance forcing file, which was continuously repeated. Atmospheric CO_2 and N deposition levels representing the year 1850 were set to 285 ppm and 0.79 kg N ha-1 yr-1, respectively. The transient phase covered the period 1851 to the year prior the first year in the site data. LMs were forced with historical changes in atmospheric CO_2 and N deposition, continually recycling the meteorological inputs. The meteorological time series was repeated intact rather than in a randomised way, to avoid splitting of the observed meteorological years at the end of each calendar year. This of course does not accurately replicate the land use history of different sites, but in most cases detailed site level histories were not available.

All models participating in PLUMBER2 are shown in Table S3. While some simulation setup information is included in the *Notes* column, more detailed information is available on the Model Output profile page for each set of simulations submitted to modelevaluation.org. While modelling groups were requested to report as many variables as possible from Table S1, the breadth of contributions were highly variable, so in an attempt to include all participants, analyses here focus on latent heat flux (Qle), sensible heat flux (Qh) and Net Ecosystem Exchange of CO_2 (NEE) only.

In addition to the LMs, two 'physical benchmarks' were also included, as per Best et al. (2015) - an implementation of a Manabe bucket model (Manabe, 1969) and a Penman-Monteith model (Monteith and Unsworth, 1990) with a reference stomatal resistance and unrestricted water availability.

2.3 Empirical machine learning based benchmarks

As suggested above, empirical models are key to quantifying site predictability, and so setting benchmark levels of performance for LMs that reflect the varying difficulty or complexity of prediction at different sites, unknown issues with data quality at some sites and more broadly understanding the amount of information that LM inputs provide about fluxes. To do this meaningfully, all empirical models need to provide out-of-sample predictions. That is, every site simulation made by an empirical model here has not used that site's data to build/train the empirical model, and so cannot be overfitted to the characteristics or noise from the site. If the site is unusual, or its data is poor, the empirical models will provide a poor simulation, thus setting a lower benchmark of performance for the LMs.

A hierarchy of different empirical models was used. From the simplest, with lowest performance expectations, to highest, these are:

- ***1lin***: a linear regression of each flux against downward shortwave radiation (SWdown), using half hourly data, training on 169 sites and predicting on one, repeated 170 times, as per Abramowitz (2012) and Best et al. (2015). Two versions were created - one trained to predict raw fluxes (1lin_raw) and one trained to predict energy-balance corrected fluxes (1lin_eb).

Deleted: or not
Deleted: (
Deleted:)
Deleted: , and how this is implemented
Deleted: a simple
Deleted: carbon(
Deleted:)
Deleted: nitrogen (
Deleted:)
Deleted: phosphorous (
Deleted:)
Deleted: ,
Deleted: 1
Deleted: ¶
Deleted: **Table 1: Participating models. Land surface models that are developed as part of a coupled modelling system are denoted as 'climate' or 'weather' in their Notes, depending on their default application being climate projection or weather forecasting, despite simulations in this experiment being uncoupled/offline. In each case, leaf area index (LAI) is either prescribed, computed by the model itself, or not used (NA).¶ Model** ... [3]
Deleted: The empirical models we use as benchmarks are also listed in Table 1.
Deleted: these
Deleted: particular
Deleted: approaches
Deleted: ,
Deleted: firstly to avoid the eccentricities of a single technique or architecture, but more importantly, to give us an approximate equivalent level of performance for each LM within the hierarchy, as a way to rate its performance. Approximately f
Deleted: simple

- ***2lin***: a multiple linear regression of each flux using SWdown and air temperature (Tair) as predictors, using half hourly data, training on 169 sites and predicting on one, as per Abramowitz (2012) and Best et al. (2015).
- ***3km27***: all site-timesteps of three predictors - SWdown, Tair and relative humidity (RH) - from 169 training sites are sorted into 27 clusters using k-means, and all site-timesteps in each cluster are used to establish multiple linear regression parameters against each flux for that cluster. Time steps at the prediction site are sorted into clusters based on proximity to cluster centres, and regression parameters for each cluster are then used to make predictions at the test site, as per Abramowitz (2012) and Best et al. (2015). 27 clusters were chosen to approximately allow each predictor high, medium and low clusters: 3^3=27. Two versions were created - one to predict raw fluxes (3km27_raw) and one to predict energy-balance corrected fluxes (3km27_eb).
- ***6km729***: As per 3km27, but using six predictors - SWdown, Tair, RH, Wind, Precip, LWdown (see Table S1 for variable definitions) - and 729 k-means clusters, training on 169 sites and predicting on one, similar to Haughton et al. (2018). 729 clusters were chosen to approximately allow each predictor high, medium and low clusters: 3^6=729;
- ***6km729lag***: As per 6km729, but with lagged Precip and Tair as additional predictors. These took the form of six additional predictors: mean Precip and Tair from the previous 1-7 days, 8-30 days and 31-90 days. Training on 169 sites and predicting on one, similar to Haughton et al. (2018);
- ***RF***: A Random Forest model with Tair, SWdown, LWdown, Qair, Psurf, Wind, RH, CO2air, VPD, and LAI as predictors. These predictor variables are listed in order of variable importance. While Precip was originally included, it actually offered negative variable importance - suggesting that including Precip reliably *degraded* the empirical prediction out-of-sample. Training was on 169 sites and predicting on one out of sample, repeated 170 times. As a nominally more sophisticated empirical model than the cluster+regression approaches above, RF offers a lower bound estimate of predictability of fluxes from instantaneous conditions (no lags). Two versions were created - one each to predict raw (RF_raw) and energy-balance corrected fluxes (RF_eb).
- ***LSTM***: A Long Short-Term Memory model given as much information as the LMs. Two types of input features were used for training: dynamic features - CO2air, LWdown, Precip, Psurf, Qair, RH, SWdown, Tair, VPD, Wind and LAI - and static site attributes that are constant per site (MAT, range of annual MAT, MAP, mean LAI, range of annual LAI, elevation, canopy height, reference height, latitude, mean SWdown, PET and IGBP vegetation type). Training was on 167 sites and prediction was on the three remaining sites (randomly chosen), repeated to make out-of-sample predictions at all sites. A single LSTM was used to predict Qle, Qh and NEE simultaneously, to account for the fact that the three fluxes are all components of a highly coupled system. The LSTM provides a lower bound estimate of predictability of fluxes using both instantaneous and meteorological conditions and internal states based on them - a proxy for LM states. Two versions were created - one to predict raw (LSTM_raw) and one to predict energy-balance corrected fluxes (LSTM_eb).

2.4 Analyses

The dimensionality and complexity of the PLUMBER2 data obviously present many options to interrogate the performance of LMs. Our analysis focuses on a relatively high-level overview without any intention to be comprehensive - we anticipate that analysis of PLUMBER2 simulations will extend well beyond this paper and will take some time. Results below consider mean fluxes, variable ratios such as evaporative fraction and water use efficiency, before examining the two summative indicators detailed.

The set of metrics we use, shown in Table S4, is independent, in the sense that for a given observational time series, a change can be made to the model time series that will affect any one of these metrics without affecting the others. This is not true, for example, of RMSE and correlation. Metrics are calculated separately for each model at each site.

Deleted: one remaining unseen, out-of-sample

Deleted: fluxes

Deleted: one to predict

Deleted: (that change for each time step

Deleted: fluxes

Deleted: ¶

Deleted: As noted above, all predictions made by empirical / machine learning models are out-of-sample at the site level which means that all of the data from the site being predicted is excluded from the training set for all the empirical models above. We also note that the 1lin, 3km27m, LSTM, and RF models made two separate sets of predictions of Qle and Qh: one when trained on the raw fluxes and one trained on the energy balance corrected fluxes.¶

Deleted: resulting from this experiment

Deleted: s

Deleted: here

Deleted: these

Deleted: Here we consider a few canonical approaches:

Deleted: variability

Deleted: (although using NEE rather than GPP)

Deleted: separation by vegetation type and viewing results in a Budyko curve context

Deleted: ¶
¶
Where possible, we try to (a) be summative, avoiding using any one metric or site at a time, and instead try to present results collectively across all site and an independent metric set, and (b) present LM performance *relative to a benchmark*, rather than using raw metric values. To better understand what we mean by this, let's first look at the metrics we use, shown in Table 2.¶

Deleted: **Table 2: The independent set of metrics used to assess aggregate performance. Metrics are calculated on all time steps (*i=1,..,n*) of observations (O) and each model (M) separately for each site. *B* is the number of bins used for density estimation, in this case 1024.**¶
Mean Bias Error (MBE) ... [4]

Deleted: is

Next, we examine our two summative indicators. To do this, we first set a reference group of benchmark empirical models, and compare all LMs to this reference group. Suppose we wish to compare a given LM against 1lin, 3km27 and LSTM, for example. Then, for each metric (m), at each site and for each variable, we have metric values for the LM, 1lin, 3km27 and LSTM. We then define the normalised metric value (NMV) for this LM at this site, for this variable and metric, in one of two ways.

First, as with PLUMBER, we define LM performance relative to a range of metric values that includes the LM and empirical benchmarks. Instead of using ranks though, we normalise this range to define the *dependent* Normalised Metric Value as:

$$dNMV_{LM} = \frac{m_{LM} - min(m_{1lin}, m_{3km27}, m_{LSTM}, m_{LM})}{max(m_{1lin}, m_{3km27}, m_{LSTM}, m_{LM}) - min(m_{1lin}, m_{3km27}, m_{LSTM}, m_{LM})} \quad (1)$$

So dependent NMV simply denotes where in the metric range of these 4 models the LM was, scaled to be between 0 and 1, with lower values representing better performance. This allows us to average NMV over metrics, sites, variables, vegetation types or other groupings to get an aggregate indication of performance.

The second approach, *independent* NMV, defines the normalised metric range using only the reference benchmark models:

$$iNMV_{LM} = \frac{m_{LM} - min(m_{1lin}, m_{3km27}, m_{LSTM})}{max(m_{1lin}, m_{3km27}, m_{LSTM}) - min(m_{1lin}, m_{3km27}, m_{LSTM})} \quad (2)$$

iNMV allows us to define lower and upper performance expectations to be independent of the LM being assessed. We might expect that 1lin as the simplest model will typically have a value of 1 and LSTM 0, and the LM, if its performance lies between these two, a value somewhere in this interval. It also allows the LM to score a much lower value than zero, if it performs much better than the empirical models, and conversely, a value much larger than 1 if it is much worse.

To illustrate why such a detailed approach to analysis is necessary, we now briefly show why some common heuristic measures of performance are inadequate. Figures S4, S5 and S6 (supplementary material) show the performance results of the 1lin model at the US-Me2 site, examining latent heat flux, sensible heat flux and NEE in three different common graphical performance measures. These are: the average diurnal cycle of NEE, shown for different seasons (Fig. S4); a smoothed time series of Qh (Fig. S5); and the average monthly values of Qle showing the evaporative seasonal cycle (Fig. S6). In most contexts, if these blue curves were plots of a LM's performance, the reader would accept this as qualitative or even quantitative evidence of excellent LM performance. Yet these represent perhaps the simplest possible model - a simple linear regression against shortwave, out-of-sample (trained on other sites only). They illustrate just how much site variability can be simply driven by instantaneous shortwave radiation, and that visual closeness of curves, and an ability to capture seasonal variability, diurnal variability and even interannual variability should not *a priori* be accepted as evidence of good model performance.

As noted above, all analyses were filtered to exclude time steps at each site where observational flux data was flagged as missing or gap-filled. Analyses were half-hourly or hourly, depending on the reported time step size at each site, except for models that only reported monthly outputs, which were then analysed with monthly averages. All data management and analyses were conducted through https://modelevaluation.org (see Abramowitz, 2012), and can be repeated there. The analysis codebase used for PLUMBER2 within https://modelevaluation.org is available at https://gitlab.com/modelevaluation/me.org-r-library.

Deleted: combine the information we get from this independent metric set into a single performance measure

Deleted: and

Deleted: staying with the broader approach used by Best et al. (2015)…

Deleted: we define *dependent* NMV

Deleted: by

Deleted: ing

Deleted: in the metric range being normalised:

Deleted: This is similar to the approach used in Best et al. (2015), except using the average of normalised metric values, rather than discrete metric value ranks. Using metric value ranks effectively forces even spacing of metric values, which can give misleading results, particularly in cases when some models' metric values are clustered.…

Deleted: using an

Deleted: Using this second approach,

Deleted: In this case

Deleted: w

Formatted: Font: Not Italic

Deleted: will have

3. Results

In examining results from the PLUMBER2 experiment, we reinforce that our aim here is to demonstrate that we have created a holistic environment and methodology that allows us to fairly attribute model-observation mismatch to LMs, where appropriate. We do provide a broad overview of the many dimensions of PLUMBER2 results, but do not investigate process representations or flaws of any model – doing so would necessarily present an incomplete picture, since these kinds of findings are specific to particular models, environments and circumstances.

Figure 1 shows the average latent heat flux (Qle) versus sensible heat flux (Qh), averaged across all sites for participating models that reported both variables. Dashed lines show a proxy for observed available energy (around 69 Wm^{-2}, defined as Qle+Qh, assuming mean ground heat flux on longer time scales is zero) and observed Bowen ratio (around 0.7). Perhaps unsurprisingly, models differ most in their partitioning of surface energy (spread along the available energy axis) rather than amount of available energy (spread along the Bowen ratio axis), supporting previous findings (see Haughton et al, 2016). Those LMs that do not operate in a coupled modelling system (i.e. are not coupled to an atmospheric model; EntTBM, LPJ-GUESS, MuSICA, QUINCY, STEMMUS-SCOPE) also appear to have a much broader spread of estimates than those used in coupled models (they are furthest from the observed Bowen ratio in Fig. 1), and the unrestricted moisture store of the Penman-Monteith model makes it a clear outlier.

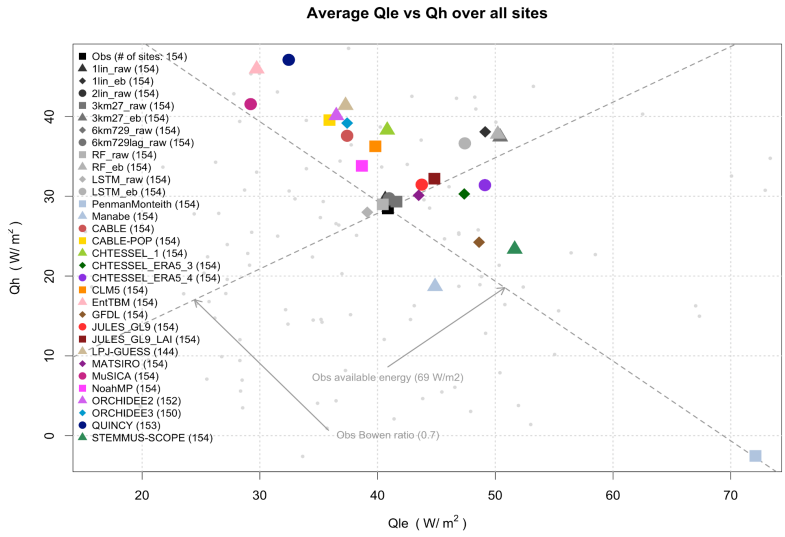


Figure 1: Latent heat flux (Qle) versus sensible heat flux (Qh) averaged over 154 sites, shown for models that submitted both quantities. Dashed lines show observed values of average available energy (Qle+Qh) and average Bowen ratio (Qh/Qle) across the sites, using raw (as opposed to energy balance corrected) flux data. Smaller light grey dots in the background represent individual site averages.

When averaged across all sites, the LMs do not appear to show any clear systematic bias in energy partitioning relative to observations across the ensemble. Note that in Fig. 1 the observations *do not* have the Fluxnet2015 energy-balance correction applied (the equivalent figure using energy balance corrected fluxes is shown in Fig. S7a). Aside from showing a little more available energy (their mean is slightly offset from the observed available energy line, by less than 10%), the LMs are relatively

Deleted: of these

Deleted: .

Deleted: Average l

Deleted: average

Deleted: all

Deleted: constant

Deleted: L

evenly spread around the observational Bowen ratio. This lends little support to an argument of systematic observational bias in the partitioning of available energy leading to apparent poor LM performance.

The empirical models trained to predict raw fluxes (those labelled *_raw) are tightly clustered around the observational crosshairs. While it is not surprising that regression-based models perform well on the mean, these models are entirely out-of-sample, demonstrating forcing meteorology alone provides enough information to predict mean fluxes accurately out-of-sample. The energy-balance corrected observations lie in amongst the empirical models trained to predict energy-balance corrected fluxes (labelled *_eb; the cluster of grey points with higher available energy in Fig. 1 - see Fig. S7a). The average Bowen ratio increases slightly to 0.73 instead of 0.7 with energy-balance correction. Perhaps more interesting is that the corrected version of flux observations contains an average of 16 Wm^{-2} additional energy across these sites, about a 23% increase, and that this value sits much further outside the spread of the mechanistic modelled estimates of available energy than the observed value in Fig. 1. So in this simple metric at least (and indeed in more below), the LMs' performance is not improved with energy-balance corrected flux data. While we present results comparing with raw fluxes in the main part of this manuscript, comparisons against energy-balance corrected data, where they qualitatively differ, are discussed and shown in Supplementary Material. Similarly, when we filter analyses to only include time steps with wind speed above 2 ms^{-1} (Fig. S7b), the scatter of models in Fig. 1 changes surprisingly little.

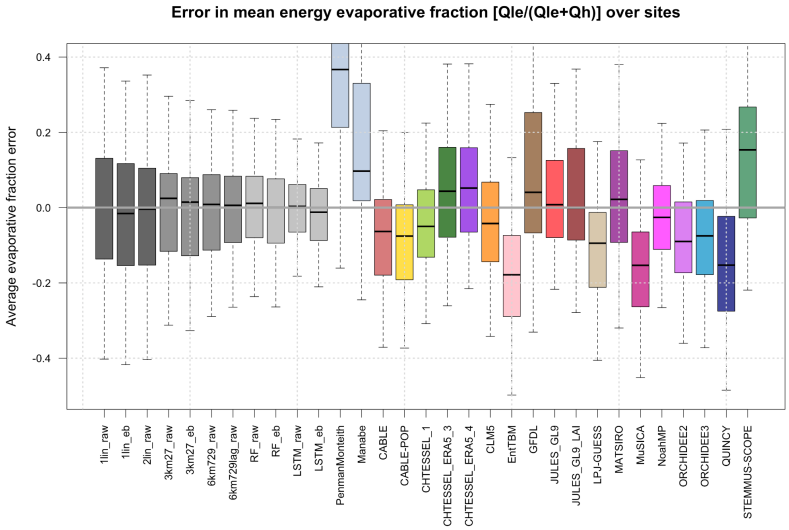


Figure 2: Box plots of error in site mean energy evaporative fraction (Qle/(Qle+Qh)) over all sites, shown separately for each model, using raw flux data across 154 sites.

Figure 2 shows boxplots of error in the average energy evaporative fraction (EF) across the same sites, shown separately for each participating model. Energy evaporative fraction is defined using average flux values at each site: Qle / (Qle+Qh). The equivalent plots using energy balance corrected data and data filtered for wind speed are almost indistinguishable from Fig. 2 and so have not been included. Consistent with what we saw in Fig. 1, the mechanistic benchmarks and ecosystem models show the largest deviation from site observations, and empirical approaches are reliably zero-centred despite having no explicit

Deleted: Perhaps unsurprisingly, observations, once e
Deleted: ,
Deleted: when
Deleted: is used
Deleted: In the discussion section we consider whether results in this paper might raise some doubt about whether the approach to energy balance correction used in Fluxnet2015 is ideal - further results below are somewhat mixed.
Deleted: W
Deleted: with
Moved (insertion) [6]
Deleted: 6
Deleted: 154
Deleted: 6
Deleted: In keeping

mechanism to constrain the ratio between Qle and Qh. The more sophisticated empirical models (6km*, RF, LSTM), as well as being zero error centred, show less spread, meaning they have fewer large errors in energy evaporative fraction. Once again, there does not appear to be any obvious reason to suspect a bias in partitioning in observations - some LMs (6) show a high EF bias, and others (11) a low bias.

An equivalent version of this figure showing water evaporative fraction, Qle / Rainf, is shown in Fig. S8a and Fig. S8b (in supplementary material), using raw and energy-balance corrected fluxes, respectively. Once again, models are well scattered about the zero error line when raw fluxes are used, and almost all appear strongly negatively biased when compared to the energy balance corrected fluxes. The equivalent plots using wind speed filtered data are qualitatively the same as Figures S8a and S8b, and so are not included here.

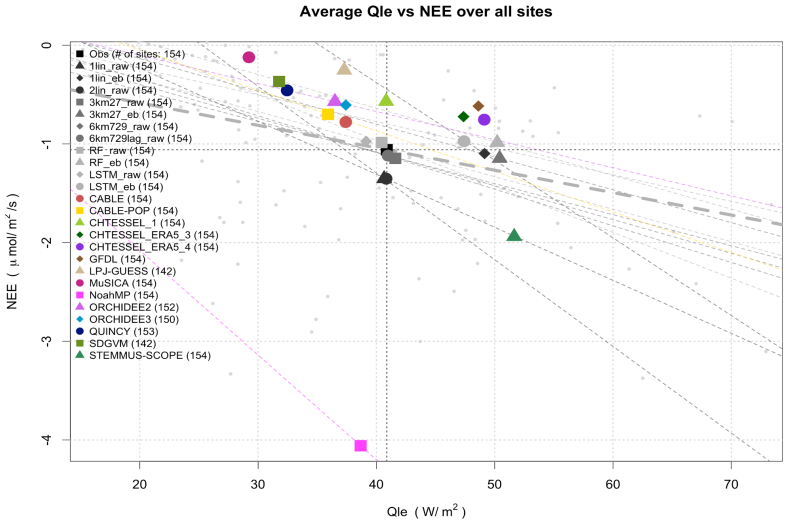


Figure 3: Latent heat flux (Qle) versus net ecosystem exchange of CO2 (NEE) averaged over 154 sites, shown for models that submitted both quantities. The observed value is shown in black with crosshairs. Light grey dots in the background represent individual observed site averages, with the linear fit between them shown in bold dashed grey. Regression lines are also shown for LMs showing a stronger fit than in the observed case (R^2=0.19).

Figure 3 is similar to Fig. 1, but shows average latent heat flux (Qle) versus Net Ecosystem Exchange of CO_2 (NEE) for LMs that reported both variables. Given the expectation that NEE is likely to be strongly dependent on site history, and that we could not reliably include this information in the modelling protocol or account for it in this plot, there is no a priori reason to expect a clear relationship here. While we might broadly expect increasing carbon uptake with increasing Qle, as shown by the observed regression line in Fig. 3, the fit is relatively weak (R^2 is 0.19). LM regressions are shown where their fit has higher R^2 than observed, although we note that aside from ORCHIDEE2, CABLE-POP and NoahMP, only empirical models meet this criterion (unsurprising, since they effectively act as data smoothers).

With the exception of Noah-MP, STEMMUS-SCOPE and some empirical models, all LMs predict less net carbon uptake than is observed. This may well be because the models were run without any site history. That is, the simulated ecosystems were

Deleted: -
Deleted: -
Deleted: 10
Deleted: 10
Deleted: 10
Deleted: 10
Deleted: ¶

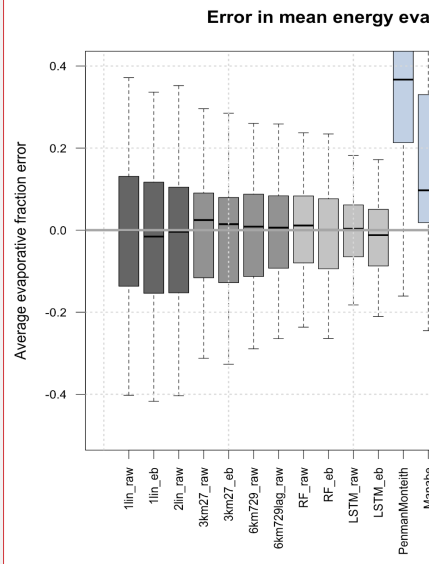


Figure 6: Box plots of error in site mean energy evaporative fraction (Qle/(Qle+Qh)) over all sites, shown separately for each model, using raw flux data across 154 sites.

Deleted: 2
Deleted: Average l
Deleted: average
Deleted: all
Deleted: 2
Deleted: those
Deleted: 2

closer to equilibrium than those in the real world. In equilibrium, vegetation and soil carbon stocks are high and thus respiration is also higher as it is generally simulated as a function of carbon stocks. Ecosystem models predict the least carbon uptake but a large range in Qle values (MuSICA, LPJ-GUESS, QUINCY, SDGVM). The equivalent plot with energy balance corrected Qle values (not shown) simply moves the 'observed' black square to the right, once again sitting amongst 1lin_eb, 3km27_eb, RF_eb and LSTM_eb. Once again, the energy-balance corrected data does not appear to match LM simulations better than raw flux data.

We also note that while LMs' spread might well be because of a lack of site history information, the empirical models show that missing this information does not actually reduce NEE predictability to a large degree (all empirical models are within 0.35 μmol/m^2/s of observations). The empirical models also do not have any site history, and indeed in most cases, do not even use any estimate of LAI. They are trained only at other sites, so they cannot infer any site history information from the meteorology-flux relationship. Despite this, they cluster quite tightly around the observations in Fig. 3, whether predicting raw Qle (cluster of grey points in the crosshairs) or energy-balance corrected Qle (cluster of grey points to the right of this). They all suggest a net uptake of C across these sites, within a narrow range spread around the observations.

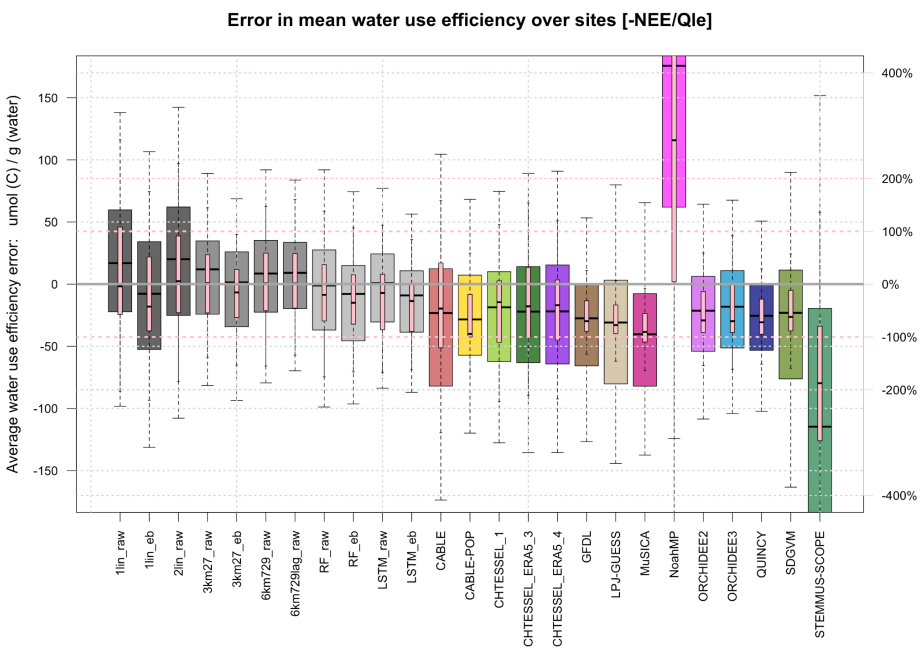
Figure 4 is similar to Fig. 2, but shows error in water-use efficiency (NEE/Qle), expressed in units of μmol of carbon gained per gram of water (left vertical axis) and error as a percentage of observed WUE (right vertical axis), with the heavy pink dashed lines representing +/- 100%. It shows that almost all LMs underestimate WUE, typically by about 50%, presumably related to the broad under-prediction of NEE by LMs evident in Fig. 3. At the other end of the spectrum, NoahMP shows a very high WUE bias, consistent with its overprediction of C uptake in Fig. 3 (due to a high dynamically predicted LAI). The empirical models, without any explicit constraint on the ratios of predicted variables (they are predicted independently), are better spread around observed values. Note that this statement applies equally to those empirical models trained on raw flux tower data and those trained on energy balance corrected data. Only the simplest empirical model – 1lin – shows 25th or 75th percentiles (across sites) outside 100% error in WUE, whereas most (8/14) LM do. The equivalent plot using energy balance corrected Qle data is shown in Fig. S9, and looks qualitatively similar to Fig. 4. For this metric, there are no discernible differences in performance across types of LM.

Deleted: (currently below RF_raw)

Deleted: 2

Formatted: Superscript

Formatted: Superscript



905 **Figure 4: Box plots of error in site mean ecosystem water use efficiency (-NEE/Qle) over all sites, shown separately for each model. WUE error is expressed both in units of umol of C gained per gram of water lost (left vertical axis, grey and multicoloured box plots) and error percentage of observed WUE (right vertical axis, pink box plots), with the heavy pink lines representing +/- 100%.**

As noted above, a range of alternative versions of the plots above are available in supplementary material, examining sensitivity to energy-
910 balance corrected data and low turbulence periods. Additional analyses, such as water evaporative fraction box plots (Figs. 8a and 8b) and variable density estimates for each model (Figs. S10a and S10b) are also in supplementary material.

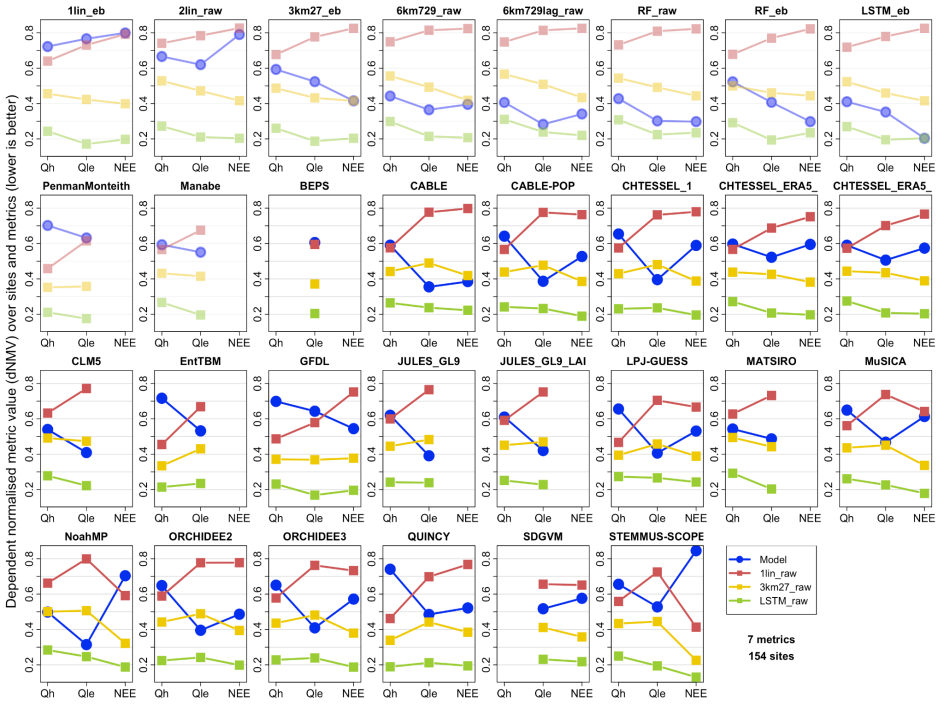


Figure 5: The average performance across all 154 sites and 7 metrics for Qh, Qle and NEE (lower is better). Average performance is the mean of dependent normalised metric values (dNMV) within the range of metric values across models being compared in each panel (4 in total, the LM (blue) - shown in plot title - and three reference benchmarks: 1lin_raw (red), 3km27_raw (yellow) and LSTM_raw (green)). The first 10 panels (faded) show empirical or physical benchmark models.

We now investigate a more direct comparison between LMs and empirical benchmarks by exploring results using our two summative indicators. Figure 5 shows modified 'PLUMBER plots', similar to Best et al. (2015), but here using the average of the dependent normalised metric values (dNMV) in the range of metric values across the four models being compared in each panel (one LM, 1lin_raw, 3km27_raw and LSTM_raw). This is as opposed to the average rank of metric values used in Best et al. (2015), which can distort results when metric values are clustered, as noted in Section 1. Each panel in Fig. 5 shows the model in the panel title in blue, with benchmark empirical models in red (1lin_raw), yellow (3km27_raw) and green (LSTM_raw). Lower values represent better performance. LMs are shown alphabetically, with the first 10 panels, faded, showing the remaining empirical models and physical benchmarks against these three benchmark models.

The out-of-sample LSTM_raw on average performs best across all fluxes, for these sites and metrics. The performance of LMs is highly variable, with half of them performing better than the 3km27_raw model for Qle, 15 of 18 worse than the out-of-sample simple linear regression (1lin_raw) for Qh, and NEE typically between the 1lin_raw and 3km27_raw levels of performance (12 of 14 LM variants). Overall, it is clear that LMs tend to perform better against the benchmarks for Qle and NEE than Qh, typically falling within the range of these three benchmarks for Qle and NEE. CLM5, MATSIRO and NoahMP are the only LMs with Qh metrics within this range. The LMs falling outside the benchmark dNMV ranges for Qle and NEE are a mixture of LSMs and ecosystem models. The equivalent plot using energy balance corrected Qle and Qh observations is

Deleted: 3

Deleted: 3

Deleted: 3

Deleted: 's

shown in Fig. S11a. The performance of the LMs against the benchmarks remains remarkably similar, with some LMs slightly better and others slightly worse against corrected data. Filtering for higher wind speed time steps (Fig. S11b, using raw flux data) also appears to make no qualitative difference, if anything making LM performance worse relative to these empirical benchmarks. While this may appear like a marked improvement in LM performance relative to results in Best et al (2015), these results are not directly comparable, something we explore further in the discussion section.

When we look at the same set of figures using *independent* normalised metric value (iNMV) instead of dependent (dNMV), the picture is very different (Fig. 6). Recall that iNMV sets the normalised metric range (0,1) based on the three reference out-of-sample empirical models (1lin, 3km27 and LSTM) only, rather than these three *and* the LM, and then compares the LM to this range. For example, if the three reference empirical models have a mean bias in Qle of 35Wm^{-2}, 28Wm^{-2}, and 25Wm^{-2} (a range of 10Wm^{-2}), and the LM has a bias of only 10Wm^{-2}, the iNMV of the reference models is 1, 0.3, and 0, respectively, and the LM has iNMV of -1.5 (remembering that lower is better). Alternatively, if the LM has a bias of 50Wm^{-2}, its iNMV would be 2.5. So iNMV values are not constrained to be in the unit interval, as they are for dNMV.

Figure 6 shows the same data as Fig. 5, but using iNMV instead of dNMV. The values of iNMV for the three reference models are now identical across all LM panels, so the values of iNMV for each LM are directly comparable. Note however that the vertical axis scale is different in each panel, so we can see the range for each LM. LM performance in iNMV clearly looks a lot worse. It tells us that when LMs perform worse than the out-of-sample linear response to shortwave, 1lin, they often perform a lot worse (at least a lot worse relative to the range between 1lin and LSTM_raw). While some LMs (CABLE, CABLE-POP, CHTESSEL, CLM, JULES, NoahMP and ORCHIDEE) perform within the range of the three empirical models for some variables, averaged over all variables, no LM outperforms the out-of-sample linear regression against SWdown. This is a sobering result. LM performance is particularly poor relative to the benchmarks for Qh with no models within the range of the benchmarks (compared to 40% of them for Qle and 29% for NEE).

Equivalent plots to Fig. 6 using energy-balance corrected fluxes (Fig. S11c) and time steps with wind speed > 2ms^{-2} (Fig. S11d) are shown in supplementary material. Again, LM performance appears remarkably similar despite the significant changes made with the target energy-balance corrected data (Fig. 1). It remains true that no LMs outperform the 1lin averaged over all fluxes. Note that in this comparison, where energy-balance corrected data are the reference target, the versions of empirical models trained for this target are used for comparison (i.e. 1lin_eb, 3km27_eb and LSTM_eb).

We also note that despite this result, some LMs do perform better than the empirical benchmarks for a subset of the metrics in Table S4 for some variables. Figs S11e – S11k are versions of Fig. 6 constructed with only one metric at a time. LMs tend to perform better in the 5th percentile and PDF overlap metrics, and worst in temporal correlation and NME. It is also apparent that RF, 6km729 and 6km729lag all outperform the LSTM in quite a few of these metrics. Despite this, we did not investigate alternatives to the LSTM as the high level empirical benchmark.

Deleted: 8
Deleted: 8
Deleted: 4
Deleted: 4
Deleted: 3
Deleted: 4
Deleted: 8
Deleted: 8
Deleted: 2
Deleted: 7
Deleted: -
Deleted: 7
Deleted: 4
Deleted: 's

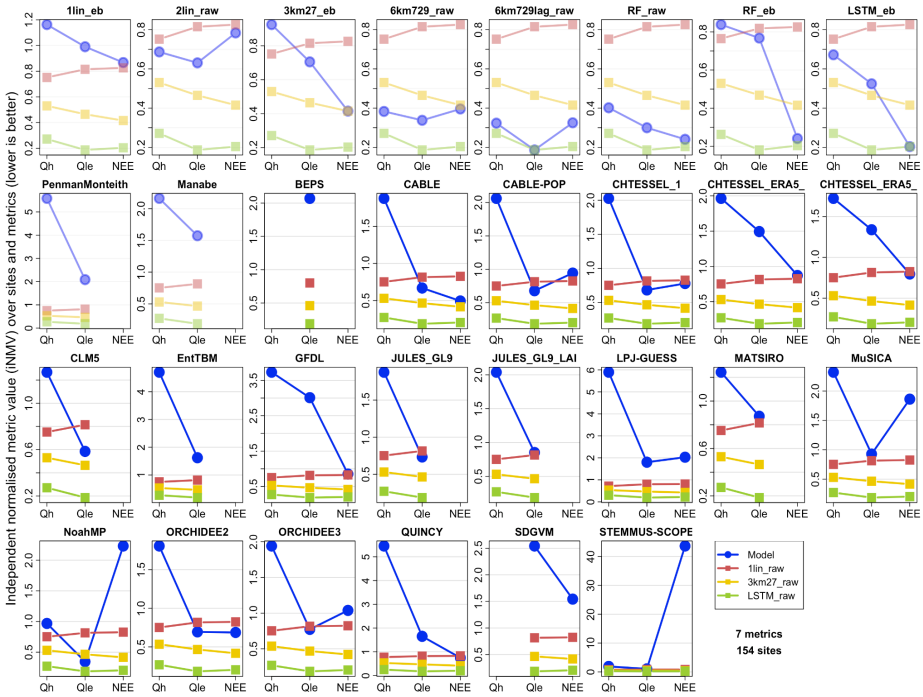


Figure 6: As per Fig. 5, but using the average of independent normalised metric values (iNMV) defined by the range of metric values across the three reference models (1lin_raw, 3km27_raw and LSTM_raw). Note that different panels have different y-axes.

We now examine the discrepancy between our best performing out-of-sample empirical models and a given mechanistic model in more detail. This defines an amount by which we *know* that the mechanistic model can improve by. This also allows us to define model performance in a way that accounts for site complexity / peculiarity / predictability, as well as observational errors particular to each site, and avoids some misleading statistics, like large RMSE values at sites that simply have larger fluxes. For this purpose, we use one of the best performing empirical model as the reference model, LSTM_raw. While it is the best performing models in this collection, it provides a *lower bound* estimate of predictability of fluxes at each site, since we can almost certainly produce better empirical models.

Deleted: 4

Deleted: 3

Moved up [6]: Figure 6 shows boxplots of error in the average energy evaporative fraction (EF) across the 154 sites, shown separately for each participating model. Energy evaporative fraction is defined using average flux values at each site: Qle / (Qle+Qh). The equivalent plots using energy balance corrected data and data filtered for wind speed are almost indistinguishable from Fig. 6, and so have not been included. In keeping with what we saw in Fig. 1, the mechanistic benchmarks and ecosystem models show the largest deviation from site observations, and empirical approaches are reliably zero-centred despite having no explicit mechanism to constrain the ratio between Qle and Qh. Once again, there does not appear to be any obvious reason to suspect a bias in partitioning in observations - some LMs (6) show a high EF bias, and others (11) a low bias.¶

¶

An equivalent version of this figure showing water evaporative fraction - Qle / Rainf - is shown in Fig. S10a and Fig. S10b (in supplementary material), using raw and energy-balance corrected fluxes, respectively. Once again, models are well scattered about the zero error line when raw fluxes are used, and almost all appear strongly negatively biased when compared to the energy balance corrected fluxes. The equivalent plots using wind speed filtered data are qualitatively the same as Figures S10a and S10b, and so are not included here.¶

Deleted: ¶

¶

Figure 5 shows how closely LMs approximate the density of observed values of Qle, Qh and NEE. The thick black line shows the observed density estimate of each variable. The thinner lines are the modelled density estimate minus the observed density estimat… [5]

Deleted: Figure 7 is similar to Fig. 6, but shows error in water-use efficiency (NEE/Qle), expressed in units of μmol of carbon gained per gram of water (left vertical axis) and error as a percentage of observed WUE (right vertical axis), with the heavy pink dashed lines representing +/- 100%. It shows that almost all LMs underestimate WUE, typically by about 50%, presumably related to the broa… [6]

Moved down [7]: While they are the best performing models in this collection, they provide a *lower bound* estimate of predictability of fluxes at each site, since we can almost certainly produce better empirical models.

Deleted: We now focus on utilising the better performing empirical models as a benchmark for the mechanistic models.

Deleted: The

Moved (insertion) [7]

Deleted: s

Deleted: they are

Deleted: they

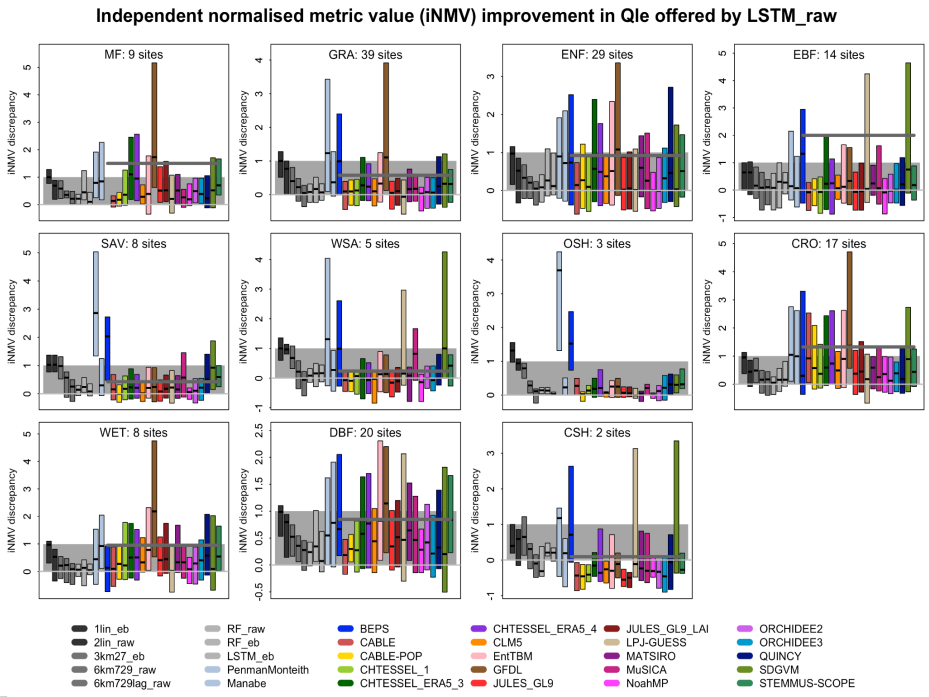


Figure 7: Independent normalised metric discrepancy between each model and LSTM_raw for latent heat flux (Qle), sorted by vegetation type. The average of all LMs for each vegetation type is shown by the bold dark grey line, and the zero line is in light grey. Lower scores are better.

In Fig. 7, we look at the discrepancy in iNMV between each mechanistic model and LSTM_raw for latent heat flux predictions. Results are shown in a separate panel for each IGBP vegetation type, and each model as a boxplot within each panel. Only the interquartile range and median estimates are shown for each boxplot. The observed vegetation types are used for each site, noting that some LMs with dynamic vegetation might represent these sites differently. Values below zero show that the LM performed better than the three benchmark empirical models (1lin, 3km27, LSTM). Values between zero and one mean that the LM performed within the range of the benchmark models (shaded grey background), and above one means that the LM was worse than 1lin. The average of all mechanistic LMs for each vegetation type is shown by the dark grey horizontal line, with the zero line in light grey. Each box plot represents the difference in independent normalised metric values across all metrics in Table S4.

There are clearly variations in performance across vegetation types, and while mean LM performance is worst for open shrubland (OSH), evergreen broadleaf forest (EBF) and mixed forest (MF), results across different LMs vary significantly. Overall, LM performance appears better for grass-dominated vegetation types (grassland and savannas) than tree ecosystems. The equivalent plots for Qh (Fig. S12b), using energy-balance-corrected data (Fig. S12a for Qle; S12c for Qh) and NEE (Fig. S12d) suggest that there is no clear differentiation of performance by vegetation type – no particular vegetation types is consistently anomalous. While some of the LM means (dark grey line) appear to change markedly for Qle after energy-balance correction (most notably for grassland sites), this seems at least partially because of significant changes to outlier LMs, rather

Moved (insertion) [9]
Deleted: First, i
Deleted: .
Deleted: 8
Deleted: ly
Deleted: and o
Deleted: Note that these are
Deleted: t
Deleted: and
Deleted: vegetation type
Deleted: s
Deleted: mechanistic model
Deleted: and number in the lower section of each panel
Deleted: 2
Moved (insertion) [8]
Deleted: 3
Deleted: 3
Deleted: 3
Deleted: 3
Deleted: eem to confirm the
Deleted: lack of
Deleted: ,
Deleted: and do not broadly show
Deleted: a
Deleted: note that
Deleted: largely

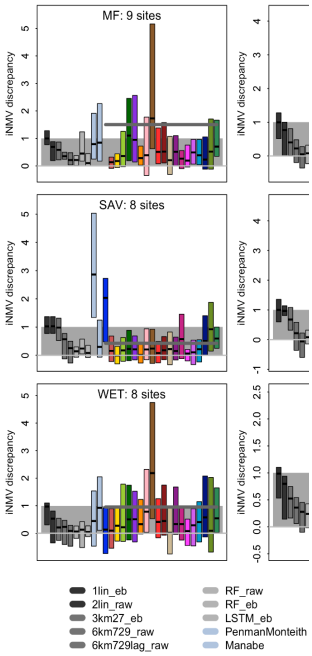
than a change in aggregate behaviour. Some LMs show improved performance using energy balance corrected data, others show degradation, although more appear to improve. There is definitely less of a change for Qh as a result of energy-balance correction. Also note that in all of these figures, the LM mean is often well above most of the 25th-75th percentile box plots. This simply reinforces the point made above that when LMs are worse than the reference benchmarks, they are often much worse (the smallest and largest 25% of values do not contribute to these box plots, obviously). Figures 13a – 13e show the same information, but sorted my model rather than vegetation type.

Finally, we examine LM performance in the context of the issue we raised in Section 2.2. A significant proportion of sites had Qle fluxes larger than incident rainfall, and since this is something that most LMs will be structurally prohibited from replicating, we explore why this might be the case, and whether the issue has biased our overall conclusions about LM performance. Figure 8 once again shows the iNMV improvement offered by the LSTM over LMs on a per-site basis, with the median difference for all LMs plotted (shown by colours). Each site's location is shown on axes of observed water evaporative fraction versus dryness index, as per Figures S2a-d. Note that the location of the 1-1 line relative to the sites is very much dependent upon our estimate of potential evapotranspiration (PET), which here is given by the Penman-Monteith model described above, so it is entirely plausible that a different estimate would see all sites (with the exception of US-Bkg) lying to the right of the 1-1 line. We might also wish to plot a curve on this figure illustrating the Budyko hypothesis (Budyko, 1974; although there is no single accepted derivation of an equation that describes the asymptotic behaviour it suggests; see Sposito, 2017; Mianabadi, 2019), but the spread of sites should make it clear why this is not particularly useful. Many sites have a water evaporative fraction above 1. This reinforces that the conceptual idealisation of the Budyko hypothesis applies only at very large spatial scales and/or in idealised circumstances of water availability. Irrigation, or landscape features like topography/hillslope, sub surface bedrock bathymetry or groundwater can mean it is entirely physically reasonable for a location to exhibit a water evaporative fraction above 1, as around 30% of these sites do. These factors are likely to still be relevant at scales of 10s of kilometres, so it seems unreasonable to suggest these effects are not also relevant for gridded simulations.

Deleted: is

Deleted: at

Moved up [8]: There are clearly variations across vegetation types, and while mean LM performance is worst for open shrubland (OSH), evergreen broadleaf forest (EBF) and mixed forest (MF), results across different LMs vary significantly. Overall, LM performance appears better for grass-dominated vegetation types (grassland and savannas) than tree ecosystems. The equivalent plots for Qh (Fig. S13b), using energy-balance-corrected data (Fig. S13a for Qle; S13c for Qh) and NEE (Fig. S13d) seem to confirm the lack of clear differentiation by vegetation type, and do not broadly show particular vegetation types as consistently anomalous. While some of the LM means (dark grey line) appear to change markedly for Qle after energy-balance correction (most notably for grassland sites), note that this seems largely because of significant changes to outlier



Moved up [9]:

Deleted: ¶
Figure S12a shows the equivalent to Fig. 8 using energy-balance corrected fluxes, and Figures S12b and S12c show the same for sensible heat flux, with raw and energy-balance corrected fluxes, respectively. As is apparent in Fig. 3, the LSTM comfortably ... [7]

Deleted: ¶
Figure 10: Independent normalised metric discrepancy between each model and LSTM_raw for latent heat flux (Qle), as per Fig. 7, but sorted by vegetation type. The average of all LMs fo ... [8]

Deleted: 11

Deleted: the same data -

Deleted: models

Deleted: -

Deleted: t'

Deleted: s

Deleted: n

Deleted: n

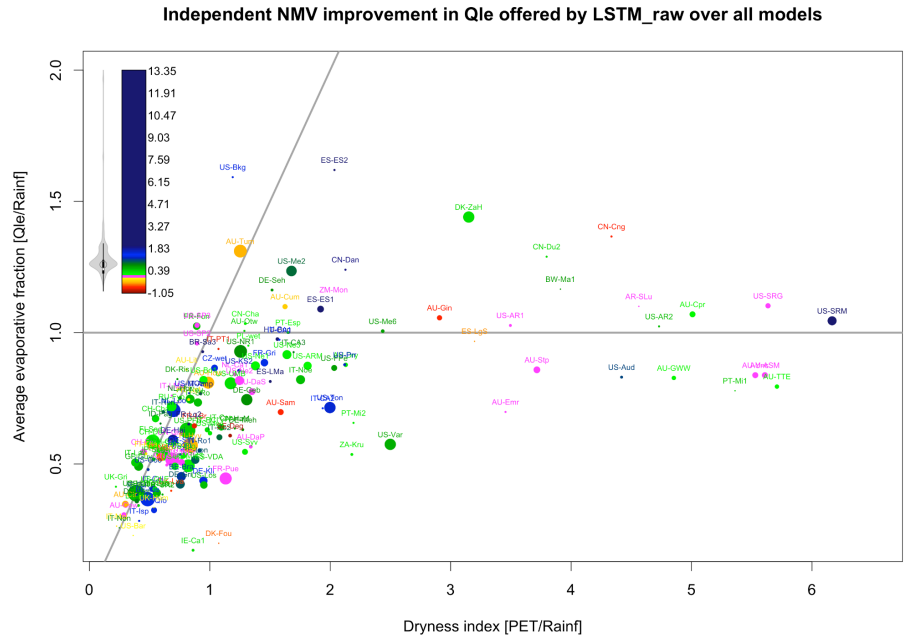


Figure 8: Independent normalised metric value (iNMV) improvement offered by LSTM over the median iNMV value of all LMs (excluding empirical and physical benchmarks), shown by colour for latent heat flux (Qle). Each site's location is shown on axes of observed water evaporative fraction versus dryness index. The prevalence of particular colour values is shown by the violin plot to the left of the colour legend. Values within [-0.1,0.1] are shown in pink, and values above 2 have constant, dark blue colour. Dot sizes indicate the length of site data, ranging from 1 (smallest) to 21 years (largest) - see Table S2 for site details.

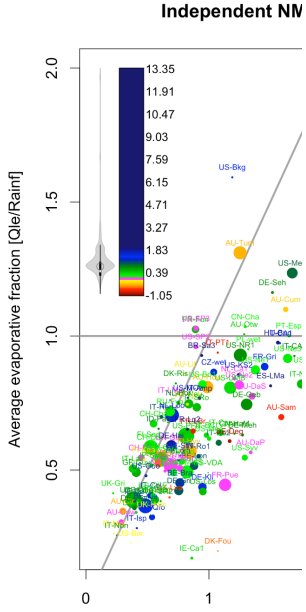
Of the sites in Fig. 8 with water evaporative fraction greater than 1, only one is irrigated (ES-ES2). Hillslope factors are quite plausibly important in four others (CN-Dan, DK-ZaH, US-SRG, US-SRM). One is affected by fire prior to the measurement period, which might mean that accumulated water was available (US-Me6). Others are sites from the La Thuile release not included in Fluxnet2015, which raises the possibility of data quality concerns (BW-Ma1, ES-ES1, RU-Zot, US-Bkg, US-SP3). But for the majority there is no immediately obvious explanation (AR-SLu, AU-Cpr, AU-Cum, AU-Gin, AU-Otw, AU-Tum, CN-Cha, CN-Cng, CN-Du2, DE-Seh, FR-Fon, US-AR1, US-AR2, US-Me2, ZM-Mon). While the data used in Fig. 8 is filtered for gap-filled and other quality control flags, we can confirm that using the entire time series for each site does not result in any qualitative change to site locations on this figure (see Fig. S2c).

The equivalent plots to Fig. 8 for corrected-Qle, Qh, corrected-Qh and NEE are shown in Figures S14a-d, respectively. None of these show a markedly higher density of poor LM performance (green-blue dots) above the 1.0 line where Qle exceeds precipitation on average. So despite there being a structural impediment to LMs simulating these sites, that impediment is clearly not the major cause of LM's poor performance. These figures also do not appear to support the community's heuristic expectation that LMs' performance decreases with dryness. While there is a cluster of energy-limited sites where LMs consistently outperform LSTM_raw (red-orange-yellow dots), there are also several water-limited sites where LMs do well, and the worst simulated sites by LMs, shown in blue, seem evenly spread throughout the figures.

Moved (insertion) [10]

Deleted: 11

Deleted: 11



Moved up [10]:

Deleted: ¶

¶

Deleted: Figure 11: Independent normalised metric value (iNMV) improvement offered by LSTM over the median iNMV value of all LMs (excluding empirical and physical benchmarks), shown by colour for latent heat flux (Qle). Each site's location is shown on axes of observed evaporative fraction versus dryness index. The prevalence of particular colour values is shown by the violin plot to the left of the colour legend. Values within [-0.1,0.1] are shown in pink, and values above 2 have constant, dark blue colour. Dot sizes indicate the length of site data, ranging from 1 (smallest) to 21 years (largest) - see Table S2 for site details.¶

Deleted: 11

Deleted: None of these

While it is clear that LSTM_raw broadly outperforms LMs at most sites, there are clearly some sites (red-orange-yellow) where LMs on aggregate outperform the LSTM. This does not appear to be the case consistently across all three fluxes for any particular site, however, or indeed any clear signal about the type of sites (in terms of vegetation type, dryness or available energy) that are better simulated. This probably suggests that these outcomes may be more stochastic than the result of any structural advantage the LMs might have.

4 Discussion and conclusions

In addressing our overall goal of fairly assessing the fidelity of land models, we aimed to create an evaluation framework that met two criteria: (1) a simulation environment that offered enough observational constraint to attribute model-observation mismatch to a model, where appropriate, and (2) a benchmarking approach within that environment that could ensure this attribution was fair by quantifying reasonable expectations of performance. Below we discuss the extent to which this was achieved, what we learned from applying the framework to the PLUMBER2 experimental results, caveats, and implications for future research.

Simulation environment, observational constraint and data quality

There are several important findings in terms of observational constraint. The first is that we can conclusively say that this simulation environment does offer enough observational constraint to diagnose model performance. The fact that a broad range of empirical models, trained at sites other than where they are tested, can reliably outperform LMs tells us that enough information is available to LMs to do better. What is less clear is whether the lack of constraint on LM parameter specification is causing poor LM performance – something we discuss in more detail below.

In terms of observational data quality, despite raw and energy-balance corrected fluxes clearly being very different, our conclusions about model performance are relatively similar with or without the correction. This qualitative similarity is only clear because of the ability to define performance expectations using empirical benchmarks separately in each case. Similarly, restricting analyses to low turbulence periods does not result in a qualitative change to the performance assessment of LMs, so it is clear that while there may well be issues with night time flux tower data, they are not the primary cause of the poor agreement between LMs and tower fluxes.

Next, the use of empirical benchmarks trained separately to predict raw and energy-balance corrected fluxes also gives us some insight into how appropriate the Fluxnet2015 energy balance correction process might be. Figure 1 and its equivalents in the supplementary material show that in general, available energy in LMs is indeed higher than in raw observations (noting that *a priori* this is not evidence that the observations are wrong). However, the energy balance corrected versions of this plot show an even larger discrepancy. Similarly, the differences between corrected and uncorrected water evaporative fraction (Figures S8a and S8b) show that corrected Qle fluxes look markedly different to almost all models. The plots based on iNMV do seem to show that the correction process helps improve overall performance for several LMs. There is, however, more subtle evidence in the performance of the empirical models that gives us other, contradictory information. LSTM_raw is the best performing reference model in Fig. 6, and as expected, LSTM_eb, trained to match qualitatively different (energy-balance corrected) target data, does not perform as well against raw flux data as LSTM_raw. This is what we would expect. However, when we look at the reverse situation, using LSTM_eb as the reference model, and energy-balance corrected fluxes as the target data (shown in Fig. S11c), the situation is quite different. LSTM_raw performs worse for Qh, as expected, but it performs better than LSTM_eb for Qle. This tells us that unlike for Qh, a sophisticated ML model trained on the corrected Qle flux has

Deleted: 's

Deleted: ¶
Versions of this plot for each individual submitted model (and indeed any future submission) are available on https://modelevaluation.org in the PLUMBER2 workspace.¶

Formatted: Font: Bold

Moved (insertion) [1]
Deleted: The question of whether
Deleted: is appropriate is not clear
Deleted: do seem to suggest
Deleted: although
Deleted: ,
Deleted: but
Formatted: Font: Italic
Deleted: 10
Deleted: 10
Deleted: Nevertheless t
Deleted: also
Deleted: 4
Deleted: 'd
Deleted: 8

no advantage predicting corrected Qle than the same ML model trained on raw fluxes - in fact it has a *disadvantage*. A similar result can be seen for 6km729lag. It is less sophisticated than LSTM_eb, and trained to predict the raw fluxes, yet it outperforms LSTM_eb. This suggests that the correction to Qle makes these fluxes *less* predictable. This suggests that the correction to Qle is categorically incorrect, whereas the correction to Qh may well add some value. This may suggest that the missing energy in uncorrected fluxes might be more likely to be in Qh fluxes (agreeing with other proposed correction approaches - see Charuchittipan et al., 2014).

Finally, we discuss the structural assumption in most LMs that horizontal transport of water between grid cells is negligible. A significant number of sites show a water evaporative fraction greater than 1, which, despite being entirely physically plausible, is simply not possible for most current LM process representations to replicate. It tells us that either (a) access to groundwater beyond gravity drainage is common, (b) below surface bedrock structure has a significant local hydrological effect, and/or (c) horizontal advection of moisture in soil (and locally on the surface) plays a significant role in moisture availability at the ~1km^2 spatial scale (i.e. flux tower fetch). Very few global coupled models include any of these effects. It is very likely that almost all sites and indeed much larger spatial scales are affected by this same issue to varying degrees, even if their water evaporative fractions do not appear to be anomalously high. This may well include all spatial scales below river and groundwater basin scales. Despite this revelation, it is remarkable that LM performance is apparently not any worse for sites where evapotranspiration exceeds precipitation (Figures 8, S14a-d). It suggests that despite this structural assumption being violated in the LMs here, other aspects of process representation are more detrimental to overall LM performance.

Benchmarking methodology

It should be clear that choices made in how we assess model performance can result in markedly different conclusions. The difference in apparent LM performance between dNMV and iNMV as a summative indicator is stark. By excluding the LM we are evaluating from the criteria that define good or bad performance (the set of the three empirical models) we define benchmark levels of performance that are independent of the LM being evaluated. It means that when the LM is *much* better, or *much* worse than a priori expectations, it will get a score that is proportionally much better or much worse. Using metric ranks or dNMV instead limits the cost of poor performance in the cumulative metrics shown in PLUMBER style plots (Best et al, 2015), and so gives an artificially positive indication of LM performance relative to the reference benchmark models.

We suggest that the framework we present provides a way to assess the significance of proposed improvements to LM performance that is relatively insensitive to metric choice, and critically, is based on demonstrated capacity for improvement. That is, when a LM is worse than an out-of-sample empirical model given the same predictors, we *know* that there is enough information provided to the LM to do better. We suggest that the summative analysis we present here using iNMV is a fairer, more comprehensive representation of LM performance than either the original PLUMBER paper or the dNMV versions of the same analyses.

Beyond a lower bound estimate of potential improvement, the hierarchy of empirical models we examined also provides more nuanced information about performance expectations. The difference in performance between 6km729 and 6km729lag, for example, quantifies the improvements in flux simulation we should expect from adding in model states such as soil moisture and temperature, rather than simply having an instantaneous response to meteorology (see Figures S13a-S13e). The same is also true of the RF and LSTM, although they had slightly different predictor sets and architectures. The simplest model, 1lin, also makes it clear that much of what we might heuristically regard as high model fidelity is a simple linear response to shortwave forcing (Figures S4, S5, S6 and perhaps most importantly Fig. 6). It should be abundantly clear that simple diagnostics can be very misleading and that defining 'good' model performance is inherently complicated. Without the

Deleted: It is likely

Deleted: The analyses also highlight the inherent complexity of real world simulation. We see a

Deleted: with

Deleted: n

Deleted: This is supported by

Deleted: the

Deleted: result

Deleted: 11

Moved (insertion) [2]

Deleted: The importance of

Deleted: metric

Deleted: choices

Deleted: becomes clear when we consider t

Deleted: 're

Deleted: land model

Deleted: ed

Deleted: a

Deleted: While this might be confronting, it

Deleted: provides the community with a

Deleted: with which

Deleted: ,

Deleted: in a way

empirical model hierarchy detailed here, judgements about LM performance would almost certainly be susceptible to confirmation bias.

PLUMBER, PLUMBER2 and implications for LMs

It might appear from Fig. 5 that many LMs (CABLE, CHTESSEL, CLM, JULES, MATSIRO, MuSICA, ORCHIDEE, NoahMP) perform better than the 3km27 model here for Qle, something that could represent progress since the original PLUMBER experiment (where no models outperformed the 3km27 model for standard metrics - see Best et al, 2015). There are however some differences here that mean PLUMBER and PLUMBER2 results are not directly comparable. First, the single set of metrics we are using here is a combination of the 'standard', 'distribution' and 'extremes' based metrics used in PLUMBER, and the worst LM performance in PLUMBER was for the standard metrics set alone. Next, Fig. 5 uses (dependent) normalised metric range, rather than ranks. We also have fewer models, and different models, in each panel that is used to calculate the metric range, and results are calculated over 154 instead of 20 sites. It nevertheless remains true that Qh is much more poorly predicted than Qle.

While of a similar performance standard to Qle prediction overall, NEE was notably underpredicted by LMs in a way that Qle was not. While it seems obvious that a lack of site history in LM setup (noting that this information was not available) is the cause for this, it is intriguing to see that empirical models (also not given this information) were able to predict NEE without this bias, in most cases without any LAI information at all (Figs. 3, 4). These empirical models were out-of-sample (they did not use any data from the sites they predicted in their training). This is an indication that importance of site history and leaf area is overstated in our LMs, and not as important as we may believe for flux prediction.

These results raise the question of whether LMs are too complex for the level of fidelity they provide. It is at least theoretically possible, for example, that an LM is perfect, but because we are unable to precisely prescribe its parameters for these site simulations (and global simulations) we are actively hindering its ability to get the right result. What the out-of-sample empirical models show is that the information available in LMs' meteorological variables *alone* - without any description of what type of vegetation or soil might be at a given site, or indeed the reference height of the measurements - is enough to outperform all of the LMs. This is not to say that LMs could not perform better if more detailed site-specific information were available, but the way that they were run here was designed to mimic their application at global scales, and for that job they are considerably more complicated than is justified by their performance. A more detailed examination of how well LMs perform when given detailed site information would not simply require showing that metric scores for LMs improved when given this information, it would require that LMs come closer to outperforming ML approaches also provided with similar site-specific information.

There are of course other reasons why we might want complexity in a LM beyond improved performance, like the ability to infer the impacts of particular decisions on a broader range of processes within the land system. But it is nevertheless important to know the degree of predictability that's possible with the increasing amount of information that our models are provided with - what we're missing out on that is categorically achievable. The fact that it has been found that increasing model complexity shows little relationship to performance in some circumstances, even when additional site information is provided, should be concerning (Lipson et al, 2023). We also need to recognise that the many increases in sophistication that we might want to include to improve the representativeness of LMs (see Fisher and Koven, 2020) may come at a significant cost. The more degrees of freedom we have in a model, the more and broader range of observational data we need to effectively constrain it, the less able we are to pinpoint model shortcomings, and the more susceptible we become to getting the right answer for the wrong reasons (see Lenhard and Winsberg, 2010). A very crude statistical analogy might be that if we have a model with

Formatted: Font: Bold
Deleted: It should be clear from the results above that the metric choices we make when evaluating LMs are critical.
Deleted: 3
Deleted: 're
Deleted: infamous poor
Deleted: 3
Deleted: 's
Deleted: 2
Deleted: categorical
Deleted: significantly
Deleted: nearly
Deleted: it is
Moved (insertion) [3]
Deleted: is
Deleted: s
Deleted: 's
Deleted: with
Deleted: of course
Deleted: 's
Deleted: 's

one process that is right 90% of the time, the model is 90% accurate. But if we have a model with 10 serial processes that are right 90% of the time, the model is $0.9^{10} = 35\%$ accurate.

This of course does not mean that LMs will always appear to perform badly in global scale studies, especially if performance expectations are not quantified the way we have done here. Figures S4, S5 and S6 show that we can explain a considerable amount of observed variability with very simple models, and examining results at longer timescales as is typically the case in global studies will not change this. We are only able to draw the conclusions we have here because we have clearly defined performance expectations in terms of the amount of information available to LMs about surface flux prediction, and examined this close to the process scale, rather than averaged over longer periods and spatial scales.

Next steps

As with most model comparisons, the summary statistics presented in this paper do not give us any categorical indications about how to start improving models. They nevertheless allow, perhaps for the first time, to fairly account for some of the inevitable difficulties and eccentricities associated with using observed data. By evaluating performance relative to out-of-sample empirical estimates we can quantify expectations of achievable LM improvement, and isolate the circumstances in which this potential for improvement is most apparent. We did not actively explore these circumstances in detail here, since they are particular to each LM, but have nevertheless provided an approach to achieve this. Some clear indications are already evident from the sites shown in green and particularly blue in Figures 8, S14a,b,c,d. These are sites where we *know* that LM model prediction can be substantially improved, since an out-of-sample empirical model offers substantial performance improvements using the same predictors as the LMs. These are of course the *average* discrepancy across all LMs, so the capacity for improvement at a particular site is likely to vary for different models. Equivalent figures for each individual model and variable can be found on modelevaluation.org in the PLUMBER2 workspace via the profile page for each submitted model output. Data and analysis code from this experiment are also available and we openly invite further analyses and contributions from the community.

The next steps for the community towards building LMs that better utilise the information available to them seem reasonably clear. Understanding the shortcomings of an LM is not a simple process, so moving away from in-house, ad-hoc model evaluation towards more comprehensive, community built evaluation tools, where the efforts of those invested in model evaluation are available to everyone will be key. This will allow results to be comparable across institutions and routine automated testing to become part of the model development cycle. This will need to cover both global scales (e.g. ILAMB; Hoffman et al., 2016; Collier et al, 2018) and site-based process evaluation (e.g. modelevaluation.org; Abramowitz, 2012). In both cases, inclusion of empirical performance estimates, such as those shown here, will be key to distinguishing incremental improvements from qualitative improvements in LM performance.

Finally, there is obviously much, much more to explore in the PLUMBER2 dataset. Most participants submitted many more variables than were examined in this paper (and several came close to the list in Table S2). The vast majority of submissions to PLUMBER2, as well the forcing and evaluation data, are publicly available on https://modelevaluation.org as a community resource for further analyses, and we actively invite further collaborations to utilise the data set that this experiment has produced. This paper nevertheless provides the community with a benchmarking framework that is relatively insensitive to observational errors, choices in evaluation metrics, and defines model performance in terms of demonstrated capacity for improvement, rather than model-observation mismatch alone.

Deleted: ¶

Moved up [2]: The importance of metric choices becomes clear when we consider the difference in apparent LM performance between dNMV and iNMV. By excluding the LM we're evaluating from the criteria that define good or bad performance (the set of the three empirical models) we define benchmark levels of performance that are independent of the land model being evaluated. It means that when the LM is *much* better, or *much* worse than a priori expectations, it will get a score that is proportionally much better or much worse. Using metric ranks or dNMV instead limited the cost of poor performance in the cumulative metrics shown in PLUMBER style plots, and so gave an artificially positive indication of LM performance relative to the reference benchmark models.¶
¶
While this might be confronting, it provides the community with a framework with which to assess the significance of proposed improvements to LM performance, in a way that is relatively insensitive to metric choice, and critically, is based on demonstrated capacity for improvement. That is, when a LM is worse than an out-of-sample empirical model given the same predictors, we *know* that there is enough information provided to the LM to do better. We suggest that the summative analysis we present here using iNMV is a fairer, more comprehensive representation of LM performance than either the original PLUMBER paper or the dNMV versions of the same analyses. ¶

Deleted: It's also clear that, as

Deleted: actually begin to

Deleted: pinpoint

Deleted: Given the number of LMs involved w

Deleted: 11

Deleted: 3

Moved up [1]: The question of whether the Fluxnet2015 energy balance correction process is appropriate is not clear. Figure 1 and its equivalents in the supplementary material do seem to suggest that available energy in LMs is indeed higher than in raw observations (although a priori this is not evidence that the observations are wrong), but the energy balance corrected versions of this plot show an even larger discrepancy. Similarly, the differences between corrected and uncorrected water evaporative fraction (Figures S10a and S10b) show that corrected Qle fluxes look markedly different to almost all models. The plots based on iNMV do seem to show that the correction process helps improve performance for several LMs. Nevertheless there is also more subtle evidence in the performance of the empirical models that gives us other, contradictory information. LSTM_raw is the best performing reference model in Fig. 4, and as expected, LSTM_eb, trained to match qualitatively different (energy-

Moved up [3]: This raises the question of whether LMs are too of fidelity they provide. It's at least theoretically possible, for example, that an LM is perfect, but because we are unable to precisely prescribe its parameters for these site simulations (and

Deleted: Beyond a lower bound estimate of potential improvement, the hierarchy of empirical models we examined can also provide more nuanced information about performance expectations. The difference in performance between 6km729 and 6km729lag, f... [9]

Deleted: ¶
The analyses also highlight the inherent complexity of real world simulation. We see a significant number of sites with an evaporative fraction greater than 1, which, despite being entirely physical... [10]

Deleted: ¶
We are also building the suite of analyses shown in this paper into a model development testing pipeline available via https://modelevaluation.org. This will allow the community ... [11]

Moved (insertion) [4]

Field Code Changed

Moved up [4]: The vast majority of submissions to PLUMBER2, as well the forcing and evaluation data, are publicly available on https://modelevaluation.org as a community resource for further

Deleted: a broad overview of the experimental design and preliminary results for key atmospheric fluxes. The vast majority of submissions to PLUMBER2, as well the forcing and evaluation data, are publicly available on https://modelevaluation.org as a co... [12]

Code and data availability

Flux tower data used here are available at http://dx.doi.org/10.25914/5fdb0902607e1 as per Ukkola et al. (2022), 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 ECMWF and 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 and ICOS Ecosystem Thematic Center, and the OzFlux, ChinaFlux and AsiaFlux offices. All land model simulations in this experiment are hosted in modelevaluation.org, and to the extent that participants had no legal barriers to sharing these, are available after registering with modelevaluation.org. The analyses shown here were also performed on modelevaluation.org, using the codebase publicly available at https://gitlab.com/modelevaluation/me.org-r-library.

Author contribution: Experimental design was conceived by GA with assistance from ML, MDK, AU, MC and wider community feedback. Data processing and analysis was developed and completed by GA, with input from ML, MDK, AU, and JCP. Empirical model simulations were completed by JCP, SH, GA, CB, JF and GN. Physical benchmark models were build and run by MB and HR. Mechanistic land model simulations were completed by MDK, AU, JK, XL, DF, SB, GB, KO, DL, XW-F, CO, PP, NV, HR, MB, SM, TN, HK, YZ, YW, BS, YK, KC, AW, PA, MC, JO, SC, SZ and CF. The manuscript was written by GA, with feedback and iterations between all coauthors.

Competing interests: The authors declare that they have no conflict of interest.

Acknowledgements

G.A., S.H., J.C.P., A.U. and M. dK. acknowledge the support of the Australian Research Council Centre of Excellence for Climate Extremes (CE170100023). A.U. acknowledges support from the ARC Discovery Early Career Researcher Award (DE200100086). This research was undertaken with the assistance of resources and services from the National Computational Infrastructure (NCI), which is supported by the Australian Government. B.S., Y.W. and Y.Z. acknowledge the support of the Netherlands Organisation for Scientific Research (NWO) (WUNDER project, grant no. KICH1. LWV02.20.004), and the Netherlands eScience Center (EcoExtreML project, grant ID. 27020G07). JF acknowledges the support of NOAA Cooperative Agreement, Grant/Award Number: NA19NES4320002. Contributions by K.O. and D.L. are supported by the National Center for Atmospheric Research (NCAR), sponsored by the National Science Foundation (NSF) under Cooperative Agreement No. 1852977. Computing and data storage resources for CLM5, including the Cheyenne supercomputer (doi:10.5065/D6RX99HX), were provided by the Computational and Information Systems Laboratory (CISL) at NCAR. LSTM models were run with compute resources provided by NASA Terrestrial Hydrology Program, Grant/Award Number: 80NSSC18K0982. Noah-MP simulations were funded by NASA grant 80NSSC21K1731.

References

Abramowitz, G.: Towards a benchmark for land surface models, *Geophys. Res. Lett.*, 32, L22702, 2005.

Abramowitz, G.: Towards a public, standardized, diagnostic benchmarking system for land surface models, Geosci. Model Dev., vol. 5, pp. 819 - 827, http://dx.doi.org/10.5194/gmd-5-819-2012, 2012.

Deleted: (2005)
Deleted: *ical*
Deleted: *earch*
Deleted: *ers*
Deleted: (2012) '
Deleted: '
Deleted: entific
Deleted: elopment

Abramowitz, G., Pouyanné, L. and Ajami, H.: On the information content of surface meteorology for downward atmospheric long-wave radiation synthesis, *Geophys. Res. Lett.* 39, L04808, doi:10.1029/2011GL050726, 2012.

Abramowitz, G., Herger, N., Gutmann, E., Hammerling, D., Knutti, R., Leduc, M., Lorenz, R., Pincus, R., and Schmidt, G. A.: ESD Reviews: Model dependence in multi-model climate ensembles: weighting, sub-selection and out-of-sample testing, Earth Syst. Dynam., 10, 91–105, https://doi.org/10.5194/esd-10-91-2019, 2019.

Arora, V. K., Seiler, C., Wang, L., and Kou-Giesbrecht, S.: Towards an ensemble-based evaluation of land surface models in light of uncertain forcings and observations, Biogeosciences, 20, 1313–1355, https://doi.org/10.5194/bg-20-1313-2023, 2023.

Aubinet, M., Feigenwinter, C., Heinesch, B., Laffineur, Q., Papale, D., Reichstein, M., Rinne, J., and Van Gorsel, E.: Nighttime Flux Correction, in: Eddy Covariance: A Practical Guide to Measurement and Data Analysis, edited by: Aubinet, M., Vesala, T., and Papale, D., Springer Netherlands, Dordrecht, available at: http: //link.springer.com/10.1007/978-94-007-2351-1, 2012.

Balsamo, G., Viterbo, P., Beljaars, A., van den Hurk, B., Hirschi, M., Betts, A. K., and Scipal, 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, 2009.

Beaudry, N. and Renner, R.: An intuitive proof of the data processing inequality, Quantum Information & Computation, 12 (5–6): 432–441, 2012.

Bennett, A. C., Knauer, J., Bennett, L. T., Haverd, V., & Arndt, S. K.: Variable influence of photosynthetic thermal acclimation on future carbon uptake in Australian wooded ecosystems under climate change. Glob. Change Biol., 30, e17021, https://doi.org/10.1111/gcb.17021, 2024.

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., Dong, J., Ek, M., Guo, Z., Haverd, V., van den Hurk, B. J., Nearing, G. S., Pak, B., Peters-Lidard, C., Santanello, J., Stevens, L. and Vuichard, N.: The plumbing of land surface models: benchmarking model performance, *J. Hydrometeorol.* 16, 1425–42, doi:10.1175/JHM-D-14-0158.1, 2015.

Bi, D., Dix, M., Marsland, S., O'Farrell, S., Sullivan, A., Bodman, R., Law, R., Harman, I., Srbinovsky, J., Rashid, H. A., Dobrohotoff, P., Mackallah, C., Yan, H., Hirst, A., Savita, A., Dias, F. B., Woodhouse, M., Fiedler, R., Heerdegen, A.: Configuration and spin-up of ACCESS-CM2, the new generation Australian Community Climate and Earth System Simulator Coupled Model, J. South. Hemisph. Earth Syst. Sci. 70, 225–251, 2020.

Bonan, G. B., Patton, E. G., Finnigan, J. J., Baldocchi, D. D., and Harman, I. N.: Moving beyond the incorrect but useful

Deleted: L.
Deleted: H.
Deleted: (2012),
Deleted: *ical*
Deleted: *earch*
Deleted: *ers*
Deleted: .
Deleted: (2012),
Formatted: Font: Not Italic
Formatted: Font: Not Italic
Formatted: Font: Not Italic
Deleted: a
Deleted:
Deleted: a
Deleted: 2015
Deleted: *ournal of*
Deleted: *ogy*
Deleted: ¶
Bishop CH and Abramowitz G, 2013, 'Climate model dependence and the replicate Earth paradigm', Climate Dynamics, vol. 41, pp. 885 - 900, http://dx.doi.org/10.1007/s00382-012-1610-y¶

paradigm: reevaluating big-leaf and multilayer plant canopies to model biosphere-atmosphere fluxes – a review. Agr. Forest Meteorol., 306, 108435. doi:10.1016/j.agrformet.2021.108435, 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., 118, 1–24, https://doi.org/10.1002/jgrd.50488, 2013.

Budyko, M. I.: Climate and life (p. 508). New York: Academic Press, 1974.

Buechel, M.: Understanding hydrological change with land surface models, Nat. Rev. Earth. Environ., **2**, 824, https://doi.org/10.1038/s43017-021-00241-0, 2021.

Bush, M., Boutle, I., Edwards, J., Finnenkoetter, A., Franklin, C., Hanley, K., Jayakumar, A., Lewis, H., Lock, A., Mittermaier, M., Mohandas, S., North, R., Porson, A., Roux, B., Webster, S., and Weeks, M.: The second Met Office Unified Model–JULES Regional Atmosphere and Land configuration, RAL2, Geosci. Model Dev., 16, 1713–1734, https://doi.org/10.5194/gmd-16-1713-2023, 2023.

Charuchittipan, D., Babel, W., Mauder, M., Leps, J.-P., & Foken, T.: Extension of the averaging time in Eddy-covariance measurements and its effect on the energy balance closure. Bound-Layer Meteorol., 152(3), 303–327. https://doi.org/10.1007/s10546-014-9922-6, 2014.

Chen, X., & Sivapalan, M.: Hydrological basis of the Budyko curve: Data-guided exploration of the mediating role of soil moisture, Water Resour. Res., 56, e2020WR028221. https://doi.org/10.1029/2020WR028221, 2020.

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. M., 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, 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.

Clark, M. P., Nijssen, B., Lundquist, J., Kavetski, D., Rupp, D., Woods, R., Gutmann, E., Wood, A., Brekke, L., Arnold, J., Gochis, D., and Rasmussen, R.: A unified approach to process-based hydrologic modeling. Part 1: Modeling concept. Water Resour. Res., 51, doi: 10.1002/2015WR017198, 2015a.

Deleted: (1974).

Deleted: .

Deleted: Chang, K.Y., Paw, U., K.T., Chen, S.-H., 2018. Canopy profile sensitivity on surface layer simulations evaluated by a multiple canopy layer higher order closure land surface model. Agric. For. Meteorol. 252, 192–207. https://doi.org/10.1016/j.agrformet.2018.01.027.¶

Deleted: (2014).

Deleted: .

Deleted: ary

Deleted: ogy

Deleted: (2020).

Deleted: .

Deleted: ces

Deleted: earch

Formatted: Font: Not Italic

Formatted: Font: Not Italic

Formatted: Font: Not Italic

Deleted: B.

Deleted: J.

Deleted: D.

Deleted: R.

Deleted: E.

Deleted: A.

Deleted: L.

Deleted: J.

Deleted: D.

Deleted: R.

Deleted: , 2015a

Deleted: ces

Deleted: earch

Deleted: .

Clark, M. P., Nijssen, B., Lundquist, J., Kavetski, D., Rupp, D., Woods, R., Gutmann, E., Wood, A., Gochis, D., Rasmussen, R., Tarboton, D., Mahat, V., Flerchinger, G., and Marks, D.: A unified approach for process-based hydrologic modeling: Part 2. Model implementation and example applications. Water Resour. Res., 51, doi: 10.1002/2015WR017200, 2015b.

Collier, N., Hoffman, F. M., Lawrence, D. M., Keppel-Aleks, G., Koven, C. D., Riley, W. J., Mu, M., Randerson, J. T.: The International Land Model Benchmarking (ILAMB) system: Design, theory, and implementation. J. Adv. Model. Earth Syst., 10, 2731–2754. https://doi.org/10.1029/2018MS001354, 2018.

Contractor, S., Donat, M. G., Alexander, L. V., Ziese, M., Meyer-Christoffer, A., Schneider, U., Rustemeier, E., Becker, A., Durre, I., and Vose, R. S.: Rainfall Estimates on a Gridded Network (REGEN) – a global land-based gridded dataset of daily precipitation from 1950 to 2016, Hydrol. Earth Syst. Sci., 24, 919–943, https://doi.org/10.5194/hess-24-919-2020, 2020.

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., 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 Syst., 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.

Falge, E., M. Aubinet, P.S. Bakwin, D. Baldocchi, P. Berbigier, C. Bernhofer, T.A. Black, R. Ceulemans, K.J. Davis, A.J. Dolman, A. Goldstein, M.L. Goulden, A. Granier, D.Y. Hollinger, P.G. Jarvis, N. Jensen, K. Pilegaard, G. Katul, P. Kyaw Tha Paw, B.E. Law, A. Lindroth, D. Loustau, Y. Mahli, R. Monson, P. Moncrieff, E. Moors, J.W. Munger, T. Meyers, W. Oechel, E.-D. Schulze, H. Thorgeirsson, J. Tenhunen, R. Valentini, S.B. Verma, T. Vesala, and S.C. Wofsy. 2017. FLUXNET Research Network Site Characteristics, Investigators, and Bibliography, 2016. ORNL DAAC, Oak Ridge, Tennessee, USA. https://doi.org/10.3334/ORNLDAAC/1530

Fisher, R. A., & Koven, C. D.: Perspectives on the future of land surface models and the challenges of representing complex terrestrial systems. J. Adv. Model. Earth Sys., 12, e2018MS001453. https://doi.org/10.1029/2018MS001453, 2020.

Gennaretti, F., Ogee, J., Sainte-Marie, J. and Cuntz, M.: Mining ecophysiological responses of European beech ecosystems to drought, Agr. Forest Meteorol., 280, 107780, doi: 10.1016/j.agrformet.2019.107780, 2020.

Goulden, M. L., Munger, J. W., Fan, S.-M., Daube, B. C., and Wofsy, S. C.: Measurements of carbon sequestration by long-term eddy covariance: methods and a critical evaluation of accuracy, Glob. Change Biol., 2, 169–182, https://doi.org/10.1111/j.1365-2486.1996.tb00070.x, 1996.

Deleted: Clark, M. P., B. Nijssen, J. Lundquist, D. Kavetski, D. Rupp, R. Woods, E. Gutmann, A. Wood, D. …ochis, D., R. Rasmussen, R., D. …arboton, D., V. …ahat, V., G. …lerchin … [13]
Deleted: et al. (2018).…The International Land Model … [14]
Moved down [11]: L. W.
Moved (insertion) [11]
Moved down [12]: A. J.
Moved (insertion) [12]
Deleted: P.
Moved down [13]: I. M.
Moved (insertion) [13]
Moved down [14]: J. G.
Moved down [15]: J. P.
Moved (insertion) [14]
Moved (insertion) [15]
Deleted: S. …alyshev, S., V. …aik, V., F. …aulot, F., E. … [15]
Moved down [16]: C. A.
Moved down [17]: K. A.
Moved (insertion) [16]
Moved (insertion) [17]
Deleted: N. …adeh, N., V. …alaji, V., C. … [16]
Deleted: C. …upuis, C., J. …urachta, J., R. … [17]
Moved down [18]: P. P. G.
Moved down [19]: S. M.
Moved down [20]: R. W.
Moved (insertion) [18]
Moved (insertion) [19]
Moved (insertion) [20]
Deleted: H.
Deleted: M. …arrison, M., J. …e, J., W. …urlin, W., C. … [18]
Moved down [21]: P. C. D.
Moved (insertion) [21]
Deleted: S.
Moved down [22]: D. J.
Moved (insertion) [22]
Deleted: J. …loshay, J., A. …adhakrishnan, A., K. … [19]
Moved down [23]: B. G.
Moved (insertion) [23]
Deleted: T.
Moved down [24]: D. M.
Moved (insertion) [24]
Moved down [25]: L. T.
Moved (insertion) [25]
Deleted: S. …nderwood, S., H. …ahlenkamp, H. ,M. … [20]
Moved down [26]: A. T.
Moved (insertion) [26]
Deleted: B. …yman, B., Y. …eng, Y., M. …hao, M. (202… [21]
Formatted: Font: Not Italic
Deleted: ournal… of …dv.ances…in …odeling…Earth Sy… [22]
Deleted: (2020).……Perspectives on the future of land su… [23]
Deleted: J. …gee, J., J.…ainte-Marie, J. and M. …untz, M… [24]

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. J. Hydrol., 377(1-2), 80-91. https://doi.org/10.1016/j.jhydrol.2009.08.003, 2009.

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., 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, 2018.

Haverd, V., Smith, B., Cook, G., Briggs, P., Nieradzik, L., Roxburgh, S., Liedloff, A., Meyer, C., and Canadell, J.: A stand-alone tree demography and landscape structure module for Earth system models. Geophys. Res. Lett. 40(19), 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.

Hengl, T., Mendes de Jesus, J., Heuvelink, G. B. M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S., and Kempen, B.: SoilGrids250m: Global gridded soil information based on machine learning, PLoS ONE, 12, e0169748, https://doi.org/10.1371/journal.pone.0169748, 2017b.

Hoffman, F. M., C. D. Koven, G. Keppel-Aleks, D. M. Lawrence, W. J. Riley, J. T. Randerson, A. Ahlström, G. Abramowitz, D. D. Baldocchi, M. J. Best, B. Bond-Lamberty, M. G. De Kauwe, A. S. Denning, A. Desai, V. Eyring, J. B. Fisher, R. A. Fisher, P. J. Gleckler, M. Huang, G. Hugelius, A. K. Jain, N. Y. Kiang, H. Kim, R. D. Koster, S. V. Kumar, H. Li, Y. Luo, J. Mao, N. G. McDowell, U. Mishra, P. R. Moorcroft, G. S. H. Pau, D. M. Ricciuto, K. Schaefer, C. R. Schwalm, S. P. Serbin, E. Shevliakova, A. G. Slater, J. Tang, M. Williams, J. Xia, C. Xu, R. Joseph, and D. Koch (2017), International Land Model Benchmarking (ILAMB) 2016 Workshop Report, DOE/SC-0186, U.S. Department of Energy, Office of Science, Germantown, Maryland, USA, doi:10.2172/1330803.

Formatted: Font: Not Italic

Formatted: Font: Not Italic

Formatted: Font: Not Italic

Deleted: &

Deleted: ¶
(2013)

Deleted: .

Deleted: ical

Deleted: earch

Deleted: ers

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

Deleted: ¶
¶
Kashinath K., Mustafa M., Albert A., Wu J-L., Jiang C., Esmaeilzadeh S., Azizzadenesheli K., Wang R., Chattopadhyay A., Singh A., Manepalli A., Chirila D., Yu R., Walters R., White B., Xiao H., Tchelepi H. A., Marcus P., Anandkumar A., Hassanzadeh P. and Prabhat 2021Physics-informed machine learning: case studies for weather and climate modelling Phil. Trans. R. Soc. A.379202009320200093 http://doi.org/10.1098/rsta.2020.0093¶

Kim, Y., P.R. Moorcroft, I. Aleinov, M.J. Puma, and N.Y. Kiang, 2015: Variability of phenology and fluxes of water and carbon with observed and simulated soil moisture in the Ent Terrestrial Biosphere Model (Ent TBM version 1.0.1.0.0). Geosci. Model Dev., doi:10.5194/gmd-8-3837-2015.

Knauer, J., Cuntz, M., Smith, B., Canadell, J. G., Medlyn, B. E., Bennett, A. C., Caldararu, S., Haverd, V.: Higher global gross primary productivity under future climate with more advanced representations of photosynthesis. Sci. Adv., 9(46). https://doi.org/10.1126/SCIADV.ADH9444, 2023

Kowalczyk, E. A., Y. P. Wang, R. M. Law, H. L. Davies, J. L. McGregor, and G. S. Abramowitz, 2006: The CSIRO Atmosphere Biosphere Land Exchange (CABLE) model for use in climate models and as an offline model. CSIRO Ma- rine and Atmospheric Research Paper 013, 43 pp. [Available online at www.cawcr.gov.au/projects/access/cable/cable_technical_description.pdf.]

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, GB1015, doi:10.1029/2003GB002199, 2005.

Kumar, S. V., Peters-Lidard, C. D., Tian, Y., Houser, P. R., Geiger, J., Olden, S., Lighty, L., Eastman, J. L., Doty, B., Dirmeyer, P., Adams, J., Mitchell, K., Wood, E. F. and Sheffield, J.: Land information system: An interoperable framework for high resolution land surface modeling. Environ. Modell. Software, 21, 1402–1415, https://doi.org/10.1016/j.envsoft.2005.07.004, 2006.

Lawrence, David M.; Fisher, Rosie A.; Koven, Charles D.; Oleson, Keith W.; Swenson, Sean C.; Bonan, Gordon; Collier, Nathan; Ghimire, Bardan; Kampenhout, Leo; Kennedy, Daniel; Kluzek, Erik; Lawrence, Peter J.; Li, Fang; Li, Hongyi; Lombardozzi, Danica; Riley, William J.; Sacks, William J.; Shi, Mingjie; Vertenstein, Mariana; Wieder, William R.; Xu, Chonggang; Ali, Ashehad A.; Badger, Andrew M.; Bisht, Gautam; Broeke, Michiel; Brunke, Michael A.; Burns, Sean P.; Buzan, Jonathan; Clark, Martyn; Craig, Anthony; Dahlin, Kyla; Drewniak, Beth; Fisher, Joshua B.; Flanner, Mark; Fox, Andrew M.; Gentine, Pierre; Hoffman, Forrest; Keppel-aleks, Gretchen ; Knox, Ryan; Kumar, Sanjiv; Lenaerts, Jan; Leung, L. Ruby; Lipscomb, William H.; Lu, Yaqiong; Pandey, Ashutosh; Pelletier, Jon D.; Perket, Justin; Randerson, James T.; Ricciuto, Daniel M.; Sanderson, Benjamin M.; Slater, Andrew; Subin, Zachary M.; Tang, Jinyun; Thomas, R. Quinn; Val Martin, Maria; Zeng, Xubin. The Community Land Model Version 5: Description of New Features, Benchmarking, and Impact of Forcing Uncertainty. J. Adv. Model. Earth Syst. 11(12): 4245-4287. 2019.

Lenhard, J. and Winsberg, E.: Holism, entrenchment, and the future of climate model pluralism, Stud. Hist. Philos. Sci.: Stud. Hist. Philos. M. P., 41, 253–262, https://doi.org/10.1016/j.shpsb.2010.07.001, 2010.

Lipson, M.J., Grimmond, S., Best, M., Abramowitz, G., Coutts, A., Tapper, N., et al. (2023) Evaluation of 30 urban land surface models in the Urban-PLUMBER project: Phase 1 results. Q. J. Roy. Meteorol. Soc. 1–44. Available from: https://doi.org/10.1002/qj.4589, 2023.

Liu, J., Chen, J. M., Cihlar, J., Park, W. M.: A process-based boreal ecosystem productivity simulator using remote sensing inputs, Remote Sens. Environ. Volume 62, Issue 2, Pages 158-175, ISSN 0034-4257, https://doi.org/10.1016/S0034-4257(97)00089-8, 1997.

Formatted: Font: Not Italic

Formatted: Font: Not Italic

Formatted: Font: Not Italic

Deleted: Koster, Randal D., Suarez, Max J.. 1994. The components of a 'SVAT' scheme and their effects on a GCM's hydrological cycle. Advances in Water Resources. Vol. 17, No. 1-2, pp. 61-78. DOI: 10.1016/0309-1708(94)90024-8 ISSN: 03091708 ¶ ¶
Koster, R. D., and Suarez M. J. , 1996: Energy and water balance calculations in the Mosaic LSM. NASA Tech. Memo. 104606, Vol. 9, 59 pp.¶

Deleted: N.

Deleted: J.

Deleted: P.

Deleted: S.

Deleted: I. C.

Deleted: (2005)

Deleted: ical

Deleted: cles

Deleted: .

Deleted: and Coauthors, 2006:

Deleted: ¶

Deleted: .

Deleted: (2019)

Deleted: .

Deleted: "

Deleted: "

Deleted: ournal

Deleted: of

Deleted: ances

Deleted: in

Deleted: ing

Deleted: tems

Deleted: .

Deleted: uarterly

Deleted: ournal of the

Deleted: al

Deleted: ogical

Deleted: iety

Deleted: J.M.

Deleted: J.

Deleted: W.M.

Deleted:

Deleted: ing

Deleted: of

Deleted: nment

Deleted: , 1997

Deleted: .

Manabe, S.: Climate and the ocean circulation: I. The atmospheric circulation and the hydrology of the earth's surface. Mon. Wea. Rev., 97, 739–805, doi:10.1175/1520-0493(1969)097,0739: CATOC.2.3.CO;2, 1969.

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

Mauder, M., Foken, T. & Cuxart, J. Surface-Energy-Balance Closure over Land: A Review. Boundary-Layer Meteorol. 177, 395–426, https://doi.org/10.1007/s10546-020-00529-6, 2020.

Mianabadi, A., Coenders-Gerrits, M., Shirazi, P., Ghahraman, B., and Alizadeh, A.: A global Budyko model to partition evaporation into interception and transpiration, Hydrol. Earth Syst. Sci., 23, 4983–5000, https://doi.org/10.5194/hess-23-4983-2019, 2019.

Moderow, U., Grünwald, T., Queck, R., Spank, U., and Bernhofer, C.: Energy balance closure and advective fluxes at ADVEX sites. Theor. Appl. Climatol., 143, 761–779, https://doi.org/10.1007/s00704-020-03412-z, 2021.

Monteith, J. L., and M. H. Unsworth, 1990: Principals of Envi- ronmental Physics. 2nd ed. Edward Arnold, 241 pp.

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. doi: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., 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., 116, D12109, doi: 10.1029/2010JD015139, 2011.

Ogée, J., Brunet, Y., Loustau, D., Berbigier, P., and Delzon, S.: MuSICA, a CO2, water and energy multilayer, multileaf pine forest model: evaluation from hourly to yearly time scales and sensitivity analysis, Glob. Change Biol., 9, 697–717, doi:10.1046/j.1365-2486.2003.00628.x, 2003.

Pan, S., Pan, N., Tian, H., Friedlingstein, P., Sitch, S., Shi, H., Arora, V. K., Haverd, V., Jain, A. K., Kato, E., Lienert, S., Lombardozzi, D., Nabel, J. E. M. S., Ottlé, C., Poulter, B., Zaehle, S., and Running, S. W.: Evaluation of global terrestrial evapotranspiration using state-of-the-art approaches in remote sensing, machine learning and land surface modeling, Hydrol. Earth Syst. Sci., 24, 1485–1509, https://doi.org/10.5194/hess-24-1485-2020, 2020.

Polcher, J., McAvaney, B., Viterbo, P., Gaertner, M.-A., Hahmann, A., Mahfouf, J.-F., Noilhan, J., Phillips, T., Pitman, A.J., Schlosser, C.A., Schulz, J.-P., Timbal, B., Verseghy D., and Xue, Y.: A proposal for a general interface between land-surface schemes and general circulation models. Global and Planet. Change, 19:263-278, 1998.

Polcher, J., and Coauthors, 2000: GLASS: Global Land- Atmosphere System Study. GEWEX News, Vol. 10, No. 2, International GEWEX Project Office, Silver Spring, MD, 3-5.

Deleted: , 1969
Deleted: .
Deleted: (2020).
Deleted: et al
Deleted: (2021)
Deleted: .
Deleted: &
Deleted: (2018).
Deleted: ournal Of
Deleted: ogy
Deleted: et al. (2011),
Deleted: .
Deleted: ¶
Pitman, A. J. and G. Abramowitz (2005) What are the limits to statistical error correction in land surface schemes when projecting the future, *Geophysical Research Letters*, vol. 32, pp. L14403, http://dx.doi.org/10.1029/2005GL023158¶
Deleted: (1998)
Deleted: ary
Deleted: .
Deleted: ¶
Shen, C., Appling, A.P., Gentine, P. et al. Differentiable modelling to unify machine learning and physical models for geosciences. Nat Rev Earth Environ 4, 552–567 (2023). https://doi.org/10.1038/s43017-023-00450-9¶

Schmidt, A., Hanson, C., Chan, W. S. and Law, B. E.: Empirical assessment of uncertainties of meteorological parameters and turbulent fluxes in the AmeriFlux network, J. Geophys. Res., 117, G04014, doi:10.1029/2012JG002100, 2012.

Seiler, C., Melton, J. R., Arora, V. K., Sitch, S., Friedlingstein, P., Anthoni, P., Goll, D., Jain, A. K., Joetzjer, E., Lienert, S., Lombardozzi, D., Luyssaert, S., Nabel, J. E. M. S., Tian, H., Vuichard, N., Walker, A. P., Yuan, W. and Zaehle S.: Are terrestrial biosphere models fit for simulating the global land carbon sink? J. Adv. Model. Earth Syst., 14, e2021MS002946. https://doi.org/10.1029/2021MS002946, 2022.

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., Yan Yu, Zeng, F., 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 Syst., 16, e2023MS003922. https://doi.org/10.1029/2023MS003922, 2023,

Smith, B., Wårlind, D., Arneth, A., Hickler, T., Leadley, P., Siltberg, J., and Zaehle, S.: Implications of incorporating N cycling and N limitations on primary production in an individual-based dynamic vegetation model, Biogeosci., 11, 2027–2054, https://doi.org/10.5194/bg- 11-2027-2014, 2014.

Sposito, Garrison. 2017.: Understanding the Budyko Equation Water 9, no. 4: 236. https://doi.org/10.3390/w9040236

Stoy, P.C., Mauder, M., Foken, T., Marcolla, B., Boegh, E., Ibrom, A., Arain, M. A., Arneth, A., Aurela, M., Bernhofer, C., Cescatti, A., Dellwik, E., Duce, P., Gianelle, D., van Gorsel, E., Kiely, G., Knohl, A., Margolis, H., McCaughey, H., Merbold, L., Montagnani, L., Papale, D., Reichstein, M., Saunders, M., Serrano-Ortiz, P., Sottocornola, M., Spano, D., Vaccari, F., Varlagin, A.: A data-driven analysis of energy balance closure across FLUXNET research sites: the role of landscape scale heterogeneity, Agric. For. Meteorol. 171–172:137–152. https://doi.org/10.1016/j.agrformet.2012.11.004, 2013.

Thum, T., Caldararu, S., Engel, J., Kern, M., Pallandt, M., Schnur, R., Yu, L., and Zaehle, S.: A new model of the coupled carbon, nitrogen, and phosphorus cycles in the terrestrial biosphere (QUINCY v1.0; revision 1996), Geosci. Model Dev., 12, 4781–4802, https://doi.org/10.5194/gmd-12-4781-2019, 2019.

Ukkola, A. M., Haughton, N., De Kauwe, M. G., Abramowitz, G., and Pitman, A. J.: FluxnetLSM R package (v1.0): a community tool for processing FLUXNET data for use in land surface modelling, Geosci. Model Dev., 10, 3379–3390, https://doi.org/10.5194/gmd-10-3379-2017, 2017.

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. J. J. M., Viterbo, P., Beljaars, A. C. M., and Betts, A. K.: Offline validation of the ERA-40 surface scheme. ECMWF Tech. Memo. No. 295, 2000.

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

Formatted: English (AUS)

Formatted: Font: Not Italic

Formatted: Font: Not Italic

Formatted: Font: Not Italic

Formatted: Font: Not Italic

Moved down [27]: K. A.

Moved down [28]: J. P.

Moved (insertion) [27]

Moved (insertion) [28]

Deleted: Shevliakova, E., S. Malyshev, I. Martinez-Cano, P. C. D. Milly, S. W. Pacala, P. Ginoux,

Deleted: C.

Moved down [29]: K. L.

Moved down [30]: T. R.

Moved down [31]: J. P.

Moved (insertion) [29]

Moved (insertion) [30]

Moved (insertion) [31]

Deleted: K.

Deleted: L.W.

Deleted: V.

Deleted: P.

Deleted: N.

Deleted: F.

Deleted: Y.

Deleted: (2023)

Deleted: ournal of

Deleted: ances

Deleted: in

Deleted: ing

Deleted: ems

Deleted: in review.

Deleted: ences

Deleted: (2013)

Deleted: .

Deleted: G.

Deleted: M. G.

Deleted:

Deleted: (2022)

Deleted: .

evaluation of gross primary production, Geosci. Model Dev., 12, 4751–4779, https://doi.org/10.5194/gmd-12-4751-2019, 2019.

Walker, A. P., Quaife, T., van Bodegom, P. M., De Kauwe, M. G., Keenan, T. F., Joiner, J., Lomas, M. R., MacBean, N., Xu, C. G., Yang, X., J., and Woodward, F. I.: The impact of alternative trait-scaling hypotheses for the maximum photosynthetic carboxylation rate (V-cmax) on global gross primary production, New Phytol., 215, 1370–1386, 2017.

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., 116, G01034, doi:10.1029/2010JG001385, 2011.

Wang, Y., Zeng, Y., Yu, L., Yang, P., Van der Tol, C., Yu, Q., Lü, X., Cai, H., and Su, Z.: Integrated modeling of canopy photosynthesis, fluorescence, and the transfer of energy, mass, and momentum in the soil–plant–atmosphere continuum (STEMMUS–SCOPE v1.0.0), Geosci. Model Dev., 14, 1379–1407, https://doi.org/10.5194/gmd-14-1379-2021, 2021.

Wartenburger, R., Seneviratne, S. I., Hirschi, M., Chang, J., Ciais, P., Deryng, D., Elliott, J., Folberth, C., Gosling, S. N., Gudmundsson, L., Henrot, A.-J., Hickler, T., Ito, A., Khabarov, N., Kim, H., Leng, G., Liu, J., Liu, X., Masaki, Y., Morfopoulos, C., Müller, C., Schmied, H. M., Nishina, K., Orth, R., Pokhrel, Y., Pugh, T. A. M., Satoh, Y., Schaphoff, S., Schmid, E., Sheffield, J., Stacke, T., Steinkamp, J., Tang, Q., Thiery, W., Wada, Y., Wang, X., Weedon, G. P., Yang, H., Zhou, T.: Evapotranspiration simulations in ISIMIP2a-Evaluation of spatio-temporal characteristics with a comprehensive ensemble of independent datasets. Environ. Res. Lett., 13(7):075001, Jun 2018.

Whitley, R., Beringer, J., Hutley, L. B., Abramowitz, G., De Kauwe, M. G., Evans, B., Haverd, V., Li, L., Moore, C., Ryu, Y., Scheiter, S., Schymanski, S. J., Smith, B., Wang, Y.-P., Williams, M., and Yu, Q.: Challenges and opportunities in land surface modelling of savanna ecosystems, Biogeosci., 14, 4711–4732, https://doi.org/10.5194/bg-14-4711-2017, 2017.

Wilson, K., Goldstein, A., Falge, E., Aubinet, M., Baldocchi, D., Berbigier, P., Bernhofer, C., Ceulemans, R., Dolman, H., Field, C., Grelle, A., Ibrom, A., Law, B. E., Kowalski, A., Meyers, T., Moncrieff, J., Monson, R., Oechel, W., Tenhunen, J., Valentini, R., Verma, S.: Energy balance closure at FLUXNET sites. Agric. For. Meteor., 113, 223–243, 2002.

Woodward, F.I., Smith, T. M., and Emanuel, W.R.: A global land primary productivity and phytogeography model, Global Biogeochem. Cy., 9, 471– 490, 1995.

Woodward, F.I., and Lomas, M. R.: Vegetation dynamics – simulating responses to climatic change, Biol. Rev., 79, 643– 670, 2004.

Deleted: T.
Deleted: P.M.
Deleted: M.G.
Deleted: T.F.
Deleted: J.
Deleted: M.R.
Deleted: N.
Deleted: C.G.
Deleted: X.J.
Deleted: F.I.
Deleted: .
Deleted: “
Deleted: ”.
Deleted: (
Deleted:)
Deleted: E.
Deleted: R.
Deleted: G.
Deleted: M. R.
Deleted: B.
Deleted: E.
Deleted: A.
Deleted: 2011
Deleted: .
Deleted: Wang, Y., Zeng, Y., Yu, L., Yang, P., Van der Tol, C., Yu, Q., Lü, X., Cai, H. and Su, Z., 2021. Integrated modeling of canopy photosynthesis, fluorescence, and the transfer of energy, mass, and momentum in the soil–plant–atmosphere continuum (STEMMUS–SCOPE v1. 0.0). *Geoscientific Model Development*, *14*(3), pp.1379-1407.
Formatted: Font: Not Italic
Formatted: English (AUS)
Deleted: ences
Deleted: T.M.
Deleted: W.R.
Deleted: .
Deleted: “
Deleted: ”.
Deleted: mical
Deleted: cles
Deleted: (
Deleted:)
Deleted: M.R.
Deleted: .
Deleted: “
Deleted: ”.
Deleted: ogical
Deleted: iew
Deleted: (
Deleted:)
Deleted: ¶
Yang, Z.-L., G.-Y. Niu, K. E. Mitchell, F. Chen, M. B. Ek, M. Barlage, K. Manning, D. Niyogi, M. Tewari, and Y. Xia, 2011: The Community Noah Land Surface Model with Multi-Parameter ... [25]

Page 4: [1] Deleted	Gab Abramowitz	04/06/2024 14:12:00
Page 4: [2] Deleted	Gab Abramowitz	04/06/2024 13:52:00
Page 7: [3] Deleted	Gab Abramowitz	03/06/2024 13:18:00
Page 8: [4] Deleted	Gab Abramowitz	04/06/2024 13:03:00
Page 17: [5] Deleted	Gab Abramowitz	08/06/2024 09:28:00
Page 17: [6] Deleted	Gab Abramowitz	08/06/2024 08:59:00
Page 19: [7] Deleted	Gab Abramowitz	09/06/2024 08:27:00
Page 19: [8] Deleted	Gab Abramowitz	09/06/2024 08:27:00
Page 24: [9] Deleted	Gab Abramowitz	05/06/2024 14:48:00
Page 24: [10] Deleted	Gab Abramowitz	07/06/2024 10:37:00
Page 24: [11] Deleted	Gab Abramowitz	07/06/2024 10:40:00
Page 24: [12] Deleted	Gab Abramowitz	07/06/2024 11:02:00
Page 28: [13] Deleted	Gab Abramowitz	25/06/2024 08:39:00
Page 28: [13] Deleted	Gab Abramowitz	25/06/2024 08:39:00
Page 28: [13] Deleted	Gab Abramowitz	25/06/2024 08:39:00
Page 28: [13] Deleted	Gab Abramowitz	25/06/2024 08:39:00
Page 28: [13] Deleted	Gab Abramowitz	25/06/2024 08:39:00
Page 28: [13] Deleted	Gab Abramowitz	25/06/2024 08:39:00
Page 28: [13] Deleted	Gab Abramowitz	25/06/2024 08:39:00
Page 28: [13] Deleted	Gab Abramowitz	25/06/2024 08:39:00

Page 28: [13] Deleted	**Gab Abramowitz**	**25/06/2024 08:39:00**
Page 28: [14] Deleted	**Gab Abramowitz**	**25/06/2024 08:42:00**
Page 28: [14] Deleted	**Gab Abramowitz**	**25/06/2024 08:42:00**
Page 28: [14] Deleted	**Gab Abramowitz**	**25/06/2024 08:42:00**
Page 28: [14] Deleted	**Gab Abramowitz**	**25/06/2024 08:42:00**
Page 28: [14] Deleted	**Gab Abramowitz**	**25/06/2024 08:42:00**
Page 28: [14] Deleted	**Gab Abramowitz**	**25/06/2024 08:42:00**
Page 28: [14] Deleted	**Gab Abramowitz**	**25/06/2024 08:42:00**
Page 28: [15] Deleted	**Gab Abramowitz**	**25/06/2024 08:46:00**
Page 28: [15] Deleted	**Gab Abramowitz**	**25/06/2024 08:46:00**
Page 28: [15] Deleted	**Gab Abramowitz**	**25/06/2024 08:46:00**
Page 28: [15] Deleted	**Gab Abramowitz**	**25/06/2024 08:46:00**
Page 28: [16] Deleted	**Gab Abramowitz**	**25/06/2024 08:47:00**
Page 28: [16] Deleted	**Gab Abramowitz**	**25/06/2024 08:47:00**
Page 28: [16] Deleted	**Gab Abramowitz**	**25/06/2024 08:47:00**
Page 28: [17] Deleted	**Gab Abramowitz**	**25/06/2024 08:47:00**

Page 28: [17] Deleted	**Gab Abramowitz**	**25/06/2024 08:47:00**
Page 28: [17] Deleted	**Gab Abramowitz**	**25/06/2024 08:47:00**
Page 28: [18] Deleted	**Gab Abramowitz**	**25/06/2024 08:48:00**
Page 28: [18] Deleted	**Gab Abramowitz**	**25/06/2024 08:48:00**
Page 28: [18] Deleted	**Gab Abramowitz**	**25/06/2024 08:48:00**
Page 28: [18] Deleted	**Gab Abramowitz**	**25/06/2024 08:48:00**
Page 28: [18] Deleted	**Gab Abramowitz**	**25/06/2024 08:48:00**
Page 28: [19] Deleted	**Gab Abramowitz**	**25/06/2024 08:49:00**
Page 28: [19] Deleted	**Gab Abramowitz**	**25/06/2024 08:49:00**
Page 28: [19] Deleted	**Gab Abramowitz**	**25/06/2024 08:49:00**
Page 28: [20] Deleted	**Gab Abramowitz**	**25/06/2024 08:50:00**
Page 28: [20] Deleted	**Gab Abramowitz**	**25/06/2024 08:50:00**
Page 28: [20] Deleted	**Gab Abramowitz**	**25/06/2024 08:50:00**
Page 28: [21] Deleted	**Gab Abramowitz**	**25/06/2024 08:51:00**
Page 28: [21] Deleted	**Gab Abramowitz**	**25/06/2024 08:51:00**

Page 28: [21] Deleted	**Gab Abramowitz**	**25/06/2024 08:51:00**
Page 28: [21] Deleted	**Gab Abramowitz**	**25/06/2024 08:51:00**
Page 28: [21] Deleted	**Gab Abramowitz**	**25/06/2024 08:51:00**
Page 28: [22] Deleted	**Gab Abramowitz**	**25/06/2024 08:44:00**
Page 28: [22] Deleted	**Gab Abramowitz**	**25/06/2024 08:44:00**
Page 28: [22] Deleted	**Gab Abramowitz**	**25/06/2024 08:44:00**
Page 28: [22] Deleted	**Gab Abramowitz**	**25/06/2024 08:44:00**
Page 28: [22] Deleted	**Gab Abramowitz**	**25/06/2024 08:44:00**
Page 28: [22] Deleted	**Gab Abramowitz**	**25/06/2024 08:44:00**
Page 28: [23] Deleted	**Gab Abramowitz**	**25/06/2024 08:52:00**
Page 28: [23] Deleted	**Gab Abramowitz**	**25/06/2024 08:52:00**
Page 28: [23] Deleted	**Gab Abramowitz**	**25/06/2024 08:52:00**
Page 28: [23] Deleted	**Gab Abramowitz**	**25/06/2024 08:52:00**
Page 28: [23] Deleted	**Gab Abramowitz**	**25/06/2024 08:52:00**
Page 28: [23] Deleted	**Gab Abramowitz**	**25/06/2024 08:52:00**
Page 28: [23] Deleted	**Gab Abramowitz**	**25/06/2024 08:52:00**

Page 28: [24] Deleted	**Gab Abramowitz**	**25/06/2024 08:53:00**

Page 28: [24] Deleted	**Gab Abramowitz**	**25/06/2024 08:53:00**

Page 28: [24] Deleted	**Gab Abramowitz**	**25/06/2024 08:53:00**

Page 28: [24] Deleted	**Gab Abramowitz**	**25/06/2024 08:53:00**

Page 28: [24] Deleted	**Gab Abramowitz**	**25/06/2024 08:53:00**

Page 28: [24] Deleted	**Gab Abramowitz**	**25/06/2024 08:53:00**

Page 33: [25] Deleted	**Gab Abramowitz**	**09/06/2024 10:34:00**