The Plumbing of Land Surface Models: Benchmarking Model Performance

J. Best, G. Abramowitz, H. Johnson, J. Pitman, G. Balsamo, A. Boone, M. Cuntz, B. Decharme, P. Dirmeyer, J. Dong, et al.

The Plumbing of Land Surface Models: Benchmarking Model Performance

M. Best, G. Abramowitz, H. Johnson, A. Pitman, G. Balsamo, Aaron Boone, M. Cuntz, B. Decharme, P. Dirmeyer, J. Dong, et al.

The Plumbing of Land Surface Models: Benchmarking Model Performance

M. J. BEST,[a] G. ABRAMOWITZ,[b] H. R. JOHNSON,[a] A. J. PITMAN,[b] G. BALSAMO,[c] A. BOONE,[d] M. CUNTZ,[e] B. DECHARME,[d] P. A. DIRMEYER,[f] J. DONG,[g] M. EK,[g] Z. GUO,[f] V. HAVERD,[h] B. J. J. VAN DEN HURK,[i] G. S. NEARING,[j] B. PAK,[k] C. PETERS-LIDARD,[j] J. A. SANTANELLO JR.,[j] L. STEVENS,[k] AND N. VUICHARD[l]

[a] *Met Office, Exeter, United Kingdom*
[b] *ARC Centre of Excellence for Climate System Science, University of New South Wales, Sydney, New South Wales, Australia*
[c] *ECMWF, Reading, United Kingdom*
[d] *CNRM-GAME, Météo-France, Toulouse, France*
[e] *Helmholtz Centre for Environmental Research–UFZ, Leipzig, Germany*
[f] *Center for Ocean–Land–Atmosphere Studies, George Mason University, Fairfax, Virginia*
[g] *NOAA/NCEP/EMC, College Park, Maryland*
[h] *Oceans and Atmosphere Flagship, CSIRO, Canberra, Australian Capital Territory, Australia*
[i] *KNMI, De Bilt, Netherlands*
[j] *Hydrological Sciences Laboratory, NASA GSFC, Greenbelt, Maryland*
[k] *Oceans and Atmosphere Flagship, CSIRO, Aspendale, Victoria, Australia*
[l] *Laboratoire des Sciences du Climat et de l'Environnement, UMR 8212, IPSL-LSCE, CEA-CNRS-UVSQ, Gif-sur-Yvette, France*

(Manuscript received 27 August 2014, in final form 19 December 2014)

ABSTRACT

The Protocol for the Analysis of Land Surface Models (PALS) Land Surface Model Benchmarking Evaluation Project (PLUMBER) was designed to be a land surface model (LSM) benchmarking intercomparison. Unlike the traditional methods of LSM evaluation or comparison, benchmarking uses a fundamentally different approach in that it sets expectations of performance in a range of metrics a priori—before model simulations are performed. This can lead to very different conclusions about LSM performance. For this study, both simple physically based models and empirical relationships were used as the benchmarks. Simulations were performed with 13 LSMs using atmospheric forcing for 20 sites, and then model performance relative to these benchmarks was examined. Results show that even for commonly used statistical metrics, the LSMs' performance varies considerably when compared to the different benchmarks. All models outperform the simple physically based benchmarks, but for sensible heat flux the LSMs are themselves outperformed by an out-of-sample linear regression against downward shortwave radiation. While moisture information is clearly central to latent heat flux prediction, the LSMs are still outperformed by a three-variable nonlinear regression that uses instantaneous atmospheric humidity and temperature in addition to downward shortwave radiation. These results highlight the limitations of the prevailing paradigm of LSM evaluation that simply compares an LSM to observations and to other LSMs without a mechanism to objectively quantify the expectations of performance. The authors conclude that their results challenge the conceptual view of energy partitioning at the land surface.

1. Introduction

Since the Project for the Intercomparison of Land-Surface Parameterizations Schemes (PILPS; Henderson-Sellers et al. 1993, 1995b) began to compare land surface models (LSMs) in 1993, the land modeling community has used a range of methods to examine how and why these models differ from each other and from observations. PILPS began with offline synthetic forcing (e.g., Pitman et al. 1999) but moved to using observational atmospheric forcing for multiple sites, including midlatitude grasslands (Chen et al. 1997; Schlosser et al. 2000), midlatitude catchments (Wood et al. 1998), high-latitude sites (Bowling et al. 2003), and the urban environment (Grimmond et al. 2010, 2011; Best and Grimmond 2013, 2014). The success of PILPS also led to regional-scale experiments

Corresponding author address: Martin Best, Met Office, Fitzroy Road, Exeter, Devon EX1 3PB, United Kingdom.
E-mail: martin.best@metoffice.gov.uk

DOI: 10.1175/JHM-D-14-0158.1

such as the African Monsoon Multidisciplinary Analysis (AMMA) Land Surface Model Intercomparison Project (ALMIP; Boone et al. 2009), which produced a multimodel ensemble of land surface states for regional-scale hydrological and meteorological studies. The spatial extent of most PILPS experiments was limited to point locations or catchments, and the need to broaden this scope was recognized by the Global Soil Wetness Project (GSWP). GSWP was global in extent, using atmospheric forcing at a 1° resolution to compare models and to produce a global soil moisture product (Dirmeyer et al. 1999). Like PILPS, GSWP led to a very large suite of science outcomes (Dirmeyer 2011), but both projects were limited by being uncoupled with the atmosphere and therefore lacking possible land–atmosphere feedbacks. This led to model intercomparison projects like the Global Land–Atmosphere Coupling Experiment (GLACE), which introduced the concept of coupling strength (Koster et al. 2004) and in turn led to major analyses of land processes coupled with atmospheric models (Koster et al. 2006; Guo et al. 2006). Most recently, the GLACE methodology has been used within phase 5 of the Coupled Model Intercomparison Project (CMIP5) to examine how soil moisture feedbacks might evolve into the future under changing climate and increasing greenhouse gas concentrations (Seneviratne et al. 2013).

These projects have led to a growth in our understanding of land surface processes and land–atmosphere interactions in the recent past and the future. However, there has also been a growing recognition of how profoundly challenging it is to compare LSMs given their varied complexity (Henderson-Sellers et al. 1995a) and that intercomparisons do not necessarily readily provide answers as to why LSM simulations differ from observations or each other.

Over the last few years, recognition of a major change has emerged within the land modeling community. The community has been developing a growing understanding of the distinctions between ''evaluation,'' ''comparison,'' and ''benchmarking.'' This paper focuses on these distinctions to arrive at a new mode of intercomparison that should catalyze a long-term revolution in how LSMs and perhaps natural system models in general, are evaluated, compared, and benchmarked. A schematic emphasizing the difference between evaluation, comparison, and benchmarking is shown in Fig. 1. We describe each in turn below.

1) *Evaluation.* Model outputs are typically compared to observations to derive an error measure (Fig. 1a). The metrics used to do this can involve a number of variables from the model, various locations, or different statistical measures that may focus on mean values, variability, or properties of variable distributions. Metrics with errors deemed to be large are usually identified as important markers for development programs. For example, in Fig. 1a, metrics 4 and 11, plotted along the x axis, have the largest relative errors and might be target metrics for model development.
2) *Comparison.* In this case, a model is not just compared to observations, but also to alternative models. In addition to identifying the metrics that have the largest relative errors, this type of analysis also identifies metrics for which one model performs better than another, or where errors in multiple models are systematic. This has the advantage over evaluation of giving a clear indication that performance improvements are achievable for those metrics where another model already performs better. For example, in Fig. 1b, metrics 4 and 11 apparently have the largest relative errors for both models A and B (and are hence likely to be flagged as development priorities). Note that while we might expect to be able to improve the models' performance for these metrics, there is no categorical guarantee that improvements are in fact achievable. We can, however, see that model B has substantially larger errors for metrics 2, 8, and 9 than model A, meaning that improvements to model B for these metrics are attainable. The same can obviously be said for model A for metric 12. We might also deduce from this type of analysis that model A performs better than model B (since it performs better in a larger number of metrics), working under the assumptions that these metrics are of equal weight and that both models are essentially designed for the same purpose. A rarely discussed but worrisome aspect of model comparison is the risk that a model is developed to be more similar to other models without necessarily understanding the causes of differences. Models therefore become more similar, but not necessarily because they are becoming more like the observations.
3) *Benchmarking.* The fundamental characteristic of benchmarking is that performance expectations—in this case, benchmark values for each error metric—are defined and perhaps prioritized a priori. There are several ways performance expectations might be defined before running a model.
   (i) *Better than another model.* The most common, and perhaps weakest, approach is to set the results from a different model as the performance benchmark. This could either be a previous version of the same model or an alternative

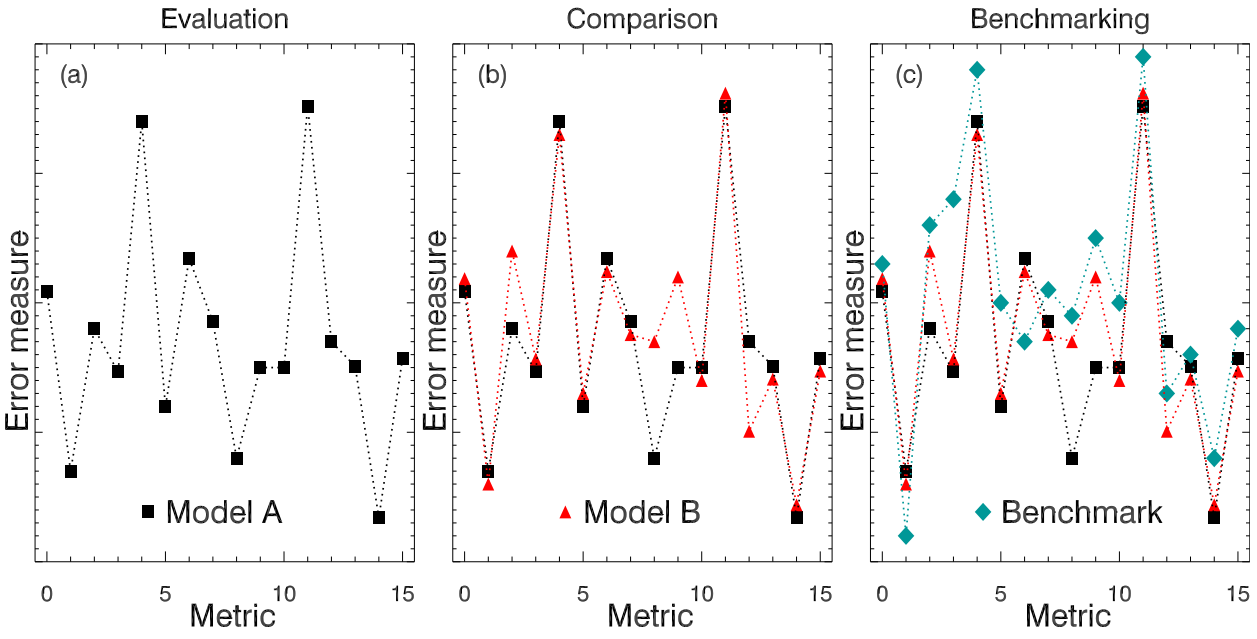

FIG. 1. Conceptual figure showing the performance of models through (a) evaluation, (b) comparison, and (c) benchmarking. The *x* axis represents a series of metrics a modeler might use to evaluate the model, and the *y* axis represents the normalized error from each metric used to assess model performance. The dotted lines are a visual guide and have no scientific relevance.

model or ensemble. The distinction here between benchmarking and comparison is subtle, but important nevertheless: the analysis focuses on what is required from the model rather than just a relative error assessment. We describe setting the performance of an alternative model as the benchmark as weak because it neglects the possibility that both models are poor, or the possibility that both models are already within observational error.

(ii) *Fit for a particular application.* A stronger application of benchmarking is to define the levels of performance required for a model to be fit for a particular application. For example, in fluvial flood forecasting applications, metrics might focus on set tolerance criteria for both the timing and volume of water that breaches a riverbank, defining the severity of the event. This can lead to very different conclusions for both model performance and development requirements. For example, in Fig. 1c, metrics 4 and 11, which were previously identified as foci for development because they had the largest relative errors, are now within the benchmark expectations and hence need no further development for this particular purpose. On the other hand, metrics 1 and 6 have errors that are greater than benchmark expectations and are now a development priority, even though metric 1 apparently has one of the smallest relative errors. Note also that model A fails to beat the benchmark for metric 12 and that there are no other metrics where the benchmark discriminates between the two models. In this case, we can draw the opposite conclusion to the one we outlined when describing comparison: model B actually performs better than model A, as it passes more of the benchmarks and is hence more suitable for this application.

(iii) *Effectively utilizes available information.* A third example of benchmarking defines a priori expectations based on the complexity of a model and the amount of information provided to it. For example, an LSM that is given information about vegetation and soil at a location in addition to time-varying meteorology should be expected to perform better than one that is not. Similarly, a model that allows for nonlinear relationships between its input and output variables should be expected to perform better than one that prescribes linear relationships. This approach tries to quantify how well a model utilizes information available in the input data when generating its output variables (Abramowitz 2005, 2012). Imagine, for example, that the green line in Fig. 1c represented a linear model that predicted land surface carbon fluxes purely as a function of incoming radiation.

TABLE 1. Information about the 20 flux tower sites.

Name	Code	Country	Lat	Lon	Plant functional type	Duration
Amplero	Am	Italy	41.90°N	13.61°E	Grassland	2003–06
Blodgett	Bl	United States	38.90°N	120.63°W	Evergreen needleleaf	2000–06
Bugac	Bu	Hungary	46.69°N	19.60°E	Grassland	2002–06
El Saler	El	Spain	39.35°N	0.32°W	Evergreen needleleaf	2003–06
El Saler 2	E2	Spain	39.28°N	0.32°W	Cropland	2005–06
Espirra	Es	Portugal	38.64°N	8.60°W	Evergreen broadleaf	2001–06
Fort Peck	FP	United States	48.31°N	105.10°W	Grassland	2000–06
Harvard	Ha	United States	42.54°N	72.17°W	Deciduous broadleaf	1994–2001
Hesse	He	France	48.67°N	7.06°E	Deciduous broadleaf	1999–2006
Howard	Ho	Australia	12.49°S	131.15°E	Woody savanna	2002–05
Howlandm	Hl	United States	45.20°N	68.74°W	Evergreen needleleaf	1996–2004
Hyytiala	Hy	Finland	61.85°N	24.29°E	Evergreen needleleaf	2001–04
Kruger	Kr	South Africa	25.02°S	31.50°E	Savanna	2002–03
Loobos	Lo	Netherlands	52.17°N	5.74°E	Evergreen needleleaf	1997–2006
Merbleue	Me	Canada	45.41°N	75.52°W	Permanent wetland	1999–2005
Mopane	Mo	Botswana	19.92°S	23.56°E	Woody savanna	1999–2001
Palang	Pa	Indonesia	2.35°N	111.04°E	Evergreen broadleaf	2002–03
Sylvania	Sy	United States	46.24°N	89.35°W	Mixed forest	2002–05
Tumbarumba	Tu	Australia	35.66°S	148.15°E	Evergreen broadleaf	2002–05
University of Michigan	UM	United States	45.56°N	84.71°W	Deciduous broadleaf	1999–2003

The knowledge that there are metrics in which this very simple model outperforms models A and B (i.e., metrics 1 and 6) tell us that 1) both models A and B have scope for improvement and 2) that models A and B do not require any more information (i.e., more input variables/parameters) to achieve this improvement. Moreover, Gong et al. (2013) show that it is possible to establish a benchmark that can identify a potential upper bound on the best achievable performance of a model. This could be used, for example, to determine if the structure of models A and B can be improved for metrics 4 and 11 in Fig. 1.

While this conception of benchmarking is positioned broadly within ecological and Earth system model evaluation in Luo et al. (2012), efforts to resolve these issues within the land surface modeling community are led from within the Global Land–Atmosphere System Study (GLASS; van den Hurk et al. 2011). These efforts continue the history of international LSM intercomparison projects but have been extended to use a common online LSM benchmarking system, the Protocol for the Analysis of Land Surface Models (PALS; Abramowitz 2012; pals.nci.org.au/home). Most recently, the PALS Land Surface Model Benchmarking Evaluation Project (PLUMBER) was created to explore these distinctions in contemporary LSM evaluation. Simulations from 13 LSMs are compared at 20 flux tower sites using five predetermined benchmarks across a range of metrics for sensible Q_H and latent Q_E heat fluxes. We omit consideration of carbon fluxes because only a small subset of the models in PLUMBER provided data. PLUMBER is intended as a foundation experiment, isolating common features of land model performance so that the community can target areas requiring improvements common to all groups, as well as areas specific to individual modeling groups.

2. Methods

a. Datasets and experimental methods

We use observations as the basis of our experiment, obtained through the FLUXNET LaThuile free fair-use subset (fluxdata.org; see Acknowledgments). The 20 flux tower sites used here are listed in Table 1 with locations shown in Fig. 2. Further gap filling and quality control specifically focused on use by LSMs were performed before netCDF versions of LSM forcing meteorological variables were made available through PALS for PLUMBER participants. This process included 1) removing time periods where any significant LSM forcing variable was not present (e.g., downward shortwave radiation, surface air temperature, rainfall, or humidity), 2) only allowing whole years of data that satisfied the first criterion, and 3) gap filling or entirely synthesizing downward longwave radiation using the approach outlined in Abramowitz et al. (2012). Log files giving details for this process at each site are accessible through PALS. Sites were chosen to obtain a global spread, giving broad coverage of different vegetation types (Fig. 2) and a range in climates (Fig. 3; the codes for each site are given in Table 1). Of the datasets that

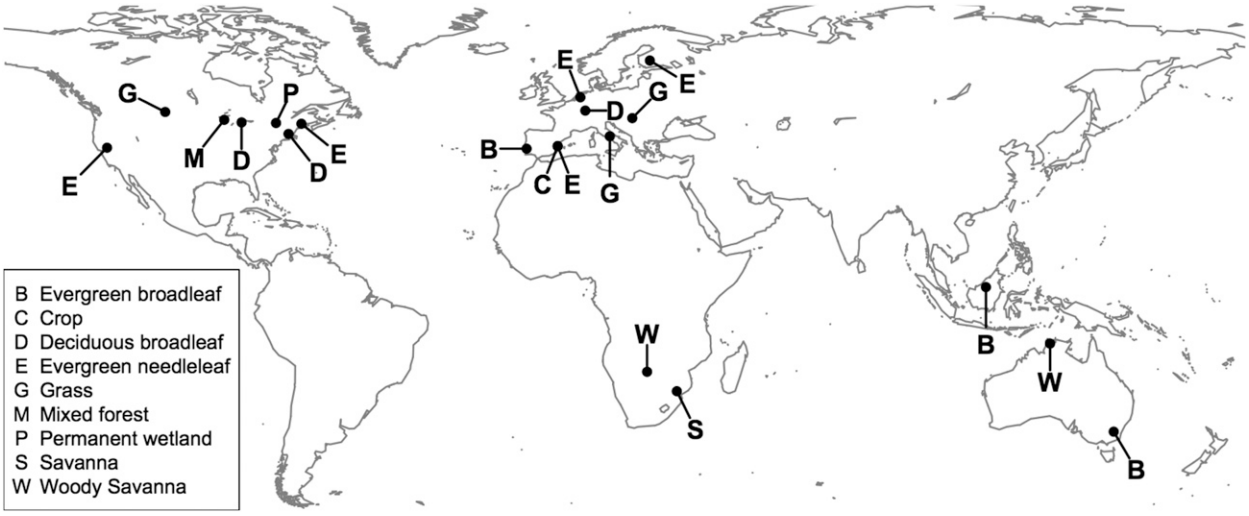

FIG. 2. Locations and biomes of the 20 flux tower sites.

fulfilled these requirements, preference was given to those with longer records.

It is known that many of the FLUXNET datasets do not have long-term energy balance closure (e.g., Wilson et al. 2002; Kidston et al. 2010), a problem that could affect our results. Unfortunately, a comprehensive list of sites with good energy balance closure is not available, and indeed knowing which sites come close to conserving energy, or the extent to which they do, would not resolve whether or not this is a cause of the empirical models' performance. We suspect energy balance closure in the FLUXNET data will not significantly affect our conclusions, but note that significant additional work will be required to determine this conclusively.

Results were returned from eight LSMs, which have been developed by different research groups. In addition, a

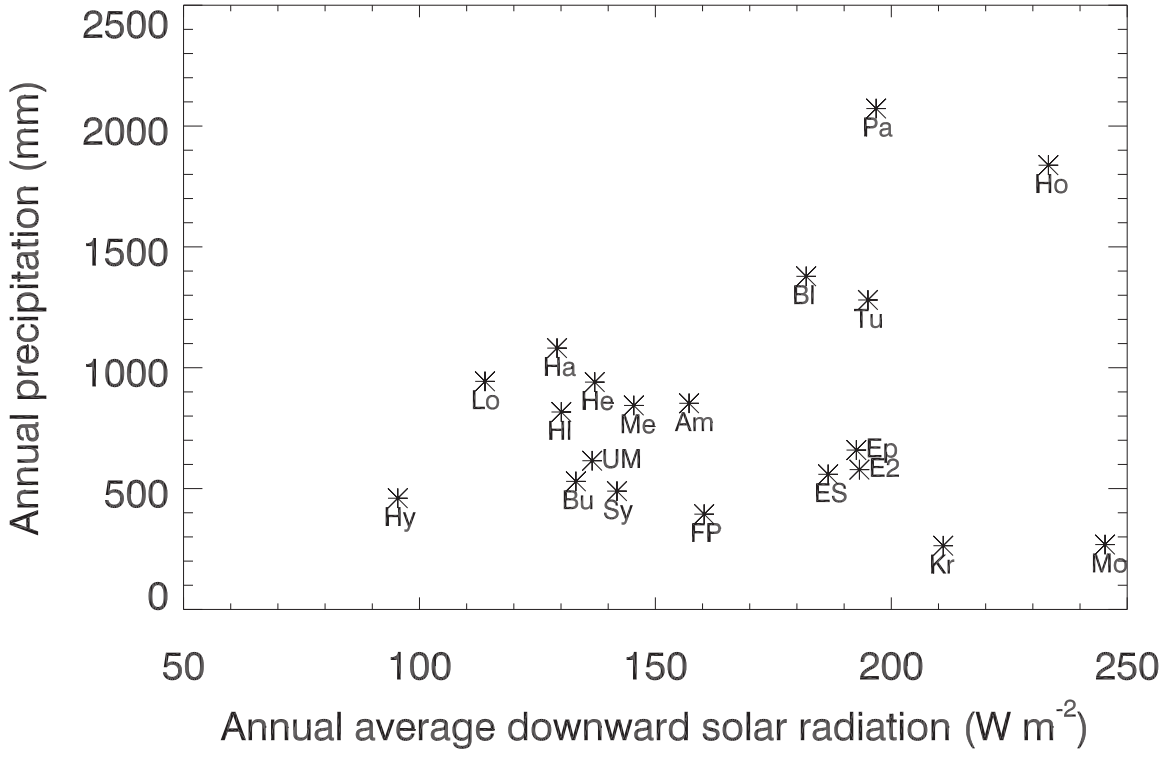

FIG. 3. Relationship between annual mean shortwave radiation and precipitation for each site. The codes for each site are given in Table 1.

TABLE 2. Participating land surface models.

Model	Model version	Reference	Notes
Community Atmosphere Biosphere Land Exchange (CABLE)	2.0	Wang et al. (2011); Kowalczyk et al. (2006)	This is identical to CABLE-2.0 tagged version (Trunk revision 304) but has extra spinup options for this study. For two sites, Kruger and Mopane, a fix for dry conditions was implemented. This fix did not have significant impact on other sites. (https://trac.nci.org.au/trac/cable/wiki)
CABLE Soil–Litter–Iso (CABLE-SLI)	2.0_SLI	Haverd and Cuntz (2010)	This uses the SLI soil model in place of the default soil scheme in CABLE-2.0.
Tiled ECMWF Scheme for Surface Exchanges over Land (TESSEL), uncoupled version (CHTESSEL)	1.0	Balsamo et al. (2009); Boussetta et al. (2013)	Carbon module only used to drive carbon allocation and carbon flux, but evaporation parameterized using the Jarvis–Stewart approach.
Center for Ocean–Land–Atmosphere Studies Simplified SiB (COLA-SSiB)	2.0	Dirmeyer and Zeng (1999); Guo and Dirmeyer (2013)	Default configuration.
ISBA-3L	Surface Externalisée, version 7.2 (SURFEXv7.2)	Boone et al. (1999); Masson et al. (2013)	Three-layer force–restore approach for the soil (superficial, root zone, and subroot zone layers).
ISBA_dif	SURFEXv7.2	Decharme et al. (2011); Masson et al. (2013)	N-layer ($N = 14$) soil moisture–temperature model (Richard's equation, heat diffusion, root profile).
Joint UK Land Environment Simulator (JULES)	3.1	Best et al. (2011)	Default configuration.
JULES_altP	3.1	Best et al. (2011)	This is identical to the default JULES 3.1 configuration, except 1) the emissivity of each surface has been reduced; 2) the spectral albedo, interactive phenology, and soil moisture heterogeneity for enhanced runoff have been turned off; 3) the van Genuchten soil hydraulic scheme is used; 4) supersaturated soil moisture is drained into lower layers; and 5) the snow canopy option has been turned off for each vegetation type except needleleaf trees.
Mosaic	—	Koster and Suarez (1992, 1994)	This is currently used in North American Land Data Assimilation System, version 2 (NLDAS-2), but is no longer used in the GSFC GCM.
Noah2.7.1	2.7.1	Ek et al. (2003)	The community Noah LSM is used operationally in NCEP models: 1) 2.7.1 currently in GFS and CFS, 2) 2.8 in NLDAS, and 3) 3.0 in NAM.
Noah3.2	3.2	www.ral.ucar.edu/research/land/technology/lsm.php	This is identical to Noah 2.7.1, except for 1) updated roughness length and snow albedo over snow-covered surfaces; 2) updated soil moisture availability; 3) added the exchange of heat required to change the temperature of falling precipitation from air temperature to skin temperature; 4) calculation of roughness and emissivity dependent to vegetation fraction; 5) added capability to use MODIS land-use dataset for vegetation categories; 6) added option to use 2D LAI; 7) significant changes to the treatment of glacial ice; 8) included multilayer urban; and 9) cold-start initialization with soil moisture initialized at 0.2 and soil temperature initialized at 290 K.

TABLE 2. (*Continued*)

Model	Model version	Reference	Notes
Noah3.3	3.3	www.ral.ucar.edu/research/land/technology/lsm.php	This is identical to Noah3.2, except for the activation of time-varying roughness length and fixes to the underground runoff. Initialized with 10-yr spinup using 1-yr recursive forcing of the first year.
ORCHIDEE	Trunk version rev. 1401	Krinner et al. (2005)	The hydrological scheme used for these simulations is a two-layer bucket model (Choisnel). (http://forge.ipsl.jussieu.fr/orchidee/browser/trunk)

further five variants or alternative versions of these eight models were also submitted, giving a total of 13 LSMs (Table 2).

All model output for PLUMBER was uploaded and analyzed in PALS. PALS hosts modeling experiments with each experiment containing downloadable driving data and experimental protocols. Resulting model output is uploaded to the system and is used in automated analyses against observational data and benchmarks associated with the experiment. PALS also automatically calculates a suite of empirically based benchmarks, described in more detail below, that provide predefined levels of performance against which LSMs can be tested. The model simulations used in PLUMBER and analyses of them are available on PALS once access to the PLUMBER workspace is requested from PALS administrators.

The initial conditions for soil moisture can have a significant impact on the surface heat and moisture fluxes from LSMs, so it is important to ensure that a consistent spinup strategy is used for all sites and all models. A challenge with some sites was that only 2 years of atmospheric forcing was available, which is not always sufficient to ensure that soil moisture is fully spun up. To ensure we could use a wide range of sites with good geographical and vegetation diversity, while also retaining consistency between all of the sites to ensure a thorough comparison, we required a spinup strategy that allowed us to use these sites with only 2 years of data. We therefore initialized all LSMs as saturated and then repeated the first year of forcing 10 times. Beginning at a saturated state accelerates spinup relative to a dry initial state because gravitational drainage helps remove excess soil moisture.

For most models and most sites this spinup procedure is more than adequate (Yang et al. 1995; Rodell et al. 2005). However, some sites remain problematic. For example, Kruger only had 2 years of forcing data, and the first year was a very dry year (274 mm of rain) relative to the climatological average (525 mm of rain; www.fluxdata.org:8080/SitePages/siteInfo.aspx?ZA-Kru). This led to very dry soil moisture, which may or may not be reflective of the previous period immediately before our chosen year. This may affect the simulation of Q_E and Q_H at this site for the 13 physically based LSMs and the Manabe bucket model (Manabe 1969) benchmark (M69) described later. However, this issue would not affect a second physically based benchmark [the Penman–Monteith (PM) model; Monteith and Unsworth 1990] or empirical benchmarks, as described below.

b. Statistical metrics

Evaluation studies typically use canonical statistical metrics such as mean bias, root-mean-square error, normalized mean error, or correlation. For PLUMBER we use four common statistical measures on half-hourly data: mean bias error MBE, standard deviation SD, correlation coefficient r, and normalized mean error NME. To obtain a metric to compare the models for SD, we use the absolute difference between 1.0 and the ratio of measured to observed standard deviations. The equations for all of the statistics used here are given in Table 3. Each of these contributes different evaluation information. The value of MBE simply represents the difference in the mean value of a variable between observations and a model; SD gives an indication of the magnitude of variability; r gives information about temporal coincidence of variability; and NME gives information about all three of the previous metrics in one, but is less sensitive to being dominated by outlier values than root-mean-square error, which is more commonly used.

While these standard statistical metrics give information about the mean and variability of a model compared to observations, they are limited in terms of identifying the skill of the model for predicting extremes. The extremes are defined as the edges or tails of the distribution of a quantity and we want to ensure LSMs capture these tails skillfully because they are increasingly central to explaining important phenomenon, including droughts and heat waves (Seneviratne et al. 2010; Hirschi et al. 2011). We therefore include two statistical measures for an analysis of the extremes, namely the 5th and 95th percentiles of the distributions. These measures define values of the flux at these points of the distribution and use the absolute distance between the modeled and observed values as a metric (Table 3).

In addition to the extremes of a distribution, there are other statistical measures that can be used to determine

TABLE 3. Statistical formulae used for the analyses of the LSMs, where M represents the model values and O represents the observed values.

Common statistical measures	
Mean bias error (MBE)	$\frac{\left[\sum_{i=1}^{n}(M_i - O_i)\right]}{n}$
Standard deviation (SD)	$\left\lvert 1 - \frac{\sqrt{\frac{\sum_{i=1}^{n}(M_i - \overline{M})^2}{n-1}}}{\sqrt{\frac{\sum_{i=1}^{n}(O_i - \overline{O})^2}{n-1}}} \right\rvert$
Correlation coefficient (r)	$\frac{n\sum_{i=1}^{n}(O_i M_i) - \left(\sum_{i=1}^{n} O_i \sum_{i=1}^{n} M_i\right)}{\sqrt{\left[n\sum_{i=1}^{n} O_i^2 - \left(\sum_{i=1}^{n} O_i\right)^2\right]\left[n\sum_{i=1}^{n} M_i^2 - \left(\sum_{i=1}^{n} M_i\right)^2\right]}}$
Normalized mean error (NME)	$\frac{\sum_{i=1}^{n}\lvert M_i - O_i\rvert}{\sum_{i=1}^{n}\lvert \overline{O} - O_i\rvert}$
Extremes of the distribution	
5th percentile statistical measure (M_5 and O_5 are values at 5th percentile of distribution of M and O, respectively).	$\lvert M_5 - O_5\rvert$
95th percentile statistical measure (M_{95} and O_{95} are values at 95th percentile of distribution of M and O, respectively).	$\lvert M_{95} - O_{95}\rvert$
Distribution statistical measures	
Skewness	$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{M_i - \overline{M}}{\mathrm{SD}}\right)^3$
Kurtosis	$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{M_i - \overline{M}}{\mathrm{SD}}\right)^4 - 3$
Overlap	$\sum_{i=1}^{n}\min(M_i, O_i)$

how well a model recreates the distribution of the observed values. These provide information about various aspects on the shape of the distribution. For PLUMBER we have used three such metrics (Table 3): the kurtosis, which is a measure of how "pointed" the distribution is; the skewness, which is a measure of how symmetrical the distribution is; and the overlap of the observed and modeled distributions (Perkins et al. 2007). The statistics for each of these were determined from probability density functions fitted to each of the modeled and observed variables.

c. *Physical benchmarks*

Our first physically based benchmark is the M69 bucket model (Manabe 1969). As the name suggests, soil moisture is represented by a simple bucket that is filled by infiltration into the soil and emptied by evapotranspiration. The M69 model has a long history in climate modeling (Manabe 1969). The first clear demonstration of the limits of this model was by Chen et al. (1997), who showed that the model tended to evaporate too rapidly because of the lack of appropriate surface resistances. However, similar to the PM benchmark, we use the M69 model expressly because the simplicity should mean that physically based LSMs should be able to beat this model. Indeed, PILPS demonstrated (Chen et al. 1997) that most land models could and should beat M69 on a long-term average, and we include it here to extend that finding to metrics other than the mean.

The PM benchmark was configured as defined by the United Nations (UN) Food and Agriculture Organization

(FAO) standard (Allen et al. 1998). The surface exchange turbulence that drives the evapotranspiration calculations in both M69 and PM was taken from the PM scheme for a grass reference surface, and the albedo was set to a constant value of 0.2 for all sites. PM assumes a standard reference crop, which is irrigated such that it is never water stressed and hence there is no requirement to model soil moisture for this scheme. This is clearly a limitation and we acknowledge this, but the key here is that physically based LSMs should be able to beat this minimum benchmark. Both of these physical benchmarks represent the first category of benchmark described in section 1.

d. Empirical benchmarks

We use three empirical benchmarks that attempt to quantify the information available in the atmospheric forcing variables for predicting, Q_E and Q_H. They are in the third category of benchmarking described in section 1. All three construct independent empirical relationships between meteorological drivers Q_E and Q_H, and all three benchmarks are used as benchmarks out of sample.

The simplest empirical benchmark (EMP1lin) is a linear regression of each of Q_E and Q_H against incoming solar radiation SWdown. The next is a multiple linear regression against SWdown and near-surface air temperature T_a (EMP2lin). The third and most complex (EMP3KM27) is a nonlinear regression against SWdown, T_a, and near-surface air relative humidity R_H. It uses a k-means clustering approach to create 27 distinct subdomains of the SWdown–T_a–R_H domain and then performs a multiple linear regression between Q_H or Q_E and the three meteorological variables. This delivers a nonlinear (piecewise linear) response to these three forcing variables. The number of clusters was chosen to give a simple conceptual representation. Imagine that each SWdown was binned into high, medium, and low values. Within each of these bins, imagine a similar discretization for T_a and R_H, so that there are nine bins with SWdown in the high range. Giving each variable three discretizations, on average, allows $3^3 = 27$ clusters.

Critically, for all three empirical benchmarks, the parameters are determined by statistical regressions using data that are out of sample, meaning that data from the site at which we are testing are not used to establish the regression parameters for that site, but are taken from the remaining 19 sites. This is in some sense analogous to not allowing the LSMs to calibrate their parameters using local site data. Meteorology at the testing site, together with these empirical parameters trained using data from other sites, is then used to make each benchmark prediction at the testing site. These three benchmarks have been used previously (Abramowitz 2012) to help determine the level of performance that can be achieved based purely on the information content in the meteorological forcing data. These benchmarks are automatically calculated within the PALS system.

It is important to note that all three empirical benchmarks represent instantaneous responses to a subset of an LSM's meteorological forcing. They have no internal state variables and no information about components that may have memory of past conditions, such as soil moisture, soil temperature, or any vegetation or soil properties. These empirical benchmarks are not constrained by the surface energy balance or by sharing a common surface temperature for Q_H or Q_E, which are constraints that do apply to the physical models.

3. Results

At each site, for Q_E and Q_H separately, the statistics for all of the LSMs and all physical and empirical benchmarks are determined. Each LSM is ranked relative to all of the benchmarks, with the best performing sample element given a score of 1 and the worst given a score of 6. These rankings are then averaged over all statistics and all sites to give an average ranking for both Q_E and Q_H separately:

$$\overline{R_i} = \frac{1}{n_s n_t} \sum_{j=1}^{n_s} \sum_{k=1}^{n_t} R_{i_{jk}},$$

where i represents the LSM, physical model benchmarks, or empirical benchmark being evaluated; $\overline{R_i}$ is the average ranking for i; n_s is the number of sites; n_t is the number of statistical measures; and $R_{i_{jk}}$ is the ranking of LSM or benchmark (i) at site j for statistical measure k.

As each LSM is compared only to the benchmarks and not to other LSMs, it is possible to obtain different average rankings for the model and benchmarks when each of the LSMs is considered. Furthermore, because each statistic at each site is given a limited value between 1 and 6, it is not possible for one site or one statistic to substantially influence the overall average rankings through a particularly good or poor performance, as is the case with some of the statistical measures themselves. Hence, the results are reasonably robust.

Figure 4 shows that all models display similar average ranking compared to the benchmarks for the standard statistical metrics. This means that while the structures and the physical parameterizations of the models vary, all models utilize the information available in their forcing data to a similar degree. Figure 4 also shows that all models perform better for Q_E than for Q_H. Note that in the context of benchmarking, better performance means that a model meets more of the metrics and not

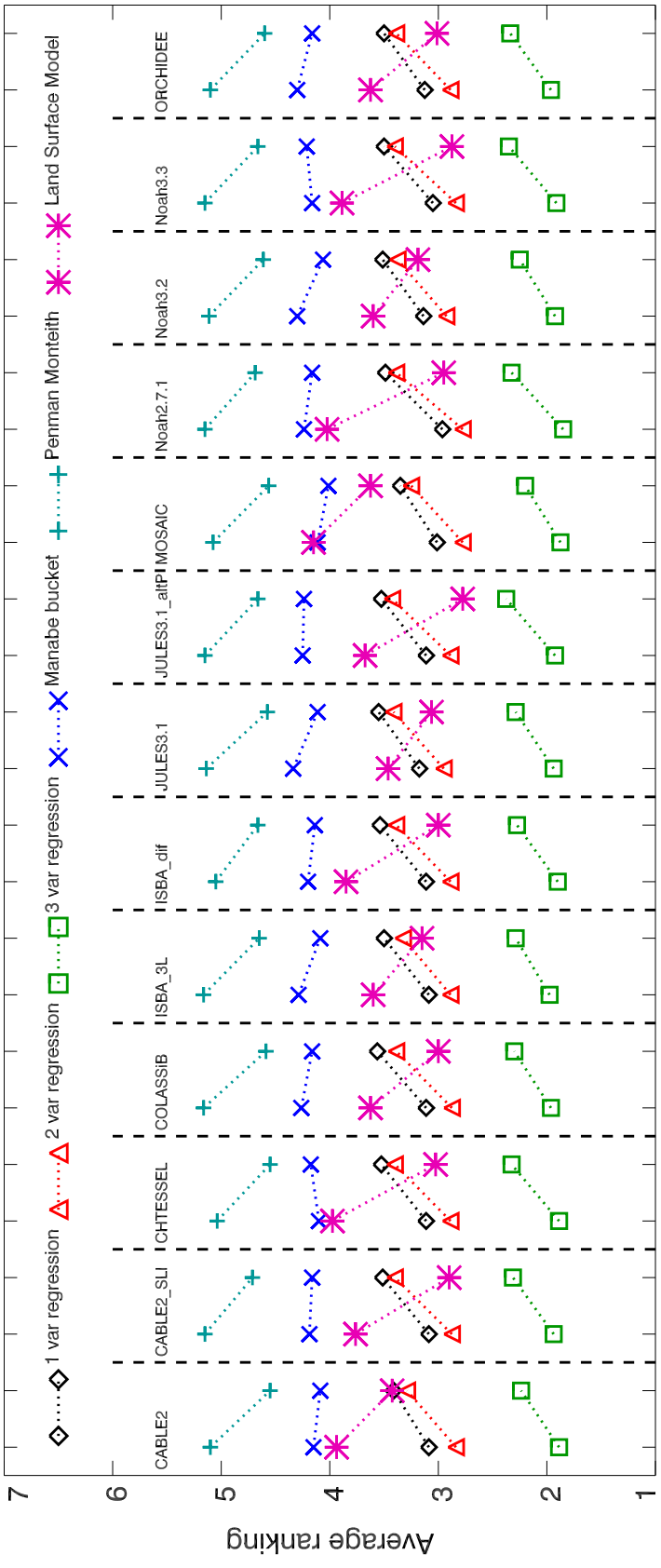

FIG. 4. Ranking of benchmarks and each model for the standard statistics (MBE, NME, SD, r) across all 20 sites. A ranking of 1 corresponds to the best performance. The dotted lines are a visual guide and have no scientific relevance.

that the absolute errors in a chosen statistic are necessarily smaller. This result is counterintuitive since the conceptual representation of Q_H is considerably less complex than Q_E (Pitman 2003). Simulating Q_H depends only on temperature gradients and atmospheric turbulence while Q_E depends not only on the turbulence, but also requires a model to represent water availability in order to determine the moisture gradients. This occurs through stomatal conductance, canopy interception, and soil water availability for transpiration and bare soil evaporation.

Note that the physically based benchmarks (M69 and PM) also show better performance in Q_E than Q_H. Performance in this sense does not mean that errors for Q_E are smaller than Q_H, but rather that the utilization of information for predicting Q_E is better. Referring back to our discussion on the third type of benchmark in the introduction, we would expect a physical model to be better at simulating the more complex Q_E compared to the empirical benchmarks, as Q_E is influenced by additional information on soil moisture that is not used by the regressions. This is essentially illustrated by the empirical benchmarks that have the opposite behavior in the average rankings compared to the LSMs and physically based benchmarks, showing better performance for Q_H than Q_E.

Comparing LSMs to physical benchmarks shows that all LSMs beat the PM and M69 for both Q_E and Q_H, as might be expected. The considerable history of LSM evolution has clearly delivered notable performance improvements (Pitman 2003). This evolution is also evident in the increased performance delivered by the simplistic water limitation that M69 provides, as opposed to the unrestricted water availability in the reference crop PM. However, all LSMs are outperformed by EMP3KM27 for both Q_H and Q_E. Even more striking is that all models are beaten by all empirical benchmarks for Q_H, including an out-of-sample linear regression against SWdown. For some LSMs, this is also true for Q_E. We provide some comments on why this might be the case in section 4.

A comparison of the LSM ranks against the benchmarks for the metrics on the extremes of the distribution is shown in Fig. 5, while Fig. 6 shows the average rankings when derived from the statistical measures for the shape of the distributions. The empirical benchmarks by nature act as a smoother (since they are regression based), so it would be expected that these benchmarks would not be good at predicting the extremes of the distribution. They are likely to be more peaked around the mean values. This will especially impact the 5th and 95th percentile metrics in Fig. 5 and the kurtosis and overlap metrics used in Fig. 6.

The results in Fig. 5 do indeed show that the LSMs perform well compared to the empirical benchmarks, with most LSMs beating all of them, particularly for Q_E. There is still an indication that the LSMs are better at simulating the extremes of Q_E than they are for Q_H, as some of the LSMs are outperformed by the two and three variable regressions for Q_H. The LSMs also have a better ranking than the physically based benchmarks, although M69 performs better than the single variable regression for the extremes. This is understandable: the M69 model tends to dry out too fast and since many extremes are associated with dry landscapes, M69 captures this, though not necessarily for the right physical reasons.

While the average rankings for the standard statistics and the 5th and 95th percentile metrics are similar between the models (Figs. 4, 5), their performance for the statistics based on the shape of the distribution of the fluxes does not show a clear signal for all of the LSMs. Some of the models perform better for Q_E compared to the benchmarks, whereas others perform better for Q_H. Despite the empirical regressions being more peaked around the mean of the distribution, some of the LSMs perform worse than all three regressions for Q_E while one model is also worse for Q_H. In addition, several of the models are worse than the physical benchmarks for either Q_H or Q_E.

The results for the models compared to the benchmarks for the extremes of the distribution shown in Fig. 5 are an average over all of the sites. However, as the empirical benchmarks are determined out of sample, we might expect that the LSMs should have the best rankings for the extremes at the sites with lowest and highest SWdown and annual mean precipitation (Fig. 3), that is, the climatic extremes from our sample of sites. Figure 7 shows the rankings of the LSMs and the benchmarks for the extremes of the distribution at each of the sites listed in Table 1. There is a figure for both Q_H and Q_E for sites ordered in terms of their average downward solar radiation and in terms of their annual mean precipitation. Shown in each figure is a box plot for all sites showing the range of rankings for the LSMs, along with the median of the LSMs and each of the benchmarks. The codes used for each site are given in Table 1.

The rankings for Q_H from the LSMs are relatively worse compared to the benchmarks at sites with the largest downward shortwave radiation (Fig. 7a). This suggests that the LSMs use the information content from the atmospheric variables inappropriately at these sites. There is no discernible change in the rankings for Q_E across the sites ordered by the downward solar radiation (Fig. 7b).

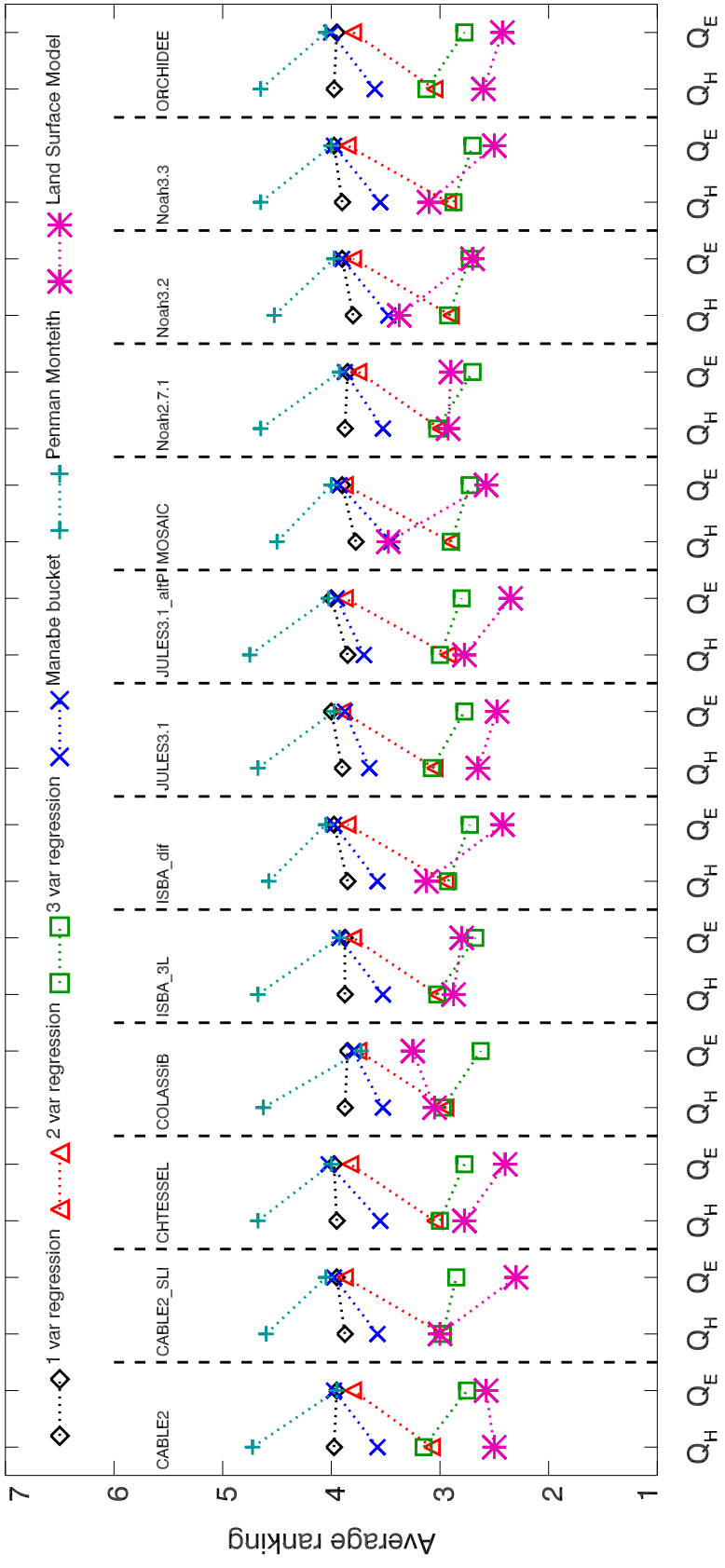

FIG. 5. As in Fig. 4, but for the 5th and 95th percentile metrics.

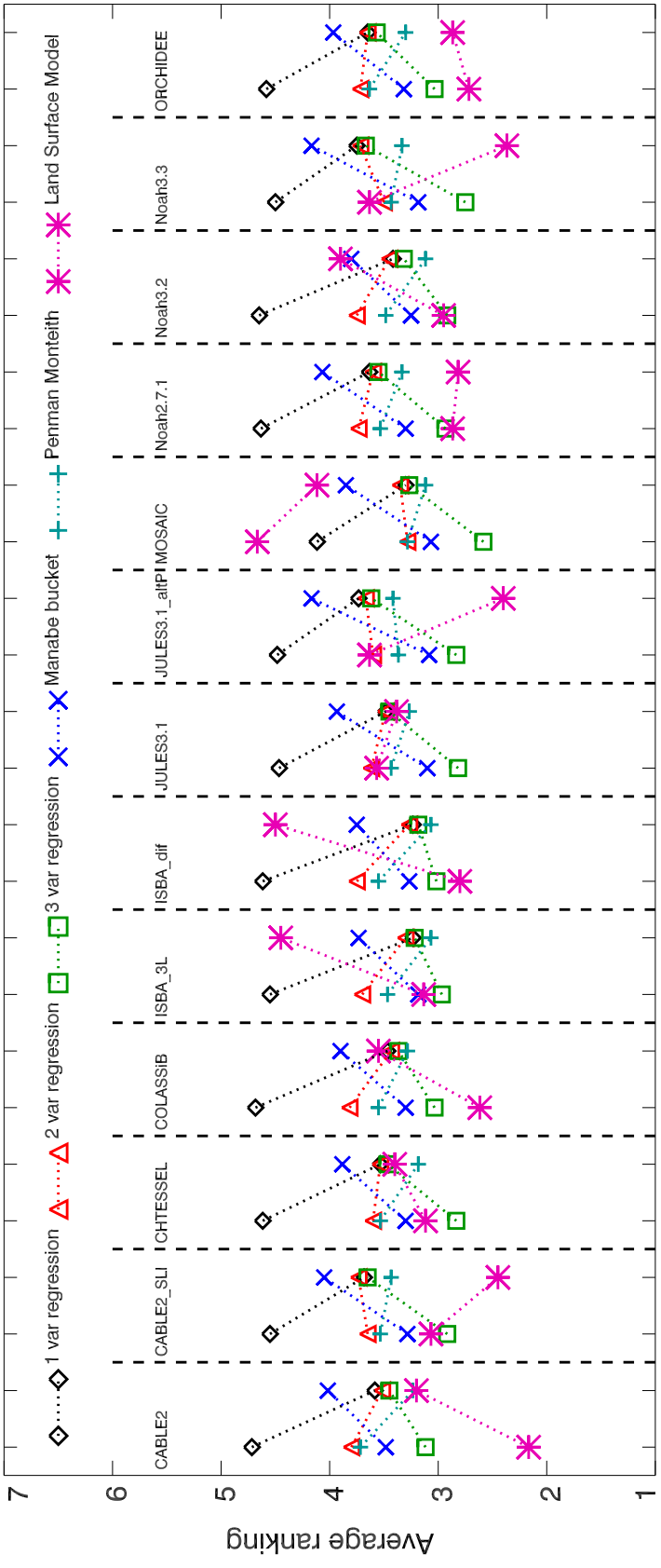

FIG. 6. As in Fig. 4, but for the distribution metrics (skewness, kurtosis, overlap).

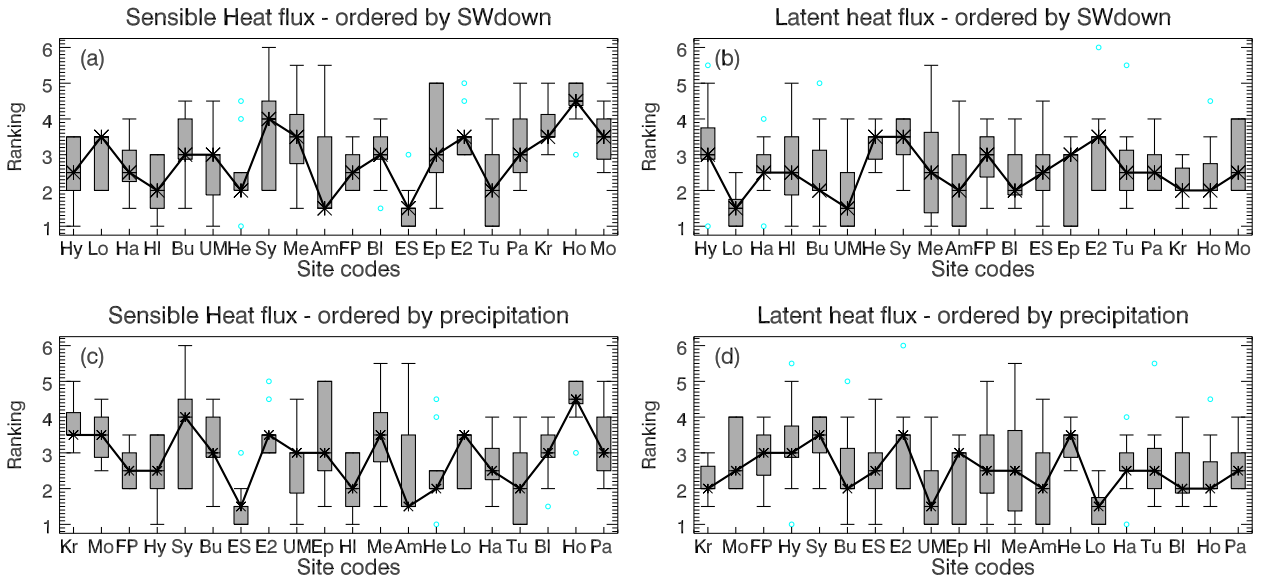

FIG. 7. Box-and-whisker plot showing the variations in ranking for the extremes of the distribution across the models for all sites. The whiskers extend to the maximum and minimum values of the data, or 1.5 times the interquartile range if this is smaller. Outlying data beyond this range are identified by circles. The codes for each site are given in Table 1. Also shown by the line are the median rankings for the LSMs across all sites. (a) Sensible heat flux and (b) latent heat flux ordered by mean annual downward shortwave radiation across the sites; (c) sensible heat flux and (d) latent heat flux ordered by mean annual precipitation across the sites.

When ordered by annual mean precipitation, there is no discernible change to the LSMs' rankings across the sites for Q_H (Fig. 7c). The LSMs rankings for Q_E at the driest sites compare favorably to the other sites (Fig. 7d), but there is a suggestion that some of the sites toward the dry end have worse rankings for the LSMs compared to the benchmarks. These are the sites where soil moisture might have its greatest impact on restricting transpiration, but not completely preventing it.

4. Discussion

One of the aims of this study was to introduce the concept of benchmarking and to identify the benefits of this approach compared to the more traditional evaluation or comparison studies. Many previous model studies have presented results on the evaluation of models and have helped to reduce the errors for given statistical metrics, while community projects have concentrated mainly on model comparison studies that have in addition helped to identify aspects of LSMs' performance that can be improved. However, neither of these approaches can tell us whether any of the LSMs are actually good models: that is, whether they adequately utilize the information about Q_E and Q_H available in the meteorological forcing. While the comparison with M69 and PM shows that these LSMs are significantly better than their predecessors, the performance against the out-of-sample empirical models shows that there is significant scope for improvement without the need for more parameter or time-varying input data.

We also note that simply evaluating performance using metrics such as RMSE alone without benchmarks can be misleading. It might well be true, for example, that in RMSE terms, one flux variable is inherently easier to predict than another. In this case, reporting a lower RMSE for that flux might mislead a researcher into thinking that their model is better at predicting that flux. By using benchmarks as we have done here, the empirical models reflect how much information is available to an LSM about each flux—that is, how inherently difficult it is to predict that flux. Thus, beating the same empirical model in each flux represents the same level of performance in each flux, regardless of the RMSE values.

An additional advantage of the benchmarking approach is that, as an international community, we can identify a stable set of benchmarks that could be used to assess progress within the community over a number of decades. Experience shows that there is limited motivation to rerun a model with old forcing datasets, or indeed to rerun old versions of a specific model. However, simple established benchmarks, such as M69, PM, or the empirical benchmarks, could easily be maintained to demonstrate and quantify future advances. For example, if within several years all LSMs reach the ranking of the three-variable nonlinear regression for both Q_H and Q_E, then the community can demonstrate success and quantify the scale of progress.

As this study was designed to identify the merits of benchmarking, we have adopted a simplified approach by only considering the turbulent fluxes of heat and moisture from the surface. We believe this is a good start, but these fluxes alone do not entirely characterize the performance of LSMs. Clearly, the concept of benchmarking should be extended to other metrics, such as additional variables with equivalent in situ measurement networks (e.g., soil moisture), or specific phenomenon (e.g., a drought) that will help to give a more balanced picture of the suitability of these LSMs for a range of applications.

We reiterate that this project does not attempt to compare the LSMs—we are not attempting to identify which is the ''best'' model. All of the models have a similar level of skill compared to the benchmarks for the standard statistics, but this is less true for the statistics based on the distribution of the fluxes. This might suggest that all of the models have been involved in similar evaluation and comparison studies that have used these standard statistical measures for their assessment and identifying subsequent development priorities. However, it is unlikely that these standard statistics adequately define the purpose of the model.

Using the basic physical models of M69 and PM, the results show that all of the models pass the benchmark for most of the statistical measures. This shows that progress has been made throughout the community over the last couple of decades in terms of the development of LSMs. However, relative to the empirical benchmarks, the models do not pass all of the metrics. This suggests that there is more information content in the atmospheric forcing data to determine the Q_H and Q_E than is currently used by the LSMs. This is especially true for Q_H, where none of the models pass the single variable regression benchmark, suggesting that there is enough information in SWdown alone to predict Q_H with a higher degree of precision than LSMs currently do. The result is reemphasized when the sites are ordered by their mean annual SWdown. Here we find that the performance of the models decreases compared to the benchmarks for the sites with the highest annual mean SWdown.

The variable Q_E is strongly influenced by moisture availability information. This can be confirmed by the fact that the three-variable nonlinear regression is the only benchmark that is not passed by the models for Q_E. This is the only empirical benchmark to contain any moisture information, although this is through atmospheric humidity rather than soil moisture that controls Q_E in the LSMs. The suggestion is that in this case, there is more information in the instantaneous atmospheric humidity about the control of evaporation than there is from the memory of soil moisture control in the LSMs (or that inappropriate soil moisture values may in fact be hindering prediction). This suggests that our physical understanding of how soil moisture influences evaporation may not be correct. The performance of the models relative to the benchmarks at sites ordered by their total annual precipitation suggests that in the situations where we may expect the soil moisture to have a dominant control on stomata and hence Q_E, the LSMs tend to be slightly worse. Meanwhile, Koster and Mahanama (2012) suggest still more can be gained from improvement of the soil moisture–runoff relationship in LSMs, which are not considered in this study in either the benchmarks or the regression models.

The combination of these results provides us with an opportunity to challenge our conceptual view of energy partitioning at the surface. The traditional view is that there is an available amount of energy set by radiative processes, and the role of the surface energy balance in the LSMs is to distribute this energy between Q_H and Q_E. However, if there is sufficient information in SWdown to determine Q_H, then perhaps the role of the surface energy balance in reality is to distribute the remaining energy between Q_E and the energy flux exchange with the underlying soil. The implications of this might be that our equation set used to solve for the surface fluxes is not correct. For instance, the assumption that both radiative, turbulent fluxes and soil fluxes share the same physical surface temperature might need to be reconsidered. We do not discuss soil heat flux here or the influence that the bare soil fraction has on our results. More information concerning the footprint of the observations is required to determine the distribution of bare soil and vegetation and the resulting measured soil and turbulent fluxes. Furthermore, the indifferent performance of the models at sites with restricted soil moisture questions the current methods used in the LSMs for representing the stomatal control on transpiration.

Finally, we note that we have not tested the significance of these results because of the small sample size. However, we would expect that the statistics for the higher-order moments to be progressively noisier among the models just based on the nature of variance.

The intercomparison presented here was designed to be a stand-alone study of LSMs, that is, there is no feedback between the surface fluxes and the atmospheric driving data. A fully coupled system contains errors from all model components and sensitivity to feedbacks, and as such is a complex system. Developing benchmark metrics that can assess the whole coupled system would be an ultimate objective, but remains a challenge that is beyond the scope of this paper.

5. Conclusions

We used 13 LSMs with 20 observational sites to examine the utility of benchmarks to inform us about the ability of existing LSMs. Benchmarking is a fundamentally different way of assessing the skill of a model compared to evaluation or comparison, because an expected level of performance for a particular metric is set a priori. Benchmarking can help to identify future development criteria not based on the largest errors from a standard statistical metric, but by the demonstrated capacity for improvement in a metric without the need for additional driving or parameter data. Although this study has been limited to offline surface schemes at single point locations, future benchmarking will evolve both horizontally (distributed) and vertically (coupled).

While our results for the LSMs vary according to the statistical measures used, a key finding is that LSMs perform better across all of the sites than the simple physical models that were used as benchmarks. This demonstrates the progress that has been made by the community over the last couple of decades. However, the LSMs are outperformed by the nonlinear three-variable empirical regression for Q_E and all of the empirical regressions for Q_H, including a linear regression between downward shortwave radiation and Q_H. This suggests that the LSMs do not appropriately use the information available in the atmospheric forcing data when estimating Q_H and Q_E. A second key result is that the LSMs perform worse compared to the benchmarks for Q_H at sites with the largest downward shortwave radiation. Clearly, the community should investigate the relationship between shortwave radiation and surface sensible heat flux more thoroughly. For Q_E, the models are worse at some of the sites with low annual mean precipitation, but not the driest sites. This suggests that the community should also investigate the relationship between soil moisture and transpiration to determine the limitations of current LSMs.

Long-term energy balance issues at some of the observational sites mean that the results presented in this study require further work to ensure that energy balance closure is not significantly affecting our results. A comprehensive list of sites with a good range of climates and biomes that conform to energy balance constraints are unlikely to be available in the near future, so alternative approaches may be required to address these issues.

Our results also demonstrate the ability of the PALS web-based system as a benchmarking tool for the community. We suggest that this should be developed into an international standard for land surface benchmarking, with metrics agreed by the user community. Such a tool would rapidly advance the science and deliver measureable improvements in our understanding and modeling capabilities.

In conclusion, our results challenge our traditional conceptual view of the surface energy balance where available energy is partitioned into the sensible and latent heat fluxes. Our first attempt to use benchmarking across the land modeling community has highlighted some uncertainty in the fundamental conceptual understanding of LSMs. While evaluation of models will remain a valuable tool for helping to quantify development requirements, we suggest that a more systematic use of benchmarking across the community should be encouraged. The benchmarking approach is likely to identify more serious challenges to land modeling and thereby accelerate improvements in our science.

Acknowledgments. M. Best and H. Johnson were supported by the Joint DECC/Defra Met Office Hadley Centre Climate Programme (CA01101). We acknowledge the support of the Australian Research Council Centre of Excellence for Climate System Science (CE110001028). This work used eddy covariance data acquired by the FLUXNET community and in particular by the following networks: AmeriFlux [U.S. Department of Energy, Biological and Environmental Research, Terrestrial Carbon Program (DE-FG02-04ER63917 and DE-FG02-04ER63911)], AfriFlux, AsiaFlux, CarboAfrica, CarboEuropeIP, CarboItaly, CarboMont, ChinaFlux, FLUXNET-Canada (supported by CFCAS, NSERC, BIOCAP, Environment Canada, and NRCan), GreenGrass, KoFlux, LBA, NECC, OzFlux, TCOS-Siberia, and USCCC. We acknowledge the financial support to the eddy covariance data harmonization provided by CarboEuropeIP, FAO-GTOS-TCO, iLEAPS, Max Planck Institute for Biogeochemistry, the National Science Foundation, Tuscia University, Université Laval and Environment Canada, and the U.S. Department of Energy and the database development and technical support from Berkeley Water Center; Lawrence Berkeley National Laboratory; Microsoft Research eScience; Oak Ridge National Laboratory; University of California, Berkeley; and University of Virginia.

REFERENCES

Abramowitz, G., 2005: Towards a benchmark for land surface models. *Geophys. Res. Lett.,* **32,** L22702, doi:10.1029/2005GL024419.

——, 2012: Towards a public, standardized, diagnostic benchmarking system for land surface models. *Geosci. Model Dev.,* **5,** 819–827, doi:10.5194/gmd-5-819-2012.

——, L. Pouyanné, and H. Ajami, 2012: On the information content of surface meteorology for downward atmospheric long-wave

radiation synthesis. *Geophys. Res. Lett.,* **39,** L04808, doi:10.1029/2011GL050726.

Allen, R. G., L. S. Pereira, D. Raes, and M. Smith, 1998: FAO Penman–Monteith equation. Crop evapotranspiration: Guidelines for computing crop water requirements, FAO Irrigation and Drainage Paper 56, 17–28. [Available online at www.fao.org/docrep/X0490E/X0490E00.htm.]

Balsamo, G., P. Viterbo, A. Beljaars, B. J. J. M. van den Hurk, A. Betts, and K. Scipal, 2009: A revised hydrology for the ECMWF model: Verification from field site to terrestrial water storage and impact in the Integrated Forecast System. *J. Hydrometeor.,* **10,** 623–643, doi:10.1175/2008JHM1068.1.

Best, M. J., and C. S. B. Grimmond, 2013: Analysis of the seasonal cycle within the first international urban land surface model comparison. *Bound.-Layer Meteor.,* **146,** 421–446, doi:10.1007/s10546-012-9769-7.

——, and ——, 2014: Importance of initial state and atmospheric conditions for urban land surface models' performance. *Urban Climate,* **10,** 387–406, doi:10.1016/j.uclim.2013.10.006.

——, and Coauthors, 2011: The Joint UK Land Environment Simulator (JULES), Model description—Part 1: Energy and water fluxes. *Geosci. Model Dev.,* **4,** 677–699, doi:10.5194/gmd-4-677-2011.

Boone, A., J.-C. Calvet, and J. Noilhan, 1999: Inclusion of a third soil layer in a land-surface scheme using the force–restore method. *J. Appl. Meteor.,* **38,** 1611–1630, doi:10.1175/1520-0450(1999)038<1611:IOATSL>2.0.CO;2.

——, and Coauthors, 2009: The AMMA Land Surface Model Intercomparison Project (ALMIP). *Bull. Amer. Meteor. Soc.,* **90,** 1865–1880, doi:10.1175/2009BAMS2786.1.

Boussetta, S., and Coauthors, 2013: Natural land carbon dioxide exchanges in the ECMWF integrated forecasting system: Implementation and offline validation. *J. Geophys. Res.,* **118,** 5923–5946, doi:10.1002/jgrd.50488.

Bowling, L. C., and Coauthors, 2003: Simulation of high latitude hydrological processes in the Torne–Kalix basin: PILPS Phase 2(e): 1: Experiment description and summary intercomparisons. *Global Planet. Change,* **38,** 1–30, doi:10.1016/S0921-8181(03)00003-1.

Chen, T. H., and Coauthors, 1997: Cabauw experimental results from the Project for Intercomparison of Land-Surface Parameterization Schemes. *J. Climate,* **10,** 1194–1215, doi:10.1175/1520-0442(1997)010<1194:CERFTP>2.0.CO;2.

Decharme, B., A. Boone, C. Delire, and J. Noilhan, 2011: Local evaluation of the Interaction between Soil Biosphere Atmosphere soil multilayer diffusion scheme using four pedotransfer functions. *J. Geophys. Res.,* **116,** D20126, doi:10.1029/2011JD016002.

Dirmeyer, P. A., 2011: A history and review of the Global Soil Wetness Project (GSWP). *J. Hydrometeor.,* **12,** 729–749, doi:10.1175/JHM-D-10-05010.1.

——, and F. J. Zeng, 1999: An update to the distribution and treatment of vegetation and soil properties in SSiB. COLA Tech. Rep. 78, Center for Ocean–Land–Atmosphere Studies, Calverton, MD, 27 pp.

——, A. J. Dolman, and N. Sato, 1999: The pilot phase of the Global Soil Wetness Project. *Bull. Amer. Meteor. Soc.,* **80,** 851–878, doi:10.1175/1520-0477(1999)080<0851:TPPOTG>2.0.CO;2.

Ek, M. B., K. E. Mitchell, Y. Lin, E. Rogers, P. Grummann, V. Koren, G. Gayno, and J. D. Tarpley, 2003: Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model. *J. Geophys. Res.,* **108,** 8851, doi:10.1029/2002JD003296.

Gong, W., H. V. Gupta, D. Yang, K. Sricharan, and A. O. Hero III, 2013: Estimating epistemic and aleatory uncertainties during hydrologic modeling: An information theoretic approach. *Water Resour. Res.,* **49,** 2253–2273, doi:10.1002/wrcr.20161.

Grimmond, C. S. B., and Coauthors, 2010: The international urban energy balance models comparison project: First results from phase 1. *J. Appl. Meteor. Climatol.,* **49,** 1268–1292, doi:10.1175/2010JAMC2354.1.

——, and Coauthors, 2011: Initial results from phase 2 of the international urban energy balance model comparison. *Int. J. Climatol.,* **31,** 244–272, doi:10.1002/joc.2227.

Guo, Z., and P. A. Dirmeyer, 2013: Interannual variability of land–atmosphere coupling strength. *J. Hydrometeor.,* **14,** 1636–1646, doi:10.1175/JHM-D-12-0171.1.

——, and Coauthors, 2006: GLACE: The Global Land–Atmosphere Coupling Experiment. Part II: Analysis. *J. Hydrometeor.,* **7,** 611–625, doi:10.1175/JHM511.1.

Haverd, V., and M. Cuntz, 2010: Soil–Litter–Iso: A one-dimensional model for coupled transport of heat, water and stable isotopes in soil with a litter layer and root extraction. *J. Hydrol.,* **388,** 438–455, doi:10.1016/j.jhydrol.2010.05.029.

Henderson-Sellers, A., Z.-L. Yang, and R. E. Dickinson, 1993: The Project for Intercomparison of Land-Surface Schemes. *Bull. Amer. Meteor. Soc.*, **74**, 1335–1349, doi:10.1175/1520-0477(1993)074<1335:TPFIOL>2.0.CO;2.

——, A. J. Pitman, B. Henderson-Sellers, and J. Verner, 1995a: Applying software engineering metrics to land surface parameterization schemes. *J. Climate,* **8,** 1043–1059, doi:10.1175/1520-0442(1995)008<1043:ASEMTL>2.0.CO;2.

——, ——, P. K. Love, P. Irannejad, and T. Chen, 1995b: The Project for Intercomparison of Land Surface Parameterization Schemes (PILPS): Phases 2 and 3. *Bull. Amer. Meteor. Soc.,* **76,** 489–503, doi:10.1175/1520-0477(1995)076<0489:TPFIOL>2.0.CO;2.

Hirschi, M., and Coauthors, 2011: Observational evidence for soil-moisture impact on hot extremes in southeastern Europe. *Nat. Geosci.,* **4,** 17–21, doi:10.1038/ngeo1032.

Kidston, J., C. Brümmer, T. A. Black, K. Morgenstern, Z. Nesic, J. H. McCaughey, and A. G. Barr, 2010: Energy balance closure using eddy covariance above two different land surfaces and implications for CO_2 flux measurements. *Bound.-Layer Meteor.,* **136,** 193–218, doi:10.1007/s10546-010-9507-y.

Koster, R. D., and M. J. Suarez, 1992: Modeling the land surface boundary in climate models as a composite of independent vegetation stands. *J. Geophys. Res.,* **97,** 2697–2715, doi:10.1029/91JD01696.

——, and ——, 1994: The components of a 'SVAT' scheme and their effects on a GCM's hydrological cycle. *Adv. Water Resour.,* **17,** 61–78, doi:10.1016/0309-1708(94)90024-8.

——, and S. P. P. Mahanama, 2012: Land surface controls on hydroclimatic means and variability. *J. Hydrometeor.,* **13,** 1604–1620, doi:10.1175/JHM-D-12-050.1.

——, and Coauthors, 2004: Regions of coupling between soil moisture and precipitation. *Science,* **305,** 1138–1140, doi:10.1126/science.1100217.

——, and Coauthors, 2006: GLACE: The Global Land–Atmosphere Coupling Experiment. Part I: Overview. *J. Hydrometeor.,* **7,** 590–610, doi:10.1175/JHM510.1.

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 Marine and Atmospheric Research Paper 013, 43 pp. [Available online at www.cawcr.gov.au/projects/access/cable/cable_technical_description.pdf.]

Krinner, G., and Coauthors, 2005: A dynamic vegetation model for studies of the coupled atmosphere–biosphere system. *Global Biogeochem. Cycles,* **19,** GB1015, doi:10.1029/2003GB002199.

Luo, Y. Q., and Coauthors, 2012: A framework for benchmarking land models. *Biogeosciences,* **9,** 3857–3874, doi:10.5194/bg-9-3857-2012.

Manabe, S., 1969: 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.

Masson, V., and Coauthors, 2013: The SURFEXv7.2 externalized platform for the simulation of Earth surface variables and fluxes. *Geosci. Model Dev.,* **6,** 929–960, doi:10.5194/gmd-6-929-2013.

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

Perkins, S. E., A. J. Pitman, N. J. Holbrook, and J. McAneney, 2007: Evaluation of the AR4 climate models' simulated daily maximum temperature, minimum temperature and precipitation over Australia using probability density functions. *J. Climate,* **20,** 4356–4376, doi:10.1175/JCLI4253.1.

Pitman, A. J., 2003: The evolution of, and revolution in, land surface schemes designed for climate models. *Int. J. Climatol.,* **23,** 479–510, doi:10.1002/joc.893.

——, and Coauthors, 1999: Key results and implications from phase 1(c) of the Project for Intercomparison of Land-Surface Parameterization Schemes. *Climate Dyn.,* **15,** 673–684, doi:10.1007/s003820050309.

Rodell, M., P. R. Houser, A. A. Berg, and J. S. Famiglietti, 2005: Evaluation of 10 methods for initializing a land surface model. *J. Hydrometeor.,* **6,** 146–155, doi:10.1175/JHM414.1.

Schlosser, C. A., and Coauthors, 2000: Simulations of a boreal grassland hydrology at Valdai, Russia: PILPS phase 2(d). *Mon. Wea. Rev.,* **128,** 301–321, doi:10.1175/1520-0493(2000)128<0301:SOABGH>2.0.CO;2.

Seneviratne, S. I., T. Corti, E. L. Davin, M. Hirschi, E. B. Jaeger, I. Lehner, B. Orlowsky, and A. J. Teuling, 2010: Investigating soil moisture–climate interactions in a changing climate: A review. *Earth Sci. Rev.,* **99,** 125–161, doi:10.1016/j.earscirev.2010.02.004.

——, and Coauthors, 2013: Impact of soil moisture–climate feedbacks on CMIP5 projections: First results from the GLACE-CMIP5 experiment. *Geophys. Res. Lett.,* **40,** 5212–5217, doi:10.1002/grl.50956.

van den Hurk, B., M. Best, P. Dirmeyer, A. Pitman, J. Polcher, and J. Santanello, 2011: Acceleration of land surface model development over a decade of glass. *Bull. Amer. Meteor. Soc.,* **92,** 1593–1600, doi:10.1175/BAMS-D-11-00007.1.

Wang, Y. P., E. Kowalczyk, R. Leuning, G. Abramowitz, M. R. Raupach, B. Pak, E. van Gorsel, and A. Luhar, 2011: Diagnosing errors in a land surface model (CABLE) in the time and frequency domains. *J. Geophys. Res.,* **116,** G01034, doi:10.1029/2010JG001385.

Wilson, K., and Coauthors, 2002: Energy balance closure at FLUXNET sites. *Agric. For. Meteor.,* **113,** 223–243, doi:10.1016/S0168-1923(02)00109-0.

Wood, E. F., and Coauthors, 1998: The Project for Intercomparison of Land-Surface Parameterization Schemes (PILPS) phase 2(c) Red–Arkansas River basin experiment: 1. Experiment description and summary intercomparison. *Global Planet. Change,* **19,** 115–135, doi:10.1016/S0921-8181(98)00044-7.

Yang, Z.-L., R. E. Dickinson, A. Henderson-Sellers, and A. J. Pitman, 1995: Preliminary study of spin-up processes in land surface models with the first stage data of Project for Intercomparison of Land Surface Parameterization Schemes phase 1(a). *J. Geophys. Res.,* **100,** 16 553–16 578, doi:10.1029/95JD01076.