Geosci. Model Dev. Discuss., 5, 549–570, 2012
www.geosci-model-dev-discuss.net/5/549/2012/
doi:10.5194/gmdd-5-549-2012

Geoscientific Model Development Discussions

Towards a public, standardized, diagnostic benchmarking system for land surface models

G. Abramowitz

University of New South Wales, Sydney, Australia

Received: 2 February 2012 – Accepted: 8 February 2012 – Published: 20 February 2012

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

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

Abstract

We examine different conceptions of land surface model benchmarking and illustrate the importance of internationally standardized evaluation experiments that specify data sets, variables, metrics and model resolutions. We additionally show how essential the definition of a priori expectations of model performance can be, based on the complexity of a model and the amount of information being provided to it, and give an example of how these expectations might be quantified. Finally, we introduce the Protocol for the Analysis of Land Surface models (PALS), a free, online land surface model benchmarking application, and show how it is structured to meet both of these goals.

1 Introduction

Land surface models (LSMs) simulate the exchange of water, heat and carbon between the land surface and atmosphere and represent these processes within climate models. Climate models in turn have evolved from extremely simplified tools used to gain a conceptual understanding of broad-scale climate features – such as continental boundary effects (e.g., Manabe, 1969) – to something more akin to operational weather forecasting tools. Climate projections now inform multi-million dollar decisions, and this is reflected in the pressures that research scientists face to provide a "*comprehensive representation of the four major components of the climate system*" (Gordon et al., 2002) "*for simulating past, present, and future climates*" (Collins et al., 2006). This change of focus has driven a commensurate transition in the nature of model evaluation from qualitative to quantitative analysis.

While LSM evaluation increasingly relies on a broad collection of data sources (e.g. in-situ gas exchange measurements, streamflow and satellite-based measurements) the limited nature of their availability and quality control historically has meant that the transition to qualitative analysis in LSM evaluation has been ad hoc. Although the term "benchmarking" has recently increased in popularity in the LSM community (e.g., Abramowitz et al., 2008; Blyth et al., 2011), there is apparent confusion regarding its meaning. In its weakest and perhaps most common usage, benchmarking is simply synonymous with model evaluation of any sort and so apparently only reflects a change in language rather than practice. Benchmarking has also been used to refer to a single institution's LSM evaluation program (e.g., Blyth et al., 2011), which would usually define previous model versions as the performance standard. The third usage, and one that we will discuss in Sect. 2, defines benchmarking as a coordinated effort to define community-wide reference data sets, spatial and temporal resolutions, variables and metrics for evaluation. We will refer to these as *standardised experiments*.

In Sect. 3 we illustrate the importance of an additional constraint on standardised experiments – the a priori specification of expectations of model performance. That is, given the complexity of a model, and the amount of information provided to it in its time-independent parameters and time-dependent input variables, how well should we *expect* it to perform? We give one possible answer to this question that recognizes that some environments are more difficult to simulate than others. We then use this solution to show how we might construct a hierarchy of performance benchmarks that could be used to rank models.

Finally, in Sect. 4, we introduce Protocol for the Analysis of Land Surface models (PALS), a web-based LSM evaluation tool that is structured to meet these goals. It acts both as a data set repository and automated evaluation tool, to be used as either a model development facility or framework for model comparison experiments, and keeps a complete version history of all the data it contains.

While we focus our discussion on LSMs designed for use in high resolution climate model simulations, we note that much of what is presented here is equally applicable to hydrological modelling or ecological modelling in areas where appropriate data sources are available.

GMDD
5, 549–570, 2012

2 Benchmarking using internationally standardised experiments

The benefits of internationally accepted standard experiments – prescribing LSM driving data, evaluation data, variables, metrics and possibly surface parameter information – are many. They allow different research teams to immediately compare results, identify shared weaknesses or strengths in LSMs and provide a fast cost-benefit analysis of any proposed modifications to a modeling system. Equally importantly, this definition of benchmarking minimizes the potentially very serious impact of the seemingly trivial modelling problems shown in Table 1. These issues, while well recognized in commercial software development, are relatively new to researchers working in science where funding sources and performance metrics rarely if ever recognize the importance and resource requirements associated with model development and management. One might speculate that the increasingly operational nature of climate projection will mean that these standards, so essential in other software development environments, cannot continue to be ignored by research managers in future.

To gauge the importance of the model traits in Table 1, try asking yourself whether the red or blue model is: more likely to be reliable; more likely to contain critical bugs; more likely to be used inappropriately; more like the model you use? We suggest that a benchmarking environment defined and maintained by a single research group is more likely to allow coding bugs or unrecognized weaknesses to propagate through successive model generations than an internationally agreed benchmarking system where evaluation against other LSMs is commonplace.

By sharing the investment required in benchmarking experiments, an internationally defined benchmarking experiment set also allows a greater depth of LSM analysis as shared experiments accrue. The process of defining this type of benchmark for the LSM community is the goal of the International Land Model Benchmarking (ILAMB) group (http://www.ilamb.org).

GMDD
5, 549–570, 2012

3 Benchmarking using a priori expectations of performance

An equally important aspect of LSM benchmarking, and one that is rarely addressed, is an assessment of the level of performance we should *expect* of LSMs. Given a variable, spatial scale, temporal scale and metric, can we specify a priori how close a model should be to observations? Put simply, what constitutes a "good" model?

For a single variable and metric, intuition might suggest that choosing the "best" model is easy – it performs best in the given metric. Yet there are several critical caveats to this response that indicate it is not a satisfactory answer to what defines a "good" model. If nothing else, it rules out the very real possibility that the "best" model is in fact a poor model. Ginzburg and Jensen (2004) give the excellent example of Ptolemy's epicycle model explaining the motion of the solar system's planets through night sky as well as Newtonian mechanics, despite its absurd physical representation. We suggest four criteria by which to judge a good model. Performance is just one of these:

1. *The simplicity of a model*. This criterion is essentially the principle of parsimony or Occam's razor – a simpler model is preferred to a complicated model where they perform to a similar standard. Simpler in this case can refer to the functional representation of relationships between quantities or the number of internal parameters. Simpler, more succinct models are preferred as they are easier to understand and diagnose when they behave in unexpected ways.

2. *The amount of information provided to a model*. A model that requires fewer time-dependent driving variables and fewer parameters describing its operating environment is preferred over one that requires more, where it performs to a similar standard. It should be clear that (1) and (2) are both essentially principles of parsimony, applied to different aspects of modelling. The motivation for their separation will be made clear below.

3. *Identifiability or physical representativeness of a model*. This is the property that separates so called "physically-based" models from empirical/statistical models. The internal variables and structure of a purely empirically fitted model, such as a feed-forward neural network, bear no resemblance to variables measurable in the physical system. Models whose internal variables purport to be quantities associated with the physical system are generally known as "physically-based" models, although when these quantities are unmeasured or unmeasurable the distinction between these two categories of models becomes ambiguous. It would seem sensible to define a physically-based model as one that requires no calibration, as a calibration data set tunes a model to the time, location and circumstances of the calibration data set, rendering it at least partly empirically-based. Even a physically-based model by this definition, used in an environment where key parameter values are not available, must at least in part be considered an empirical model. Needless to say physically based models are preferred.

4. *How well a model performs out of sample*. Model performance in a given metric must be assessed out of sample. That is, the data used to assess the model must not have been used in the model's calibration or development. Performance on calibration data should not be used for evaluation.

A "good" model therefore need not be the best performing model – it may be "good" because of it's ability to provide adequate simulations with very little input data, the simplicity of its algorithms, or the ability of its constituent variables to be unambiguously identified with those in the natural system it simulates. We therefore suggest that a priori expectations of performance in out of sample experiments should in some way take these considerations into account. One should have lower expectations of a simple model than of a complicated one.

One approach to doing this that explicitly considers three of these four criteria is empirical benchmarking (e.g., Abramowitz et al., 2008). This essentially involves training an entirely empirical model (such as regression or neural network based approach) to

GMDD
5, 549–570, 2012
A benchmarking system for land surface models
G. Abramowitz

do the job of a LSM, and testing the empirical model out of sample on the LSM evaluation data set. We can then manipulate which input variables the empirical model uses as well as its complexity in order to gauge the level at which our LSM is performing.

An example is shown in Fig. 1. It shows smoothed four-year time series of latent heat flux (LH) at a single flux tower site (Tumbarumba; see Leuning et al., 2005). Observations are shown in black and a LSM simulation (the Community Atmosphere Biosphere Land Exchange model (CABLE); see Wang et al., 2011), driven with site meteorology, vegetation type, soil type, reference height and vegetation height, is shown in blue. In most circumstances, an author would suggest this is a competent, even very good LSM simulation. The two curves are clearly highly correlated and regularly overlap.

The red curve in Fig. 1 shows a simple empirical model simulation of LH. First, we performed a linear regression between downward shortwave radiation (SW) and LH, using data from 30 sites globally that did not include the Tumbarumba site – around 2 500 000 time steps of data. Then these two regression parameters were used to predict LH at Tumbarumba, based solely on SW, at a half-hourly time step. While we can see in Fig. 1 that this extremely simple empirical model benchmark has low variance (as we might expect from a linear regression), it nevertheless outperforms the LSM in root mean square error (not shown) and normalized mean error (NME, shown in Fig. 1), both in the smoothed time series shown in Fig. 1 ("Score_mooth") and the original half-hourly time series ("Score_all").

An identically structured empirical model is used in Fig. 2 to predict net ecosystem exchange of CO_2 (NEE), again only as a function of SW. It shows the average diurnal cycle across several years of a single site, divided into four seasonal panels. This time we see that this commonly used qualitative metric again appears convincingly simulated by this simple regression model, with NME reflecting this in all seasons. In these two metrics, at least in this instance, this LSM is performing comparably to a linear regression against sunlight.

Not all examples are this revealing of course, these were deliberately chosen to highlight the utility of this approach, but it does illustrate the importance of what we

GMDD
5, 549–570, 2012

might call *a priori benchmarking*. Qualitative similarity between modelled and observed curves, so often passed as rigorous model evaluation, may in fact tell us very little about model performance. Using an empirical model in this way reveals: the extent to which LH is predictable from SW alone; how a very simple functional relationship appears in familiar diagnostic measures; and how predictable LH is, out-of-sample, at the Tumbarumba site. Since empirical model performance will be poorer at sites that exhibit unusual or unique behaviour, this approach implicitly recognizes that some environments are more difficult to simulate than others. It also gives a model-like time series and so can provide a benchmark level of performance in any chosen metric.

Using this approach, we can construct a hierarchy of benchmarks that can be used to assess how well a model is performing relative to its complexity and the amount of information provided to it in its inputs and parameters. By making a comparison similar to that shown in Figs. 1 and 2, using empirical models that vary in their complexity and the variables that are provided to them, we can rank a LSM's performance. Figure 3 gives an example. It shows probability density functions of sensible heat flux (SH) as observed (black) at Tumbarumba, as simulated by a LSM (blue), and as predicted by three increasingly complex empirical models (red, yellow, green). These empirical models are: (1) the linear regression against SW discussed above, (2) a multiple linear regression of SH against both SW and surface air temperature (T), and (3) a k-means clustering of the time series of SW, T and wind speed (W), with a multiple linear regression between (SW, T, W) and SH performed at each cluster. In this example, 243 clusters were used. This simply creates a piecewise linear functional representation of the relationship between (SW, T, W) and SH in the training data set. More generally, this hierarchy of benchmarks could also use energy conservation and observational uncertainty as part of its definition, as illustrated in Table 2.

We choose to use flux tower data here for three reasons. Firstly, it allows us to construct an empirical model that operates at the same time step size and using the same input data as the LSM. The types of functional relationships between inputs and outputs seen in the empirical model should therefore be very similar to those of the LSM.

GMDD
5, 549–570, 2012

Next, flux tower data has directly measured meteorological drivers at the same time and spatial scale as the measured fluxes used for evaluation. While we acknowledge the significant uncertainties associated with flux tower data, particularly surrounding energy conservation (e.g., Wilson et al., 2002; Kidston et al., 2010), using coincident driving and evaluation products that involve little or no interpolation or additional modelling means that this data source offers unparalleled model constraint. We are as close as is possible with current data availability to having a error free driving data and so as close as is currently possible to true diagnostic model evaluation. Finally, flux towers are one of the very few data sources that provide data in quantities that allow for the construction of robust empirical models. While the results above use 30 flux tower sites (around 2 500 000 model time steps), the La Thuile Fluxnet release contains around 500 sites. It is a goal of the PALS project described below to continue to process flux tower data for LSM evaluation as it is made available.

While we feel this approach offers the best option for a priori benchmarking, we also acknowledge that evaluation at flux towers does not by any means constitute complete LSM evaluation. Larger spatial and temporal scale features produced by LSMs in coupled models are a key aspect of climate prediction, and this is undoubtedly the ultimate purpose for most LSMs. These features are, however, emergent properties of LSMs and their atmospheric model counterparts (or forcing data), so that untangling cause and effect in circumstances of uncertain or error-prone forcing data can be extremely difficult. Accordingly, model evaluation for the diagnosis of model deficiencies can also be very difficult in this context. While it is also commonly argued that LSMs are designed to simulate grid cells rather than point-scale data, we note that LSMs have no explicit length scale, and that LSMs rarely if ever undergo fundamental change when run at different resolutions within a coupled model environment.

The process above gives us an idea of how good a model is relative to its complexity and the information it is provided with, but it cannot answer the somewhat more subjective question *how good is good enough*? The validity of a model in the sense that Oreskes et al. (1994) and Medlyn et al. (2005) describe is entirely user-dependent.

GMDD
5, 549–570, 2012

A user may not care that a complex LSM performs on par with a linear regression against meteorological variables if their purpose is simply to resolve diurnal flux cycles. It is perhaps even unusual that the complexity of a model is tailored to its application. Indeed this is arguably the state of LSM use within climate models today. While most current generation LSMs have 30–40 vegetation and soil description parameters, almost all are provided only with a vegetation type and soil type for each location (typically from a choice of 20 possible types globally). Put differently, the parameter information required for these relatively complex LSMs is not available at the global scale, so parameter values are fitted to effective "types" and calibrated with available data belonging to each type. This over parameterized approach risks calibrating LSMs to the particular variables, metrics, time and spatial scales used in their calibration.

4 The Protocol for the Analysis of Land Surface models (PALS)

PALS (http://www.pals.unsw.edu.au) is an automated web application for the diagnostic evaluation of LSMs that tries to meet the two goals outlined in the two previous sections. We spend some time below outlining the general structure of the PALS before detailing its first phase of implementation and future developments.

PALS performs several functions simultaneously. First, it acts as a repository for quality controlled, standardised-format LSM driving and evaluation data sets, and maintains a complete version history of each data set. Subsets of PALS data sets are aggregated into *experiment* structures, each of which may contain LSM forcing data sets, information for constraining LSM parameters and evaluation data sets.

PALS also allows upload of model output data files associated with a PALS *experiment*. Each time a LSM output is uploaded, ancillary files associated with it may also be uploaded. For example, one may wish to upload simulation log files, namelist files, control files, parameter files or even the model code associated with a particular simulation as a way of ensuring the reproducibility of a simulation. Unless a user decides to delete their model output contributions to an experiment, PALS will maintain the

GMDD
5, 549–570, 2012

complete version history of Model output experiment submissions.

Once LSM output is uploaded, PALS automatically runs a range of analyses comparing LSM output and observed data. Particular types of analysis are associated with particular types of experiments – examples of analyses associated with a single flux tower based experiment are shown in Figs. 1 through 4. For this to work, of course, LSM output files need to be in a standardised format. PALS currently reads ALMA-based netcdf output files (CABLE – Wang et al., 2011; ORCHIDEE – Krinner et al., 2005; JULES – Blyth et al., 2006) as well as CLM's netcdf format (Levis et al., 2004; Oleson et al, 2004). Currently all automated analyses use R (http://www.r-project.org/) to generate graphics, and the PALS R package containing all analysis scripts is available upon request.

Uploaded model outputs are labeled as either "public" or "private" by the user. Analyses of "private" outputs are available only to the submitting user, who then effectively uses PALS as a private model development tool. New versions of a model are run, results are uploaded and assessed on PALS. Analyses of "public" model outputs are available to all PALS users. While not yet implemented, a structure to allow a higher-order set of analyses associated with each experiment is being developed. These show the aggregate behaviour of all public model simulations associated with a given experiment, somewhat like an automated, ongoing Project for the Intercomparison of Land-surface Parameterization Schemes (PILPS; Henderson-Sellers et al., 1996) experiment.

Where possible, a single scalar metric is associated with each analysis type. This is intended to aid decision making when comparing two or more model versions across a wide range of metrics. While not yet implemented, a report generation facility is in development to give a multiple-page document summarizing metrics from several models or model versions, specifically for this purpose.

Additionally, PALS allows users to nominate up to three benchmark time series to help evaluate the performance of their LSM output. These can be toggled on/off most existing analysis graphs, with scalar metrics shown for benchmarks in addition to model

GMDD
5, 549–570, 2012

results. By default, these three benchmarks are three empirical models, as described in Sect. 3, applied to the user's current experiment. As well as empirical models, public LSM outputs from any user associated with the same experiment can also be nominated as benchmarks, although this extension of a priori benchmarking is yet to be implemented.

All of the above features can be accessed in PALS using two different modes. The first is simply within the main PALS database, where a "public" LSM output's analyses are available to all users. The second mode is within a PALS *workspace*. A workspace can be created by any user, who becomes the workspace owner. The workspace owner can then invite a subset of PALS users to participate in a workspace, and all data sets, models and public LSM outputs are viewable only to the workspace users. Private LSM outputs remain entirely private in both modes.

Phase 1 of PALS' implementation centres on flux tower data, for the reasons described in Sect. 3 above. PALS currently hosts data from more than 50 flux towers in around 20 countries, with all data taken from the Fluxnet La Thuile free-usage release (http://www.fluxdata.org) and some additional quality control and gap-filling performed. Details about additional processing for each site are available at each site's webpage within PALS. Only consecutive whole years of data were considered, and years with large sections of missing meteorological or flux data were not used. Both meteorological driving data and flux evaluation data are available on the PALS site as ALMA formatted netcdf files.

Currently, the complete analysis set on single flux tower experiments includes the generation of around 50 graphs across 6 LSM output variables and takes around one minute of server processing time to complete. These graphs include: probability density function overlap with observations; smoothed time series; model vs. observed scatter plots; Taylor diagrams; seasonally discrete average diurnal cycles; average annual cycle; smoothed evaporative fraction; and conservation of energy checks. Most include a scalar metric as described above. Where data and model output are available, these analyses operate on: net radiation; net shortwave radiation; latent heat flux;

GMDD
5, 549–570, 2012

sensible heat flux; ground heat flux; and net ecosystem exchange of CO_2. Benchmarks in Phase 1 are restricted to comparison with prescribed empirical model time series.

In addition to these LSM-focused features, Phase 1 of PALS maintains the ability for flux tower investigators to directly maintain their data on the PALS site. When a new flux tower data set version is uploaded, PALS runs a suite of automated analyses scripts that explore the properties of uploaded data, including energy conservation and the timing of gap-filling, where meta-data has been included. Data sets are uploaded in a standardised spreadsheet format.

Phase 2 of PALS is likely to include coarse gridded global analysis of variables such as albedo, snow cover, runoff from a selection of catchments globally as well as a comparison of continental-scale water and carbon budgets. Experimental protocols for these are being developed through the International Land-Atmosphere Model Benchmarking project (http://www.ilamb.org).

While PALS is still in development, as a community-based project feedback of any nature is welcomed. Contributions in the form of additional analyses, features, or programming support (in R, Java or Flash) are actively encouraged. Both the analysis and website code are available on request through palshelp at gmail dot com.

5 Conclusions

We have illustrated the importance of both international standardization of LSM evaluation and the definition of a priori performance benchmarking. In particular, we showed that apparently excellent LSM performance may in fact be poor, and that without quantitative understanding of what should be expected of a LSM in a given experiment, qualitative comparisons may give very little insight. Finally, we introduced a community-based automated online evaluation tool, the Protocol for Analysis of Land Surface models (PALS), and showed how it addresses both of these issues.

GMDD
5, 549–570, 2012

Acknowledgements. PALS is supported by the Australian Terrestrial Ecosystem Research Network (TERN) and the University of New South Wales. It has been developed in cooperation with the Global Energy and Water Cycle Experiment's Global Land-Atmosphere System Study (GLASS) panel. This work used eddy covariance data acquired by the FLUXNET community and in particular by the following networks: AmeriFlux (US Department of Energy, Biological and Environmental Research, Terrestrial Carbon Program (DEFG0204ER63917 and DEFG0204ER63911)), AfriFlux, AsiaFlux, CarboAfrica, CarboEuropeIP, CarboItaly, CarboMont, ChinaFlux, FluxnetCanada (supported by CFCAS, NSERC, BIOCAP, Environment Canada, and NRCan), GreenGrass, KoFlux, LBA, NECC, OzFlux, TCOSSiberia, USCCC. We acknowledge the financial support to the eddy covariance data harmonization provided by CarboEuropeIP, FAOGTOSTCO, iLEAPS, Max Planck Institute for Biogeochemistry, National Science Foundation, University of Tuscia, Université Laval and Environment Canada and US 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, University of Virginia.

References

Abramowitz, G., Leuning, R., Clark, M., and Pitman, A. J.: Evaluating the performance of land surface models, J. Climate, 21, 5468–5481, 2008.

Blyth, E. M., Best, M., Cox, P., Essery, R., Boucher, O., Harding, R., Prentice, I. C., Vidale, P.-L., and Woodward, I.: JULES: a new community land surface model, IGBP newsletter, 6, 9–11, 2006.

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

Collins, W. D., Bitz, C. M., Blackmon, M. L., Bonan, G. B., Bretherton, C. S., Carton, J. A., Chang, P., Doney, S. C., Hack, J. J., Henderson, T. B., Kiehl, J. T., Large, W. G., McKenna, D. S., Santer, B. D., and Smith R. D.: The Community Climate System Model version 3 (CCSM3), J. Climate, 19, 2122–2143, 2006.

Ginzburg, L. R. and Jensen, C. X. J.: Rules of thumb for judging ecological theories, Trends Ecol. Evol., 19(3), 121–126, 2004.
Gordon, H. B., Rotstayn, L. D., McGregor, J., Dix, M. R., Kowalczyk, E. A., Farrell, S. P. O., Waterman, L. J., Hirst, A. C., Wilson, S. G., Collier, M. A., Watterson, I. G., and Elliott, T. I.: The CSIRO Mk3 climate system model. Technical Report 60, CSIRO Atmospheric Research, Aspendale, Melbourne, 2002.
Henderson-Sellers, A., McGuffie, K., and Pitman A. J.: The Project for Intercomparison of Land-surface Parameterization Schemes (PILPS): 1992 to 1995, Clim. Dynam., 12, 849–859, 1996.
Kidston, J., Brümmer, C., Black, T. A., Morgenstern, K., Nesic, Z., McCaughey, J. H., and Barr, A. G.: Energy balance closure using eddy covariance above two different land surfaces and implications for CO_2 flux measurements, Bound.-Lay. Meteorol., 136, 193–218, doi:10.1007/s10546-010-9507-y, 2010.
Krinner, G., Viovy, N., de Noblet-Ducoudre, N., Ogee, J., Polcher, J., Friedlingtein, 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.
Leuning, R., Cleugh, H. A., Zegelin, S. J., and Hughes, D.: Carbon and water fluxes over a temperate Eucalyptus forest and a tropical wet/dry savanna in Australia: measurements and comparison with MODIS remote sensing estimates, Agr. Forest Meteorol., 129, 151–173, 2005.
Levis, S., Bonan, G., Vertenstein, M., and Oleson, K.: The community land model's dynamic global vegetation model (CLM-DGVM): technical description and user's guide, NCAR Tech. Rep. TN-459+IA, 50 pp., 2004.
Manabe, S.: Climate and the ocean circulation: 1, the atmospheric circulation and the hydrology of the Earth's surface, Mon. Weather Rev., 97, 739–805, 1969.
Medlyn, B. E., Robinson, A. P., Clement, R., and McMurtrie, R. E.: On the validation of models of forest CO_2 exchange using eddy covariance data: some perils and pitfalls, Tree Physiol., 25, 839–857, 2005.
Oleson, K. W., Dai, Y., Bonan, G., Bosilovich, M., Dickinson, R., Dirmeyer, P., Hoffman, F., Houser, P., Levis, S., Niu, G.-Y., Thornton, P., Vertenstein, M., Yang, Z.-L., and Zeng, X.: Technical description of the community land model (CLM), NCAR Tech. Rep. TN-461+STR, 174 pp., 2004.

GMDD
5, 549–570, 2012

A benchmarking system for land surface models

G. Abramowitz

Oreskes, N., Shrader-Frechette, K., and Belitz, K.: Verification, validation, and confirmation of numerical models in the earth sciences, Science, 263, 641–646, doi:10.1126/science.263.5147.641, 1994.
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.
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., and Verma, S.: Energy balance closure at FLUXNET sites, Agr. Forest Meteorol., 113, 223–243, 2002.

GMDD
5, 549–570, 2012

Table 1. A polarised description of two approaches to research modelling. The right hand column details a model with a collection of development issues commonly ignored both by researchers and research funding sources that affect the reliability, usability and reproducibility of LSMs, while the left hand column illustrates an exemplary model. Problems resulting from these kinds of issues are less likely to propagate through successive model versions in an environment of standardised international LSM evaluation.

Model has technical documentation Technical documentation matches what is in the model code	Model has no technical documentation Technical documentation related to what was in the code 5 yr ago
Model is open source, community oriented and has hundreds of users	Model is proprietary and only used by a few people in one organisation
All development of the model is contained in a version control system	Individuals maintain and manage multiple versions in home directories/desktop
Model has a clear user interface and user guide	Model has no user guide and no specific interface
Code is clearly commented, and logically structured	Code is not commented at all and structure is ad hoc
Variable names are consistent throughout the code and relate to their function	Variable names change in each subroutine call and are meaningless
Model changes meet prescribed performance/realism/functionality checks	Changes are accepted purely on the basis of personal preference.

GMDD
5, 549–570, 2012

GMDD

5, 549–570, 2012

A benchmarking system for land surface models

G. Abramowitz

Table 2. A hierarchy of a priori levels of benchmark performance for LSMs, with tiers defined by increasingly complex empirical models provided with more meteorological and site description variables.

Conservation of mass and energy (**weakest**)
Linear regression against shortwave radiation (**weak**)
....
Complex empirical model as a function of meteorology and vegetation and soil type (**strong**)
Within observational uncertainty (**strongest**)

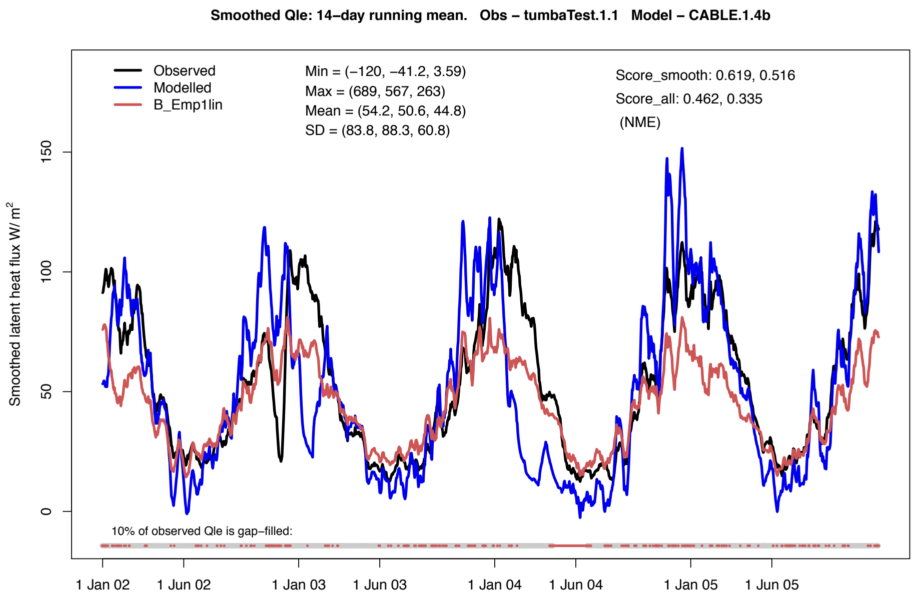

Fig. 1. A 14-day running mean latent heat flux time series at a single flux tower site. While model performance (blue) relative to observations (black) looks very good, many metrics on this time series show that an out-of-sample linear regression (red) of latent heat against short-wave radiation performs similarly.

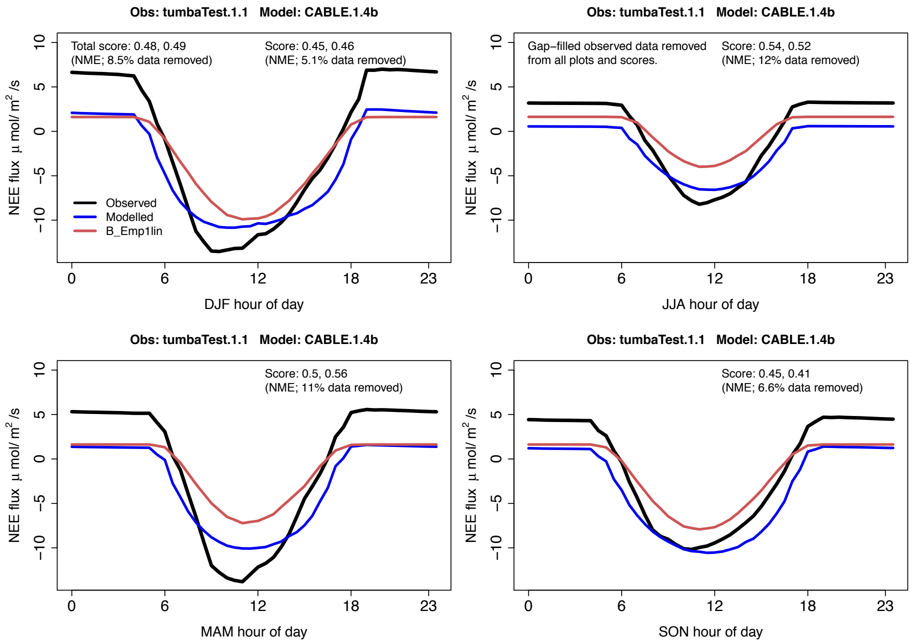

Fig. 2. Average diurnal cycle of net ecosystem exchange of CO_2 (NEE) at a single flux tower site, shown in a separate panel for each season. As in Fig. 1, an out of sample linear regression of NEE against downward shortwave radiation (shown in red) performs comparably to a LSM in this instance (blue). Normalised mean error of the average diurnal cycle is used as the scalar metric, shown separately for each panel and combined in the DJF panel.

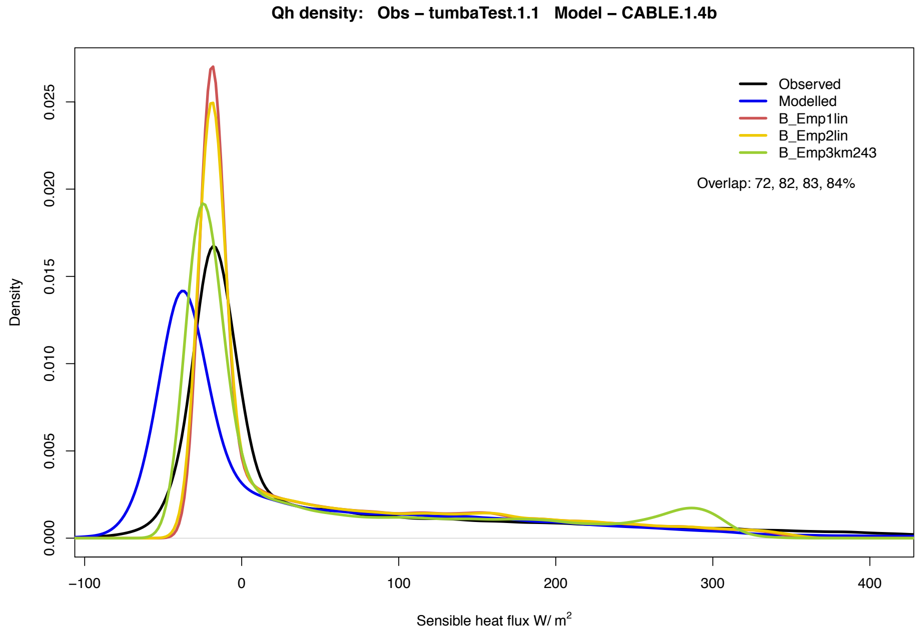

Fig. 3. Probability density functions (PDFs) of observed (black) and modelled (blue) sensible heat flux at a single flux tower site. Three additional PDFs representing a hierarchy of empirical models are also shown. The proportion of overlap of observed and modelled (or a benchmark PDF), expressed as a percentage, is used as the scalar metric for this analysis type. Metric values are listed in the same order as the legend.

GMDD
5, 549–570, 2012

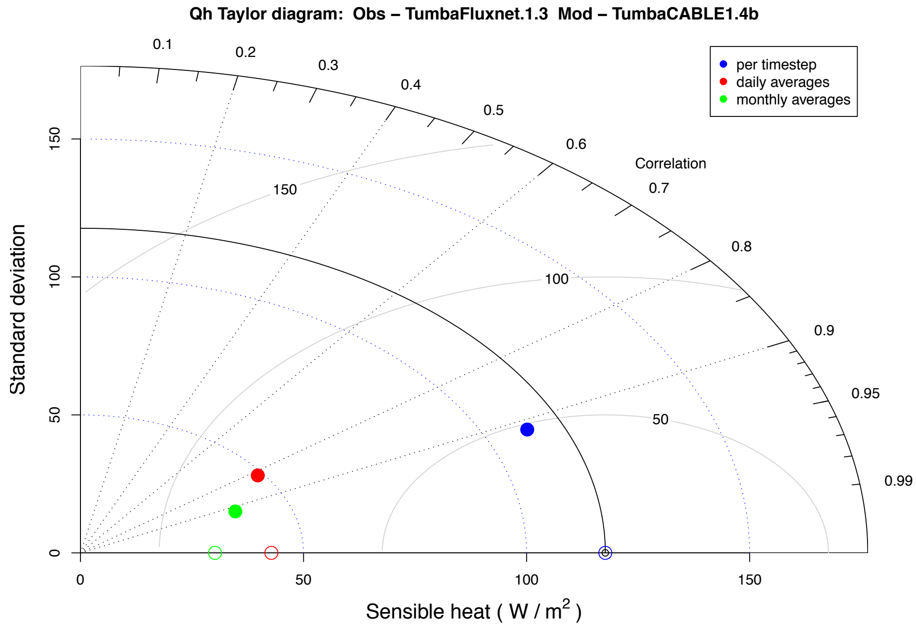

Fig. 4. Taylor diagram of sensible heat flux at a single flux tower site. Per-timestep standard deviation, correlation with observations and root mean square error are shown in blue, daily values in red and monthly values in green.

GMDD
5, 549–570, 2012

A benchmarking system for land surface models

G. Abramowitz