

Towards a benchmark for land surface models

Gab Abramowitz

Department of Physical Geography, Macquarie University, Sydney, New South Wales, Australia

Received 19 August 2005; revised 3 October 2005; accepted 11 October 2005; published 17 November 2005.

[1] This paper addresses the question of how well we should expect a land surface model to perform. A statistically-based artificial neural network is used as a de facto land surface model and its results used to benchmark the performance of a traditional physically-based land surface model. This provides us with a measure of land surface model performance relative to the information contained in the meteorological forcing about the surface fluxes. Further, it is a benchmark that is independent of the measure of model performance. The technique is used to benchmark three models at three observational sites, with results showing that for the most part, the models under-utilise the information available to them. This suggests that there are considerable opportunities for model improvement. **Citation:** Abramowitz, G. (2005), Towards a benchmark for land surface models, *Geophys. Res. Lett.*, 32, L22702, doi:10.1029/2005GL024419.

1. Introduction

[2] Land surface models (LSMs) represent the land surface component of global climate models (GCMs) which are used to make climate change projections. They partition available energy into sensible and latent heat, available water into runoff and evaporation and simulate carbon exchange, based on meteorological forcing, vegetation and soil characteristics. Techniques for assessing the performance of land surface models usually involve comparing simulations with satellite or ground based measurements. While there are many quantitative measures (e.g., daily RMSE or annual deviation from observed values) which help decide whether a particular model's performance is 'good', the threshold or benchmark for performance is usually another LSM, commonly an earlier version of the same LSM. We would like to move away from this somewhat circular approach to using a benchmark which reflects properties of the biophysical system. We would like to know the level of skill we might reasonably expect from a land surface model under specific conditions. Ultimately, we want to know how much information there is in the meteorological forcing about the surface fluxes at a particular site, and how well the LSM has utilised this information in simulating surface fluxes.

[3] To ascertain this, LSM performance is compared with that of a simple statistical model. The relationship between the meteorological forcing and surface fluxes captured by the statistical model is based solely on observational data. This de facto LSM is used as a benchmark to test the physically-based LSM. The statistical model operates on the model time step, and so effectively provides an objective, site-specific threshold for whichever measure of perfor-

mance a modeller may choose. While the ability of the statistical model clearly depends on the data on which it is built, we show that there is enough information in available flux measurements to provide a valuable critique of LSM performance. Details about methodology are given in section 2; results and discussion are presented in section 3; Conclusions are made in section 4.

2. Methodology

[4] We benchmark the performance of a physically based LSM against that of a statistically based Artificial Neural Network (ANN). We do this at three sites and consider three fluxes: latent heat (λE), sensible heat (H) and net ecosystem exchange of CO_2 (NEE). We use the first half of the data to establish a functional relationship between the meteorological forcing and surface fluxes ('train' the ANN). The second half is used to compare the ANN with the LSM.

[5] One may think of an ANN as a tool, which via an iterative process, establishes a functional relationship between one set of variables (inputs) and another (outputs). In this case, since we want to compare the ANN to a LSM, the relationship is between meteorological forcing and the three output fluxes we are considering. The ANN knows nothing about biophysics or soil moisture and temperature evolution; it has a completely instantaneous, statistically-based response to the meteorological forcing operating on a per-time step basis. It is provided with a training set from which it establishes the above relationship. A separate testing set is used to assess both the ANN and the LSM.

[6] The ANN we use is the Self Organising Linear Output map (SOLO) [see Hsu *et al.*, 2002; Abramowitz *et al.*, 2005]. SOLO 'learns' the relationship between its inputs and output by using a training data set as follows. Firstly, the input data are classified into groups or nodes using a Self Organising Feature Map (SOFM) [Kohonen, 1989], so that each node represents a distinct region of the input space. Then a linear regression is performed between the input data and their corresponding output data at each node. The result is a piecewise linear approximation of the training data. Unlike a feedforward ANN, this structure affords us some insight into the underlying processes [Hsu *et al.*, 2002], although we do not make use of this here.

[7] We focus on the performance of a single LSM, the CSIRO Biosphere Model (CBM), but include simulations from the ORCHIDEE LSM [Krinner *et al.*, 2005] and the Common Land Model (CLM) [see Oleson *et al.*, 2004; Levis *et al.*, 2004] to demonstrate that CBM is representative of leading LSMs. CBM uses a two-leaf canopy model consisting of a radiation model which calculates radiation absorbed by sunlit and shaded leaves, as well as a coupled model of stomatal conductance, photosynthesis and partitioning of absorbed net radiation into H and λE [Leuning *et*

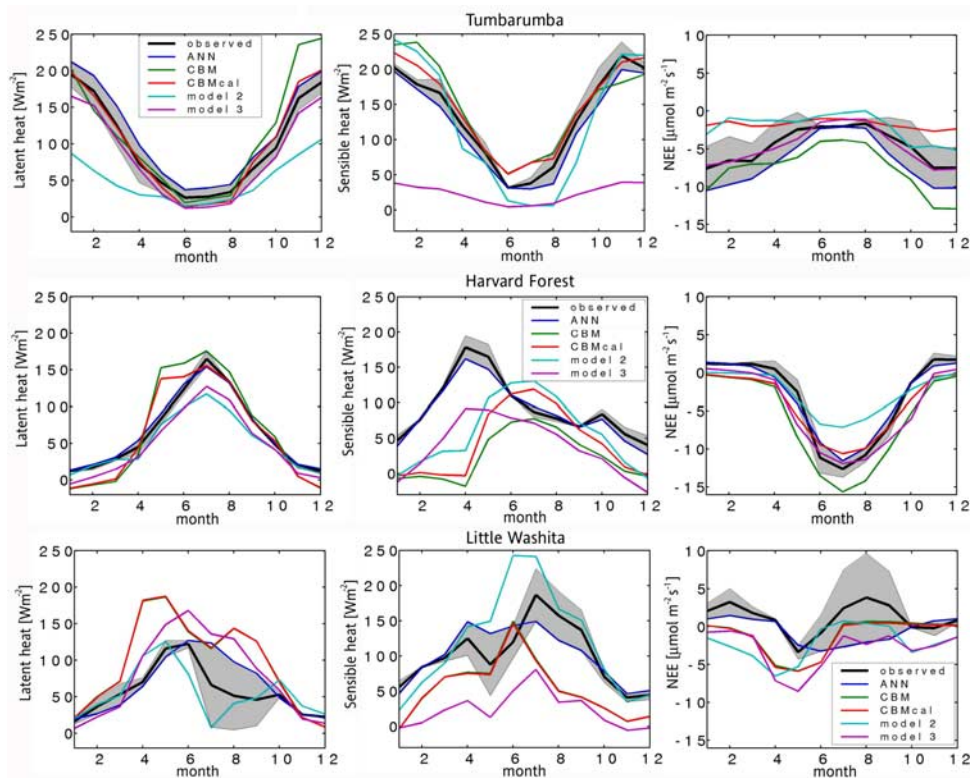


Figure 1. Average monthly fluxes for the three sites. The heavy black lines show observed values, the green lines CBM, and the blue lines the ANN. Red lines show the LSM ‘calibrated’ (see section 2) on the particular flux in shown in each plot. Cyan and magenta lines represent the performance of two other reference LSMs. The shaded region represents the absolute difference between observed and ANN values and represents the region of ‘good’ LSM performance.

al., 1998; Wang and Leuning, 1998]. The soil component calculates Richards’ equation moisture transport and heat conduction in six layers, includes soil freeze and thaw cycles, and simulates three snowpack layers. ORCHIDEE was used with plant and soil carbon activated and vegetation characteristics prescribed. For CLM, each site was represented as a 0.5 by 0.5 degree grid cell with all plant functional types and dynamic vegetation.

[8] We compare results at three sites, each using a one hour time step. Tumbarumba (35°39’S, 148°09’E) is a wet sclerophyll eucalypt forest in south eastern Australia. The canopy height is around 40 m and mean annual precipitation 1000 mm. Around three years of data (27575 time steps) were used from 2001–2004. Harvard Forest (42°32’N, 72°10’W) is a cool moist temperate deciduous forest site in Massachusetts which consists of a mixture of hardwoods and conifers. The canopy height is around 25 m and mean annual precipitation is 1050 mm. The eight years (70080 time steps) 1992–1999 were used here. Little Washita (34°58’N, 97°59’W) is a temperate grassland site in Oklahoma with mean annual precipitation 760 mm. Two full years of data (1997–1998, 17520 time steps) were used.

[9] At each of these three sites, the first half of the time series was used both to train the ANN and ‘spin-up’ the LSM. LSM spin-up involves running the model repeatedly on a data set until soil moisture and temperature stabilise. The results we discuss below show the LSM and ANN performance on the second half of the time series, with nighttime flux data excluded due to concerns with their reliability.

[10] For CBM and the two comparison LSMs, we use the default set of parameters for the grid cell containing the flux site, taken from the global fields prescribed by each of the model developers. This is intended to represent model behaviour as it would be in a GCM. For comparison, we also show CBM ‘calibrated’ (using the parameter optimisation technique described by *Vrugt et al.* [2003]) at each site. The entire data set was used for these calibrations; the 6 (of 36) calibrated vegetation and soil parameters and their allowed ranges were suggested by the model developers. We choose three measures of performance to demonstrate the benchmark, monthly and annual averages as well as standard deviation. We emphasise that the technique presented here provides a benchmark regardless of the measure of model performance chosen.

[11] Results for the SOLO map were relatively insensitive to its most critical parameter, the resolution of the SOFM, in the range of 100–1024 nodes. We report results from a resolution in the middle of this range, 400 nodes.

3. Results and Discussion

[12] Figure 1 shows the average monthly results for the three fluxes at all three sites during daytime hours. The heavy black lines show observed values, green lines CBM and the blue lines the purely statistical ANN. Red lines show the CBM ‘calibrated’ on the particular flux shown in each plot. The cyan and magenta lines show the performance the two reference LSMs (CLM and ORCHIDEE; intentionally not identified). The grey shaded region repre-

Table 1. Average Annual Fluxes and Per-Timestep Standard Deviation of λE , H and NEE for the Three Sites^a

	Latent, Wm ⁻²	σ	Sensible, Wm ⁻²	σ	NEE, $\mu\text{m m}^{-2}\text{s}^{-1}$	σ
<i>Tumbarumba</i>						
Obs	116.22	(113.9)	150.91	(157.3)	-5.29	(5.4)
CBM	132.50	(120.7)	164.66	(146.0)	-8.38	(4.7)
ANN	133.68	(106.9)	137.04	(129.5)	-7.66	(5.5)
<i>Harvard Forest</i>						
Obs	64.57	(84.5)	92.29	(122.1)	-3.21	(7.4)
CBM	66.22	(113.5)	24.54	(61.5)	-6.08	(7.0)
ANN	66.62	(77.7)	86.88	(106.8)	-3.12	(6.3)
<i>Washita</i>						
Obs	56.71	(52.4)	100.52	(100.5)	1.09	(3.3)
CBM	94.83	(121.7)	53.41	(76.0)	-1.20	(2.8)
ANN	65.28	(62.7)	100.73	(96.8)	-0.53	(3.1)

^aFigures are shown for observed fluxes (Obs), CBM, and ANN. CBM used default parameters.

sents the benchmark region based on the ANN performance, calculated by adding (or subtracting) the absolute difference of the observed and ANN value to the observed value for each monthly average calculation. We define ‘good’ model performance as lying within this region. Table 1 shows the annual average fluxes as well as the per-timestep standard deviation of each flux throughout the (daytime) testing period. Default parameters for CBM were used for this table.

[13] At Tumbarumba, the average monthly λE performance of CBM (top left in Figure 1) is largely within the grey shaded region, especially when calibrated against λE alone (red line). This suggests that the model performs well at this site, for this flux, at this timescale. It has utilised the information in the meteorological forcing as well as the statistical model. A similar appraisal applies for annual λE and standard deviation at Tumbarumba (Table 1).

[14] Tumbarumba H is similar except for a late summer and mid winter bias. This resulted in a good annual average and standard deviation result for CBM. Sensitivity of NEE flux to parameters here (top right in Figure 1) has probably biased CBM’s result, with reasonably poor performance in all three measures.

[15] At Harvard Forest, aside from CBM’s annual λE flux, it’s performance is consistently poor relative to the ANN in all fluxes and measures. This is true even when the model parameters are calibrated to best match observed fluxes (red line). It is tempting, given the similar performance of the two reference LSMs (cyan and magenta lines) to suggest Harvard Forest is simply a “difficult site” to simulate. However, the ANN demonstrates that clear, discernible relationships exist between the meteorological forcing and the fluxes. If the statistically based ANN is significantly better than the LSM, then the LSM’s ‘structure’ or ‘parametrisations’ are demonstrated to be inadequate in this case.

[16] At Little Washita results are similar. While there are a few months of the year when CBM’s and NEE performance compares well with that of the ANN, for the most part it lies well outside the shaded region. Annual averages and standard deviation results are also poor. Note that the two reference LSMs are similar in terms of performance.

[17] An obvious and legitimate criticism of this approach is its sensitivity to the quality of observed data. While noise in observed data should not be cause for concern (given the regression-based structure of the SOLO ANN), any systematic error in observations will weaken the legitimacy of the benchmark. Results will also be sensitive to the length (i.e., climatological coverage) of the ANN training set. If we have few data, we risk training on an ‘unrepresentative’ period and the benchmark not being strict enough. This is likely the case with the Tumbarumba and Washita sites, where we trained the ANN with only 1.5 and 1 year of data respectively. The shaded region is clearly wider at these two sites than at Harvard Forest. As we increase the number of training data, the closer the ANN will come to producing ‘average’ (instantaneous) flux responses for a site, and the more a LSM will need to utilise it’s internal states (soil moisture and temperature) to out-perform the ANN.

[18] We might also make the benchmark stricter. The ANN had no time dependence in its inputs (e.g., we could have used the previous time step’s value of a variable as an input). There may also be better statistical models for this purpose.

[19] How fair is this benchmark? At any particular site, the ANN has an advantage over the LSM because it is trained at that site alone, while the LSM is built to perform globally. We have not shown that a single ANN could serve as a global benchmark, but the transparency of the SOLO ANN and its robustness in terms of climate simulation [Pitman and Abramowitz, 2005] mean it may be possible with appropriate flux observations. Also, the model is restricted by mass and energy conservation; the ANN and observations (particularly Fluxnet - Harvard and Washita), are not. The ANN may be fitting bad data. Conversely, the LSM has the advantage of internal state variables which provide a time dependence lacking in the ANN. How we weigh these considerations against each other to decide whether the benchmark is fair is not immediately clear, and may be model and site dependent. We have however come closer to defining objectively ‘good’ LSM performance, based on the biophysical characteristics of a particular site.

4. Conclusions

[20] The use of a statistically-based model as a benchmark for a physically-based LSM guides us in deciding how much model deviation from observational data is acceptable. The ANN presented here recognised the relationship between meteorological forcing and surface fluxes better than the LSM with some measures and worse with others. LSM performance was accordingly described as “poor” and “good” in these measures respectively. Since the ANN provides a time series similar to that of the LSM, this benchmark technique does not restrict the choice of measure of LSM performance. The approach is likely to be sensitive to systematic errors in observed data, but could also be improved with time-dependent inputs to the ANN and perhaps a more appropriate statistical model in place of the ANN used here.

[21] Despite these caveats and possible improvements, we present a test for land surface modellers. Rather than ‘is your LSM better than an earlier version’ or ‘does it capture a particular aspect of the observations’, we ask ‘does it

outperform a simple statistical model with no time dependency?'. Our conclusion, using three sites, is that CBM, ORCHIDEE and CLM in most cases do not. This suggests that there is enough information in model input data for the models to do a better job; there is room for improvement. The results also underline the importance of the collection of high quality observational datasets of the type used here.

[22] **Acknowledgment.** Thanks to Kuo-lin Hsu for the SOLO code; CSIRO Atmospheric Research for the Tumberumba data; Fluxnet and Steve Wofsy for the Little Washita and Harvard Forest data; Andy Pitman and Ray Leuning for advice and review.

References

- Abramowitz, G., H. Gupta, A. J. Pitman, Y. Wang, R. Leuning, and H. Cleugh (2005), Neural Error Regression Diagnosis (NERD): A tool for model bias identification and prognostic data assimilation, *J. Hydro-meteorol.*, in press.
- Hsu, K.-l., H. V. Gupta, X. Gao, S. Sorooshian, and B. Imam (2002), Self-Organizing Linear Output map (SOLO): An artificial neural network suitable for hydrologic modeling and analysis, *Water Resour. Res.*, 38(12), 1302, doi:10.1029/2001WR000795.
- Kohonen, T. (1989), *Self-Organization and Associative Memory*, Springer, New York.
- Krinner, G., N. Viovy, N. de Noblet-Ducoudré, J. Ogée, J. Polcher, P. Friedlingstein, P. Ciais, S. Sitch, and I. C. Prentice (2005), A dynamic global vegetation model for studies of the coupled atmosphere-biosphere system, *Global Biogeochem. Cycles*, 19, GB1015, doi:10.1029/2003GB002199.
- Leuning, R., F. X. Dunin, and Y. P. Wang (1998), A two-leaf model for canopy conductance, photosynthesis and partitioning of available energy. II. Comparison with measurements, *Agric. Forest Meteorol.*, 91, 113–125.
- Levis, S., G. Bonan, M. Vertenstein, and K. Oleson (2004), The Community Land Model's Dynamic Global Vegetation Model (CLM-DGVM): Technical description and user's guide, *Tech. Rep. TN-459+IA*, Natl. Cent. for Atmos. Res., Boulder, Colo.
- Oleson, K., et al. (2004), Technical description of the Community Land Model (CLM), *Tech. Rep. TN-461+STR*, Natl. Cent. for Atmos. Res., Boulder, Colo.
- Pitman, A. J., and G. Abramowitz (2005), What are the limits to statistical error correction in land surface schemes when projecting the future?, *Geophys. Res. Lett.*, 32, L14403, doi:10.1029/2005GL023158.
- Vrugt, J., H. V. Gupta, L. A. Bastidas, W. Bouten, and S. Sorooshian (2003), Effective and efficient algorithm for multi-objective optimization of hydrologic models, *Water Resour. Res.*, 39(8), 1214, doi:10.1029/2002WR001746.
- Wang, Y. P., and R. Leuning (1998), A two-leaf model for canopy conductance, photosynthesis and partitioning of available energy. I. Model description, *Agric. Forest Meteorol.*, 91, 89–111.

G. Abramowitz, Department of Physical Geography, Macquarie University, Sydney, NSW 2109, Australia. (gabramow@els.mq.edu.au)