Evaluating the Performance of Land Surface Models

GAB ABRAMOWITZ
Climate Change Research Centre, University of New South Wales, Sydney, New South Wales, Australia

RAY LEUNING
CSIRO Marine and Atmospheric Research, Canberra, Australia

MARTYN CLARK
National Institute for Water and Atmospheric Research, Christchurch, Australia

ANDY PITMAN
Climate Change Research Centre, University of New South Wales, Sydney, New South Wales, Australia

(Manuscript received 17 December 2007, in final form 16 April 2008)

ABSTRACT

This paper presents a set of analytical tools to evaluate the performance of three land surface models (LSMs) that are used in global climate models (GCMs). Predictions of the fluxes of sensible heat, latent heat, and net CO₂ exchange obtained using process-based LSMs are benchmarked against two statistical models that only use incoming solar radiation, air temperature, and specific humidity as inputs to predict the fluxes. Both are then compared to measured fluxes at several flux stations located on three continents. Parameter sets used for the LSMs include default values used in GCMs for the plant functional type and soil type surrounding each flux station, locally calibrated values, and ensemble sets encompassing combinations of parameters within their respective uncertainty ranges. Performance of the LSMs is found to be generally inferior to that of the statistical models across a wide variety of performance metrics, suggesting that the LSMs underutilize the meteorological information used in their inputs and that model complexity may be hindering accurate prediction. The authors show that model evaluation is purpose specific; good performance in one metric does not guarantee good performance in others. Self-organizing maps are used to divide meteorological "forcing space" into distinct regions as a mechanism to identify the conditions under which model bias is greatest. These new techniques will aid modelers to identify the areas of model structure responsible for poor performance.

1. Introduction

Land surface models (LSMs) are the components of global climate models (GCMs) that simulate land surface processes, such as the absorption and partitioning of radiation, moisture, and carbon. They are typically provided with meteorological conditions as inputs (from a boundary layer atmospheric model) and produce outputs that include latent and sensible heat fluxes, CO₂ fluxes, reflected solar and emitted long-wave radiation, surface runoff, and deep soil drainage. They have internal state variables such as soil moisture and temperature, vegetation and soil carbon pools, and snow and ice volume and density. To describe the conditions of a simulation, LSMs have time-invariant (for the period of simulation) parameters, which describe soil and vegetation properties. For the purposes of climate simulation on decadal to century time scales, the most important outputs arguably are the turbulent fluxes, particularly latent heat and net ecosystem exchange of CO₂ (NEE). Moisture and heat fluxes are clearly essential drivers of precipitation (Koster et al. 2004) and longer time scales, and NEE plays a fundamental role in determining global CO₂ concentrations. Therefore, while we agree that any comprehensive
LSM evaluation must include all LSM outputs, for the purposes of demonstrating the techniques we present here, we focus on three turbulent fluxes: latent heat ($Q_l$), sensible heat ($Q_h$), and NEE, since these are instrumental in the surface energy, water, and carbon balances.

Deardorff (1978) presented one of the first LSM evaluations using fluxes estimated from a few meteorological observations during two summer days for a wheat crop in the United Kingdom. While subsequent flux evaluations have increased in statistical rigor to some degree (e.g., Sellers and Dorman 1987), sample size remained an issue. It was not until the beginnings of projects such as FLUXNET (Baldocchi et al. 2001) and the Project for the Intercomparison of Land-Surface Parameterization Schemes (PILPS) (Henderson-Sellers et al. 1996; Chen et al. 1997) that enough high temporal resolution data were available to characterize LSM biases statistically. While Chen et al. (1997) were the first to define performance that was “good enough” (using the model ensemble mean as a benchmark), no justification was offered as to why this was a satisfactory level of performance. Indeed, even more recent evaluations [e.g., the Community Land Model (CLM) (Dai et al. 2003) and Organizing Carbon and Hydrology in Dynamic Ecosystems (ORCHIDEE) (information available online at http://www.ipsl.jussieu.fr/~ssipsl/) have essentially not advanced in sophistication since Sellers and Dorman’s (1987) assertion that their LSM was “in general agreement with observations.”

In this paper we define a minimal benchmark level of LSM performance by comparing LSM predictions to those of two statistical models that establish empirical relationships between four meteorological variables and the three fluxes mentioned above, using data from several eddy-covariance flux towers. The statistical models have no mechanism to distinguish between vegetation and soil differences, and no internal state variables that allow quantities to evolve with time—they simply provide an instantaneous flux estimate based on a set of meteorological variable values. We show these comparisons in a wide range of performance measures and determine the extent to which different performance measures may be used as proxies for each other, if at all. We try to illustrate where LSM weaknesses may lie rather than presenting a quantitative evaluation framework for total model uncertainty. We introduce tools that help isolate the sources of bias in LSM model predictions by exploring the notion of conditional bias, whereby model performance is worse in some regions of bioclimatic and parameter space than in others. It is hoped that the approach presented will contribute to a formal LSM evaluation framework.

2. Tools and methodology

In this section we first describe the three LSMs and how they were set up for the simulations described here. Next, we detail the eddy covariance flux data used to drive and evaluate the model simulations. Last, we detail the two statistically based benchmarking techniques.

We compare three LSMs designed for use in climate models: the Community Atmosphere Biosphere Land Exchange (CABLE) (Kowalczyk et al. 2006); CLM (Levis et al. 2004; Oleson et al. 2004); and ORCHIDEE (Krinner et al. 2005). For CABLE, we examine a range of techniques to select parameter sets. First, default parameter sets are chosen by identifying one of 13 vegetation types (from Potter et al. 1993) and one of nine soil types (from Zobler 1988) appropriate for each site. This is essentially the approach used for GCM simulations, where a global $2^\circ \times 2^\circ$ grid of these vegetation and soil types is used. Next, a perturbed-parameter ensemble simulation is used to investigate output flux uncertainty resulting from parameter uncertainty. Eight of the CABLE 35 parameters were given several discrete possible values, and a simulation was produced for each possible combination of parameter values. These eight parameters were chosen in consultation with CABLE’s builders as uncertain and sensitive parameters: details of the parameter values used for the 20 736 simulations are shown in Table 1. Results are presented as probability density functions (PDFs). Last, for a “calibrated” parameter set, we chose the CABLE parameter set from the 20 736 ensemble simulation that minimized root-mean-square error (RMSE) when compared against eddy covariance measurements of each of the three turbulent fluxes under consideration. This appears to be a reasonable approach given the recognized difficulty in estimating more than eight LSM parameters using flux measurements (Wang et al. 2001). For CLM and ORCHIDEE only default parameter sets were examined, as though prescribed during a GCM simulation. All three used their own default (time varying) leaf area index (LAI) values based on latitude and longitude. All LSMs also had internal states initialized by conducting a spinup on each of the datasets. That is, they were repeatedly forced with multiple-integer-year meteorological data until variables such as soil moisture and temperature converged. While this does not guarantee true values of states, particularly if the meteorological data poorly represents the preceding period (e.g., a drought), it ensures that LSM heat fluxes are not
The three LSMs are compared to two statistically based models: both are empirically established relationships between four meteorological variables (a subset of the LSM inputs) and the three fluxes in question. Neither have any internal states (i.e., no knowledge of soil moisture or temperature) nor do they distinguish between vegetation and soil types; they both simply produce an instantaneous response to meteorological forcing. The first is a multiple linear regression between downward shortwave radiation, air temperature, specific humidity, wind speed, and NEE $Q_{Le}$ and $Q_h$. The second is the self-organizing linear output map (SOLO) artificial neural network (Hsu et al. 2002), using the same four inputs as predictors for these three fluxes. SOLO uses a self-organizing map (SOM) (Kohonen 1989) to divide all time-step data of these four inputs into 100 distinct groups, or “nodes.” For each of these nodes, a multiple linear regression is performed between the four meteorological variables and each of the output fluxes, using all time steps belonging to the node. Thus, SOLO produces a piecewise linear reconstruction of any functional relationship between the two sets of variables and is relatively insensitive to noise. This is the technique used by Abramowitz (2005). These two statistical models provide a time series at the same time step as the meteorological forcing and LSM output data, and so can give a benchmark level of performance in any performance metric that we may choose to examine. They represent a minimal estimate of the amount of information available to the LSms in their input data about their output fluxes.

We use these benchmark models in two modes, differing in the relationship between training and testing datasets. In the first mode, we train the benchmark models with data from five of our sites, and test them on the sixth, which we dub a “global benchmark.” For

### Table 1. Values of the eight parameters used to construct the perturbed parameter CABLE ensemble at Tumbarumba.

<table>
<thead>
<tr>
<th>Name</th>
<th>Units</th>
<th>Default</th>
<th>Ensemble values</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>LA1add</td>
<td>—</td>
<td>0.0</td>
<td>−1.0, 0.0, 1.0</td>
<td>Additive adjustment to LAI</td>
</tr>
<tr>
<td>HcAdd</td>
<td>m</td>
<td>0.0</td>
<td>−5.0, 0.0, 5.0</td>
<td>Additive adjustment to vegetation height</td>
</tr>
<tr>
<td>Bch</td>
<td>—</td>
<td>7.1</td>
<td>6.5, 7.1, 7.9, 8.7</td>
<td>Parameter $b$ in Campbell equation</td>
</tr>
<tr>
<td>Hyds</td>
<td>m s$^{-1} \times 10^{-6}$</td>
<td>4</td>
<td>2, 4, 6, 10</td>
<td>Hydraulic conductivity at saturation</td>
</tr>
<tr>
<td>Sfc</td>
<td>m$^3$ m$^{-3}$</td>
<td>0.301</td>
<td>0.26, 0.301, 0.35</td>
<td>Volume $H_{2O}$ at soil field capacity</td>
</tr>
<tr>
<td>Sucs</td>
<td>m</td>
<td>−0.59</td>
<td>−0.35, −0.45, −0.59, −0.62</td>
<td>Suction at saturation</td>
</tr>
<tr>
<td>Vemax</td>
<td>mol m$^{-2}$ s$^{-1} \times 10^{-5}$</td>
<td>8.5</td>
<td>1, 5, 8, 5, 12</td>
<td>Maximum RuBP carboxylation rate top leaf</td>
</tr>
<tr>
<td>Rp20</td>
<td>—</td>
<td>1.1342</td>
<td>0.1, 1.1342, 3.0</td>
<td>Plant respiration scaling factor</td>
</tr>
</tbody>
</table>

dominated by unequilibrated state variables. The appropriateness of such a spinup is perhaps least appropriate for carbon pool equilibrium (affecting NEE). A relatively short spinup dataset is unlikely to represent the longer term complexity of system disequilibrium in the period preceding the simulation; yet few other estimation techniques are available.

The observed meteorological and eddy-covariance flux data from the six sites used are outlined in Table 2. For all sites, downward longwave radiation was synthesized, as it was not measured. For Tumbarumba, New South Wales, Australia (Leuning et al. 2005), where we perform a more detailed analysis of LSM results, we used a cloudiness-corrected clear-sky approximation (Brutsaert 1975). The ratio of measured shortwave radiation to maximum anticipated shortwave radiation (based on latitude, longitude, and time of day) was used as a cloudiness index to increase downward longwave radiation. For the other five sites we used a clear-sky approximation based solely on temperature (Swinbank 1963). The Tumbarumba data also went through strict quality control of nighttime fluxes to account for underestimation of eddy flux due to advection (van Gorsel et al. 2008). Issues of bias and energy balance closure in eddy flux data are discussed after the experimental results.

The three LSMs are compared to two statistically based models: both are empirically established relationships between four meteorological variables (a subset of the LSM inputs) and the three fluxes in question. Neither have any internal states (i.e., no knowledge of soil moisture or temperature) nor do they distinguish between vegetation and soil types; they both simply produce an instantaneous response to meteorological forcing. The first is a multiple linear regression between downward shortwave radiation, air temperature, specific humidity, wind speed, and NEE $Q_{Le}$ and $Q_h$. The second is the self-organizing linear output map (SOLO) artificial neural network (Hsu et al. 2002), using the same four inputs as predictors for these three fluxes. SOLO uses a self-organizing map (SOM) (Kohonen 1989) to divide all time-step data of these four inputs into 100 distinct groups, or “nodes.” For each of these nodes, a multiple linear regression is performed between the four meteorological variables and each of the output fluxes, using all time steps belonging to the node. Thus, SOLO produces a piecewise linear reconstruction of any functional relationship between the two sets of variables and is relatively insensitive to noise. This is the technique used by Abramowitz (2005). These two statistical models provide a time series at the same time step as the meteorological forcing and LSM output data, and so can give a benchmark level of performance in any performance metric that we may choose to examine. They represent a minimal estimate of the amount of information available to the LSMs in their input data about their output fluxes.

We use these benchmark models in two modes, differing in the relationship between training and testing datasets. In the first mode, we train the benchmark models with data from five of our sites, and test them on the sixth, which we dub a “global benchmark.” For
example, the global benchmark models at Tumbarumba would be trained with data from Tharandt, Germany; Bondville, Illinois; Metolius, Oregon; Little Washita, Oklahoma; and Weiden Brunnen, Germany. We argue that this is a reasonable benchmark for LSMs using default parameter sets: Neither the LSMs nor the benchmark models have been allowed to be tuned with any local data. In the second mode, we allow both the LSMs and benchmark models to utilize locally measured flux data. As described above we allow calibration of the LSM parameters, and train the benchmark models on the same data on which it is tested.

3. Results and discussion

This section is divided into three parts. The first examines the performance of all of the models at all six sites using default parameter sets that do not utilize any site-specific calibration data. The second part focuses on CABLE at Tumbarumba. The performance of CABLE using both default and calibrated parameter sets is compared to the two benchmark models and examines the relationship between calibrated and default parameter set simulations. The third part investigates notions of LSM bias, and suggests techniques that may help us to identify weaker sections of a LSM code.

a. Part 1—All models and sites without site-specific information

We begin by examining a range of traditional performance measures for our five models. Table 3 shows the mean flux, RMSE, gradient, and intercept of the model versus observed linear regression, and the squared correlation coefficient for NEE, $Q_{le}$, and $Q_h$ at all six sites. The best performing model for each flux in each measure is in bold. The first point that we emphasize is the good performance in one of these measures is no guarantee of good performance in any other. There is not a single case of one model performing better than the others in all five measures, even for a single flux at a single site. This clearly illustrates the difficulty and complexity of the model evaluation process, and the need to reemphasize that evaluation is necessarily specific to the modeler’s objectives (Medlyn et al. 2005). For the purposes of climate change research, while some potential feedback mechanisms have been identified, as a community we are unsure of both which measures of performance are important and how accurately the models need to be. This leaves us with the somewhat unfocused task of minimizing error in all measures for which we have information from observed data. In Table 3 we simply declare the “best” model as that which has the most best performances in individual measures, and set it in bold for each of the six sites. A trivial sum of these shows that CABLE is best in 12 instances, ORCHIDEE in 17, and CLM in 15.

The next point noted from Table 3 is that the two statistically based models [multiple linear regression (MLR) and artificial neural network (ANN)] generally outperform the three physically based LSMs (MLR is best in 25 cases and ANN in 21 cases). In most cases, if we were to remove the best-performing empirical model, the next best-performing model would be the other empirical model: Neither has been trained with data from the site at which they are tested. At the very least this tells us that there is a high degree of consistency in the eddy-covariance flux data across different sites, continents, and vegetation types as measured by independent research teams. There is a chance that these data are consistently biased, especially given known issues with energy balance closure (Wilson et al. 2002); yet this does not appear to be the case. Figure 1 shows the average seasonal cycle (top three rows) and average diurnal cycle (bottom three rows) of observed fluxes, the three physically based LSMs, and the ANN benchmark for four of the six sites. If it is true that the eddy-covariance data were uniformly biased, we might expect that the observed data curve (black) and ANN prediction (yellow) would be consistently separated from the three physically based LSMs. This is clearly not the case; for any one of the measures in Fig. 1 or Table 3, the range of LSM behavior provides an envelope for the observations and all of the models across the sites considered. There is also recent evidence that the lack of vegetation heat storage accountability (usually not measured) plays a significant role in the energy budget closure issues of eddy-covariance data (Haverd et al. 2007).

Figure 1 also gives us useful information about the LSMs. ORCHIDEE, for example, appears to consistently underpredict $Q_h$, evident both on seasonal and diurnal plots. Similarly, CLM consistently underpredicts $Q_{le}$. None of the models managed to reproduce the relatively extreme diurnal cycle of NEE at the forested Tumbarumba and Tharandt sites, while most overpredicted the diurnal range of NEE behavior at the Bondville crop site.

Figure 2 shows flux values across all sites as a single PDF for all five models. Notable are the CABLE and CLM very high density of near-zero NEE predictions, with CLM having an apparent bias toward carbon uptake overall. The ORCHIDEE distribution appears considerably better, albeit with a slight respiration-based bias. The latent heat panel in Fig. 2 shows the negative latent heat bias of CLM, with a slightly smaller
negative bias for CABLE. The sensible heat panel shows markedly different behavior between the three LSMs. While ORCHIDEE, CLM, and the statistically based models show a clear peak of slightly negative values, the CABLE response is relatively flat with a clear negative bias. ORCHIDEE also shows a considerable negative bias, consistent with the results shown in Fig. 1.

b. Part II—CABLE at Tumbarumba with and without site-specific information

In this section, we use a variety of techniques to examine the advantage of allowing models to be tuned with data from the site at which they are being tested. We wish to ascertain the size and nature of the performance improvement due to the provision of testing site data, for both the LSM and statistically based ANN benchmark, as well as how easy it is to get the improvement from the LSM once these data are available. Owing to the number of measures of performance considered, we limit the study to a single LSM—CABLE—at a single site—Tumbarumba.

Figure 3 shows seasonal (left-hand-side column) and diurnal (right-hand-side column) cycles of NEE, $Q_v$, and $Q_H$. Results are shown for the two modes of the
Fig. 1. Average (top) seasonal (monthly averages) and (bottom) diurnal (hourly averages) cycle of net ecosystem exchange (NEE) of CO$_2$, latent heat ($Q_{le}$), and sensible heat ($Q_{h}$) for four flux tower sites (columns). Results are shown for observations (obs), the three LSMs, and the ANN benchmark.
ANN benchmark model, as described in the methodology section above: "global" (trained at the five sites other than Tumbarumba) shown in green and "local" (trained with Tumbarumba data only) shown in blue. Observed values are shown by the black line. The default simulation by CABLE that was used in section 3a is shown in pink. The parameter set from the ensemble runs that produced the lowest RMSE for each flux is shown as "CABLE calibrated" (red). The distribution of the CABLE ensemble is represented using two techniques: For the seasonal cycle, the distribution of CABLE’s behavior is represented by violin plots. That is, the monthly average values from each combination of parameters in the ensemble for a given month are used to construct a PDF, which is then rotated by 90° and reflected about a standard box plot. The gray region represents the range and density of values, the white dot the median value, and the heavy black vertical line the 25th–75th percentiles. For the average diurnal cycle plots, the CABLE ensemble density is shown using grayscale shading for each hourly average value, with heavier shading indicating higher densities.

One of the most surprising results from Fig. 3 is that the calibrated-CABLE run appears in some cases not to perform as well as the default run, despite being optimized for each individual flux (e.g., diurnal latent heat flux curve). The reason for this apparent contradiction that the choice of cost function for calibration (per time step RMSE) does not guarantee performance in other cost functions (i.e., those shown in Fig. 3). This independence of cost function is again clearly demonstrated in Fig. 4, which shows RMSE, the linear regression gradient, and $R^2$ values of CABLE predictions versus observations for a range of averaging window sizes. That is, “7” on the abscissa represents statistics for weekly averages, and “30” represents statistics for monthly averages. Shading shows 95% confidence intervals for the gradient estimates. The default CABLE simulation is shown in black dots, calibrated-CABLE in solid black, the global benchmark in gray dots, and the local benchmark in solid gray. It is clear in all three fluxes that the CABLE calibration (from dotted to solid black lines, based on RMSE) reduces RMSE. Note however, for example, that the slope for NEE is considerably less than unity (desirable) for the calibrated run across all of these time scales, whereas the slope approaches unity for both $Q_e$ and $Q_h$. This reinforces the results in Table 3 that good performance in one measure does not necessarily translate to others.

Another interesting feature in Fig. 3 is the very wide range of NEE behavior that CABLE displays within given parameter uncertainty ranges. Despite this, and even though the uncertainty range of the respiration scaling parameter was broad enough to encompass 11 of the 13 default vegetation types prescribed for CABLE, it is clear that at Tumbarumba CABLE underestimates nighttime respiration [especially noting the recent work on nighttime NEE observations by van Gorsel et al. (2008)]. A similar result is evident in both latent and sensible heat, although to a lesser degree. During months 5 to 9 (winter), despite the considerable allowed parameter ranges, all of the ensemble runs underestimate observed monthly average values, possibly caused by an overestimation of albedo and hence an underestimation of net absorbed radiation (Fig. 5).

While the inclusion of more parameters in the ensemble may have improved model performance, these results highlight a serious problem. Even for a site for which we have a great deal of information (Leuning et al. 2005; van Gorsel et al. 2008), it is very difficult to choose “good” (i.e., in terms of performance) parameter values a priori. Wang et al. (2001) demonstrated with a very similar LSM that only three to four parameters are clearly constrained using flux data. Taken together with the results shown here, this suggests that, even when we have good flux evaluation data, getting the best from a LSM can be a matter of guesswork. This lack of parameter certainty is less of a problem for
FIG. 3. Spread and density of the CABLE perturbed parameter ensemble at Tumbarumba: (left) Average seasonal cycle density is shown using violin plots and (right) average diurnal cycle density is shown using shading. For reference, CABLE default parameter runs and those using the RMSE-calibrated parameter set are also shown.
localized model applications, especially at well studied and measured sites. In the hydrological literature this lack of parameter identification gives rise to the notion of the “equifinality” of different “behavioral” (well performing) parameter sets (Beven and Freer 2001), where each set is argued to contribute valuable independent information toward the prediction problem. In the case of a LSM designed for global simulation, however,
Table 4. Seven performance measures on predictions of the three fluxes for the five models at Tumbarumba: mean and standard deviation, RMSE, gradient and intercept of model vs observed linear regression, square of correlation coefficient. Bold as in Table 3.

<table>
<thead>
<tr>
<th>Flux</th>
<th>CABLE calibrated</th>
<th>CABLE default</th>
<th>ORC default</th>
<th>CLM default</th>
<th>MLR global</th>
<th>Bench local</th>
<th>Bench global</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean (obs)</td>
<td>$-1.32 (-0.14)$</td>
<td>$-2.30 (-0.14)$</td>
<td>$-0.50 (-0.14)$</td>
<td>$-1.19 (-0.14)$</td>
<td>$-1.53 (-0.14)$</td>
<td>$-0.14 (-0.14)$</td>
<td>$-1.32 (-0.14)$</td>
</tr>
<tr>
<td>SD (obs)</td>
<td>5.56 (7.30)</td>
<td>5.53 (7.30)</td>
<td>4.80 (7.30)</td>
<td>2.96 (7.30)</td>
<td>4.81 (7.30)</td>
<td>6.44 (7.30)</td>
<td>4.77 (7.30)</td>
</tr>
<tr>
<td>RMSE</td>
<td>4.90</td>
<td>5.48</td>
<td>5.16</td>
<td>6.13</td>
<td>4.40</td>
<td><strong>3.44</strong></td>
<td>4.65</td>
</tr>
<tr>
<td>Grad</td>
<td>0.57</td>
<td>0.54</td>
<td>0.47</td>
<td>0.24</td>
<td>0.55</td>
<td><strong>0.78</strong></td>
<td>0.52</td>
</tr>
<tr>
<td>Int</td>
<td>$-1.24$</td>
<td>$-2.22$</td>
<td>$-0.43$</td>
<td>$-1.15$</td>
<td>$-1.45$</td>
<td><strong>-0.03</strong></td>
<td>$-1.25$</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.57</td>
<td>0.52</td>
<td>0.50</td>
<td>0.35</td>
<td>0.70</td>
<td><strong>0.78</strong></td>
<td>0.64</td>
</tr>
</tbody>
</table>

Latent

<table>
<thead>
<tr>
<th>Flux</th>
<th>CABLE calibrated</th>
<th>CABLE default</th>
<th>ORC default</th>
<th>CLM default</th>
<th>MLR global</th>
<th>Bench local</th>
<th>Bench global</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean (obs)</td>
<td>47.1 (52.0)</td>
<td>49.1 (52.0)</td>
<td>45.78 (52.0)</td>
<td>26.1 (52.0)</td>
<td>45.2 (52.0)</td>
<td><strong>51.9 (52.0)</strong></td>
<td>42.3 (52.0)</td>
</tr>
<tr>
<td>SD (obs)</td>
<td>77.2 (87.7)</td>
<td>80.1 (87.7)</td>
<td><strong>82.6 (87.7)</strong></td>
<td>52.4 (87.7)</td>
<td>60.0 (87.7)</td>
<td>78.5 (87.7)</td>
<td>55.2 (87.7)</td>
</tr>
<tr>
<td>RMSE</td>
<td>47.8</td>
<td>55.9</td>
<td>58.1</td>
<td>72.9</td>
<td>47.6</td>
<td><strong>39.0</strong></td>
<td>49.9</td>
</tr>
<tr>
<td>Grad</td>
<td>0.74</td>
<td>0.71</td>
<td>0.73</td>
<td>0.38</td>
<td>0.59</td>
<td>0.80</td>
<td>0.54</td>
</tr>
<tr>
<td>Int</td>
<td>8.6</td>
<td>12.0</td>
<td>8.1</td>
<td>6.5</td>
<td>14.5</td>
<td>10.3</td>
<td>14.1</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.71</td>
<td>0.61</td>
<td>0.59</td>
<td>0.40</td>
<td>0.74</td>
<td><strong>0.80</strong></td>
<td>0.74</td>
</tr>
</tbody>
</table>

Sensible

<table>
<thead>
<tr>
<th>Flux</th>
<th>CABLE calibrated</th>
<th>CABLE default</th>
<th>ORC default</th>
<th>CLM default</th>
<th>MLR global</th>
<th>Bench local</th>
<th>Bench global</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean (obs)</td>
<td>46.2 (54.1)</td>
<td>37.8 (54.1)</td>
<td>14.3 (54.1)</td>
<td>55.9 (54.1)</td>
<td>42.7 (54.1)</td>
<td><strong>54.1 (54.1)</strong></td>
<td>40.5 (54.1)</td>
</tr>
<tr>
<td>SD (obs)</td>
<td>129.8 (131.1)</td>
<td>114.8 (131.1)</td>
<td>23.0 (131.1)</td>
<td>136.2 (131.1)</td>
<td>97.7 (131.1)</td>
<td>123.6 (131.1)</td>
<td>113.8 (131.1)</td>
</tr>
<tr>
<td>RMSE</td>
<td>49.2</td>
<td>58.3</td>
<td>118.1</td>
<td>84.5</td>
<td>54.9</td>
<td><strong>43.7</strong></td>
<td>51.4</td>
</tr>
<tr>
<td>Grad</td>
<td><strong>0.92</strong></td>
<td>0.79</td>
<td>0.16</td>
<td>0.83</td>
<td>0.69</td>
<td>0.89</td>
<td>0.81</td>
</tr>
<tr>
<td>Int</td>
<td>$-3.6$</td>
<td>$-5.0$</td>
<td>6.0</td>
<td>10.9</td>
<td>5.2</td>
<td>6.0</td>
<td><strong>-3.0</strong></td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.87</td>
<td>0.82</td>
<td>0.78</td>
<td>0.64</td>
<td>0.87</td>
<td><strong>0.89</strong></td>
<td>0.86</td>
</tr>
</tbody>
</table>

where comparatively little spatially explicit information is available, having many ill-defined or poorly constrained parameters is not only computationally wasteful but is actually a serious hindrance to accurate prediction.

It should be clear from the two statistically based benchmark runs (green and blue) that LSMs should, indeed, be capable of more accurate predictions. Recall that both of these ANN models are simply empirically derived relationships between meteorological conditions and fluxes. The LSM, which incorporates knowledge of around 35 spatially varying vegetation and soil properties, as well as internal states (such as soil moisture and soil temperature), should have a clear competitive advantage. The locally calibrated benchmark ANN (blue) arguably gives a minimal estimate of how much of a competitive advantage local flux data should give. A similar trend is apparent in the aggregated measures of model performance shown in Table 4. While the empirically derived models are not restrained by energy conservation (as the LSMs are), this does not appear to be the reason for their success. At Tumbarumba, the values of $Q_{le} + Q_{h}$ for the empirical models lie within the LSM range of $Q_{le} + Q_{h}$ for 92% of the time steps, with 4% above and 4% below.

Even though this section has only examined the effect of calibration on one LSM, Fig. 1 and Table 3 remind us that this issue is common to other LSMs as well. In practical climate applications, using default parameter sets—the distinction between a poor or heavily biased model and one that is overparameterized is meaningless—we are unable to make LSMs perform as well as a multiple linear regression, and this is a problem.

c. Part III—Isolating LSM biases

Given the issues identified in the two previous subsections, how might we go about isolating problem areas in a given LSM? We discuss the nature of LSM bias and suggest a few possible ways forward. Any successful approach will require both analytical techniques and the process expertise of those who built the LSM. Whether we want to simply correct a biased parameterization or actually simplify a collection of processes so that fewer parameters need to be identified, the procedure will be similar: isolate the specific set of conditions under which the LSM produces biased output. We formalize the idea of conditional bias as follows.

Figure 6 shows nine panels of PDFs of NEE values at Tumbarumba, each similar to the first panel of Fig. 2. The data used to construct the PDFs in each panel are a discrete subset of the Tumbarumba time series. That is, any one time step from the Tumbarumba time series will belong to just one of these nine panels. Each panel (hereafter node) represents a collection of time steps (not necessarily consecutive) belonging to each
node are shown in Fig. 7. The three bars in each node of Fig. 7 represent the means of downward shortwave radiation, air temperature, and specific humidity, with distributions represented by violin plots, as described for Fig. 3. The time series was sorted into these nodes using a SOM, although there are clearly many ways one could choose to make these divisions.

The three nighttime nodes in Fig. 7 (those with the first column essentially zero) correspond to those in Fig. 6 with respiration-based distributions (note the different x axis scales). It is clear that CABLE, whether calibrated or not, does not predict enough respiration at night at Tumbarumba, reinforcing the results in Figs. 1 and 3. What is also clear from the global benchmark ANN simulation is that nighttime respiration is lower at the five other sites than at Tumbarumba. In the daytime nodes, the observed PDFs in Fig. 6 are broader than any of these models, including the locally trained benchmark. While this is likely due to the noisiness of the eddy covariance data (say, from advection or turbulence), we can only say for certain that it is not predictable from the meteorological inputs provided to the ANN. This is less evident in the aggregated distributions (across all sites) shown in Fig. 2.

By illustrating the notion of “conditional bias,” Figs. 6 and 7 can also provide us with information about why
CABLE is not performing as well as it might. We can see, for example, that calibration mainly affects warm, humid nighttime conditions. That is, for the three nighttime nodes in the SOM (top left, bottom left, and middle right, indicated by a zero shortwave column in Fig. 7), only the two associated with warmer, wetter conditions show divergence between the pink (CABLE default) and red (CABLE calibrated) lines. We can also see that, of the daytime nodes, those with less shortwave radiation and cooler temperatures (middle left, center, and bottom right) show CABLE with a bimodal response that is not shown in the observed data or the statistical models. This may be indicative of a division between the behavior of morning and evening time steps, possibly because of the sensitivity to soil temperature or soil moisture. This type of analysis can clearly give insight into weaker areas of a LSM, especially when large datasets are available and the number of SOM nodes can be large.

There are of course many other useful diagnostic tools one might use. Recent techniques that assess parameter sensitivity temporally (Wagener et al. 2003) and the overfitting tests of Ginzburg and Jensen (2004) are examples that should provide valuable insights.

4. Further discussion and conclusions

The above land surface model evaluations underscore the need for a wide variety of performance measures to make conclusive judgment about performance. For all three fluxes considered, at all six sites, no single model performed best in all of the measures considered. This clearly makes the extent to which model evaluation is problem specific—hence our avoidance of
the term “validation.” Overall, we have seen that the LSMs presented here have not captured the extent to which land surface fluxes are driven by meteorology. In most measures of performance they were consistently outperformed by simple statistical models without any mechanisms to distinguish between different vegetation or soil types, even when these statistical models were trained at sites other than those at which they were tested. We also have no reason to believe these LSMs are not representative of LSMs as a whole. We suggest that whether the LSMs could have done better with site-specific flux data and intensive calibration is not relevant, as these models are designed for global climate simulation. Even with the provision of testing site flux data and a simple eight parameter calibration for one of the LSMs, we were unable to choose parameters that allowed it to consistently outperform the statistically based models provided with the same data. While there are known issues with eddy-covariance flux data, we argue that they do not affect the findings. For most measures of performance shown, the LSMs provided a range of behavior that enveloped the empirically based models. We also suggested a technique that may help to find which areas of a LSM parameterization are responsible for its biases. By developing the notion of “conditional bias,” isolating the particular set of conditions that lead to biased LSM behavior, we provide a mechanism to direct efforts to improve LSMs.

Acknowledgments. We thank Eva van Gorsel for Tumbarumba data, FLUXNET for the provision of the data from other sites, Kuo-lin Hsu for the SOLO code, and Yingping Wang and Eva Kowalczyk for help with CABLE parameter selection. GA and AP were supported by a grant from the Australian Research Council.

REFERENCES


