Systematic Bias in Land Surface Models

GAB ABRAMOWITZ

University of New South Wales, Sydney, Australia, and CSIRO Marine and Atmospheric Research, Victoria, Australia

ANDY PITMAN

University of New South Wales, Sydney, Australia

HOSHIN GUPTA

Department of Hydrology and Water Resources, The University of Arizona, Tucson, Arizona

EVA KOWALCZYK AND YINGPING WANG

CSIRO Marine and Atmospheric Research, Aspendale, Victoria, Australia

(Manuscript received 15 August 2006, in final form 28 February 2007)

ABSTRACT

A neural network–based flux correction technique is applied to three land surface models. It is then used to show that the nature of systematic model error in simulations of latent heat, sensible heat, and the net ecosystem exchange of CO_2 is shared between different vegetation types and indeed different models. By manipulating the relationship between the dataset used to train the correction technique and that used to test it, it is shown that as much as 45% of per-time-step model root-mean-square error in these flux outputs is due to systematic problems in those model processes insensitive to changes in vegetation parameters. This is shown in the three land surface models using flux tower measurements from 13 sites spanning 2 vegetation types. These results suggest that efforts to improve the representation of fundamental processes in land surface models, rather than parameter optimization, are the key to the development of land surface model ability.

1. Introduction

A land surface model (LSM) is used in a climate model to represent the interaction between the atmosphere and land surface. It simulates radiation, water, heat, and carbon exchanges, with explicit representation of vegetation and soil types (see Pitman 2003). LSMs are commonly evaluated using observed values of three key model outputs: latent heat flux (Qle), sensible heat flux (Qh), and Net Ecosystem Exchange (NEE) of CO_2 from eddy covariance flux measurements (e.g., Sellers and Dorman 1987; Chen et al. 1997; Dai et al. 2003). While several eddy covariance–based studies have attempted to reduce model bias associated with inappropriate parameter values (e.g., Wang et al. 2001; Braswell et al. 2005; Drewry and Albertson 2006), few have attempted to examine bias resulting from the LSM itself (Dekker et al. 2001 is a notable exception). In the broader context of geophysical fluid dynamics, longer-term trends have been subtracted from the model output as a way of removing bias in reanalysis (e.g., Klinker and Sardeshmukh 1992; Saha 1992). Also, more recently, bias identification techniques have been developed for state-based data assimilation (e.g., Dee and Todling 2000; Keppenne et al. 2005). While these tools are useful for improving numerical weather predictions, they offer little benefit for long-term prognostic simulations or insight into how bias identification can aid model development.

Corresponding author address: Gab Abramowitz, University of New South Wales/CSIRO Marine and Atmospheric Research, Aspendale, VIC 3195, Australia.
E-mail: gabsun@gmail.com

DOI: 10.1175/JHM628.1

Abramowitz et al. (2006) attempted to identify bias in an "uncoupled" LSM (not coupled to a climate model) using eddy covariance data from two flux tower sites. They showed that flux predictions could be significantly improved in a variety of measures using an Artificial Neural Network (ANN) correction to Qle, Qh, and NEE. Ensembles from multiple criteria parameter estimation pareto sets (see Vrugt et al. 2003) were used to remove bias associated with inappropriate parameter values. In that paper, a Self-Organizing Feature Map (SOFM; Kohonen 1989) based ANN was used to correct model flux predictions at each time step, using meteorological variables (LSM inputs) and LSM flux predictions as inputs to the ANN. The ANN, by means of supervised training, established a functional relationship between these inputs and the magnitude of model error (simply the observed minus modeled value) in a particular flux. Using a dataset unseen by the ANN, predictions of LSM error made by the ANN were used to correct LSM output.

In this paper, this methodology is extended so that ANN training and testing sets incorporate several sites simultaneously. In the first instance, this allows us to examine the potential for applying the correction regionally or globally. More importantly, however, by using the correction technique as a tool to characterize and quantify LSM systematic error, we can understand the extent to which LSM systematic error is shared between simulations of different environments. Such an understanding clearly helps to define directions for LSM improvement. We also note the known ability of ANNs to simulate moisture and carbon fluxes regionally (e.g., Papale and Valentini 2003).

We investigate whether systematic LSM bias is shared in different environments by examining how dissimilar ANN training and testing sets need to be before the nature of LSM error learned by the ANN from the training set is no longer useful in predicting LSM error on the testing set. Tests for the "robustness" of the correction by changing the relationship between training and testing sets have been undertaken for the ANN correction trained and tested at a single site, both in terms of temporal robustness (Abramowitz et al. 2006) and in terms of a temperature-based climate change scenario (Pitman and Abramowitz 2005). This multiple-site examination, however, allows us to compare the size and nature of the LSM systematic error at a range of sites with different vegetation types. In particular, we show that the considerable systematic bias identified in simulations of temperate grassland sites is of a similar nature to that identified at coniferous forest sites. That is, an ANN trained to correct a LSM at one vegetation type competently corrects the same LSM at another.

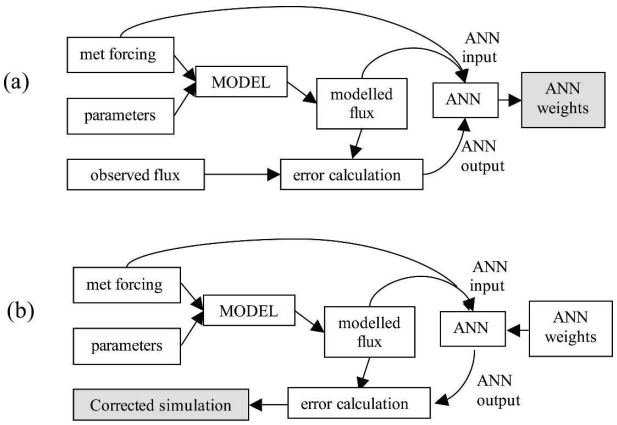

FIG. 1. The two phases of ANN operation. (a) The training phase seeks to find ANN weights that represent the relationship between ANN inputs and outputs and (b) the testing phase, when this is used to correct model output.

This was true, to some extent, of all three LSMs examined. This implies that the nature of the LSM error is shared across different LSM parameter sets representing the different vegetation types, so that those model processes responsible for the systematic error in output are largely insensitive to vegetation parameters. We therefore suggest that efforts to improve existing processes within LSMs, using the increasing wealth of in situ observational data, would be fruitful.

We begin by outlining the correction technique and the five configurations we use to train and test it. Results are then discussed for each configuration in turn. We then discuss the implications of the results in terms of directions for LSM improvement as well as possible protocols that may assist further improvement.

2. Methodology

We wish to examine the improvement in LSM performance afforded by an ANN-based statistical correction. The experimental design is essentially that of Abramowitz et al. (2006), which we now explain. The ANN-based correction is made independently to three LSM output fluxes, Qle, Qh and NEE, on a per-time-step basis. Figure 1 represents the two-stage correction process, with shaded boxes representing the goal of each stage. Ultimately, the ANN will make a prediction of the model's error in a particular flux (the ANN output), based on meteorological conditions and the model's output (the ANN inputs; Fig. 1b). This prediction of error will then be used to correct the LSM's simulation. By LSM "error," we simply mean the difference between the simulated and observed value. To do this, the ANN must first "learn" the relationship between these

inputs and this output. The first stage therefore involves providing the ANN with a time series of input–output training data from which it establishes this relationship. This information is stored in the ANN's internal weights (Fig. 1a).

For this work, Qle error for a particular model time step was simulated by the ANN as a function of downward shortwave, air temperature, and *modeled* Qle at that time step; the Qh error was simulated by the ANN as a function of downward shortwave, air temperature, and modeled Qh; the NEE error as a function of downward shortwave, air temperature, and modeled NEE. While a systematic sensitivity study to determine the optimal set of ANN inputs in each case would deliver the greatest correction, this single set of three inputs better demonstrates the broad, universal nature of the models' systematic bias and thus the universal applicability of the correction. We also note this choice of inputs is broadly consistent with van Wijk and Bouten (1999).

The ANN we use is the Self-Organizing Linear Output map (SOLO; Hsu et al. 2002). Readers familiar with ANNs might assume we use a feed-forward ANN. However, unlike a feed-forward ANN, SOLO uses a SOFM to classify input data into a two-dimensional matrix of groups, or "nodes." Here, each member of a group is a time step of downward shortwave, temperature, and the modeled flux. Time steps with similar characteristics (in terms of these three variables) will belong to neighboring groups in the SOFM. Once the classification of all time steps into groups is complete, a multiple linear regression is performed between all the members of a particular group (a collection of three-dimensional vectors) and their corresponding output values in the time series from which they came. The SOLO algorithm therefore establishes a piecewise linear relationship between the three input variables and the output, LSM flux error. If we use just one node in the SOFM, SOLO corrects the LSM with a single multiple linear regression between the three input variables mentioned above and the flux error. The regression-based structure of SOLO therefore avoids some of the pitfalls of feed-forward ANNs, such as problems with gradient descent convergence and overtraining (e.g., van Wijk and Bouten 1999). More details about SOLO can be found in Hsu et al. (2002) and Abramowitz et al. (2006). The only SOLO parameter we manually adjust in this work is the resolution (or number of groups) of the SOFM.

To avoid results being model specific we used three LSMs, chosen based on availability, documentation, and ease of modification of the input–output format. We also wished that the LSMs broadly reflected the range of approaches to land surface modeling used in climate modeling applications. The three LSMs we use are the Commonwealth Scientific and Industrial Research Organisation (CSIRO) Biosphere Model (CBM; Leuning et al. 1998; Wang and Leuning 1998), the Organising Carbon and Hydrology in Dynamic Ecosystems model (ORCHIDEE; Krinner et al. 2005), and the Community Land Model (CLM; Levis et al. 2004; Oleson et al. 2004), which includes dynamic vegetation. Throughout this paper, these models are executed offline (*not* coupled within a climate model).

To compare LSM bias in different environments, we must be satisfied that the above correction does indeed correct for systematic bias. If we represent a LSM functionally as

$$Y_t = M(I_t, \phi, \zeta_{t-1}), \tag{1}$$

where Y_t are the LSM outputs for a given time step t, I_t are the LSM inputs, ϕ are the (time invariant) parameters, and ζ_{t-1} are the previous time step's states, we can then express the LSM *error* for this time step as

$$E(\phi, I_t, O_t, \zeta_{t-1}, M) = Y_t - O_t, \tag{2}$$

where O_t are the observations of the LSM outputs made concurrently with the LSM inputs. This suggests four sources of LSM simulation error: parameter values, ϕ, that do not best represent the system being simulated; measurement error in the LSM inputs and outputs, I_t and O_t; misprescription of initial states, ζ_{t-1}; and flaws in the LSM's representation of physical processes, M (note that some authors may also refer to this as "model structure error" or "model parameterization error"). While we deal with each of these sources in turn, our argument differs from Abramowitz et al. (2006) only in the method parameter prescription, so we will focus on this process.

For all three LSMs, for each site simulated, we used the default set of parameters provided by the model developers. These were derived from global fields, as though each LSM were running inside a climate model, with each site represented by a relatively large grid cell. This clearly means that agreement between the LSMs and individual site measurements will not be as good as if we had optimized model parameters at each of these sites individually. However, Abramowitz (2005) and Abramowitz et al. (2006) showed that LSM bias was considerably larger than gains afforded by parameter optimization. Results below showing the transferability of the ANN correction from one vegetation type to another (with very different parameter values) suggest that the same is true in the cases presented here. We also note that this is how parameter values are com-

TABLE 1. The 13 flux observation sites.

	Vegetation type	Lat	Lon	Country	Annual rain	Yr/length
Aberfeldy	Coniferous	56°37′N	03°48′W	Scotland	1200	12 Mar 1997–31 Dec 1998
Bondville	Cropland	40°0′N	88°18′W	U.S.-IL	760	1 Jan 1997–31 Dec 1999
Bordeaux	Coniferous	44°42′N	00°46′W	France	950	12 Jul 1996–31 Dec 1998
Flakaliden	Coniferous	64°07′N	19°27′E	Sweden	590	8 Oct 1996–31 Dec 1998
Hyytiala	Coniferous	61°51′N	24°17′E	Finland	640	2 Apr 1996–31 Dec 2003
Little Washita	Grassland	34°58′N	97°59′W	U.S.-OK	830	14 May 1996–31 Dec 1998
Loobos	Coniferous	52°10′N	05°45′E	Netherlands	790	1 Jan 1997–31 Dec 2002
Metolius	Coniferous	44°30′N	121°37′W	U.S.-OR	710	1 Jan 1996–31 Dec 1997
Norunda	Coniferous	60°05′N	17°28′E	Sweden	530	1 Jan 1996–31 Dec 1998
Ponca City	Grassland	36°46′N	97°08′W	U.S.-OK	800	1 Jan 1997–31 Dec 1997
Shidler	Grassland	36°56′N	96°41′W	U.S.-OK	830	1 Jan 1997–31 Dec 1997
Tharandt	Coniferous	50°58′N	13°38′E	Germany	820	1 Jan 1996–31 Dec 2000
Weiden Brunnen	Coniferous	50°09′N	11°52′E	Germany	890	12 Jun 1996–31 Dec 1999

monly selected for these LSMs in global coupled simulations. Our only departure from this regime was choosing vegetation types for CBM and ORCHIDEE to best match the vegetation found at the flux tower sites. CLM incorporated dynamic vegetation while CBM and ORCHIDEE did not. We discuss CLM's representation of the two vegetation types below when discussing state initialization.

The 13 flux tower sites are shown in Table 1. These sites are from the Fluxnet network (available online at http://www.fluxnet.ornl.gov/fluxnet/index.cfm) with gap filling described by Falge et al. (2001). Where observational teams had flagged time steps that had been gap filled, we excluded these, where possible, from the analysis. The sites represent essentially two vegetation types within the LSMs, nine are coniferous forest and four are grassland or cropland. While we would have liked to explore a wider range of vegetation, these were the two vegetation types available from the Fluxnet site that had sufficiently long datasets, contained the range of gap-filled variables we required and were represented by several sites.

With a large range of sites from different observational teams, observational error (in I_t and O_t above) is inevitable. We wish to distinguish, however, between random and systematic error. With hundreds to many thousands of data points for each regression node in the SOFMs used in the ANN, truly random error in observations should not prevent the ANN from capturing systematic LSM bias. Systematic bias in observations, however, would hinder the ability of the correction technique. Given known issues with energy closure in eddy correlation data (Twine et al. 2000; Wilson et al. 2002), we cannot dismiss this as a possibility. We will return to this issue in the discussion, where we argue that the spread of LSM simulations suggests that LSM bias is significant relative to observational bias.

State initialization issues in long-term prognostic simulation are usually dealt with by "spinning up" the LSMs concerned. The LSM runs a simulation dataset repeatedly until model states, such as the soil moisture, temperature, and vegetation distribution equilibrate. All LSM simulations performed here involved a spinup period, some as long as 200 yr due to the CLM using dynamic vegetation. This is no guarantee, however, that for a given time step, the previous time step's state, ζ_{t-1}, is representative of the natural system. The internal feedback of errors resulting from LSM inadequacy, M, in time steps 1, . . . , $t-1$ mean that LSM states may "drift" from measured values, if they exist. For the short integration times of numerical weather forecasts, ζ_{t-1} is sensitive to ζ_1, which is the initial state value. The availability of state observations therefore mean that there are real rewards, through Kalman filter or variational approaches, for state-estimation data assimilation. For this work, however, since the length of integration means that this sensitivity does not exist, we consider ζ_{t-1} as functionally dependent on LSM's representation of processes, or structure, M. This is particularly important for CLM's representation of dynamic vegetation at these sites, since there is no mechanism that ensures that the vegetation at a flux tower site will match CLM's prediction. We note, however, that grassland sites were indeed dominated by grass plant functional types (see Levis et al. 2004; Table 1) with vegetation heights ranging from 0 to 5.5 m and coniferous forest sites similarly represented by tall forest at 3.5–10.8 m. This considerably looser definition of vegetation types for CLM is particularly important for the last two of the five training-testing cases we present below.

We examine the ability of the ANN to correct model output in five configurations. The differences between the five configurations of the ANN are essentially

changes in the relationship between the data used to train the ANN and the data used to test it:

- Case 1: Train a single ANN on the first half of the time series of data from *all* 13 sites and test on the second half. This is a test of the *temporal* transitivity of the ANN correction.
- Case 2: Train a single ANN on the *entire* time series from half of the coniferous sites and test it on the entire time series from the other half of the coniferous sites. Then use a second ANN to do the same for grassland sites. This is a test of *spatial* transitivity of the ANN correction within a vegetation type, for two vegetation types.
- Case 3: Train a single ANN on the entire time series from half of *all* the sites, both conifer and grass, and test it on the entire time series from the other half of the sites. This is a test of the spatial transitivity across both vegetation types.
- Case 4: Use the ANN from the first instance of case 2, trained to correct LSMs at coniferous sites, and test its ability to correct at grassland sites. Then the reverse, using an ANN trained on grassland to correct coniferous site simulations. This is a test of *vegetation-type* transitivity.
- Case 5: Use an ANN from case 1, trained to correct one LSM, and test it on another. Given there are only three models, we look at all six permutations. This is a test of *model* transitivity.

Cases 4 and 5 should make it clear that we are not only interested in using the correction technique regionally, but that we also wish to examine the similarity in the nature of LSM systematic error across vegetation types and the three models. We now look at the results from each of these five configurations in turn.

3. Results

a. Case 1: Temporal transitivity (both vegetation types)

In this case we use a single ANN to correct each LSM's simulations at all 13 sites. We train the ANN using the first half of the time series from all sites, and test on the second. This is essentially a multiple-site version of the technique described in Abramowitz et al. (2006). As noted above, for the correction of each of the three fluxes, we use downward shortwave, air temperature, and the *model output value* of the flux as the ANN inputs. In case 1, we also include a fourth input: vegetation type. This is an integer input that represents the dominant vegetation type, which in CBM and ORCHIDEE is either coniferous forest or grassland. For CLM, since dynamic vegetation was enabled, there was no restriction. The ANN output is simply error in the flux under consideration. Case 1 is the only case of the five that uses *all* time steps of observed data, including those flagged as unreliable in the quality control procedure mentioned above. That is, the *entire* first half of the time series from all sites forms the training set and the *entire* second half forms the testing set. This was simply to facilitate analysis and would likely make the correction by the ANN less effective.

Figure 2a shows the reduction in the per-time-step root-mean-square error (RMSE) of each of the three fluxes (columns) afforded by the ANN correction for the testing period in case 1. The y axis shows RMSE in each flux (μmol m^{-2} s^{-1} for NEE, W m^{-2} for Qle and Qh), while the x axis shows increasing resolution of the SOFM used in the ANN. This can be interpreted as increasing complexity or sophistication of the ANN used to make the correction. For example, the number "8" on the x axis implies a $8^2 = 64$ node SOFM. The RMSE value at $x = 0$ represents the RMSE of the uncorrected model simulation. The result shown is an average over all 13 sites. It shows a clear relationship between increasing SOFM resolution and decreasing per-time-step RMSE.

The results of case 1 are summarized in Table 2. The ANN afforded a correction of 10%–40% in per-time-step NEE RMSE with an appropriate choice of SOFM resolution, bringing the RMSE of *all* corrected LSM simulations to 4.0–4.2 μmolm^{-2} s^{-1}. The per-time-step Qle RMSE was reduced by 17%–41%, bringing each model's corrected simulation to a RMSE of around 37 W m^{-2}. The Qh reductions were on the order 19%–28% RMSE, bringing all corrected simulations to about 45 W m^{-2} of RMSE. We can also begin to see the clear systematic nature of model error. Even a single node SOFM ("1" on the x axis), which represents a linear correction to the LSMs, removes a significant proportion of the error. In all three fluxes in case 1, noting the lowest corrected values (Table 2 and the right-hand side of Fig. 2a), we see that the ANN acts as a convincing "model equaliser." That is, once corrected the three LSMs have nearly identical per-time-step RMSE. Whether or not this truly represents a theoretical limit or not is unclear.

This is, however, only one measure of performance. It gives us an idea of the amount each model deviates from the observation at each time step. We will consider other measures after detailing the next two possible configurations for the correction technique.

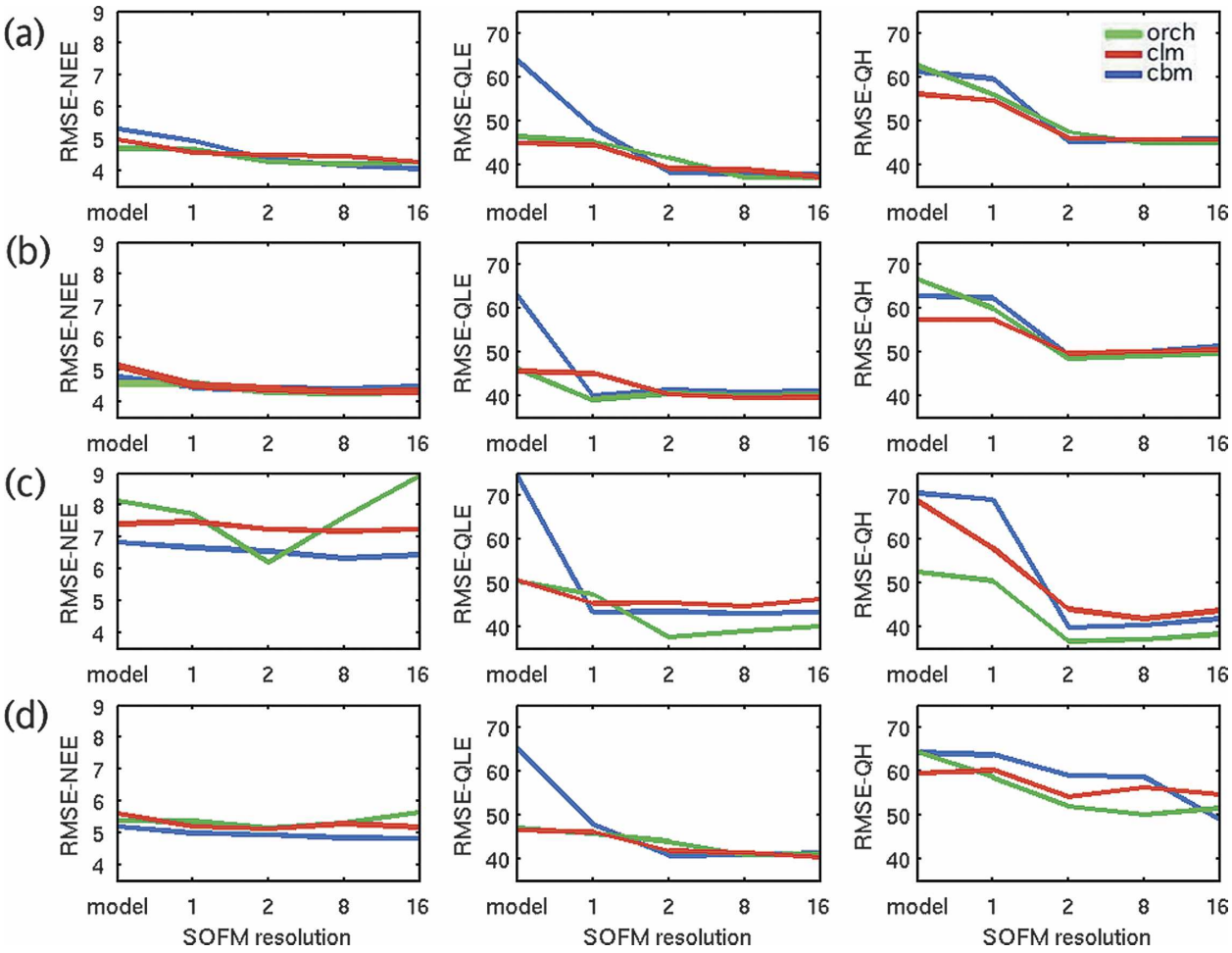

FIG. 2. The per-time-step RMSE of ANN-corrected LSM simulations: (a) case 1, temporal transitivity, all sites; (b), (c) case 2, spatial transitivity across coniferous and grassland sites, respectively; and (d) case 3, spatial transitivity across both vegetation types. Blue lines represent CBM, green represent ORCHIDEE, and red represent CLM. The x axis moves from uncorrected LSM simulation ($x = 0$), through a linear correction ($x = 1$), to a correction by a $16^2 = 256$ node SOLO ANN. Results are shown for NEE (μmol m^{-2} s^{-1}), Qle, and Qh (W m^{-2}).

b. Case 2: Spatial transitivity within vegetation type

Case 2 uses two ANNs to correct each LSM: one for coniferous and one for grassland sites. Whereas in case 1 we divided our total data into training and testing sets temporally, in case 2 we do so spatially. That is, for each vegetation type we train the ANN correction on the entire time series from some sites, and test it on the entire time series from other sites.

For the coniferous ANN correction, we train the ANN using the meteorological, model, and model error time series from four sites: Aberfeldy, Scotland; Bordeaux, France; Hyytiala, Finland; and Weiden Brunnen, Germany (see Table 1). We test the spatial transitivity of the ANN correction learned at these four sites by testing it at the other five: Flakaliden, Sweden; Loobos, Netherlands; Metolius, Oregon; Norunda, Sweden; and Tharandt, Germany. For the grassland/

TABLE 2. The per-time-step RMSE in NEE, Qle, and Qh for case 1. Corrections are shown for all three LSMs using optimal SOFM resolution (refer to Fig. 2a).

CASE 1	CBM	CBM correction	ORCH	ORCH correction	CLM	CLM correction
NEE (μmol m^{-2} s^{-1})	5.34	4.06 (24%)	4.71	4.22 (10%)	5.01	4.28 (15%)
Qle (W m^{-2})	63.94	37.85 (41%)	46.40	37.05 (20%)	44.94	37.29 (17%)
Qh (W m^{-2})	61.29	45.32 (26%)	62.73	44.88 (28%)	56.21	45.72 (19%)

TABLE 3. The per-time-step RMSE in NEE, Qle, and Qh for case 2. Corrections are shown for all three LSMs using optimal SOFM resolution (refer to Figs. 3b,c). Results are shown for the ANN trained and tested on (i) coniferous forest sites and (ii) grassland sites.

CASE 2 (i)	CBM	CBM correction	ORCH	ORCH correction	CLM	CLM correction
NEE (μmol m^{-2} s^{-1})	4.78	4.37 (9%)	4.56	4.28 (6%)	5.14	4.31 (16%)
Qle (W m^{-2})	62.90	39.87 (37%)	46.11	39.08 (15%)	45.57	39.49 (13%)
Qh (W m^{-2})	62.82	49.25 (22%)	66.61	48.43 (27%)	57.36	49.73 (13%)
CASE 2 (ii)						
NEE (μmol m^{-2} s^{-1})	6.84	6.34 (7%)	8.16	6.21 (24%)	7.42	7.20 (3%)
Qle (W m^{-2})	74.65	43.01 (42%)	50.41	37.61 (25%)	50.65	44.65 (12%)
Qh (W m^{-2})	70.43	39.80 (43%)	52.51	36.73 (30%)	68.73	41.85 (39%)

cropland correction, we train the ANN using the entire time series from two sites, Little Washita, and Ponca City, Oklahoma, and test on the entire time series from the other two, Bondville, Illinois, and Shidler, Oklahoma. This choice of sites was arbitrary, we did not examine any other combinations but believe that other configurations are unlikely to affect the nature of the conclusions. In both of these ANNs, we use downward shortwave, air temperature, and modeled flux as ANN inputs.

Figures 2b,c show the results of the coniferous and grassland ANN corrections, respectively. For the coniferous case, the nature of the correction are similar, although the magnitudes are slightly smaller. Results are summarized in Table 3 [case 2(i)]. With an appropriate choice of SOFM resolution, NEE RMSE is reduced by 6%–16% depending on LSM, bringing all corrected simulations down to around 4.3 μmol m^{-2} s^{-1} RMSE. The Qle RMSE is reduced by 13%–37%, bringing all runs to about 39 W m^{-2} RMSE. The Qh RMSE is reduced by 13%–27%, which brings runs to about 49 W m^{-2} RMSE. Note that the best correction to Qle for both CBM and ORCHIDEE in case 2 is a linear correction (i.e., a correction made by a one-node SOFM).

The behavior for the grassland site correction is different. For ORCHIDEE's NEE correction, there is a clear disadvantage to using higher-resolution SOFM ANNs, with the correction actually making the per-time-step RMSE worse for SOFMs with resolution $16^2 = 256$ nodes. This result indicates that the extra information about ORCHIDEE's error learned by the higher-resolution SOFM is specific to the training sites, and not applicable or transferable to the testing sites. It implies that ORCHIDEE does capture some of the detailed differences between sites, since if it could not, we would expect the RMSE to decrease with increasing SOFM resolution. With lower resolutions we can however decrease per-time-step NEE RMSE by between 3% and 24%, the Qle RMSE by between 12% and 42%, and the Qh RMSE by between 30% and 43% [see Table 3(ii)].

c. *Case 3: Spatial transitivity (both vegetation types)*

Here we use a single ANN, trained with the entire time series from half the sites and tested with the entire time series of the other half of the sites, with some of each vegetation type in both the training and testing sets. The training set consists of Aberfeldy, Bordeaux, Hyytiala, Little Washita, Ponca City, and Weiden Brunnen; the testing set consists of Bondville, Flakaliden, Loobos, Metolius, Norunda, Shidler, and Tharandt (Table 1). The inputs to the ANN are same as case 1: downward shortwave, air temperature, the model output value of the flux, and the vegetation type.

The results for case 3, shown in Fig. 2d and summarized in Table 4, again suggest that the ANN correcting ORCHIDEE's NEE simulation learns too much information specific to the training sites if the SOFM resolution is high. On the whole, however, with an appropriate choice of SOFM resolution, the correction still provides clear improvements in all fluxes. Note that here, as in the previous two cases, Qh requires a higher-resolution SOFM than Qle to achieve the best correction. This suggests that the nature of LSM systematic error in Qh is significantly more complex than in Qle.

TABLE 4. The per-time-step RMSE in NEE, Qle, and Qh for case 3. Corrections are shown for all three LSMs using optimal SOFM resolution (refer to Fig. 3d).

CASE 3	CBM	CBM correction	ORCH	ORCH correction	CLM	CLM correction
NEE (μmol m^{-2} s^{-1})	5.21	4.83 (7%)	5.39	5.14 (5%)	5.61	5.13 (9%)
Qle (W m^{-2})	65.23	40.79 (37%)	46.93	40.91 (13%)	46.53	40.35 (13%)
Qh (W m^{-2})	64.26	49.10 (24%)	64.30	50.02 (22%)	59.56	54.18 (9%)

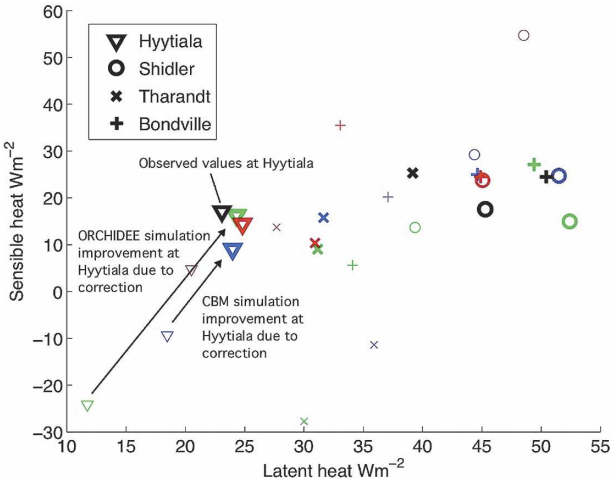

FIG. 3. Scatterplot of average latent heat (Qle) vs sensible heat (Qh) at four flux tower sites. Black symbols represent observed values, blue represent CBM, green represent ORCHIDEE, and red represent CLM. Bold colored symbols represent each LSM corrected by a 256-node SOFM ANN, while normal symbols represent uncorrected model values.

We now examine the effect of the case-3 correction on mean annual flux values and average diurnal cycle. In doing so, we will address the very real possibility that these corrections stem from the systematic *observation* error rather than the LSM error.

Figure 3 is a scatterplot of annually averaged Qle versus Qh for all LSMs and their corrected values at four sites, each represented by a symbol: Hyytiala (∇), Shidler (O), Tharandt (×), and Bondville (+). Black bold symbols represent the observed fluxes, while the three colors represent the three models (CBM in blue, ORCHIDEE in green, and CLM in red). Corrected LSM values (by an ANN with SOFM resolution of $16^2 = 256$ nodes), are shown as bold, while normal-type symbols represent uncorrected LSM simulations. The 4 sites are chosen to be broadly representative of the 13, both in terms of climate and the effect of the case-3 correction. If the correction has been successful, we would expect a symbol to "move" (from normal to bold) closer to its black counterpart. This is the case, for example, for CBM at Hyytiala—the bold blue triangle is closer to the black triangle than the normal type triangle. A quick visual inspection suggests that overall, the correction affords real improvements in this measure, consistent with the drop in per-time-step RMSE shown in Fig. 2d.

Figure 4 shows the average diurnal cycle for the three fluxes at the same four sites. LSM simulation values *without correction* are shown as dashed lines, while *corrected* LSM simulations are represented by solid lines. Observed values are again in black. As with Fig. 3, results in this measure are broadly consistent with per-time-step RMSE: while in some cases the correction does not improve LSM simulations (e.g., Shidler and Tharandt NEE), overall it does clearly enhance the LSM's predictive ability, particularly in the two energy fluxes.

d. Case 4: Vegetation-type transitivity

We now investigate whether the nature of the systematic model error learned by the ANN with one vegetation type is transferable to the other. That is, we use the ANN from case 2, trained to correct coniferous sites, and test its ability to correct the grassland sites from the other case-2 testing set, and vice versa. As in case 2, we only have the standard three ANN inputs; the ANN has no mechanism to recognize the change in vegetation type.

Figure 5a shows the results of the ANN correction trained on *coniferous* sites tested on *grassland* sites. This is the same testing set as in Fig. 2c; note that the uncorrected model values (the RMSE values at $x = 0$) are the same in both plots. In all three fluxes, for all three models, the ANN is able to produce a positive correction despite the apparent disparity in its training and testing sets. Results for Qle and Qh are summarized in Table 5. This demonstrates that the nature systematic error (a considerable portion of total error) in these LSMs is relatively insensitive to vegetation type, at least for the two vegetation types considered here.

Figure 5b shows the results of the ANN correction trained on *grassland* sites tested on *coniferous* sites. This is the same testing set as in Fig. 2b. Immediately we can see that the information learned about NEE LSM error at the grassland sites is not transferable to the coniferous sites; corrections to all LSMs at all SOFM resolutions decrease performance. Correction to Qle and Qh, however, are still strong (see Table 5). This suggests that the LSMs have some skill in simulating vegetation-specific NEE but not Qle and Qh, at least for these vegetation types. The asymmetry of vegetation-type transitivity presumably means that the range of NEE prediction error behavior at coniferous sites is larger and contains the range of behavior seen at grassland sites to some extent. Why this should be so is not immediately clear.

e. Case 5: Model transitivity

We now attempt to gauge whether the nature of the systematic error revealed in previous cases is common between these LSMs. We use an ANN trained to correct one LSM and use it to correct another. We use the training and testing sets from case 1, and therefore have

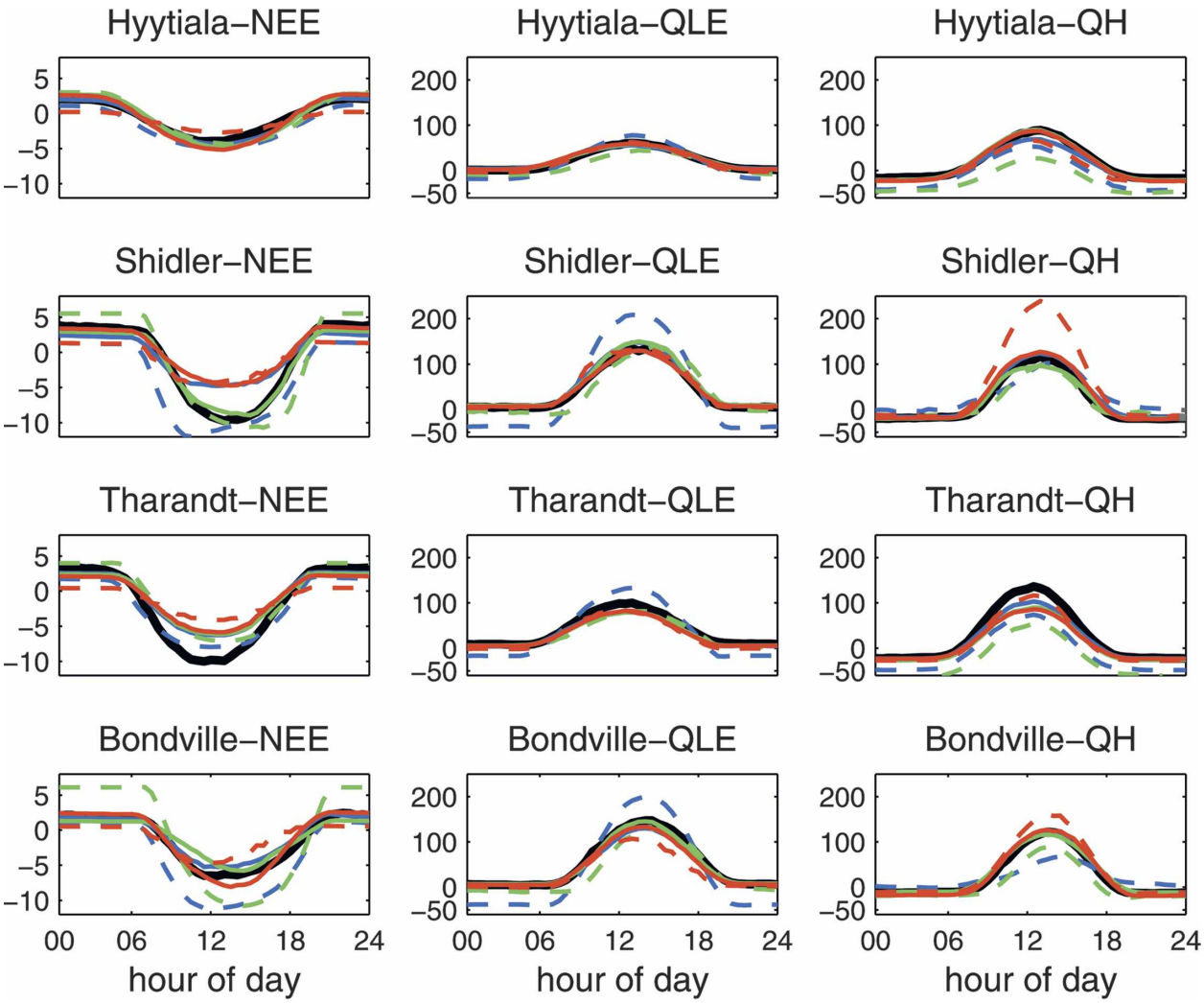

FIG. 4. Average diurnal cycles for NEE, Qle, and Qh at the same four sites as Fig. 3. Black lines represent observed values, dashed lines represent uncorrected LSM simulations, and solid lines represent each LSM corrected by a 256-node SOFM ANN. Blue represents CBM, green represents ORCHIDEE, and red represents CLM. NEE fluxes are in μmol m^{-2} s^{-1}; Qle and Qh are in W m^{-2}.

the same four ANN inputs as case 1. We examine all six possible permutations.

The two rows of Fig. 5c show the same results using two different scales. The top row uses the same scale as in previous cases, and the bottom row allows the full range of behavior to be shown. We use the shortened "CBM2ORC" to mean we have trained the ANN to make a correction to CBM but tested its ability to correct ORCHIDEE. The legend is as follows: CBM2ORC (black), CBM2CLM (dark blue), ORC3CBM (green), ORC2CLM (red), CLM2CBM (light blue), and CLM2ORC (pink). In considering Fig. 5c, it is useful to think of the relationship between pairs of models: CBM and ORCHIDEE (black and green), CBM and CLM (dark blue and light blue), and ORCHIDEE and CLM (red and pink). Note that the six lines converge to the three uncorrected LSM simulation values at the *y* axis, and that these three values are those of case 1 in Fig. 2a.

These results demonstrate higher sensitivity to the SOFM resolution than the previous cases. In all three fluxes, corrections to CLM by other model ANNs (red and dark blue lines) result in severe degradations at certain SOFM resolutions. Also, CLM-trained ANNs have little capacity to correct the other two models (light blue and pink). This suggests that the nature of CLM's systematic error is significantly different to that of the other two LSMs. This contrasts with the relationship between CBM and ORCHIDEE (green and black lines). The ORCHIDEE-trained ANN reduces CBM NEE RMSE from 5.34 to 4.66 μmol m^{-2} s^{-1} (13%); and vice versa from 4.71 to 4.50 μmol m^{-2} s^{-1} (4%). The ORCHIDEE-trained ANN reduces CBM Qle

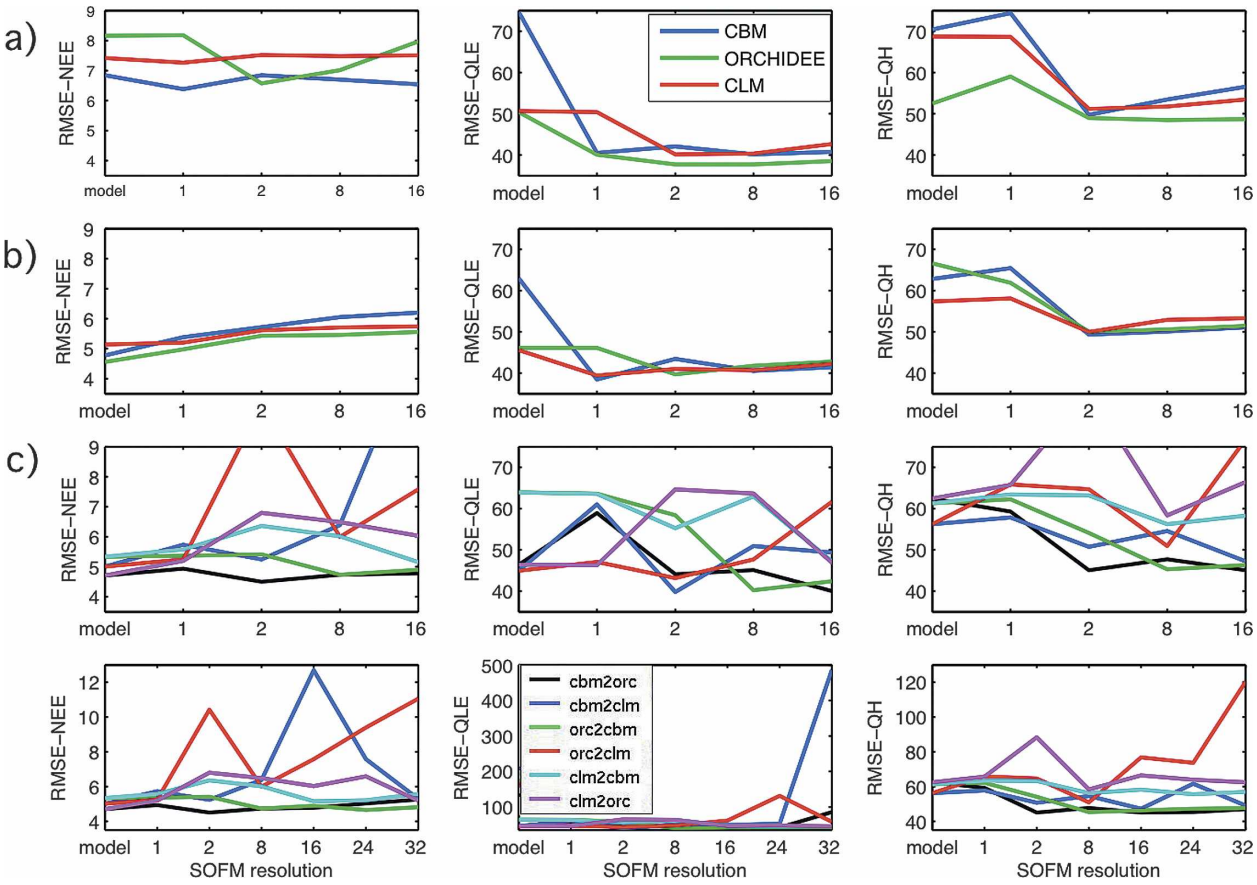

FIG. 5. The per-time-step RMSE in NEE, Qle, and Qh of corrected LSM simulations. Case 4 represents the ANN correction (a) trained on coniferous sites and tested on grassland sites, and (b) trained on grassland sites and tested on coniferous sites. (c) Case 5 represents the ANN correction trained to correct one model applied to another model, in all six possible permutations: CBM2ORC (train on CBM, test on ORCHIDEE, represented in black), CBM2CLM (dark blue), ORC2CBM (green), ORC2CLM (red), CLM2CBM (light blue), and CLM2ORC (pink). The two rows in (c) are the same plots on two different scales, to allow comparison with earlier plots.

63.94 to 44.64 W m^{-2} (30%) and vice versa from 46.40 to 40.05 W m^{-2} (14%). The ORCHIDEE-trained ANN reduces CBM Qh from 61.29 to 45.30 W m^{-2} (26%) and vice versa from 62.73 to 45.06 W m^{-2} (28%). This grouping of CBM and ORCHIDEE is also evident in the diurnal averages in Fig. 4, particularly for NEE and Qh. We suspect that these differences in behavior are likely to be the result of CLM using dynamic vegetation

TABLE 5. Summary of per-time-step Qle and Qh RMSE improvements due to a 256-node SOFM ANN. Results are shown for case 2 (spatial transitivity within vegetation types) and case 4 (vegetation-type transitivity).

Case	Qle (W m^{-2})	CBM	CBM correction	ORCH	ORCH correction	CLM	CLM correction
2(i)	Train grass, test grass	74.65	43.01 (42%)	50.41	37.61 (25%)	50.65	44.65 (12%)
4(i)	Train conifer, test grass	74.65	40.17 (46%)	50.41	37.71 (25%)	50.65	40.17 (21%)
2(ii)	Train conifer, test conifer	62.90	39.87 (37%)	46.11	39.08 (15%)	45.57	39.49 (13%)
4(ii)	Train grass, test conifer	62.90	38.45 (39%)	46.11	39.71 (14%)	45.57	39.41 (14%)
	Qh (W m^{-2})	CBM	CBM correction	ORCH	ORCH correction	CLM	CLM correction
2(i)	Train grass, test grass	70.43	39.80 (43%)	52.51	36.73 (30%)	68.73	41.85 (39%)
4(i)	Train conifer, test grass	70.43	49.73 (29%)	52.51	48.46 (8%)	68.73	51.18 (26%)
2(ii)	Train conifer, test conifer	62.82	49.25 (22%)	66.61	48.43 (27%)	57.36	49.73 (13%)
4(ii)	Train grass, test conifer	62.82	49.30 (22%)	66.61	50.00 (25%)	57.36	49.97 (13%)

where the other models did not have this activated, although a larger LSM sample size would be required to confirm this.

4. Discussion

Cases 1–3 demonstrate the potential of this type of correction to be used in regional simulations, at least in a LSM not coupled to a GCM. They also tell us something about the proportion of LSM error that is systematic error (and thus may be removed through improving the LSM). We saw in Fig. 2 that despite a wide range of per-time-step RMSE performance across the three LSMs ($x = 0$ values), once the correction was applied their performance was similar. Whether these values represent a theoretical limit imposed by the quality of observed data or the nature of the natural system is unclear.

How best to choose the SOFM resolution is also not particularly clear. Although in the extended analysis of case 3 we chose a 256-node SOFM, there is no reason to believe that this will always be appropriate. From Figs. 2 and 5a,b it should be clear that there may be a point at which increasing resolution does not provide any additional benefit and may in fact make some simulations worse. There are however several possibilities for improving this uncertainty. The first would be to ensure that the training set has a wider range of climatic conditions than the testing set (e.g., by making a more careful selection of training sites). This may well mean that all of the curves in these figures continue to decrease with increasing SOFM size. The second would be to improve the mapping ability of SOLO, for example, by adding a dimension reduction technique to the input data.

We must also consider the possibility that the correction's success is due to observational bias rather than LSM bias. That is, that the ANN is correcting reasonable LSM simulations to better match biased observations. If this were the case, in Figs. 3 and 4 we would expect *uncorrected* LSM values to be grouped together and observations to be relatively separate. While this occurs in some instances, it does not appear to be a consistent pattern. In Fig. 4, for example, uncorrected model simulations (dashed lines) of all three fluxes are consistently below observed values at some sites, consistently above at some sites, as well as spread around observations at others. In terms of annual averages (Fig. 3), some sites have uncorrected model simulations spread around the observed values (e.g., Shidler, Bondville, etc.), and others do not. While it is likely that some observational bias exists (especially given the energy closure issues discussed by Wilson et al. 2002), it still appears that model bias dominates. Looking at Fig. 4, it seems more plausible that the ANN has, for example, corrected considerable systematic bias CBM's prediction of Qle and CLM's prediction of Qh, particularly at Shidler. The incompatibility of CLM with either ORCHIDEE or CBM in case 5 also supports the idea that the bias in question predominantly originates from the LSMs. We also note that if these results were largely due to observational bias, the bias would have to be consistent across the many (and relatively independent) observational teams that collected the data.

Case 4 gives us considerable insights into where inside the LSMs these problems may occur. In Table 5 we can see that for Qle, the success of the correction is entirely independent of the vegetation type on which it was trained. That is, systematic error in Qle is independent of the parts of the model code that distinguish between these two vegetation types. The results for Qh tell us something different. While it is clear for grassland that a grassland-trained ANN is best for correction (as we would expect), it matters very little for the coniferous sites. This suggests that the much larger range of Qh behavior at the grassland has enabled the grassland-trained ANN to capture the nature of the models' coniferous error.

Case 5 confirms that the nature of systematic error is shared by two of the three LSMs. In general, this approach could be used as a technique for deciding which models have similar behavioral characteristics, beyond simply comparing the average values of their outputs. LSMs that have similar bias characteristics (e.g., CBM and ORCHIDEE) might, for example, be considered "dependent" and unsuitable for co-inclusion in a model ensemble. LSMs which have differing characteristics in this respect (e.g., CBM and CLM) we might consider as "independent," and be ideal for an ensemble simulation.

One of the major criticisms of ANNs is that they offer little insight into the processes they simulate. This is certainly true of the correction technique as we have described it so far. However, while the weights of nodes in feed-forward ANNs have very little concrete meaning (e.g., as used in van Wijk and Bouten 1999; Dekker et al. 2001), we suggest that the SOFM-based SOLO ANN used here could provide insight into the causes of LSM error. Each SOFM node represents a subset or "climatological regime" of the training set; those with large outputs (i.e., large LSM systematic error) should indicate the subset of conditions under which the LSM behaves poorly. While such an investigation is beyond the scope of this paper, this may be another possible tool for understanding which processes in an LSM require further development.

The existence of considerable systematic error in LSMs (albeit in a sample of only three models) suggests a need for refinement of the existing processes in LSMs and brings us to the question of how we can develop testing regimes that help us to avoid these problems. In this manuscript we considered three measures of performance. While there is much argument over what constitutes validation (Berk et al. 2002; Medlyn et al. 2005), for a general LSM validation we would hope for a more comprehensive set of tests than this. However, even if we had many performance measures, we still would have no mechanism to define how well an LSM *should* perform. Qualitative descriptions of model ability may be convincing, but the meaning of any quantitative description is not clear, as we have no benchmark with which to define "good" performance. Without having seen the solid colored lines in Fig. 4, we may have agreed that all three LSMs (dotted lines) were performing well. Given that a simple multiple linear regression has the ability to capture seasonal and diurnal cycles, we need to be clear that broad qualitative agreement of long-term averages should not qualify as model validation.

One possible solution to this problem is to use a "benchmark time series" to prescribe the level of ability we expect from a LSM. By providing a time series with the same time step size as the model under analysis, the benchmark delineates superior/inferior performance in whichever measure of performance is appropriate for a particular study. For example, Abramowitz (2005) used the SOLO ANN presented here as a statistical model to produce a benchmark time series. Based on site-by-site flux tower data, this biophysically based benchmark time series provided a stricter test at highly predictable sites (e.g., energy-limited sites), since it was derived from observed data at each site in question.

5. Conclusions

We have demonstrated the existence of a considerable systematic error in the flux outputs of three LSMs. This systematic error was as much 45% of the per-timestep RMSE. We have also demonstrated the ability of a neural network–based technique to correct this systematic error. The technique was demonstrated with three fluxes: latent heat, sensible heat, and net ecosystem exchange of CO_2, at 13 flux tower sites. The method's success was most marked for latent and sensible heat, but also evident in the net ecosystem exchange of CO_2.The technique also offered some insight into which LSM processes may be improved to reduce this systematic error, although the potential for using the self-organizing map structure of the ANN for this purpose was not explored.

The success of the ANN correction vindicates the work done by flux observation groups. The fact that the correction technique trained at some sites was successfully tested at others, even on other continents, implies that the quantities measured by these independent groups are consistent and directly comparable. The real possibility that models' transferability resulted from systematic *observation* bias rather than *model* bias appears unlikely as model values were regularly scattered around the observed values (see Figs. 3 and 4). Unfortunately our quality control process and the need for both meteorological and flux gap-filled data meant this study could only source data to adequately represent two vegetation types: coniferous forest and grassland.

By illustrating the considerable proportion of LSM error that is removable, the technique alerts us to problems in land surface modeling that might otherwise go unnoticed. When testing or validating a model, each user will have particular measures of performance that they deem important. Yet aside from *qualitative* agreement with observations, we have no objective measure to decide how well a model *should* perform in a quantitative sense. How good is good enough? In many of the cases presented here, a simple linear regression correction ("1" on the x axes of Figs. 2 and 5) to the LSMs provided considerable improvements in performance at sites not included in the regression calculation. These included sites with different vegetation types, on different continents, and in different climatic zones. On this basis we suggest the use of a benchmarking technique at a variety of sites, such as that presented in Abramowitz (2005), to aid the process of LSM refinement. Validation of a model would then require a level of model performance that reflected the complexity and variability of the sites being simulated.

Acknowledgments. We would like to thank Fluxnet. We would also like to thank Timo Vesala and Nuria Altimir for the Hyytiala data; Eddy Moors and Wilma Jans for the Loobos data; and Ray Leuning and Helen Cleugh for their advice and constructive critique.

REFERENCES

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

——, H. Gupta, A. J. Pitman, Y. Wang, R. Leuning, H. Cleugh, and H.-L. Hsu, 2006: Neural Error Regression Diagnosis (NERD): A tool for model bias identification and prognostic data assimilation. *J. Hydrometeor.,* **7,** 160–177.

Berk, R. A., and Coauthors, 2002: Workshop on statistical approaches for the evaluation of complex computer models. *Stat. Sci.,* **17** (2), 173–192.

Braswell, B. H., W. J. Sacks, E. Linder, and D. S. Schimel, 2005:

Estimating diurnal to annual ecosystem parameters by synthesis of a carbon flux model with eddy covariance net ecosystem exchange observations. *Global Change Biol.,* **11,** 335–355.

Chen, T. H., and Coauthors, 1997: Cabauw experimental results from the Project for the Intercomparison of Land-surface Parameterization Schemes. *J. Climate,* **10,** 1194–1215.

Dai, Y., and Coauthors, 2003: The Common Land Model. *Bull. Amer. Meteor. Soc.,* **84,** 1013–1023.

Dee, D. P., and R. Todling, 2000: Data assimilation in the presence of forecast bias: The GEOS moisture analysis. *Mon. Wea. Rev.,* **128,** 3268–3282.

Dekker, S. C., W. Bouten, and M. G. Schaap, 2001: Analyzing forest transpiration model errors with artificial neural networks. *J. Hydrol.,* **246,** 197–208.

Drewry, D. T., and J. D. Albertson, 2006: Diagnosing model error in canopy-atmosphere exchange using empirical orthogonal function analysis. *Water Resour. Res.,* **42,** W06421, doi:10.1029/2005WR004496.

Falge, E., and Coauthors, 2001: Gap filling strategies for long term energy flux data sets. *Agric. For. Meteor.,* **107,** 71–77.

Hsu, K., H. V. Gupta, X. Gao, S. Sorooshian, and B. Imam, 2002: Self-Organizing Linear Output (SOLO): An artificial neural network suitable for hydrological modeling and analysis. *Water Resour. Res.,* **38,** 1302, doi:10.1029/2001WR000795.

Keppenne, C. L., M. M. Rienecker, N. P. Kurkowski, and D. A. Adamec, 2005: Ensemble Kalman filter assimilation of temperature and altimeter data with bias correction and application to seasonal prediction. *Nonlinear Processes Geophys.,* **12,** 491–503.

Klinker, E., and P. D. Sardeshmukh, 1992: The diagnosis of mechanical dissipation in the atmosphere from large-scale balance requirements. *J. Atmos. Sci.,* **49,** 608–627.

Kohonen, T., 1989: *Self-Organization and Associative Memory*. 3d ed. Springer Series in Information Sciences, Vol. 8, Springer-Verlag, 312 pp.

Krinner, G., and Coauthors, 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. For. Meteor.,* **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. NCAR Tech. Rep. TN-459+IA, 50 pp.

Medlyn, B. E., A. P. Robinson, R. Clement, and R. E. McMurtrie, 2005: On the validation of models of forest CO_2 exchange using eddy covariance data: Some perils and pitfalls. *Tree Physiol.,* **25,** 839–857.

Oleson, K. W., and Coauthors, 2004: Technical description of the community land model (CLM). NCAR Tech. Rep. TN-461+STR, 174 pp.

Papale, D., and R. Valentini, 2003: A new assessment of European forests carbon exchanges by eddy fluxes and artificial neural network spatialization. *Global Change Biol.,* **9,** 525–535.

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

——, 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.

Saha, S., 1992: Response of the NMC MRF model to systematic-error correction within integration. *Mon. Wea. Rev.,* **120,** 345–360.

Sellers, P. J., and J. L. Dorman, 1987: Testing the Simple Biosphere model (SiB) using point micrometeorological and biophysical data. *J. Climate Appl. Meteor.,* **26,** 622–651.

Twine, T. E., and Coauthors, 2000: Correcting eddy-covariance flux underestimates over a grassland. *Agric. For. Meteor.,* **103,** 279–300.

van Wijk, M. T., and W. Bouten, 1999: Water and carbon fluxes above European coniferous forests modeled with artificial neural networks. *Ecol. Modell.,* **120** (2-3), 181–197.

Vrugt, J., H. V. Gupta, L. A. Bastidas, W. Bouten, and S. Sorooshian, 2003: Effective and efficient algorithm for multiobjective optimization of hydrologic models. *Water Resour. Res.,* **39,** 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. For. Meteor.,* **91,** 89–111.

——, ——, H. A. Cleugh, and P. A. Coppin, 2001: Parameter estimation in surface exchange models using nonlinear inversion: How many parameters can we estimate and which measurements are most useful? *Global Change Biol.,* **7,** 495–510.

Wilson, K. B., and Coauthors, 2002: Energy balance closure at FLUXNET sites. *Agric. For. Meteor.,* **113,** 223–243.