Biogeosciences Discuss., 8, C501–C514, 2011
www.biogeosciences-discuss.net/8/C501/2011/
© Author(s) 2011. This work is distributed under
the Creative Commons Attribute 3.0 License.

Biogeosciences Discussions

Interactive comment on "How errors on meteorological variables impact simulated ecosystem fluxes: a case study for six French sites" *by* Y. Zhao et al.

Anonymous Referee #1

Received and published: 8 April 2011

General Comments

This study attempted to evaluate the error introduced to model simulations of ecosystem-atmosphere carbon and energy fluxes by using gridded meteorology products rather than locally measured meteorology. The strategy of the study was to compare modeled fluxes at six field sites in France using the ORCHIDEE ecosystem model driven with locally measured meteorology with model simulations driven by a number of gridded meteorology products. Contributions to error in modeled fluxes were derived from differences between simulated fluxes across model runs. The authors have identified an important scientific issue: since coupled atmosphere-biosphere climate models

use large-scale gridded meteorology to drive ecosystem flux predictions, the effects of errors in the meteorological drivers on flux predictions could have important impacts on the reliability of contemporary flux estimates as well as future climate predictions. However, there are some fundamental issues with the implementation of this study and with the presentation of the results. The major issues are as follows:

1. The model was not well parameterized for the sites, and as a result model error dominates the total error. The authors note in section 4.1 that "correlations between modeled and observed fluxes are always rather low" and "ORCHIDEE needs to be further calibrated." In the results, model errors (as identified from simulations driven by "observed" meteorology) almost always dominate the total error. In several cases, the correlation with measured fluxes is lower for runs forced by local measurements than for those forced by gridded datasets. This raises serious questions about the usefulness of the results. Is it possible to accurately identify additional contributions to error when the baseline simulations already do a poor job of replicating observed fluxes? This is a problem when interpreting the results presented in section 5. Results are presented as ratios between model error and forcing error (sec. 5.2), or between forcing error and total error (sec. 5.3), so the amount of model error is central to understanding the implications of these results. The results are therefore only really applicable to other modeling studies with high model error.

2. The authors represent the site-scale observed meteorology as "truth", with no discussion of the uncertainty in these time series, or whether the scale of tower-based measurements is the proper match for the footprint of eddy-covariance measurements. The authors state that "site observed meteorology will be considered in the study as the truth against which meteorological model products can be benchmarked" (p 2471, line 11-12). There are problems with this approach:

A. Measurements contain uncertainty from factors such as sensor placement, gaps, equipment failure, and random variability. The uncertainty in measurements is never estimated or addressed in the manuscript. Gap-filling was done as described in sec.

BGD
8, C501–C514, 2011

2.3.1, but there is no discussion of the uncertainty added by this procedure. Since the bulk of the results focus on an analysis of contributions to error, it is crucial to include measurement error and uncertainty in this discussion.

B. Site-level measurements may not actually be the most representative scale for eddy covariance measurements. Eddy covariance measurements are representative of a footprint area, and there is no guarantee that the footprints of flux measurements and site meteorology observations match. Site heterogeneity could cause a mismatch. For example, topographical variations could be related to differences in temperature over the flux footprint that are not captured in tower-based air temperature measurements. Precipitation measurements can be especially problematic, because of the high spatial heterogeneity of precipitation from individual weather events. Soil moisture at sites could be affected by runoff from a large area, while site measurements of precipitation cover only a very small area. In such cases, interpolated products such as SAFRAN may in fact be more representative of the actual drivers than site measurements.

C. If site observations are in fact "truth", one would expect simulations driven by observed meteorology to be significantly better at replicating observed fluxes. In fact, "forcing ORCHIDEE with OBS meteorology compared to gridded products delivers only a small reduction of MAE" (p. 2485, line 11-12). For some sites and modeled fluxes, driving the model with observed meteorology actually made simulated fluxes worse relative to using gridded products. This suggests that observed meteorology is not in fact much "better" than gridded products when it comes to driving the model.

Because of these factors, I don't think it is appropriate to refer to observed meteorology as "truth." Rather, observed meteorology should be included as an alternate driver to gridded products, and the results should be re-evaluated from the perspective of *differences* between simulations with different drivers and sensitivity of results to variations in drivers, not *errors* compared to a benchmark.

3. One of the meteorology products, SAFRAN, was also used for gap-filling the "ob-

BGD

8, C501–C514, 2011

served" time series, making evaluation of errors introduced by using that dataset as a driver problematic. In the gap-filling process described in sec. 2.3.1, missing periods of meteorology were filled using the SAFRAN model. This means that the "OBS" and "SAFRAN" simulations are not really independent data sets, since they share some driver data. Were gap-filled periods excluded from comparisons for SAFRAN? Since model results would be affected by earlier periods even if those periods were excluded from direct statistical analysis, even a comparison where gap-filled periods were excluded would perhaps not be truly independent. The authors need to prove that gap-filling the observed datasets using SAFRAN does not bias the results of the study toward agreement between "OBS" and "SAFRAN" simulations, or they need to use an independent dataset for gap filling.

4. One of the gridded meteorology data products, REMO, has a daily time step. The method the authors used to produce half-hourly values for use in driving the ecosystem model, and for comparison with hourly observed meteorology, is never described in the manuscript and produces results that the authors themselves describe as problematic. Because of this, it is not clear how the REMO part of this study can produce useful results. The process used to produce half-hourly values from the other gridded products, which have time steps from one hour to six hours, is also never described.

5. The paper lacks a discussion section that puts the results in the context of existing literature. The authors do cite some literature in the introduction section to establish the purpose of the study in the context of existing literature, but there should also be a discussion section at the end where the authors cite literature to place the *results* of the study in context. Do the results support or contradict other studies? What were the results of previous studies that compared gridded meteorology products against site measurements? What is the potential of the results of this study to change or challenge the results of earlier modeling studies that used gridded data sets for meteorology?

Specific Comments

BGD

8, C501–C514, 2011

p. 2470, line 3: a better wording would be "potential positive feedbacks." Model comparisons show a wide range of predicted changes (e.g. Friedlingstein et al., J. Climate, 2006).

p. 2471, line 10: what is meant by "online"?

line 11-14: Whether local measurements should be used as "truth" and whether they are really the "best possible driver", especially when being compared to regional scale studies, is problematic, as I discussed above.

p. 2472: Description of flux processing and gap filling in appendix A should be moved to this section. Flux processing is central to the study.

p. 2473, line 9-10: What is a "livestock unit"? One animal? And what kind of animals are they?

p. 2474: It would be helpful to list required meteorological inputs for the model in this section, or include a reference to the section where they are listed.

p. 2475, line 7-8: Is this soil type accurate? What data is this based on? Do all the sites really have identical soil characteristics?

The process described in Appendix D should be included here in the methods section, as it seems to be a fairly important feature of the study setup. Does the model have a way to explicitly include removal of biomass? Is simply scaling NEE a realistic way to account for biomass removal? Was forest thinning at the managed forest sites included in the model setup?

p. 2476, line 13-18: Since SAFRAN is an hourly product, why were only daily values used? Is the phase of the sine wave accurate for each site? Rather than assuming times for max and min temperature, could a fit to a sine wave including phase be done for each site?

p. 2477: In sec. 2.3.2, it would help to summarize differences in temporal resolution

BGD

8, C501–C514, 2011

between gridded data products. These are included in the table, but highlighting that REMO is a daily product is important since this is an issue later in the manuscript.

ORCHIDEE requires half-hourly input data, yet none of the gridded data products has a half-hourly resolution. How were half-hourly values produced? This is especially important for the 6-hour and daily products.

p. 2478, line 1-6: The purpose of this calculation is completely unclear. Needs be to explained better.

line 10: Why are other sites only included in the appendix? This is one of the major results of this paper, and all of it should be included here in the results section. Analysis of how gridded data products perform at different sites around the region could be really interesting.

Sec. 3.1: Since only one of the gridded products has an hourly temporal resolution, how can the others be compared to obs at an hourly time scale? Maybe a 6-hour time scale would be more appropriate, since that is the temporal scale of most of the gridded data products. The process for conversion of REMO to hourly values is never described. Was this done for other products as well, or just REMO? Whatever process was used to convert REMO to hourly values produced results for Qair that are obviously inconsistent with observations and other data products. If the results are obviously wrong, a different method should be used. REMO is not being tested in any meaningful way if a method that is known to be flawed was used to produce the actual driver time series.

Sec. 3.2: Results are described for sites that are only shown in the appendix figures. Results and figures for all of those sites should be included in this section where they are described and interpreted.

p. 2479, line 22-24: Problems with observed precipitation data reinforce the issues with assuming that observed data is "truth." More complete information on uncertainty

BGD

8, C501–C514, 2011

CC BY

and quality of observed meteorology is needed in the methods section.

p. 2480, line 11-12: What r2 qualifies data as "correct"?

line 14-16: What is the statistical significance of the underestimate? Were gap-filled points excluded from these comparisons?

p. 2481, Sec. 3.4: In previous page, an r2 of 0.55 was characterized as correct, while here an r2 of 0.48 is described as "poorly captured by models". Is there a specific cutoff? There needs to be a consistent and justified basis for determining whether models are "correct" if they are going to be described in those terms.

line 10-11: Of the three sites being described, only one has observed LWdown at the inter-annual scale in the figures. Modeled LWdown should not be compared with values that resulted from the empirical gap-filling equation.

line 13-16: "badly simulated": Here it sounds like the measurements, and not the simulations, may be the problem

p. 2482, line 17-18: Again, conversion of REMO values to hourly resolution was never described, and the method produced poor results. A better method should be used, or perhaps this dataset should not be included at all. Running the ecosystem model using drivers based on a flawed conversion of a dataset does not test the effect of using that dataset in any meaningful way.

p. 2483, line 4-5: Again, raises issues with using tower measurements as "truth". Tower measurements do not sound representative of the footprint.

line 19-20: For hourly variations, it seems that the problem was the conversion to hourly resolution rather than the performance of the REMO model. REMO itself does not produce hourly predictions, and thus it is meaningless to evaluate its performance against observations at an hourly scale.

line 24-25: Rainfall is often highly spatially heterogeneous, so a likely explanation is

BGD

8, C501–C514, 2011

that tower-based precipitation measurements are not representative of rainfall at the scales represented by the models. It is not clear that the tower measurements are any more "true".

line 26: This method needs to be described somewhere! Appendix B refers to a nonexistent Sec. 1.3, and the gap-filling process in sec. 2.3.1 describes only a process for Tair, not Qair.

p. 2484, line 13-15: Why are those descriptions in the appendix? These results seem central to the results presented in the manuscript and should be included here.

line 20-22: Model structural errors are fundamentally different from model calibration errors. The calibration procedure for each site needs to be described in the methods section. The authors identify a need for further calibration; why wasn't this done before continuing with the study? If poor calibration was the problem, this could and should be fixed. Figure 5 shows that at several sites some correlations were actually lower using observed meteorology than using gridded data products, which raises serious questions about the basis of this study: if OBS "truth" produces worse results than gridded data products with "errors", there is a major problem.

p.2485, line 7-8: With correlations this low, is it possible to separate error compensation effects from random variations? What is the statistical significance of these differences in R^2?

line 11-12: In several cases driving ORCHIDEE with OBS delivers an increase in MAE. This is inconsistent with the idea that OBS is "truth" In this section or in the figure, it would be useful to see a comparison of MAE to magnitude of fluxes or expressed as a percentage error.

p. 2486, line 10: Given the fact that OBS-based simulations do not produce significantly better results than simulations based on modeled drivers, this entire approach to model error seems flawed. There is not adequate justification for characterizing distance from

BGD

8, C501–C514, 2011

CC BY

OBS-based simulations as "errors," since OBS-based simulations are not convincingly "true."

line 11-16, 21: Model structural error, calibration error, and errors in initial conditions are all fundamentally different features of model error. These cannot all be referred to as "structural error."

p. 2487, line 13: What are the units here? Is share of structural error referring to the ratio emod/etot? As I noted before, structural error and calibration error are different. Say "model error." Are these numbers referring to medians of the box plots in Fig. 7? Or means? Is this for all models, all sites? The variability is very high in some cases, which needs to be mentioned.

p. 2488, line 1-2: This assertion should be supported with references to literature.

line 10: structural, calibration, initial conditions errors

line 25-28: Forcing errors seem to be much smaller than model errors at inter-annual time scale for C fluxes in Fig. 7, so the data contradicts this statement.

p. 2489, line 1-3: Forcing error certainly doesn't look like it has a magnitude of 0.8 of model error for any of the fluxes at annual time scale. Is this statement referring to Fig. 7? Does not seem consistent with the figure.

line 10-11: Error numbers would be more useful if compared to total flux magnitude for each site.

line 16-17: Does the data really support this statement? Only one site with "difficult" meteorology is included, and it doesn't seem to be that much worse than the other sites.

p. 2490, line 12: Based on Fig. 8, GRI appears to have larger forcing error than LQE. It may have a larger ratio e_F/e_{mod}, but that says more about the model error than it does about the forcing error.

BGD

8, C501–C514, 2011

CC BY

Section 5.4: Why was only SAFRAN used here? Probably each forcing product would have different ratios based on the different mismatches between modeled and observed meteorology. It might be useful to run a comparison where each flux is increased by a set percentage, rather than replaced by a gridded product. This would give a defined, quantitative sensitivity of each modeled flux to a known change in the specific driver or combination of drivers.

p. 2491, line 13-17: The "concave" and "convex" terminology is somewhat ambiguous. A clearer description of what is meant here, or a demonstrative figure, would be useful. Is there support for the assertion that model fluxes respond in a nonlinear/concave way to drivers? Multiple model runs using a range of driver variations could be used to test this, and provide visual evidence to the reader of nonlinear effects. It is difficult to understand how to reach this interpretation just from looking at the numbers in Fig. 9.

p. 2492, line 19-20: Be precise about sources of model error (parameterizations, structure, initial conditions). The relative contributions of model and forcing error show here are really only relevant for other models with similar magnitude of model error. Since this study had high model error, are the results applicable to studies with well parameterized models or different levels of model error?

Appendix A: This could be moved to the methods section

p. 2493, line 21: The magnitude of these uncertainties should be shown in the manuscript, and included in the overall analysis of uncertainty and errors. How do model and forcing errors compare to measured flux uncertainty? This is fundamental to the interpretation of the results.

p. 2494, line 8: Temporal resolution should be included here along with spatial resolution. line 24: was linear interpolation used for all driver time series?

p. 2495, line 15-17: The text refers to Sect. 1.3, which does not exist. Is this meant to refer to 2.3.1? That section discusses Tair and SWdown, but does not describe how

BGD
8, C501–C514, 2011

other drivers would be converted to half-hourly temporal resolution. There needs to be a full explanation for all of this.

Appendix C: This section needs written analysis to be included, and should be merged with the rest of the analysis in the results section.

Appendix D1: This should be part of the results section. These results are interesting and very important for the interpretation of the rest of the study's results.

Appendix D2: The description of the optimization process should be in the methods section of the manuscript. A description of how the model performed at the various sites and time scales similar to the descriptions in D1 and D3 should be added, and included in the manuscript's results section.

It's not clear that this procedure is appropriate or realistic for sites where biomass is removed by harvesting. Shouldn't this be accounted for by changing the modeled biomass pools rather than scaling the ecosystem respiration? Maybe this is a source of some of the problems at crop sites

p. 2498, line 13-14: The math here is wrong. If TERobs/TERsim=0.67, then the overestimate = (TERsim - TERobs)/TERobs = (TERsim/TERobs) - 1 = 49%

Appendix D3: Move to results section.

At LBR, OBS and SAFRAN simulations seem quite a bit worse than other gridded driver simulations at the monthly scale. This should be discussed.

Appendix D4: Move to results section.

p. 2500, line 9: what is CV?

Technical Corrections

This manuscript would benefit from additional proofreading for word choice, grammar, and English usage. Some specific corrections follow, but this list is not exhaustive.

BGD

8, C501–C514, 2011

p. 2470, line 6-7: carbon flux and pool variability

line 17: one of the first studies

p. 2471, line 7-8: This allows explicit calculation of . . . and is consistent with . . .

p. 2474, line 5: use "predicts" or "simulates", not "describes"

line 14: are modeled

p. 2475, line 6: driven by meteorological. . .

p. 2476, line 7: "discarded in the follow": say "following discussion", or "are therefore not included in our analysis"

p. 2482, line 15: Figure 4a shows the MAE. . .

line 19: gridded data products

line 20: "null model"

line 27: dataproducts

p. 2490, line 21: driver

p. 2492, line 3: SAFRAN, which also has. . .

line 19-20: as much by meteorology as by imperfect. . .

p. 2498, line 16-17: overcome, not "overcomed"

line 23: slightly too small

p. 2500, line 1: seems to have had

line 11-12: tends to produce

line 13: overestimated

Table 1: Add definitions of vegetation class acronyms to caption

BGD

8, C501–C514, 2011

For figures including different time scales, the figure heading is "Annual" but the text refers to "Interannual" time scales. If these are referring to the same parts of the figure, they should use the same terminology throughout to avoid confusion.

Fig. 1: LQE is shown as LQI in the figure

Fig. 2 caption: Dashed lines

Fig. 3: LQE shown as LQI in figure

Fig. 4: Caption refers to three columns, but figure has five columns. Some plotted boxes go out of vertical scale. Caption should explain how box plots were calculated, and which sites and/or models were included in the calculations. Explain difference between A and B plots. LQE is shown as LQI in figure.

Fig. 5: Some numbers are unreadable due to low resolution of figure. LQE is shown as LQI

Fig. 6: Resolution of image is too low to see details of some boxes. Change LQI to LQE. Explain what dots outside whiskers mean (outliers?)

Fig. 7: Which model(s) are included in this figure? An estimate of the magnitude of fluxes would help put the errors in context.

Fig. 8: Explain which models were included in box plots. It is difficult to see the TER boxes because of the scale.

Fig. 9: Numbers are difficult to read because of low image resolution. Which sites were included in this part of the analysis?

Fig C: Combine with Fig. 2.

Fig D: Include a color legend on the first page of this figure. The caption does not match the color legend (LOCAL in caption, ORCHIDEE_OBS in legend). Instead of "water fluxes" say "sensible and latent heat fluxes".

BGD

8, C501–C514, 2011

CC BY

Interactive comment on Biogeosciences Discuss., 8, 2467, 2011.

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Discussion Paper