Biogeosciences Discuss., 8, C6214–C6243, 2012
www.biogeosciences-discuss.net/8/C6214/2012/

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.

Y. Zhao et al.

yan.zhao@lsce.ipsl.fr

Received and published: 2 April 2012

BGD Biogeosciences Discussions

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

Dear Editor,

We are grateful to the two anonymous reviewers for their constructive comments. We have revised the manuscript with corresponding to the reviewers' comments.

Yours sincerely,

Yan ZHAO on behalf of co-authors

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 ïňĆuxes by using gridded meteorology products rather than locally measured meteorology. The strategy of the study was to compare modeled ïňĆuxes at six ïňĄeld 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 ïňĆuxes were derived from differences between simulated ïňĆuxes across model runs. The authors have identiïňĄed an important scientiïňĄc issue: since coupled atmosphere-biosphere climate models use large-scale gridded meteorology to drive ecosystem ïňĆux predictions, the effects of errors in the meteorological drivers on ïňĆux predictions could have important impacts on the reliability of contemporary ïňĆux 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 ïňĆuxes are always rather low" and "ORCHIDEE needs to be further calibrated." In the results, model errors (as identiïňĄed from simulations driven by "observed" meteorology) almost always dominate the total error. In several cases, the correlation with measured ïňĆuxes 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 ïňĆuxes? 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 understand-

ing the implications of these results. The results are therefore only really applicable to other modeling studies with high model error.

The seemingly Âń poor performance Âż of ORCHIDEE are due to our filtering of daily time scale. On hourly and monthly scales, the performance of ORCHIDEE is comparable to previously published studies and could be called “good” (see table). Further, the same model version was intensively used in several site-level comparisons (Santaren et al.2007; Jung et al. 2007; NACP papers) and demonstrated performances similar to those of other process models designed for global applications.

Keenan et al. (2009) compared ORCHIDEE modeled daily GPP with flux data at 4 Mediterranean sites in Europe, showing good performance (their Fig 7). But the seasonal cycle was not removed. In our study, daily data have been de-seasonalized (Section 2.4 : Equ. 2) by removal of a centered 31-day running mean. Then simulated and measured de-seasonalized daily flux anomalies were compared with each other, to estimate the mismatch on synoptic scale, related to the ecosystem responses to changing weather regimes. In their study, the skill (R2) of ORCHIDEE to reproduce seasonal cycle of GPP is 0.68 in old version and 0.88 in an improved version, R2 of seasoanl cycle of LH is 0.39 at old version and 0.48 in the improved version. Their estimations are very close to our

Therefore, the “ poor “ model performance revealed on daily time scale outlined by the reviewer is principally due to our specific time domain filtering. A similar large model data disagreement on “daily to inter-monthly” scale was found by Mahecha et al. (2010) at *four flux sites using time-frequency localized decomposition. The same conclusion was also reached by Dietze et al. 2011 using spectral analysis of the model-data misfit at 9 North American flux sites.

Finally, we evaluated another terrestrial biosphere model ISBA-Ag-s from Meteo France (Zhao et al. in preparation) against the same site-level measurements using the same protocol, and found similar performances: low skill at daily compared to good

skills at annual and diurnal time scales. So it seems that the poor model-data misfit on daily scales is not specific to ORCHIDEE, but also applies to other models.

We have re-worded section 3.5 (lines 420-440) accordingly.

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-ïňĄlling was done as described in sec. 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.

We rewrote section 2.3 to address uncertainty in tower meteorological measurement. We have chosen the period of 2004-2007 which has less than 0.01% missing values to be re-gap-filled, so that the uncertainty caused by our gap-filling procedure can be neglected. We’ve added uncertainty discussion (for example conversion from daily to half-hourly values for Qair and LWdown) in the last section.

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 ïňĆux 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 ïňĆux 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.

We agree that there is a spatial scale difference (bias) between site meteorology observations and the footprints of flux measurements. A discussion of this effect was added in section 6 and Appendix B, but we have no quantification of this bias at each site. Yet, local tower meteorology must be closer to any 'true' meteorology controlling the fluxes in the tower footprint compared to interpolation of station data or numerical weather prediction model results.

The reviewer comment raises an important issue, that is the different spatial scale of the forcing products used in the study. To further investigate this question, we created a low-resolution version (80 by 80 km instead of 8 by 8 km) of the SAFRAN data, referred to as SAFRAN_LOW. Comparison of the modeled fluxes driven by SAFRAN_LOW compared to the original SAFRAN data set allowed us to show that there was no statistically difference between SAFRAN_LOW and SAFRAN fluxes at hourly, daily and monthly scales. At annual scale, the impacts of spatial resolution in meteorological forcing is significant, comparable to the difference between different met. Data set, but were found not to be significant on simulated carbon and water fluxes, thus not changing our conclusions. This was added in the discussion of section 6 (the last second paragraphe).

C. If site observations are in fact "truth", one would expect simulations driven by observed meteorology to be signiïňĄcantly better at replicating observed ïňĆuxes. 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 ïňĆuxes, driving the model with obser ved meteorology actually made simulated ïňĆuxes 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.

We followed the reviewer suggestion and replaced 'truth' by 'local' forcing in the revised manuscript, and treated as an alternate driver. However, the fact that local forcing does not brings a great improvement of the model result for CO2 and H2O fluxes is due to the dominance of model structural and parameter errors, more likely than to the fact that local forcing is far from the truth.

3. One of the meteorology products, SAFRAN, was also used for gap-ïňĄlling the "observed" time series, making evaluation of errors introduced by using that dataset as a driver problematic. In the gap-ïňĄlling process described in sec. 2.3.1, missing periods of meteorology were ïňĄlled 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-ïňĄlled 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-ïňĄlled periods were excluded would perhaps not be truly independent. The authors need to prove that gap-ïňĄlling 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 ïňĄlling.

Comparison was made during 4 years (2004-2007) when there is very little missing meteorological data gap-filled by SAFRAN (less than 0.01%). In particular during the growing season when comparison is made, there is nearly no missing values to be filled with SAFRAN. This has been clarified in revised manuscript (page 8 line 239-240) We clearly state the quasi-independence of OBS and SAFRAN forcing in section

2.3.

For the simulated fluxes, the impacts of forcing during previous period generally decreases to small effects after 2-3 months in ORCHIDEE, thus has rather small impacts on the growing season as we focus. Moreover, our comparison is based on time series of Âń anomalies Âż (Equ.2), so that the trend caused by previous forcing is removed from our analysis.

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.

We documented the procedure of conversion from daily to hourly REMO forcing and from 6-hour to half-hour in is described in 2.3.1 and 3.1 in the revised manuscript, respectively.

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?

Thanks for this suggestion. We've added a discussion at the end of this paper. We also inserted more references to previous works throughout the text.

Special comments

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).

Corrected

p. 2471, line 10: what is meant by "online"? This word was unnecessary, so we have removed it.

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.

We've re-written the sentence indicating the uncertainty in the observed, as well as discussing the scale mismatch issue(line 97-98).

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

Thank for this suggestion, we re-organized the manuscript as suggested and moved flux processing and gap-filling sections in section 2.1 (line 158-167).

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

Animals are cows. We modified the sentence to indicate that the extensively managed Laqeuille grassland is lightly grazed.

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.

We described the required meteorological inputs for running ORCHIDEE in section 2.3.1

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?

Sensitivity tests have shown that the simulated carbon and water fluxes are not sensitive to the soil characteristics . The reason is possibly that the hydrology process in the present version of ORCHIDEE is relatively simple so that it can't make much difference between different soil characteristics. We removed this sentence from the text to avoid confusion.

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?

We have taken this process to the method section. Harvest of biomass is included at crop sites; forest management (thinning during the model-data comparison period) is not included in the model, which may degrade the performance of modeling interannual variability as we discussed in the case of HES, where thinning events occurred in winter of 1995/1996, 1999/2000 and 2004/2005 (section 2.1, Line 130-131).

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 ïňĄt to a sine wave including phase be done for each site?

We use daily values rather than simply replacing missing hourly values with SAFRAN in order to keep consistency on daily time resolution. Our analysis shows that the sine wave assumption does possibly produce about 1 hour phase shift depending the season. However, phasing of hourly scale is given less attention than for other time scales in this study. The possible uncertainty caused by this conversion procedure is discussed in revised manuscript (section 6; page*).

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

Thank you for this comment. We rewrote sections 2.3.1 and 3.1 accordingly.

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

We rewrote section 2.4 to explain the procedure of this calculation and the rationale behind.

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.

We did this in order to limit the length of main text part. The revised manuscript follows the reviewer suggestion

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 ïňĆawed was used to produce the

actual driver time series.

More detailed explanation has been given in section 3.1. We've also made the comparison statistics based on 6-hour time scale (not shown). The conclusion of the study remains valid despite difference in the values in terms of R2 and MAE using 6-hourly instead of hourly (including extrapolated hourly from 6-hourly). The conversion procedure of REMO from daily to hourly values has been clarified.

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

Results and figures showing all the sites are now 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 and quality of observed meteorology is needed in the methods section.

We indicated that the poor quality of rainfall data at this specific site is due to instrumental malfunction.

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

We modify “correct” to “good” when r2 value >0.5 and significant at 95% t-test.

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

We have given the statistical signiïňĄcance and excluded gap-filled values (yet only 1%).

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 speciïňĄc cutoff? There needs to be a consistent and justiïňĄed basis for determining whether models are "correct" if they are going to be described in those terms.

Thank you for this comment. We indicated the R2 significance level in the text. We consider rounded R2 >= 0.5 as good in the entire manuscript if the correlation is significant (p = 0.05)

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

We agree with the reviewer and removed comparison between gap-filling equation and modeled LWdown

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

This sentence was rewritten

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 ïňĆawed conversion of a dataset does not test the effect of using that dataset in any meaningful way.

We included REMO because this forcing was widely used as driver to TBMs for the Carboeurope project (Vetter et al. 2009) including conversion to hourly data. Conversion of daily to hourly forcing is unfortunately common practice for ecosystem models, so we included it in this analysis in order to show that this method can produce poor results, raising caution on simulated fluxes. This has been clarified in revised manuscript

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

Changed. We've discussed with the PI at this site the overestimated Swdown is likely translated from the overestimated Tair in the gridded data sets due to the coarsely-resolved topography in the atmospheric analyses. The mismatch between tower mea-

surements and representive of the footprint is discussed in section 6.

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 obser vations at an hourly scale.

We agree with the reviewer and we restricted the comparison to daily data using REMO model.

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

We have modified this sentence. We agree that rainfall has the nature of high heterogeneity and even fine resolution data set as SAFRAN which contains synoptic station data can not solve this problem (Quintana-Segui et al., 2008). The problem of spatial mismatch when we use tower measurements as “benchmarks” is discussed in section 6.

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

Changed

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.

We put the descriptions in section 4.1

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 ïňĄxed.

The model was not calibrated for its parameter values at each site, for comparing the effects of forcing. The goal is to take the same model version as the one used for regional applications. We’ve deleted this sentence not to make further confusion.

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.

It is not that OBS is a worse forcing, but that there are error compensation from model structure that degrades the performance when OBS is used. Was clarified in the text.

p.2485, line 7-8: With correlations this low, is it possible to separate error compensation effects from random variations? What is the statistical signiïňĄcance of these differences in R2 ?

The improvement is not statistically significant in carbon fluxes but significant in water fluxes, nevertheless, indicating error compensation in water fluxes.

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 ïňĄgure, it would be useful to see a comparison of MAE to magnitude of ïňĆuxes or expressed as a percentage error.

It is not that OBS is a worse forcing, but that there must be error compensation from model structure that degrades the performance when OBS is used. Was clarified in the text.

p. 2486, line 10: Given the fact that OBS-based simulations do not produce signiïňĄcantly better results than simulations based on modeled drivers, this entire approach

to model error seems ïňĆawed. There is not adequate justiïňĄcation for characterizing distance from OBS-based simulations as "errors," since OBS-based simulations are not convincingly "true."

We cannot prove that site-observed meteorology is true, but it must be more realistic than any modeled large scale forcing because based on local observations. There is NO reason for a large scale forcing to better capture meteorological heterogeneity effects in the 1 km2 footprint of the flux tower than OBS. So we agree not to call OBS 'truth' but we argue that the most realistic simulations with perfect model should be obtained with OBS rather than with any other forcing.

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."

We modified the sentence as "model 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 ver y high in some cases, which needs to be mentioned.

Yes it refers to ratio of median model error to median total error (Emod/Etot). These numbers are for all models and all sites in Fig. 7. We re-plot Fig7 and 8 in a more straightforward way. in the new figs, units are (kgC/m2/year) and (W/m2) instead of variance units

p. 2488, line 1-2 (i.e. : The most interesting result is that the forcing error is not negligible, compared to other model errors. This comes a bit as a surprise because meteorology is generally assumed in vegetation modelling to be well enough known not to create a misïňĄt in modeled ïňĆuxes).: This assertion should be supported with

references to literature.

One reference (Spadavecchia et al., 2011) is added to show that forcing uncertainty is estimated as low (about 10% of total flux)

line 10: structural, calibration, initial conditions errors

Changed as "model error"

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

We re-plot Fig.7 to make this more clear (it is renamed Fig. 8 in revised manuscript).

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 ïňĆuxes at annual time scale. Is this statement referring to Fig. 7? Does not seem consistent with the ïňĄgure.

see revised Fig 7

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

Thanks for the suggestion, We indicated the forcing error in percentage of total flux magnitude.

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

We've deleted this statement.

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

Thanks for this reminder. LQE has a larger ratio Ef/Emod. Given the absolute val-

ues of Ef, Emod and Etot vary among sites, it's more meaningful to show the relative contribution.

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 ïňĆux is increased by a set percentage, rather than replaced by a gridded product. This would give a deïňĄned, quantitative sensitivity of each modeled ïňĆux to a known change in the speciïňĄc driver or combination of drivers.

We used SAFRAN here because that: 1) SAFRAN is shown to be the closest to OBS. Using SAFRAN can give the test how the lest difference in meterological driver can impact on the simulation of fluxes. 2) SAFRAN is the gridded data set used to drive ORCHIDEE and another model over France in another paper by Lafont et al. 2011. We thus check what are the possible uncertainties in these simulations. Percentage change of each flux is beyond scope of this analysis.

p. 2491, line 13-17: The "concave" and "convex" terminology is somewhat ambiguous. A clearer description of what is meant here, or a demonstrative ïňĄgure, would be useful. Is there support for the assertion that model ïňĆuxes 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 difiňĄcult to understand how to reach this inter pretation just from looking at the numbers in Fig. 9.

This sentence has been changed. We've abandoned the wording of Âń concave Âż and Âń convex Âż

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 parame-

terized models or different levels of model error?

We've changed. The sources of model error is not specified in our study. We've also added a reference (Jung et al., 2007). In their study, the simulated interannual variability of GPP in three examined ecosystem models (Biome-BGG, LPJ and ORCHIDEE) is striking sensitivity to the different meteorological driver data sets. Appendix A: This could be moved to the methods section

Changed

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 ïňĆux uncertainty? This is fundamental to the interpretation of the results.

The estimation of these uncertainties ($\sim$25 g Cm-2yr-1) is added here. Both forcing and model errors are considerably larger than the measured flux uncertainty. The comparion is made in section 5.3.

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?

Changed in revised manuscript

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

Changed It is in sec. 2.3."flux tower meterology"

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

Written analysis was added and reviewers suggestion was followed

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

Changed. This part was moved as section 4.1.

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.

We've re-organized the paper. The description of the optimization process is now in the method section (section 2.5). Description of TER has been added.

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

Thanks for this suggestion. We agree that this optimized procedure is far from perfect and it needs to be improved in the future. The problems at crop sites is expected to be some reduced by running a crop-specified version of ORCHIDEE which is not used in this study because it is still underway to improve. With this crop-included version of ORCHIDEE, the information on site management can be taken into account more realisitically than the current version.

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%

Changed

Appendix D3: Move to results section.

Changed

At LBR, OBS and SAFRAN simulations seem quite a bit worse than other gridded

driver simulations at the monthly scale. This should be discussed.

We made an error in the figure of monthly TER and NEE (Fig. D2, D3) where the LOCAL and SAFRAN simulations were not optimized as the other gridded driver simulations, which make them look quite a bit "worse". We've corrected this mistake in revised version (Figure 6). The simulations based on OBS and SAFRAN are not significantly worse than with other gridded data in terms of R2 and MAE.

Appendix D4: Move to results section.

Changed

p. 2500, line 9: what is CV?

CV = Coefficient of variation, defined in revised text

Anonymous Referee #2 Received and published: 4 May 2011 Zhao et al. present a model study testing the driver sensitivity of a well known terres- trial ecosystem model, ORCHIDEE. They compare model runs at six different eddy- covariance sites in France, driving the model both with observed in-situ meteorology, and meteorology derived from regional climate products. The main conclusions are that inaccuracies and biases in data from the large scale meteorology leads to differences in estimated ïňĆuxes of carbon and water from the ecosystem model. The authors argue that driver error is thus an important source of error in large scale simulations, particularly at longer time-scales.

The experiment is well conceived in that the authors clearly identify climate products in which they are interested, and suitable sites within the region of availability of those products. The comparison of the climate products is rigorous and should be of interest both to people using those products, and, to a lesser extent, to those with a broader interest in the extent of variability between different climatic data sets.

It study is poorly conceived, unfortunately, in that the ecosystem model is not well parameterized for the sites, leading to a model which poorly characterizes the ecosystem

sensitivity to climatic drivers. The baseline ORCHIDEE model performance is very poor, and seems much poorer than other works previously presented (e.g. Krinner et al., 2005; Jung et al., 2007a), particularly at the water stressed sites (where it has previously been shown to perform quite well (Keenan et al., 2009)). This leads to a common occurrence in the analysis where the model performs better when using regional climate data instead of climate data measured at the site. This detracts greatly from the signiïňĄcance of all statistical comparisons made between model output driven by local or regional based climate.

The authors therefore are forced to spend much of their analysis on teasing out the difference between what they call model structural error, and forcing error. I ïňĄnd this unfortunate, as I would argue and sensitivity in model output to changes in drivers is fundamentally erroneous as the model sensitivity to climate was not properly characterized by the model in the ïňĄrst place (as evidenced by the poor model performance using site observed driver data). The results therefore are not necessarily generalizable to other models. The focus on trying to separate driver error from model error prevents a detailed analysis of the (far more interesting) relative sensitivities of GPP and RE to errors in climate drivers.

I see two different ways in which these concerns could be addressed: 1. The authors could argue that the model uses PFT parameters and therefore should not be 'tuned' to a site (i.e. site level parameters should not be introduced). In this case, all comparisons with site level ïňĆuxes should be omitted, allowing the authors to focus on a more regional analysis. This would greatly beneïňĄt the comparison of the climate data products as independent meteorological stations could be incorporated (currently the six sites available to test your climate products far underrepresent the spatial heterogeneity of the region). Regional model runs could then be performed using each of the climate products to test sensitivity of model outputs to differences in drivers (as in Jung et al., 2007b, but focusing on the relative sensitivity of GPP and RE, and thus NEE).

2. The authors could consider the site level measured values of NEE and derived values of GPP and RE to be important baselines, in which case ORCHIDEE simulations would need to be rerun, ensuring that the model is accurately reproducing the observations when driven by the observed meteorology. This should not be too difïňĄcult, given that the model is already set up to run at these sites.

We thank the reviewer for this comment. The goal of the study was to use the model without site parameters calibration. Model-data comparison on daily unfiltered data shown in fig below shows good agreement as shown in previous studies. However, time scale separation highlights poor performances in particular on daily filtered time scale. âĂŤâĂ
Abstract: Page 2469 Line 6: please clarify what 7-14 degrees celcius refers to.

Sentence was corrected

Line 11: 'The seasonal cycle of air temperature, humidity and shortwave downward radiation is reproduced correctly by all meteorological models (average R2 = 0.90).' It is unclear what the authors mean by this. Do they mean the magnitude of change between seasons? Please be more speciïňĄc.

We mean the phase of seasonal cycle of meterological variables is reproduced correctly. Was corrected in the text.

Line 13-15: "At sites located near the coast and inïňĆuenced by sea-breeze, or located in altitude, the misïňĄt of meteorological drivers from gridded dataproducts and tower meteorology is the largest" This statement makes the inference that sea-breeze and altitude are responsible for the errors in the gridded data products. Such vague statements are not necessary and could be misleading. If you have shown that sea-breeze and altitdude are the culprits then state so clearly. Otherwise, omit such statements.

These speculative statements have been removed.

Line 17-18: "R2 between modeled grid point and measured local meteorology going

from 0.35 (REMO model) to 0.70 (SAFRAN model)" This statement could be read as the SAFRAN model is much better than the REMO model. Again, this is misleading, as here you are reporting the limits? This is not clear and should be reported more precisely.

Sentence rewritten

Line 23-25: "The magnitude of this forcing error is compared to that of the model error deïňĄned as the modeled-minus-observed ïňĆux, thus containing uncertain parameterizations, parameter values, and initialization" There is a break in the ïňĆow of logic here. There reader has the impression that you are reporting results (not methods) and are saying that model error is comparable (or of a similar magnitude to) forcing error. Please rephrase.

Sentence rewritten

Line 25: "The forcing error is the largest on a daily time scale, for which it is as large as the model error." It is to be assumed then that there is no bias in the forcing error? If there were to be a bias, this would lead to small errors in modeled NEE on the daily scale accumulating to larger errors in the annual sums. The ïňĄrst line of the abstract states that biases in forcings are to be analyzed, but there is no statement on the result of this analysis of biases other than that using different forcing products gives you different modeled NEE. Please clarify this and revise the whole abstract.

We've added a discussion on the bias of observed meterology which is used as benchmark(section 6, line843-856). The biases of each analysis data set are given as conclusion 2. We've rewritten the abstract.

Page 2471 Line 8: "which is not the norm in many biosphere models". Please remove. It could equivalently be stated as "which is the norm in many biosphere models", and thus has no information content.

Changed

Line 21: "uncertainty" Do you mean error?

Changed

Line 19-25. Perhaps bullet point these questions to aid the reader.

Changed, as suggested.

Line 25-28 These are two separate questions here. Please separate.

Changed, as suggested.

Page 2473, Line 12: "The six sites cover" -> "The six sites (Table 1) cover"

Changed

Page 2474. Line 16: "Stomatal conductance is reduced by soil water stress (McMur trie et al., 1990)" This was shown to be an ineffective approach for simulating soil water stress in ORCHIDEE by Keenan et al. (2009). That probably explains the very poor simulation of GPP by ORCHIDEE with observed climatology at the water stressed sites included in the comparison (PUE, AVI, and GRI).

Thank you for this comment, which was added in the text.

Page 2475, Line 6-7: "The last three drivers have no impact on the model output, and are thus discarded in the follow." Surface pressure and wind speed have no effect on ORCHIDEE output? Then why are they in the model?

Sentence changed

Line 7: " in the follow." in the following.

Changed

Line 10-11: "with data-gaps of 1% for Tair, Qair, rainfall and SWdown and of up to 12-days in duration at few sites." Please rephrase.

Changed

Section 2.4: Given the length of the manuscript, this section could be omitted.

This is a key section to illustrate the method we have used in model-data comparison. Cannot be omitted

Page 2482: Line 27. “other dataproducts”

Changed

Page 2483: Line 4. “This is possibly due to a slope exposure of 5 the LQE tower”. Please clarify – a slope exposure would justify a lower than normal Swdown, but higher than normal?

Changed. The problem is in the analyses rather than the measurements.

Page 2483: Line 7. “REMO is worse than the other 3 models” Please be careful when making such statements. REMO does not perform as well as the other 3 models at your sites, and that is all you can say. Each model is designed differently and may perform better or worse for different conditions. A model may be generally worse than another, but you can not say that conclusively with your limited number of test sites. This is an impor tant point and the manuscript should be revised throughout to account for it. e.g Line 18; “SAFRAN is superior to other models ” ... at our sites.

Thank you for this comment ; the revised manuscript was changed accordingly

Line 24: “This shor tcoming is however rather unimpor tant for CO2 ïňĆux modeling.” Such statements should either be proven by the results presented in the manuscript or backed up with citations. In this case, hourly rainfall variability may well be impor tant if you are interested in modeling hourly CO2 ïňĆux.

We’ve deleted this statement. We agree with the reviewer that hourly rainfall variability may be important for modeling hourly carbon fluxes.

Page 2484, Line 6: “and pointing out to local feedbacks of the vegetation” Variability in SAFRAN performance does not suggest local vegetation feedbacks.

Statement deleted

Line 10: “given their coarser resolution”. This is unclear. Is the LQE grassland site not representative of the broader area contained within the coarse resolution grid cell of the gridded products? Again, this is an example of seemingly unsubstantiated statements.

LQE site is not representative of average elevation and climate of coarse resolution grid point. We verified that the altitude of the grid cell containing LQE in the gridded dataproducts is lower than the real altitude. Thus air temperature in all gridded products is systematically warmer than at the site. Sentence changed in the text.

Line 21: “This suggests that model structural errors largely explain the small values of R2”. This is not true. The ORCHIDEE model has been shown to perform well at these sites (too many papers to cite here). The problem most likely is that the model was not parameterized correctly for these sites. The assumption seems to be that given PFT general parameters the model should perform well ’out-of-the-box’ so to speak. But there is little reason to believe that a particular site is representative of it’s corresponding PFT.

We explained the Âń poor performance Âż by the time scale decomposition used in this study, especially at daily time scale. Using daily data, the model-data comparison is as good as in former studies. We have explained this in revised manuscript.

Line 22: “This also suggests that for site-level study, ORCHIDEE needs to be further calibrated.” This does seem to be the case, and in fact it would appear to me to be the main problem with this manuscript. The authors attempt to use the measured ïňĆux data to test the impact of using different climatic drivers on NEE, but make no attempt to ensure that the model is ’behaving’ well at the sites when driven with the observed met data. Such an approach invalidates any statistical comparison of ORCHIDEE driven with different climate data to ORCHIDEE driven with observed met data. What is the meaning of better or worse in this case when ORCHIDEE can perform better with met data from a gridded product?

We've rephrased this section given the explanation above. We agree that if ORCHIDEE was better calibrated at the sites, the discussion in this study could be somewhat easier. But on other hand, regional simulations are carried by an un-calibrated version. On the other hand, we show that the modeled flux depends on the forcing. So any calibration of parameters at the sites may mask forcing errors, which is not desired to meet our objectives of assessing forcing errors.

Page 2485, Line 4-8: "On average for daily scale, driving ORCHIDEE with OBS meteorology gives higher correlations than 5 when using atmospheric analyzed meteorology, except for the LQE mountain grassland where using a modeled meteorology improves the value of daily R2 over ïňĄve ïňĆuxes from 0.16 to 0.28. This indicates error compen- sation in ORCHIDEE, where a biased forcing compensates for a structural bias." That ORCHIDEE with OBS met is only better than ORCHIDEE with gridded met makes for a ver y poor comparison. With only 6 sites, it would not have been much effor t to make sure ORHCIDEE was preforming as well as it could at each site. If ORCHIDEE is not parameterized correcltly for the sites, then it's climatic sensitivity must be poorly char- acterized (bad parameterization) and so any quantiïňĄcation in terms of gC sensitivity to different climatic drivers (as presented later on in the ms) will be invalid.

See explanation above.

Page 2485: line 20: "This indicates that poorly captured ecosystem processes that control the model-data misïňĄt differ at each site". The authors are to be commended for the openness in which they present the poor performance of ORCHIDEE at the sites. That said, I would strongly encourage them to rerun the analysis with better model runs. It is well known ORCHIDEE can do better than this, and I am sur prised that 'team ORCHIDEE' have an interest in releasing such poor results. Can ORCHIDEE really not do better?

We've declared the Âń poor performance Âż is attributed to rigorous validation ap-

proach, the performance to produce diurnal and seasonal cycles are comparable with previous works.

Page 2487, line 12: "model structural error". Be careful with the use of this term throughout the manuscript. What you are talking about is a mix of model structural error and parameter error, which are two very different things. Model error may be a better term to use.

Thanks for this suggestion. We've maken the changes.

Page 2487, Line 15: "For TER, the forcing error εF (orange) on hourly scale is negligible, because soil temperature and soil humidity that control soil respiration in ORCHIDEE exhibit no diurnal variability." So there is no bias in the temperature in the gridded drivers?

We've changed. The forcing error on hourly scale is small (75 gCm-2month-1) because the diurnal variability of modelled and measured TER is weak (35 gCm-2month-1).

Page 2488, Line 3: "1.8 and 3.0 g C m -2 year-1" 3gC year-1 is a ver y ver y small amount. There's something I'm not getting here? Is this gC day-1? Please report errors on the scale you are referring to.

Thanks for this examination. The unit is kg C m-2 and we've meken the correction.

Page 2489, Line 10: "εF = 716, 286 and 644 g C m -2 year-1" Again, I would really question what these values mean given that the model performs so poorly. If the modeled climate response to OBS met is not correct, then the model must be either over or undersensitive to some climatic driver(s). You then swap that climate driver for one from a 'gridded product', and are interested in how sensitive model output is to the change in driver, but model sensitivity is poorly characterized.

We've explained the Âń poor performance Âż by the time scale decomposition used in this study. Thus, this version of ORCHIDEE can used as other TBMs to make sensitivity experiments (see section 5.5). As suggested by reviwer 1, these values of εF have

been replaced by ratio εF/ εtot.

Page 2492: Line 20: “by meteorology than by ” as by

Changed

Page 2493, Line 8-10: “Maybe when calculating regional budgets, there are spatial error compensations in meteorological forcing that will make the situation better, and diminish the contribution of forcing errors”. The situation of compensating errors can not be considered to be something positive! Please remove this statement.

We agree that errors compensation is not a positive point. Statement was removed.

Page 2495, Line 7: “REMO refers”, REMO here refers to...

Changed

Page 2511; Fig 2. Please revise Axis labels and tick values – it is impossible, or at least very uncomfortable, to read this, and I have good eye sight!

We’ve redrawn the figures.

Fig 4. Same comment as above. Bear in mind that a lazy reader will skip over a graph if they meet any difïňĄculty in interpreting it.

We’ve redrawn the figures.

Fig 5. Similar comment as above – for example there is lots of room within each box, the number font could be doubled and still ïňĄt in the same space. These are small details but important if you want people to pay attention to your work. Please considered these comments when revising the other Figures in the manuscript. Also please use SI units for all axis – e.g., "gC m-1s -1", not "gC/m/s", Fig. D. Please insert a legend, each ïňĄgure should be independently interpretable.

We’ve redrawn the figures and used SI units.

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