Anonymous Referee #1
Received and published: 23 December 2010

The authors use the spatial autocorrelation of the net ecosystem exchange predicted by three biospheric models to compare and evaluate the models. This is done by a combination of methods from spatial statistics and multimodel inference. Given the sparsity of carbon flux measurements, models which work on a regional or global scale cannot be compared against data. It is an important and difficult question to assess the different models. Several studies addressed this problem with different techniques. The use of the spatial autocorrelation of the predicted fluxes as starting point of the analysis is new.

The methods used are variogram analysis to quantify and model spatial autocorrelation, variable selection based on Bayes Information Criterion, and geostatistical regression with generalized least square. These techniques are well known and standard in spatial modelling and comparing against real data. If this is an adequate and valid methodology also for the problem of model comparison in this article needs additional proof (see below). With some exception (see below) the paper is clearly written and well structured. The presentation is good and the scientific quality is solid. I am not so positive about the relevance: given models with given simulations runs are compared with known techniques. The starting question is interesting and important, but the results are mainly indications to the main differences between the models.

As an example of the results let me take Section 3.2.1. The only variables selected in all three models are evapotranspiration and MDBF. On the other hand these two variables explain only a very small part of the spatial variability. This means that model specific variables have a large influence and thus point to fundamental differences in the models. This is certainly an important finding, but it may not come to a big surprise for people acquainted with the models. In the course of the analysis the found differences have to be explained by expert knowledge and model understanding. All this is said and done correctly, but it leaves me with the comprehension that the presented method can only point to potential differences between the models.

The primary goal of the manuscript and the study was to compare model estimates of flux. More specifically, the goal of the manuscript is to examine the environmental variables that appear most significant in explaining the spatial variability of modeled fluxes when you do not have access to the model code, driving variables, etc. Often times, the approach taken is to look at the correlation of estimated fluxes to a given (single) environmental factor, such as precipitation or temperature. Here, we are taking a more systematic approach, by using a combination of spatial autocorrelation, variable selection, and geostatistical regression. Combined, these methods allow for a more complete analysis of what environmental factors are most correlated with a model estimate of land-atmosphere surface flux. We agree that such an analysis is no substitute for a detailed sensitivity study using a single model or a small group of models. However, such sensitivity tests require new model runs, as well as access to model code and all of the associated driver data used by the model. The methods presented here could be used to compare model estimates when sensitivity tests are not possible or practical (e.g., code not available, many models to compare, etc.).

The discussion and conclusions will be revised accordingly to better clarify the benefits and limitations of this type of approach for model intercomparisons.

I have some points of criticism concerning the methodology.

(1) Variogram analysis is certainly the appropriate tool to study spatial autocorrelation, however it may not be adequate to compare spatial autocorrelations. At least mathematically two completely different random fields can have the same spatial covariance structure.

Variogram analysis was used to assess the degree of spatial variability in the model estimates of flux. We agree that two different random fields could have the same or identical spatial covariance structure. By extension, two maps of flux that appear at first glance very similar (visually), might have two different and distinct covariance structures. Variograms were used as a tool to assess the degree of spatial variability in the model estimates. We do not use the variograms to assess whether models are similar. In fact, two models may estimate fluxes with very similar variogram. This does not mean, however, that the two models that generated these fluxes are themselves similar. However, if two models yield fluxes with very different variograms, then we can conclude that the models that generated those fluxes are likely very different. We have clarified this point in the text, along with the general usefulness of the variogram analysis.

(2) More important is the connection between spatial features of the covariates and those of the modelled fluxes. The authors attribute the spatial differences to the different models. However, since the modes are driven with different data sets derived from different sources and hence expressing different spatial pattern, I don't think this is valid without a test of the spatial properties of the driving data. The authors should take the following steps to show that the differences can be attributed to the models and are not solely caused by the resolution and spatial properties of the driving data: In the table listing the driving data, there should be information added about the native spatial resolution of the data and information which model is using which data set. Table 1 is misleading since it does not list the modis active fire product as a driving data which it certainly is and similar for the NDVI, since here is no mentioning that two different variants derived from different sensors are used as mentioned in the text.

We agree with the reviewer that the models are driven by different data sets, from difference sources, and as a results, these differences will influence the spatial pattern and variability of the model's estimate of flux. When we use the term "models", we are referring to the model in its entirety (model structure, driving variables, assumptions, etc.). We can make this point clearer in the revised manuscript.

As suggested by the reviewer, we can include additional information about the driving data of each model in Table 1.

(3) Also the authors should make a similar analysis (list variograms) for the sets of driving data. (Some of this could be done in an appendix.) I actually expect that the variograms listed in figure 5 are similar to the variograms of their driving data sets (the ones where the models differ). If this is the case the whole study of limited scientific value, since it does not introduce new methods. Until the authors have not proven the effect of the

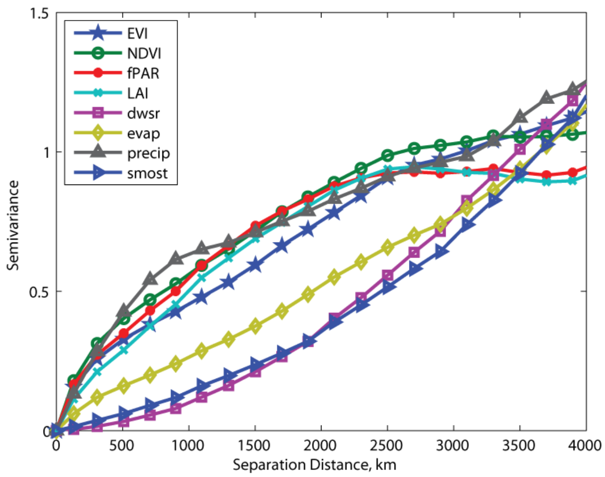
models on the spatial variability, this effect remains at least questionable. For example for the two CASA variants it would be sufficient to investigate the spatial pattern of the modis active fire product, and the two different NDVI products used in the analysis. I would find it actually very interesting to compare the spatial properties of the driving data directly with the spatial properties of the models.

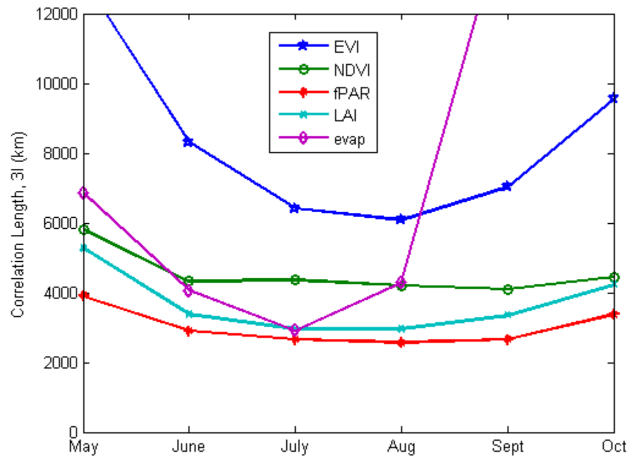
We believe that the reviewer is asking whether the spatial scales of variability in the fluxes are simply reflective of the spatial scales of variably of the driving data; or, conversely, whether the model structure itself actually influences the spatial scales of variability in the estimated fluxes. We are not sure that one could completely isolate the impact of the model structure on spatial variability compared to the variability that results from the environmental data used to drive the model. Nor, are we entirely sure how important this separation really is. However, we did construct variograms of a subset of environmental variables used in this analysis. Because the environmental variables differ in units as well as the magnitude and range of values, we cannot compare variances of the environmental drivers to variances of flux estimated from the models. We did compare the spatial scales of variability (i.e., correlation lengths) among the different environmental variables to the correlation lengths estimated for model estimates of NEE and GPP.

Many of the environmental variables are non-stationary (e.g., precip, soil moisture, radiation) and thus correlation lengths cannot be calculated. Using the same approach as described in the manuscript, we did however calculate the summertime correlation lengths of EVI, NDVI, fPAR, LAI, and evap, and these range between about ~3000km and 6000km. These correlation lengths are significantly longer than those calculated for model estimates of GPP (range between 1100km to 1400 km). Thus, the scales of spatial variability of the environmental variables alone cannot be used to explain the scales of spatial variability in the models.

We do agree with the reviewer; however that, the scale of variability in environmental variables, such as those that reflect phenology, likely explain some of the spatial variability seen in the models. In fact, determining which variables are more important at explaining the patterns of flux estimated from the models (and comparing these variables across models) was one of the goals of this manuscript. The point of the approach presented is to evaluate which auxiliary variable(s) "explain" the variability in the response variable (flux). We believe that looking at the individual driving variables (and their individual variograms) would not be informative (just like using a linear regression approach with a single variable as described in the manuscript). This is because it is difficult to isolate the explanatory power of the individual environmental variables by looking at them in isolation (as discussed and shown in the manuscript). Any unexplained variability may be a result of model choices (model specific input environmental data), assumptions, scaling approaches, and model formulations. Thus the comparison provides information about which variables appear to explain some of the variability in modeled flux among the models, whether these variables are consistent or different among models, and how much of the variability remains unexplained (as a result of environmental variables not considered, model assumptions, scaling approaches, model formulations, etc.).

Based on the reviewer's comments it appears that the goals of, and information gained from, this type of approach for model comparisons need to be made clearer in the manuscript. We have revised the text to address this reviewer's concerns and clarify any misconceptions about the goals of the work.





A minor point is: spatial autocorrelation can be a complicated matter and we hope to capture its features with the variogram. To reduce everything to one single number h0 may be possible in some cases, but I see no proof that this is the case here. Furthermore the classical parameters of the variogram have a natural interpretation, which is not the case for h0. In addition the choice of max makes its value rather arbitrary.

We do examine the varigoram parameters (variance, range) independently, in addition to comparison of h_o. Refer to Section 3.1 of the manuscript.

(4) Regarding the presentation, I have three objections. (1) The method section occupies 10 of the 25 text pages and is too long. It is appropriate to give a short review of the three models, as this is needed to understand the discussion section. However, much of the description of the statistical techniques is dispensable; e.g. the formula for the drift coefficients is well-known to experts and of no help for laymen. On the other hand things like the used variogram model or probability assumptions for the calculation of the likelihood have to be mentioned. The combination of variogram analysis, variable selection, and geostatistical regression was explained in several papers of the University of Michigan group and need not be repeated — just cite it.

We have attempted to shorten the manuscript as much as possible, referencing other's work when appropriate to reduce the length of the methods section., However, we have left relevant details in the manuscript so that the paper to stand on its own.

(5) The goal of the paper "[..] introduce a set of methods [..] and adapt them for use as tools for comparing model estimates" is too vague and as such not sufficient. Is it an assessment, comparison, and evaluation of models or of methods? Both is said/done in the text and for both I see deficiencies. For the moment I read it as an approach with one combination of methods, that gives indications about potential differences of the models,

which in the following have to be explained with expert knowledge and model understanding.

We agree with the reviewer, and have revised the manuscript to state the goals of the work more clearly in the introduction. The objectives have been revised to read: "In this paper, statistical tools that account for the spatial autocorrelation in land-atmosphere carbon fluxes are applied in order to compare flux estimates across models. These tools can be applied in light of inherent differences among these models, and without the need for new model simulations. The goal of this work is to evaluate these statistical methods in terms of their ability to assess the overall similarities and differences in modeled flux, as well as identify the environmental drivers that appear to have the greatest control over the spatial variability of predicted fluxes. Thus, the objective of this study is to assess the methods presented in terms of their ability to: (1) quantify the degree of spatial variability or autocorrelation of modeled carbon exchange across North America; (2) identify the environmental variables that appear most significant in explaining the spatial variability or patterns of modeled fluxes; and (3) quantify the relationship between these variables and modeled flux. Thus, the goal of the manuscript is not to perform a detailed intercomparison of biospheric models, but instead to evaluate the methods using a small set……"

(6) The emphasis on a quantitative approach is a bit misleading. Of course the statistical methods are quantitative, but the results are qualitative and give only indications to the main differences between the models. I am also not happy with the title: most readers will expect a comparison of the fluxes (totals or spatial distribution) and not a study 'only' of their spatial autocorrelation pattern. It would be desirable, as also said by the authors, to have a joint analysis of the temporal and spatial autocorrelation. But this would mean asking too much in a review.

We agree with the reviewer that while the approach is quantitative, the results are more qualitative. We have revised the manuscript to switch the emphasis from quantitative to a more "objective" and "systematic" approach of comparing model estimates to one another and potential explanatory environmental variables.

We also agree with the reviewer that we can revise the title slightly to reflect the revised manuscript.

Technical

p. 7916, l. 24: Do you really want to refer to (Eq. 5) here?

No, this should be equation 4, not 5. Thank you for pointing out this error.

p. 7918, l. 7: I see the minimum in September.

Yes, this should read September, not October

p. 7919, l. 2: From where do you get these lengths?

These lengths are from the h0 parameter defined in equation 3.

p. 7943: Add a unit for separation distance in the axis label.

Yes, we can add a unit (km) for separation distance here.