Biogeosciences Discuss., 7, C1210–C1217, 2010
www.biogeosciences-discuss.net/7/C1210/2010/

Biogeosciences Discussions

Interactive comment on "A geostatistical synthesis study of factors affecting gross primary productivity in various ecosystems of North America" *by* V. Yadav et al.

V. Yadav et al.

vineety@umich.edu

Received and published: 31 May 2010

The authors thank the reviewer for his very thoughtful suggestions to improve the quality of the manuscript. Responses to individual comments are included below. (Original referee comments are in italicized text.)

Flux towers observe NEE not GPP. While the authors use a consistent algorithm to infer GPP using the MDS method, there are assumptions made by the MDS method that is likely to modify the estimates of GPP. While MDS does not use functional forms as other environmental variable regression methods (which would certainly complicate

interpretation of results here), it does assume that the NEE data can be partitioned by a variety of environmental factors. My worry is that these assumptions bias your results of environmental controls, perhaps over or under estimating an effect. I recognize that the authors are not flux tower operators, but I think at least a discussion of the role of flux partitioning algorithm on the robustness of the relationships is worth discussion. The paper does mention Desai et al (2008), which is more about differences in GPP estimates by method and the importance of using a consistent method when estimating GPP across sites; that paper did not directly address this issue of regression environmental variables by method

We accept that the choice of flux partitioning algorithm to an extent may affect the presented results, although we did select the MDS method because we believe that it minimizes the risk of bias, as the reviewer points out. We had explored this concern in more detail in a previous publication (Mueller et al. GBC 2010, currently available at http://www.agu.org/contents/journals/ViewPapersInPress.do?journalCode=GB). However, in response to the reviewer's concern, we have now incorporated a discussion of the flux partitioning algorithm and its effect on the strength of identified relationships. However, we also want to reiterate that, as part of this study, we did compare the independent variables selected at each site to explain GPP (as estimated using the MDS method) to those selected for GPP derived from another method (i.e., artificial neural network), and found no difference in the selected best model. Furthermore, the regression coefficients (betas) of the independent variables associated with these two GPP products were within the 2 standard deviations of one another. This point has been clarified in the revised manuscript.

My other concern, and perhaps this comes from a lack of understanding, is the insistence on using a linear additive model as the only type for model selection. I understand the nature of model selection process somewhat forces this constraint here, since this

BGD

7, C1210–C1217, 2010

*is in essence a sophisticated multiple linear regression, but we know that factors that influence GPP are not additive, but rather multiplicative (limiting factors). The Montieth light use efficiency style equations are GPP = LUE * f(temperatre) * f(light) * f(etc...),and here instead the authors fit GPP = f(temperature) + f(light) + ... I'm not convinced the two are interchangeable, but I think it might be worth convincing (dense?) people like me that there is some level of equivalence between the two.*

The reviewer is correct that linear additive relationships are not equivalent to multiplicative (interactive) relationships. However, it is generally recommended that factors representing interactions among multiple variables should only be included if there is a scientific rationale (process-based understanding) for including them and/or if they are needed to explain the observed variability. We did include interactions between auxiliary variables in the superset of factors on which model selection was performed (e.g. temperature $\times$ Vapor pressure deficit, EVI $\times$ Soil Moisture $\times$ Vapor pressure deficit), but these factors were never selected through the BIC model selection , and hence we decided to use only additive relationship between variables. Additionally we also want to emphasize that if interactions among variables are included in a regression model, then the interpretation of the regression coefficients (betas) becomes more difficult. In the discussion paper, the reason for excluding interactions among auxiliary variables based on preliminary testing was not mentioned, but we have included this clarification in the revised manuscript.

While discussion of differences between sites are discussed, I'm perplexed why the authors didn't go to the next step - namely trying to jointly fit sets of sites with the same approach. It would seem to be a natural thing to do and allow one to estimate what are the larger scale controls on GPP. In one sense, this comes from my feeling that if one works hard enough, there is always some variable somewhere that "looks" like GPP. Why not include CO2 concentration or tower-observed fPAR, for example?

BGD

7, C1210–C1217, 2010

We thank the reviewer for this valuable suggestion. In the revised manuscript we have included results obtained by jointly fitting the data for all six sites using the method outlined in the paper. For simplicity, we limited this additional analysis to the daily scale. At this scale, the best model on the basis of BIC included EVI , Global Radiation, Air Temperature, Soil Temperature, Vapor Pressure Deficit and Precipitation. The R2 between GPP across the six sites and these variables was 0.72. A discussion of this additional analysis have been included in the revised manuscript.

Similar to my comment about flux tower, there is some importance to discussion of "know thy data" that is missing in the discussion of the environmental variables. Granted, this is reasonable given the nature of the analysis here, but I'm curious about the gap-filling of environmental data and the strong correlations of LAI, EVI, NDVI, as they are all derived from the same satellite, using a similar combination of spectral bands. For example, I don't really understand a mechanistic explanation for why one site may be better explained by EVI and another by NDVI. I thought EVI was NDVI on steroids? I would suggest discussing a bit more on the provenance and inherent self correlation among the data.

The reviewer is of course correct that LAI, EVI and NDVI are derived from same satellite and a similar combination of spectral bands, and for some sites and scales these indices are also highly correlated to one another ($0.70<\rho<0.85$). We agree with the reviewer that there is no mechanistic reason why one site would be better explained by EVI and another by NDVI. It might just be due to the better 'statistical performance' of a particular vegetation index in explaining the phenology of a particular site. In the interpretation of the results, we have clarified that the interpretation of selecting LAI, EVI, and NDVI should be similar, and that one should not interpret the choice of one of these variables over another without first identifying a plausible mechanistic explanation for the difference.

BGD

7, C1210–C1217, 2010

We do note, however, that the model selection process never selects more than one of these variables for any of the sites, which is also a positive result. In other words, the model selection step is able to identify that these variables are providing very similar (and therefore redundant) information for the examined sites.

Does the model selection and regression account for self-correlation of data? I think it does, but I suggest the authors make the case in the paper.

To an extent, correlation in the independent variables is accounted for by the model selection routine, in the sense that highly correlated data will provide redundant information in terms of explaining the observed variability. As a result, adding more than one of these variables will increase the size of the candidate model, but will not significantly improve the fit of the model (i.e. reduce the size of the residuals). As a result, the model selection step is unlikely to select highly multiple correlated variables.

However, to avoid such possible problems, it is good practice to either remove highly multi-collinear independent variables before performing model selection, or to group independent variables according to the physical properties they measure and only allow at most one variable representative of each property to be chosen by the model selection routine. More details about the latter approach for flux tower sites is discussed in Mueller et al (2010). As described in the reply to the previous comment, we did not encounter problems with multiple very similar variables being selected for any one site, and such a formal process was therefore not implemented.

In addition, we looked at the condition number of the correlation matrix of the auxiliary variables, and also at the correlation between the estimated regression coefficients (betas) for the best model (equation 7). Whenever the correlation between betas was > 0.70, we presented the combined contribution of the independent variables rather than their individual contribution.

BGD

7, C1210–C1217, 2010

As suggested by the reviewer, in the revised manuscript we have included a more detailed description of the benefits of model selection in accounting for multi-collinearity among the data.

What role does error in data play - how can it be incorporated into the maximum likelihood approach.

The equations presented in the paper only explicitly account for error in the dependent variable (i.e. GPP). Error in both dependent and independent variables can be incorporated explicitly in the presented setup. However the equations mentioned in the paper have to be framed in total least squares framework.

In the current setup, errors in the auxiliary (i.e. independent) variables play a similar role as they would in multiple linear regression. If the independent variables have errors, this will weaken the relationship between them and the dependent variable, such that (a) fewer variables will be selected as being significant in explaining the variability observed in the dependent variable, and (b) the regression coefficients (betas) of the selected variables will have higher uncertainty due to the "noise" in the independent variables. Therefore, the results will still be "accurate" in the sense that the uncertainties associated with the statistical model will be correct, but the uncertainties on the regression coefficients and the fraction of the variability not explained by the model will be larger.

To truly make this geostatistical approach applicable to the research community, I suggest the authors consider including pseudo-code or sample scripts as a supplement. While well written, it is not trivial to replicate/implement these equations from scratch.

We completely agree with the reviewer, and have recently started making our data and codes publicly available at http://puorg.engin.umich.edu/code.php. The code for this paper will be posted once the manuscript is accepted in its final form. We have added

BGD

7, C1210–C1217, 2010

a link to this site in the revised manuscript.

Wavelet analysis is one approach that has been used. I believe Dennis Baldocchi also has some earlier papers on Fourier-style analysis of flux tower observations. I suggest digging those up and citing.

We have reviewed these references, and have included a brief description of them in the background section of the revised manuscript.

P 1458, line 15 - The authors mention desirability to reduce the number of candidate models and describes a nice method for doing so. I'm not sure why precisely this was needed for this study, given the number of candidate models and sites.

As the number of independent variables increases, the total number of models resulting from performing all possible regressions (only considering linear additive relationships) is equal to 2p, where *p* is the number of independent variables (For example, for 11 covariates, the total number models possible is equal to 211 = 2048). A considerable number of these 2p models do not have any basis of support on the Jeffrey' s scale (please see page 1456 line 20-27 in the discussion paper), and hence it is desirable to eliminate these models before arriving at a subset of models that can explain the dependent variable. The method presented in the paper is an objective way to achieve this end.

Can you discuss the computational needs/time of this approach that requires this pruning, or is there another reason?

Computational demands for performing pruning of the 2p models to determine the candidate set of models that explain the dependent variable is minimal if the covariance matrix [**Q**] is fixed. In our algorithm, once the covariance matrix [**Q**] is inverted, we can

BGD

7, C1210–C1217, 2010

perform all possible regressions for the 11 examined variables in a couple of minutes, and then arrive at a candidate set of models that explain the dependent variable in seconds.

References

Mueller, K. L., V. Yadav, P. S. Curtis, C. Vogel, and A. M. Michalak (2010) Attributing the Variability of Eddy-Covariance CO2 Flux Measurements across Temporal Scales Using Geostatistical Regression for a Mixed Northern Hardwood Forest, Global Biogeochemical. Cycles, doi:10.1029/2009GB003642, in press, temporarily available at http://www.agu.org/contents/journals/ViewPapersInPress.do?journalCode=GB