Biogeosciences Discuss., 9, C8550–C8564, 2013
www.biogeosciences-discuss.net/9/C8550/2013/

Biogeosciences Discussions

Interactive comment on "Improving terrestrial CO_2 flux diagnosis using spatial structure in land surface model residuals" *by* T. W. Hilton et al.

T. W. Hilton et al.

hilton@unm.edu

Received and published: 27 February 2013

Author responses to Anonymous Referee #2
Referee comments in boldface, author responses in normal typeface.

Overall, this paper addresses an important and difficult question of how to improve our ability to diagnose and predict regional CO2 fluxes. However, there is not enough effort made to tie the specific analyses performed and results of the study back to the ultimate goals of the science. First and foremost, how does specifying a good land surface model residual covariance matrix help us to improve models like VPRM, or improve flux estimates from kriging maps or atmospheric inversions? What's the marginal benefit of these covariance matri-

ces in light of the ultimate goals? Is improving diagnosis and prediction of CO2 fluxes at regional scales a tractable problem given land surface heterogeneity?!! (You should at least discuss/ mention this.)

A land surface model residual covariance matrix is one of a number of pieces of information which, taken together, constrain terrestrial carbon flux estimates. Any regional atmospheric inversion that uses prior flux estimates, for example, makes an assumption about the correlation structure of prior flux errors between regions. This assumption may be explicit (e.g. Rodenbeck et al 2003) or implicit (e.g. Pacala et al. (2001), Peylin et al. (2002), Gurney et al. (2002) and Peters et al. (2005)) , but it is always there.

A model residual covariance length scale also places an upper bound on the size of a region that a particular model may be expected to constrain, thereby suggesting a model pixel size for a regional study.

Both of these uses for model residual covariance structures provide a new piece of unique information to constrain NEE estimates from diagnostic modeling studies.

Our study uses observed NEE to directly estimate the spatial covariance structure for one model, VPRM. To our knowledge this approach is unique in the literature.

Regional land surface heterogeneity certainly makes the flux diagnosis problem more difficult. We believe that in spite of this difficulty, land surface models do have information to contribute to the puzzle. Successes in modeling specific eddy covariance site observed NEE using models with more general parameter estimations (e.g. Kuppel et al, 2012) reinforce this belief. As we see it, one of the most important goals of diagnosing terrestrial carbon dioxide NEE is to further our understanding of land-atmosphere interactions toward the goal of understanding how these interactions might contribute to future climate.

We have added text to section 1 (Introduction) to clarify these points.

In terms of the specific analyses, it is also not clear from the outset what are the hypotheses or overall questions that you're trying to address. Is this study primarily an evaluation of VPRM? Or are these results more widely applicable to other land surface models? What is the goal in terms of trying to figure out how many parameters to estimate? (What are we trying to optimize? Is it the flux maps that you mention in the 2nd to last paragraph in the conclusions?)

The last paragraph of the introduction (p. 7078 lines 3 to 11) addresses this point. We will revise the introduction and abstract to make this more prominent.

You should also give some theoretical introduction as to how we expect the four estimated parameters in VPRM to vary in space and time. What do plot-level studies and physiological understanding tell us?

Mahadevan et al (2008) estimate parameters for 11 different North American eddy covariance towers and validate those estimates at a further 11 sites. Their values differ from ours: for example their light use efficiency (lambda) estimates for cropland sites are much lower than ours. Similarly, Lasslop et al (2010) constrain their LUE estimates to smaller values (0.0 to 0.22) than some of our estimates. Schaeffer et al (2012) note that among LUE-based models the LUE parameter has a larger influence than other GPP parameters in determining model GPP fit to observations. Our empirically-derived LUE parameter estimates are sometimes outside the physically-realistic bounds determined by observations; we chose to allow these LUE parameter values due to its importance to simulating GPP and the relative simplicity of VPRM structure.

We have revised the text to include these points.

Why do you assume that the 3 scaling factors (for water, phenology and temperature) don't vary in space and time?

These three scaling parameters are mathematical formulations of observed quantities, specified by equations in VPRM's structure. They do vary in space and time, although

they are not optimized from data. Mahadevan et al (2008) provide detailed justifications of those choices.

How much model complexity (in terms of # of parameters) is needed in order to improve the ultimate goal of improved CO2 flux model prediction or diagnosis?

This is an important question, and is the focus of ongoing research (e.g. Kuppel et al (2012), Schaeffer et al (2012)). We have developed thoughts on this topic using the VPRM parameterizations presented here; these are described in a separate manuscript to be submitted shortly. In this study we focus on the hypothesis that VPRM residuals are correlated at spatial scales larger than the footprint of a single eddy covariance tower, but smaller than a continent.

How do we evaluate the benefits and costs associated with parameter lumping/ separation?

Further work is needed in this area. An optimization method such as Markov Chain Monte Carlo that provides probability density functions for the parameters will quantitatively determine whether parameter differences are statistically significant.

You mention multiple times that we expect, a priori, that VPRM parameters should vary by plant functional types (PFT's). You need to back this up with more with ecological understanding and references, rather than assume that the reader is familiar with this argument.

We thank reviewer 2 for raising this question. We will expand the text on p. 7086 (lines 11 to 22) to include references to several studies that use PFTs as a classification method for model parameterization.

Some of the writing is overly-abstract and hard to follow. One example starts on p3, line 94 with "Michalak et al. (2004) point out that spatial structure, if existent, contains information that constrains fluxes and suggests weights for fluxes to identify and remove redundant information." I've read this paper multiple times,

but I still don't understand this sentence!

We have revised this text to read "Michalak et al. (2004) point out that spatial structure, if existent, contains information that constrains fluxes. A spatial covariance matrix also identifies regions whose errors are strongly correlated. These correlated areas may then be assigned lesser weights than uncorrelated areas to quantitatively acknowledge that their results are, to some degree, redundant.

Also, there are a few abrupt transitions where the flow of arguament is lost. For example, the paragraph starting at line 35 is quite abrupt, as compared to the previous paragraph. You mention a number of issues associated with poor diagnostic and predictive ability in land surface models. Then, you launch into the data sources available to constrain the problem. Why do carbon cycle scientists struggle with weak performance in land surface models? Is this a tractable problem? Is there reason to believe that more observations will help to solve this problem? How is your approach helping to bring data to bear in order to improve models?

We have added a sentence to the paragraph cited above to ease the transition. With respect to the larger questions raised here, modeled fluxes are not the only information provided by flux models. Residuals contain information as well. A quantitative error covariance matrix makes land surface models more useful as prior flux estimates in atmospheric inversions, for example. This is separate from, for example, making structural improvements to a model to better represent ecology, but also contributes to the community's ability to diagnose fluxes.

How did you choose the 9 different groupings for parameter estimation (lines 288- 290)? Spell out clearly what these are in the text, and your justification for choosing these groupings. Also, talk more generally about the results beyond just the finding that the PFT's don't seem to matter. Does site-level or seasonal (monthly) variability matter at all? Where do you see significant differences?

(Isn't there any way to get uncertainties on your ML parameter estimation?)

Allowing parameters to vary too frequently in space or time risks overfitting the data, while dividing observations too coarsely results in a poorer fit of modeled NEE to observations. There is not a clear consensus in the literature regarding the "optimal" space-time grouping of observations for parameter estimation; indeed, this almost certainly varies according the modeling goal. Our main purpose in this study is to determine whether NEE model-observation residuals covary in space at a length scale more than a few kilometers; demonstrating this is true across several parameterization choices is this sufficient for our purposes here. We have developed some conclusions regarding parameterization choices for model-based diagnostic upscaling of eddy covariance NEE observations, and we plan to submit these as a separate manuscript.

It is possible to estimate uncertainties in model parameters. The Markov Chain Monte Carlo (MCMC) approach, for example, delivers a full parameter probability density function. Because of the substantially increased computational expense and the lack of a statistically robust residual likelihood function to provide to an MCMC estimation we chose the simpler point estimates provided here.

What do you make of the fact that the 9 different groupings give remarkably different sill variances (Fig. 4)? How important is a reduction in sill variance in light of your ultimate goal (to create flux maps from VPRM, improve inversions, etc.)? What's the trade-off in terms of reducing the sill variance (explaining more of the variability in the residuals) and estimating more parameters? Did you try any cross-validation, excluding some of the sites from the parameter estimation? For example, you could test which grouping of parameters best minimize the model-data residuals, with a penalty for the number of estimated parameters. There should be an appropriate statistical test for this.

Comparing the VPRM parameterizations derived in this study is a useful task, but is more important to the problem of upscaling eddy-covariance NEE using VPRM. We

have evaluated a separate set of 27 North American eddy covariance towers for parameter cross validation. We intend to submit those results soon as part of separate manuscript focusing on upscaling tower fluxes. We do not feel that the choice of optimized parameter sets among the nine considered is central to the main question of the analysis reported here, which is to determine whether or not VPRM NEE residuals exhibit spatial covariance at length scales greater than an eddy covariance tower footprint, as the results here suggest spatial correlation is likely present in each of the parameterizations considered here.

Finally, is the 400km length scale for VPRM residuals model-dependent? Would we expect other land-surface model residuals to have similar length scales, or are the spatial scale of model errors dependent on the model itself? If so, what have you learned about VPRM model formulation?

The estimated length scale of any model's residual spatial covariance is a function of both the model structure's capacity to estimate real NEE as well as the capacity of the North American flux tower network to constrain modeled fluxes. A more complex model structure, for example, should remove sources of structural error from the model NEE. Each source of error should have a characteristic length scale, and its removal should therefore change the residual length accordingly.

We therefore expect that the magnitude of residual spatial covariance length scale to vary somewhat between different models.

Specific comments

Make a distinction between running VPRM with grid-scale remote sensing data (wall-to-wall estimation), vs. point-based with site-specific data collected at eddy covariance sites, as in this study. What's the representation error associated with extrapolating from the site-specific to the grid-scale form of estimation? (Does this matter for your analysis or ultimate goal of producing maps?) As a corollary, you should also mention in the introduction that residual correlation

length scales are likely scale-dependent. For example, correlation lengths of CO2 fluxes at a 1km resolution are likely to beshorter than at a 500km resolution. Given the poor diagnostic ability of flux models like VPRM, there is reason to believe that the correlation lengths of model residuals from VPRM are also likely scale-dependent.

We agree that moving from point estimates at specific tower locations to a regional grid-based diagnosis will incur representation error. As Reviewer 2 notes, this issue is more pertinent to the task of grid-based "mapping" diagnoses. We are preparing a second manuscript to be submitted separately that focuses on upscaling tower fluxes along these lines.

This study focuses on the question of whether there is detectable spatial covariance in flux model residuals. That question is most directly applicable to regional atmospheric inversions, where the correlation length scale can provide weights for the priors for different regions. By addressing the question using measured NEE at single points, we avoid the representation error problem here.

Does it matter that the exponential distribution was only selected by AIC in 74 out of 1000 GRF's in the pseudo-data test, and that the median covariance range of these exponential distributions is 936km (almost double the "true" length of 402km)? That seems like a relatively low ability to detect the correct distribution, and estimate the correct correlation lengths. What do you make of the fact that the exponential covariance function was chosen for 92 of 252 of the observed residuals, a much higher proportion than in the pseudo-data test?

We believe that the difference in the detection rate between the exponential covariance structure pseudodata and the actual residuals is consistent with a situation where the true underlying covariance structure is stronger than the median results reported here, but the realization of eddy covariance tower placements available to us (Fig 1) is minimally adequate to detect it some of the time. This is also consistent with our finding that

spurious exponential covariance structures were identified in only 2.5% of pseudodata realizations tested, far below the detection rate of 36% for the real residuals.

How much confidence do you have in the median length scale of 400km, given the large range in your estimates (from 100 to 900km)? In terms of recommending 400km for an estimation scale in inversions, this may still be too large directly around measurement locations, where there is a risk of misinterpreting small wiggles in the data, and attributing local influences to much larger regions.

We believe that our results demonstrate that it is quite likely that the proper length scale for model error covariance is larger than an eddy covariance tower footprint; this implies that atmospheric inversions should revisit the assumption that regional prior flux estimate errors are independent and identically distributed (i.i.d.); that is, that the covariance length scale is zero.

In the review of inversions and correlation length scales, please also mention geostatistical inversion techniques that don't use priors (on page 3): Michalak et al., 2004; Gourdji et al, 2008, 2008, 2012 & Mueller et al, 2008. These are terrestrial flux papers, which may be more relevant than an ocean inversion (the Jacobson et al., 2007a,b references). In these approaches, where fluxes are estimated at fine scales without prior flux estimates from a land surface model, the flux covariance structure is critical.

We agree that these references are relevant to our work, as they make use of the correlation structures we seek to quantify. We have added discussion of Michalak (2004) and the work of Mueller and Gourdji to the introduction.

Also, mention the use of RML for estimating correlation length scales (of fluxes or flux residuals) from atmospheric CO2 data in an inversion framework (Michalak et al, 2005; Gourdji et al., 2010, 2012) in the last paragraph on p. 3.

We have added a brief discussion of this work. It is an interesting comparison, as

Michalak (2005) uses atmospheric CO_2 concentration measurements to estimate prior flux error covariances, whereas our work presented here uses surface flux observations.

Finally, these papers are relevant for your discussion of the optimal estimation scale for inversions (lines 477 to 484), as they estimate at fine scales and post-aggregate fluxes to larger, more meaningful scales. Concerns about aggregation error should be acknowledged (Kaminski et al., 2001; Engelen et al., 2002; Schuh et al., 2009) as a concern when choosing the appropriate estimation scale for inversions.

We will added a brief discussion of aggregation errors to the text in lines 25 to 30 on page 7093 which discusses constraints on choices of regions or atmospheric inversions.

As a minor note: how is Pscale different from EVI? Doesn't EVI get at phenology? This should be folded into a few more sentences on the scaling parameters, and how they are derived, and why we don't expect them to vary by site or season.

Mahadevan et al (2008; see paragraph 13) discuss in detail the structure of VPRM. Pscale includes in VPRM the effects of leaf expansion and senescence on canopy-scale photosynthesis. We have revised the sentence describing Pscale to note this.

The paragraph beginning at line 369, introducing the analysis of inter-annual variability, is very confusing! First, you should introduce the concept of inter-annual flux and residual variability, and why we expect this to vary at larger spatial scales than NEE itself. Then, you can move on to how you calculated anomalies from the long-term mean of both NEE (VPRM & observed), and model residuals in order to investigate this question. The ordering in that paragraph currently makes it hard for the reader to follow. Also, is VPRM structured well enough to get at drivers of inter-annual variability (e.g. El Nino/ La Nina, disturbances, extreme climate events, volcanoes, etc.)? If not, is this analysis mean-

ingful at all?

We thank reviewer 2 for pointing out this lack of clarity. We have expanded the paragraph to note that NEE interannual variability is largely controlled by large-scale, climatic drivers such as temperature and moisture availability (Desai et. al, 2008; Ricciuto et al, 2008; Law et al (2002)).

Analyses prepared for a separate manuscript to be submitted shortly show that annually integrated VPRM residuals follow roughly a normal distribution, with a mean of 1.6 gC m^{-2} yr^{-1} and a standard deviation of 168.5 gC m^{-2} yr^{-1}. This mean annual residual is small relative to the 100 to 300 gC m^{-2} yr^{-1} sink typical of a productive North American ecosystem. These results suggest that VPRM performs well at the annual scale. This seems reasonable given that VPRM is driven largely by climate-driven quantities such as temperature, EVI, and radiation.

Finally, the outcome of the investigation into length scales of inter-annual variability is unclear (paragraph beginning with line 377). The main result seems to be: "Of the seven years examined, NEEobs anomalies show correlation at scales of roughly 1000 km only for 2006." I would add another phrase, saying that the AIC chose a nugget distribution for all other years, indicating no spatial structure at all, and that this result with the observations is roughly consistent with that from VPRM estimates. The reader shouldn't have to go to Fig. 6 to draw this conclusion for him or herself.

We have revised the text to note this.

Also, in the following paragraph you discuss how the residuals seem to have more spatial structure than NEE estimates themselves. I would add an aside that this occurs mainly with more lumped parameters, e.g. annual, but not monthly, and all sites as compared to individual sites or PFT's.

We have added text describing this connection to the parameterizations.

In the Caveats section (4.1), you mention that VPRM doesn't capture long-term drivers of NEE well, e.g. disturbance and carbon pools. Then, you conclude the paragraph with "These simplifications caution us against attempting detailed ecological interpretation of the VPRM NEE results and VPRM residuals." Does this invalidate the whole study then?!! If you can't learn anything about nature from your analysis, what have you learned that helps to advance the science and the ultimate goals of the science?

Thank you for pointing out that this sentence is unclear.

We have expanded the sentence in question to read "VPRM is able to credibly partition the contributions of photosynthesis and respiration to observed NEE (Mahadevan 2008). However, these simplifications caution us against attempting site-specific ecological interpretation of short-term fluctuations in VPRM parameter values, fluxes, and residuals."

You mention that disturbance history is likely a driver of carbon cycle dynamics and land surface heterogeneity (lines 448 to 458). Can VPRM indirectly capture this at all? For example, this likely affects EVI, and perhaps the estimated parameters or the scaling factors, which might vary by stand age?

The question of how site disturbance history impacts VPRM's parameters and scaling terms is deserving of further investigation. When parameters are estimated for specific sites by optimizing model NEE to observed NEE, the parameter values themselves must contain some information about all of the drivers of NEE, including disturbance history. As Reviewer 2 notes, EVI is probably impacted also (certainly in the short term). Plant Functional Type, though not included directly in VPRM structure, is also determined in part by species succession following a disturbance. Thus, VPRM NEE diagnoses and estimated parameters must contain some information about disturbance history. That said, through the lens of VPRM that information is convoluted with other drivers of NEE. This makes it difficult to assess disturbance directly through

VPRM.

We have added a sentence to the conclusions to note this.

The last sentence in the conclusions should tie this paper and analysis back to the ultimate goal of the science. E.g. "With an accurate land surface model residual covariance matrix, CO2 flux estimates from inversions can be improved, for the ultimate goal of diagnosing regional CO2 fluxes and improving predictive skill in land surface models. . .". This is kind of rough, but for someone who only reads the introduction and conclusions, they should come away with why your paper is useful, and how it fits into the larger goals of the science.

Thank you for this suggestion. We have included this text:

A quantitative land surface model error covariance matrix can help to improve atmospheric inversion-derived ecosystem-atmosphere CO2 flux estimates as well as estimate accompanying uncertainties more accurately. This, in turn, can help improve mechanistic understanding of the terrestrial carbon cycle, furthering the goal of increasing the predictive skill of land surface models.

Technical corrections

In the abstract, you say that spatial structure exists in data-model residuals at a length scale of âĹij1000km. The body of the manuscript says 400km. Correct or clarify.

The abstract has been corrected to 400 km.

Think of a better way to rephrase the last sentence in the abstract. What do you mean by "carbon cycle participation"? Maybe say that VPRM model parameters do not differentiate by plant functional type?

We have reworded the sentence to read "This calls into question whether PFTs can successfully partition ecosystems' fundamental ecological drivers when the viewing

lens is a simple model."

First sentence in introduction: update emissions from fossil fuel combustion, which are now around 9 PgC/yr in 2010 (http://www.globalcarbonproject.org/carbonbudget/10/hl-compact.htm).

We have updated the fossil fuel emissions estimate.

Lines 141 to 143: I assume that temperature and PAR are collected at the flux tower sites? Don't these count as meteorological driver data? (Clarify the statement "It can thus be run globally, with no need to compile temporally-filled meteorological driver data.")

We have added a clarifying sentence: "Temperature and PAR can come from site-level observations when modeling a point or, if the model is to be run globally, from gridded reanalysis products."

*** Lines 163-167: Please explain why site phenology, land surface water, and land surface cover type are needed in terms of the VPRM model formulation (eqs. 1 & 2). In general, after introducing the equations for VPRM (after line 137), specify which are the 4 estimated parameters, and how you get the 3 scaling factors.**

We have revised the text to refer to lambda, PAR0, alpha, and beta as "user-supplied parameters", not "model parameters". We have also revised the descriptions of Pscale, Wscale, and Tscale to specify them mathematically.

Site phenology and land surface water determine Pscale and Wscale, respectively. Including the mathematical specifications of the scaling terms thus clarifies the role of these remote sensing products.

*** Line 381: shouldn't this be 74 of 1000 attempts?**

This has been corrected to 74 of 1000 attempts.

*** Add a reference to the statement on lines 445-446, starting with "Second, PFTs**

are commonly assumed. . .".

We have added references for this statement.

REFERENCES

Desai, A. R., B. R. Helliker, P. R. Moorcroft, A. E. Andrews, and J. A. Berry (2010), Climatic controls of interannual variability in regional carbon fluxes from top-down and bottom-up perspectives, J. Geophys. Res., 115, G02011, doi:10.1029/2009JG001122.

Kuppel, S., Peylin, P., Chevallier, F., Bacour, C., Maignan, F., and Richardson, A. D.: Constraining a global ecosystem model with multi-site eddy-covariance data, Biogeosciences, 9, 3757-3776, doi:10.5194/bg-9-3757-2012, 2012.

Law, B. E., et al. (2002), Environmental controls over carbon dioxide and water vapor exchange of terrestrial vegetation, Agric. For. Meteorol., 113, 97–120, doi:10.1016/S0168-1923(02)00104-1.

Ricciuto, D. M., M. P. Butler, K. J. Davis, B. D. Cook, P. S. Bakwin, A. Andrews, and R. M. Teclaw (2008), Causes of interannual variabil- ity in ecosystem–atmosphere CO2 exchange in a northern Wisconsin forest using a Bayesian model calibration, Agric. For. Meteorol., 148(2), 309–327, doi:10.1016/j.agrformet.2007.08.007.

Schaefer, K., et al. (2012), A model-data comparison of gross primary productivity: Results from the North American Carbon Program site synthesis, J. Geophys. Res., 117, G03010, doi:10.1029/2012JG001960.