Dear Dr. Bahn,

We have carefully considered the comments from both referees (in plain text) and respond to all comments in a point-by-point manner in italics. We additionally made a number of minor changes to improve readability, to correct minor errors, and to improve the display of the wavelet half planes in many figures.

Referee #1:

This article evaluates twenty ecosystem models against measurements of the NEE from the ten eddy covariance research sites provided by the NACP program. The authors found that:

1) Models with prescribed phenology often fit NEE observations better on annual to interannual times scales in short vegetation

2) Models that calculate NEE as GPP-ER showed better fit on monthly to seasonal time scales in two coniferous forests

3) Models that incorporated foliar nitrogen data were successful at capturing NEE variability on multiple year time scales at Howland Forest, Maine

4) Few NACP models correctly predicted fluxes on seasonal and interannual time scales.

I believe that the contents in this article are meaningful to the community and will be of interest to audience of "Biogeosciences". I also enjoyed reading this article and the ideas in this article. I hope that several issues below make much better readability for its contribution to the community.

We thank the referee for support of the manuscript and for the insightful comments, which greatly improved the manuscript.

1. Wavelet coherence analysis needs continuous data without any data missing. This indicates that the results of wavelet coherence and their interpretation may (or may not) be sensitivity to the gap filling methods. First, if the authors applied the gap-filling of NEE based on Barr et al. (2004), daytime gaps in NEE were estimated by GPP minus ER and GPP and ER estimated independently from the light-response curve and temperature dependency of ER. The authors concluded that the models calculating NEE as GPP-ER were superior only on monthly to seasonal time scales in two coniferous forests, and I want to know this result depends on the gap filling methods of the observed NEE (e.g., look-up table of NEE). Second is related to gap filling of climatological conditions. As the article showed clearly, the models showed different responses to the climate conditions and the climatological conditions have been used for driving the model. Therefore, I wonder if the gap filling method can result in some uncertainty in interpreting the variations among the models. So I just want to know if the authors can test these sensitivity issues by comparing wavelet coherence results of different gap-filling methods.

Here, we explore the relationships among time series that are modeled against time series that are both measured and gapfilled. The gapfilling approach used here is empirical, data-driven, and independent of all ecosystem models. Previous studies have shown that the orthonormal wavelet coefficients of measured NEE time series from multiple ecosystems closely matched the orthonormal wavelet coefficients of NEE time series from multiple gapfilling models (Figure 4 in Stoy et al., 2006)*, except at the sub-daily time scales that we do not analyze in the present manuscript and that are prone to variability in the flux footprint. We now cite the findings of Stoy et al. (2006) more explicitly when discussing the analytical approach.*

Regarding the specific point that the models that use NEE = GPP-ER tend to be superior only on monthly to seasonal time scales, we argue that this finding further suggests that the gapfilling approach is independent to the models. If our result was dependent upon the gapfilling method, *one might expect that significant wavelet coherence would also be observed at diurnal, weekly, and annual/interannual frequencies. From a practical standpoint, multiple gapfilled observational products were not made available for the NACP Site-Level Interim Synthesis, but were explored by Stoy et al. (2006) as noted. Multiple gapfilled NEE estimates will be available in future versions of the FLUXNET database. Exploring the impacts of multiple gapfilling models in the spectral domain across the entire flux database may make for an interesting future analysis.*

Regarding the second point, gapfilling the climatological conditions certainly adds uncertainty in the time and spectral domains, but gaps in the meteorological drivers are infrequent compared to the eddy covariance observations. Understanding uncertainty due to meteorological drivers is an important part of the NACP Site-Level Interim Synthesis exercise, and was studied in detail in previous analyses (Ricciuto et al., 2009). *From personal experience, gaps in meteorological drivers are easy to fill because the relationships with nearby sensors are usually linear. Properly gapfilled meteorological time series are not likely to introduce anomalous patterns in the frequency domain apart from the potential for diurnal hysteresis, but again sub-daily periodicities are not emphasized in the present analysis. Bias error in meteorological observations is a more serious concern* (see e.g. Ricciuto et al., 2009)*, but is beyond the scope of this analysis.*

2. In Abstract, the authors wrote that "‥, multi-year oscillations in climatological drivers, and…are known to be important for determining ecosystem function." I fully agree with this opinion and so I wonder if the authors can do wavelet coherence of climatological drivers and peaks of their wavelet coherence can be related to the peaks in the figures. Especially, if there are peaks in the wavelet coherence of the modeled or observed NEE that are not seen in the model outputs, it tells that the modeled or observed NEE are sensitive to small changes in climate conditions, thus indicating that biology may be important.

Wavelet half-planes of wavelet coefficients can be created for climatological drivers, but two time series are required to calculate the wavelet coherence. Orthonormal wavelet transforms of climatological drivers are displayed for example in Stoy et al. (2005) *and similar manuscripts. These analyses show that spectral energy in climatological drivers tends to reside at seasonal and annual frequencies, except for the case of precipitation, which displays spectra similar to white noise. The median orthonormal wavelet spectra of all climatological observations in the FLUXNET database are shown in Figure 4 of* (Stoy et al., 2009). *In light of comment 4, we now discuss these existing findings briefly to better set the stage for the interpretation of the wavelet coherence results.*

3. NEE is the composition of two independent processes, GPP and ER. I wonder if separate analysis of wavelet coherence of GPP and ER can provide more insight on the model performances.

This is an interesting point. The wavelet coherence between GPP or ER and climatological drivers could be calculated. A challenge arises however because GPP is entirely modeled using the flux partitioning approach, and ER is mostly modeled, except during nighttime periods with sufficient turbulent exchange such that ER can be measured. The (orthonormal) wavelet spectra of ER time series also tend to match observations poorly (see Figure 4 in Stoy et al., 2006). NEE on the other hand is mostly observed, sometimes modeled, and can be gapfilled with little alteration in the wavelet frequency domain as noted. By analyzing NEE gapfilled using a validated algorithm, we can compare a time series with information mostly from observations against time series that are entirely modeled.

4. As the authors pointed out in Introduction (from line 18 page 2044), a few papers reported the NACP model performances regarding phenology, better fit from the model calculating NEE from GPP minus ER and etc. These findings are also emphasized again in this article and it needs some clear description on which more can be obtained from wavelet coherence. In particular, the authors said that "less certain is how models match measurements on multiple time scales as they respond to climatic and biological forcings that act on multiple time scales". Therefore, I think that analysis in time scales of climatic forcings must be discussed before moving to wavelet coherence of the models.

We feel that it is important to ground our observations in the context of existing analyses. The additional benefit of wavelet coherence is its ability to identify the times and time scales at which these previous findings hold. For example, the finding that 'models need improved phenology' certainly holds for the models of the NACP site-level interim synthesis. But blanket statements like this, regardless of their accuracy, may omit years, ecosystems, or models for which the variability of NEE due to phenology is simulated well. Further analyses may even indicate that phenology is in these situations simulated correctly for the correct reasons. By learning about the patterns of model-measurement mismatch across time and time scale, we can (hopefully) improve model fidelity.

5. Figure 1: I wonder what was happing in the early measurement period (1991 and 2004) and described as blue color.

The blue spaces correspond to periods when measurements were not available. From this figure and others it should be clear that periods of missing data have diminishing effect at lower frequencies, depending on the size of the gap. The explanation for the periods with zero wavelet coherence (essentially the coherence between a variable model and an invariable gap) is now expanded upon in the text. For the case of the Harvard Forest time series, the impacts of missing data early in the measurement record are minimal, and we do not focus on these periods in the discussion of results.

6. P3050 / L10 – 19: I want some discussion on the performance of ED2 in Figure 3 compared to the others

Figure 3 presents significant wavelet coherence half planes for four models against the Harvard Forest time series. We chose to highlight the results of the LoTEC model in the text rather than ED2 because of the unique data assimilation approach employed by LoTEC that results in significant coherence about a much larger region of the wavelet half plane. ED2 tends to perform poorly on diurnal to monthly time scales when confronted with observations from US-Ha1, and focusing on the ED2 would amount in our mind to unnecessary criticism of this model when other models are likewise simulating patterns at these time scales poorly. It is for this reason, subjecting models to undue criticism, that we interpret model attributes, rather than individual models (with the exception of LoTEC), as described in Table 2.

7. P3052 / L5: $10^{3.5}$ (i.e. 3 months) may be revised to $10^{3.3}$ (i.e. 3 months) or $10^{3.5}$ (i.e. 4 months)

The error in the manuscript has been corrected to read '$10^{3.5}$ (i.e. 4 months)', thank you for the careful read.

8. P3052 / L21 – 27: Please indicate which figure supports these sentences.

'Figures 5 and 6' is stated on line 25.

9. P3052 / L27 – P3053 / L2: The authors said that "Remote sensing is often unsuccessful for capturing grassland phenology (Reed et al., 1994), due in part to the fact that the shift from green to brown biomass is critical for modeling NEE but can be subtle and difficult to ascertain remotely (Sus et al., 2010)" It seems to me that this sentence is saying that the prescribed phenology is not good for modeling in grassland. However, the authors found that the models that prescribed phenology showed better performance at the cold, non-forest sites. I am curious about consistence between these two sentences in the article.

We decided to re-word the passage as it gave a cursory overview of a challenging issue described in more detail by (Richardson et al., 2012) *and related references.*

10. I want to know the exact meaning of white-colored area in Figs 5-9. If I correctly understand, this white-colored area can be when two types of models are either significant or not significant simultaneously. Sometimes, it seems important to distinguish these two different conditions on the model performance.

The referee is correct. White spaces refer to regions in time/time scale in which at least 50% of models with each model structure are adjudged to be significantly related to observations, or neither type of model structure is significantly related to observations. In other words, regions where there is no reason to believe that one model structure is superior to another. We agree that it can be important to note both cases separately, but we decided upon the present approach to reduce the dimensionality of the discussion.

11. It seems to me that Section 3.8 did not show any important results but discuss some issues in interpreting wavelet coherence. So, I suggest this subsection is moved to conclusion, rather than the independent subsection in Results and Discussion.

The major purpose of Section 3.8 is to note similar publications in order to place our findings in the broader context of spectral analyses of flux models and measurements. We feel that the subsection is mis-labeled instead of misplaced as it is not a conclusion of our results but rather a discussion of similar studies. We now call the subsection 'Multi-scale analysis of measurements and models' We also briefly expanded this section in light of new publications by van Gorsel et al. (2013) *and Vargas et al.* (2013).

12. It seems to me that paragraphs in Conclusion are saying what is not discussed in Results and Discussion. In particular, because the data assimilation is discussed in Appendix A, any comments on LoTEC are not relevant in Conclusion. I suggest that Appendix A is moved to Results and Discussion. Also, I could not find any evidence to support the paragraph from Line 2 to 10 in Page 3056.

We agree with the referee and have moved Appendix A into the main body of the text, also in accordance with the Biogeosciences *preferences. The paragraph from Line 2 to 10 discussed findings of the data assimilation comparison, which is now in the main body of the text.*

13. There is one previous study to use wavelet coherence for evaluating ecosystem model and to discuss results from wavelet coherence with phenology (Hong and Kim, 2011, Impact of the Asian monsoon climate on ecosystem carbon and water exchanges: A wavelet analysis and its ecosystem modeling implication, Glob Chan. Biol, 17, 1900-1916). Proper citation of this study should be done.

We thank the referee for pointing out this study. We now cite Hong and Kim (2011) in the discussion section.

14. I want any explanations in the spacious white-colored area in Figs 6G, 6I, 7G, 7I, 8G, 8I, 9G, and 9I, if possible.

The (spacious) white-colored areas refer to areas of the measured time series where observations were not available. We excluded the analysis of these periods in the revised manuscript and revised the figures. The time axis has been changed, and as a consequence we now analyze different time periods that have different model/measurement frequency characteristics.

15. The title of Detto et al. (2012) in Reference is wrong. Correct title is Detto et al. (2012) Causality and persistence in ecological systems: A nonparameteric spectral Granger causality approach, Am. Nat., 179, 524-535.

The title cited referred to a previous version of Detto et al., and has been corrected. Thank you for the careful read.

Referee #2

The paper of Stoy and colleagues with considerable interest, as it aims at providing a new tool for an important aspect of earth system sciences: the evaluation of model performance with the objective to understand structural deficiencies in models. Overall the paper is well written, the scope of the study clearly stated and the results are thoroughly discussed.

Methodological Issues:

As the paper does propose a new methodology for diagnostic model evaluation it is crucial that all techniques are employed in a rigorous manner, with a strong grounding in the relevant methodological literature. At this point I am sorry to say that the paper suffers from severe methodological issues (listed below) that prevent me from recommending it for publication in its present form.

We thank the referee for pointing out the Maraun & Kurths (2004) and Maraun et al. (2007) manuscripts cited below. A careful read of Maraun & Kurths (2004) revealed that their recommendations agree with our analysis. A major finding of Maraun and Kurths (2004) is that wavelet coherence (called coherency in their manuscript), rather than the wavelet cross spectrum, should be employed in significance testing. The reason for this recommendation is that spurious peaks in the wavelet cross spectrum may result if one, but not necessarily both, of the time series under investigation exhibits high power at a given time and time scale such that erroneous significance determinations may be made. Because wavelet coherence normalizes the wavelet cross spectrum by each wavelet power spectrum, such spiky behavior is avoided. We now cite Maraun and Kurths (2004) to further justify the use of wavelet coherence for our application.

The major findings of Maraun et al. (2007) is that pointwise and areawise significance testing in the wavelet domain may give false positive results that result from autocorrelation in time or frequency when interpreting autoregressive models versus *white noise. They recommend a series of tests on individual points in the wavelet half-plane, areas of significant coherence, and bootstrap-based methods against random coherence to ensure that coherence is not overestimated for each region.*

There are a number of reasons for which we feel that the additional tests recommended by Maraun et al. (2007) do not add value to the present analysis. First and foremost, Maraun et al. (2007) are investigating a different case by studying white noise time series that should rarely exhibit coherence with autoregressive time series at the 95% confidence level. We investigate models and measurements that one would hope would share common power at all scales and frequencies. In other words, regions in time and time scale where coherence between models and measurements are not observed are more interesting than regions where common power is observed, because models are more interesting when they fail. The emphasis on model-measurement mis-fit (i.e. model-data disagreement) is stated as early as line 9 in the abstract which reads 'Wavelet coherence can quantify the times and frequencies at which models and measurements are significantly different. From this standpoint, regions in time and time scale where coherence may be overestimated does not represent a false positive coherence, rather a false negative lack of coherence, and therefore a more conservative interpretation of times and scales at which models fail. When re-reading the manuscript, we noted that a number of passages did not clearly state that model mis-fit is more interesting and important than correct model fit. We comprehensively revised the manuscript accordingly.

A second important difference between Maraun et al. (2007) and our study is that we are averaging multiple results from multiple wavelet half-planes. Isolated incidents of (arguably) erroneously large regions of significant coherence are likely to be averaged out. We further note that regions of false positive coherence are anticipated some 5% of the time such that overinterpretation of regions that may not show significant coherence overstep the stated bounds of the significance test.

Application of wavelet analysis to time series with missing values

This is a recurrent feature for several analyzed time series (see e.g. "constant" areas in Figures 1, 6, 7, 8, 9). I assume that the corresponding time steps have been filled with constant values (e.g. zero). This leads to spurious patterns in wavelet coherence.

For high frequencies (small scales) these artifacts are easily visible and the authors do avoid interpretation. However, for low frequencies these artifacts are more difficult to separate from real effects, rendering interpretation difficult. For more information see the arguments on "the cone of influence" in the relevant literature. Thus, the authors must either avoid the analysis of time series with missing values or mask the areas where wavelet coherence is influenced by such effects.

Referee #1 made a similar point, and we agree with both referees that the analysis should be performed differently in this case. The revised manuscript avoids periods when observations are not available for interpretation in the NACP site-level interim synthesis and the discussion of results has been revised accordingly. The importance of the cone of influence was stated on page 3047. From the original figures in question, missing data have a diminishing impact at lower frequencies. Regardless, they have now been removed from the interpretation.

Significance testing

The significance-testing of wavelet coherence in this study does closely follow the suggestion of Grinsted et al. (2004). (Based on simulation of stochastic processes and subsequent application of wavelet-coherence). Unfortunately recent methodological research (Maraun & Kurths, 2004 and

Maraun et al., 2007) has demonstrated that this approach is prone to discover “large significant areas” even if pure noise is analyzed.

(For an impressive illustration of this effect see Figure 8 in Maraun et al., 2007). As the presented study relies heavily on the identification of significant locations in the wavelet space, I cannot recommend this study for publication without this issue being resolved.

(For possible approaches to this issue see Maraun et al., 2007). Note also, it is not sufficient to state that only “large” areas are considered for interpretation, this renders a formal significant test arbitrary, making it essentially useless (as it then depends on the investigators will and not on transparent rules).

The dissimilarity between Maraun et al. (2007) and the present analysis was detailed above. Regarding the approach to discuss only ‘large’ areas of significant wavelet coherence, we note that each comparison includes information from some 200 wavelet half planes (roughly the sites in table 1 multiplied by the models in table 2), and each half-plane includes over a million points in time/time scale space. Delving into the details of each model/measurement combination at each time and time scale would not lend itself to a succinct analysis. As discussed in more detail below, we never seek to interpret the precise dimensions of any region adjudged to be significant, in part to interpret our results conservatively. We also note some number of regions are adjudged to be significant by chance as a consequence of the uncertainty bounds determined by the Monte Carlo analysis, and testing each region somewhat reduces the purpose of stating clear uncertainty bounds.

General comment on wavelet techniques in this context
Overall I am not sure if wavelet-coherence is the best available too to solve the objective of the study (model evaluation with respect to distinct frequency bands). The main advantage of wavelet approaches compared to conventional spectral analysis is the ability to track changes in spectral properties over time. However, the authors do mainly interpret coherence between observations and models for distinct spectral bands (e.g. seasonal, annual), without considering time varying phenomena. Therefore I wonder whether the same findings (identifying in which frequency band models are (not) close to observations) could have been obtained with (conventional) cross-spectral analysis. In comparison to wavelet-coherence cross spectral analysis has a more robust theoretical grounding, avoiding many of the mentioned issues related to significance testing.

Thus I suggest to the authors to reconsider their choice of methods, possibly replacing wavelet-coherence with conventional cross-spectral (cross spectra, coherence) methods. This would likely yield more stable results and make the analysis more transparent. If the results presented in the manuscript are robust, a re-evaluation should not change the findings. This would allow for a change of methods with minimal writing effort.

Regarding conventional cross-spectral analyses, the recommendations of Maraun and Kurths (2004) are clear. Wavelet coherence, not the wavelet cross spectrum, should be used to explore significance testing across times and time scales. Resorting to conventional methods limits the analysis by excluding the possibility to interpret models and measurements at different times and scales in time. From a novelty standpoint, multiple other manuscripts have used wavelet analysis to interpret only the time scale domain, and not the time domain (Dietze et al., 2011; Siqueira et al., 2006; Stoy et al., 2005, 2006a, 2009). *We do not focus on particular times; rather, groupings in time like the growing season or the leaf on/leaf off transition periods.*

We note that other techniques to interpret patterns in multiple and/or gappy time series, including the Lomb-Scargle periodigram tend to only investigate the frequency dynamics, rather than the time variability, of the data series. In other words, wavelet coherence offers a richer view of model/measurement mismatch than conventional methods, which as we note are all-too-often not instructive and note that models are 'incorrect', which is a truism. Despite the inherent complications of interpreting eddy covariance time series, we argue that a cautious interpretation of wavelet coherence offers one of the best opportunities for modelers to critique model output in the time and time scale domains. Our analysis was implicitly cautious in the previous version of the manuscript; at no time was a particular time or time scale (beyond the broad classification of 'weeks', 'months', 'years', etc. with the exception of an anomalously warm year in Canada) discussed, in part because of inherent uncertainty. We now state explicitly the basis for the cautious interpretation (Maraun et al., 2007) and suggest that future analyses likewise not over-analyze significant wavelet coherence patterns.

References

Dietze, M. C., Vargas, R., Richardson, A. D., Stoy, P. C., Barr, A. G., Anderson, R. S., Arain, M. A., Baker, I. T., Black, T. A., Chen, J. M., Ciais, P., Flanagan, L. B., Gough, C. M., Grant, R. F., Hollinger, D., Izaurralde, R. C., Kucharik, C. J., Lafleur, P., Liu, S., Lokupitiya, E., Luo, Y., Munger, J. W., Peng, C., Poulter, B., Price, D. T., Ricciuto, D. M., Riley, W. J., Sahoo, A. K., Schaefer, K., Suyker, A. E., Tian, H., Tonitto, C., Verbeeck, H., Verma, S. B., Wang, W. and Weng, E.: Characterizing the performance of ecosystem models across time scales: A spectral analysis of the North American Carbon Program site-level synthesis, Journal of Geophysical Research, 116(G4), doi:10.1029/2011JG001661, 2011.

Van Gorsel, E., Berni, J. A. J., Briggs, P., Cabello-Leblic, A., Chasmer, L., Cleugh, H. A., Hacker, J., Hantson, S., Haverd, V., Hughes, D., Hopkinson, C., Keith, H., Kljun, N., Leuning, R., Yebra, M. and Zegelin, S.: Primary and secondary effects of climate variability on net ecosystem carbon exchange in an evergreen Eucalyptus forest, Agricultural and Forest Meteorology, in press, -, doi:10.1016/j.agrformet.2013.04.027, 2013.

Ricciuto, D. M., Thornton, P. E., Schaefer, K., Cook, R. B. and Davis, K. J.: How uncertainty in gap-filled meteorological input forcing at eddy covariance sites impacts modeled carbon and energy flux, in AGU Fall Meeting Abstracts, vol. 1, p. 3., 2009.

Richardson, A. D., Anderson, R. S., Arain, M. A., Barr, A. G., Bohrer, G., Chen, G., Chen, J. M., Ciais, P., Davis, K. J. and Desai, A. R.: Terrestrial biosphere models need better representation of vegetation phenology: results from the North American Carbon Program Site Synthesis, Global Change Biology, 18(2), 566–584, 2012.

Siqueira, M. B., Katul, G. G., Sampson, D. a., Stoy, P. C., Juang, J.-Y., Mccarthy, H. R. and Oren, R.: Multiscale model intercomparisons of CO2 and H2O exchange rates in a maturing southeastern US pine forest, Global Change Biology, 12(7), 1189–1207, doi:10.1111/j.1365-2486.2006.01158.x, 2006.

Stoy, P. C., Katul, G. G., Siqueira, M. B. S., Juang, J.-Y., McCarthy, H. R., Kim, H.-S., Oishi, A. C. and Oren, R.: Variability in net ecosystem exchange from hourly to inter-annual time scales at adjacent pine and hardwood forests: a wavelet analysis, Tree Physiology, 25, 887–902, 2005.

Stoy, P. C., Katul, G. G., Siqueira, M. B. S., Juang, J.-Y., Novick, K. A. and Oren, R.: An evaluation of methods for partitioning eddy covariance-measured net ecosystem exchange into photosynthesis and respiration, Agricultural and Forest Meteorology, 141, 2–18, 2006a.

Stoy, P. C., Katul, G. G., Siqueira, M. B. S., Juang, J.-Y., Novick, K. a., Uebelherr, J. M. and Oren, R.: An evaluation of models for partitioning eddy covariance-measured net ecosystem exchange into photosynthesis and respiration, Agricultural and Forest Meteorology, 141(1), 2–18, doi:10.1016/j.agrformet.2006.09.001, 2006b.

Stoy, P. C., Richardson, A. D., Baldocchi, D. D., Katul, G. G., Stanovick, J., Mahecha, M. D., Reichstein, M., Detto, M., Law, B. E., Wohlfahrt, G., Arriga, N., Campos, J., McCaughey, J. H., Montagnani, L., Paw U, K. T., Sevanto, S. and Williams, M.: Biosphere-atmosphere exchange of CO_2 in relation to climate: a cross-biome analysis across multiple time scales, Biogeosciences, 6, 2297–2312, 2009.