Referee 1

We thank the referee for the comments on our manuscript, which helped improving our study. We hope that our answers and the modifications are satisfactory.

Page 2, Line 9: I don't understand "substitute observed by modeled fluxes". Substitute modeled fluxes for observed fluxes?

We reformulated: "In addition we create synthetic observations using modeled fluxes instead of the observed ones, to explore the potential to infer prior uncertainties from model-model residuals".

Page 2, Line 11: Was the random error added to observed or modeled "tower" fluxes? If this was added to the observed fluxes, why? There is already random error in the measurements.

We clarified: " a random measurement noise was added to the modeled tower fluxes".

Page 2, Lines 14-15: "This difference… " isn't clear. Do the large biases exist with respect to 5PM, or the other models? And how does a large bias cause a long temporal autocorrelation?

We reformulated: "This difference is caused by a few sites with large biases between the data and the 5PM model."

Regarding the second question: we also computed the temporal autocorrelation time excluding those sites with a large model-data mismatch bias (see section 5.1, page 9408), and this was found to be less than half of the temporal autocorrelation time using all sites.

Note that unfortunately the abstract contained the wrong numbers which were inconsistent with those in table 2 and in the text. We corrected this, and page 9394 lines 13,14 now reads: 30 and 70 days.

Page 3, line 10. I don't understand the term "regularized." Can this be defined? It is used in more than one place, and I don't recognize what concept is being communicated.

"Regularization" is a standard term in statistics and refers to a process of introducing additional constraints in order to solve an ill-posed problem. See for instance:

Hansen PC, Oleary DP, The use of the L-curve in the regularization of discrete ill-posed problems. SIAM journal of scientific computing, 14 (6) 1487-1503, 1993.

We added: "In this way, the solution of the otherwise ill-posed problem is regularized in the sense that the optimization problem becomes one with a unique solution."

Page 4, lines 4-6. Coarser scale inversions may not explicitly utilize correlation lengths, but they are implicitly imposing a large correlation length (which may be entirely inappropriate). Please clarify.

We clarified: "…(Houweling et al., 2004; Rödenbeck et al., 2003b). For the former case large correlation scales are implicitly assumed since fluxes within a grid-cell are fully correlated. For regional…".

Page 4, lines 13-14. I don't understand this description. How can you derive a spatial correlation in the prior flux error from a coarse resolution inversion?

We reformulated: "In some regional studies the same correlations are used as in large scale inversions…".

Page 4, lines 22-23. This is not correct. The pattern of fluxes was not used to evaluate the spatial correlation length. Lauvaux et al, 2012 tested the spatial correlation length scale by cross-validation of the posterior CO2 mole fractions. CO2 observations were reserved from the inversion, and the correlation length that provided the best fit to the reserved CO2 data was identified as the best choice.

We clarified: "…based on cross-validation of the simulated against observed CO2 mole fractions. The simulated mole fractions were derived using the influence functions and the inverted fluxes".

Page 5, lines 1-4. This sentence is unintelligible.

We clarified: "A recent study by Broquet et al. (2013) obtained good agreements between the statistical uncertainties as derived from the inversion system and the actual misfits calculated by comparing the posterior fluxes to local flux measurements at the European and 1-month scale".

Page 4, lines 12-13. Make it clear that flux measurement sensitivities are areas, not lengths. You are describing dimensions of an area measurement. This is not clear as written. I also recommend that you note that the resolution of an inversion system is not necessarily the same as the true resolution of an inverse flux estimate.

We clarified: "While typical inversion systems have a resolution ranging from tens of kilometers up to several degrees (hundreds of km), with the true resolution of the inverse flux estimates being even coarser, the spatial representativity of the flux observations typically covers an area with a radius of around a kilometer".

Page 4, line 15. There are many, many studies of flux upscaling with towers and spatial databases. The paper would benefit from a somewhat expanded review of this literature.

We added: "…one applied by Chevallier et al., (2012). Typical approaches to up-scale site level fluxes deploy for example model tree algorithms, a machine learning algorithm which is trained to predict carbon flux estimates based on meteorological data, vegetation properties and types (Jung et al., 2009, Xiao et al., 2008), or neural networks (Papale and Valentini 2003). Nevertheless eddy covariance measurements provide…".

References added:

Jung, M., Reichstein, M., and Bondeau, A.: Towards global empirical upscaling of FLUXNET eddy covariance observations: validation of a model tree ensemble approach using a biosphere model, Biogeosciences, 6, 2001-2013, doi:10.5194/bg-6-2001-2009, 2009.

Papale, D. and Valentini, R., A new assessment of European forests carbon exchanges by eddy fluxes and artificial neural network spatialization. Global Change Biology, 9: 525–535. doi: 10.1046/j.1365-2486.2003.00609.x, 2003

Xiao, J. F., Zhuang, Q. L., Baldocchi, D. D., et al.: Estimation of net ecosystem carbon exchange for the conterminous United States by combining MODIS and AmeriFlux data, Agr. Forest Meteorol., 148(11), 1827–1847, 2008.

Page 7, line 25. What does "across entire daily course" mean?

We mean flights were made at various times of the day to cover the daily course on average.

We clarified: "… flown 52 and 54 times respectively, covering the daily course. Exact routes…".

Page 9, lines 13-14. Why is the year of EC data used to optimize respiration parameters singled out? What about the other parameters and years of EC data?

For the heterotrophic respiration no calibration with many years was ever done here; so we followed an ad hoc procedure and chose the year 2007 to calibrate, which is the year for which the calculations presented in the paper have been done, to prevent any bias caused by a systematic error in the respiration. The calibration pertains by the way only to the "amplitude" of the respiration.

For other parameters, the calibrations come from Groenendijk et al., (2011) and are based on 4.5 years per site on average.

We modified : "The "5 parameter model" (5PM) (Groenendijk et al., 2011), also used in atmospheric inversions (Tolk et al., 2011, Meesters et al., 2012), is a physiological model describing transpiration, photosynthesis, and respiration. It uses MODIS LAI (leaf area index) at 10km resolution, meteorological data (temperature, moisture, and downward shortwave radiative flux, presently from ECMWF at 0.25 degrees resolution), and differentiates PFTs for different vegetation types and climate regions. 5PM fluxes are pertained to station locations at hourly temporal resolution. The optimization has been done with EC-data from Fluxnet as described (except for heterotrophic respiration) in Groenendijk et al., 2011. Regarding the heterotrophic respiration, an ad hoc optimization using Fluxnet EC-data from 2007 was performed since no previous optimization was availabe."

Page 10, line 3. English. I suggest, "it is the only model with spatial resolution (10 km) comparable to …"

We corrected: "VPRM was used since it is the only model with spatial resolution (10 km) comparable to …".

Page 10, lines 6-8. Number of flights is repeated. Resolution of aircraft flight data contradicts earlier text that stated 2 km resolution.

Repetition is deleted: "Each grid point was sampled 52 times in forests, and 54 in agricultural land".

We clarified: "… short flight distances. Aircraft NEE data, natively at 2 km resolution along the track, have been aggregated into 10 km segments, to maximize the overlap with the VPRM grid…".

Page 10, lines 15-16. Do you neglect the impact of the observation errors? Or is it just that you cannot separate these errors from errors in the model? The observation errors are already part of the observations.

As the observation errors are already part of the observations we do consider them in the model-observation analysis. The word "neglect" here only means that the impact on the correlation lengths is not studied here. The impact is studied later in the model-model analysis where we consider or neglect the observation error by adding or not a random measurement error to the fluxes used as reference.

We furthermore clarified: "therefore e-folding correlation length estimations do include the observation error term.".

Page 11, equations (2) and (3) and the "nugget" effect: Mathematically, when k=0, equation (2) = 1. What is this nugget effect? How can it exist for k=0? The numerator and denominator are identical when k=0.

"Nugget" is a standard term in spatial statistics. The nugget effect can be attributed to measurement errors or spatial sources of variation at distances smaller than the sampling interval or both. See also Wackernagel, H., 2003, Multivariate Geostatistics, Springer.

Page 9403 line 15. We clarified: "… also known as the nugget effect. The nugget effect is driven by measurement errors and variations at distances (spatial or temporal) smaller than the sampling interval. For this…".

Page 11, line 19. What does "distributed along the entire daily course" mean? And how can aircraft data span 36 days? Are the times in between the aircraft flights neglected?

We removed that unclear wording: "distributed along the entire daily course".

36 days is the duration of the flight campaign. Within this period, flights were planned and made to equally sample each time of the day, along the campaign. Since flights are obviously discontinuous in time, they constitute a sub-sampling thus are insufficient to compute a daily average at each grid-cell. Our approach was therefore to keep all individual fluxes to constrain the temporal autocorrelation, assuming that fluxes are overall not biased toward specific times of the day, given the sub-sampling is equally distributed.

And yes, temporal gaps that are not sampled are not used in the analysis.

Page 12, equation (4). Again, I don't see how a nugget is relevant with the normalized correlation equation (2).

We think that this is answered with the explanation of the nugget effect.

Page 12-13. "... assuming that the involved prior errors for each model are identical in a sense that they share the same statistics and (are?) not correlated." What does this mean?

We reformulated: ", assuming that models have independent prior errors".

Page 13, line 8. Richardson et al., 2008, only deals with random sampling error. More recent papers add gap filling errors and friction velocity screening errors to the observational error assessment.

We do not use gap filled data. The Richardson et al. (2008) reference was used to provide a somewhat realistic estimation of the observation error to be added in the model-model comparison. Including other sources of error such as resulting from gap filling and friction velocity screening would lead too far for this context.

There are many years of flux tower data. Only 2007 observations are used. Why? Could the results be strengthened with additional years of observations?

As we are interested in the daily time scale we used only 1 year data. Nevertheless by using additional years we would not expect to gain more information or significantly different correlation scales. For example Chevallier et al., 2012 used multiple years and we still have comparable results.

page 9397 line 18-19. We have clarified in this regard, also in response to the second reviewers comments by adding the following paragraph:

“Hilton et al., (2012) studied also the spatial model – data residual error structure using a geostatistical method. Hilton’s study is focused on the seasonal scale, i.e. investigated residual errors of seasonally aggregated fluxes. However, the state space (variables to be optimized considering also their temporal resolution) of current inversion systems is at high temporal resolution (daily or even three-hourly optimizations). Further, the statistical consistency between the error covariance and the state space is crucial. Thus the error structure at the daily time-scale is of interest here, and can be used in atmospheric inversions of the same temporal resolution. Similar to Hilton’s study we select an exponentially decaying model to fit the spatial residual autocorrelation.”

Page 14, lines 3-6. Your equations state (but do not define explicitly in the equations) that you are going to evaluate model-data, and model-model differences. The results, however, begin by stating standard deviations in observed NEE. I cannot tell what you have computed. Please clarify the methods and the associated results so that these values can be interpreted by the reader. I think you are presenting time-constant, spatially varying standard deviations across sites, then summed over all times. But the paper should not make me guess what you are computing here, and the methods say nothing about this computation.

This is just the standard deviation of the daily fluxes across all stations and all times to show how observed and modeled fluxes vary spatially and temporally. As this is a very general metric, we disagree that this should be stated explicitly in the methods section.

Page 14, line 11. Why is this “in line” with the spatial standard deviation? Model-data differences at a given site are not necessarily related in any way to differences in observed fluxes across sites.

We mean that the fact that the standard deviation of the residuals (modeled - observed) is only slightly smaller than the standard deviation of the fluxes themselves is in line with the small r-square values.

We have clarified the corresponding sentence: “Those values are only slightly smaller than the standard deviations of the observed or modeled fluxes themselves. This fact is in line with the generally low fraction of explained variance with r-square values …”.

Page 14, line 13-14. What are “site specific correlations” that are presented as a single value? If these are site specific correlations, what site is being presented? Again, the methods are not sufficiently clear.

We clarified: “When using site-specific correlations (correlations computed for each site, then averaged over all sites), …” .

Page 14, lines 15-18. This sentence’s English needs work. Further, the statistical comparison (see above comment) isn’t clear. Finally, the model-data difference, if I understand it, also includes a temporal component since it is summed over time. As best I can tell, the same data go into both calculations, so I don’t understand how the authors can draw this fuzzy conclusion about ability to simulate temporal variability better than spatial variability. Clear, targeted work on this topic has been published elsewhere but has been neglected in the introduction to this paper. This is also not clearly a main focus of this paper. I would suggest that you either expand the paper to address this topic properly, or delete this discussion.

We added a clarification of the computation in Page 9406, line 13, and we also reformulate regarding also comment from reviewer 2:

“When using site-specific correlations (correlations computed for each site, then averaged over all sites), the average fraction of explained variance increases to 0.38, 0.36, 0.35, and 0.42, for VPRM10, VPRM1, ORCHIDEE and 5PM, respectively. Note that for deseasonalized time-series (using a 2nd order harmonic, not shown) the same picture emerges with increased averaged site specific correlation compared to correlations using all sites. This indicates better performance for the models to simulate temporal changes (not only seasonal, but also synoptic) at the site level. Further, the differences between site-specific…”.

Figure 2 is not clear. Is each point a different site? How is this figure related (or not) to the “site specific correlations” noted on lines 13-14 of Page 14?

We clarified the “site-specific” term in the previous comment. This figure refers to the averaged, site-specific correlation, for sites that have the same vegetation type. The x-axis lists the different vegetation types (7 in total). Each site is characterized according to its representative vegetation. The total number of sites sharing a common vegetation type is shown under the vegetation type. Hence each of the bars are vegetation (according to x position in the plot) and model (according to color code) specific.

We also clarified the caption: “… Box and whisker plot for for site-specific correlation..”.

Figure 3 has the same problem as Figure 2. Please specify what distribution (sites?) are being illustrated by the box and whisker plots. In addition, the sign of the bias is never defined. Finally, it would be useful to provide a conversion to gC m-2 yr-1.

We clarified as in figure 2.

Page 9406, line 24 With respect to the second comment we clarified: "Figure 3 shows the distribution of bias (defined as modeled – observed fluxes) for different…".

Regarding the conversion, we added in the figure caption: "(for conversion to gC m^{-2} yr^{-1} reported values in y axis should be multiplied by 378,7694)".

Figure 4 has too many lines of similar color and tiny size to be read clearly.

We changed the color for the sites which were excluded from the analysis to better contrast with the remaining sites. However we note that this figure is not meant to be analyzed for each site individually (thin red and blue lines) but we rather want to show the average characteristics of the sites. Then the reader should concentrate on the all-site data (thick black and grey lines) and on the exponential fit (thick green and dark green lines).

Page 14, line 28. All site? Flux site? Sub-site? The terminology surrounding Figure 4 needs to be cleaned up. I cannot tell what is being plotted. The text appears to contradict the figure caption, and the terminology changes enough to be quite confusing.

"All-site" temporal autocorrelation is explained in page 9403 line 6. The "site data" as given in page 9402 equation 1 refers to the autocorrelation for each individual site and this is exactly what is plotted in figure 9 with the thin red lines. These show the temporal autocorrelations for each site (53 in total). Sub-site is explained in the caption of figure 4. It is computed according to "all-site" but excluding those sites with large model data bias. To make this more clear, we have reformulated the figure caption:

"Temporal lagged autocorrelation from model-data daily averaged NEE residuals for all models. Thin red lines correspond to different sites, while the blue thin lines reveal the sites with a bias larger than +/-2.5 μmol m^{-2} s^{-1}. The thick black line shows the all-site autocorrelation, and the thick grey line indicates the all-site autocorrelation but for a sub-set that excludes sites with large model-data bias ("sub-site"). The dark green line is the all-site exponential fit, and the light green line shows the all-site autocorrelation excluding the sites with large bias. The exponential fits use lag times up to 180 days."

Page 15, lines 2-10. Do not describe the figures in the text. The figures present these results. Discuss the significance of the figures in the text.

We have shortened the text by deleting: "The all-sites correlation for the VPRM model at 10 km resolution remains positive for lags < 104 days and for lags > 253 days. Weak negative correlations were found in between with minimum value -0.03. In contrast we found only positive correlation for VPRM at 1 km resolution for the whole year with a minimum value of 0.002. Similarly, ORCHIDEE follows the same patterns with positive correlations for lags < 76 days and for lags > 291. Minimum correlation was found to be -0.09. For 5PM model we also found only positive correlations. The minimum value was found to be 0.08."

However, some detail is needed in order to compare with the corresponding results by Chevallier et al., (2012).

The exponential fit appears to be a poor choice. Based on Figure 4, most of the site autocorrelations are below the fit lines at lag times of 30-70 days. Thus the model used to fit these curves is quite biased. It is also more biased for some model-data comparisons than for others. This makes the comparison of decay times misleading, to the point perhaps of being meaningless. I would not publish results based on such a biased approximation of the site-based results. I think this functional fit must be changed. At minimum, the quality of fit must be made very clear, and the logic for keeping this function, despite its relatively poor fit, articulated.

We note that the fit is performed on the all-site and sub-site curves (black and grey lines). The referee should not compare the fit (light and dark green lines) with the thin red lines.

We find r^2 values between the all-site autocorrelation and the exponential fit of 0.94, 0.94, 0.92, and 0.89 for VPRM10, VPRM1, ORCHIDEE and 5PM respectively. The standard deviations of the residuals (or RMSE for root mean square error) are 0.040, 0.036, 0.059, and 0.043 for the different models. Expressed as NRMSE (normalized RMSE, i.e. RMSE divided by the range of the autocorrelation), this results in values ranging from 0.061 to 0.092, which indicates relative errors in the fit of less than 10%. We therefore disagree with the reviewer that the fit is poor. Further we highlight also the importance of this fitting model, which is quite simple and is using only few parameters, which is a critical point for proper implementation into the inversion systems.

Following the suggestion of the reviewer for a better articulation of our arguments, we added the following indications:

Page 9407 line 14 We added: "However the correlogram exhibits a nugget effect (values ranging from 0.31 to 0.48 for the different models)"

Page 9407 line 17. We added: "The fit has a root mean square error ranging from 0.036 to 0.059 for the different biosphere models. The normalized RMSE (i.e. RMSE divided by the range of the autocorrelation) results in values ranging from 0.061 to 0.092 indicating relative errors in the fit of less than 10%."

Page 15 lines 13-15. If the correlation at zero lag is not 1, either your calculations are flawed or equation (2) does not represent your methods.

In spatial statistics nugget effect is a standard term which describes the sharp decrease of the correlation for infinitesimal temporal separation distances. The value which drops is 1-a and this is the nugget as described in eq. 3.

Page 15. A root mean square error of a functional fit to an autocorrelation curve is not very meaningful. Evaluating the quality of multiple potential fits to find the best fit to the data would be meaningful and improve the analysis.

RMS error is a measure of agreement. Considering the relatively simple exponential model which contains only 2 parameters to be optimized, the agreement is satisfactory good.

We would like to refer the referee to Chevallier et al. (2012), where a polynomial model was fitted with much more complexity (5^{th} order polynomial with 6 parameters), and the corresponding RMSE of 0.01 was not much smaller than the RMSE of 0.036 to 0.059 of our simpler (2 parameter) model.

Page 16, line 5. “not applicable”

We corrected: “…measurements are not applicable”.

Page 16, lines 9-11. The sites were screened because the bias was greater than 2.5 umol m-2 s-1. But now the text says that the bias for these sites was not greater than 2.5 umol m-2 s-1, just “larger than average.” I am confused. Please clarify.

This threshold value is exceeded for all of those sites only for 5PM modeled fluxes. For the rest of the models the value was over the threshold only for some sites. Nevertheless the bias even though not exceeding always the threshold value yet, was larger than the averaged bias.

We clarified: “… threshold of 2.5 μmol m^{-2} s simultaneously for each individual model…”.

There is no functional fit to the aircraft data (figure 5). Given the poor quality of the fits in Figure 4, and I am not convinced that there really is a difference between the two data sets. I would be much more convinced by a comparison of the mean or median values, binned by lag time.

We replaced the wrong figure with the appropriate one which contains the model fit. The fit found to have 13 days e-folding length with values between 10 and 16 days within 95% confidence interval. Hence we disagree that this difference is not significant. We also disagree that the fit is biased.

Page 16, line 29. What is the purpose of the root mean square error?

The use of RMSE is very common as a general purpose error metric for numerical predictions. As RMSE has the useful property of being in the same units as the response variable, we can then evaluate how good the model performs.

We also added the normalized RMSE (divided by the autocorrelation range) estimations:

"… 5PM, respectively. The normalized RMSE is found to have values ranging from 0.05 to 0.084 indicating relative errors of the fit less than 9%.".

Figure 6. The exponential fit is consistently below the median at distances of 200-400km. I would argue that your correlation computation shows consistently positive values out to approximately 200-400km, which is consistent with Hilton et al., (2013), who performed a similar calculation for North American flux tower sites and model-data differences using VPRM. Again, your exponential fit appears to be biased. I do not believe that quoting the results of a biased fit is sound.

The reviewer is right for the case of VPRM model which might be an exception. A careful look to the other two models (ORCHIDEE and 5PM) shows that autocorrelation values are well centered around the exponential for distances longer than 200km. We note also in Hilton's paper that the 1st bin is at 500km with no information for the smaller scales. In our study we used bins of approximately 100km.

Hilton et al., (2013), published a paper using very similar methods using North American flux towers, a much longer time series of data, and more evaluation of the robustness of the resulting length scales. This was published in Biogeosciences. The results contradict the results presented here in that Hilton et al (2013) found significantly larger length scales for their variogram fits. The similarity is so great that the Hilton et al (2013) paper really should be cited and evaluated with respect to these results.

Hilton et al. (2013) calculated the length scales by considering seasonal mean residuals. In our study we used daily averaged residuals since this is the temporal scale used in the state space for regional inverse models. This largely coarser time resolution used in Hilton et al. (3013) is likely the driver of the differences on the spatial scales.

Regarding the robustness of the fit, Hilton et al., 2013 compared with the AIC criterion whether the exponential fit or the pure nugget (which means no spatial coherence) is better. We note that they did not fit different error models to evaluate which model was fitting better. We added also:

Page 9397 line 18-19 "Hilton et al., (2012) studied also the spatial model – data residual error structure using a geostatistical method. Hilton's study is focused on the seasonal scale, i.e. investigated residual errors of seasonally aggregated fluxes. However, the state space (variables to be optimized considering also their temporal resolution) of current inversion systems is at high temporal resolution (daily or even three-hourly optimizations). Further, the statistical consistency between the error covariance and the state space is crucial. Thus the error structure at the daily time-scale is of interest here, and can be used in atmospheric inversions of the same temporal resolution. Similar to Hilton's study, we select an exponentially decaying model to fit the spatial residual autocorrelation."

Page 9413 line 17 additional paragraph: "Only weak spatial correlations for model-data residuals were found, comparable to those identified by Chevallier et al. (2012) limited to short lengths up to 40 km without any significant difference between the biospheric models (31 - 40 km). Hilton et al. (2012) estimated spatial correlation lengths of around 400km. However we note that significant differences exist between this study and Hilton et al. (2012) regarding the methods that were used and the landscape heterogeneity of the domain of interest. With respect to the first aspect the time resolution is much coarser (seasonal averaged flux residuals) compared to the daily averaged residuals used here. Furthermore spatial bins of 300 km were used for the autocorrelation analysis, which is far larger than the approximate bin width of 100 km that were used in our study. Regarding the second aspect North America has a more homogenous landscape compared to the European domain. The scales for each ecosystem type (e.g. forests, agricultural land etc.) are drastically larger than those in Europe as can be seen from MODIS retrievals (Friedl et al., 2002).".

Figure 7. How is the confidence interval computed? This is not a simple case of computing the standard deviation of a Gaussian. Please explain the methods. I am still dubious of the value of the exponential fit, but in any case the methodology for the confidence interval estimate must be explained.

Page 9404 line 19 We clarified by adding following paragraph: "Confidence intervals for the estimated model parameters were computed based on the profile likelihood (Venzon and Moolgavkar, 1987) as implemented within the "confint" function from MASS package inside the R statistical language.".

Figure 7 points out something that is lacking from the primary results reported in the abstract – uncertainty bounds. These results suggest that the uncertainties in the computed correlation lengths are very large. This should be reported in the abstract.

Page 9394 line 16 We added : "… up to few tens of km but with uncertainties up to 100% of this estimation".

The standard spatial statistical method for Figures 6 and 7 would be a variogram. Why have the authors chosen a different approach?

The choice of the correlogram over the variogram was made since a) Chevallier et al. (2012) also used the correlogram for a similar analysis, and b) it was simpler to implement in the code.

Page 18, lines 5-6. English. ‘it difficult to determine … where the asymptote lies” perhaps?

Corrected, now reads : “making it difficult to identify where the asymptote lies.”.

Figure 8 illustrates again how poorly the exponential model fits the data. And the exponential model is not shown on the figure, which is inconsistent with figures 6 and 7. D=35km with a 95% confidence interval of 26-46 km is clearly biased given that none of the aircraft data reaches 1/e of the zero correlation anywhere within that range. The exponential model is poor and should not be used, or only with serious caveats about the biased nature of the fit.

We corrected: the figure now shows the model fit. However, we disagree in this point with the reviewer, the fit is not biased when looking at the residuals in Fig. 8.

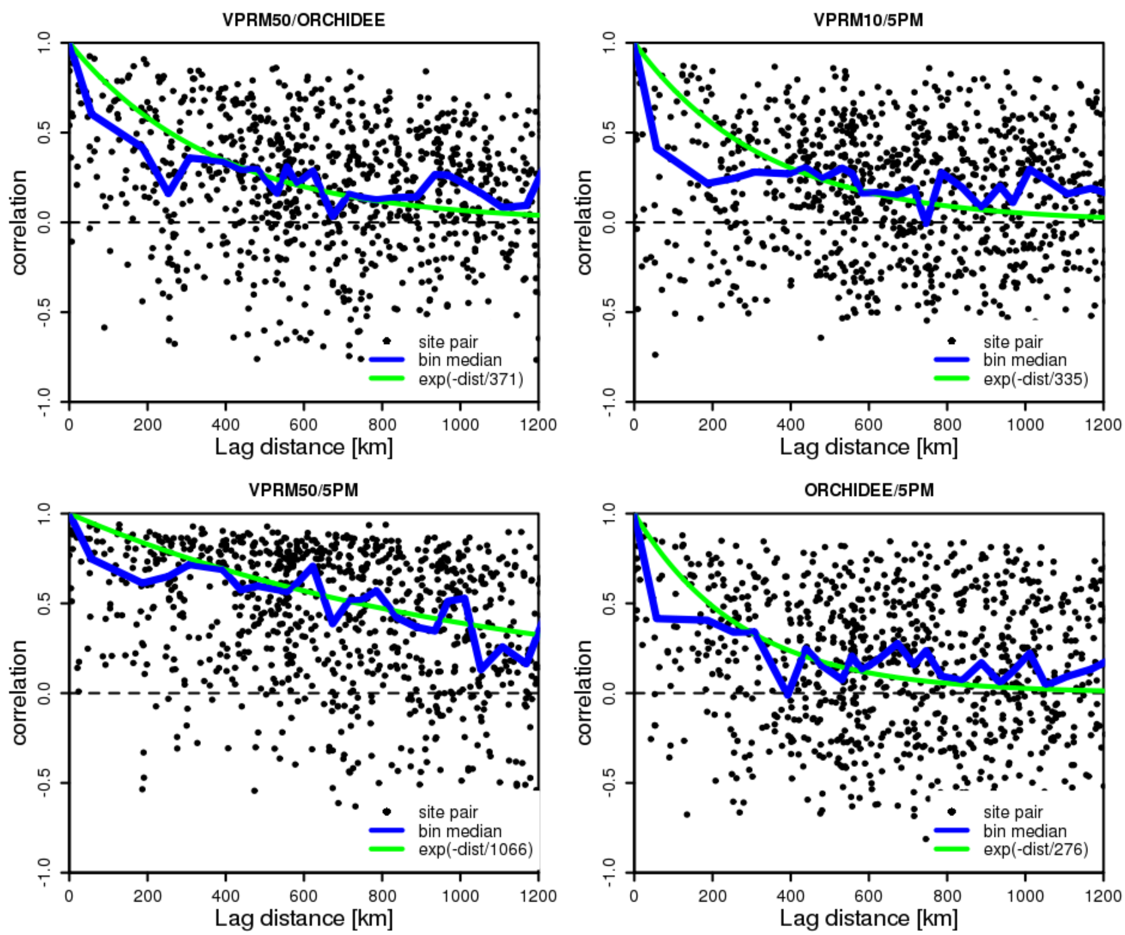
Some of the colors in Figure 9 are nearly indistinguishable when used to plot very thin lines. Please either reduce the content of the figure or find a way to distinguish the different model-model pairs more clearly.

We reduced the information and we added also comparisons between VPRM50/ORCHIDEE following also the comment from referee #2 regarding the incompatible model resolution.

Figure 9: Why are the individual points not shown, as for the model-data comparison? I can understand reducing the information shown, but I am concerned about the quality of the exponential fit, and it is impossible to evaluate from this figure.

In the figure below, which is similar to figure 7 we present the correlations of differences for all the different model-model combinations (model-data for figure 7) together with the respective exponentially decaying model fits. Figure 9 is not meant to present the goodness of fit neither to show all paired correlations.

Nevertheless we plotted the spatial autocorrelation for the paired models in order to evaluate the exponential fit (referred to as figure 11 in the attached files).



Above: Distance correlogram for the daily net ecosystem exchange (NEE) differences between pairs of models using all sites. Black dots represent the different site pairs; the blue line represents the median value of the points per 100-km bin, and the green line shows an exponential fit. Results are shown for differences between VPRM at a resolution of 50 km vs. ORCHIDEE (top left), between VPRM at a resolution of 1 km and 5PM (top right), between VPRM at a resolution of 50 km and 5PM (bottom left), and between ORCHIDEE and 5PM (bottom right).

Figure 10. Again, the blue-green-purple lines are difficult to distinguish (all dark), and the red-green lines will be indistinguishable for those who are red-green colorblind.

We corrected the plot and further we reduced the amount of information. We also added the VPRM50/ORCHIDEE comparisons.

Discussion and conclusions. The results are compared only to Chevallier's publications, and that comparison is limited to the temporal autocorrelation. There is insufficient effort to put these results into the context of prior work on this topic.

Page 9413, lines 17 We added a new paragraph: "Only weak spatial correlations for model-data residuals were found, comparable to those identified by Chevallier et al. (2012) limited to short lengths up to 40 km without any significant difference between the biospheric models (31 - 40 km). Hilton et al. (2012) estimated spatial correlation lengths of around 400km. However we note that significant differences exist between this study and Hilton et al. (2012) regarding the methods that were used and the landscape heterogeneity of the domain of interest. With respect to the first aspect the time resolution is much coarser (seasonal averaged flux residuals) compared to the daily averaged residuals used here. Furthermore spatial bins of 300 km were used for the autocorrelation analysis, which is far larger than the approximate bin width of 100 km that were used in our study. Regarding the second aspect North America has a more homogenous landscape compared to the European domain. The scales for each ecosystem type (e.g. forests, agricultural land etc.) are drastically larger than those in Europe as can be seen from MODIS retrievals (Friedl et al., 2002)."

Page 21, lines 26-28. The observational errors are in your calculations. It is not neglected. It cannot be isolated and removed, but it is not neglected.

We deleted part of the sentence, which now reads: "Of note is that the eddy covariance observation error has no significant impact on the error structure, as the addition of an observation error to the analysis of model-model differences had only minor influence on the error structure.,.

End of page 24 beginning of page 25. This is an interesting discussion. Again, the methods used for this interesting calculation are opaque. Please explain. Propagating these error estimates is not trivial. How was this done?

The theoretical approach for this calculation is based on Rodger (2000) by introducing an aggregation operator. We explain with the following equation where "×" represents matrix multiplication notation:

$u \times Q_c \times u^T$. Q_c is the full prior error covariance matrix with dimensions equal to the product between number of regions (grid-cells) and number of timesteps. With our setup this translates into $184 \cdot 104 \cdot 8 \cdot 365 = 55877120^2$ elements (lon·lat·timesteps·days). The main diagonal contains the model-data difference scaled down to account for the difference in spatial resolution of the state space. The off-diagonal elements contain the spatial and temporal correlations. u is a scalar operator that aggregates the full covariance to the target quantity (i.e. domain-wide and full year).

Page 9416 line 24. We added: "… over longer time periods. To aggregate the uncertainty to large temporal and spatial scales, we used the following equation (after Rodgers, 2000):

$$Ua = u \times Q_c \times u^T \qquad (7)$$

Where "×" denotes matrix multiplication, *Qc* is the prior error covariance matrix and *u* a scalar operator that aggregates the full covariance to the target quantity (e.g. domain-wide and full year)."

Reference added:

"Rodgers, C., D. Inverse methods for Atmosphere Sounding: Theory and Practice, World Sci., River Edge, N. J., 2000" page 30 line 20

Page 22 line 25-28 Same section: Comparing your aggregated error estimate to the range of existing continental-scale flux estimates (e.g. Peylin et al, 2013) would be more useful than the very limited analysis presented.

We agree with the reviewer and we added: "This value is also 8 times smaller when comparing it to the variance of the signal between 11 global inversions reported in Peylin et al., (2013) which was found to be 0.45 GtC/y, proving that the aggregated uncertainties are unrealistically small."

Same section: I agree that this analysis (pending evaluation of the unknown methods) would strongly suggest that the total continental-scale, annual flux errors are seriously underestimated, and I agree that this is an important issue to point out. This should be part of the abstract, as it leads to significant uncertainty regarding the validity of the correlation lengths. The current abstract suggests no such uncertainty regarding the conceptual model promoted in this paper.

Page 9394 line 16. We added: "Propagating this error structure to annual continental – scale yields an uncertainty of 0.06 Gt C and strongly underestimates uncertainties typically used from atmospheric inversion systems, revealing the existence of another potential source of errors. Long spatial e-folding correlation lengths up to several…".

Page 25, paragraph starting with "Exponentially decaying …". This paragraph begins to give reasons for using an exponential model for the correlations. Some notes below about this discussion:

1) **This discussion belongs earlier in the paper. It presents the logic for using this fit.**

We moved this paragraph and now is located in page 9412 early in the discussion.

2) The reasons given are entirely reasons of simplicity and convenience, not accuracy of the fit. I would suggest that the best job of describing the correlations should be the primary goal of this paper. Considering how to simplify these correlation functions to make them convenient is another problem. I have already noted that I believe the exponential fit is so poor that it is significantly misleading. The analysis would be improved by evaluating different fits and finding what fits are best given the data.

Regarding the simplicity and convenience we will refer the reviewer to Hilton et al., 2013. In that study two models were used. An exponentially decaying model and one that uses a pure nugget effect. The nugget only model is equivalent to an exponentially decaying model with the length scale of zero, which means no spatial correlation can be detected. So the assessment of whether the nugget only model or the exponentially decaying model in Hilton is appropriate, could simply be done, by assessing if the length scale in the exponentially decaying model is significantly different from zero.

Further this study tries to give insights in the error structure targeting to describe prior uncertainties with a relevant way that the atmospheric inverse modeling community may benefit from it. The exponential model is widely used by this community. Describing the correlations with a pure mathematical way which makes a convenient fit but is not being used from the inverse modeling community is also of less importance. However we do not believe that the fits are poor.

3) Lines 20-25. Computational simplicity is not a good reason to use the wrong correlation length. This is a disturbing discussion.

We disagree with the reviewer at this point that the correlation lengths are wrong, see also responses to previous comments. The exponential model might be simple but performs satisfactory good. Further despite the hesitation of the reviewer regarding the importance of computational efficiency, this is a major issue for regional scale inverse modeling.

Equation (7) does not exist in the paper.

Page 9417 line 23 Corrected, the sentence now reads "Using the same hyberbolic equation for the spatial correlation…".

Page 26, lines 3-4. English needs work. And I'm not sure what is meant by "for the short spatial scales." And what studies have already used the correlation lengths derived from this study? I'm not sure that the 'future work' needs to be part of this paper

With this sentence " …short spatial scales" we mean that numerous studies in atmospheric inverse modeling use significantly larger spatial correlation scales than those derived in this study. However most of the inversion systems already use temporal correlation scales of around 1 month, which is in line with our findings.