Point-by-point response to the reviews including relevant changes made in the manuscript:

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.

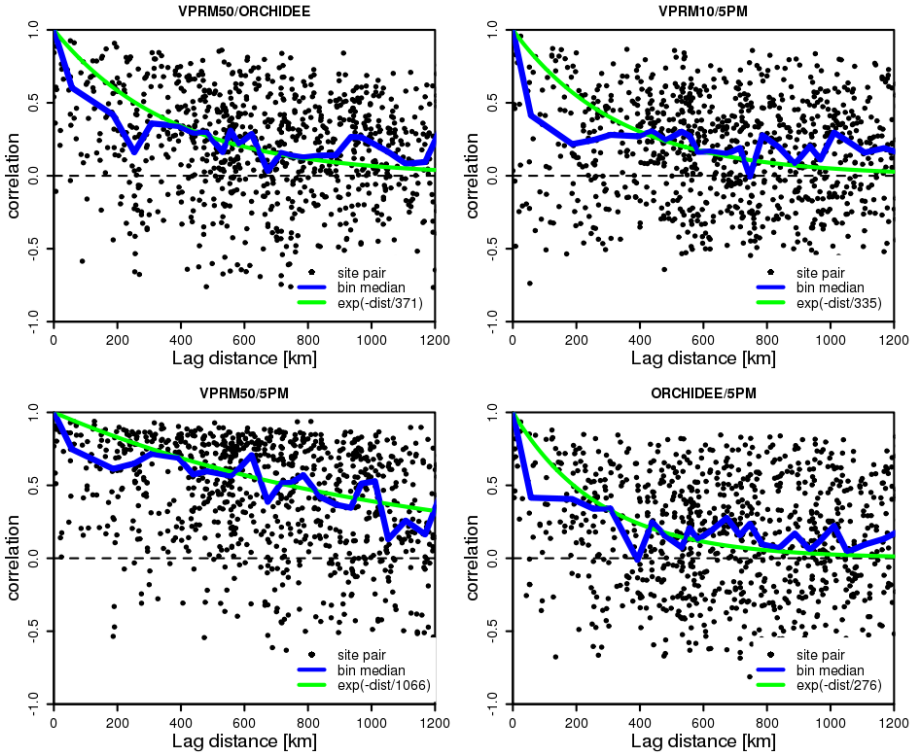
Some of the colors in Figure 9 are nearly indistinguishable when used to plot very thin lines. Please either reduce the content of the figure or find a way to distinguish the different model-model pairs more clearly.

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):

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

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.

Referee 2

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

MAJOR CONCERN:

My largest concern with the study is that the spatial resolution / support of the flux observations is substantially finer than the spatial resolution of the model simulations, making flux values at these disparate scales fundamentally incompatible. The authors

acknowledge as much in p. 9397 lines 11-13 "While typical inversion systems have a resolution ranging from tens of kilometers up to several degrees (hundreds of km), the spatial representativity of the flux observations is typically around a kilometer." In the Chevallier et al. studies that the authors cite, the analysis of errors was conducted by comparing km-scale flux observations with "a site-scale configuration of the ORCHIDEE model," thereby leading to compatible spatial scales. The resulting error statistics were then upscaled to be representative of the scales estimated by typical inversions.

We agree with the reviewer that some of the scales seem incompatible, and have made the following changes in the manuscript:

With respect to the model-model comparisons we produced fluxes also at 50 km resolution for VPRM model. VPRM now provides 3 different resolutions at 50 10 and 1 km (VPRM50 VPRM10 and VPRM1 respectively).We compare VPRM at 50 km with ORCHIDEE (which has also 50 km spatial resolution). We add this in plot 9 and 10. VPRM10-5PM comparisons are also in line as both models have the same spatial resolution.

We withdraw VPRM1-VPRM10 comparisons from plot 9 and we have also deleted following sentence:

Page 9415 line 21-27 "A special case in the context of the model-model study is the comparison between VPRM1 and VPRM10, which is the only case that produced short spatial correlation scales. These two models only differ in the spatial resolution of MODIS indices EVI and LSWI (1 vs. 10 km). Thus differences between those two models are only related to variability of these indices at scales below 10 km, which is not expected to show any spatial coherence. Indeed the results show only very short correlation scales (Fig. 9) with an exception during fall, however there the uncertainty is also large."

With respect to the model-data analysis VPRM1 fluxes are obtained at a spatial scale comparable with the flux observations. Further we found no significantly different spatial scales from the model-data residual autocorrelation analysis from the rest of the models (5PM, ORCHIDEE and VPRM10/50. Therefore we do not expect our results to be biased.

We added the following sentences/paragraphs:

page 9400 line 12 "… at hourly temporal resolution and at three spatial resolutions of 1, 10 and 50 km (referred to as VPRM1,VPRM10 and VPRM50)".

We added a discussion about the scale mismatch:

page 9402 line 2 "… over the year 2007. Simulated fluxes from the different models are at different spatial resolution, which makes comparisons difficult to interpret. For the model-data residual analysis, the models VPRM1, VPRM10, ORCHIDEE and 5PM were used. We note that VPRM1 with 1km resolution is considered compatible when comparing with local measurements. For the model-model analysis we use VPRM50 at 50km resolution when comparing with ORCHIDEE fluxes, as both models share the same resolution. VPRM10 is considered also appropriate for comparisons with 5PM model as they both share same resolution (MODIS LAI resolution of 1 km aggregated to 10 km and meteorological resolution at 0.25 degrees). Following we compare VPRM50 with 5PM to investigate if the different spatial resolution influences the correlation scale as a measure of how trustful might be the derived scales from ORCHIDEE – 5PM comparisons.",

page 9405 line 1 : "For the model-model analysis fluxes derived from the model pairs VPRM50-ORCHIDEE and VPRM10-5PM share the same spatial resolution and therefore are fully comparable. Similar to the …"

page 9408 line 1 "The e-folding correlation lengths show no dependence on the modeled flux resolution as same results yielded from all models. Further we examined also the spatial autocorrelation from VPRM50-data residuals with no significant difference compared to previous results.".

3.2 Section is modified and refers to the new model-model pairs:

"We investigate the model-model error structure of NEE estimates by replacing the observed fluxes

which were used as reference, with simulated fluxes from all the biosphere models. Note that for consistency with the model-data analysis, the simulated fluxes contained the same gaps as the observed flux time series. The e-folding correlation time is found to be slightly larger compared to the model-data correlation times, for most of the cases. An exception is the 5PM-VPRM10 pair which produced remarkably larger correlation time (Table 2). Specifically, VPRM50-ORCHIDEE and VPRM10-5PM residuals show correlation times of 28 days (range between 24-32 days within 95% confidence interval) and 131 (range between 128-137 days within 95% confidence interval), respectively. Significantly different e-folding correlation times are found for VPRM50-5PM compared to VPRM10-5PM with correlation times of 52 days (range between 49-56 days within 95% confidence interval).Repeating the analysis excluding sites with residual bias larger than 2.5μmol/m2s, correlation times of 28 and 100 days for VPRM50-ORCHIDEE and VPRM10-5PM are found, respectively. If we use ORCHIDEE-5PM pair the e-folding correlation time found to be 38 days (range between 35-41 days within 95% confidence interval).

Although the e-folding correlation times show but minor differences compared to the model-data residuals, this is not the case for the spatial correlation lengths (Fig. 9). The standard case (S) was applied for the annual analysis, with no minimum number of days with overlapping non-missing data for each site within the pairs. Taking VPRM50 as reference, much larger e-folding correlation lengths of 371 km with a range of 286-462 km within 95% confidence interval yielded for VPRM50-ORCHIDEE comparisons, and 1066 km for VPRM50-5PM were found. However VPRM10-5PM analysis which is also considered appropriate in terms of the spatial resolution compatibility contrary to the VPRM50-5PM pair, is in good agreement with VPRM50-ORCHIDEE spatial scale (230-440 km range within 95% confidence interval with the best fit being 335 km). With ORCHIDEE as reference the e-folding correlation length for the ORCHIDEE-5PM comparison is 276 km with a range of 183-360 km within 95% confidence interval. However the later correlation length might be affected by the different spatial resolution as the difference between VPRM10 and VPRM50 against 5PM suggests. Seasonal e-folding correlation lengths, using a minimum of 20 days overlap in the site-pairs per season (Fig. 9), are also significantly larger compared with those from the model-data analysis.

When we add the random measurement error to the modeled fluxes used as reference (crosses in Fig. 9), we observe only slight changes in the annual e-folding correlation lengths, without a clear pattern. The correlation lengths show a random increase or decrease but limited up to 6%. Interestingly, the seasonal e-folding correlation lengths for most of the cases show a more clear decrease. For example, the correlation length of the VPRM10-5PM residuals during winter, decreases by 22% or even more for spring season. Despite this decrease, the e-folding seasonal correlation lengths remain significantly larger in comparison to those from the model-data analysis. Overall, all models when used as reference show the same behavior with large e-folding correlation lengths that mostly decrease slightly when the random measurement error is included. Although the random measurement error was added as “missing part” to the modeled fluxes to better mimic actual flux observations, it did not lead to correlation lengths similar to those from the model-data residual analysis. To investigate if a larger random measurement error could cause spatial correlation scales in model-model differences, we repeated the analysis with artificially increased random measurement error (multiplying with a factor between 1 and 15). Only for very large random measurement errors did the model-model e-folding correlation lengths start coinciding with those of the model-data residuals (Fig. 10).”

Page 18 line 14-18 We added: “Whilst fluxes from ORCHIDEE model are at much coarser resolution compared to the representative area from the flux measurements, VPRM1 fluxes (1 km resolution and only the meteorology at 25 km) are considered appropriate for the comparisons. Despite the scale mismatch results are in good agreement across all model-data pairs.”.
Table 2 is also changed.

Reference	VPRM10 [days]	VPRM1 [days]	ORCHIDEE [days]	5PM [days]
OBSERVATION	32 (27)	33 (29)	26 (24)	70 (34)
VPRM50	-	-	28 (28)	52 (46)
VPRM10	-	-	-	131 (100)
ORCHIDEE	-	-	-	38 (32)
5PM	-	-	-	-

For all of the analysis, it would be important to more explicitly discuss the time scales for which the analyses are conducted, and emphasize that the error statistics computed therein are therefore only valid for that same (i.e. daily) temporal resolution. Both the spatial and temporal correlation lengths will be affected by the temporal resolution of the analyzed data.

We made two additions to better clarify this:

Page 9397 line 18 in response also from referee 1 comment we added "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."

Page 9414 line 2. "… on the error structure. We note that the current analysis focuses to daily time scale and therefore the error statistics with respect to the estimated spatial and temporal correlation lengths are valid for such scales.".

Throughout the manuscript, the terms "correlation length" / "correlation time" (approximately 3*tau and 3*d in the authors' notation in eqns. 3 and 4) and the terms "e-folding time" (tau) and "e-folding correlation length" (d) and their variants are used, but due to the number of variations, it is not always clear when the authors are referring to 3*tau vs. tau, and to 3*d vs. d. This should be made completely clear throughout to avoid confusion. Please also pay close attention to this when comparing your numbers to those from earlier studies.

We corrected and we refer to all lengths throughout the revised manuscript as "e-folding correlation lengths" following also the notation from Chevallier et al., (2012).

For the airborne analysis, the authors find correlation lengths of approximately 39 days (3 * e-folding time of 13 days, page 9408 line 22). Given that there are only 36 days of data, correlation lengths of much beyond ~18 days (half the maximum separation distance) cannot be reliably identified. This should, at a minimum, be discussed.

We disagree at this point, and argue that e-folding times of 13 days can in fact well be fitted with time series data in which time differences are up to 35 days. This is also obvious from the 95% confidence interval which we added to the paper:

page 9408 line 22 "… correlation time of 13 days (range of 10 – 16 days within the 95% confidence interval). Whilst the …".

In terms of the overall correlation lags, the authors need to make a fundamental choice as to whether they are trying to represent errors at synoptic scales, or errors at seasonal scales. While the numbers that come out of their analysis represent errors at the seasonal-scale, it is important to note that this means that they are assuming that errors at the synoptic scale are very highly correlated. This may not be a valid assumption. Although I understand how these numbers come out of the analysis as it has been designed, some thought should be given

to whether these are indeed the scales that are relevant to whatever atmospheric inversions the authors have in mind

This is obviously a misunderstanding. We did not intend to estimate the error structure at synoptic scales but rather to study if the error structure has a seasonal dependence. We have made the manuscript more clear by adding:

Page 9404 line 8. "… observations was applied. We note that we do not intend to investigate the errors at the seasonal scale but rather to study if different seasons trigger different error correlation structures. "

p. 9396 lines 11-12 This statement is not entirely correct. Objective approaches were proposed earlier by Michalak et al. (2004, 2005), and have been applied in a number of studies since. The authors distinguish the Michalak et al. (2004) study as applying a "geostatistical" approach, but fundamentally both inversion approaches rely on characterizing the statistical characteristics of prior errors. I note that the Michalak et al. (2005) study was also for a classical Bayesian approach.

We agree and clarified: "This is because only recently an objective approach to define prior uncertainties based on mismatch between modeled and observed fluxes has been developed (Chevallier et al., 2006 and 2012)."

- Airborne flux observations: 10km spatial windows, but no indication of the "width" of the window (p. 9402 line 6), i.e. 10km x ?km.

The width of the windows was indeed computed with footprint modeling. Each individual flux determination the footprint distance depends on atmospheric conditions and extends upwind the measurement transect. On average for the entire campaign, a peak footprint distance was computed at 514 m, while the 90% footprint distance (i.e including 90% of observed flux) was computed at 3.9 km.

Page 9402 first paragraph. We added: "Footprint areas of aircraft fluxes were computed with the analytical model of Hsieh et al. (2000) yielding an average footprint width containing 90% of the flux of 3.9 km. Averaging also over the different wind directions (perpendicular or parallel to the flight direction), and taking into account the 10 km length of the segments, the area that the aircraft flux data corresponds to, is around 23.5 km^2 +- 12 km^2."

p. 9402 lines 5-7: I disagree with this statement. Even if the aircraft observations were "grouped" into 10km segments, this still does not match the VPRM grid, as the airborne segments are not representative of a 10km "width," just "length" along the flight path.

The reviewer is correct. This recalls to the footprint analysis comment. 90% footprint width was computed at 3.9 km thus fluxes are not representative of entire VPRM grid-cells, but still the 10-km grouping is the best strategy adoptable.

Page 9402 first paragraph. We corrected: "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, obtaining 6 grid points in forest transects and 8 in agricultural land transects.".

p. 9403 eqn. 3 and associated text: A nugget parameter would typically be defined as one minus alpha in the notation used by the authors, as it represents the portion of the variability that is not spatially (or temporally) correlated.

We corrected "(1-a)" Equation 3.

p. 9406 line 13-18: I wonder whether the better correlations at the site scale are simply due to the fact that the models and towers agree as to the overall seasonality of the fluxes. A more representative analysis might be to calculate the correlations after removing an average seasonality.

We did the analysis again with deseasonalized timeseries. For that we fit a 2-rank sinusoidal

equation to the flux data and we subtract it from them. This results to the following correlation values for VPRM1, VPRM10, ORCHIDEE and 5PM respectively:
All site correlations: 0.12, 0.10, 0.06, 0.14
And for site scale: 0.18, 0.18, 0.16, 0.22
Page 9406 line 15. We clarified by adding: "… 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."

Manuscript version with marked-up changes

An objective prior error quantification for regional atmospheric inverse applications

Kountouris, P.,[1] Gerbig, C.,[1], Totsche, K.-U.,[2] Dolman, A.J.,[3] Meesters, A.G.C.A.,[3] Broquet, G.,[4] Maignan, F.,[4] Gioli, B.,[5] Montagnani, L.,[6] and Helfter, C.[7]

[1]{Max Planck Institute for Biogeochemistry, Jena, Germany}

[2]{Institute of Geosciences, Department of Hydrology, Friedrich-Schiller-Universitaet, Jena, Germany}

[3]{VU University Amsterdam, Amsterdam, Netherlands}

[4]{Laboratoire des Sciences du Climat et de l'Environnement, CEA-CNRS-UVSQ, UMR8212, IPSL, Gif-sur-Yvette, France}

[5] Institute of Biometeorology, IBIMET CNR, Firenze, Italy

[6]{ Faculty of Science and Technology, Free University of Bolzano, Piazza Università 5, 39100 Bolzano, Italy; Forest Services, Autonomous Province of Bolzano, Via Brennero 6, 39100 Bolzano, Italy}

[7]{Centre for Ecology and Hydrology, Edinburgh, UK}

Correspondence to: P. Kountouris (pkount@bgc-jena.mpg.de)

Abstract

Assigning proper prior uncertainties for inverse modeling of CO_2 is of high importance, both to regularize the otherwise ill-constrained inverse problem, and to quantitatively characterize the magnitude and structure of the error between prior and "true" flux. We use surface fluxes derived from three biosphere models VPRM, ORCHIDEE, and 5PM, and compare them against daily averaged fluxes from 53 Eddy Covariance sites across Europe for the year 2007, and against repeated aircraft flux measurements encompassing spatial transects. In addition we create synthetic observations using modeled fluxes ~~to substitute~~instead of the observed ~~by modeled fluxes~~ones, to explore the potential to infer prior uncertainties from model-model residuals. To ensure the realism of the synthetic data analysis, a random measurement noise was added to the modeled tower fluxes which were used as reference. The temporal autocorrelation time for tower model-data residuals was found to be around ~~35~~30 days for both VPRM and ORCHIDEE, but significantly different for the 5PM model with ~~76~~70 days. This difference is caused by a few sites with large ~~model data~~ biases between the data and the 5PM model. The spatial correlation of the model-data residuals for all models was found to be very short, up to few tens of km~~. Long spatial~~ but with uncertainties up to 100% of this estimation. 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 hundreds of km were determined when synthetic data were used. Results from repeated aircraft transects in south-western France, are consistent with those obtained from the tower sites in terms of spatial autocorrelation (35 km on average) while temporal autocorrelation is markedly lower (13 days). Our findings suggest that the different prior models have a common temporal error structure. Separating the analysis of the statistics for the model data residuals by seasons did not result in any significant differences of the spatial e-folding correlation lengths.

1 Introduction

Atmospheric inversions are widely used to infer surface CO_2 fluxes from observed CO_2 dry mole fractions with a Bayesian approach (Ciais et al., 2000, Gurney et al., 2002, Lauvaux et al., 2008). In this approach a limited number of observations of atmospheric CO_2 mixing ratios are used to solve for generally a much larger number of unknowns, making this an ill-posed problem. By using prior knowledge of the surface-atmosphere exchange fluxes and by using an associated prior uncertainty, the information retrieved in the inversion from the observations is spread out in space and time corresponding to the spatiotemporal structure of the prior uncertainty. 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. This prior knowledge typically comes from process-oriented or diagnostic biosphere models that simulate the spatiotemporal patterns of terrestrial fluxes, as well as from inventories providing information regarding anthropogenic fluxes such as energy consumption, transportation, industry, and forest fires.

The Bayesian formulation of the inverse problem is a balance between the a priori and the observational constraints. It is crucial to introduce a suitable prior flux field and assign to it proper uncertainties. When prior information is combined with inappropriate prior uncertainties, this can lead to poorly retrieved fluxes (Wu et al., 2011). Here, we are interested in biosphere-atmosphere exchange fluxes and their uncertainties, and make the usual assumption that the uncertainties in anthropogenic emission fluxes are not strongly affecting the atmospheric observations at the rural sites that are used in the regional inversions of biosphere-atmosphere fluxes.

Typically inversions assume that prior uncertainties have a normal and unbiased distribution, and thus can be represented in the form of a covariance matrix. The covariance matrix is a method to weigh our confidence of the prior estimates. The prior error covariance determines to what extent the posterior flux estimates will be constrained by the prior fluxes. Ideally the prior uncertainty should reflect the mismatch between the prior guess and the actual (true) biosphere-atmosphere exchange fluxes. In this sense it needs to also have the corresponding error structure with its spatial and temporal correlations.

A number of different assumptions of the error structure have been considered by atmospheric CO_2 inversion studies. Coarser scale inversions often neglect spatial and temporal correlations as the resolution is low enough for the inverse problem to be regularized (Bousquet et al., 1999, Rödenbeck et al., 2003a) or assume large spatial correlation lengths (several hundreds of km) over land (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 scale inversions, with higher spatial grid resolutions which are often less than 100 km, the spatial

correlations are decreased (Chevallier et al., 2012) and the error structure need to be carefully defined. A variety of different assumptions exist. This is because only recently an objective approach to define prior uncertainties ~~was followed~~based on mismatch between modeled and observed fluxes has been developed (Chevallier et al., 2006 and 2012). In some regional studies, the same correlations are ~~derived from~~used as in large scale inversions in order to regularize the problem, although the change of resolution could lead to different correlation scales (Schuh et al., 2010). Alternatively, they are defined with a correlation length representing typical synoptic meteorological systems (Carouge et al., 2010). In other cases, ad-hoc solutions are adopted, where the correlation lengths are assumed to be smaller than in the case of global inversions (Peylin et al., 2005), or derived from climatological and ecological considerations (Peters et al., 2007) where correlation lengths only within the same ecosystem types have a value of 2000 km. In addition some studies use a number of different correlation structures in order to analyze which seems to be the most appropriate one based on ~~some evaluation~~cross-validation of the ~~resulting~~simulated against observed CO_2 mole fractions. The simulated mole fractions were derived using the influence functions and the inverted fluxes (Lauvaux et al., 2012). Michalak et al., (2004) applied a geostatistical approach based on the Bayesian method, in which the prior probability density function is based on an assumed form of the spatial and temporal correlation and no prior flux estimates are required. It optimizes the prior error covariance parameters, the variance and the spatial correlation length by maximizing the probability density function of the observations with respect to these parameters.

A recent study by Broquet et al~~.~~, (2013) obtained good agreements between the ~~statistics at the European and 1-month scale of both the prior and posterior~~ statistical uncertainties as derived from ~~their inversions of~~ the ~~biosphere fluxes~~inversion system and ~~that of~~ the ~~average~~actual misfits ~~of~~calculated by comparing the ~~prior and~~ posterior ~~estimates of the~~ fluxes to ~~the~~ local flux measurements~~.~~ at the European and 1-month scale. These good agreements relied in large part on their definition of the prior uncertainties based on the statistics derived in an objective way from model-data mismatch by Chevallier et al., (2006) and Chevallier et al., (2012). In these studies, modeled daily fluxes from a site scale configuration of the ORCHIDEE model are compared with flux observations made within the global FLUXNET site network, based on the eddy covariance method (Baldocchi et al., 2001), and a statistical upscaling technique is used to derive estimates of the uncertainties in ORCHIDEE simulations at lower resolutions. 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 ~~is~~ typically covers an area with a radius of around a kilometer. Considering also the scarcity of the observing sites in the flux network, the spatial information they bring is limited

without methods for up-scaling such as the one applied by Chevallier et al., (2012). ~~Nevertheless these~~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 a unique opportunity to infer estimates of the prior uncertainties by examining model-data misfits for spatial and temporal autocorrelation structures.

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 often 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.

In this study, we augment the approach of Chevallier et al., (2006 and 2012), to a multi-model - data comparison, investigating among others a potential generalization of the error statistics, suitable to be applied by inversions using different biosphere models as priors. This expectation is derived from the observation that the biosphere models, despite their potential differences typically have much information in common, such as driving meteorological fields, land use maps, or remotely sensed vegetation properties, and sometimes even process descriptions. We evaluate model – model mismatches to (I) investigate intra-model autocorrelation patterns and (II) to explore whether they are consistent with the spatial and temporal e-folding correlation lengths of the model – data mismatch comparisons. Model comparisons have been used in the past to infer the structure of the prior uncertainties. For example, Rödenbeck et al., (2003b) used prior correlation lengths based on statistical analyses of the variations within an ensemble of biospheric models. This approach is to a certain degree questionable, as it is unclear how far the ensemble of models actually can be used as representative of differences between modeled and true fluxes. However, if a relationship between model – data and model – model statistics can be established for a region with dense network of flux observations, it could be used to derive prior error structure also for regions with a less dense observational network.

Moreover, to improve the knowledge of spatial flux error patterns, we make use of a unique set of aircraft fluxes measured on 2-km spatial windows along intensively sampled transects of several tens of km, ideally resolving spatial and temporal variability of ecosystem fluxes across the

landscape without the limitation of the flux network with spatial gaps in between measurement locations. Lauvaux et al., (2009) compared results of a regional inversion against measurements of fluxes from aircraft and towers, while this is the first attempt to use aircraft flux measurements to assess spatial and temporal error correlation structures.

This study focuses on the European domain for 2007 (tower data) and 2005 (aircraft data) and uses output from high-resolution biosphere models that have been used for regional inversions. Eddy covariance tower fluxes were derived from the FLUXNET ecosystem network (Baldocchi et al., 2001), while aircraft fluxes were acquired within the CarboEurope Regional Experiment (CERES) in southern France. The methods and basic information regarding the models are summarized in Section 2. The results from model-data and model-model comparisons are detailed in Section 3. Discussion and conclusions are following in Section 4.

2 Data and Methods

Appropriate error statistics for the prior error covariance matrix are derived from comparing the output of three biosphere models which are used as priors for regional scale inversions with flux data from the ecosystem network and aircraft. We investigate spatial and temporal autocorrelation structures of the model-data residuals. The temporal autocorrelation is a measure of similarity between residuals at different times but at the same location as a function of the time difference. The spatial autocorrelation refers to the correlation, at a given time, of the model-data residuals at different locations as a function of spatial distance. With this analysis we can formulate and fit an error model such as an exponentially decaying model, which can be directly used in the mesoscale inversion system to describe the prior error covariance.

2.1 Observations

A number of tower sites within the European domain, roughly expanding from -12^{o} E to 35^{o} E and 35^{o} N to 61^{o} N (see also Fig. 1), provide us with direct measurements of CO_2 biospheric fluxes using the eddy covariance technique. This technique computes fluxes from the covariance between vertical wind velocity and CO_2 dry mole fraction (Aubinet et al., 1999). We use Level 3, quality checked, half hourly observations of net ecosystem exchange fluxes (NEE), downloaded from the European Flux Database (www.europe-fluxdata.eu), and listed by site in Table 1. Each site is categorized into different vegetation types (Table 1). A land cover classification is used to label the sites as crop (17 sites), deciduous forest (6), evergreen forest (17), grassland (8), mixed forest (3), savannah (1 site), and shrub land (1). For the current study we focus on observations from these 53

European sites during the year 2007 (Fig. 1).

Additionally, aircraft fluxes are used, obtained with an eddy covariance system installed onboard a SkyArrow ERA aircraft (Gioli et al., 2006). Flights were made in southern France during CERES (CarboEurope Regional Experiment) from May 17 to June 22, 2005. Eddy covariance fluxes were computed on 2-km length spatial windows along transects of 69-km above forest and 78-km above agricultural land, flown 52 and 54 times ~~across entire~~respectively, covering the daily course~~, respectively~~. Exact routes are reported in Dolman et al., 2006.

2.2 Biosphere models

We simulate CO_2 terrestrial fluxes for 2007 with three different biosphere models described in the following. The "Vegetation Photosynthesis and Respiration Model" (VPRM) (Mahadevan et al., 2008), used to produce prior flux fields for inverse studies (Pillai et al., 2012), is a diagnostic model that uses EVI - enhanced vegetation index and LSWI – land surface water index from MODIS, a vegetation map (Synmap, Jung et al., 2006) and meteorological data (temperature at 2m and downward shortwave radiative flux extracted from ECMWF short term forecast fields at 0.25 degrees resolution) to derive gross biogenic fluxes. VPRM parameters controlling respiration and photosynthesis for different vegetation types (a total of four parameters per vegetation type) were optimized using eddy covariance data for the year 2005 collected during the CarboEuropeIP project (Pillai et al., 2012). For this study, VPRM fluxes are provided at hourly temporal resolution and at ~~two~~three spatial resolutions of 1, 10 and ~~10~~50 km (referred to as VPRM1~~ and ~~, VPRM10 and VPRM50). The difference between the 1,10 and ~~10~~50 km resolution version is the aggregation of MODIS indices to either 1, 10 or ~~10~~50 km, otherwise the same meteorology and VPRM parameters are used. At 10 km resolution VPRM uses a tiled approach, with fractional coverage for the different vegetation types, and vegetation type specific values for MODIS indices. For the comparison with the aircraft data VPRM produced fluxes for 2005 at 10 km spatial resolution.

The "Organizing Carbon and Hydrology In Dynamic Ecosystems", ORCHIDEE, model (Krinner et al., 2005) is a process based site scale to global land surface model that simulates the water and carbon cycle using meteorological forcing (temperature, precipitation, humidity, wind, radiation, pressure). The water balance is solved at a half-hourly time step while the main carbon processes (computation of a prognostic LAI, allocation, respiration, turnover) are called on a daily basis. It uses a tiled approach, with fractional coverage for 13 Plant Functional Types (PFT). It has been extensively used as prior information in regional and global scale inversions (Piao et al., 2009, Broquet et al., 2013). For the present simulation, we use a global configuration of the version 1.9.6 of ORCHIDEE, where no parameter has been optimized against eddy covariance data. The model is

forced with 0.5° WFDEI meteorological fields (Weedon et al., 2014). The PFT map is derived from an Olson land cover map (Olson 1994) based on AVHRR remote sensing data (Eidenshink and Faundeen 1994). The fluxes are diagnosed at 3-hourly temporal resolution and at 0.5 degree horizontal resolution.

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 ~~provided~~ at ~~0.25 degrees spatial and~~ 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 ~~optimization of the heterotrophic respiration,~~ EC-data from 2007 was performed since no previous optimization was available ~~were used here~~.

Formatted: Not Highlight

Modeled fluxes for all above mentioned sites have been provided by the different models by extracting the fluxes from the grid cells which encompass the EC station location using vegetation type specific simulated fluxes, i.e. using the vegetation type within the respective grid cell for which the eddy covariance site is assumed representative. For most of the sites the same vegetation type was used for model extraction as long as this vegetation type is represented within the grid-cell. As VPRM uses a tile approach, for two cases ("IT-Amp", "IT-MBo") the represented vegetation type (crop) differ from the actual one (grass). For these cases, the fluxes corresponding to crop were extracted. Fluxes were aggregated to daily fluxes in the following way: first, fluxes from VPRM and 5PM as well as the observed fluxes were temporally aggregated to match with the ORCHIDEE 3-hourly resolution; in a second step we created gaps in the modeled fluxes where no observations were available; the last step aggregated to daily resolution on the premise that a) the gaps covered less than 50% of the day, and b) the number of gaps (number of individual 3-hourly missing values) during day and during night were similar (not different by more than a factor two) to avoid biasing.

Spatial and temporal correlation structures and the standard deviation of flux residuals (model-observations) were examined for daily fluxes over the year 2007. Simulated fluxes from the different models are at different spatial resolution, which makes comparisons difficult to interpret. For the model-data residual analysis, the models VPRM1, VPRM10, ORCHIDEE and 5PM were used. We note that VPRM1 with 1 km resolution is considered compatible when comparing with local measurements. For the model-model analysis we use VPRM50 at 50km resolution when comparing with ORCHIDEE fluxes as both models share the same resolution. VPRM10 is

Formatted: Not Highlight

Formatted: Not Highlight

considered also appropriate for comparisons with 5PM model as they both share same resolution (MODIS LAI resolution of 1 km aggregated to 10 km and meteorological resolution at 0.25 degrees). Following we compare VPRM50 with 5PM to investigate if the different spatial resolution influences the correlation scale as a measure of how trustful might be the derived scales from ORCHIDEE – 5PM comparisons.

For the aircraft analysis, only the VPRM ~~model~~ was used since it is the only ~~having a sufficiently high~~model with spatial resolution (10 km) comparable with aircraft flux footprint and capable of resolving spatial variability in relatively short flight distances. Aircraft NEE data~~-~~, natively at 2 km resolution along the track, have been ~~grouped~~aggregated into 10 km segments ~~along the track~~, to ~~match~~maximize the overlap with the VPRM grid, obtaining 6 grid points in forest transects and 8 in agricultural land transects. ~~Each grid point~~Footprint areas of aircraft fluxes were computed with the analytical model of Hsieh et al. (2000), yielding an average footprint width containing 90% of the flux of~~was sampled 52 times~~ 3.9 km. Averaging also over the different wind directions (perpendicular or parallel to the flight direction), and taking into account the 10 km length of the segments, the area that the aircraft flux data corresponds to, is around ~~in forests, and 54 in agricultural land~~ 23.5 $km^2 \pm 12$ km^2. VPRM fluxes at each aircraft grid cell were extracted, and then linearly interpolated to the time of each ~~flight~~flux observation.

2.3 Analysis of model-observation differences

Observed and modeled fluxes are represented as the sum of the measured or simulated values and an error term, respectively. When we compare modeled to observed data this error term is a combination of model (the prior uncertainty we are interested in) and observation error. Separating the observation error from the model error in the statistical analysis of the model-observation mismatch is not possible; therefore ~~we first neglect the impact of~~e-folding correlation length estimations do include the observation error term ~~on the correlation lengths~~. Nevertheless later in the analysis of model-model differences we assess the impact of the observation error on estimated e-folding correlation lengths.

The tower temporal autocorrelation is computed between the time series of model-observations differences $x_{l,i}$ at site *l* and the same series lagged by a time unit k (Eq. 1), where $\bar{x}$ is the overall mean and N the number of observations:

Formatted: Not Highlight

$$r_l(k) = \frac{\sum_{i=1}^{N-k}(x_{l,i} - \overline{x_l}) \cdot (x_{l,i+k} - \overline{x_l})}{\sum_{i=1}^{N}(x_{l,i} - \overline{x_l})^2} \quad (1)$$

In order to reduce boundary effects in the computation of the autocorrelation at lag times around one year, the one-year flux time series data (model and observations) for each site was replicated four times. This follows the approach of Chevallier et al., (2012), where sites with at least three consecutive years of measurements have been used.

In the current analysis we introduce the all-site temporal autocorrelation by simultaneously computing the autocorrelation for all the observation sites, with M the number of the sites:

$$r(k) = \frac{\sum_{l=1}^{M}\sum_{i=1}^{N-k}(x_{l,i} - \overline{x_l}) \cdot (x_{l,i+k} - \overline{x_l})}{\sum_{l=1}^{M}\sum_{i=1}^{N}(x_{l,i} - \overline{x_l})^2} \quad (2)$$

Temporal correlation scales τ were derived by fitting an exponentially decaying model:

$$r = (1-\alpha) \cdot e^{-\frac{t}{\tau}} \quad (3)$$

Here t is the time lag. For the exponential fit, lags up to 180 days were used (thus the increase in correlations for lag times larger than 10 months is excluded). At zero lag time the correlogram has a value of one (fully correlated), however for even small lag times this drops to values smaller than one, 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 we include the nugget effect variable α.

The aircraft temporal autocorrelation was similarly computed according to Eq. 1 using VPRM, and the same exponentially decaying model (Eq. 3) was used to fit the individual flight flux data~~, distributed along the entire daily course.~~. The temporal interval was limited at 36 days by the experiment duration.

For the spatial analysis the correlation between model-observation residuals at two different locations (i.e sites or aircraft grid points) separated by a specific distance was computed in a way similar to the temporal correlation, and involved all possible pairs of sites and aircraft grid points. Additional data treatment for the spatial analysis was applied to reduce the impact of tower data gaps, as it is possible that the time series for two sites might have missing data at different times. Thus in order to have more robust results, we also examined spatial structures by setting a minimum

threshold of 150 days of overlapping observations within each site pair. Furthermore spatial correlation was investigated for seasonal dependence, where seasons are defined as summer (JJA), fall (SON), winter (DJF for the same year), and spring (MAM). In those cases a different threshold of 20 days of overlapping observations was applied. We note that we do not intend to investigate the errors at the seasonal scale but rather to study if different seasons trigger different error correlation structures.

To estimate the spatial correlation scales, the pairwise correlations were grouped into bins of 100 km distance for towers and 10 km for aircraft data, respectively (*dist*). Following the median for each bin was calculated, and a model similar to Eq. 3 was fitted, but omitting the nugget effect variable:

$$r = e^{-\frac{dist}{d}} \quad (4)$$

The nugget effect could not be constrained simultaneously with the spatial correlation scale *d*, given the relatively coarse distance groups, the fast drop in the median correlation from one at zero distance to small values for the first distance bin combined with ~~the~~ somewhat variations at larger distances. Note that this difference between the spatial and the temporal correlation becomes obvious in the results section 3.

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.

As aircraft fluxes cannot obviously be measured at the same time at different locations, given the relatively short flight duration (about one hour) we treated aircraft flux transect as instantaneous 'snapshots' of the flux spatial pattern across a landscape, neglecting temporal variability that may have occurred during flight.

2.4 Analysis of model-model differences

We evaluate both model-data flux residuals and model-model differences in a sense of pairwise model comparisons, in order to assess if model-model differences can be used as proxy for the prior uncertainty, assuming that ~~the involved prior errors for each model are identical in a sense that they share the same statistics and not correlated.~~models have independent prior errors. For the model-model analysis fluxes derived from the model pairs VPRM50-ORCHIDEE and VPRM10-5PM share the same spatial resolution and therefore are fully comparable. Similar to the model-observation analysis, the statistical analysis gives a combined effect of both model errors. We assess

Formatted: Space Before: 12 pt

the impact in the error structure between model-observation and model-model comparisons caused by the observation error by adding a random measurement error to each model-model comparison. This error has the same characteristics as the observation error which is typically associated with eddy covariance observations; the error characteristics were derived from the paired observation approach (Richardson et al., 2008). Specifically, we implement the flux observation error as a random process (white noise) with a double-exponential probability density function. This can be achieved by selecting a random variable u drawn from the uniform distribution in the interval (-1/2, 1/2), and then applying Eq. 5 to get a Laplace distribution (also referred to as the double-exponential)

$$x = \mu - \frac{\sigma}{\sqrt{2}} \cdot \mathrm{sgn}(u) \cdot \ln\left(1 - 2 \cdot |u|\right) \quad (5)$$

Here μ=0 and σ is the standard deviation of the double-exponential. We compute the σ according to Richardson et al., (2006) as

$$\sigma = \alpha_1 + \alpha_2 \cdot |F| \quad (6)$$

where F is the flux and α_1, α_2 are scalars specific to the different vegetation classes. Lasslop et al., (2008) found that the autocorrelation of the half hourly random errors is below 0.7 for a lag of 30 min, and falls off rapidly for longer lag times. Thus we assume the standard deviation for hourly random errors to be comparable with the half hourly errors. Hourly random errors specific for each reference model are generated for each site individually. With ORCHIDEE as reference with fluxes at 3-hourly resolution, a new ensemble of 3-hourly random noise was generated with σ for the 3-hourly errors modified (divided by the square root of three to be coherent with the hourly σ). As both modeled and observed fluxes share the same gaps, the random errors were aggregated to daily resolution, with gaps such to match those of observed fluxes. Finally the daily random errors were added to the modeled fluxes.

3 Results

3.1 Model-data comparison for tower and aircraft fluxes

Observed daily averaged NEE fluxes, for all ground sites and the full time-series, yield a standard deviation of 3.01 μmol m^{-2} s^{-1}, while the modeled fluxes were found to be less spatially varying and with a standard deviation of 2.84, 2.80, 2.53, 2.64 μmol m^{-2} s^{-1} for VPRM10, VPRM1, ORCHIDEE and 5PM respectively.

The residual distribution of the models defined as the difference between simulated and observed daily flux averages for the full year 2007 was found to have a standard deviation of 2.47, 2.49, 2.7 and 2.25 μmol m^{-2} s^{-1} for VPRM10, VPRM1, ORCHIDEE and 5PM respectively. Those values are

only slightly smaller than the standard deviations of the observed or modeled fluxes themselves~~, which~~. This fact is in line with the generally low fraction of explained variance with r-square values of 0.31, 0.27, 0.12, and 0.25 for VPRM10, VPRM1, ORCHIDEE and 5PM respectively. 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~~, and~~. Further, the differences between site-specific to the overall r-square values indicate limitation of the models to reproduce observed spatial (site to site) differences. Figure 2 shows the correlation between modeled and observed daily fluxes as a function of the vegetation type characterizing each site. All models exhibit a significant scatter of the correlation ranging from 0.9 for some sites to 0 or even negative correlation for some crop sites, with the highest correlation coefficients for deciduous and mixed forest.

The distribution is biased by -0.07, 0.26, 0.92 and 0.25 µmol m^{-2} s^{-1} for VPRM10, VPRM1, ORCHIDEE and 5PM, respectively. Figure 3 shows the ~~bias~~ distribution of bias (defined as modeled – observed fluxes) for different vegetation types. Bias and standard deviation seem to depend on the vegetation type for all models, without a clear general pattern.

The temporal autocorrelation was calculated for model-data residuals for each of the flux sites ("site data" in Fig. 4), but also for the full dataset ("all-site" in Fig. 4). The "all site" temporal autocorrelation structure of the residuals appears to have the same pattern for all models. It decays smoothly for time lags up to 3 months and then remains constant near to 0 or to some small negative values. The temporal autocorrelation increases again for time lags > 10 months, which is caused by the seasonal cycle. ~~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.~~ These temporal autocorrelation results agree with the findings of Chevallier et al., (2012).

The exponentially decaying model in Eq. 3 was used to fit the data. At zero separation time (t=0) the correlogram value is 1. However the correlogram exhibits a nugget effect (~~i.e. a value of~~

Formatted: Superscript

~~0.39~~values ranging from 0.31 to 0.48 for ~~VPRM10~~the different models) as a consequence of an uncorrelated part of the error. For the current analyses we fit the exponential model with an initial correlation different from 1. The fit has a root mean square error ~~of 0.041.~~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%. The e-folding time (defined as the lag required for the correlation to decrease by a factor of e (63% of its initial value) ranged between 26-70 days for the different models (see Table 2). Specifically, for VPRM10 and VPRM1 the e-folding time is 32 and 33 days respectively (30-34 days within 95% confidence interval for both). Confidence intervals for the e-folding time were calculated by computing the confidence intervals of the parameter in the fitted model. For ORCHIDEE best fit was 26 days (23-28 days within 95% confidence interval). In contrast, 5PM yields a significantly longer correlation time between 65-75 days (95% confidence interval) with the best fit being 70 days.

For a number of sites a large model-data bias was found. In order to assess how the result depends on individual sites where model-data residuals are more strongly biased the analysis was repeated under exclusion of sites with an annual mean of model-data flux residuals larger than 2.5μmol/m^{2}s. This threshold value is roughly half of the most deviant bias. In total 9 sites ("CH-Lae", "ES-ES2", "FR-Pue", "IT-Amp", "IT-Cpz", "IT-Lav", "IT-Lec", "IT-Ro2", "PT-Esp") across all model-data residuals were excluded. From these sites "CH-Lae" appears to have serious problems related to the steep terrain, where the basic assumptions made for eddy covariance flux measurements are not ~~well~~ applicable (Göckede et al., 2008). The rest of the sites are located in the Mediterranean region, and suffer from summer drought according to the Köppen-Geiger climate classification map (Kottek et al., 2006); in those cases a large model - data bias is expected as existing models tend to have difficulties to estimate carbon fluxes for drought prone periods (Keenan et al., 2009). The model-data bias at those sites does not necessarily exceed the abovementioned threshold of 2.5μmol/m^{2}s simultaneously for each individual model, but a larger bias than the average was detected. After exclusion of those sites the temporal correlation times were found to be between 33-35 days within 95% confidence interval for 5PM with the best fit value being 34 days. The rest of the models had temporal e-folding times of 27, 29 and 24 days (1st row of Table 2), while the all-site correlation remains positive for lags <76, <79, <66 days for VPRM10, VPRM1 and ORCHIDEE respectively. Some weak negative correlations exist, with a minimum value of -0.06, -0.02, -0.09, -0.005 for VPRM10, VPRM1, ORCHIDEE and 5PM respectively.

The temporal correlation of differences between VPRM10 and aircraft flux measurements could be computed for time intervals up to 36 days (Fig. 5) corresponding to the duration of the campaign.

The correlation shows an exponential decrease, and levels off after about 25 days with an e-folding correlation time of 13 days (range of 10 – 16 days within the 95% confidence interval). Whilst the general behavior is consistent with results obtained for VPRM-observation residuals for flux sites, the correlation time is two times smaller.

Regarding spatial error correlations, results for all models show a dependence on the distance between pairs of sites. The median correlation drops within very short distances (Fig. 6). Fitting the simple exponentially decaying model (Eq. 4) to the correlation as a function of distance we find an e-folding correlation length d of 40, 37, 32 and 31 km with a root mean square error (RMSE) of 0.14, 0.09, 0.05 and 0.07 for VPRM10, VPRM1, ORCHIDEE and 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%. Spatial correlation scales are also computed for a number of different data selections (cases) in addition to the standard case shown in Fig. 6 (case S): using only pairs with at least 150 overlapping days of non-missing data (case S*), using only pairs with identical PFT (case I), using only pairs with different PFT (case D), and using only pairs with at least 150 overlapping days for the D and I cases (cases D*, I*). The results for these cases are summarized in Fig. 7. Also 95% confidence intervals were computed, and the spread spatial correlation was found to be markedly more critical than for the time correlations. Note that for some cases the 2.5%-ile (the lower bound of the confidence interval) hit the lower bound for correlation lengths (0 km). The e-folding correlation lengths are similar for each of the models: this also means that no dependence on the spatial resolution was detectable. Further we examined also the spatial autocorrelation from VPRM50-data residuals with no significant difference compared to previous results.

Interestingly, if we restrict the analysis to pairs with at least 150 overlapping days between site pairs, larger correlation scales are found (case S* in Fig. 7). Considering only pairs with different PFT (case D), consistently, all e-folding correlation lengths are found to be smaller compared to the standard case (S). This is expected to a certain degree, as model errors should be more strongly correlated between sites with similar PFTs than between sites with different PFTs. By considering only pairs within the same vegetation type (case I) we observe a significant increase of the e-folding correlation length relative to case S for VPRM at 10 and 1 km resolution to values of 432 km and 305 km, respectively. The ORCHIDEE and 5PM models show some (although not significant) increase in e-folding correlation length. Restricting again the analysis to pairs with at least 150 overlapping days for the D and I cases (D*, I*) we observe an increase of the e-folding correlation lengths that is however significant only for VPRM at 10 and 1 km.

Seasonal dependence of the e-folding correlation lengths for at least 20 overlapping days per season and for all site-pairs is also shown in Fig. 7. VPRM showed somewhat longer correlation lengths

during spring and summer, ORCHIDEE had the largest lengths occurring during summer and fall, and 5PM e-folding correlation lengths show slightly enhanced values during spring and summer. However, none of these seasonal differences are significant with respect to the 95% confidence interval.

The spatial error correlation between VPRM10 model and aircraft fluxes measured during May-June along continuous transects at forest and agriculture land use (Fig. 8) shows an exponential decay up to the maximum distance that was encompassed during flights (i.e. 70 km). Of note is that only two measurements were available at 60 km distance and none for larger distances making it difficult to identify where the asymptote ~~lying~~lies. Nevertheless fitting the decay model (Eq. 4) leads to d = 35km (26 – 46 km within the 95% confidence interval), which is in good agreement with the spatial correlation scale derived for VPRM10 using flux sites during both spring and summer (Fig. 7).

3.2 Model-model comparison

We investigate the model-model error structure of NEE estimates by ~~substituting~~ replacing the observed fluxes which were used as reference, with simulated fluxes from all the biosphere models. Note that for consistency with the model-data analysis, the simulated fluxes contained the same gaps as the observed flux time series. The e-folding correlation time is found to be slightly larger compared to the model-data correlation times, for most of the cases. An exception ~~are~~ is the 5PM-VPRM10 ~~and 5PM VPRM1~~ pair~~s~~ which ~~they~~ produced remarkably larger correlation time~~s~~ (Table 2). Specifically, VPRM~~1~~50-ORCHIDEE and VPRM10-5PM residuals show correlation times of ~~30~~28 days (range between 2~~7~~4-3~~3~~2 days within 95% confidence interval) and 131 (range between 128-137 days within 95% confidence interval), respectively. Significantly different e-folding correlation times are found for VPRM50-5PM compared to VPRM10-5PM with correlation times of 52 days (range between 49-56 days within 95% confidence interval). ~~-~~Repeating the analysis excluding sites with residual bias larger than 2.5µmol/m^{2}s, correlation times of ~~31~~ 28 and 100 days for VPRM~~1~~50-ORCHIDEE and VPRM10-5PM are found, respectively. If we use ORCHIDEE-5PM pair ~~as reference~~ the e-folding correlation time~~s~~ ~~are~~found to be ~~30, 28 and~~ 38 days (range between 35-41 days within 95% confidence interval) ~~with respect to VPRM10, VPRM1 and 5PM comparisons respectively~~.

Formatted: Not Highlight

Although the e-folding correlation times show but minor differences compared to the model-data residuals, this is not the case for the spatial correlation lengths (Fig. 9). The standard case (S) was applied for the annual analysis, with no minimum number of days with overlapping non-missing data for each site within the pairs. Taking VPRM50 as reference, much larger e-folding correlation

lengths of 3712 km with a range of 2086-4621 km within 95% confidence interval yielded for VPRM510-ORCHIDEE comparisons, and 31066 km for VPRM510-5PM were found. However VPRM10-5PM analysis which is also considered appropriate in terms of the spatial resolution compatibility contrary to the VPRM50-5PM pair, is in good agreement with VPRM50-ORCHIDEE spatial scale (230-440 km range within 95% confidence interval with the best fit being 335 km). With ORCHIDEE as reference the e-folding correlation length for the ORCHIDEE-5PM comparison is 278 276 km with a range of 183-360 km within 95% confidence interval. SeasonalHowever the later correlation length might be affected by the different spatial resolution as the difference between VPRM10 and VPRM50 against 5PM suggests. Seasonal e-folding correlation lengths, using a minimum of 20 days overlap in the site-pairs per season (Fig. 9), are also significantly larger compared with those from the model-data analysis.

When we add the random measurement error to the modeled fluxes used as reference (crosses in Fig. 9), we observe only slight changes in the annual e-folding correlation lengths, without a clear pattern. The correlation lengths show a random increase or decrease but limited up to 6%. Interestingly, the seasonal e-folding correlation lengths for most of the cases show a more clear decrease. For example, the correlation length of the VPRM10-5PM residuals during winter, decreases by 22% or even more for spring season. Despite this decrease, the e-folding seasonal correlation lengths remain significantly larger in comparison to those from the model-data analysis. Overall, all models when used as reference show the same behavior with large e-folding correlation lengths that mostly decrease slightly when the random measurement error is included. Although the random measurement error was added as "missing part" to the modeled fluxes to better mimic actual flux observations, it did not lead to correlation lengths similar to those from the model-data residual analysis. To investigate if a larger random measurement error could cause spatial correlation scales in model-model differences, we repeated the analysis with artificially increased random measurement error (multiplying with a factor between 1 and 15). Only for very large random measurement errors did the model-model e-folding correlation lengths start coinciding with those of the model-data residuals (Fig. 10).

4. Discussion and conclusions

We analyzed the error structure of a-priori NEE uncertainties derived from a multi-model – data comparison by comparing fluxes simulated by three different vegetation models to daily averages of observed fluxes from 53 sites across Europe, categorized into 7 land cover classes. The different models showed comparable performance with respect to reproducing the observed fluxes; we found mostly insignificant differences in the mean of the residuals (bias) and in the variance. Site-specific

Formatted: Not Highlight

Formatted: Not Highlight

Formatted: Not Highlight

correlations between simulated and observed fluxes are significantly higher than overall correlations for all models, which suggest that the models struggle with reproducing observed spatial flux differences between sites. Furthermore, the site-specific correlations reveal a large spread even within the same vegetation class, especially for crops (Fig. 2). This is likely due to the fact that none of the models uses a specific crop model that differentiates between the different crop types and their phenology. The models using remotely sensed vegetation indices (VPRM and 5PM) better capture the phenology; ORCHIDEE is the only model that differentiates between C_3 and C_4 plants, but shows the largest spread in correlation for the crop. Differences in correlations between the different vegetation types were identified for all the biosphere models, however it must be noted that the number of sites per vegetation type is less than 10 except for crop and evergreen forests.

Model-data flux residual correlations were investigated to give insights regarding prior error temporal scales which can be adopted by atmospheric inversion systems. Whilst fluxes from ORCHIDEE model are at much coarser resolution compared to the representative area from the flux measurements, VPRM1 fluxes (1 km resolution and only the meteorology at 25 km) are considered appropriate for the comparisons. Despite the scale mismatch results are in good agreement across all model-data pairs.

Exponentially decaying correlation models are a dominant technique among atmospheric inverse studies to represent temporal and spatial flux autocorrelations (Rödenbeck et al., 2009, Broquet et al., 2011, Broquet et al., 2013). However, regarding the temporal error structure we need to note the weakness of this model to capture the slightly negative values at 2-10 months lags and, more importantly, the increase in correlations for lag times larger than about 10 months. Error correlations were parameterized differently by Chevalier et al., (2012) where the prior error was investigated without implementing it to atmospheric inversions. Polynomial and hyperbolic equations were used to fit temporal and spatial correlations respectively. Nevertheless, we use here e-folding lengths not only for their simplicity in describing the temporal correlation structure with a single number, but also because this error model ensures a positive definite covariance matrix (as required for a covariance). This is crucial for atmospheric inversions as otherwise negative, spatially and temporally integrated uncertainties may be introduced. In addition it can keep the computational costs low; this is because the hyperbolic equation has significant contributions from larger distances: for the case of the VPRM1 model, at 200 km distance the correlation according to Chevallier et al., hyperbolic equation is 0.16, compared to 0.004 for the exponential model. As a consequence, more none-zero elements are introduced to the covariance matrix, which increases computational costs in the inversion systems. Using the same hyberbolic equation for the spatial correlation, d-values of 73, 39, 12 and 20 km were found with a RMSE of 0.11, 0.07, 0.05, 0.07 for

VPRM10, VPRM1, ORCHIDEE and 5PM respectively. A similar RMSE was found when using the exponential (0.14, 0.09, 0.05 and 0.07), indicating similar performance of both approaches with respect to fitting the spatial correlation.

Autocorrelation times were found to be in line with findings of Chevallier et al., (2012). The model-data residuals were found to have an e-folding time of 32 and 26 days for VPRM and ORCHIDEE respectively, and 70 days for 5PM. This significant difference appears to have a strong dependence on the set of sites used in the analysis. Excluding nine sites with large residual bias, the autocorrelation time from the 5PM-data residuals drastically decreased and became coherent with the times of the other biosphere models. The all-models and all-sites autocorrelation time was found to be 39 days, which reduces to 30 days (28-31 days within 95% confidence interval), when excluding the sites with large residual bias, coherent with the single model times. From the model-model residual correlation analysis, the correlation time appear to be consistent with the above-mentioned results, and lies between 28 and 46 days for most of the ensemble members. However model-model pairs consisting of the VPRM and 5PM models produced larger times up to 131 days; omitting sites with large residual biases this is reduced to 100 days (99-105 days within 95% confidence interval). This finding could be attributed to the fact that despite the conceptual difference between those models, they do have some common properties. Both models were optimized against eddy covariance data although for different years (2005 and 2007 respectively), while no eddy covariance data were used for the optimization of ORCHIDEE. In addition, VPRM and 5PM both use data acquired from MODIS, although they estimate photosynthetic fluxes by using different indices of reflectance data. Summarizing the temporal correlation structure, it appears reasonable to a) use same error correlation in atmospheric inversions regardless which biospheric model is used as prior, b) use an autocorrelation length of around 30 days.

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

Formatted: Not Highlight

Formatted: Not Highlight

from MODIS retrievals (Friedl et al., 2002).

Although the estimated spatial scales are shorter than the spatial resolution that we are solving for (100 km bins), the autocorrelation analysis of aircraft measurements made during CERES supports the short scale correlations. These measurements have the advantage of providing continuous spatial flux transects along specific tracks that were sampled routinely (in this case over period of 36 days at various times of the day), thus resolving flux spatial variability also at small scales, where pairs of eddy covariance sites may not be sufficiently close. On the other hand, aircraft surveys are necessarily sporadic in time. Of note is that the ~~impact of the~~ eddy covariance observation error ~~on the estimated prior error and its structure had to be initially neglected as it is not possible to subtract the unknown error from the observations. However we do not expect this to have a~~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. We note that the current analysis focuses to daily time scale and therefore the error statistics with respect to the estimated spatial and temporal e-folding correlation lengths are valid for such scales.

Model-data residual e-folding correlation lengths show a clear difference, between the cases where pairs only with different (D) or identical (I) PFT were considered, with the latter resulting in longer correlation lengths, but only identified for the VPRM model at both resolutions. The "D" case has slightly shorter lengths for all models than the standard case (S). One could argue that as VPRM uses PFT specific parameters that were optimized against 2005 observations, the resulting PFT specific bias could lead to longer spatial correlations. However ORCHIDEE and 5PM also show comparable biases (Fig. 3), but long correlation scales were not found. Moreover we repeated the spatial analysis after subtracting the PFT specific bias from the fluxes, and the resulting correlation lengths showed no significant change. The impact of data gaps was also investigated by setting a threshold value of overlapping observations between site pairs. Setting this to 150 days results in an increase for the "S" case up to 60 km, but only for the VPRM model. For the "D" and "I" cases when setting the same threshold value (D^* and I^*) we only found an insignificant increase, indicating that data gaps are hardly affecting the "D" and "I" cases. These findings suggest that high-resolution diagnostic models might be able to highlight the increase of the spatial correlation length between identical PFTs vs. different PFTs. Note that the Chevallier et al., (2012) study concluded that assigning vegetation type specific spatial correlations is not justified, based on comparisons of eddy covariance observations with ORCHIDEE simulated fluxes. The current study could not further investigate this dependence, as the number of pairs within a distance bin is not large enough for statistical analyses, when using only sites within the same PFT. With respect to the seasonal analysis, spatial correlations are at the same range among all models and seasons.

Although in some cases (VPRM10 and VPRM1 spring) the scales are larger, they suffer from large uncertainties. Hence, implementing distinct and seasonally dependent spatial correlation lengths in inversion systems cannot be justified.

The analysis of model-model differences did not reproduce the same spatial scales as those from the model-data differences, but instead spatial e-folding correlation lengths were found to be dramatically larger. Adding a random measurement error to the modeled fluxes used as reference slightly reduced the spatial correlation lengths to values ranging from ~~86~~ 278 to ~~320~~ 1058 km. Even when largely inflating the measurement error, the resulting spatial correlation lengths (Fig. 10) still do not approach those derived from model-data residuals. Only when the measurement error is scaled up by a factor of 8 or larger (which is quite unrealistic as this corresponds to a mean error of 1.46 µmol m^{-2} s^{-1} or larger, which is comparable to the model-data mismatch where a standard deviation of around 2.5 µmol m^{-2} s^{-1} was found), the e-folding correlation lengths are consistent with those based on model-data differences. Whilst the EC observations are sensitive to a footprint area of about 1 km^2, the model resolution is too coarse to capture variations at such a small scale. This local uncorrelated error has not been taken into account by the analysis of model-data residuals as the error model could not be fitted with a nugget term included, favoring therefore smaller correlation scales. The analysis of differences between two coarser models ~~(excluding VPRM at 1 km for the reason mentioned in the next paragraph)~~ does not involve such a small scale component, thus resulting in larger correlation scales. This would suggest that for inversion studies targeting scales much larger than the eddy covariance footprint scale, the statistical properties of the prior error should be derived from the model-model comparisons.

~~A special case in the context of the model-model study is the comparison between VPRM1 and VPRM10, which is the only case that produced short spatial correlation scales. These two models only differ in the spatial resolution of MODIS indices EVI and LSWI (1 vs. 10 km). Thus differences between those two models are only related to variability of these indices at scales below 10 km, which is not expected to show any spatial coherence. Indeed the results show only very short correlation scales (Fig. 9) with an exception during fall, however there the uncertainty is also large.~~

The large e-folding correlation lengths yielded from this model-model residual analysis suggest that the models are more similar to each other than to the observed terrestrial fluxes, at least on spatial scales up to a few hundred kilometers regardless of their conceptual differences. This might be expected ~~at~~to some extent due to elements that the models share. Respiration and photosynthetic fluxes are strongly driven by temperature and downward radiation, respectively, and those meteorological fields have significant commonalities between the different models. VPRM and

5PM both use temperature and radiation from ECMWF analysis and short-term forecasts. Also the WFDEI temperature and radiation fields used in ORCHIDEE are basically from the ERA-Interim reanalysis, which also involves the integrated forecasting system (IFS) used at ECMWF (Dee et al., 2011). Regarding the vegetation classification all models are site specific and therefore are using the same PFT for each corresponding grid-cell. Photosynthetic fluxes are derived with the use of MODIS indices in VPRM (EVI and LSWI) and in 5PM (LAI and albedo).

Using full flux fields from the model ensemble (rather than fluxes at specific locations with observation sites only) to assess spatial correlations in model-model differences is not expected to give significantly different results, as the sites are representative for quite a range of geographic locations and vegetation types within the domain investigated here.

The current study intended to provide insight on the error structure that can be used for atmospheric inversions. Typically, inversion systems have a pixel size ranging from 10 to 100 km for regional and continental inversions, and as large as several degrees (hundreds of km) for global inversions. If a higher resolution system assumes such small-scale correlations (as those found in the current analysis), in the covariance matrix, of note is that this leads to very small prior uncertainties when aggregating over large areas and over longer time periods. To aggregate the uncertainty to large temporal and spatial scales, we used the following equation (after Rodgers, 2000):

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

Field Code Changed

Where "×" denotes matrix multiplication, Q_c 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). For example, with a 30 km spatial and a 40 day temporal correlation scale, annually and domain-wide (Fig. 1) aggregated uncertainties are around 0.06 GtC. This is about a factor ten smaller than uncertainties typically used e.g. in the Jena inversion system (Rödenbeck et al., 2005). 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. In addition, the aggregated uncertainties using the VPRM10-ORCHIDEE error structure (32 days and 320 km temporal and spatial correlation scales) are found to be 0.46 GtC/y which is also much smaller than the difference between VPRM10 (NEE= - 1.45 GtC/y) and ORCHIDEE (NEE= - 0.2 GtC/y), when aggregated over the domain shown in Fig. 1. Although this analysis does capture the dominating spatiotemporal correlation scale in the error structure, it fails in terms of the error budget, suggesting that also other parts of the error structure are important as well. Therefore additional degrees of freedom (e.g. for a large-scale bias) need to be introduced in the inversion systems to fully describe the error structure.

Field Code Changed

Formatted: Font: Italic

Formatted: Font: Italic

~~Exponentially decaying correlation models are a dominant technique among atmospheric inverse studies to represent temporal and spatial flux autocorrelations (Rödenbeck et al., 2009, Broquet et al., 2011, Broquet et al., 2013). However, regarding the temporal error structure we need to note the weakness of this model to capture the slightly negative values at 2-10 months lags and, more importantly, the increase in correlations for lag times larger than about 10 months. Error correlations were parameterized differently by Chevalier et al., (2012) where the prior error was investigated without implementing it to atmospheric inversions. Polynomial and hyperbolic equations were used to fit temporal and spatial correlations respectively. Nevertheless, we use here e-folding lengths not only for their simplicity in describing the temporal correlation structure with a single number, but also because this error model ensures a positive definite covariance matrix (as required for a covariance). This is crucial for atmospheric inversions as otherwise negative, spatially and temporally integrated uncertainties may be introduced. In addition it can keep the computational costs low; this is because the hyperbolic equation has significant contributions from larger distances: for the case of the VPRM1 model, at 200 km distance the correlation according to Eq. 7 is 0.16, compared to 0.004 for the exponential model. As a consequence, more none-zero elements are introduced to the covariance matrix, which increases computational costs in the inversion systems. Using the parameterization from Eq. 7 for the spatial correlation, d values of 73, 39, 12 and 20 km were found with a RMSE of 0.11, 0.07, 0.05, 0.07 for VPRM10, VPRM1, ORCHIDEE and 5PM respectively. A similar RMSE was found when using the exponential (0.14, 0.09, 0.05 and 0.07), indicating similar performance of both approaches with respect to fitting the spatial correlation.~~

Whilst temporal scales found from this study have already been used in inversion studies, this is not the case to our best knowledge for the short spatial scales. The impact of the prior error structure derived from this analysis, on posterior flux estimates and uncertainties will be assessed in a subsequent paper. For that purpose, findings from this study are currently implemented in three different regional inversion systems aiming to focus on network design for the ICOS atmospheric network.

Acknowledgements.

The research leading to these results has received funding from the European Community's Seventh Framework Program ([FP7/2007-2013]) under grant agreement n°313169 (ICOS-INWIRE) and under grant agreement n°283080 (GEOCARBON). The EC data used in this study were funded by the European Community's sixth and seventh framework program and by national funding agencies. In the following sites are listed, sorted by project/funding agency: “BE-Bra”, “BE-Lon”, “BE-Vie”,

"CH-Lae", "CH-Oe1", "CH-Oe2", "CZ-BK1", "DE-Geb", "DE-Gri", "DE-Hai", "DE-Kli", "DE-Tha", "DK-Lva", "ES-ES2", "ES-LMa", "FI-Hyy", "FR-Fon", "FR-Hes", "FR-LBr", "FR-Lq1", "FR-Lq2", "FR-Pue", "IT-SRo", "NL-Dij", "NL-Loo", "PT-Esp", "PT-Mi2", "SE-Kno", "SE-Nor", "SE-Sk1", "SK-Tat", "UK-AMo", "UK-EBu" funded by CarboEuropeIP (grand agreement n° GOCE-CT-2003-505572); "UK-AMo", "UK-EBu" also co-funded by EU FP7 ECLAIRE project and NERC-CEH; "IT-BCi", "IT-Cas", "IT-Lav", "IT-LMa" funded by CarboItaly (IT-FISR), "IT-Amp", "IT-Col", "IT-Cpz", "IT-MBo", "IT-Ren", "IT-Ro2" co-funded by CarboEuropeIP and by CarboItaly, "CH-Cha", "CH-Dav", "CH-Fru" co-funded by CarboExtreme (grant agreement n° 226701) and GHG-Europe (grant agreement n° 244122), "ES-Agu" funded by GHG-Europe, "FR-Mau" co-funded by CNRM/GAME (METEO-FRANCE, CNRS), CNES and ONERA. We also acknowledge funding agencies for the sites "FR-Aur", "FR-Avi", "HU-Mat".

The authors would like to thank Anna Michalak and the anonymous reviewer whose insightful comments helped the development of the manuscript.

References

Albergel, C., Calvet, J.-C., Gibelin, A.-L., Lafont, S., Roujean, J.-L., Berne, C., Traullé, O., and Fritz, N.: Observed and modelled ecosystem respiration and gross primary production of a grassland in southwestern France, Biogeosciences, 7, 1657-1668, doi:10.5194/bg-7-1657-2010, 2010.

Allard, V., Ourcival, J.-M., Rambal, S., Joffre, R., and Rocheteau, A.: Seasonal and annual variation of carbon exchange in an evergreen Mediterranean forest in southern France, Global Change Biology, 14(4), 714-725, 2008.

Ammann, C., Spirig, C., Leifeld, J., and Neftel, A.: Assessment of the nitrogen and carbon budget of two managed grassland fields, Agriculture, Ecosystems and Environment, 133, 150–162, 2009.

Aubinet, M., Grelle, A., Ibrom A., Rannik Ü., Moncrieff J., Foken T., Kowalski A.-S., Martin P.-H., Berigier P., Bernhofer C., Clement R., Elbers J., Granier A., Grünwald T., Morgenstern K., Pilegaard K., Rebmann C., Snijders W., Valentini, R. and Vesala, T.: Estimates of the Annual Net Carbon and Water Exchange of Forests: The EUROFLUX Methodology, Adv. Ecol. Res. 30 , 113-175, 2000.

Aubinet, M., Chermanne, B., et al.,: Long term carbon dioxide exchange above a mixed forest in the Belgian Ardennes, Agricultural and Forest Meteorology 108(4): 293-315, 2001.

Baldocchi, D., et al., FLUXNET: A new tool to study the temporal and spatial variability of ecosystem–scale carbon dioxide, water vapor, and energy flux densities, Bull. Am. Meteorol. Soc., 82, 2415–2434, 2001.

Barcza, Z., Weidinger, T., Csintalan., Zs., Dinh, N.-Q., Grosz, B., Tuba, Z.: The carbon budget of a semiarid grassland in a wet and a dry year in Hungary, AGR ECOSYST ENVIRON 121: (1-2)21-29, 2007.

Bates, D.-M. and Watts, D.-G.: Nonlinear Regression Analysis and Its Applications, Wiley, 1988

Bousquet, P., Ciais, P., Peylin, P., Ramonet, M. and Monfray, P.: Inverse modeling of annual atmospheric CO_2 sources and sinks: 1. Method and control inversion, Journal of Geophysical Research: Atmospheres (1984--2012) 104, 26161-26178, 1999.

Broquet, G., Chevallier, F., Breon, F.-M., Kadygrov, N., Alemanno, M., Apadula, F., Hammer, S., Haszpra, L., Meinhardt, F., Morgui, J.-A. and et al.,: Regional inversion of CO_2 ecosystem fluxes from atmospheric measurements: reliability of the uncertainty estimates, Atmospheric Chemistry and Physics 13, 9039-9056, doi: 10.5194/acp-13-9039-2013, 2013.

Broquet, G., Chevallier, F., Rayner, P., Aulagnier, Cé., Pison, I., Ramonet, M., Schmidt, M., Vermeulen, A.-T. and Ciais, P.: A European summertime CO_2 biogenic flux inversion at mesoscale from continuous in situ mixing ratio measurements, Journal of Geophysical Research: Atmospheres (1984--2012) 116, 2011.

Carouge, C., Bousquet, P., Peylin, P., Rayner, P.-J. and Ciais, P.: What can we learn from European continuous atmospheric CO_2 measurements to quantify regional fluxes - Part 1: Potential of the 2001 network, Atmos. Chem. Phys 10, 3107–3117, 2010.

Casals, P., Lopez-Sangil, L., Carrara., A., Gimeno, C., and Nogues, S.: Autotrophic and heterotrophic contributions to short-term soil CO_2 efflux following simulated summer precipitation pulses in a Mediterranean dehesa, Global Biogeochemical Cycles, 25 (3), doi: 10.1029/2010GB003973, 2011.

Chevallier F., Viovy N., Reichstein M., and Ciais, P.: On the assignment of prior errors in Bayesian inversions of CO_2 surface fluxes, doi: 10.1029/2006GL026496, 2006.

Chevallier F., Wang T., Ciais P., Maignan F., Bocquet M., Altaf A.-M., Cescatti A., Chen J., Dolman A.-J., Law B,-E. and et al.,: What eddy-covariance measurements tell us about prior land flux errors in CO_2 flux inversion schemes, Global Biogeochemical Cycles 26, doi: 10.1029/2010GB003974, 2012.

Chiesi M., Fibbi L., Genesio L., Gioli B., Magno R., Maselli F., Moriondo M., Vaccari F.,-P: Integration of ground and satellite data to model Mediterranean forest processes. International Journal of Applied Earth Observation and Geoinformation, 13, 504–515, doi: 10.1016/j.jag.2010.10.006, 2011.

Ciais, P., Peylin, P., and Bousquet, P.: Regional biospheric carbon fluxes as inferred from atmospheric CO_2 measurements, Ecological Applications 10, 1574-1589, 2000.

Dee, D.-P., Uppala, S.-M., Simmons, A.-J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M.-A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A.-C.-M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A.-J., Haimberger, L., Healy, S.-B., Hersbach, H., Hólm, E.-V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., Mcnally, A.-P., Monge-Sanz, B.-M., Morcrette, J.-J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N. and Vitart, F.: The ERA-Interim reanalysis: configuration and performance of the data assimilation system, Q.J.R. Meteorol. Soc., 137(656), 553–597, doi:10.1002/qj.828, 2011.

Delpierre, N., Soudani, K., François, C., Köstner, B., Pontailler, J.-Y., Nikinmaa, E., Misson, L., Aubinet, M., Bernhofer, C., Granier, A., Grünwald, T., Heinesch, B., Longdoz, B., Ourcival, J.- M., Rambal, S., Vesala, T., and Dufrêne, E.: Exceptional carbon uptake in European forests during the warm spring of 2007: a data-model analysis, Glob. Change Biol., 15, 1455–1474, 10 doi:10.1111/j.1365-2486.2008.01835.x, 2009.

Dietiker, D., Buchmann, N., Eugster, W.: Testing the ability of the DNDC model to predict CO_2 and water vapour fluxes of a Swiss cropland site, Agriculture, Ecosystems and Environment, 139, 396-401, 2010.

Dolman, A.-J., Noilhan, J., Durand, P., Sarrat, C., Brut, A., Piquet, B., Butet, A., Jarosz, N., Brunet, Y., Loustau, D., Lamaud, E., Tolk, L., Ronda, R., Miglietta, F., Gioli, B., Magliulo, V., Esposito, M., Gerbig, C., Korner, S., Glademard, R., Ramonet, M., Ciais, P., Neininger, B., Hutjes, R-W.-A., Elbers, J.-A., Macatangay, R., Schrems, O., Perez-Landa, G., Sanz, M.-J., Scholz, Y., Facon, G., Ceschia, E., Beziat, P.: The CarboEurope Regional Experiment Strategy. Bulletin of the American Meteorological Society, 87(10), 1367-1379. (2006).

Eidenshink, J. C. and Faundeen, J. L.: The 1 km AVHRR global land data set: first stages in implementation, Int. J. Remote Sens., 15(17), 3443–3462, 1994.

Etzold, S., Buchmann, N., and Eugster, W.: Contribution of advection to the carbon budget measured by eddy covariance at a steep mountain slope forest in Switzerland, Biogeosciences, 7, 2461-2475, doi: 10.5194/bg-7-2461-2010, 2010.

Farquhar, G., G., von Caemmerer, S., Berry, J.-A.: A biochemical model of photosyntetic CO_2 assimilation in leaves of C_3 species. Planta 149, 78-90, 1980.

Friedl, M. A., McIvera, D. K., Hodgesa, J. C. F., Zhanga, X. Y., Muchoneyb, D., Strahlera, A. H., Woodcocka, C. E., Gopala, S., Schneidera, A., Coopera, A., Baccinia, A., Gaoa, F., and Schaafa, C.: Global land cover mapping from MODIS: algorithms and early results, Remote sensing of

environment 83, 287-302, doi:10.1016/S0034-4257(02)00078-0, 2002.

Gabriel, P., Gielen, B., Zona, D., Rodrigues, A., Rambal, S., Janssens, I., Ceulemans, R.: Carbon and water vapor fluxes over four forests in two contrasting climatic zones, Agricultural and Forest Meteorology 180, 211– 224, 2013.

Garbulsky, M.-F., Penuelas, J., Papale, D., Filella, I.: Remote estimation of carbon dioxide uptake by a Mediterranean forest, GLOBAL CHANGE BIOLOGY, vol. 14, 2860-2867, ISSN: 1354-1013, doi: 10.1111/j.1365-2486.2008.01684.x, 2008.

Garrigues, S., Olioso, A., Calvet, J-C., Martin, E., Lafont, S., Moulin, S., Chanzy, A., Marloie, O., Desfonds, V., Bertrand, N., Renard, D.: Evaluation of land surface model simulations of evapotranspiration over a 12 year crop succession: impact of the soil hydraulic properties, Hydrology and Earth System Sciences Discussion, 11, 11687-11733. doi:10.5194/hessd-11-11687-2014, www.hydrol-earth-syst-sci-discuss.net/11/11687/2014/, 2014.

Gielen, B., B De Vos, Campioli, M., Neirynck, J., Papale, D., Verstraeten, A., Ceulemans, R., Janssens, I.: Biometric and eddy covariance-based assessment of decadal carbon sequestration of a temperate Scots pine forest. Agricultural and Forest Meteorology 174, 135-143, 2013.

Gioli, B, Miglietta, F, Vaccari, F.-P, Zaldei, A., De Martino, B.: The Sky Arrow ERA, an innovative airborne platform to monitor mass, momentum and energy exchange of ecosystems. Annals of Geophysics, 49, n. 1, pp 109-116, 2006.

Göckede, M., Foken, T., Aubinet, M., Aurela, M., Banza, J., Bernhofer, C., Bonnefond, M.-J., Brunet, Z., et al.,: Quality control of CarboEurope flux data – Part 1: Coupling footprint analyses with flux data quality assessment to evaluate sites in forest ecosystems, Biogeosciences, 5, 433-450, 2008.

Groenendijk, M., Dolman A.-J., van der Molen, M.-K., Leunung, R., Arneth, A., Delpierre, N., Gash, J.-H.-C., Lindroth, A., Richardson, A.-D., Verbeeck, H., and Wohlfahrt., G.: Assessing parameter variability in a photosynthesis model within and between plant functional types using global Fluxnet eddy covariance data, Agr. Forest Meteorol. 151, 22-38 , doi: 10.1016/j, 2011.

Guidolotti, G., Rey, A., D'Andrea, E., Matteucci, G., De Angelis, P.: Effect of environmental variables and stand structure on ecosystem respiration components in a Mediterranean beech forest, Tree Physiology 33: 960-972 (doi:10.1093/treephys/tpt065), 2013.

Gurney, K.-R., Law, R.-M., Denning, A.-S., Rayner, P.-J., Baker, D., Bousquet, P., Bruhwiler, L., Chen, Y.-H., Ciais, P., Fan, S., and et al.,: Towards robust regional estimates of CO_2 sources and sinks using atmospheric transport models, Nature 415 , 626-630, 2002.

Helfter, C., Campbell, C., Dinsmore, K.-J., Drewer, J., Coyle, M., Anderson, M., Skiba, U., Nemitz, E., Billett, M.-F., and. Sutton, M.-A.: Drivers of long-term variability in CO2 net ecosystem exchange in a temperate peatland. Accepted for publication in Biogeosciences (March 2015).

Hilton, W-T., Davis, J-K., Keller, K., and Urban, M-N.: Improving terrestrial CO_2 flux diagnosis using spatial structure in land surface model residuals, Biogeosciences,

Houweling, S., Breon, F.-M., Aben, I., Rödenbeck, C., Gloor, M., Heimann, M., and Ciais, P.: Inverse modeling of CO_2 sources and sinks using satellite data: A synthetic inter-comparison of measurement techniques and their performance as a function of space and time, Atmospheric Chemistry and Physics 4 , 523-538, 2004.

Hsieh, C. I., Katul, G. G., & Chi, T. W. (2000). An approximate analytical model for footprint estimation of scalar fluxes in thermally stratified atmospheric flows. Advances in Water Resources, 23, 765−772.

Jans, W.-W.-P., Jacobs, C.-M.-J., Kruijt, B., Elbers, J.-A.., Barendse, S., and Moors, E.-J.: Carbon exchange of a maize (Zea mays L.) crop: influence of phenology, Agriculture, Ecosystems & Environment, 139, 316-324, 2010.

Elbers, J.-A., Jacobs, C.-M.-J., Kruijt, B., Jans, W.-W.-P., and Moors, E.-J.: Assessing the uncertainty of estimated annual totals of net ecosystem productivity: A practical approach applied to a mid latitude temperate pine forest, Agricultural and Forest Meteorology, Volume 151, Issue 12, 15 December 2011, Pages 1823-1830, ISSN 0168-1923, 10.1016/j.agrformet.2011.07.020.

Jarosz, N., Brunet, Y., Lamaud, E., Irvine, M., Bonnefond, J.-M. and Loustau, D.: Carbon dioxide and energy flux partitioning between the understorey and the overstorey of a maritime pine forest during a year with reduced soil availability, Agricultural and Forest Meteorology 148 1508-1523, 2008.

Jongen, M., Pereira, J., Saires., L.-M.-I., and Pio, C.-A.: The effects of drought and timing of precipitation on the inter-annual variation in ecosystem-atmosphere exchange in a Mediterranean grassland. Agricultural and Forest Meteorology.151, 595-606. http://dx.doi.org/10.1016/j.agrformet.2011.01.008, 2011.

Jung, M., Henkel, K., Herold, M., and Churkina, G.: Exploiting synergies of global land cover products for carbon cycle modeling, Remote Sensing of Environment, 101(4), 534–553, doi:10.1016/j.rse.2006.01.020, 2006.

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.

Keenan, T., Garcia, R., Friend, A.-D., Zaehle, S., Gracia, C., Sabate, S.: Improved understanding of drought controls on seasonal variation in Mediterranean forest canopy CO_2 and water fluxes through combined in situ measurements and ecosystem modeling, Biogeosciences, 6, 2285-2329, 2009.

Klumpp, K., Tallec, T., Guix, N., Soussana, J.-F.: Long-term impacts of agricultural practices and climatic variability on carbon storage in a permanent pasture, Global Change Biology, 17, 3534–3545, 2011.

Knohl, A., Schulze, E.-D., Kolle, O., Buchmann, N.: Large carbon uptake by an unmanaged 250-year-old deciduous forest in Central Germany, Agricultural and Forest Meteorology 118, 151-167, 2003.

Kottek, M., Grieser, J., Beck, C., Rudolf, B., Rubel, F.: World Map of the Köppen-Geiger climate classification updated, Meteorologische Zeitschrift, 15, 259-263, 2006.

Krinner, G., Viovz, N., Ogee, J., Polcher, J., Friedlingstein, P., Ciais, P., Sitch, S., and Prentice, C.-I.: A dynamic global vegetation model for studies of the coupled atmosphere-biosphere system, Global Biogeochemical Cycles 19, doi: 10.1029/2003GB002199, 2005.

Kutsch, W.-L., Aubinet, M., Buchmann, N., et al.,: The net biome production of full crop rotations in Europe, Agriculture Ecosystems & Environment, 139, 336-345, 2010.

Lasslop, G., Reichstein, M., Kattge, J., and Papale, D.: Influences of observation errors in eddy flux data on inverse model parameter estimation, Biogeosciences Discussions 5, 2008.

Lauvaux, T., Schuh, A.-E., Bocquet, M., Wu, L., Richardson, S., Miles, N., and Davis, K.-J.: Network design for mesoscale inversions of CO_2 sources and sinks, Tellus B 64, doi: 10.3402/tellusb.v64i0.17980, 2012.

Lauvaux, T., Gioli, B., Sarrat, C., Rayner, P.-J., Ciais, P., Chevallier, F., Noilhan, J., Miglietta, F., Brunet, Y., Ceschia, E., Dolman, H., Elbers, J.-A., Gerbig, C., Hutjes, R., Jarosz, N., Legain, D., Uliasz, M.: Bridging the gap between atmospheric concentrations and local ecosystem measurements. Geophysical Research Letters, 36, Art. No. L19809, 2009.

Lauvaux, T., Uliasz, M., Sarrat, C., Chevallier, F., Bousquet, P., Lac, C., Davis, K., Ciais, P., Denning, A., and Rayner, P.: Mesoscale inversion: first results from the CERES campaign with synthetic data, Atmospheric Chemistry and Physics 8 , 3459-3471,2008.

Longdoz, B., Gross, P., Granier, A.: Multiple quality tests for analysing CO_2 fluxes in a beech temperate forest. Biogeosciences. 2008, 5, 719–729.

Mahadevan, P., Wofsy, S.-C., Matross, D.-M., Xiao, X., Dunn, A.-L., Lin, J.-C., Gerbig, C., Munger, J.-W., Chow, V.-Y., and Gottlieb, E.-W.: A satellite-based biosphere parameterization for net ecosystem CO_2 exchange: Vegetation Photosynthesis and Respiration Model (VPRM), Global Biogeochemical Cycles 22 , n/a-n/a , doi: 10.1029/2006GB002735, 2008.

Marcolla, B., Cescatti, A., Manca, G., Zorer, R., Cavagna, M., Fiora, A., Gianelle, D., Rodeghiero, M., Sottocornola, M., Zampedri, R.: Climatic controls and ecosystem responses drive the inter-annual variability of the net ecosystem exchange of an alpine meadow, Agric. For. Meteorol., 151, 1233–1243, 2011.

Marcolla, B., Cescatti, A., Montagnani, L., Manca, G., Kerschbaumer, G., Minerbi, S.: Importance of advection in the atmospheric CO_2 exchanges of an alpine forest. Agric. For. Meteorol., 130, 193–206, 2005.

Marcolla, B., Pitacco, A., Cescatti, A.: Canopy architecture and turbulence structure in a coniferous forest, Boundary-Layer Meteorol., 39–59, 2003.

Matteucci, M., Gruening, C., Goded., B., I., Cescatti, A.: Soil and ecosystem carbon fluxes in a Mediterranean forest during and after drought. Agrochimica LVIII, 91–115, 2014.

Meesters, A.-G.-C.-A., Tolk, L.-F., Peters, W., Hutjes, R.-W.-A., Vellinga, O.-S., Elbers, J.-A., Vermeulen, A.-T., van der Laan, S., Neubert, R.-E.-M., Meijer, H.-A.-J., and et al.,: Inverse carbon dioxide flux estimates for the Netherlands, Journal of Geophysical Research: Atmospheres 117 , 1984-2012, doi: 10.1029/2012jd017797, 2012.

Meijide, A., Manca, G., Goded, I., Magliulo, V., di Tommasi, P., Seufert, G., Cescatti, A.: Seasonal trends and environmental controls of methane emissions in a rice paddy field in Northern Italy, Biogeosciences 8, 3809–3821, 2011.

Michalak, A.-M., Bruhwiler, L., and Tans, P.-P.: A geostatistical approach to surface flux estimation of atmospheric trace gases, Journal of Geophysical Research, 109, D14109, doi: 10.1029/2003JD004422, 2004.

Montagnani, L., Manca, G., Canepa, E., Georgieva, E., Acosta, M., Feigenwinter, C., Janous, D., Kerschbaumer, G., et al.,: A new mass conservation approach to the study of CO_2 advection in an alpine forest, Journal of Geophysical Research – Atmospheres, 114, D07306, DOI: 10.1029/2008JD010650, 2009.

Moors E.-J., Jacobs, C., Jans, W., Supit, I., Kutsch, W.-l, Bernhofer, C., Bezat, P., Buchmann, N., Carrara, A., Ceschia, E., Elbers, J., Eugster, W., Kruijt, B., Loubet, B., Magliulo, E., Moureaux, C., Olioso, A., Saunders, M., Soegaard, H.: Variability in carbon exchange of European croplands. Agriculture, Ecosystem & Environment, 139(3): 325-335, 2010

Moureaux, C., Debacq, A., Bodson, B., Heinesch, B., Aubinet, M.: Annual net ecosystem carbon exchange by a sugar beet crop. Agricultural and Forest Meteorology 139, 25-39, 2006.

Nagy, Z., Pintér, K., Czóbel, Sz., Balogh, J., Horváth, L., Fóti, Sz., Barcza, Z., Weidinger, T., Csintalan, Zs., Dinh, N., Q., Grosz, B., Tuba, Z.: The carbon budget of a semiarid grassland in a wet and a dry year in Hungary. AGR ECOSYST ENVIRON 121:(1-2)21-29, 2007.

Olson, J.-S.: Global ecosystem framework-definitions, USGS EROS Data Cent. Intern. Rep. Sioux Falls SD, 37, 1994, 1994.

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

Peters, W., Jacobson, A.-R., Sweeney, C., Andrews, A.-E., Conway, T.-J., Masarie, K.-B., Miller, J., Bruhwiler, L.-M.-P., Petron, G., Hirsch, A.-I., Worthy, D.-E.-J., van der Werf, G.-R., Wennberg, J.-T.-R.-P.-O., Krol, M.-C. and Tans, P.-P.: An atmospheric perspective on North American carbon dioxide exchange: CarbonTracker, Proceedings of the National Academy of Sciences, doi: 10.1073pnas.0708986104, 2007.

Peylin, P., Rayner, P., Bousquet, P., Carouge, C., Hourdin, F., Heinrich, P., Ciais, P., and et al.,: Daily CO_2 flux estimates over Europe from continuous atmospheric measurements: 1, inverse methodology, Atmospheric Chemistry and Physics 5, 3173-3186, 2005.

Piao, S., Fang, J., Ciais, P., Peylin, P., Huang, Y., Sitch, S., and Wang, T.: The carbon balance of terrestrial ecosystems in China, Nature 458, 1009-1013, doi: 10.1038/nature07944, 2009.

Pillai, D., Gerbig, C., Kretschmer, R., Beck, V., Karstens, U., Neininger, B., and Heimann, M.: Comparing Lagrangian and Eulerian models for CO_2 transport- a step towards Bayesian inverse modeling using WRF/STILT-VPRM, Atmospheric Chemistry and Physics 12 , 8979-8991 , doi: 10.5194/acp-12-8979-2012, 2012.

Pita, G., Gielen, B., Zona, D., Rodrigues, A., Rambal, S., Janssens, I., Ceulemans, R.: Carbon and water vapor fluxes over four forests in two contrasting climatic zones, Agricultural and Forest Meteorology, 180, 211-224, 2013.

Prescher, A., K., Grünwald, T., Bernhofer, C.: Land use regulates carbon budgets in eastern Germany: From NEE to NBP, Agricultural and Forest Meteorology, 150, 1016-1025, 2010.

Rey, A., Belelli-Marchesini, L., Were, A., Serrano-Ortiz, P., Etiope, G., Papale, D., Domingo, F., and Pegoraro, E.: Wind as a main driver of the net ecosystem carbon balance of a semiarid Mediterranean steppe in the South East of Spain, Global Change Biology, 18, 539-554, 2012.

Richardson, A- D., Mahecha, M.-D., Falge, E., Kattge, J., Moffat, A.-M., Papale, D., Reichstein, M., Stauch, V.-J., Braswell, B.-H., Churkina, G., and et al.,: Statistical properties of random CO_2 flux measurement uncertainty inferred from model residuals, Agricultural and Forest Meteorology 148 , 38-50 , doi: 10.1016/j.agrformet.2007.09.001, 2008.

Richardson, A.-D., Hollinger, D.-Y., Burba, G.-G., Davis, K.-J., Flanagan, L.-B., Katul, G.-G., William M.-J., Ricciuto, D.-M., Stoy, P.-C., Suyker, A.-E., and et al.,: A multi-site analysis of random error in tower-based measurements of carbon and energy fluxes, Agricultural and Forest Meteorology 136 , 1-18, 2006.

Rödenbeck, C., Gerbig, C., Trusilova, K., and Heimann, M.: A two-step scheme for high-resolution regional atmospheric trace gas inversions based on independent models, Atmospheric Chemistry and Physics 9, 5331-5342, 2009.

Rödenbeck, C.: Estimating CO_2 sources and sinks from atmospheric mixing ratio measurements using a global inversion of atmospheric transport, Jena: Max Planck Institute for Biogeochemistry, technical report 6, 2005.

Rödenbeck, C., Houweling, S., Gloor, M., and Heimann, M.: Time-dependent atmospheric CO_2 inversions based on interannually varying tracer transport, Tellus B 55 , 488-497, 2003a.

Rödenbeck C., Houwelling S., Gloor M., and Heinmann, M.: CO_2 flux history 1982-2001 inferred from atmospheric data using a global inversion of atmospheric transport, Atmospheric Chemistry and Physics 3, 1919-1964, 2003b.

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

Schuh, A.-E., Denning, A.-S., Corbin, K.-D., Baker, I.-T., Uliasz, M., Parazoo, N., Andrews, A.-E., and Worthy, D.-E.-J.: A regional high-resolution carbon flux inversion of North America for 2004, Biogeosciences 7 , 1625-1644 , doi: 10.5194/bg-7-1625-2010, 2010.

Soussana, J.-F., Allard, V., Pilegaard, K., Ambus, P., Amman, C., Campbell, C., Ceschia, E., Clifton-Brown, J., Czöbel, Sz., Domingues, R., et al.,: Full accounting of the greenhouse gas (CO_2 , N_2O, CH_4) budget of nine European grassland sites, Agriculture, Ecosystems & Environment, 121, 121-134, 2007.

Skiba, U., Jones, S.-K., Drewer, J., Helfter, C., Anderson, M., Dinsmore, K., McKenzie, R., Nemitz, E., and Sutton, M.-A.: Comparison of soil greenhouse gas fluxes from extensive and intensive grazing in a temperate maritime climate, Biogeosciences, 10, 1231-1241, 10.5194/bg-10-1231-2013, 2013.

Suni, T., Rinne, J., Reissel, A., Altimir, N., Keronen, P., Rannik, Ü., Dal Maso, M., Kulmala, M., and Vesala, T.: Long-term measurements of surface fluxes above a Scots pine forest in Hyytiälä, southern Finland, 1996-2001. Boreal Environment Res. 4, 287-301, 2003.

Tallec, T., Béziat, P., Jarosz, N., Rivalland V., and Ceschia E.: Crops water use efficiencies: comparison of stand, ecosystem and agronomical approaches, Agricultural and Forestry Meteorology, 168, 69 – 168, 69-81, 2013.

Taufarova, K., Havrankova, K., Dvorská, A., Pavelka, M., Urbaniak, M., Janous, D.: Forest ecosystem as a source of CO_2 during growing season: relation to weather conditions. International Agrophysics, 28, 239-249, doi:10.2478/intag-2014-0013, 2014.

Tolk, L.-F., Dolman, A.-J., Meesters, A.-G.-C.-A., and Peters, W.: A comparison of different inverse carbon flux estimation approaches for application on a regional domain, Atmospheric Chemistry and Physics 11 , 10349-10365 , doi: 10.5194/acp-11-10349-2011, 2011.

Weedon, G.-P., Balsamo, G., Bellouin, N., Gomes, S., Best, M.-J., and Viterbo, P.: The WFDEI meteorological forcing data set: WATCH Forcing Data methodology applied to ERA-Interim reanalysis data, Water Resources Research, 2014.

Wei, S, Yi, C., Hendrey, G., Eaton, T., Rustic, G., Wang, S., Liu, H., Krakauer, N.-Y., Wang, W., Desai, A.-R., et al.,: Data-based perfect-deficit approach to understanding climate extremes and forest carbon assimilation capacity, Environmental Research Letters. 9(6):065002, 2014.

Wu, L., Bocquet, M., Lauvaux, T., Chevallier, F., Rayner, P., and Davis, K.: Optimal representation of source-sink fluxes for mesoscale carbon dioxide inversion with synthetic data, Journal of Geophysical Research: Atmospheres (1984--2012) 116, doi: 10.1029/2011JD016198, 2011.

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.

Zeeman, M., J., Hiller, R., Gilgen, A.-K., Michna, P., Plüss, P., Buchmann, N., Eugster, W.: Management and climate impacts on net CO_2 fluxes and carbon budgets of three grasslands along an elevational gradient in Switzerland, Agric. For. Meteorol., 150, 519-530, doi: 10.1016/j.agrformet.2010.01.011, 2010.

Zweifel, R., Eugster, W., Etzold, S., Dobbertin, M., Buchmann, N., Häsler, R.: Link between continuous stem radius changes and net ecosystem productivity of a subalpine Norway spruce forest in the Swiss Alps, New Phytol., 187, 819-830, doi: 10.1111/j.1469-8137.2010.03301.x, 2010.

Table 1: Eddy covariance sites measuring CO_2 fluxes that were used in the analysis. The land cover classification which is used, is coded as follows; CRO, DCF, EVG, MF, GRA, OSH, SAV for crops, deciduous forest, evergreen forest, mixed forest, grass, shrub and savanna respectively.

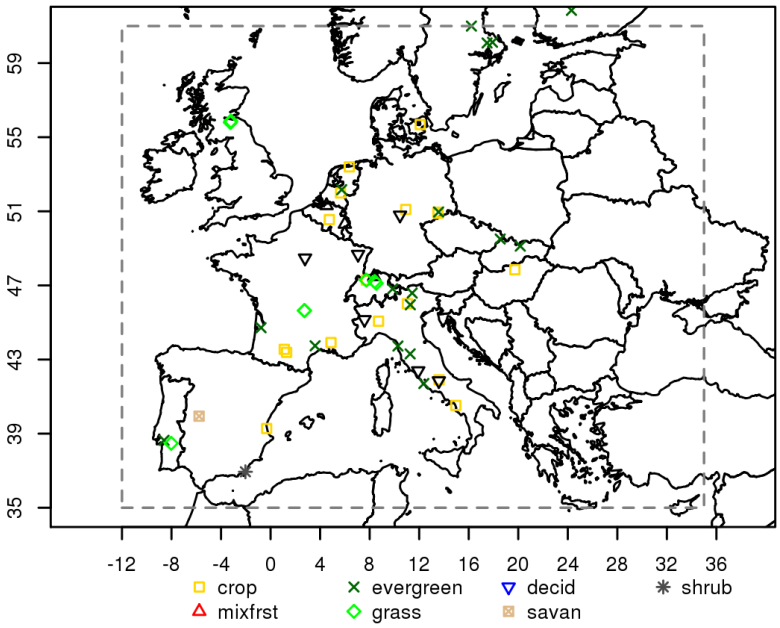
Site code	Site name	Land cover classification	Latitude	Longitude	Citation
BE-Bra	Brasschaat	MF	51.31	4.52	Gielen et al., 2013
BE-Lon	Lonzee	CRO	50.55	4.74	Moureaux et al., 2006
BE-Vie	Vielsalm	MF	50.31	6.00	Aubinet et al., 2001
CH-Cha	Chamau	GRA	47.21	8.41	Zeeman et al., 2010
CH-Dav	Davos	ENF	46.82	9.86	Zweifel et al., 2010
CH-Fru	Frebel	GRA	47.12	8.54	Zeeman et al., 2010
CH-Lae	Laegern	MF	47.48	8.37	Etzold et al., 2010
CH-Oe1	Oensingen grassland	GRA	47.29	7.73	Ammann et al., 2009
CH-Oe2	Oensingen crop	CRO	47.29	7.73	Dietiker et al., 2010
CZ-BK1	Bily Kriz forest	ENF	49.50	18.54	Taufarova et al., 2014
DE-Geb	Gebesee	CRO	51.10	10.91	Kutsch et al., 2010
DE-Gri	Grillenburg	GRA	50.95	13.51	Prescher et al., 2010
DE-Hai	Hainich	DBF	50.79	10.45	Knohl et al., 2003
DE-Kli	Klingenberg	CRO	50.89	13.52	Prescher et al., 2010
DE-Tha	Tharandt	ENF	50.96	13.57	Prescher et al., 2010
DK-Lva	Rimi	GRA	55.68	12.08	Soussana et al., 2007
ES-Agu	Aguamarga	OSH	36.94	-2.03	Rey et al., 2012

ES-ES2	El Saler-Sueca (Valencia)	CRO	39.28	-0.32	-
ES-LMa	Las Majadas del Tietar (Caceres)	SAV	39.94	-5.77	Casals et al., 2011
FI-Hyy	Hyytiälä	ENF	61.85	24.30	Suni et al., 2003
FR-Aur	Aurade	CRO	43.55	1.11	Tallec et al., 2013
FR-Avi	Avignon	CRO	43.92	4.88	Garrigues et al., 2014
FR-Fon	Fontainebleau	DBF	48.48	2.78	Delpierre et al., 2009
FR-Hes	Hesse	DBF	48.67	7.07	Longdoz et al., 2008
FR-LBr	Le Bray	ENF	44.72	-0.77	Jarosz el al., 2008
FR-Lq1	Laqueuille intensive	GRA	45.64	2.74	Klumpp et al., 2011
FR-Lq2	Laqueuille extensive	GRA	45.64	2.74	Klumpp et al., 2011
FR-Mau	Mauzac	GRA	43.39	1.29	Albergel et al., 2010
FR-Pue	Puechabon	EBF	43.74	3.60	Allard et al., 2008
HU-Mat	Matra	CRO	47.85	19.73	Nagy et al., 2007
IT-Amp	Amplero	GRA	41.90	13.61	Barcza et al., 2007
IT-BCi	Borgo Cioffi	CRO	40.52	14.96	Kutsch et al., 2010
IT-Cas	Castellaro	CRO	45.07	8.72	Meijide et al., 2011
IT-Col	Collelongo	DBF	41.85	13.59	Guidolotti et al., 2013
IT-Cpz	Castelporziano	EBF	41.71	12.38	Garbulsky et al., 2008
IT-Lav	Lavarone	ENF	45.96	11.28	Marcolla et al., 2003

IT-Lec	Lecceto	EBF	43.30	11.27	Chiesi et al., 2011
IT-LMa	Malga Arpaco	GRA	46.11	11.70	Soussana et al., 2007
IT-MBo	Monte Bondone	GRA	46.01	11.05	Marcolla et al., 2011
IT-Ren	Renon	ENF	46.59	11.43	Marcolla et al., 2005
IT-Ro2	Roccarespampani 2	DBF	42.39	11.92	Wei et al., 2014
IT-SRo	San Rossore	ENF	43.73	10.28	Matteucci et al., 2014
NL-Dij	Dijkgraaf	CRO	51.99	5.65	Jans et al., 2010
NL-Loo	Loobos	ENF	52.17	5.74	Elbers et al., 2011
NL-Lut	Lutjewad	CRO	53.40	6.36	Moors et al., 2010
PT-Esp	Espirra	EBF	38.64	-8.60	Gabriel et al., 2013
PT-Mi2	Mitra IV (Tojal)	GRA	38.48	-8.02	Jongen et al., 2011
SE-Kno	KnottŒsen	ENF	61.00	16.22	-
SE-Nor	Norunda	ENF	60.09	17.48	-
SE-Sk1	Skyttorp 1	ENF	60.13	17.92	-
SK-Tat	Tatra	ENF	49.12	20.16	-
UK-AMo	Auchencorth Moss	GRA	55.79	-3.24	Helfter et al., 2015
UK-EBu	Easter Bush	GRA	55.87	-3.21	Skiba et al., 2013

Table 2: Annual temporal autocorrelation times in days, from model-data and model-model residuals. The number within the brackets shows the correlation times when excluding sites with large model-data bias from the analysis.

Reference	VPRM10 [days]	VPRM1 [days]	ORCHIDEE [days]	5PM [days]
OBSERVATION	32 (27)	33 (29)	26 (24)	70 (34)
VPRM50	-	-	28 (28)	52 (46)
VPRM10	-	~~47 (46)~~	~~30 (31)~~	131 (100)
~~VPRM1~~	~~-~~	~~-~~	~~28 (28)~~	~~116 (85)~~
ORCHIDEE	-	-	-	38 (32)
5PM	-	-	-	-



1

Figure 1. Eddy covariance sites used in the study. The dashed line delimits the exact domain used to calculate the aggregated fluxes.

2

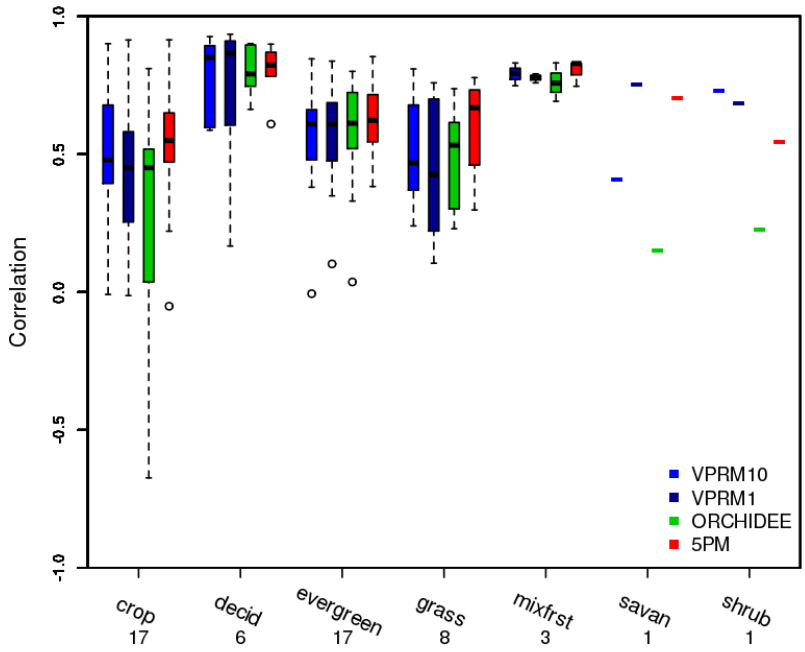


Figure 2. Box and whisker plot for site-specific correlation coefficients between modeled and observed daily fluxes as a function of the vegetation type. The numbers beneath the x-axis indicate the number of sites involved. The bottom and the top of the box denote the first and the third quartiles. The band inside the box indicates the central 50% and the line within is the median. Upper and lower line edges denote the maximum and the minimum values excluding outliers. Outliers are shown as circles.

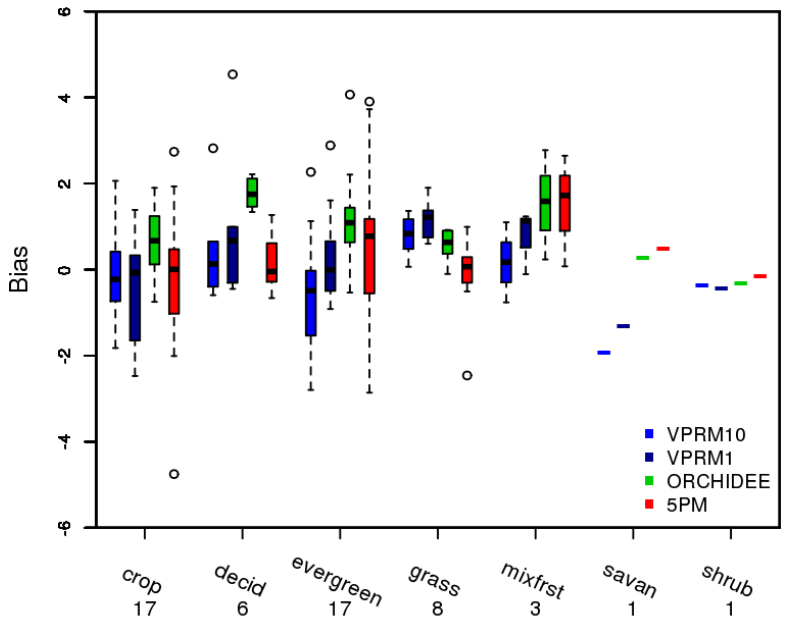


Figure 3. Box and whisker plot for the annual site–specific biases of the models differentiated by vegetation type. Units at y-axis are in µmol m^{-2} s^{-1}. (for conversion to gC m^{-2} yr^{-1} reported values in y axis should be multiplied by 378,7694).

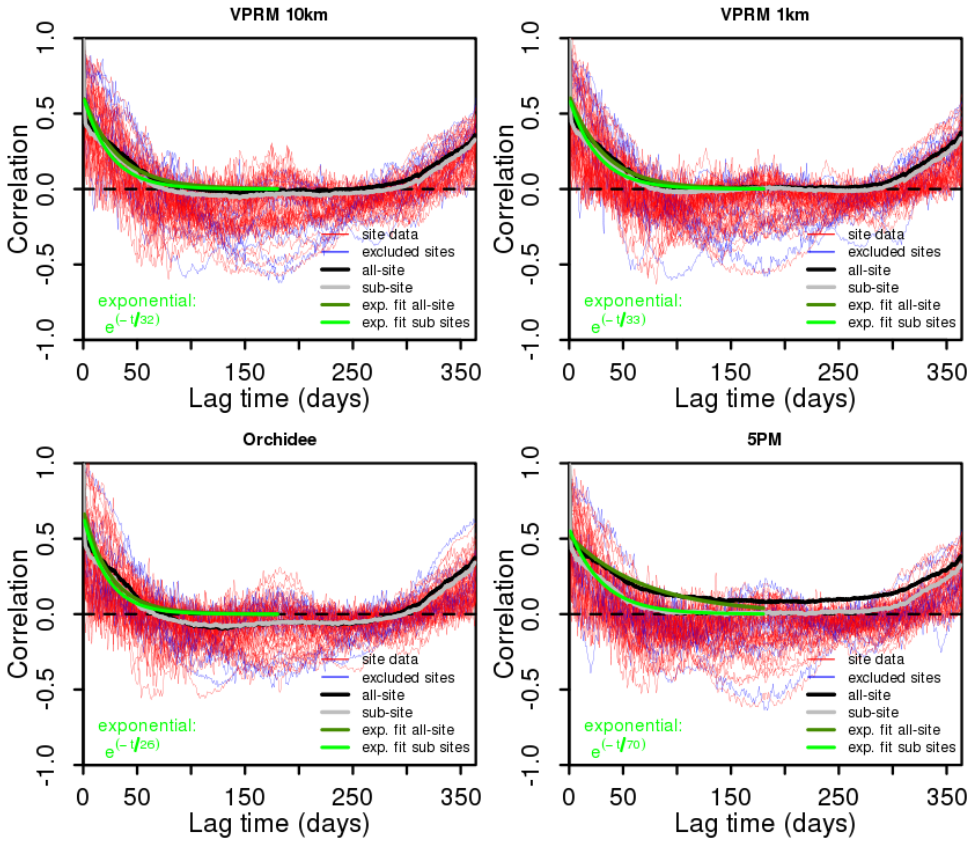


Figure 4. Temporal lagged autocorrelation from model-data daily averaged NEE residuals for all models. ~~Red~~Thin red lines correspond to different sites, while the ~~dark magenta color reveals~~blue thin lines reveal the sites with a bias larger than +/-2.5 µmol m^{-2} s^{-1}. ~~Black~~The thick black line shows the all-site autocorrelation, and the thick grey line indicates the all-site autocorrelation ~~excluding~~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 ~~using lags up to 180 days~~, 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.

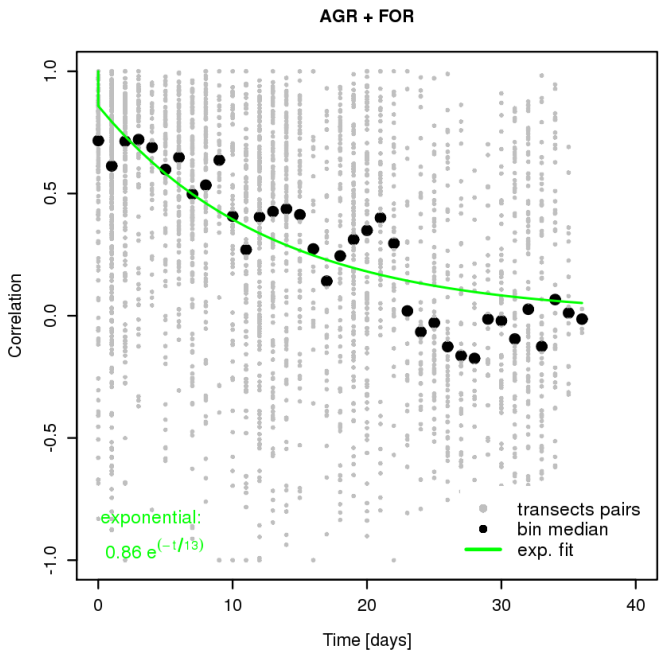


Figure 5. Temporal autocorrelation for VPRM10 – aircraft NEE residuals. Black dots represent individual flux transects pairs sampled at different times as function of time separation. Black circles represent daily scale binned data.

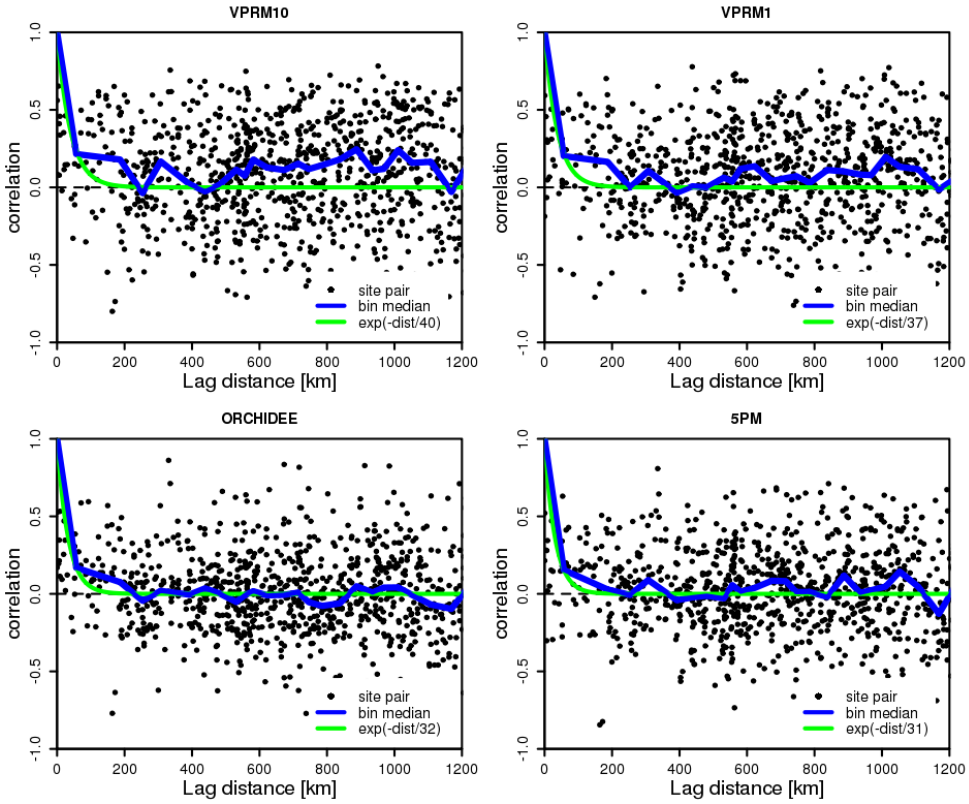


Figure 6. Distance correlogram for the daily net ecosystem exchange (NEE) residuals using all sites. Black dots represent the different site pairs; the blue line represents the median value of the points per 100-km bin and the green an exponential fit. Results are shown for residuals of VPRM at a resolution of 10 km (top left) and 1 km (top right), ORCHIDEE (bottom left), 5PM (bottom right).

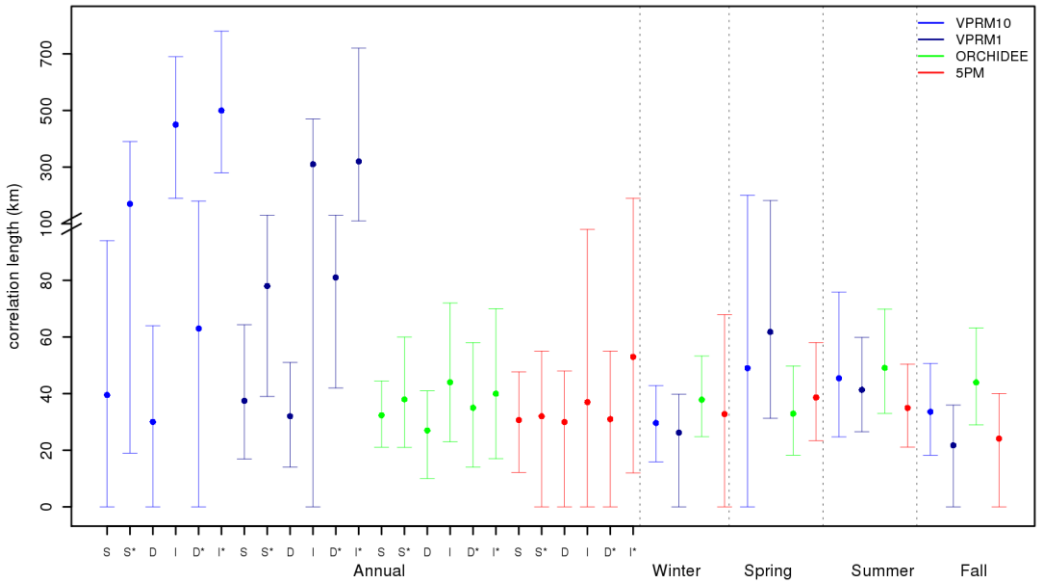


Figure 7. Annual and seasonal e-folding correlation length of the daily averaged model-data NEE residuals for VPRM at 10 and 1 km resolution, ORCHIDEE and 5PM. "S" refers to the standard case where all pairs were used, “D” refers to the case where only pairs with different vegetation types were used, "I" denotes the case in which only pairs with identical vegetation type were considered, and “*” denotes that in addition 150 days of common non-missing data are required for each pair of sites. The dot represents the best-fit value when fitting the exponential model. The upper and the lower edge of the error bars show the 2.5 and 97.5 percentiles of the length value. Note the scale change in the y-axis at 100 km.

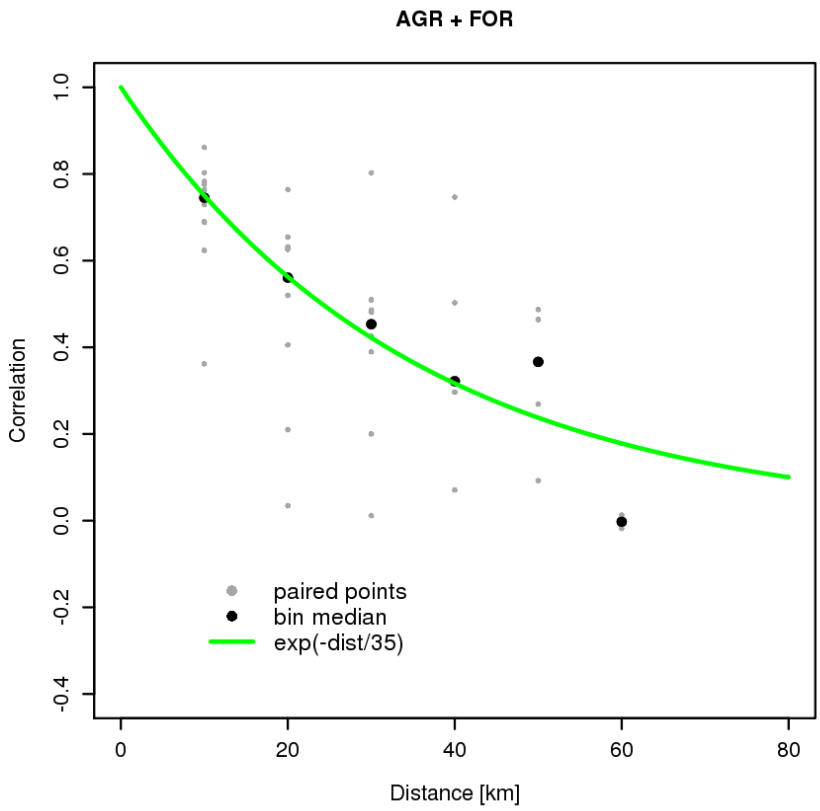


Figure 8. Distance correlogram between VPRM10 and aircraft NEE measurements. Black dots represents the different aircraft grid points pairs; black circles represent 10 km scale binned data.

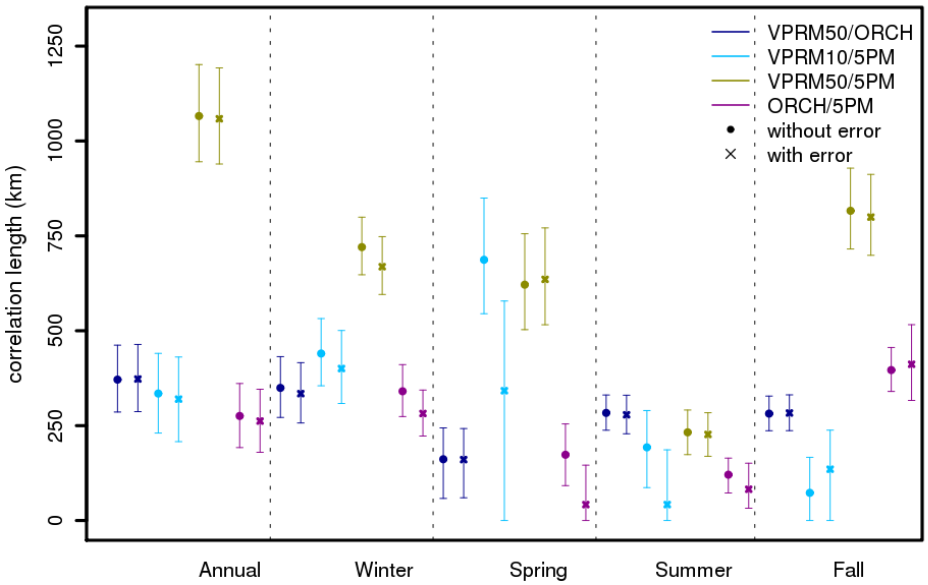


Figure 9. Annual and seasonal e-folding correlation length for an ensemble of daily averaged NEE differences between two models without (filled circle) and with random measurement errors added to the modeled fluxes used as reference (crosses). The symbols represents the best fit value when fitting the exponential model, and the upper and lower edge of the error bars show the 2.5 and 97.5 percentiles of the correlation length. The first acronym at the legend represents the model used as reference and the second the model which was compared with. Note that for the VPRM10/VPRM1 case during spring (with and without random error), the 97.5 percentile of the length value exceeds the y-axis and has a value of 1073, 1626 km respectively.

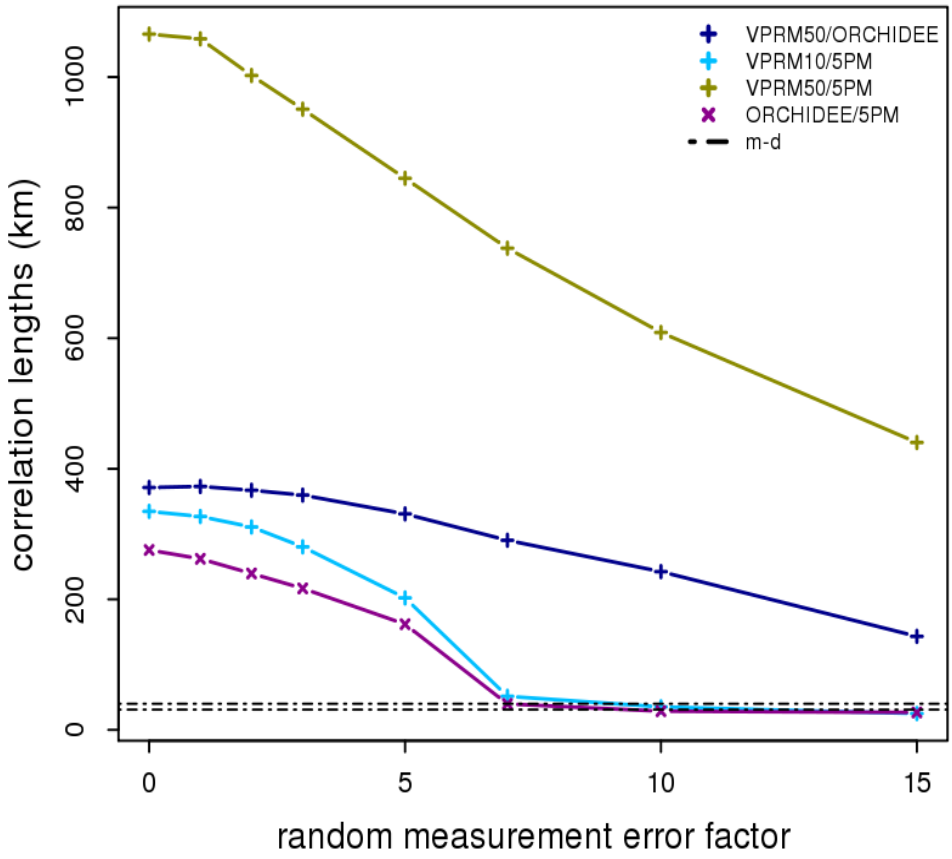


Figure 10. Annual e-folding correlation lengths as a function of the factor used for scaling the random measurement error, for all model-model combinations. The black dot-dash lines reveal the range of the spatial correlation lengths generated from the model-data comparisons.