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."