We thank the reviewer for her/his comments that give us the opportunity to offer the following explanations regarding the objectives, key results and conclusions of our study.

This paper evaluates the methods of estimating/assigning uncertainties to the fluxes derived using inversion modelling. A posteriori fluxes are compared with those estimated from the flux tower network over the European domain. They find that the a posteriori uncertainties are somewhat large as estimated here and that interannual variations in fluxes cannot be determined statistically significantly at the European scale given the measurement network and inversion modelling systems employed for this work.

These two limitations do not entirely reflect the main conclusions of our paper. Regarding the first limitation: yes, the posterior uncertainties are a bit larger than the posterior misfits. However, (1) the posterior misfits also include flux tower measurement errors and representativeness differences (between the model grid cells and flux tower sites), and (2) posterior uncertainties are larger but close to posterior misfits. We invite the reviewer to check, in the synthesis of global inversions by *Peylin et al., 2013 BGD*[1] (http://www.biogeosciences-discuss.net/10/5301/2013/bgd-10-5301-2013.html; this paper will be cited in our paper; see for example the table 2), that such an agreement between posterior misfits and uncertainty in our inversion study is in fact remarkable given that the spread of results from state-of-the-art inversions is generally well larger than the estimate of uncertainties produced by these systems.

The paper is well written, and addresses an important issue in inverse modelling. However, I have some reservations that their system is a bit too specific to be widely applicable for all inverse modelling systems in general.

- If the system is suitable for addressing the science questions described in our paper, we have no other multipurpose objectives. Most of inversion systems are global. Therefore, we do not necessarily aim at generalizing all our conclusions to other existing inversion systems but rather at giving confidence in the results which we obtain with our regional system over Europe. Regarding that objective, we demonstrate (using comparison to EC data) that the inversion yields a significant improvement in the estimate of the fluxes (which is far from easy to get: see the spread of results from state of the art global inversions, and thus the typical amplitude of errors from a basic global inversion, in *Peylin et al., 2013 BGD*), and a significant decrease of uncertainty in these estimate. We also show that the inversion yields a reliable estimate of the seasonal cycle.
- On the other hand, our study shows that a validation of the estimates of uncertainties, using independent cross validation data, is necessary while few inversion results document the degree of reliability for their uncertainties. Our study also shows that we have data to support such a validation. And the positive results obtained when comparing theoretical estimates of uncertainties with misfits to EC data gives some general confidence in the Bayesian estimate of uncertainties based on the Gaussian statistical framework, which is used by nearly all the inversion systems.
- Finally, our conclusions regarding the confidence in the inverted seasonal cycle but not in the Inter-Annual Variability (IAV) of NEE seems to apply to global inversions according to figures (cf Fig. 8 & 10 for Europe) from *Peylin et al., 2013 BGD* even though “Europe” in this paper is a far larger area than Europe in our paper and even though we focus more on the IAV of monthly estimates rather than on that of annual estimates while *Peylin et al., 2013, BGD* provide results for the IAV at annual scale only. Therefore, these 2 papers raise identical insights based on independent analysis. These analysis are critical since many inversion products are used to study the IAV of fluxes. See also the two discussions on the IAV in answer to two of the reviewer's next comments.

1Peylin, P., Law, R. M., Gurney, K. R., Chevallier, F., Jacobson, A. R., Maki, T., Niwa, Y., Patra, P. K., Peters, W., Rayner, P. J., Rödenbeck, C., and Zhang, X.: Global atmospheric carbon budget: results from an ensemble of atmospheric CO2 inversions, Biogeosciences Discuss., 10, 5301-5360, doi:10.5194/bgd-10-5301-2013, 2013.

We intend to add some text to clarify these points in the introduction and in the conclusion of our paper.

Please find my concerns below. I do not demand the authors to resolve any or all of the issues raised below, but at the least limitations of the approach should be clarified before publications in ACP.

Detailed comments: p.5771, l.15: I do not understand "respectively" as used here and elsewhere in this para.

This refereed to the previous two "respectively" (line 13 and 14): we provide statistics of significance levels for the prior and the posterior. The terms respectively l.20 and 21 were independent to that at lines 13-15 and work together. **We will correct the use of respectively everywhere** to avoid using "respectively" from the first occurrence of each duality.

p.5772, l. 1: "atmospheric CO2 measurements" may sound better

We will apply this correction.

p.5772, l. 6: What is the difference between inventory and climatology?

The term "climatology" refers to a long-term mean which can be applied for any year (the term "climate" should relate to long time scales).

The term inventory would rather apply to estimates for specific years / seasons (for relatively short time scales compared to the idea of climatology).

The sentence will be modified to clarify the distinction between the two words in terms of time scale.

Furthermore, the terms climatology and inventory, due to their original definitions or due to usages, traditionally refer to different types of product (type of measurement or type of models) or of quantities (type of fluxes). We will not use the word climatology for anthropogenic emissions because their estimates are rather called "inventories", while we would not use "inventories" for the climatological mean ocean fluxes used, e.g., as our prior estimate of the ocean fluxes.

p.5778,#l.10: This is difficult to believe. I cannot say what is wrong, unless detailed values are given or the correlation lengths too large? In principle the prior flux unc in NEEs should be large in the summer than winter.
Yes: here, this is true "on average" at any spatial scale.
Can you comment on this?
The uncertainty reduction is defined as 1-posterior uncertainty / prior uncertainty.
As said above, the prior uncertainty is always (i.e. at any spatial scale) larger in summer than in winter. Let us now focus on the posterior uncertainty, which is function of the prior uncertainty, of the observation error and of the atmospheric transport.
The observation error (which includes the atmospheric transport error) is larger in winter than in summer. However, the atmospheric transport (i.e. the factor that scales a given flux into a given mixing ratio increment; in other terms: the Jacobian or the sensitivity of mixing ratios to fluxes) is usually "larger" in winter than in summer due to thinner PBL. Therefore, it is not easy to anticipate which term dominates in the change of uncertainty reduction from July to Dec: larger prior uncertainty + smaller obs error in July vs "larger" atmospheric transport in December ? The result is

even more complicated to anticipate when considering uncertainty reduction at the European scale since the weight of spatial covariances in the posterior uncertainties will be critical and difficult to derive with a simple hand calculation or by intuition.

Here are the results from the analysis that have been conducted at LSCE using the system described in this paper (these results will be part of a paper "Kadygrov et al." which should be submitted soon): at the pixel scale (0.5° resolution), the uncertainty reduction is larger in July than in December for most of the domain used here. However, smaller spatial correlations in the posterior uncertainty in December than in July (especially in the core of the observation network, around Germany) make the uncertainty reductions in July and December converge toward a similar value when increasing the spatial scale.

Since this analysis is a bit out of the scope of our paper and requires some detailed discussions which will be conducted in Kadygrov et al., here we just intend to extend the list of parameters influencing the estimate of uncertainty reduction that vary from year to year or from winter to summer (mentioning the correlations of the posterior uncertainty) highlighting that these different parameters can have opposed effect for the increase/decrease of uncertainty reduction and that the overall effect at the European scale is neutral.

Or the system has large number of dipoles, when integrated over the whole domain the results look similar!

We should speak about correlations in the posterior uncertainties rather than about "dipoles" to analyze how the uncertainty reduction evolves as a function of the spatial scale, even though "dipoles" are phenomena that are usually connected to negative correlations in posterior uncertainties (caused by gaps in observation networks). We have some negative correlations in the posterior uncertainty both in December and July, and, as stated above, these correlations are generally smaller (i.e. "more negative") in December than in July. The discussion above details how this contributes to getting similar uncertainty reductions at the European scale in December and July.

How are the results at country scale, say, Germany or France or at site scale?

This will be analyzed in Kadygrov et al. The scores of uncertainty reduction in July and December converge at scales ranging between the country and Western Europe scales (i.e. for France, the uncertainty reduction is still a bit higher in July than in Dec). Germany (which is located at the heart of the observation network and which is a large country) is a prominent exception with higher uncertainty reduction in Dec than in July. At "site scale", strictly speaking (i.e. checking uncertainty reduction for model pixels containing atmospheric stations), one can have larger or smaller uncertainty reduction in July than in Dec (depending on the weight of larger prior uncertainty + smaller obs error in July vs "larger" atmospheric transport in December; cf above) even though on average (including all the pixels), the uncertainty reduction at pixel scale is larger in July than in Dec.

p.5780,#l.26: I am curious to see the results, if you make four divisions of the western Europe. Could you show a four column figure? Please provide the figure in your reply, if not in the main paper.

We attach such a figure to this discussion. Given the configuration of our small European domain in which nearly all the atmospheric stations are located in the "North", we have defined 3 regions (south of 42°N, and 2 regions north of 42°N: west and east of 10°E) rather than 4. The inversion in the North-West part of our domain gets information from the major part of the atmospheric stations

used in the study due to their locations and due to winds blowing mostly from the West. Therefore, the uncertainty reduction there is very high, while it is very low in Southern Europe.
Since the analysis of these results is quite out of the scope of the paper (which does not study the spatial variations of the results from the inversion), we do not intend to discuss them in this paper.

p.5782, l. 1: It is strange that all the figures and tables are cited before the results section

Section 3.1 explains and justifies the analysis of the results based on these figures and tables in a synthetic way. We believe that these explanations are quite easier to conduct with the support of the figures, especially since we use several levels of statistics which may confuse the reader without any visual support (anomalies, misfits in anomalies, RMS of the STD of monthly uncertainties etc.).

Another reason for presenting the figures before the result section is that the results are themselves presented in a synthetic way, e.g., the distribution of prior misfits is analyzed based on fig 2 and 4 and then the distribution of posterior misfits is analyzed based on fig 2 and 4.

We consider that the large number of statistics that we produced and analyzed could have made the paper very difficult to follow if going through statistics and figures one by one. Therefore we feel that our general presentation of diagnostics prior to the result section is a simpler and more elegant way to proceed.

p.5786, l.23: I thought that was one of the main targets of this paper?

It is true that the paper aims at evaluating uncertainties from the inversion and that this evaluation at annual scale cannot be conducted rigorously based on our comparison to EC data. However, the main target of the paper is clearly related to the uncertainties at the 30-day scale and most of the analysis are dedicated to the evaluation of results at the 30-day scale, especially since we have not tried to derive posterior uncertainties at annual scale. Furthermore, the analysis at annual scale still raises some insights about the low confidence in the inverted fluxes at annual scale which is an objective underlying the evaluation of the uncertainties.

We intend to remind in the conclusion that even though EC data have not helped us evaluating prior uncertainties, our study shows that the confidence in annual anomalies is very low.

Have not such conclusions already well documented in published lieratures?

Getting proper estimates of uncertainties at annual scale is obviously difficult. Previous studies at global scale may have raised such a concern but they have rather focused on the difficulty to derive the flux at annual scale themselves, which is a different object than the uncertainty at annual scale that is discussed at lines 23-25 p5786.

Our discussions/conclusions on the low confidence in the estimates of the IAV from the inversion at lines 13-15 p5786 for the annual scale and in section 5.2 for the monthly to annual scale are quite in line with recent results from *Peylin et al, 2013, BGD* (cf their Fig 8). However:
- many papers have studied the IAV of the NEE based on global inversions which shows that the literature often has a high confidence in such estimates
- *Baker et al. 2006, GBC*[2] (that will be cited in our paper) concluded that the IAV estimates for most of the large regions defined by the TRANSCOM projects (see their figure 1) were not significant

2 Baker, D. F., et al. (2006), TransCom 3 inversion intercomparison: Impact of transport model errors on the interannual variability of regional CO2 fluxes, 1988 – 2003, Global Biogeochem. Cycles, 20, GB1002, doi:10.1029/2004GB002439.

enough compared to their estimate of posterior uncertainties. However their conclusion was that for their European domain (which is far larger than ours), estimates of the IAV was significant compared to "these" posterior uncertainties. Additionally, the robustness of their estimate of posterior uncertainties (much of their efforts focused on simulating atmospheric transport errors based on the spread of different global transport models forced with large scale fluxes) could be questioned and was not evaluated.

Here, we have applied an objective analysis using independent cross validation flux tower measurements to check the confidence in the inversion posterior uncertainties, which, to our knowledge, has not been conducted recently, and which brings new insights for the discussions on the robustness of the IAV estimates.
Furthermore, we use a regional system instead of a global system in order to improve the estimate of fluxes regionally. Therefore, we could have expected more robust estimates of the IAV when using our system than when using global systems. This study shows that despite an increase in the spatial / temporal resolution of the inversion, and despite using a mesoscale atmospheric transport model, we still have difficulties in deriving the IAV over Europe, and this conclusion is definitely new and worth stating clearly. Many papers have written that bottom up annual mean NEE estimates, for instance from inventories, can be verified by top down inversions. It is important to outline that with the present network and our system, this is not true. See also the discussion regarding the IAV in answer to one of the reviewer's next comments.

We will illustrate with the recent paper by *Peylin et al, 2013, BGD* the low reliability in the IAV from the state of the art global inversion systems. We will also indicate that previous analysis by *Baker et al. 2006, GBC*, had been rather positive regarding the inversion of the IAV for Europe and, thus, that this topic still needs some investigation. The introduction will better states that our analysis checks whether our regional system with increased spatial / temporal resolution and using more atmospheric data has the ability to raise more robust estimates of the fluxes at monthly to annual scale. In the conclusion, the paper will state that despite a high confidence in the set-up of the regional system, and despite the assets of regional inversion, we observe that such a system does not seem to be able to raise more reliable estimates of IAV than global systems (recognizing a limitation of the inversion method, but not of the study conducted here).

p.5788, l.15: Whilst talking about the europe wide fluxes, it may be good to use TgC/yr or /mon units?

First, our European domain is quite small and it may confuse the reader if not giving fluxes per unit of area (the number given in TgC could be mistakenly compared to usual estimates for the "whole" Europe).

Secondly, we tried to keep as few units as possible to make comparisons easier for the reader. Since we describe the uncertainties at the daily/pixel scale, and to avoid any confusion with the /month unit (due to the ambiguity between exact months and 30-day months) we decided to work with fluxes in gC/m2/day or gC/m2/year when dealing with annual fluxes, which is a unit familiar to the scientific community (see for example the papers for the synthesis of the CarboEurope project: Ciais et al. 2010 GCB and Luyssaert et al. 2010 GCB which provide estimates of annual fluxes in gC/m2/year).

p.5788,#l.25: For these conclusions that IAVs in monthly or annual fluxes have to be greater than a posteriori uncertainty for the flux IAVs to be significant, I think most of the model/data errors assiged to the sites are systematic, and only partly random, which would cancel out for sufficient number of model realisations. Much of the systematic components will keep the a

posterior uncertainty high, but the mean flux value will change due to the signals in atmospheric CO2 data anomaly.

Given that the estimate of posterior uncertainty here is based on Monte Carlo ensembles (with 60 members) with "errors assigned to model and data" that are purely random, the reviewer may assume that the misfits to EC data for a given month (e.g. February) are quite similar from year to year and that actual errors are quite "systematic" (and thus that we would have validated the STD of our random distribution by comparing it to a bias).
In a more general way, the reviewer may assume that the posterior error is fully correlated from one year to the other, and so that the posterior error for the IAV is very small. However:

(1) The analysis of the distribution of posterior misfits to eddy covariance measurements for a given month shows that these misfits have significant variations and we can have positive and negative posterior misfits depending on the year for most of the months. This invites us to believe that actual posterior errors are highly variable from year to year too.

(2) Systematic errors from year to year are likely in the prior estimates. But since the increments from inversion are large, and since processes that are highly variable from year to year, interact in the inversion, the assumption that posterior estimates have errors that do not vary from year to year is very unlikely.

We will clarify the discussion on that point in the revised paper.

It is true that uncertainty in IAV is different from uncertainty at annual scale and a robust knowledge of the correlations for lags=1 year is required (but presently impossible) to derive a rigorous estimate of the ratio signal/noise for the IAV. However, the fact that uncertainties on monthly estimates are higher than their IAV, points (1) and (2) above, the analysis of the anomaly in summer 2003 and results from *Peylin et al 2013, BGD* definitely raise a need for caution and lead to believe that, presently, the IAV from the inversion has a low reliability.

As traditionaly done in the CO2 inverse modelling, one has to run sensitivity inversions to estimate the uncertainties for flux IAVs at monthly/annual scales.

We do not believe "traditional sensitivity studies" to be a more robust approach than that of our paper to evaluate the uncertainty in the IAV.

Our estimates of posterior uncertainties based on a Monte Carlo approach is in fact "a sensitivity study" pushed to its furthest extent (i.e. we do not run only 1 new simulation to test the impact of each source of error that we have identified; we run 60 simulations to get the full statistical structure of the impact of all sources of errors that we have identified).
The major point of this paper is to evaluate results from the Monte Carlo approach (i.e. to what we could get from a lot of sensitivity studies) before stating that these results can be used to draw conclusions.

Sensitivity studies by themselves (without validation) are definitely far from objective (results are fully driven by the assumptions made to choose and to run each sensitivity test). They consist in assuming that uncertainties are related to a limited number of sources of errors that can be characterized quite easily. A lot of sources can easily be missed or poorly represented when running few sensitivity tests which results in under-estimating the actual errors. An "optimistic" selection of weakly sensitive settings would yield for instance to the conclusion that a result is "robust" whereas it may not have been so with a larger range of sensitivity tests.

Therefore, even though it is true that our Monte Carlo approach at monthly scale is not fully adapted to estimate the structure of the uncertainty for multi-year periods,
(1) it sounds more robust and more feasible than ~5 6-year inversions (which are computationally extremely expensive) used to test the sensitivity to a change in ~4 parameters
(2) the present lack of knowledge on the prior uncertainty for multi-year periods, (presently, no inversion system include any prior correlation for lags ≥1 year between monthly to annual uncertainties) prevents from deriving a robust prior or posterior uncertainty in IAV.

p.5789, l.15: The "remarkable agreement" comes from the inversion setup, say, a priori dependence. A priori meaning not the a priori fluxes only, but also including the correlation lengths etc., which controls your inversion results Can you reduce the correlation lengths to a few forward model grids around the measurement sites, and perform the same analysis only for the grids of measurement sites?

The agreement depends on any single piece of input information for the inversion system or for the comparison to eddy covariance measurements: prior fluxes, uncertainty in the prior fluxes A, atmospheric transport model, inversion framework (i.e. the practical implementation of the Bayesian formula), observation errors R, atmospheric measurements and eddy covariance measurements. A change in any component could break this agreement.

The reviewer is likely asking to know what is the sensitivity of the agreement to prior uncertainties in order to check whether it is so "remarkable".

Estimates of the uncertainty reduction when reducing the correlation length (from 250 to 150 km) in B has been tested and will be analyzed and discussed in Kadygrov et al (note, however, that this specific test was conducted for an atmospheric measurement network larger than the one used here). This yields smaller uncertainty reduction at the European scale, lower prior uncertainties at European scale (the prior uncertainty for the whole European domain is divided by ~3/2) and slightly smaller posterior uncertainties at the European scale (division by ~1.2 but this rescaling works for a network larger than the one used in our paper, while the value given for B is independent of the obs network). Note that at pixel scale A is generally even larger when decreasing the correlations in B since a given station provides information to a smaller area due to smaller spread of the increments through B.

Since the prior uncertainty when using 150 km correlation length in B would be 1.5 smaller than that in this paper, i.e. equal to ~0.46g/m2/day (instead of 0.69g/m2/day here[3]), but since the prior NEE would not change, i.e. since the prior misfits would be the same ~0.64g/m2/day, we would definitely decrease significantly the agreement between the prior uncertainty and the prior misfits that we show in our paper if changing spatial correlations in B.

We would unlikely keep the agreement shown in our paper for posterior uncertainties too but it is difficult to anticipate since we need to rerun the 6-year inversion to get the resulting posterior misfits to EC data (estimates by hand calculations are quite impossible). At least, we know that while A would decrease by a factor ~1.2 (again, this factor is valid for a network denser than that in the paper, but qualitatively we know that A would decrease) from 0.33g/m2/day to 0.28g/m2/day, the posterior misfits would increase to values higher than 0.4g/m2/day since increments to prior fluxes from inversion would be smaller (the "Kalman gain" of the inversion is smaller i.e. corrections applied to fit with concentration measurements are smaller, when B is smaller which is

3 These numbers have been slightly revised since there used to be a little inconsistency between the different type of "monthly" estimates (some were based on actual months -e.g. January 1 – 31 for month 1- while the others were based on 30-day periods -e.g. January 1-30 for month 1; for month 12 this had significant impacts because the inversions end on December 26). Now, following what was already said in the first manuscript, all results are based on 30-day periods. This has no consequence for any major discussion of the paper.

reflected by smaller uncertainty reduction estimates) and subsequently, misfits to EC data would be larger. Therefore, we should significantly decrease the agreement between the posterior uncertainty and the posterior misfits.

We hope that this reasoning convinces that the sensitivity to correlation lengths in B is necessarily high for prior uncertainties vs prior misfits, and likely high for posterior uncertainties vs posterior misfits. Running a new 6-year inversion based on a B matrix with 150km correl length to get the exact sensitivity would have a huge computational cost. That would be very expensive for a secondary discussion which may not bring far more insights about the stability of our "remarkable agreement" than the simple demonstration given above since a true rigorous check of the stability of the agreement would require to rerun 6-year inversions to check the sensitivity to STD in R, in B, temporal correlations in B, atmospheric transport model, spatial and temporal resolution of the control vector for the inversion, inversion or Monte Carlo ensemble frameworks etc.
Note that one could feel that such sensitivity studies could help to get an even better agreement between estimates of uncertainties and misfits to EC data. However, we do not claim that we should look for the best agreement as possible since misfits to EC data encompass EC measurement errors and differences of representativeness. The B and R matrices used for our study are based on our best knowledge of these errors at 0.5° resolution (see *Broquet et al.,2011, JGR*), independent of the resulting inversion products, and here, we just aim at evaluating them, not at "optimizing" them.

The "analysis only for the grids of atmospheric(?) measurement sites" cannot apply to our comparisons of uncertainties vs EC data which have to be conducted at EC measurement locations only.

In such a system you will be handling mainly the a priori and and posteriori fluxes, I presume, constrained by CO2 measurements, without other external influences.

Atmospheric inversion constrains the posterior fluxes based on the prior fluxes and on CO2 atmospheric data. Presently, we do not use other source of information but in principle we could easily (mathematically and technically) assimilate other types of data such as seasonal budgets. As discussed in the paper, we could also invert other parameters than CO2 fluxes such as the boundary conditions but this has not been applied in our study. If the reviewer ask about what could be the parameters influencing the agreements between prior/posterior uncertainties and prior/posterior misfits, please, refer to the discussion above.