We would like to thank the 3 reviewers for their helpful comments. The manuscript has been revised according to the referees' comments. We also added details about the uncertainties of instruments, and a new comparison between the 2010 CO2 posterior estimates in LA County and the inventory ODIAC. The difference is 18%.

Referee #1

(1) The authors highlight significant changes in CO and NOx emissions between 2002 and 2010 (40% decreases). Yet the posterior estimates in 2010 were lower than the NEI05 by a similar magnitude (30-40%). Are the differences driven by systematic errors in the bottom-up inventory, or expected changes from air quality management efforts and recession in 2008? Based on your results, it would seem that the latter is more likely. Please address this issue more fully in the results and discussion. It may help to describe year-to-year posterior results first (section 3.4), and then make comparisons to the bottom-up inventories (section 3.1-3.3).

The fact that the differences between the 2002 and 2010 posteriors and between the NEI 2005 prior and 2010 posterior are similar is coincidental. While the difference between NEI 2005 and the 2010 posterior is driven by errors in the NEI as well as changes to emissions between 2005 and 2010, the difference between 2002 and 2010 posteriors is driven a) by differences in the observations in the 2 field experiments and b) potentially by uncertainties in the model. We used NEI 2005 as a prior for the inversion of the 2002 flight. The NEI 2005 prior inventory didn't have large changes compared to the posterior in 2002, indicating that the NEI 2005 is a reasonable inventory for surface emissions in 2002. We didn't discuss the differences between the 2002 posterior and the NEI 2005 inventory to avoid repetitive text and because it did not provide novel information. We added comments in section 3.4 and the conclusion to address this issue.

(2) In the last line of the Abstract, mention is made to the importance of spatial patterns of emissions on atmospheric forecasts. Yet the connection could be strengthened on pages 31461-31462 where WRF-Chem modeling is discussed. Are the down-scaled CARB10 emission estimates the same as the inversion fluxes (CARB10 not shown in Table 3), and is this why inferences on spatial effects can be made? How did O3 performance improve in different parts of the basin? The spatial effects on O3 are one of the more interesting aspects of this study.

The total flux and the spatial distribution of ozone precursors are the 2 variables that differ between the different WRF-Chem runs. In terms of the average bias between simulated and observed concentrations of CO and NOx (or NOy), the WRF-Chem runs with gridded CARB 2010 emissions are close to those using the CO and NOx posteriors. Hence, we cannot explain the differences in correlation between simulated and observed CO and NOx in WRF-Chem based on differences in total emission flux. Therefore, the spatial distribution should play an important role in the correlations. Furthermore, while the gridded CARB inventory has a smaller bias in CO and NOx than NEI 2005, WRF-Chem simulations using CARB 2010 emissions have a larger bias in ozone. The lowest ozone bias in WRF-Chem is obtained by using the posterior estimates, which also produce the best correlations between simulated and observed ozone mixing ratios. Therefore we suggest that the spatial distribution should play an important role in the

simulated ozone chemistry. However, we decided not to go into details of the chemistry because it would be beyond the scope of this paper. Our intention in presenting these results is to highlight the important relationship between emission estimates and air quality.

We clarified this point in the discussion. We also use the word "suggest" instead of "show" in the abstract since we don't give a detailed explanation of the WRF-Chem results.

(3) The use of CARB08 in bottom-up comparisons and CARB10 for WRF-Chem modeling is confusing (section 2.2). Why is CARB10 not used throughout the analysis?

We used CARB 2008 in the paper because it is an official CARB product, and we wanted to compare the posterior to official regulatory data. The CARB 2010 is a projection based on CARB 2008 and therefore might include additional uncertainties. The difference in emissions between CARB 2008 and the projection in 2010 is only a 9% reduction in CO and NOx emissions. Our conclusions about differences between CARB 2008 and the posterior inventory are identical if CARB 2010 had been used instead. The county-level CARB 2010 inventory had to be gridded and combined with other source of information to provide VOCs emissions in WRF-Chem. We chose the CARB 2010 projection for this purpose, as this should be the best product available from CARB for the CalNex time period. However, since this product has undergone several modifications (notably downscaling to a 4x4km grid), we should have talked about a "gridded CARB" inventory instead of "CARB 2010" to avoid confusion with CARB 2008. In the revised version of the manuscript, we refer to "a gridded CARB inventory", meaning the CARB 2010 projection product gridded for WRF-Chem. Those corrections and explanations have been added in section 2.2 and the discussion.

(4) A section on how anthropogenic VOC emissions are estimated is needed earlier in the manuscript, preferably in the Methods section. First mention is in the second to last sentence of the Conclusions.

We added details about the VOC emissions used in section 2.2

Specific comments

(1) On page 31445, line 19, how is NOy calculated when NO3- is not available? As mentioned, this appears to be an important consideration in the eastern part of the basin.

If NO3- is missing, then NOy uses the remaining components (NOx, HNO3, organic nitrates) to calculate a concentration. In this case the NOy concentration is underestimated but will mostly affect the eastern part of the basin. Each flight used in the analysis included NO3-. We didn't examine the detailed impact that missing NO3- data in portions of any given flight have on the inversion results. However, missing NO3- data can result in the inversion calculation underestimated NOx emissions by up to 40% in the eastern part of the basin. NO3- data was not available for the 2002 flight, which might contribute to an underestimate of the decrease in NOx emissions between 2002 and 2010. We clarified this point in section 2.1 and 3.4.

(2) On page 31446, line 6, what are the other sources of uncertainties? Please list.

The major sources of uncertainties come from WRF and FLEXPART: uncertainties on the wind speed, wind direction, parameters in the PBL, linear interpolation in FLEXPART, and FLEXPART's treatment of turbulence in the PBL. We added to the sentence mentioning all of these potential uncertainities.

(3) On page 31447, line 14, why are 3 different meteorological configurations needed? Purpose not clearly stated.

3 different meteorological configurations are used in the paper to have a better estimate of the uncertainty introduced by the meteorological models on the inversion process. It is particularly difficult to estimate this uncertainty if only one model is used. Table 1 shows that each meteorological model configuration produces time series of CO with about the same level of correlation between models as that between any given model and the observations. We clarified this point in the text in section 2.3.

(4) On page 31451, line 18, does the flux ratio inversion method account for differences in the relation between CO and NOy with CO2 by source category? More description would be helpful.

Based on the results of Brioude et al. (2012a), the answer is yes. The ratios at the surface are not homogeneous, but vary with the predominant source category within each grid cell. However, the method cannot distinguish different activity sectors collocated in the same grid cell.

(5) On page 31453, line 26, language is confusing and seems to imply that CO emissions are higher on weekends than weekdays, though Pollack et al. (2012) found the opposite.

We have corrected the sentence in the revised version.

Referee #2

1) It is not clear why the authors chose to use a particle dispersion model in the inversion while having access to a full chemical transport model (at various configurations). It seems like NO/NO2 ratios are dramatically different between emissions and ambient concentrations, and general reactivity of these species should be important. In particular, this assumption seems like it needs more support: "changes in NOy are interpreted as changes in NOx emissions."

The first reason is that there is no existing adjoint model for WRF-Chem. Secondly, using an off-line model like FLEXPART allows the use of configurations that are not necessarily available in WRF-Chem. Using FLEXPART also costs much less computation time. Using a hypothetical WRF-Chem adjoint would have been more difficult technically for 6 flights for 3 different configurations, in addition to the test runs. We added a sentence about this issue in section 2.3

2) What is the reason for subtracting background levels from measurements? What were the background levels? Where they constant?

As explained line 11-12, page 31450, FLEXPART doesn't assume any background value. Therefore, background values have to be subtracted from each flight. The background values are calculated for each flight and therefore vary from one flight to another.

3) It is not clear if inversions were done on each flight independently. If so, it might be worth commenting on the variability in the posteriors from each.

Yes, the inversions are applied to each flight independently. This is explained on lines 1 to 12, page 31452. This is why we were able to derive the uncertainty from a single-flight based inversion in the paper.

4) I would like to see more discussion on the applicability of this method to larger scale. The authors briefly mention that single flight inversion are possible on mesoscale given careful flight path planning (and probably favorable met conditions), but I am not convinced that this method picks up anything more than strong local sources. In fact, it might overestimate the influence of local sources depending on how exactly the "background" values came out. The interaction of local and transported air pollution is very important at current lower levels of emissions and seems to be a weakness of this methodology

The uncertainty in defining chemical background values for each flight is taken into account in the inversion by adding this uncertainty to the uncertainty from the instruments. However, selecting a background value from each flight, based on screening the measured chemical data, is probably more reliable than using a background value from a chemical transport model, because the model can also be impacted by the uncertainties from surface emissions at larger scales. The derived background values are rather homogeneous within the LA Basin. The pollution plumes were sampled on the aircraft only a few hours downwind of the source. We didn't see any clear signature of background value gradients during a given flight. Furthermore, changes in background values can be recognized by differences in chemical signatures (i.e., relative concentrations of related tracers).

We agree that the uncertainties for weak emitters at the surface are larger than those for strong emitters. However, the present method has been applied successfully for Houston. Brioude et al (2011) have shown that it is possible to extract information at mesoscale, even for weak emitters, using such techniques (for instance, trends in highway emissions over Houston in Brioude et al., 2011).

We are able to select a background value for each flight because the flight patterns are highly localized in the basin and always sample air masses upwind of the sources. Of course, the same approach cannot be used for an inversion at larger scale using satellite data, for instance. Either a large scale FLEXPART run, or background value from a third party chemical transport model, would be necessary to apply the inversion method at larger scales. We added a comment about this issue in the discussion.

Referee #3

General comments:

. The NOx emissions constraints are based upon the assumption that daytime NOy is a conservative tracer, whose variations can be exclusively ascribed to errors in NOx emissions from the preceding day. This is a convolution of a few assumptions that warrant further consideration. That the effect of NOx lasts only a day is addressed in a separate comment below. Now, consider the remaining assumption that NOy is (a) conserved and (b) governed exclusively by NOx emissions. Given that these assumptions are repeated from earlier work (Brioude et al., 2011), but not tested nor supported by literature citations there nor in the present manuscript, I strongly recommend that it be evaluated at this stage. The modeling effort would be minimal – simply perturb the NOx emissions on one day in the LA Basin in WRF-Chem, and then track how NOy changes.

The recent paper of Pollack et al. (2012) has thoroughly analyzed changes in NOx emission using NOy measurements. Since the aircraft sampled pollution plumes only a few hours downwind of the sources, NOy is nearly completely conserved. We found good agreement in our weekday vs weekend NOy emissions with the results from Pollack et al. (2012). Therefore, the assumption that variations in NOy concentrations are the consequence of NOx emission variability is a valid assumption. We added a comment in section 2.1

The validation effort is a bit limited. The reasons for restricting the comparisons to specific times and locations doesn't really seem justified given that the authors extrapolate their emissions constraints to much broader temporal (e.g., daily to annual average trends) and spatial scales (e.g., 31454.9) than they allow to be included in the validation tests using only aircraft data from 10 am to 6 pm LT between 200 and 700 m. If we are to believe the emissions constraints are valid at broad scales, then they should be evaluated at such scales as well. Are there other observations from CALNEX (e.g., surface monitoring) that could provide additional assessments of the top-down emissions estimates? Using the same observations that were used in the inversion itself is not as stringent of a test. Even testing the inversion results from one flight relative to observations from a different flight would not reveal some forms of systematic bias.

We assume the reviewer is referring to the comparison in the discussion section with WRF-Chem. No restrictions were applied on the flights used in the inversion. All the points from Figure 1 were used in the inversion. Furthermore, Table 2 presents the comparison for the 3 weekday flights and shows similar differences as the localized comparisons presented in the Discussion section. Comparisons were restricted to a specific region of the basin in order to reduce any bias from the spatial distribution sampled during each particular flight. The specific times were chosen for chemistry purposes when the inversion results were compared within WRF-Chem. The fluxes reported in Table 3 and 5 are for daytime only, the time period used in the inversion. However, besides the daytime-only fluxes in Table 3 and 5, we also extrapolated the CO2 fluxes to a yearly average in Table 6. We think that annual average values are the easiest to compare to other estimates, either from existing inventories or published studies.

We agree with the reviewer that the validation between the simulated and observed time series of CO and NOx cannot be considered independent as we used those observations in the inversion. Therefore, we compared the simulated and observed time series for CO and NOy as follows:

For a given weekday flight, we averaged the posteriors found from the 2 remaining weekday flights and used this average posterior (based on 2 flights) to simulate time series of CO and NOy for the given flight. In this way, the simulated time series and the measured time series are completely independent. We applied the same procedure for the other 2 other weekday flights.

The errors found are:

CO error(ppbv)	WRF-Chem 3.1	WRF-Chem 3.4	WRF 3.3	Ensemble
Posterior	-2.4	6.9	14.4	6.3
NOy error (ppbv)	WRF-Chem 3.1	WRF-Chem 3.4	WRF 3.3	Ensemble
Posterior	-2.2	-0.88	0.17	-0.97

The average error found for CO using the posterior from 2 flights is similar in magnitude to the one found using the posterior based on 3 flights (13 ppbv) (compare table above to Table 2 in the paper). The error in NOy is negative (-0.97 ppbv) when using 2 flights, instead of positive (0.9 ppbv) when using 3 flights. However, this procedure shows that the error using a posterior based on 2 flights is still lower than the error in NOy using NEI 2005. We added a comment in the discussion section of the revised version.

The 4 and 12km WRF-Chem runs using the posteriors were also validated using measurements from the CalTech site. This analysis will be presented in a separate paper (Ahmadov et al., 2013, in preparation). It is particularly difficult to simulate the meteorology at this site. Using the WRF-Chem 4km run, the bias in CO was 28 ppbv or an overestimation of 11%, the bias in NOx was 4ppb or an overestimation of 33%, and a bias in ozone of -6.1ppb or an underestimation of 10%. Using the WRF-Chem 12km run, the biases are 5.3ppb (+7%), 1.6ppbv (+12%) and -4.2 ppbv (-6%) respectively. Those results are based on CalTech daytime measurements from May 14th to June 15th. 30% of the NOx measurements were missing. Those results show that the measurements at the CalTech site were in good agreement with the 4 and 12km WRF-Chem runs using the CO and NOx posteriors.

Specific comments:

• 31444.1: This seems a bit misleading, as remote sensing can also sample pollutants at different distances downwind of a source. There have been many papers assessing plumes from power plants (e.g., Valin et al., AMT, 2011; Wang, Streets et al., ERL, 2010).

This sentence was not a comparison with satellites. We moved this sentence before any reference to satellite measurements.

• 31446.6: Can the authors more quantitatively assess, rather than assert, the relative magnitudes of the sources of error?

The variability among the WRF runs is about 15%. This variability comes mainly from uncertainties in the wind and PBL height. FLEXPART can add additional uncertainties due to the spatial/temporal interpolations within the model, and uncertainties in the turbulent mixing scheme. Based on a recent study on HYSPLIT, STILT and FLEXPART, the variability between Lagrangian models is about 5% on average (Heggarty et al., manuscript in preparation). Therefore, we estimate the overall uncertainty from the models to be on the order of 20%, larger than the uncertainties from assumptions made about the observations. Since we didn't have any way to evaluate precisely the uncertainty from FLEXPART, we decided to remove any reference to uncertainty estimates from the abstract and conclusions.

• 31448.27: Is comparing surface fluxes estimated using 24 h and 48 h sufficient to quantify the error? Instead, aren't at least three tests necessary to show that the results are converging? Otherwise, we might just as easily conclude that each additional day back considered would change the result by an additional 5%. It also seems hard to rationalize the conclusions. For CO and CO2, wouldn't it depend upon the meteorological conditions surrounding the individual flight? It just seems that in stagnant conditions, emissions within the domain from much further back than 24 h could have an influence.

Yes, stagnant conditions could affect the 24h inversion used in this analysis. We extended our test to 72 hours using the WRF 3.3 configuration for which 72h time periods were available for 5 of the 6 flights. We found that restricting the inversion to 24h trajectories might overestimate surface flux by 6.3% after 48h and 7.7% after 72h. 72h trajectories don't modify significantly the results using the 48h trajectories. Stagnant conditions were limited during the flights used in the inversion. We added a comment in Section 2.3 and in the conclusion. We also explained that we constrained the trajectories to 24h for technical reasons, as the WRF-Chem runs used in this paper were limited to 36h.

• 31449.6: An adjoint is unnecessary, or is FLEXPART equivalent to the adjoint of a Lagrangian model? I think it is really the latter, as time-reverse Lagrangian models are considered to be exactly that (e.g., Pillai et al., ACP, 2012, and references therein). So it is perhaps more precise to say "adjoint of WRF-Chem".

We agree that this sentence is confusing. We have modified the sentence as " … so that an adjoint model of WRF-Chem is unnecessary to apply an inverse modeling technique.

• Section 2.4: Unless I missed it, there doesn't appear to be much about the prior or observation error statistics. How are these determined? Are they assumed to be uncorrelated? Are observations randomly sampled to ensure this as in previous works by Brioude?

We added details about the uncertainty used in Section 2.4. The observation uncertainty comes from the uncertainty of the measurements and the uncertainty in defining the background value. We assumed a prior uncertainty of 100% before applying the L-curve criterion. Unlike Brioude et al. (2011), we used all the measurements in the inversion process. However, we didn't find any significant difference by combining the measurements and using a random subsample like in Brioude et al. (2011).

• Section 2.4: Restriction of the inversion to only consider sources for which their is already a significant emission could bias the inversion. It would not be possible to use this technique to infer a source that was not present in the initial inventory. This should be recognized, and if the authors can assume that the only errors in the inventories worth discovering are adjustments to the magnitudes of known sources, that should also be justified and explained further.

It is correct that completely new sources within the basin would not be discovered by this technique. The most likely sporadic sources in a well-developed area like the LA Basin are wildfires. No fires were observed during the 6 flights. Furthermore, the grid cells used in the inversion cover a large area in the basin. We think that all the major anthropogenic sources were taken into account in the inversion. We added those comments in section 2.4

• 31454.15: The wording here is a bit odd. I think it may be clearer to say "modified the spatial distribution of the CO surface fluxes compared to the prior," because the prior itself has not actually changed. If the spatial distributions are shifting though, it does raise some concern about ruling out the possibility of a missing source, as mentioned in the previous comment.

We agree and we used the correction suggested by the reviewer in the revised version

• 31455.23: It would also be useful to compare these slopes to the a priori simulations, to see what the improvement has been following the inversion.

This is a good suggestion. Using the NEI-05 inventory, the simulated CO/NOy slope is 10.4 for weekday and 10.3 for weekend. The slope is 11.8 in the 2002 measurements. These results confirm that the CO/NOy ratios have also improved compared to the ratio in the NEI-05. We added these details to the revised version

• 31462.9: Why is the correlation higher for the 12 km case than the 4 km case?

The correlation is higher for ozone only. We don't have a definitive explanation. However, the fact that only the ozone correlation is higher for the 12 km case might be related to chemistry processes and how these interact with complex terrain. At 12km grid spacing, the terrain is smoothed out and has less impact on chemistry than the 4km case. It is possible that the 4 km case underestimates mixing.

• 31457.9: Could the authors discuss a bit further the sectors contributing to CO2 and why there is less of a weekend effect for these species relative to CO and NOx?

Beside mobile and residential sources, CO2 is also emitted by industrial sources. The weekend effect is less pronounced for CO2 because the mobile source portion of the CO2 flux is less important than it is for CO and NOx emissions.

Figure 5: Can the authors comment on the apparent increase in weekend emissions of CO2 in the San Diego region?

The San Diego area was taken into account because a small portion of a flight offshore was sampling air masses downwind of the San Diego area. However, since this area was under sampled, it is also associated with large uncertainties in the CO2 posterior (see Figure 5, for instance). Therefore, weekday/weekend variations over this area are highly uncertain.

• 31463.2: NH3 is not a conservative tracer, so extension of these methods to this species are not clear.

This is correct. NH3 will require a special treatment to be inverted. We will remove NH3 from the list.

• Abstract and discussion: mention forecast, but not sure if really mean forecast or reanalysis.

We use the term “simulations” in the revised manuscript.

Editorial comments

The editorial comments were taken into account.