Our responses to reviewers are in blue.

REFEREE #1
Received and Published: 11 April 2014

General Comments (GC):
This manuscript investigates on estimating the methane emissions in California by applying inverse modeling technique and utilizing atmospheric CH_4 observations from the CalNex aircraft campaign. The results are compared with other studies which use different inversion methodologies, and are summarized. Additionally the study uses the satellite (GOSAT) observations to assess its ability to constrain methane emissions in California; particularly in the Los Angeles Basin. The study is further extended to assess the efficacy of future satellites, using observation system simulation experiment (OSSE) results. The model described in the paper is GEOS-Chem with 1/2 degree x 2/3 degree horizontal resolution and uses EDGAR v4.2 emission inventory. The manuscript is organized well and is written concisely and clearly; hence easy to follow in most of the cases. The topic of the study is certainly within the scope of ACP.

However, I do have certain comments. The main concern is the robustness of the inversion results that is sensitive to the choice of a priori and its uncertainty. Unfortunately, there is also no real discussion on potential reasons for seeing such a large discrepancy in emission estimates (between a priori and optimized fluxes + between the inventories). Changing the confidence in a priory by 25 % alone has resulted about 7.5 to 10 % change in estimated California emissions of 2.8 Tg yr^{-1} (see Section 3.2, 2nd paragraph). Transport related errors constitute another issue anyway (see the Specific Comment). What happens if using CARB as a priori in the same set-up? In that case, I don't have the reason to believe that the estimated emission will be as high as reported here.
We add two paragraphs to section 3.2 discussing the sensitivity of the inverse results to a priori specification.

I am curious to see the improvements (i.e., reduction in model-observation discrepancy) when using optimized fluxes in the GEOS-Chem forward model. I assume that these simulations are already performed (see Section 5).
We include optimized model concentrations as an additional panel in figure 2.

An independent evaluation (other aircraft or satellite data) will be of great help here to support the robustness of the results.
We don't have data for this purpose other than those used in the paper.

Another concern is regarding two citations which are not yet in the stage of "accepted" or "published" – Wecht et al., 2014 (the year is also wrongly cited in the text) and Santoni et al., 2014 – the issue here is that these citations are largely used in the present manuscript to compare the results and the methodologies. I recommend the paper to be published after considering the above and following comments.
The Wecht et al. and Santoni et al. manuscripts are included in this response to referee comments.

Specific Comments (SC):

p. 4121: "In Wecht et al. (2013), we present . . ." + "Santoni et al. (2014) previously.." Please see the comment above (GC) and update the citation + text accordingly. Since these citations are used many times in the manuscript, it is worth to check the entire text and modify accordingly.
We include Wecht at al. and Santoni et al. manuscripts as part of this response to referee comments.

p.4124: "CARB only reports statewide totals. A gridded version of the CARB .." Do this study use gridded version? If so, please specify the resolution.
We clarify, "CARB only reports statewide totals. A gridded inventory scaled to match CARB totals is available…"
We write at the beginning of section 2.2 that "A priori anthropogenic emissions in GEOS-Chem are from the EDGAR v4.2 global inventory at 0.1^ox0.1^o resolution…"

p.4126: "The middle panel of Fig. 2 shows . . . discrepancies in patterns that point to errors in the EDGAR emissions." I am less convinced here. By looking at the patterns in Fig 1 (top left panel) and Fig 2 (left + middle panel), I have the strong feeling that (model) transport related errors are more pronounced here rather than flux errors, provided that the prevailing wind could be from west. This could be the reason on seeing better model to observation match in some pixels in the South (Fig. 2). I highly recommend authors to comment on this.
We add discussion of model transport errors, particularly associated with vertical mixing in the planetary boundary layer, to section 3.1. In section 3.2, we now include a discussion of model errors and a priori sensitivity in the inverse solution.

p.4130: "The MLR best fit has an R2 of .." I am a bit confused here. Are you talking about R2 averaged over all grid squares? Please clarify.
We remove the MLR discussion from the paper.

p.4130: "..we examine their combined value for constraining.." As far as I understood, this study does not use GOSAT+TES combined observations, as observation vector (y), for the inversion. TES data are used only for the tropospheric background correction. Please clarify.
We clarify that TES observations are excluded from the observation vector, y, and we add performing a free tropospheric background correction "is necessary to ensure that a free tropospheric model bias does not impact the inverse solution."

p.4132: "Figure 5 (right) shows the optimized correction . . . GOSAT observations" It is much helpful if you also include the "GEOS-Chem a priori" (forward) simulations on interpreting these inversion results. The middle panel (TES observations) can be omitted if it is not as a part of the observation vector, y and only used for subtracting the mean bias.
We retain the TES observations to show their spatial distribution. As noted in the text, the background correction supplied by TES "is necessary to ensure that a free tropospheric model bias does not impact the inverse solution."
In figure 5, we add a priori GEOS-Chem concentrations corresponding to each GOSAT observation.

p.4132-4133: Section 5. Please include figures to support your results, particularly the spatial

plot of synthetic observations representing true atmosphere.
We include images of DOFS from the OSSEs (Figure 4). We note that synthetic observations are generated from the same model fields that produced the concentrations in Figure 2 (bottom left). We feel that showing the synthetic concentrations would overstate what is done in the OSSE, which is mainly a statistical exercise.

REFEREE #2
Received and Published: 12 April 2014

General Comments (GC):
This paper presents an Eulerian methane inversion study for the state of California at a high spatial resolution for the period May 1-June 22 2010, based largely upon measurements from the CalNex aircraft campaign. The resultant methane fluxes are found to be significantly higher than those predicted by either EDGAR4.2 or CARB emission inventories, a finding consistent with multiple previous studies. Flux estimates are also carried out based on satellite data (GOSAT and TES) from the same period, which are found to be unable to significantly constrain the fluxes. An OSSE is carried out to assess the applicability of the planned satellite sensor TROPOMI and the proposed geostationary mission GEO-CAPE, both of which were found to be able to constrain the methane fluxes as well or better than a dedicated aircraft campaign. Overall the paper is very well written, and the arguments are clear and well laid out. Despite this, I have two significant reservations about the paper in its current form.

The first has to do with its heavy reliance upon and reference to not-yet-published results. Without being able to refer to the more detailed methodology of Santoni et al. (2014), which has been submitted to JGR, it is difficult to assess the results. The data upon which the entire study depends, namely the measurements of the CalNex aircraft campaign, are introduced only briefly, and never really shown. (Figure 2 doesn't really give an idea of the density or timeline of the measurements - it would be nice to see a plot of the flight paths.) Furthermore, the description of the modelling system refers heavily to Wecht et al., 2013/2014 (<- this should be changed consistently to 2014 Fixed), which apparently has been submitted somewhere, but certainly cannot be found at this point for further information. This may well resolve itself over the course of the editorial process, but at the moment it is troublesome.
The Santoni et al. (2014) and Wecht et al. (2014) manuscripts have been included in this response to referee comments. Wecht et al. (2014) has been accepted for publication in JGR-Atmospheres.

The second problem relates to possible errors with the transport in the model. Although the model is being run at a fairly high resolution (~50-60 km), this is not necessarily sufficient to resolve many mesoscale transport effects. My first thought upon comparing the distributions in Fig. 1 and Fig. 3 was that the main inland red area most likely corresponded to a topographical feature. Not being overly familiar with the geography of California, I consulted an elevation map and found it to be a near-perfect match with the Central Valley. In mesoscale modelling it's

common to see "lakes" of tracers pooling in valleys, and persisting for quite some time under some conditions. This can be difficult to reproduce with a coarser model, and may partially explain the high RSD values near the surface in this region. But more telling than having a higher standard deviation between model and measurement, an inability to represent the transport over such complex terrain would likely result in a systematic offset, which would be interpreted as a mismatch in the fluxes. If the model were unable to simulate (for example) the pooling of tracers in the Central Valley, the inversion would respond by increasing the posterior fluxes in this region, which is exactly what we see in Figure 3. There is no assessment presented to convince the reader that the simulation of the transport over such complex terrain is actually sufficient to allow for flux inversion: perhaps here some comparison of simulated and measured meteorological parameters would be warranted. Surely CalNex measured more than just methane?
We expand and clarify the discussion of even pressure-weighted sampling, noting that, "This even sampling mitigates the impact of vertical transport errors, such as bias PBL height, that lead to systematically biased model concentrations near the surface."

Related to this (and transport errors in general), the error in the simulation of the planetary boundary layer is discussed in some detail, and the use of weighting of data points to ensure that the region from 0-2 km is evenly represented seems valid. I presume this even sampling is pressure-weighted rather than altitude-weighted? The explanation at the end of section 3.1 does not make this entirely clear - some explanation of the methodology is lacking.
We add phrases to the paragraph in question clarifying that the even sampling is pressure-weighted and explaining that, "However, 79% of the observations between 0-2 km are in fact below 1 km altitude and a PBL bias would cause a model underestimate unrelated to emissions."

In general, it would be nice to have some (graphical) idea of the distribution of the flight data. What does it mean that "most" of the observations were under 1 km - is that 55%? 80%? Again, I wanted to see some sort of plot of the measurement locations, but this was lacking. I have access to the EDGAR emission inventories, but found it helpful to see Figure 1 to help understand the results. I do not have access to the CalNex flight paths, but I find this information similarly necessary in order to interpret the results.
Data is described and visualized in detail by Santoni et al. (2014). We include the Santoni et al. (2014) manuscript with this response to referee comments.

Regarding the robustness of the results: the posterior total flux was surprisingly sensitive to the prior flux uncertainty. The fact that the total posterior fluxes increased even further when allowed that latitude implies that the optimized fluxes still have a systematic (low) offset. It might be instructive to see how the model-measurement mismatch looks, before and after optimization (based on a forward run of the optimized fluxes). What about repeating the experiment with the gridded version of the CARB dataset as the prior? If the spatial distribution and/or category breakdown of the posterior result remained consistent, it would certainly lend credence to the conclusions. Once these points are addressed, the manuscript would be suitable for publication in ACP. The subject matter is certainly fitting to the journal, and the study addresses important challenges related to the verification of emissions by atmospheric measurements.
We add two paragraphs to the end of section 3.2, discussing the sensitivity of the inverse results

to a priori specification. We also add a panel to Figure 2 that shows model methane concentrations using the optimized inversion fluxes.

Specific Comments (SC):
p4121 (18-19): Should be rephrased, of course there aren't really observations from future satellite instruments, but rather simulations using pseudo-data representing the expected measurement characteristics of future spaceborne sensors.
Changed to read "…synthetic observations from future satellite instruments."

p4124 (21): inconsequent -> inconsequential
Changed.

p4125 (5-6): How important is the timing of the rice growing season to your results? The flight campaign straddles the onset of the growing season. Can this onset be seen clearly in the measurements? If you're solving for the total flux over the whole time period it may sort of cancel out, but the step function is unlikely to represent reality.
At the end of section 3.3, we add that, "Rice paddies in the Sacramento Valley are sampled by two flights on 11 May and 14 June that straddle the onset of rice emissions."

Figure 1: Please put total flux units on the maps themselves, not just in the caption. Also, the colour scale is in rather a strange unit: why in molecules instead of mass (mg m^-2 day^-1 is often used for methane...)?
Total flux units are included next to the colorbar on each image. The flux unit used is commonly used

p4126 (10-14): An example of where I need to read Santoni et al. (2014) to understand the data selection and free troposphere correction. How big was this correction? How noisy? Perhaps it is presented there, but it is not clear.
The Santoni et al. (2014) manuscript is included in this response to referee comments.

p4126 (24): underestimate -> underestimation
Underestimate is used properly here.

p4127: see PBL discussion above.
See response above.

p4129 (2): nstate -> n state
Addressed. Bring to attention of publisher. Correct in my manuscript.

p4131 (16): Are there spaces between number and unit (km)? (Here and elsewhere - hard to tell, but I think not.)
Addressed. Bring to attention of publisher. Correct in my manuscript.

p4132 (1): This is the first time that the specific dates of the campaign are mentioned - this information should appear much earlier in the paper.
We now include specific dates at the beginning of section 3.1: Observations and error

characterization.

p4132(2-3): Awkward sentence, rephrase.
We change the sentence in question to read, "There are 257 GOSAT and 133 TES observations on the GEOS-Chem grid."

p4133: OSSE is overly optimistic in several ways, not all of which are pointed out. The random removal of clouds (rather than correlated, bunched, persistent patterns) is almost a best-case scenario for cloud screening. (Why not use MODIS or similar?) Dividing measurement errors by the square root of the number of measurements assumes that the measurement errors are uncorrelated, which is unlikely to be the case. The assumption that there would be no significant (and hard to detect) bias between a TIR sensor used to correct the free troposphere and the SWIR sensor is also rather optimistic. Nonetheless, this optimism is somehow the nature of OSSEs, and not the primary focus of this study. Still, some further discussion should be added.
In section 5, we add text stating our assumption that satellite instrument errors are uncorrelated. We note existing text in section 5 that states, "We assume no background bias in the model or observations as this could be corrected through other observations such as a TIR instrument (e.g., TES for GOSAT) or by iterative adjustment of emissions and boundary conditions in the inversion (Wecht et al., 2014)."
We also add a sentence that reads, "These OSSEs therefore represent an optimistic assessment of the capabilities of future satellites."

p4135 (1): I think this should be Santoni et al. (2014)?
All instances change to Santoni et al. (2014).

p4136 (9): underestimate -> underestimation
Underestimate is used properly here.

REFEREE #3
Received and Published: 29 April 2014

General Comments (GC):
This study investigates the use of inverse modeling for verifying existing emission inventories of methane for the state of California. It is found that emission inventories underestimate the emissions. Results using aircraft measurements are compared with those using space borne measurements from the GOSAT and TES satellite instruments. These currently operational missions impose a weaker constraint on the sources than the aircraft data. Significant improvements in the performance are expected for future mission such as TROPOMI and GEO-SCAPE. The manuscript is very well structured and written, which makes it easy and the fast to read. Existing methods are used, that seem to perform fairly well for the case that is studied. Useful results are obtained confirming conclusions drawn in earlier studies that emissions inventories tend to underestimate the Californian emissions, although the attribution of this difference to specific sources remains uncertain. Provided that authors manage to adequately

address the issues that are raised below, I see no reason to uphold publication in ACP.

In my opinion the most critical assumption underlying the results and the conclusions is that the extrapolation of inversion-derived bimonthly fluxes to annual totals introduces errors that are small enough for the difference with the emission inventories to remain significant. A few sentences are spent on this extrapolation step, discussing the seasonal cycle of biological sources. However, the role of the seasonality in energy production, and corresponding emissions from fossil fuel use receives too little attention. It shouldn't take much effort to look up statistics on domestic heating vs air conditioning, and the difference in energy use between 2008 and 2010.
We add text to section 3.2 discussing our extrapolation of emissions during May/June to annual totals.

A related problem is in the comparison between the use of aircraft measurements and GOSAT and TES. The advantage of the latter is that data are available for the whole year. Therefore it doesn't seem fair to use only 2 months of data to conclude that the derived constraints on annual sources are only weak. To estimate the performance of future missions using two months worth of data also doesn't seem defensible. The easiest fix seems to address the performance of the inversion on the monthly time scale (with 2 estimates from 2 months), rather than the annual time scale.
Indeed. We now comment in section 4 on the increased information from the satellite observations if extended to a full year (or multiple years).

The analysis of the capability of the inversion to resolve process specific emissions highlights the difficulty to trace back the diagnosed problem in the inventories to the process level. In the conclusion section, however, it is mentioned that the limited correlation between prior and posterior emission patterns per processes points to problems in the spatial pattern of emissions in the inventories. Some sentences are needed here to put the ability of the inversion to resolve such patterns and to separate between processes in better perspective.
We remove the discussion of the multiple linear regression from the paper.

In addition to information about DOFs and how they compare between the different inversions, it would be useful to quantify the uncertainty reduction that is achieved. As an advantage of the matrix inversion approach this information should readily be available.
In our discussion of DOFS in section 3.2, we add, "Equation (4) shows that **A** reflects the degree to which uncertainty has been reduced in the vector of optimized emissions. Higher DOFS, or larger values on the diagonal of **A**, means that more information is available to constrain the spatial distribution of emissions"

The discussion about the concentration boundaries requires some more attention to the eastern boundary of the domain. If the dominant wind direction is from the west, then the observed east to west concentration gradient should provide a reliable constraint on the emissions in the coastal zone. If winds from the east are important, however, the increased emissions could be a compensation for underestimated eastern boundary conditions, which are more difficult to constrain than the oceanic background. Some further model analysis and discussion is needed to assess the possibility of such a mix up between emissions and boundary conditions.
We add, "Santoni et al. (2014) show that prevailing winds during the CalNex period are from the

west and north-west, where methane emissions around California are small. The free tropospheric background correction therefore effectively accounts for boundary conditions." The Santoni et al. (2014) paper is included in the response to referee comments.

Specific Comments (SC):
P4125, line 23: The spatial pattern of fossil fuel emissions is used to infer which component is most important. However, it should be realized that population density is often used as a proxy to disaggregate emissions in emission inventories. Therefore finding that this is the case need not say much about the contribution of specific processes.
We modify the lines in question to read, "EDGAR spatially allocates emissions using both extraction and distribution data, yet the correlation with population in gas/oil emissions suggests that it is mostly from distribution rather than extraction, which is concentrated in the southwestern end of the Central Valley."

P4130: Does the ranking of process specific correlation coefficients correspond to the ranking of the importance of each process using fixed contributions per grid box?
We remove the multiple linear regression discussion from the paper.

P4132: It is mentioned that the DOFs of GOSAT for the Los Angeles basin are dominated by three near by measurements. Has this been tested by using only these measurements?
Results in the Central Valley, not Los Angeles, are dominated by three observations, "Central Valley correction factors are driven by just three observations located at the southern end of the Valley..."

P4133: What is the basis for the 80% cloud cover of satellite retrievals. Shouldn't it be different for Tropomi and GEO-SCAPE given the difference in footprint size?
We add, "In reality, cloud free observations will not be random, and the different pixel sizes of TROPOMI and GEO-CAPE observations will lead to different fractions of cloud-free observations. However, we use 80% for each as a rough estimate in this study."

P4133: When emission estimates using TROPOMI and GEO-SCAPE are compared with the truth it becomes important whether or not the synthetic data have been perturbed randomly according to So. Has this been done?
We clarify in section 5, "Each element of the observation vector **y** represents the average methane column mixing ratio observed over a GEOS-Chem grid square, including measurement error. When multiple synthetic observations exist in the same $1/2^o$x$2/3^o$ GEOS-Chem grid square, we average them into one single observation with square root decrease of the measurement error following the central limit theorem."

Table 1: The unit in the top row is shifted between columns
Addressed. Bring to attention of publisher. Correct in my manuscript.

Equation 5: I recommend changing the notation such that measurements can easier be distinguished from model results (right now X and omega mean could mean both)
We retain the current notation because it was established in previous publications.

Equation 6: z-hat is missing in the left hand side of the equation.
Addressed. Bring to attention of publisher. Correct in my manuscript.