Point-by-point response to comments by reviewer #1

General Comments: This paper investigates on different filtering mechanisms which can be used to identify regionally representative observations to be utilized in the inverse models. This is highly important especially in case of measurements from a complex terrain such as mountain top. Hence the topic is scientifically relevant and the inverse modelers can benefit from this area of research.

Thank you for the kind note. We have significantly revised the manuscript in light of your comments.

However I have a serious concern about authors' choice on model simulations (a global model- CarbonTracker) to construct the filters although they are aware about the deficiency of global models to represent complex regions- i.e mainly transport (this is later discussed in Section 5). I would consider this as a major drawback of this paper, but I appreciate the attempt (via filtering methods) to exclude observations that are difficult to model and are not regionally representative in order to improve the regional flux estimates.

Inverse models all have problems with vertical mixing and lapse rates may be untrustworthy. CT/TM5 is one of a class of global inverse models that have similar constraints in application and so we feel it is a good model system, and most importantly, the only one that regularly ingests data from the Rocky RACCOON network.

As such, we feel that the lapse rate filter is still a viable direction to investigate. We added some discussion (see 2nd paragraph of intro) that treats this issue. We also included some new lapse rate uncertainty sensitivity tests. See the 3rd paragraph of the Discussion ("The SVLR protocol however...")

I suggest authors to comment on high-resolution modeling efforts towards this direction.

Thank you. We added some discussion of this in several places in the Introduction beginning with the 2nd paragraph. Yes we (authors) did discuss this issue as mentioned in our interactive reply that we posted. In short there are subtleties in using model data mismatch as a benchmark that we thought justified its use in a follow-up study instead. Principally, models tend to give less statistical weighting to measurements from complex terrain, and also successful assimilation of an observation but with incorrect winds is a misleading and frequent issue that needs careful evaluation (the answer might be "better" but wrong winds could easily cause CO2 fluxes to be optimized 180 degrees in the wrong direction). See last paragraph of Discussion.

Also I strongly recommend including a comparison of model simulations and observations with different filters; then one can assess the potential of different filters (to judge whether filter is over or less selective). Besides these I do suggest authors to work a bit on the readability of the paper (sentences are sometimes rather long and difficult to follow); also the sections 2.1 and 2.2 can be shortened (many repetitions). With all these recommendations/ suggestions/comments incorporated, the paper can be published in ACP.

We systematically went through the paper and reduced many of the redundancies in the Introduction (especially section 2.1) and Discussion. More discussion of comparisons to

model is included and a companion paper is focusing on model simulations in response to the filters.

Specific Comments:
Section 1: comment on high-resolution modeling efforts especially for complex terrain. See Pillai et al., 2011 and van der Molen and Dolman, 2007.

Done. See paragraph 2 of Intro.

Section 1 and 2: mainly here is my comment about the readability of the paper

Thank you. We did significantly improve this by cutting out/simplifying much of the text there.

Section 3: Could you please clarify or explain a bit more on how these filters do account for synoptic variability? In case of synoptic events, I would assume that 1 ppm standard deviation criteria would not work. Please comment.

In addition to our interactive reply we added a new paragraph to the end of Sect. 4.3 (synoptic case studies) to clarify this. A 1ppm standard deviation alone is the "SV" filter and does permit for a large majority of synoptic observations to pass as in Fig. 6. Note that these are standardized test. The regular SV filter allows most of those cold front CO2 obs through.

Section 3 and 4: The filters based on a global model with a typical resolution of 1o x 1o can very well exclude observations which contain lots of important and regionally relevant atmospheric information. This is simply because of the transport model deficiency due to its coarse resolutions. Then the filter is over selective and avoids most of the observations (the atmospheric "wealth"). This is a serious issue.

Our results actually show that over-selectivity is actually an issue for the windowing filters (WM, SI) even when the criteria are relaxed/standardized. The model-specific filter (SVLR), as we set it up for this paper, uses only max & min lapse rates from the model, which are typically much larger than measured lapse rates at the station. About 90% of RACCOON measurements were retained using CT lapse rates for the SVLR filter. The more likely issue with CT/TM5 is probably that the lapse rates are way too big (as you mentioned in this review). Our fundamental assumption is that model lapse rates are indiviative of processes or information that can be assimilated at the transport scale.

Technical Comments:
pp 25330: please rephrase the sentence – "Our goal in this study is … carbon cycle inversion models."

Fixed

pp 25342: please indicate clearly – "The 0–4 subset…"- you may have to write 0:00–

Fixed. This was already done in the final ACPD version that is online.

Figure legends are missing for Fig. 3 and 8.

I'm not sure why they didn't show up in your version. I made sure they are in the current version however.

Suggested References:
Pillai, D., Gerbig, C., Ahmadov, R., Rödenbeck, C., Kretschmer, R., Koch, T., Thompson, R., Neininger, B., and Lavrié, J. V.: High-resolution simulations of atmospheric CO2 over complex terrain – representing the Ochsenkopf mountain tall tower, Atmos. Chem. Phys., 11, 7445 7464, doi:10.5194/acp-11-7445-2011, 2011.

van der Molen, M. K. and Dolman, A. J.: Regional carbon fluxes and the effect of topography on the variability of atmospheric CO2, J. Geophys. Res.-Atmos., 112, D01104, doi:01110.01029/02006JD007649, 2007.

Thanks. I was not aware of these papers but read them and added some references to them and others on regional inversions.

Point-by-point response to comments by reviewer #2

General Comments:
This paper is relevant (and practical) to individuals performing atmospheric inversions of biological trace gases such as CO2. It expands available CO2 observations for use in atmos inversions by providing filtering methodology (and comparisons) and does this in an area for which "regionally representative" CO2 measurements are very difficult to get. I would say that the most novel thing about this paper is the attempt to perform model-specific filtering techniques. This is a very interesting direction of research and I think there is a lot of research still to come on this topic.

Unfortunately, I think CarbonTracker and TM5 might not be the best example to be used for this paper, or alternatively, the right metrics of comparison between the model and the observations haven't been identified. Many coarse models have a very difficult to time with vertical transport, let alone in complex terrain, and therefore I think the authors should continue to look for more representative metrics of evaluation that can be used in their "model-specific" filter (lapse-rate in this paper).

Thank you for the review. Clearly all global inverse models will have issues similar to CT/TM5, and we use this example simulation as one that is widely used and relatively easy to modify. The vertical transport deficiency in many models is an issue for us since the SVLR filter uses a metric that is affected directly by vertical transport. Nonetheless we did consider other metrics like horizontal gradients and 'time-step gradients' but among them vertical lapse rates were the best indicator of model discretization. We addressed this point you made by including a new test on the SVLR filter's sensitivity to lapse rate uncertainty. See more below.

Additionally, for researchers attempting to employ the methods, the equations and notation need to be error-free and much easier to follow. I'm only listing "major revisions" because of this last point (math notation) and the fact that I think one figure needs to be redesigned, and I would like to make sure those two things are done.

Thank you. There were notation errors and we fixed them. We also made a supplementary spreadsheet that demonstrates each filter.

Otherwise, I have many comments, but mostly of "minor revision" nature. Overall, timely, stimulating, and most importantly, useful paper.

Thank you.

Specific Comments:
(1) the derivation of a "common" data set for comparison to aircraft data was very hard to follow. Maybe some kind of timeline or spreadsheet comparison of times and associated data (lapse rates), in order to facilitate "seeing" the intersections between the data sets used for the different methods.

We created a supplementary spreadsheet to go with the paper that provides this timeline of CO2 measurements from Carr and Niwot and refer to it at the end of the first paragraph in Section 4.2.

(2) since measurement error is likely small, the 'filtering' presented is really a way to subset observations so that model-data mismatch errors are lowered, i.e. so that obs are consistent w/ behavior that the models can reproduce. Hence future work would probably benefit from more model-obs comparisons, in addition to the aircraft comparisons which were very nice.

We partially agree. Airborne profiles are not the perfect benchmark nor is model-data-mismatch. For example CO2 coming from airflows that transport models cannot reproduce can reduce MDM. If the air comes from 180-degrees in the opposite direction of what the inversion model thinks, even though the mismatch would be smaller, the model would hypothetically have incorrectly attributed the source region of that CO2.

(3) Sections 3.4/3.5 need a review by the authors and/or a statistician. The notation is very difficult to follow and there appears to be a lack of definitions. This is a serious impediment (and serious comment) to anybody hoping to employ any of these methods. See assorted comments in Technical Corrections.

We improved the notation problems and also created a supplementary spreadsheet that demonstrates each filter and can readily be used by any reader interested in applying a filter.

(4) Section 4.2 is very valuable part of this paper. Unfortunately, it seems to read a bit awkwardly as far as the subsetting is concerned. I'm very confused for how the different "filtering" subsets relate to the 24 possible CARR vertical profiles. More details below.

We edited the text of this section to make it clearer.

(5) Section 4.2: A general comment is that the authors should emphasize that 1500 meters of well mixed air is a more stringent requirement than 50 meters (although the highest variability should be near surface) so that the fact that 218 of the 255 flights aren't "useable" doesn't NECESSARILY imply that the 218/255 % of the tower data isn't useable. This would depend on the reasons for the variability in the aircraft vertical profiles (aircraft hitting plumes/thermals of high/low co2 air, etc).

Correct. 1,500 meters of air with little difference in CO2 is more stringent that 50 m with small CO2 difference. But this is justified given that we do not expect air over the plains at Carr to be similar to ridgetop air at Niwot. To make this a more probable means for comparison we set stringent criteria for when we thought there would be high surface layer similarity.

(6) Section 4.3: A general comment on this section. This represents a good opportunity for the authors to hypothesize on further research. For example, the ability of atmospheric models to accurately model storm fronts, topography, complex weather, etc largely a function of grid resolution. Although, fine resolution doesn't imply GOOD MODELING, coarse resolution certainly implies that many features can not be resolved. Therefore, it would seem that the Lapse Rate filter concept would benefit heavily from additional filtering/weighting based upon a model's (like CarbonTracker/TM5) ability to resolve wind speed, direction, vertical gradient etc, in addition to the "onesided" lapse rate test alone. You are more likely to "accept" the observation into assimilation if the model shows it can reproduce a number of features in addition to lapse rate.

Thank you and good idea. We added some additional text referring to recent work using

variable resolution grids for atmospheric transport models (Wu et al, 2011, JGR).

(7) Discussion: I'm still horribly confused by Figure 9. The authors need to reinvestigate how to display this data. Even if is correctly displayed (which I'm not sure of), if you can't figure it out in a few minutes, nobody is ever going to pay any attention to it. I really like "dense" images but in this case, I have to recommend to spend a little bit of time on a different/simpler visualization of this data. This is the figure I'm referring to in the "general" comments.

Very right. I think we have found a better way of displaying the data. See the new figure 9.

(8) Discussion: I like the idea of using model output to help subset the data used but the way in which the CT/TM5 data is used brings up some serious considerations. The CarbonTracker/TM5 CO2 has had some historical issues with surface CO2 being VERY wrong, high CO2 I believe and not just at night under stable conditions. I'm not expert on TM5 but you mention NOAA folks in acknowledgements and thus you have access to this information.

Yes. Actually we are working on this now in a follow-up paper. One question is how much of this large CO2 mismatch between the model surface and true surface CO2 is due to the elevation mismatch between model and measurement site.

If you assume a model's lapse rate is indicative of it's ability to accurately model THAT lapse rate, and you have strange steep gradients of CO2 near the surface you really lose the ability to confidently say what you are trying to say. Furthermore, with increasingly high lapse rates coming out of models, one would have to assume that uncertainty in the modeled CO2 HAS to go up. In other words, if the model is WRONG w/ the lapse rate, you should throw the data out, but if the model is RIGHT about the lapse rate then you are trying to model VERY difficult conditions,
stability, etc.

Interestingly the gradients aren't necessarily that steep. The model CO2 mole fractions are way off (as in 500+ ppm CO2 at model surface), but this occurs for example where the elevation mismatch is 1,000+ meters. So the gradient in the model gets stretched out and reduced some.

The authors should comment first on the TM5/CarbonTracker surface CO2 issue (consult NOAA-ESRL) and possible effect on your lapse rates. I'm not sure of the answer here but I know that most people who read this, and have seen the CarbonTracker CO2 data, will think about these issues. Then the authors should certainly caveat the lapse rate filter by the fact that VERY incorrect transport in the model could essentially allow the data to come into the inversion essentially unfiltered, even when the model might be getting the dynamics very wrong. A recommendation would then be to speculate on additional model output that could be used to evaluate the model's ability to accurately model transport in complex terrain, in other words something to provide a "check" on simply using the lapse rate. I like the direction that the authors are going, but they have to be careful about specific claims and evaluation metrics.

We added some text explaining that model lapse rates may better be calculated higher in the model atmosphere in locations where terrain mismatch between the model and actual surface

is large. For example computing lapse rate at the same elevation that the data were assimilated from.

Technical Corrections:
(1) abstract: change "..terrain are difficult to measure often due to..." to "terrain are often difficult to measure due to..."

Done

(2) abstract: the phrase "standardized to common subset sizes" is too technical for abstract since I'm guessing the reader has no idea, a priori, to know why this is an issue, or what you are even referring to. Maybe the authors should drop that phrase and just talk about it in the text.

Done

(3) Refs on page 4. I would put in some of Thomas Lauvaux regional/microscale paper as well since they belong in this "set" of inversion papers. Look up Lauvaux et al. 2009 and Lauvaux 2011 (ACPD) even though this was probably accepted after your submission. Also, possibly Gourdji, et al 2011 (Biogeosciences). I'm sure there are more, these are just what I'm immediately familiar with. Goeckede 2010 and Lauvaux 2009/2011 are probably most relevant at the scale that is being looked at here.

Done. Added and discussed Lauvaux, Gourdji, and Goecekede on topic of scale representativeness.

(4) Section 3.4:
X(n): is this a sub-daily time series?, daily time series? etc. It appears that "n" is a day but it is implied that X(n) contains subdaily values so please clarify EXACTLY what X(n) is. Same for x(n), the subsample.

Equation 1 has an "i" index which doesn't appear in the expression, I'm confused. Nothing in the summation notation, N, i, appears in the expression.

The definition of x(n) around line 264 is confusing also. x(n) appears to be defined in terms of itself?
I like the rigor with which the authors attempt, but please run by an objective noninformed statistician/mathematician in order to make sure the definitions and equations are clarified.

Fixed. X(n) now represents the original hourly values and X(N) the daily medians of those hourly values. Also changed summation notation to go from i=1 to the total number of days in the series (14).

(5) Section 3.5
Same comments apply from 3.4 w/ respect to X(n) and x(n) notation.

It would also be helpful to officially define X "sub" i. It is indicated that X "sub" i. It is implied that X(n) "sub" i is somehow the hourly mix ratios belonging to day "n" and excluding some set of hours. Please clarify.

Fixed

(6) Section 4.1
line 320: It would appear that the SI filter under all the different hourly subsamples is higher variability, not just under the "complete obs" as written.

Fixed

line 321: Although technically correct, I would change the sentence "Also time-of-day sampling generally has little ..." to "Also time-of-day sampling *alone* generally has little ..." just to emphasize that the authors are only subsetting the times here and no other filtering is occurring. This is probably an important "benchmark" for readers to understand.

Fixed

line 333: Change "on the order of -0.4ppm from the complete set." to "on the order of -0.4 ppm from the complete set, implying a slightly weaker seasonal amplitude than complete set" or something to that effect. I'm worried that "LARGER DIFFERENCE" implies to "LARGER SEASONAL CYCLE" to somebody reading this quickly.

Fixed

line 337: Change "SVLG and SI subsets ..." to "Largely due to the way they are defined, SVLG and SI ..." or something to that effect. They are defined as mechanisms to filter based on stratifications, so this should be no surprise.

Fixed

(7) Section 4.2
line 354: I'm assuming that the WM, SI were tuned (criteria) until the authors got "around" 20 or so common profiles? I'm not clear though how the hourly-stat filter subsets were handled. Were the filtered data sets FURTHER reduced to 20 common hourly profiles? based on individual CO2 gradients of each filtered data set? Might want to clarify this a bit.

Fixed. Specified clearly that the windowing filters were scaled-up and statistical filters scaled-down.

line 352: I'm also a bit puzzled about the need for "standardizing" the comparisons. Although, I assume the authors have legitimate reasons for them, it would be nice to hear them. It would seem that the filters as set up (default) would not admit enough observations to be compared to aircraft? and so the filter had to be loosened in order to have ANY data to make the comparisons with. If this is correct, it would be better to explain the rationale for the filter adjustments in that fashion.

Fixed. See first few sentences of 2nd paragraph in section 4.2 where we now explain that standardizing is done to remove bias from subjectively chosen filter limits.

(8) Discussion

line 454: "were able to identify and retain CO2...". Sort of nit picky here but I don't like "identify", it seems to imply that the authors are identifying CORRECT observations or something similar. All these observations are presumably correct. The filter simply is trying to subset some that represent regional scale variability as opposed to local. Change "were able to identify and retain" to "retained", or something similar.

Fixed

line 457: "diurnal"? Do these case studies represent DIURNAL or SYNOPTIC variability? I would assume synoptic might be more appropriate here?

Synoptic. Changed some of the text in the second to last paragraph of section 4.3 to clarify that diurnal oscillations are overprinted on the synoptic change.

line 470: where does the 86% come from?

Fixed. That number was not updated with the last analysis revision.

line 479: What about night of 17th? big drop there too, was that not synoptic?

Fixed. The dates/time of the last synoptic change was listed incorrectly. Thank you.

line 490; I don't recall, can the authors tell me whether 0-4LT filtering reproduces the seasonal cycle?

We added a new line to Table 3 that shows that 0-4 LT filtering reproduces the weakest seasonal cycle of all filtering methods. I have thought a good bit about this. There is concern about bias in the 0-4 LT filter when used alone because it represents data measured when CO2 respiration is at its peak, particularly during the summertime. Perhaps this may offset or bias the downward descending air.

lines 512:513: Does this mean there was only 1 lapse rate from model for each time of day? So, a 12LT lapse rate from model for winter and summer were combined? It does seem that this needs to be considered more in the future as far as future lapse rate / transport model comparison filters are created. One would envision that there could be very strong differences seasonally. The authors might even elaborate on the actual differences in the lapse rates between winter and summer in the discussion, simply a couple averages of day/night over winter/summer would probably suffice.

Correct. We experimented with diurnally, seasonally, and annually varying lapse rate filters (and combinations of them) and found that they tended to reject too many observations perhaps due to limited sample size from the model data. We discuss this issue about half-way through the Discussion: "we might instead have used lapse rate limits that were seasonally specific ..."

lines 530:531: This comment is important. I have to say that most modelers will look at comparisons of NWR CO2 to 1x1 degree TM5 results in the mountains and say that there is no way you can compare these things. The authors do caveat with these comments, but they should not be taken lightly.

True, but those modelers we've communicated with specifically about this section and passage have indicated that comparing lapse rates and not CO2 mole fractions directly gives us some leeway for this kind of comparison.