To the authors:

I don't mean to be unfriendly but please address the comments sequentially like they have been written out in the review. You've addressed a few particulars of importance in your "general" response but it would make the process go much quicker if you simply addressed each one of my comments one by one. Copy and paste my comments and fill in responses to each of them. Otherwise, I'm forced to take your general response and somehow see how each addresses each of my sequential comments. If/when you are a reviewer, you will see why this saves an enormous amount of time for both the reviewer and author. The underlined comments are those which I do not feel have been addressed at all or not well enough. In particular, the mathematical notation and equations still needs to be cleaned up.

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.

**This has been promised but I haven't seen it yet. I will assume this will be included in final version which is fine with me. This will assist greatly w/ following the methods.**

(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.

**Comment, no response needed.**

(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.

**This still needs review by math/stat grad student or faculty. What is X(n=1) and how does it differ from X(N=1), etc? I would suggest rewriting the notation and possibly adding another variable to represent the median function, to keep the time series separate from the median which is a function of it.**

(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.

**Details following, no response needed.**

(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).

**I'm willing to let this slide a bit as well, more of a comment than a steadfast suggestion of more work.**

(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 "one-sided" 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.

**You seem to have added one paragraph to end of 4.3. This is fine but I think you missed my point. It isn't simply the addition of filters on top of each other to further refine the observation selection. It is the selection of observations by filtering against different met fields than simply lapse rate, based upon the case scenario you are looking at, i.e. wind speed, PBL depth, etc. Lapse rate does capture a number of these but for example, if the lapse rate is too steep, it is assumed that it can correctly capture it since it can simply model it. I can produce a model which does this but is reasonably worthless. Question is, what met fields would you compare to evaluate an overprediction (or underprediction for that matter) of lapse rate?**

(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.

**The new way to display the data shown in the "example" figure looks reasonable to me and assuming the data is replotted in this fashion, I'm fine with it.**

(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. 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. The authors should comment first on the TM5/CarbonTracker surface CO2 issue (consult NOAA-ESRL if ?s) 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.

**I'm satisfied that the author is aware of the potential issues and some caveats are included in the paper.**

Technical Corrections:

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

**GOOD, changed.**

(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.

**GOOD, changed.**

(3) Refs on page 4. I would put in some of Thomas Lauvaux regional/microscale papers 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.

**GOOD, changed.**

(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.

**This appears to have been changed somewhat and is just as confusing this time. Please have somebody mathematically oriented, preferably a mathematics or statistics grad student or faculty, review the notation.**

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.

**Good, the "i" has been removed, however I'm not sure if the expression was really double checked? "I" represents the limit around WHAT median value? X(15) I assume? Since the indices will go backwards from there? That would mean that the weight on the most recent median is 2^-14 and the weight on the oldest would be 2^-1 ? Correct me if I'm wrong but the expressions and equations in this section appear to have been pretty hastily written.**

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 non-

informed statistician/mathematician in order to make sure the definitions and equations are clarified.

**This appears to have been changed somewhat and is just as confusing this time. Please have somebody mathematically oriented review the notation.**

(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.

**This appears to have been changed somewhat and is just as confusing this time. Please have somebody mathematically oriented review the notation.**

(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.

**Good, changed.**

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.

**Good, changed.**

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.

**Good, changed.**

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.

**Good, changed.**

(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.

**Again, a bit tricky to follow still. I'm hoping that a spreadsheet of the data would assist all these sequential filterings. I'm assuming here that each filter was "relaxed" until it was able to generate 20 "common" hourly CO2 measurements w/ the 37 from the aircraft. If this is the case, the wording could probably be improved.**

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.

**I will assume that this was necessary and that further explanation would be a detour in the paper. I'm hoping the spreadsheet will supply the data for reproduction anyways, so I'm happy with that.**

(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.

**Again, details. I would like to see "…identify and retain CO2 measurements despite…" changed to "…identify and retain spatially homogeneous CO2 measurements despite…"**

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

**The point here is that I don't see why you are saying that certain filters have problems with these synoptic case studies BECAUSE OF DIURNAL VARIABILITY. Isn't it the synoptic variability that causes problems? not the diurnal? I don't recall seeing the word diurnal anywhere in your discussion of these case studies.**

line 470: where does the 86% come from?

**No response? And also, it seems that the observations actually sit in the "white" and not the magenta? I could be wrong but it doesn't look like "…lapse rates the model can represent (see magenta colored region in Fig. 9). The remaining 14% constitute measurements with lapse rates the model can not obtain. Are the colors reversed here?**

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

**No response?**

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

**No response?**

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.

**No response?**

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.

**I can see this as being the main point of contention in the paper but I'm fine with the author response and caveats included in paper.**