The Reviewers provided very constructive comments on the paper (Atmos. Meas. Tech. Discuss., 8, C3214–C3216, 2015), so we would first like to thank them for their efforts in improving the manuscript. The following are detailed responses to comments from both of the Reviewers. Note that the corresponding manuscript was also updated to reflect the comments.

## Anonymous Referee #1

The manuscript by Shephard et al. presents detailed validations of the Aura TES retrievals (including ammonia, CO, formic acid, and methanol) using airborne vertical profiles over the Canadian oil sands region. There have been few direct validations of these TES products due to various in-situ measurement challenges. The aircraft observations reported in this study were performed both coincidentally and downwind of the TES footprints, providing good validation opportunities. In addition, the validated satellite ammonia/CO datasets were intercompared with GEM-MACH air quality model. In general, this manuscript is well-written, the validation methodology is appropriate, and it should be published after addressing the comments below.

### General comments:

1. The uncertainties in the aircraft estimated comparison profile were neglected in the validation (page 9519, lines 23-25). While the instrumentation errors presented in section 2 could be averaged out, the representation errors should be more carefully discussed. Vertically, mapping the high resolution aircraft profile to the satellite retrieval levels could induce representation errors due to the turbulent nature of the atmosphere, even assuming that the measurements were accurate.

   Yes, we agree completely that the representativeness of profiles with fine vertical structure profiles (likely captured by the aircraft profiles) will not be captured in the coarser vertical resolution. This actually brings up a key point and the merits of this method of performing the comparisons. The main goal of this comparison is to validate the satellite observations themselves given the known coarse resolution of the observations, not to validate whether or not the satellite can capture the fine vertical structure (even though this is interesting in its own right). Also, there are a couple of nice aspects of an optimal estimation retrieval that take into consideration the vertical resolution and representativeness. First, the vertical resolution and sensitivity of the satellite observations are computed from the averaging kernels for each profile, so both the fine resolution aircraft observations and the satellite observations are same coarse vertical resolution to directly validate what the satellite can measure. In addition, the optimal estimation retrieval does provide an estimated vertical smoothing or representative error with the retrieved product (generally generated from an ensemble of model or measurement profiles) (the Shephard et al., 2011 reference has more details).

   In the paper we have the following, but if the point needs to be made stronger please let us know.

   - On page 9518 it states: "Since the main goal here is to validate just the retrieved information provided by satellite measurements it is often desirable to perform a profile

comparison using the satellite observation operator, especially for species with limited information content. This approach provides direct comparisons of the satellite retrieved quantities by taking into consideration the reduced vertical resolution of the retrieved values, as well as removing the influence of the a priori information (i.e. profile shape, etc.) used in the inversion of the satellite observed radiances to concentration values at each level."

Horizontally, the majority of the validated satellite pixels was not collocated with aircraft spirals, and could be up to 35km away (more for ammonia). There could be errors due to the spatial variation of concentrations. The author briefly mentioned that the spatial variability was small for ammonia according to TES data (page 9526, lines 10-13). Is it possible to be more quantitative and estimate the related errors? The aircraft data may be a better choice to estimate the spatial scale of variability than historically satellite data.

The Review again brings up a valid point in terms of horizontal variability, especially in terms of short lived species. As we learnt from this study, over the study region there appears to be more of a "background" (possibly biogenic) source that does not appear to be a very localized so the spatial sampling is not very important over the temporal and spatial sampling of the comparison flights. In addition to the using the TES satellite data themselves over the 240 km transect one can also get an idea of the horizontal spatial variability by looking aircraft flights (and pointed out by the Reviewer). Examples of this are already provided for the comparison flights and as a function of time in plot (b) of each of the single example comparison plots. Again, it is somewhat surprising how little variability there is over individual aircraft flights over the region during this time period. To provide even further insight on this topic, we removed the distance criteria and including all the TES pixels along the transect for each species and repeated the summary comparison analysis. The overall summary results did not change significantly providing further indication that under these particular conditions over this region and within the aircraft afternoon flight time, the horizontal variability is not very important.

2. During the comparison, all TES pixels seem to be treated equally regardless the spatial/temporal differences from airborne observations. The comparison errors are presumably larger for profile D (aged air mass, large spatial/temporal mismatch) than for profile A (ideal spatial/temporal match). Are there larger discrepancies (TES - aircraft) for larger spatial/temporal mismatch?

Yes, we agree that intuitively one might think this is the case, especially for some of these short-lived species. However, as noted in the response to comment 1, the spatial variability during over the flight duration and spatial coverage is not very large, as shown in both the aircraft flight data and the TES transect data. Thus, there is really not much of an increase in TES-aircraft differences for larger spatial or temporal mismatches as one might have expected.

3. Are the oil sands strong sources of CO and ammonia? If so, the spatial variation long the TES transect should be substantial, because pixel 11 and 12 are immediately downwind of sources, while the other pixels cover boreal vegetation. The dispersion and losses (for ammonia) should to be considered when comparing TES profiles and downwind/aged airborne profiles.

This is a very good question. Before we started looking at the satellite (and now the aircraft) data we really did not have a good idea on how strong the industry specific emissions over the oil sands compared with the "background" boreal forest region. We have performed some very preliminary "climatologies" using 2-years of TES special observations that cover a 240-km transect centered on the oil sands (see slide below), and so far we don't see any significant increases in the NH3 or CO concentrations directly over the oil sands mining region (maybe hints of a deficit for NH3 possible due to aerosol formation with the presence of large NOx and SO2 emissions around). This would support the argument that there is not a strong oil sands signal for NH3 and CO.



## "Climatological" Oil Sand Region Satellite Observations:
### Period 2012/07/14 – 2014/04/03   (Warm Season: M-J-J-A-S)

| | Ammonia ($NH_3$) | Methanol ($CH_3OH$) | Formic Acid ($HCOOH$) | Carbon Monoxide ($CO$) |
|---|---|---|---|---|
| VMR (ppbv) | ~1-2 | ~4-5 | ~2-3 | ~150-200 |
| Pressure (hPa) | 850-900 | 825 | 825 | 680 |
| DOFS | 0.65 | 0.6 | 0.75 | 1.1 |

## Specific comments:

Page 9507, line 3: what are these processes of ammonia evaluated in the model-satellite intercomparison?

This is a general introduction statement to cover any processes. An example process would be the vertical transport of NH3 in the model as the satellite is sensitive to NH3 above the surface in the middle of the boundary layer.

Page 9508, lines 20-21: formic acid is the dominant source of what? Please clarify this sentence.

Good catch. I must have lost the text when doing some cut-and-pasting. The sentence in the paper now reads, "Formic acid (HCOOH) is a dominant source of atmospheric acidity, and is the dominant contributor (60-80 %) to acid rain over boreal forest regions (i.e. surrounding the oil sands operations) (Stavrakou et al., 2012)."

Page 9514, lines 9-11: the authors have the data to quantify the dispersion/aging processes of the air mass. Is it possible to assess the error induced by spatial/temporal mismatch?

As noted above in response to questions 2 and 3, given the lack of spatial and temporal variability over this region during the flight of the aircraft, the error will be small. For example, there is little difference between the results with and without applying the distance/temporal criteria.

Page 9515, line 25: does the "delay time" mean response time? If it is only a constant time delay, the measurements should be just shifted by 2s.

Yes, "response" time is a better term. The "delay time" has been changed to "response time" in the revise paper.

Page 9517, lines 7-10: the propagation of all these errors (10-15%, 5-10%, 10-15%,10%, and < 5%) seems to be larger than 20-25%. Is it possible to show how it is propagated?

This is a good point. Looking at it in more detail, it is not really possible to propagate the uncertainties and trying to quantify like we originally did just adds confusion. In order to be a little more conservative what we decided to do was to provide the biggest uncertainty from the calibration of the response factor and just list that there might be other biases introduced that cannot be reliably estimated. The text was changed to:
"The uncertainty (1-sigma) in the CIMS HCOOH is primarily contributed by the observations is contributed to by several factors including; uncertainty in derived response factors (± 10-15 %), variations in flow, pressure and temperature (5-10 %), transmission through lines (10-15 %), degradation of standards (10 %) and uncertainty in fitting mass spectral peak in software (<5 %). , although other factors may introduce unknown systematic biases that have not been fully quantified. Such factors include, variations in flow, pressure and temperature, transmission through lines, degradation of calibration standards and uncertainty in fitting mass spectral peaks in software. The overall uncertainty is estimated to be ~20-25 %."

Table 1 and 2: are these volume mixing ratios of the a priori or from observations? Please clarify.

These are the observed/retrieved volume mixing ratios.