Response to Reviewers
A chemical transport model study of plume rise and particle size distribution for the Athabasca oil sands

Original Reviewer comments are in normal font, responses in *italics.*

Anonymous Referee # 1

The manuscript by Akingunola et al. presents an interesting set of sensitivity simulations to plume rise modelling from stacks in the framework of the Canadian mesoscale chemistry-transport eulerian model GEM-MACH. The study is timely since simulation of the subgrid plume dynamic processes are still affected by significant uncertainties, as also confirmed by this work, and it is relevant for air quality applications considering the large role that emissions from elevated stacks plays nowadays and will play also in the near future. I found the manuscript generally well written and clear and I recommend publication to ACP after some minor corrections and clarifications as detailed below.

We thank the reviewer for the helpful comments; our responses to the minor corrections and clarifications follows.

1. P. 1, L. 19: ". . .reducing the magnitude of the original surface PM2.5 negative biases by 32%". Would be more clear to specify the range of change: from bias x to bias y.

The text has been modified to read "...reducing the magnitude of the original surface PM2.5 negative biases 32%, from -2.62 to -1.72 μg m^{-3}."

2. P. 1, L. 24: ". . .with 39 to 60% of predicted plume heights . . .". I suggest to specify what the given range is referring to, e.g. rephrasing the sentence adding a compact summary of what are the best and worst performing cases.

Our submitted draft referred to the companion paper's results in the abstract – rather than take examples from that work, we've carried out additional analysis from our scatterplot figure (Figure 8 in the revised manuscript), and based on that analysis, we have modified the sentence in the abstract to read:
"As in our companion paper (Gordon et al., 2018), we found that Briggs algorithms based on estimates of atmospheric stability at the stack height resulted in under-predictions of plume rise, with 116 out of 176 test cases falling below the model:observation 1:2 line, 59 cases falling within a factor of 2 of the observed plume heights, and an average model plume height of 289 m compared to an average observed plume height of 822 m. We used a high resolution meteorological model to confirm the presence of significant horizontal heterogeneity in the local meteorological conditions driving plume rise. Using these simulated meteorological conditions at the stack locations, we found that a layered buoyancy approach for estimating plume rise in stable to neutral atmospheres, coupled with the assumption of free rise in convectively unstable atmospheres, resulted in much better model performance relative to observations (124 out of 176 cases falling within a factor of two of the observed plume height, with 69 of these cases above and 55 of these cases below the 1:1 line and within a factor of two of observed values). This is in contrast to our companion paper, wherein this layered approach (driven by meteorological

observations not co-located with the stacks) showed a relatively modest impact on predicted plume heights. "
We have also included a new table (Table 7 in the revised manuscript) which includes the distribution of values about the 1:2, 1:1 and 2:1 lines of the scatterplots, as well as the predicted average plume heights, in order for this comparison to be more quantitative.

3. P. 1, L. 28: ". . .between the surface and 1km elevation". I suggest to clarify the concept. From my understanding this refers to the bias in the simulated lapse rate dT/dz as compared to observations.

We have changed the sentence to read, "Persistent issues with over-fumigation of plumes in the model were linked to a more rapid decrease in simulated temperature with increasing height than was observed. This in turn may have led to overestimates of near-surface diffusivity, resulting in excessive fumigation."

4. P. 2, L. 6: ". . .(it is not created by chemistry)". Suggest to change "chemistry" in "photochemical reactions in the atmosphere".

We've followed your suggestion; the sentence now reads: "SO_2 is a primary emitted pollutant (it is not created by photochemical reactions in the atmosphere), with the majority of anthropogenic SO_2 emissions in the study region coming from large smokestacks (Zhang et al., 2018)."

5. P. 2, L. 10-11: "Anthropogenic SO2 emissions are the main source of most atmospheric sulphur deposition". Suggest to add a reference for this statement.

We have added a reference to Mylona, 1996 in the updated manuscript to support the above statement.

6. P. 3, L. 17-18: Please specify to what conditions/cases the given ranges (34 to 52% and 0 to 11%) are referred to.

The sentence refers to the companion paper, which did not segregate the extent of over/underprediction by stability class; we have added to the original statement, viz: "There we found that the Briggs (1984) plume rise parameterization significantly underpredicted plume heights in the vicinity of the multiple large SO_2 emissions sources in the Canadian Athabasca oil sands, with 34 to 52% of the parameterized heights falling below half of the observed height, compared to 0 to 11% of predicted plume heights being above twice the observed height, over conditions ranging from neutral, through stable to unstable."

7. P. 3, L. 22: typo "Sulpher" should be "Sulphur"
Corrected.

8. P. 4, L. 15: typo "as-phase" should be "gas-phase"
Corrected.

9. P. 5, L. 2: Would be more informative to add the height of the levels in the bottom 1

km of the model.

The model uses a hybrid coordinate system so the heights above ground change with location. However, the revised manuscript includes a new Figure (Figure 2 in the revised manuscript) which shows the model heights at a number of locations in the portion of the 2.5km domain studied here.

10. P. 5: Moreover, given the relevance for the results, I recommend to add a description of the parameterizations adopted in the model for the PBL and the surface layer turbulence.

The revised manuscript contains a new table (Table 1) which includes the main parameterizations in the meteorological model, including the Moist TKE scheme used for turbulence.

11. P. 8, eq. 6: Please check the second condition "0.5 < H < 1.5" since the range seems to refer to a unitless quantity, but here only H is given.

This was a typo on our part, and has been corrected in the revised manuscript.

12. P. 9, L. 6: "top of the atmosphere" is confusing: is it perhaps the top of the PBL?

An error on our part – the portion of the sentence should have read (and has been corrected to) "between h_t and h_b".

13. P. 9, L. 8: "value of hs was assumed", perhaps is "value AT hs was assumed". Moreover, the "s" of "hs" should be a subscript.

The phrase was modified to read "centered on h_s"

14. P. 10, L. 5: typo "and = hs", please check the left-hand side.

The sentence has been modified to "At the stack height, $F_{j=0} = F_b$, and $z_j = 0$ (that is, the vertical distances are relative to the top of the stack)."

15. P. 11, L. 21: "modstat" should be "modStat"

Done.

16. P. 12, L. 27: ". . .negative bias has decreased by 34%" it is not perfectly clear here and in the following if these bias changes are actual relative changes or absolute changes of the normalized mean bias. Please clarify.

The sentence has been changed to ". For example, the magnitude of the mean bias has decreased from -2.623 to -1.725 μg m^{-3}, a reduction of 34%, indicating that a sizeable fraction of particulate under-predictions in 2-bin simulations may be due to poor representation of particle *microphysics through the use of the 2-bin distribution, despite sub-binning being used in some microphysics processes. "*

17. P. 12, L. 31-32: "Figure 2 shows that . . . less than 5 um diameter . . .". Please double check this statement. The figure shows the PM2.5 concentrations binned as a function of CONCENTRATION not SIZE.

Thanks for catching this – the reviewer is quite right; this has been corrected.

18. P. 15, Table 2: Please check the values that should be given in Italics, since not all the rows seem to contain it.

Corrected.

19. P. 16: referred to the discussion of SO2 overestimation and SO4 underestimation: can the two things be linked? E.g. by slow SO2 to SO4 conversion in the model, perhaps by slow aqueous chemistry?

The aircraft observations were conducted under clear-sky conditions, so the potential for aqueous chemistry being the main issue is unlikely. Rather, noting that both predicted SO2 and SO4 aloft had negative biases, while predicted SO2 at the surface was biased high, it seems more likely that the main cause of the SO2 and SO4 negative biases aloft was a tendency for the model to overpredict fumigation to the surface, as noted in the original and revised manuscript.

20. P. 17, L. 4-7: the paragraph seems to imply the presence of at least a (b) point, but only (a) is given. Please check or rephrase.

The (a) has been removed (holdover from an earlier version of the manuscript, missed on checking prior to submission).

21. P. 19, L. 11: ". . .took place between 16:30 and 20:30 on Aug 24th. . .". Although I am assuming the intervals are given in local time and not in UTC, it would be useful to have a confirmation in the paper. Here and also at least in the caption of the first figure showing time series (Figure 5).

Corrected. The first pair of times are actually UTC (local times have been added in the revised manuscript), and Figure 5 (now Figure 6 in the revised manuscript) were also in UTC; this has been mentioned explicitly in the revised figure caption.

Anonymous Referee # 2:

This paper introduces a very interesting and potentially highly useful field campaign. It also provides some important insights into the performance of the operational Canadian model. However, the paper has some weaknesses listed below.

We thank the reviewer for the detailed comments on the paper – in addressing these, we feel that these comments have resulted in a significantly improved manuscript. The reviewer's comments and our responses (in italics) follow:

1. In the same special issue for which this paper is submitted, there is another paper by the same group of authors, Gordon et al. 2018, which is also devoted to the plume rise topic, and it is said to have found opposite results. Neither is the reason for that clearly resolved, nor does it become clear why the plume rise topic is split between two papers.

In our observation companion paper (Gordon et al, 2018) we saw that the Briggs algorithms, including the layered approach, tended to significantly underestimate plume rise. However, in that work we also noted that the observations themselves showed significant horizontal heterogeneity in the meteorological data used to drive the plume rise equations, with the corollary that the conditions at the actual location of the stacks may be sufficiently different from the surrounding meteorological towers to influence the predicted heights. No observations are available at the stack locations themselves – we therefore investigate this potential local influence further, using the high resolution on-line chemical transport model, GEM-MACH. As a demonstration of the model's ability to capture the local heterogeneity, we show in Figure 2(a) of the revised manuscript a snapshot of the PBL height at Saturday August 24, 2013 at 1 pm local time, and Figure 2(b) the corresponding model generated temperature profiles at the local meteorological towers AMS03, AMS05, the windRASS instrument and at three of the stacks examined in the companion paper. The model shows a significant variation in the temperature profiles between the tower and windRASS locations where the observations are available, and the stack locations. The temperature profiles suggest strong differences in both the strength of the inversion and its vertical location. This confirms the potential for spatial variability to have a significant influence on predicted plume heights relative to the meteorological observation locations in our companion paper. We therefore investigated the plume rise algorithms again within the current work, in order to determine the extent to which this local variability may influence predicted plume heights. In contrast to the companion paper, we found that the "layered" approach of calculating local stability residuals through successive model layers resulted in significantly improved plume heights relative to the more standard Briggs approaches which employ stability estimates at the top of the stacks.

We have included the new Figure 2 in the revised manuscript, as well as some discussion in the Introduction section of the manuscript:

"…Our companion paper made use of different sources of meteorological observations to drive the Briggs (1984) plume rise algorithms, as well as CEMS data and aircraft observations of SO_2 plumes from multiple sources over a 29-day period. There we found that the Briggs (1984) plume rise parameterization significantly underpredicted plume heights in the vicinity of the multiple large SO_2 emissions sources in the Canadian Athabasca oil sands, with 34 to 52% of the parameterized heights falling below half of the observed height, compared to 0 to 11% of predicted plume heights being above twice the observed height, over conditions ranging from neutral, through stable to unstable.

However, in our companion paper we also noted the presence of considerable spatial heterogeneity in the meteorological observations used for the algorithm tests. Temperature profiles and other data used to define the input parameters for the Briggs algorithms were taken from two tall meteorological towers, a windRASS, and a research aircraft, and showed a substantial variation in the resulting plume height predictions, despite relatively close physical proximity of these sources of meteorological data (e.g. 8 km distance between the two meteorological towers). The region under study is subject to complex meteorological conditions due to the nature of the terrain (a river valley with up to 800 m of vertical relief, and open pit mines and settling ponds which may each be tens of km^2 in spatial extent). This heterogeneity cast some uncertainty on the results of the companion paper, in that the best application of the plume rise algorithms would be driven by the meteorology at the location of the stacks, rather than the location of the available meteorological instruments, and the latter suggested substantial local changes in meteorological conditions. As we show in the sections which follow, the spatial heterogeneity of meteorological conditions has a controlling factor on the predicted plume rise, and, in contrast to our companion paper, an approach making use of local temperature gradients between individual model layers has greatly improved accuracy in comparison to those inferring atmospheric stability conditions from the conditions at the top of the emitting stacks."

The new Figure 2 is described via the following discussion in the revised text:

"We noted in our companion paper (Gordon et al., 2018) that meteorological observations varied substantially in the study region depending on location, citing this as a possible confounding factor on the results of tests of the plume rise algorithms. This spatial heterogeneity was well captured by the high resolution GEM-MACH simulations, as is demonstrated by the example depicted in Figure 2, which shows the typical local variation in planetary boundary layer height (Figure 2(a)), ranging from about 1200m to 400m, the lower values corresponding to the main cleared areas (open pit mines, settling ponds) of the industrial facilities. The corresponding temperature profiles in several locations marked in Figure 2(a) are given in Figure 2(b): These show a substantial difference in model predicted stability at the three meteorological observation locations of Gordon et al. (2018) (windRASS, AMS03, and AMS05), and substantial differences between these and the locations of the main stacks of some of the facilities (Syncrude 1, CNRL, and Suncor). The temperature profiles show that the height and strength of the inversion may vary by over 100m in the vertical, and that the profiles do not merge with the larger scale flow until an elevation of 750m asl (450m agl) is reached. Given this level of variation, we might expect potential errors in calculated plume heights when applying the meteorological observations to plume rise at the stack locations, in turn suggesting that a re-examination of plume rise using the model results is worthwhile."

2. The model performance is not only influenced by the aspects forming the focus of this paper, but also by the accuracy of the meteorological part of the model, and by the numerics of

transport, notably the vertical diffusion and the handling of the point sources in the Eulerian framework. Their role is discussed only at the very end and, in my opinion, not sufficently in depth. In order to evaluate specific model aspects, one first needs to understand the performance of the model in general, with its strengths and weaknesses.

The focus of this work is not to evaluate the overall model performance, but to evaluate how specific updates to the representation of the aerosol size distribution and the plumerise algorithm contribute to better model performance when compared to available observations. An evaluation of each aspect of a complex reaction-transport model is beyond the scope of a single paper. Nevertheless, as we already stated in the discussion section of the manuscript, we have carried out a sensitivity run which showed that variations in the magnitude of model diffusivity had a minimal impact the predicted plume behaviour and on the vertical distribution of SO_2 plumes at the point of release from the stacks, though we acknowledge that the model's tendency to overpredict the rate of decrease of air temperature with height may influence the shape of the diffusivity profile. We have also added additional references on the description and evaluation of the vertical diffusion scheme used in the meteorological portion of the model (Mailhot and Benoit 1982), as well as a more recent publication on the overall description and performance of the meteorological model as a whole (Girard et al., 2014), in Table 1 of the revised manuscript.

3. The statistical approach chosen for the evaluation of the model options relies on metrics which exclusively are based on "match in time and space" data pairs. It is well known in air-pollution modelling that for near-source conditions (which is what we find here), there is often too much "noise" in the data (be it due to the stochastic nature of the plume, be it due to unresolved meteorological variability) to give meaningful results. Correspondingly, some of the statistical parameters are not very good. Therefore, global comparisons (such as deviations from the cumulative frequency distribution, statistics of cross-wind integrated values, or average dependency on key parameters such as stability and wind speed) are often used to assess models in a more robust way.

We have added a simple table (Table 7, revised manuscript) showing the frequency distribution of the predicted versus observed plume rise from the three different variations on plume rise examined here. This new table is in agreement with the measurement statistics in that both show that the layered approach provides a better fit to observations, with a distribution more centered around the model:observation 1:1 line. While we agree with the reviewer regarding the difficulties associated with use of matching pairs for near-source conditions, we nevertheless respectfully hold that improvements in these statistics represent real improvements in model performance. For example, while a mean bias score is the average deviation between model and observed pairs, this average is over a large set of conditions, hence should be subject to less issues associated with the stochastic nature of the plume on any given hour. While near-source comparisons are often difficult due to the nature of the near-source region (as the reviewer suggests), improvements in these statistics nevertheless imply real improvements in model performance.

4. The paper is written well on the "small scale" (apart from numerous technical deficiencies as listed below), but the broad topics could be worked out more clearly. In the end, the findings are: twelve aerosol size bins are better than two (not surprising, but good to see it quantified), there is

an improvement by using the model's vertical profile information for plume rise calculation but given the model's deficiencies the overall conclusion seems to be not so clear, and no improvement was found for using hourly stack data, but it remains unresolved why. We may wonder whether the work is mature enough for publication if we consider this state of the quintessential findings.

We note that the results presented in the work show that the use of stack-location-specific meteorological information combined with the residual buoyancy calculations provides a considerably more accurate estimate of plume rise than the top-of-stack stability parameterizations often used in air-quality models. We have provided an additional table which shows that the distribution of plumes is better represented with the residual buoyancy calculation than with the top-of-stack stability plume rise calculation. The average plume rise calculated using the CEM-based data is closer to the observations than the annual totals, from both the original analysis and the additional table. While both the CEMS (using the hourly stack data) and the non-CEMS model scenarios are very close, the key point is that the revised, residual buoyancy plume rise algorithm has much better performance than the original algorithm. We have noted in the revised manuscript that "the relatively small differences between Figure 8(b) and 8(c), and between the last two columns of Table 7, imply that the residual buoyancy approach of equations 9 was relatively insensitive to the range of the initial buoyancy flux resulting from the two sets of emissions data used here, compared to the temperature gradients in equation (5)." We have also mentioned the insensitivity of the residual buoyancy calculation to the range of initial buoyancy flux in the revised conclusions; "However, the latter approach was also shown to be relatively insensitive to the range of initial buoyancy fluxes resulting from the two different emissions estimates, with the use of hourly observed (and presumed more accurate) stack parameters resulted in a slight degradation of performance relative to the use of annual reported values for these parameters."

Both sources of emissions data are limited by the model resolution and the independent verification of the accuracies of either is not available. We also point out in the revised manuscript that both sources of data have inherent errors. For example, as mentioned in the emissions paper referenced by this work (Zhang et al., 2018), and clarified/noted in the revised manuscript, the data referred to as CEMS here also contains engineering estimates of "upset conditions" wherein facility emissions are redirected to a flare stack for which direct emissions observations are not possible. That is, what we have referred to as "CEMS" data incorporates considerable associated uncertainties – this should have been included in the original manuscript and not left as a reference to the emissions paper alone. This has been mentioned in the description of the CEMS data as follows, "…and second with emissions information derived from a combination of CEMS hourly stack parameters as well as engineering estimates of emissions during "upset conditions" in which the effluent is redirected to flare stacks (the latter estimates are considerably more uncertain than the CEMS information, but are nevertheless included here since they result in substantial changes in pollutant emissions and plume characteristics, see Zhang et al, 2018)."

RC2: Specific Comments

1. Page 2, l. 18: Why are you thinking that reasons for weak performance include only those meteorological variables that are used for the plume rise calculation, but not, for example, wind direction?

We acknowledge that the model's performance is of course the result of many factors. The intent of our work is rather to evaluate the relative impact of the plume rise calculations on the results. The given sentence has been modified to "(iii) errors in meteorological forecast variables (including wind speed and direction, etc., as well as those used in calculating plume rise)".

2. The model overview section lacks information on the numerical scheme used for vertical diffusion even though this is crucial in the context of study (cf. discussion on p. 24). The main reference for the MACH model seems to be Makar et al. (2010) – an extended abstract that would not be available for most people who haven't attended the conference as it is not freely accessible. Is there no more detailed and open description of this model? Note that also the Coté et al. citation is one of those for which the reference is missing. In addition, the handling of the point sources is not described (usually, Eulerian models use some sub-model to track plumes until they match the size of the grid cells).

The revised manuscript includes a new Table (Table 1) which gives the main references for the meteorological (GEM) components for the model, including the reference for the Moist TKE approach used for calculating vertical diffusivity. Regarding the online air quality GEM-MACH model, the first overall description reference is Moran et al. (2010) (and not Makar et al. (2010) as stated by the reviewer), and hence we feel obliged to include it in published descriptions of GEM-MACH. However, this is not the only description of the model or its components, and others from the journal literature appeared in the original manuscript; we cited Gong et al (2003) for the aerosol microphysics, Gong et al (2006) for the aqueous-phase chemistry, Makar et al, (2003) for the inorganic heterogeneous chemistry, Lurmann et al. (1986) and Stroud et al. (2008) for the gas-phase mechanism , Zhang et al. (2001, 2002, 2003) and Makar et al (2018) for the gas and particle deposition. We also cited the Air Quality Model Evaluation International Initiative papers Im et al (2015a,b) and Makar et al (2015(a,b)) papers, which contain detailed descriptions of the model, its chemical and physical parameterizations, and its performance relative to other models of its type. The revised reference list has been double-checked to include all references (including the papers by Côté et al).

While some air-quality models include a form of "plume in grid" parameterization which track emitted puffs in a Lagrangian sense or employ a Gaussian dispersion model at the sub-grid scale, these approaches have not become predominant for three main reasons: (1) they ultimately rely on the driving large-scale meteorology (which may be inaccurate, as already pointed out in our work and by the reviewer, reducing their potential advantages); (2) they may add considerably to the processing time (particularly if a large number of point sources, chemical reactions and multiple species are considered), and (3) most models employ a self-nesting capability which allow the models to locally go to higher resolution, negating some of the advantages to a plume-in-grid approach. Consequently, most air-quality models have continued to rely on the handling of point sources using a combination of plume rise algorithms, and nesting to higher resolutions, as has been done in our work.

3. The model set-up description in section 2.2 is not easy to follow. It might be helpful to move some of the information into a table and to shorten the text.

A table summarizing the model description has been added as suggested.

4. Page 8, line 1: The plume's buoyancy flux is **not** dependent on the stack height (at least not directly).

This typographical error has been corrected.

5. From the sentence beginning on p. 11, l. 7, on, the text does not really belong to the section 2.2.3. It should become a section of its own, as it introduces the simulations forming the base of the rest of the paper (maybe merge with some parts of the 2.2 chapeau).

Section 2.2.3 has been renamed "Sources of Emissions Data", and the remainder of the previous section 2.2.3 past the point noted by the reviewer has been split off and renamed "2.2.4 Simulation Scenarios"

6. Page 11, Section 3.1: What is the justification for removing measurements with values exceeding some threshold? Without proper justification this would not be acceptable.

The key phrase in the original manuscript was that "extreme single-hour measurements" have been removed. That is, if the time series jumps from a background value to something greater than 150ppbv (SO_2, NO_2, and O_3) or 150 μg m^{-3} (PM2.5), then immediately back again in the next hour, that jump is assumed to be due to instrumentation error and/or calibration times in the measurement record. In contrast, a rise above these levels for more than a single hour is retained. This has been clarified in the revised manuscript, viz: "The observation data have been filtered to remove extreme single-hour measurements that are greater than 150ppbv for SO_2, NO_2, and O_3, and 150 µg m^{-3} for $PM_{2.5}$ (single-hour spikes of this nature in hourly records are assumed to correspond to instrumentation errors or calibration times for the instruments)."

7. Page 12, Section 3.2: The phrase 'spatial linearly interpolated model values at the models chemistry time resolution of 2 minutes' is awkward. If you have 10 s data as said before, why do you need to interpolate for obtaining 2 min data? Also, it would be good to know which distance corresponds to both 10 s and 2min flight data, and how this compares to the model's grid size.

The portion of the sentence has been corrected to "linearly interpolated values in time and space from the model's 2 minute time step and 2.5km resolution". We have also added the sentence: "The nominal cruise speed of the National Research Council Convair 580 used in the experiment is 550 km/hour; a 10 second time interval thus represents an observation integration distance of 1.528 km, and a two minute time interval an observation integration distance of 18.3 km."

8. Section 4 (Results and Discussion) needs to be structured into subsections.

This has been done, with subsections 4.1 Spatial Heterogeneity of Meteorological Conditions, 4.2 Two-bin versus Twelve-bin Evaluation, and 4.3 Plume Rise Algorithm Evaluation.

9. Table 1: Apart from widely used or self-explanatory metrics such as FAC2, RMSE or r , the metrics parameters need to be defined.

We have included a new Table; Table 2, which is now referenced at the start of section 3, and includes the mathematical definitions of all of the metrics.

10. Page 12, l. 31: "Figure 2 shows that the model simulations are biased high for particles less than 5 μm diameter, and biased low for the larger particle sizes." As this figure only shows results for PM2.5, a statement on larger particles can't be based on it

This was a typo; the text should have referred to 'concentration' and not 'particle sizes'; this has been corrected in the updated manuscript.

.

11. Page 13, l. 14: Information on the bin sizes belongs to the model description section, not the result section.

We have added a description of the cut sizes for both 2 and 12 bin simulations to the model description section. However, the mention of the 2 bin cut sizes is necessary here to explain why a comparison between the 2-bin model results with the aircraft observations is not appropriate (the 2-bin model lacks the size cut resolution to be able to simulate the PM1 observed by the aerosol mass spectrometer aboard the aircraft). The sentence has been changed to "The aircraft's AMS instrument measures speciated atmospheric particle concentrations for particles less than 1μm size, and therefore cannot be compared with the 2-bin model results because the smaller size bin (with upper diameter size cut 2.56 μm) will be biased high relative to the 1 μm size cut of the AMS."

12. Page 13: The second paragraph on this page contains a number of statements about results without pointing to the figures or tables which show them.

The revised manuscript references Table 4 for this paragraph.

13. Concerning the model performance for PM, it should be discussed that even though the twelve-bin version leads to significant improvements, major discrepancies to observations remain.

The last sentence of the new section 4.2 has been modified to read: "The use of the 12-bin size distribution (purple histogram bars, Figure 3) improves the fit to the observations (blue histogram bars), in comparison to the 2-bin distribution results (red histogram bars), though significant over-predictions of the frequency of low concentration events and under-prediction of high concentration events, remain.".

14. A number of tables are presented where several metrics are used to compare various model versions, with the best one being emphasised by bold print. Sometimes, differences are tiny and probably insignificant. Only those values that are significantly better should be highlighted to

avoid a wrong impression of the results (for example, in Table 3 the model version seems to have no impact for O3 but we get the impression that the simpler model is better.)

The problem with this suggestion is that different readers may have different ideas regarding what is considered a 'significant' change, what is considered to be a 'tiny' difference, etc. We feel that the readers will look at both the numbers themselves as well as the highlighting, as the reviewer did, to note the relative level of differences. We mentioned some of these relative differences in the original text as well, as a caution to the reader not to base their judgement on the highlighting alone, e.g., with reference to the ozone evaluation in Table 3: "Ozone, in contrast, is created or destroyed through secondary chemistry over relatively longer time-spans than the transport time from the sources in this comparison (spatial scales on the order of 10's of km). Accordingly, the impact of the plume rise of NO_x *on ozone formation is relatively minor, usually in the third decimal place (though first decimal place improvements occur for the mean bias with the use of the new plume rise algorithm)."*

15. Why is the use of hourly emission data beneficial for NO2 but detrimental for SO2?

One significant difference between SO_2 *and* NO_2 *for the study region is that the latter originates almost completely in major point sources, while only 40% of the latter originates in major point sources, the rest in area sources (heavy hauler fleets used by the open pit mine operators). The* NO_2 *values will thus be due to a combination of sources, with the possibility of compensating errors at the emissions level influencing the net model* NO_2 *concentration .*

16. The discussion paper does not comply with the ACP Data Policy; it does not have a "Data availability" section and says nothing about data availability

This has been added to the revised manuscript (in our other papers with ACP, this has come as a request from the Editor following the completion of the review process, sorry for not having added it to the submission): all of the data used here are publicly available on the oil sands data archive or the website of the Wood Buffalo Environmental Association. We have also added the standard Author Contributions, Competing Interests, Special Issue Statement, and Acknowledgements, to the revised manuscript.