Response to Referee’s comments, Vehicle induced turbulence and atmospheric pollution, Makar *et al*, 2021.

Please note – this updated version of the response to the Referee’s comments includes an additional section of responses to final comments by Referee number 2 – these additional comments and responses appear at the end of this document (Under “Final Referee 2 Comments and Responses”).

Also - a PDF file of this response has been attached as a supplement to this response (the PDF file may be easier to read, for formatting).

Anonymous Referee # 1

This study develops/proposes a parameterization for an additional source of atmospheric dispersion due to moving vehicles on roads (vehicle-induced turbulence, VIT) for use in 3-D chemical transport models. The topic is relevant as mobile sources are often dominant contributors to air pollution in many urban areas while grid scales typically employed by 3-D CTMs are not sufficiently fine to adequately represent those roadway sources. The manuscript is well organized, and the methodology and study outcomes are effectively presented.

We thank the referee for these positive comments.

What's missing in the current manuscript is a proper evaluation of the proposed VIT scheme. The authors evaluated performance of a 3-D CTM with and without the VIT scheme against observations and showed that the model performance was generally better with the VIT scheme. As the authors also noted, however, model performance of a 3-D CTM is affected by a number of factors, and good model performance doesn't necessarily mean that the model is right for the right reasons. For example, improved model performance could be resulted from model biases due to over-estimated vehicle emissions being reduced by increased mixing by the VIT scheme. While the 3-D CTM simulations serve well as a sensitivity analysis (it's clearly shown that the proposed VIT scheme implemented in a 3-D CTM has significant impacts on the model results), a better evaluation of the VIT scheme may be to directly compare the scheme in a simplified version of the CTM (e.g., with a single horizontal grid cell with multiple vertical layers) with a finer-resolution LES or CFD model, using a small test case with a more controlled setup (e.g., a hypothetical roadway with a predefined vehicle configuration). At least, more discussion on this should be added.

> *Simplified models of the sort described by the referee were referenced in original manuscript (Eskridge and Catalano, 1987; Eskridge et al., 1991), and have been used in local scale engineering applications in the past. Our parameterization is built in part on finer resolution LES model results to describe the decrease in VIT TKE with height and the typical mixing length associated with vehicles travelling alone and in tandem. We have modified one of the sentences in section 2.2 to clarify this point,*
>
> “We examined four datasets (the observations of Rao *et al.*, 2002, and the LES modelling of Kim *et al.*, 2016a; Woodward *et al.*, 2019; Zhang *et al.*, 2017) to evaluate the extent to which a Gaussian distribution may be used to represent the decrease in VIT with height above moving vehicles, as well as examining the expected range of mixing lengths which may result from VIT.”

We have also modified one of the sentences in the Abstract – we want to be clear there that the parameterization is intended to allow VIT to be incorporated in regional chemistry models, so that the effect can be represented for the large spatial domains those models cover, but ->not<- that regional models and this VIT parameterization should be a substitute for LES models on the local scale. We have therefore modified the abstract sentence from

"This parameterization allows vehicle-induced turbulence to be represented at the scales inherent in 3D chemical transport models, allowing its impact over large regions to be represented, without the need for the computational resources and higher resolution of large eddy simulation models."
To
"This parameterization allows vehicle-induced turbulence to be represented at the scales inherent in 3D chemical transport models, allowing this process to be represented over larger regions than is currently feasible with large eddy simulation models."

We show that the parameterization accounts for most of the variation in height created by CFD models such as Large Eddy Simulation models (with correlation coefficients of 0.54 to 0.98, in Table 1). In that sense we've gone beyond comparing LES and regional air-quality models – we've devised a parameterization for regional air-quality models which makes use of LES modelling results, providing reasonable correlations for the key mixing length variable generated from LES models. We note that the difficulty we face for regional-scale applications of VIT is that the turbulence occurs at the sub-gridscale level, and what's needed for regional air-quality simulations is a means of incorporating VIT effects without having to go to the LES scale – since that resolution would be far too computationally expensive to carry out on, for example, a North American sized domain, with currently available computer technology. We have modified a paragraph in the Introduction to discuss this latter issue further, where we also note the domain size and scale issue, and we mention the need for a computationally efficient parameterization in the Conclusions.

Modified paragraph in the Introduction:

*"*Large eddy simulation (LES) / computational fluid dynamics (CFD) models have shown the importance of VIT towards modifying local values of turbulent kinetic energy, as noted in the references above. However, these models require relatively small grid cell sizes compared to regional chemistry models (cm to tens of metres) and time steps to allow forward time stepping predictions of future meteorology and chemistry. These constraints in turn severely limit the size of the domain in which they can be applied, and the processing time for simulations for these reduced domains can be very high. For example, the FLUENT model was used by Kim et al (2016a) with an adaptive mesh with a minimum cell size of 1 cm, with a 100x20x20m domain, while Woodward et al (2019)'s implementation of FLUENT had an equivalent cell size of 50 cm, operating in a domain of 600,000 nodes (a volume of 75,000 cubic metres), and an adaptive timestep limited by a Courant number of 5. The latter criteria implies a computation timestep of less than 0.09 s for a 100 km hr^{-1} vehicle (or wind) speed, while a 1 cm grid cell size implies a computation timestep of less than 1.8x10^{-3} s timestep. Similarly, the LES model employed by Zhang et al (2017) utilized a 1m x 2m x 1m cell size and a computation timestep of 0.03 s. Other LES models have larger horizontal resolution, but are limited in horizontal domain extent relative to regional chemical transport models (example LES models incorporating gas-phase chemistry include: Vinuesa and Vil.-Guerau de Arellano (2005), with a 50m horizontal resolution, 3.2x3.2 km domain), Ouwersloot et al. (2011), with a 50m horizontal resolution and a 12.8km x 12.8 km domain, Li et al. (2016), with a 150m horizontal resolution and a 14.4km x 14.4km horizontal domain, and Kim *et al.* (2016b), with a 66.6m horizontal resolution and a 6.4x6.4 km domain. In contrast, a 3D regional chemical transport model typically operates over a domain with may be continental in extent (the simulations described here have a 10km and 2.5km horizontal resolutions with 7680x6380 km and 1300x1050km domains,

respectively). The limiting horizontal resolution for regional chemical transport models is on the order of kilometres, with a limiting vertical resolution on the order of 10's of metres, and timesteps on the order of 1 minute. These limits for regional chemical transport models are a function of the need to provide chemical forecasts over a relatively large region, within a reasonable amount of current supercomputer processing time (the chemical calculations typically taking up the bulk of the processing time). LES models are capable of capturing VIT effects (Kim et al. (2016a), Zhang et al., (2017), Woodward et al. (2019)), and their results have been used here in developing our parameterization, but are constrained by current computer capacity from being applied for the larger scale domains required in regional to continental-scale air pollution simulations. A "scale gap" exists between LES and regional chemical transport models – for regional chemical transport models, parameterizations of the physical processes such as VIT, resolvable at the high resolution of LES models, are therefore required. In return, these parameterizations allow the relative impact of the parameterized processes on the larger domain sizes of regional chemical transport models to be determined"

The addition to the Conclusions:

Our work implies that the turbulence associated with vehicle motion is capable of having a significant effect on the concentrations of key pollutants in the lower atmosphere, using a parameterization which allows these effects to be incorporated at the relatively coarse horizontal resolutions of regional chemical transport models.

Minor technical issues are listed below.

Thanks very much for catching these (and that we'd missed the minor technical issues from part of line 358 on down – a result of not copying the comments correctly to the master version of our Response to the Reviewers). Many of these were the result of this work having gone through multiple versions prior to submission to ACPD – we really appreciate the Referee catching these issues!

Line 64: "a vehicle11" ?

- *Thanks for catching this. This was from an earlier version of the manuscript which utilized numbered references; the correct reference in ACP format has been added to the text. Reference is Rao et al (2002).*

Line 282: "added via the "F" terms in (6)" -> "added via the "F" terms in (8)"

- *corrected.*

Line 302: "All six panels also show a trend of ∂K⁄∂z becoming more negative" - revise this sentence; many of the panels actually show positive ∂K⁄∂z.

- *Our wording here was imprecise – yes, the values of ∂K⁄∂z are positive in the much of the figure. However, the impact of VIT is to reduce the "positivity" of the slopes; the value of* $\frac{\partial K}{\partial z}_{VIT} - \frac{\partial K}{\partial z}_{No-VIT}$ *is usually less than zero. We've made this more explicit in the revised text "* All six panels also show a trend of ∂K⁄∂z becoming more negative (that is, near-surface positive slopes become less positive, negative slopes become more negative),…"

Line 358: Line 358: "with different values for the input coefficients of thermal turbulent transfer coefficient (K)" -> "with different values for the input coefficients of thermal turbulent transfer coefficient (K) and for the lower boundary conditions (E)"

Corrected as suggested by the reviewer

Line 479: "metrics used to here (see Methods)" -> "metrics used here"]

Corrected.

Line 497: "Figure 7(a,b)" -> "Figure 7(a,c)"; "Figure 7(c)" -> "Figure 7(e)"

Corrected.

Line 587: "S11" -> "S10"

Corrected. Note that many of the Figure numbers in the revised manuscript and the Supplemental Information have been renumbered due to the addition of additional figures with the 90% confidence interval analysis; we checked to make sure the final numbers match with the revised text.

Line 830: Figure 1 caption says "the length scale of turbulence immediately behind the leading vehicle, a large transport truck is only 3m, while the length scale immediately behind the trailing vehicle in the ensemble (an identical transport truck) is 12.73m", but Table 1 shows that the mixing length for an isolated lead diesel cargo truck is 5.13m and that for the 2nd diesel cargo truck in an ensemble is 14.64m. Explain the discrepancies.

The Figure caption has been corrected to match the Figure. The caption numbers were a hold-over from an earlier version of the Figure – we had subsequently more carefully transposed the contour lines from the reference in the revised version used in the submitted paper, including marking the cross-section locations where the contour values were extracted for fitting to the exponential decay function in the submitted paper version of the Figure – but forgot to update the values in the captions(!). The numbers in the Figure are the correct/final ones, and the caption has been corrected in the revised manuscript.

Line 836: In Figure 2 caption, "at low (a,c) and high (c,d) resolution" -> "at low (b,d) and high (a,c) resolution"

Changed to "at high (a,c) and low (b,d) resolution."

Line 845: Figure 4(b) caption says "equation (8)", but the legend says "eqn (6)".

The Figure has been corrected – this was the result of an earlier version of the paper starting off with what later became equation (8) initially as equation (6).

Anonymous Referee # 2

This study investigates the impact of kinetic energy from moving vehicles on the vertical distribution of combustion emissions and uses a VIT parameterization to account for the vertical transport of fresh mobile emissions in a 3D chemical transport model. This is an important topic as a better representation of vehicle emissions and mixing in atmospheric chemical transport models is crucial for an improved understanding of air pollutants. The manuscript is generally well written and aims to provide a way to improve mixing of mobile emissions in 3D regional modeling. Reasonable assumptions are made to parameterize vehicles with different sizes and running with distinct road conditions. However, the evaluation of the VIT parameterization is rather weak and there are a few major flaws in the manuscript.

1. In the introduction section, it states that the LES models are typically employed at centimeter or meter level resolutions, while the mixing lengths associated with VIT are on the order of tens of meters. This incomplete review of LES studies is misleading as it indicates the vertical influence of VIT is between the scales of a LES and a 3D regional model. However, the following studies all applied LES coupled with chemistry at a horizontal resolution of tens of meters, and there are more similar LES studies not listed here.

Vinuesa and Vil.-Guerau de Arellano (2005) Atmos. Environ., 39(3), 445–461

Ouwersloot et al. (2011) Atmos. Chem. Phys., 11(20), 10681–10704

Li et al. (2016) J. Geophys. Res., 121(13), 8083-8105

Kim et al. (2016) Geophys. Res. Lett., 43(14), 7701–7708

As the VIT problem is actually on a LES scale and a LES model with chemistry has already taken into account turbulent mixing in the boundary layer, it might be more convincing to illustrate the impact of VIT on the vertical mixing of vehicle emissions if a LES model is employed.

It was not our intent to imply that LES models were not suitable for VIT studies, or that the scale of VIT was between that of LES models and regional air quality models such as ours. Quite the contrary. Rather,
(1) The horizontal scale of roadways is much smaller than that associated with regional chemical transport models, hence
(2) The vertical motions associated with VIT need, at the regional model horizontal resolution, to be represented by a parameterization such as ours, also
(3) LES models are <u>very</u> suitable to capture this scale of motion. Indeed, our parameterization is in part built upon the results of LES models, as was explained in the original text, and has been highlighted further in the revised text, however,
(4) The problem we face when carrying out regional air-quality model simulations, is that they of necessity operate on much larger domains than LES models, and thus LES resolution is not possible given current computational time and memory resources for these domains. We can't afford to run North America at an LES resolution of 10's of m to a few cm. For example, in the four references quoted by the Reviewer, the horizontal resolutions and domain sizes were 50m & 3.2x3.2km, 50m & 12.8x12.8km, 150m & 14.4x14.4km, and 66.6m & 6.4x6.4km. Compare these to the sizes of our North American and PanAm domains: (10km &7680x6380 km and 2.5km & 1300x1050km). Our point in the paragraph is <u>not</u> that LES models are not suitable for simulating VIT (they definitely are!), but that the larger domain sizes of regional chemical transport models can not operate at LES resolution, and that consequently, parameterizations are needed to represent the impacts of VIT for regional scale simulations. The text has been modified to include the following, making use of the references provided by the Referee:

"Large eddy simulation (LES) / computational fluid dynamics (CFD) models have shown the importance of VIT towards modifying local values of turbulent kinetic energy, as noted in the references above. However, these models require relatively small grid cell sizes compared to regional chemistry models (cm to tens of metres) and time steps to allow forward time stepping predictions of future meteorology and chemistry. These constraints in turn severely limit the size of the domain in which they can be applied, and the processing time for simulations for these reduced domains can be very high. For example, the FLUENT model was used by Kim et al (2016a) with an adaptive mesh with a minimum cell size of 1 cm, with a 100x20x20m domain, while Woodward et al (2019)'s implementation of FLUENT had an equivalent cell size of 50 cm, operating in a domain of 600,000 nodes (a volume of 75,000 cubic metres), and an adaptive timestep limited by a Courant number of 5. The latter criteria implies a computation timestep of less than 0.09 s for a 100 km hr^{-1} vehicle (or wind) speed, while a 1 cm grid cell size implies a computation timestep of less than $1.8x10^{-3}$ s timestep. Similarly, the LES model employed by Zhang et al (2017) utilized a 1m x 2m x 1m cell size and a computation timestep of 0.03 s. Other LES models have larger horizontal resolution, but are limited in horizontal domain extent relative to regional chemical transport models (example LES models incorporating gas-phase chemistry include: Vinuesa and Vil.-Guerau de Arellano (2005), with a 50m horizontal resolution, 3.2x3.2 km domain), Ouwersloot et al. (2011), with a 50m horizontal resolution and a 12.8km x 12.8 km domain, Li et al. (2016), with a 150m horizontal resolution and a 14.4km x 14.4km horizontal domain, and Kim *et al.* (2016b), with a 66.6m horizontal resolution and a 6.4x6.4 km domain. In contrast, a 3D regional chemical transport model typically operates over a domain with may be continental in extent (the simulations described here have a 10km and 2.5km horizontal resolutions with 7680x6380 km and 1300x1050km domains, respectively). The limiting horizontal resolution for regional chemical transport models is on the order of kilometres, with a limiting vertical resolution on the order of 10's of metres, and timesteps on the order of 1 minute. These limits for regional chemical transport models are a function of the need to provide chemical forecasts over a relatively large region, within a reasonable amount of current supercomputer processing time (the chemical calculations typically taking up the bulk of the processing time). LES models are capable of capturing VIT effects (Kim et al. (2016a), Zhang et al., (2017), Woodward et al. (2019)), and their results have been used here in developing our parameterization, but are constrained by current computer capacity from being applied for the larger scale domains required in regional to continental-scale air pollution simulations. A "scale gap" exists between LES and regional chemical transport models – for regional chemical transport models, parameterizations of the physical processes such as VIT, resolvable at the high resolution of LES models, are therefore required. In return, these parameterizations allow the relative impact of the parameterized processes on the larger domain sizes of regional chemical transport models to be determined."

2. On line 635, it states that "An examination of all of the other possible sources of error in air-quality models is beyond the scope of this work." This is understandable. But 3D regional models typically have difficulties representing turbulence and vertical mixing, which cause a large portion of their model-observation discrepancies. Without considering errors related to boundary layer turbulence, it is hard to evaluate the VIT parameterization developed in this study.

   We are well aware of the issues with boundary layer turbulence and the difficulties 3D regional models have in representing turbulence and vertical mixing, some of us having published work in the past on this topic (e.g. Makar et al, 2014, which showed that a substantial portion of a regional air-quality model's error could be accounted for by the choice of magnitude of a lower limit in vertical thermal diffusivity coefficient). Here, we are not attempting to improve on the underlying meteorological model's turbulence parameterization, but rather are asking the question, "How much of the problems encountered by regional air-quality models with respect to strength of turbulence might be accounted for by VIT?" The Referee is stating that difficulties in the representation of turbulence and vertical mixing cause a large portion of model-observation *discrepancies. We agree – however, our contribution to this work is to investigate the extent to which VIT, as a component of turbulence, may add to this problem. We also explore the possibility, and present evidence, that at least some of these problems may not lie within the meteorological turbulence parameterization of the models, but in the absence of VIT as means by which surface-emitted pollutants may be mixed upwards, on the sub-gridscale. However, this is not intended to suggest that other improvements to turbulence parameterizations should not be pursued!*

The manuscript states that "We also emphasize that the work does not identify a deficiency in existing meteorological boundary layer turbulence models." Does it mean the 3D model used in this study represents turbulence very well? Please clarify whether it refers to the 3D model used in this study and how "deficiency" is evaluated.

We added this sentence out of concern that our work might be otherwise be taken to imply that the excellent work that has taken place in the past on improving meteorological turbulence parameterizations was somehow deficient, by not including VIT. Nor are we saying that the turbulence parameterization in our own model is perfect. We are examining whether VIT might account for sufficient vertical mixing of fresh at-source pollutants to influence the distribution and transport of those pollutants. We are studying this as a separate issue from meteorological model turbulence parameterizations performance. The Referee states above that a large portion of model-observation discrepancies may be attributed to meteorological turbulence parameterizations. We don't disagree with this view – however, that does not remove the possibility that VIT is also a contributing factor to discrepancies between observed and modelled concentrations, specifically within regional air-quality models. We do agree that we should place our work more clearly in that context, and we definitely don't want readers to have the impression formed by the Reviewer due to the original sentence - we have modified the sentence to read, "This work is not intended to be taken as a review or critique of existing boundary layer parameterizations within meteorological or regional air-quality models. There has been excellent work in recent years on improving these parameterizations, and there are several reviews discussing this topic in the literature (e.g. Edwards et al., 2020). Rather, we focus here on an ancillary problem specific to regional air-quality models: whether the turbulent kinetic energy associated with vehicle motion could account for sufficient sub-grid-scale vertical mixing to influence the concentrations of surface-emitted pollutants at and above roadways, and further downwind."

3. Although the manuscript states that "The use of the VIT parameterization has been demonstrated to result in decreases in air-quality model error," this is not convincing as the changes in the metrics used to evaluate model performance are inconsistent and the differences in these metrics between the VIT simulation and No VIT simulation are quite small. It is necessary to show whether adding VIT actually leads to statistically significant differences. Statistical significance can be calculated based on the differences (VIT simulation – No VIT simulation) in daily averaged NO2, PM2.5 and O3 at each site. Alternatively, estimates of vehicle km travelled can be used as a criterion to select sites, then significance could be calculated based on the selected sites with similar traffic conditions and background meteorological conditions.

   While we would not characterize a factor of 8.4 reduction in the magnitude of North American NO2 bias as being quite small, and note that the atmospheric system has many other components in addition to turbulence, which interact in a non-linear fashion (improvements associated with one chemical species will not necessarily be mirrored in others), we agree with the referee's point on statistical significance.

We thank the referee for this very good suggestion regarding statistical significance, and we have made extensive modifications to the manuscript to follow up on it. We agree with the referee that an examination of the statistical significance of the differences of the results is worthwhile, and through including that examination, we feel our revised paper better demonstrates that the incorporation of VIT does result in statistically different mean values at the 90% confidence level. In order to examine the statistical significance of the two different simulations, we generated 90% confidence limit ratios on the average concentrations generated by each approach. We calculated 90% confidence levels for all of the original figures comparing mean values in the original manuscript, and added additional figures to the revised manuscript showing the results of these calculations (revised manuscript Figures 8, 12, 13, S8, and S9). As we note in the revised manuscript:
In Section 3.3:

*"*The region over which the two simulations' mean values differ at the 90% confidence level is shown in Figure 8. The difference between the mean values of the two simulations (M_{VIT}, M_{NoVIT}) becomes significant at a confidence level c if the regions defined by $M_{VIT} \pm z^* \frac{\sigma_{VIT}}{\sqrt{N}}$ and $M_{NoVIT} \pm z^* \frac{\sigma_{NoVIT}}{\sqrt{N}}$ do not overlap (where N is the number of gridpoint values averaged, the σ values are the standard deviations of the means, and z^* is the value of the $\sqrt{c}$ percentile point for the fractional confidence interval c of the normal distribution, where $z^*=1.645$ at $c=0.90$. Grid cell values where the mean values differ at or above the 90% confidence level are thus defined as $\frac{|M_{VIT} - M_{NoVIT}|}{\frac{z^*}{\sqrt{N}}(\sigma_{VIT} + \sigma_{NoVIT})} > 1$ thus differ at greater than the 90% confidence level. The mean values at each gridpoint and their standard deviations may thus be used to determine the confidence level – these values for each of the mean differences of Figure 7 are shown in Figure 8, with red colours indicating differences significant at greater than 90% confidence.

From Figure 8, it can be seen that the continental scale model means for the VIT versus No VIT simulations for surface NO_2, surface PM2.5 and surface O_3 at night differ at 90% confidence, over much of the domain for NO2 and PM2.5, and in urban core areas for O_3. The spatial extent of 90% confidence is much greater under the stable conditions of night (Figure 8 (a,c,e)) than the less stable conditions of daytime (Figure 8(b,d,f)), as would be expected from the relative magnitude of K_T versus K_{VIT} during the day and night. While the nighttime influence of VIT on NO_2 extends over much of the continent, for O_3, the impact is primarily within the cities, where the increased mixing of NOx results in higher nighttime O_3 concentrations due to decreased NOx titration."

In Section 3.7:

"The spatial extent of the region where the wintertime mean values for the PanAm domain differ at greater than 90% confidence are shown in Figures 12 and 13 for the model's surface concentrations and the corresponding vertical cross-section, respectively. The corresponding summertime differences for this domain are shown in Figures S8 and S9. For the wintertime PanAm domain simulations, surface NO_2 and PM2.5 90% confidence regions are similar to those of the continental 10km domain, and can be seen to extend into the late morning hours (14 UT; 10 AM local time; Figure 12(b,e)). The mean values of NO_2 and to a lesser extent PM2.5 also differ at greater than 90% confidence later in the day in the urban core regions (Figure 12(c,f)). In contrast to the continental scale results (Figure 8) the influence of VIT on surface O_3 approaches but remains below the 90% confidence level at 14 UT in the urban regions (Figure 12(h)), and remains below 90% confidence at the other times shown. The vertical influence of wintertime VIT results in mean values differing at greater than 90% confidence up to ~700m altitude for NO_2 and PM2.5, and the above-ground O_3 mean values differ at greater than 90% confidence for regions between 25 and 200m altitude over specific large urban areas (e.g. New York City at 14 UT, Figure 13(h)). Regions of greater than 90% confidence in the vertical at 22 UT for NO_2 and PM2.5 are confined to the urban core regions near the surface (Figure 13(c,f)). For the summertime high resolution PanAm domain, differences at greater than 90% confidence occur for surface NO_2 and PM2.5 at night and early morning (Figures S8,S9 (a,d)) and persist until later morning over parts of the Great Lakes (Figure S8(b,e)), and isolated locations over cities (Figure S9(b,e)). Differences in the mean ozone aloft at night occur at greater than 90% confidence over the largest cities (e.g. New York, Figure S9(a)).

Taken together, Figures 8, 12, 13, S8 and S9 show that the incorporation of VIT into the model results in mean values which are statistically different at the 90% confidence level, for NO_2 and PM2.5 over large regions, and to a lesser degree for O_3 over urban areas, with a greater influence at night, in the early morning, and under the more stable conditions of winter compared to summer."

We have also added to our 5th point in in the Discussion and Conclusions section: "These differences occur at greater than 90% confidence over much of the model domains for NO_2 and PM2.5, and in urban core regions for O_3 at 10km resolution, as well as up to hundreds of metres above the surface."

4. To evaluate the VIT parameterization, using observations from surface monitoring sites only is not sufficient. Due to the limitations and uncertainties acknowledged in this study, it is actually better to develop and test the VIT parameterization based on a small domain, maybe city size, which has relatively simply traffic and meteorological conditions as well as observational vertical profiles of chemical species for evaluation. The manuscript shows vertical cross-sections in Figure 10. Without the observed vertical distributions of NO2, PM2.5 and O3, it is hard to determine whether using the VIT parameterization leads to an improvement.

We have shown that VIT leads to a statistically significant difference in the model mean values (see above response), and that VIT leads to an improvement in surface concentration predictions for the majority of the metrics used for evaluation. Surface monitoring network data are also the standard benchmark used for air-quality model performance. Measurements in the vertical for these species are a good idea – but would require a separate measurement study for data collection, and may be expensive to mount. For example, collocated Doppler LIDAR, ozone LIDAR and particulate matter LIDAR near roadways might be one way to achieve the vertical resolution needed to examine these effects (which is why we argue in section 4 for such studies to improve on the parameterization shown here). We also note that our evaluation covers the specific region most important for human health impacts - pollutants – the surface level the human population inhabits. Mounting a major measurement study is beyond the scope of the current work, but we are hoping that this work may be used as an argument that such a study should take place.

One thing that we could do, however, to address the referee's concern is examine the city-scale results of the parameterization across North America directly, through the use of an urban mask on the observation stations used for evaluation, using our human population field. We used the population field depicted in Figure S2 (a) to regenerate the model performance specifically for model grid cells in which the population was greater than 800 km^{-2}, and have included these results into a new table in the Supplemental Information. We have added the following text contrasting the two Tables, in Section 3.3:

"Following the above comparison using all available surface monitoring network data (Table 2), we carried out a further evaluation where the stations were selected based on human population within grid cells (Figure S2(a)), with only those stations in which the population exceeded 800 km^{-2} used for analysis. The results of this evaluation are shown in Table S2, which may be compared to Table 2 to show the relative influence of VIT on high population areas. We note that the magnitude of the improvement in model performance associated with VIT has increased for many statistics when high population (i.e. high vehicle traffic) areas are examined separately in this manner; for example the incremental improvement in North American NO_2 mean bias changes from 1.053 ppbv for all stations versus 1.782 for population > 800 km^{-2} stations, and the incremental improvement in PM2.5 MGE for North America changes from 0.249 to 0.665 μg m^{-3} (both numbers are differences between No VIT and VIT values in Tables 2 and S2 in each case. The number of model performance improvements with the use of VIT has increased when grid cells with populations greater than 800 km^{-2} are evaluated (62 out of 72 metrics improved with the use of VIT in Table 2, while 66 out of 72 metrics improved for stations corresponding to grid cells with populations greater than 800 km^{-2}). Most of these additional improvements were associated with better ozone prediction performance in urban regions."

5. On line 108, the manuscript states "Here we make use of both the observational and LES modelling studies to devise a parameterization for VIT." This is the last place in the manuscript that LES is referred to, so it is rather confusing how LES modeling studies are used in this study. As discussed above, please acknowledge other LES studies, and also careful elaborate how "LES modeling studies" are used here. If not used, please also clarify.

   Certainly – Large Eddy Simulations are one form of Computational Fluid Dynamics model studies, which contributed to the data appearing in Figure 1, as well as the information in 3 studies quoted in Table 1 (Kim et al., Woodward et al., Zhang et al. papers). Specifically, they have been used to provide the justification for the use of equation 1 in describing the functional form of the decrease of VIT TKE with height. We have shown that this formula usually accounting for a large amount of the variation (R^2 from 0.54 to 0.98). They have also been used to demonstrate, from that formula, the range of mixing lengths resulting from equation (1) and those studies. We have modified the text in section 2.2 to specifically identify the LES model results used in the work:

   *"*We examined four datasets (the observations of Rao *et al.*, 2002, and the LES modelling of Kim *et al.*, 2016a; Woodward *et al.*, 2019; Zhang *et al.*, 2017) to evaluate the extent to which a Gaussian distribution may be used to represent the decrease in VIT with height above moving vehicles, as well as examining the expected range of mixing lengths which may result from VIT."

Final Referee 2 Comments and Responses

Final Referee 2 comments and text additions resulting from the responses are in normal font, responses are in *italics*.

The authors provided reasonable explanations to the concerns and comments. The manuscript is clearer and improved. Please correct the following issues. Line numbers specified below are for the tracked version of the manuscript.

In the paragraphs and figures related to 90% confidence level, please specify what values are calculated and shown in the maps or the time series plots. It causes confusion as confidence level shouldn't exceed 1. A logical expression is included in the paragraph on line 552-560. But an equation leading to the values needs to be clearly defined.

Response: *This has been added to the revised manuscript – we have: (1) modified all confidence ratio values to clearly state they are confidence ratios, not confidence levels, both in the manuscript text and in the appropriate Figure captions, also mentioning in the latter that "values greater than unity (red colours) indicate the model simulation values are different at greater than 90% confidence", and (2), the text introducing this concept has been modified to include the confidence ratio as a separate equation in the text (new equation (14)), clarifying its relationship to confidence limits. One advantage of confidence ratios is that they also show areas which considerably exceed or fall below the 90% confidence level – shown in our figures as darker red and lighter blue areas. The modified text reads as follows:*

"The significance of the differences between VIT and no-VIT simulations was estimated using 90% confidence levels, expressed here as confidence ratios. The difference between the mean values of the two simulations (M_{VIT}, M_{NoVIT}) becomes significant at a confidence level c if the regions defined by $M_{VIT} \pm z^* \frac{\sigma_{VIT}}{\sqrt{N}}$ and $M_{NoVIT} \pm z^* \frac{\sigma_{NoVIT}}{\sqrt{N}}$ do not overlap (where N is the number of gridpoint values averaged, the σ values are the standard deviations of the means, and z* is the value of the $\sqrt{c}$ percentile point for the fractional confidence interval c of the normal distribution, where z*=1.645 at c=0.90. Grid cell values where the mean values differ at or above the 90% confidence level are thus defined as the confidence ratio:

$$CR = \frac{|M_{VIT} - M_{NoVIT}|}{\frac{z^*}{\sqrt{N}}(\sigma_{VIT} + \sigma_{NoVIT})} \qquad (14)$$

Where, when z* =1.645, and the other terms are as described above, a CR value greater than unity defines the difference between the model simulations at that gridpoint as being significantly different at greater than the 90% confidence level. The mean values at each gridpoint and their standard deviations may thus be used to determine the confidence ratio at each gridpoint – these values for each of the mean differences of Figure 7 are shown in Figure 8, where the colour scaling in Figure 8 and other confidence ratio Figures which follow use red colours to indicate differences which are significant at greater than 90% confidence. Gridpoint differences which exceed the 90% confidence level requirement to progressively higher degrees are shown as progressively darker red colours, while differences falling progressively further below the 90% confidence level requirement are shown as progressively lighter blue colours, in these Figures."

Line 102-107: The round bracket after "3.2x3.2 km domain" should be moved to the end of the sentence.

Response: we've modified the sentence so that spare bracket was removed altogether, and the way in which the resolution and domain size was referenced was uniform throughout the sentence.

The manuscript uses "UT". But it is better and clearer to specify whether it is UTC, UT1, or others, as these versions are different.

Response: A valid point, though we note that the difference between UTC and UT1 is kept within 0.9 seconds / year of UT1 – the difference is insignificant on the time scales and for the subject matter studied here (c.f. https://en.wikipedia.org/wiki/Universal_Time#:~:text=UTC%20(Coordinated%20Universal%20Time)%20is,which%20civil%20time%20is%20based.&text=Whenever%20a%20level%20of%20accuracy,UTC%20is%20known%20as%20DUT1). Here we are using the international atomic clock-based UTC. We have modified the manuscript and the SI so that UTC, rather than UT, is referenced throughout the manuscript