General comments:

The authors didn't provide sufficient feedback to the comments of reviewer 2 in most cases. I have summarised below some of the specific comments that deserve adequate feedback. In addition, the authors need to answer some theoretical questions underpinning air pollution exposure modeling;

1. Air pollutants are not only defined spatially but temporarily and both characteristics interact in a complex relationship that can't be disentangled especially with the approach used by the authors. They also use just one pollutant (NO2) which is very localized. This might explain the performance difference reported between global and local models.
2. The assumption behind using traffic and population-related variables to construct their models is not rooted in the literature because other land-use and temporal variables have been reported to explain variability in air pollution levels.
3. The statistical approaches used for this analysis are not contextualized in their manuscript and the supplementary materials provided are too theoretical.
4. I doubt this paper will benefit from any other revision because the issues are related to how the research question was conceptualized. Does this work have any potential to add anything new to the existing literature in this field?. My answer is NO.

Specific comments

R2 comment: The overall readability of the text is very poor: the presentation of machine/statistical learning models is very superficial (there are no formulas and no technical modeling aspects are discussed); To make the paper more readable, the formulas are added in the supplementary material.

Author's Feedback: **A reference to the equations and technical aspects of the considered models is now available in supplementary, equations.**

My comment: Although, the authors have added a document with the general equation of the models considered in their analysis. However, these equations are mostly general equation description found in textbooks or papers. The application of these model parameters to their data wasn't discussed in detail. I would have expected the authors to include some equations summarising their models in the manuscript to complement the text provided.

R2 comment: The paper is very long and confusing: reading requires continuous jumping from one section to another to understand what models and assumptions the authors are analyzing.

Author's Feedback: I changed the structure of the methodology. In the renewed methodology, the data (2.1) is introduced first, followed by a short elaboration per model used in the next subchapter (2.2); subsection 2.3 elaborates on feature selection which is an important part as it determines the relevant variables for the modeling; the last section of the methodology provides insights into how the models are evaluated and used, thereby showing the relevant models in a table overview.

My comment: In general, the paper is still confusing. I struggle to understand why the decisions to construct these models. What other information do these different models add to understanding the spatial pattern in N02 exposure?

R2 comment: Methodology: Models are presented without specifying their technical characteristics, differences and rationale for their use. No formulas explaining the structure of the models (e.g., the spatiotemporal structure of random effects) are included by the authors. A paper using statistical methodology should never assume that the reader is aware of the methods;

Author's Feedback: Information on technical characteristics, differences, and rationales for modeling can now be found in supplementary equations and supplementary parameters. The temporal aspect is neglected in the models unfortunately, as this is outside the scope of this research, however, should be addressed in future research.

My comment: The technical characteristics discussed in this section are very theoretical and largely didn't describe why and how these methods were applied to this analysis.

R2 comment: There are dozens of linear models and spatio-temporal mixed-effects models in the literature that provides a fair trade-off between interpretability and predictive ability. In the text, none of them are mentioned. I do not intend to cite specific ones, but just type in Google scholar "spatio-temporal models" to retrieve them. I would suggest starting with the spatio-temporal modeling of Wikle-Cressie (who have made history in this branch of research) and colleagues [1, 2];

Author's Feedback: Thank you for the suggestion. We agree that the spatiotemporal modelling works from Wikle Cressie is a great reference and provide inspiring perspectives. We also agree that the spatiotemporal mixed-effect models are making impressive progresses in improving both predictive ability and model interpretability. What is slightly confusing to us regarding the comment is that our study has not reached to the next milestone of spatiotemporal modelling but so far confined into spatial modelling, as many issues remain at this level. We agree that missing the temporal dimension add difficulties in interpretation, uncertainty assessment, and prediction also over space, but joint spatiotemporal modelling greatly complicated the modelling and we believe it is more illustrative and apprehensible to firstly study at a lower dimensionality. We add in the revised manuscript about the future vision of spatiotemporal mixed-effect modelling. Added: "Simultaneously, the absence of the temporal dimension poses challenges in interpretation, uncertainty assessment, and spatial prediction. Still, joint spatio temporal modeling greatly complicated the modeling and we believe it is more illustrative and reprehensible to firstly study at a lower dimenstionality."

My comment: This is a fundamental issue for a paper titled: **"A close look at using national ground stations for the statistical modeling of NO2"**. I wonder about the gap this paper is trying to fill. We know already that air pollutants are defined by their spatial and temporal characteristics. Also, the interaction between these characteristics can be complex – more reason why complex models that can account for these interactions are becoming more popular in air pollution exposure science. So a statistical modeling of NO2 at local and national levels without accounting for temporality would not tell us the full story of these complexities in air pollution exposure modeling. The title also didn't make this explicit.

R2 comment: Section 3.1.2: why did you move from a 20-fold CV for global model assessing to a N-fold (LOO) CV for local models? This choice introduces some issues when comparing models as the predictions are computed using sample sizes;

Author's Feedback: Because the local dataset has fewer data, therefore a LOO approach suits better. I removed a part of 3.1.2 and moved it to the methodology where it is more suitable. Concerns: "With the mixed-effects model, fixed and random effects are included. Fixed effects consist of the most influential predictors while random effects account for potential spatial trends in the data. The spatial trends in the data related to observations being clustered in a way. The spatial character of the observation, i.e. whether an observation is situated in an urban area, low-populated area, or far from road area, accounts for the random effect in the model. In contrast, the linear model composes all the fixed effects while neglecting the possibility of observation clustering. Additionally, two kriging methods are used for local modeling, being ordinary- and universal kriging.

My comment: The authors argued in their previous response to the comment about spatial cross-validation that spatial cross-validation is not a suitable measure of model performance because they have no theoretical underpinning. I am confused as to why the leave one out cross-validation – a spatial cross-validation method was then used to assess the local models' performance. It’s also surprising that the authors compared the global model performance assessed using random cross-validation to local models performance assessed using spatial cross-validation. What is the theoretical underpinning for this ?. This is not an appropriate way to go about it and the conclusion from this comparison is questionable.