This study addresses the challenge of accurately modeling nitrogen dioxide (NO2) concentrations, which is essential for understanding air pollution's health and environmental impacts. Given that NO2 levels vary significantly across different spatial settings, especially in urban areas, the authors investigate how various statistical and machine learning models perform across urban, suburban, and rural regions. By comparing global and local models—trained on datasets from Germany, the Netherlands, and specifically Amsterdam—the study evaluates the strengths and limitations of each model type.

The authors tackle this problem by creating spatially-defined groups based on traffic volume and population density, which allows the models to account for spatial heterogeneity and examine prediction patterns across different zones. They apply both linear and non-linear models, as well as mixed-effects and kriging methods, to see how well each approach handles the spatial intricacies of NO2 concentrations. Model performance is evaluated through standard metrics like R^2, RMSE, and MAE, revealing that ensemble-based models perform best in rural areas, while urban areas remain challenging due to data heterogeneity. This methodology highlights the critical role of spatial grouping and suggests that relying solely on prediction accuracy without considering spatial context may lead to misleading results.

The study is interesting and investigates an important aspect of using data-driven methods to predict air pollutants on a large scale. There are however some concerns I have identified before the study can be published, listed below.

Methodology

1. (section 2.1 Data) The authors refer to the two datasets as global and local. However, the global dataset only contains data originating from two neighboring countries, with similar characteristics. I suggest to change the name from global to something more appropriate as these two countries do not reflect the global stage.
2. Line 100-105: Can the authors discuss the inclusion of the distance to roads feature in clustering the regions or include a citation to a study that supports such distinction? Traditionally, the classification of urban, sub-urban and rural areas is based on population alone.
3. Line 111-103: Discuss this more, as it seems like these two statements are contradictory.
4. Line 135: 50000 estimators seems extremely excessive for the boosting algorithms. Traditionally, the number of estimators is in the hundreds. I would suggest the authors to discuss the reasoning behind this. While gradient boosting algorithms are resilient to overfit, they are not overfit-proof and including a very large number of estimators has the potential for overfitting.
5. Line 138, 141: Specific references to the supplementary material is missing. e.g. In which table are these results presented?
6. Line 150: SI table 4 shows the results for Linear, LASSO and Ridge and SI Table 5 shows the results for Random Forest, LightGBM etc.
7. Section 2.2.2 and 2.2.3: The description of the methodology for the modeling part is superficial. The authors should expand these sections significantly, especially section 2.2.3 which describes the mixed-effects model and the Kriging method.
8. Section 2.3: The SHAP values plot should be in the main text of the manuscript instead of the supplementary material, as it contains useful information that are used in the main study.

9. Section 2.3: It's not clear whether the authors have normalised the data used here for the machine learning algorithms. From the SHAP figure it seems that the target variable range is not normalized. Normalization of the input and target variables to the 0 to 1 range ensures that all the input variables are equally weighted (unless the setup of the model specifically requires asymmetric weights) and the input variables with the largest range (or absolute values) do not dominate the others.
10. Section 2.3: Consider expanding this section as it's not clear which predictors are selected and how the process is implemented. Also, a table of all the predictors used and their origin, range and name could be useful to the reader.
11. Line 192: In Kerckhoffs et. al (2019) the maps were produced by measuring for limited time periods using mobile sensors. The temporal resolution of the predictions of the models here deal with much coarser temporal resolutions.

Results

1. Line 201: How was this 20-fold validation performed? Since the number of data points is small, I suggest to perform cross-validation with a small number of folds (e.g. 5). How many data points were selected in each fold and on which set these metrics were evaluated on?
2. A graph of ground truth vs predictions would be beneficial here to identify edge cases at which the algorithms do not perform well.
3. Line 216: It's not clear what the spatial resolution of the predictions is nor how these maps were created.
4. Table 6: Another useful metric that could be used to gauge the significance of the RMSE and MAE metrics would be percentage error wrt to ground truth.
5. Line 256: When was leave-one-out cross-validation used before this? Discuss how this was implemented.
6. Line 260: A significant limitation of the study setup is the fact that the most heterogeneous group (urban) is the least represented in terms of number of data points. This should be discussed by the authors.
7. The large number of models and the use of different set of models for each group makes these comparisons very difficult. Also, why have the authors used similar models (e.g. LightGBM and XGBoost) which makes the comparisons even more difficult to follow.
8. Line 290: “These outliers were removed...” Discuss this choice more. How many points were removed and why do you think these points performed poorly?

Disussion

1. Line 323 “While a well-trained model...” Not clear what this sentence conveys to the reader.
2. LIne 334: I would argue that spatial cross-validation is essential in this kind of models, as it ensures that the model learns sufficient representations to generalise to other regions that do not have any ground stations. In this case, spatial cross-validation would be beneficial within the groups selected. i.e. ensure that the model generalises well within the urban cluster
3. Line 342-346: This is counter-intuitive. It pollutes the urban group by including areas with less population that the initial definition (upper 75% quartile) but it was necessary to expand the size of the dataset to a sufficient level. The authors should make this clear and

address it as a limitation of the study. While it's understandable this was a necessary step in the experimental setup of the study, it needs to be clearly addressed.

Conclusions

1. “In this study, we understand...” Consider changing the word understand to investigate