**Again, we appreciate your time and effort into the feedback given. Our replies to your feedback is expressed in bold font.**

# Feedback 1

Thanks for your substantial work. I think this is a good work to stimulate discussion on with predictive mapping of environmental factors. I would like to recommend a major revision to address the following comments.

**There is now a pdf including all the supplementary figures.**

**Major:**

Line 75-80. - Could you confirm if the global dataset overlaps with the local dataset? Even though they are from different sources, some of them might be the same stations but with different names.

**There is no overlap. Also see figure 4 and figure 5 of the supplementary figures.**

Line 100-105. - Consider renaming the three groups. The current version reads a bit misleading as it looks like they partially overlap.

**The three groups have been (partly) renamed to" urban", "suburban", and "rural" spatial groups so that there is no indication that they might overlap.**

Line 110-115.
- If the definition of the three groups changed between local and global models, would it still be fair to compare their model performance? For example, the threshold adjusted from 0.75 to 0.5 for the "urban" in the global and local models. It could impact the conclusion. In the result section, the different distributions of model predictions could stem not only from the levels (global or local) but also from the different definitions of the "urban" group.

**The adjustment of the threshold from 0.75 to 0.5 for defining "urban" areas in the local dataset addresses the inherent differences between global and local datasets. The local dataset has fewer samples but features higher population densities. In this context, a lower threshold for "urban" classification is necessary to accurately capture areas with the highest population concentrations. By lowering the threshold, we can better represent the spatial distribution of densely populated areas in the local context, leading to more precise and relevant modeling results.**

**This adjustment allows for a more flexible and context-sensitive definition that better reflects the unique characteristics of the local dataset. However, it's important to approach comparisons between models using different thresholds with caution. Adjusting the threshold is a methodological choice intended to improve the accuracy and relevance of the model in capturing local phenomena.**

**Your point is valid, and I have clarified this rationale in the main text.**

Line 110-115.
- Have you tried to develop models trained with the balanced numbers of instances across spatial groups? i.e., the same number of instances in each group, as the statistic learning model can be easily biased due to the unbalanced distribution of training instances in each category.

**Not necessarily, although adjusting the urban threshold for the local dataset helps to address this issue indirectly. For the global dataset, this is less relevant due to its larger sample size.**

**For the local dataset, 56 observations are classified as "urban," 46 as "suburban," and 30 as "rural." This adjustment was made to partially address the imbalance in the number of instances across these spatial groups, aiming for a more equitable distribution between "urban," "suburban," and "rural" categories. While this doesn't completely eliminate the imbalance, it does help the local model better reflect the higher population density characteristic of urban areas in the local context. We acknowledge that the unequal distribution of instances across groups could introduce bias in statistical learning models, but this threshold adjustment was an initial step to mitigate such effects. This has also been briefly noted in the main text.**

**The decision to adjust the threshold for the local dataset was justified by the need to achieve a fairer distribution of instances across groups:**

- **With a 0.75 threshold, there are 25 samples in the urban group, 56 in the suburban group, and 51 in the rural group.**
- **With the 0.5 threshold, there are 56 samples in the urban group, 46 in the suburban group, and 30 in the rural group.**

Figure 6. - Comparing the spatial variations of predictions between global and local models is challenging due to the differing algorithms used.

**Thank you for the feedback. We acknowledge that comparing spatial variations between global and local models is challenging due to the differing algorithms used. In the revised version of the paper, we have addressed this issue by highlighting how the distinct approaches of the models impact the observed spatial patterns. Figure 6 now explicitly discusses these algorithmic differences and their effects on the comparability of predictions.**

## Minor:

Abstract: 1. What is your final conclusion or the key message? Better to specify it in the final sentence in the abstract.

**Extended the last part of the abstract.**

Line 75-80.
- Please clarify the spatiotemporal resolution of the models.

**Global Model:**

- **Total Area (km$^2$): 398,087.4**
- **Total Points: 482**
- **Point Density (points per km$^2$): 0.0012**

**Local Model:**

- **Total Area (km$^2$): 196.4**
- **Total Points: 116**
- **Point Density (points per km$^2$): 0.591**

**I've included this information in the main text.**

- Add the number of stations of the two datasets.

**Added**

Line 180-185. - Please specify why the feature selection. If it is about avoiding collinearity, why not use the VIF value?

**The feature selection process utilizes Shapley values primarily to identify and prioritize predictor variables with the most significant influence on NO2 concentration levels, enhancing model performance and interpretability. Shapley values are advantageous because they provide a nuanced assessment of each feature's contribution by considering all possible combinations of features. This approach allows for a detailed evaluation of feature importance, accounting for interactions and correlations between features.**

**While the Variance Inflation Factor (VIF) is effective for detecting multicollinearity by measuring how much the variance of a regression coefficient is inflated due to collinear predictors, it does not directly address feature importance or interactions. VIF is primarily a tool for identifying redundant features rather than assessing their contribution to the target variable. In contrast, Shapley values offer a comprehensive measure of how each feature impacts the prediction, including the effect of feature interactions, which is crucial for understanding the model's behavior and improving its performance.**

**Therefore, Shapley values are chosen over VIF because they provide a more holistic view of feature importance and interactions, which aligns with the goal of enhancing model performance and interpretability in this context.**

**The VIF analysis shows that the population variables are correlated with each other, as evidenced by their high VIF values. Despite this, both population_1000 and population_3000 might be crucial predictors of NO2 levels based on the Shapley**

**values, indicating they provide significant and complementary information to the model.**

**To address the potential multicollinearity while retaining the valuable information from these features, regularization techniques such as Ridge or Lasso regression are employed. These methods can help manage the redundancy while still incorporating the features' predictive power. This is addressed in section 2.2.2 Multiple linear regression. The VIF scores for both the global and local datasets are included in supplementary table 2 and table 3 respectively.**

Line 190. - What do you mean by the out-of-sample cross-validation? Did you use external/third- party data sets (other than the global/local measurements)?

**This out-of-sample cross-validation terminology is implemented since the previous feedback round, as feedback from another reviewer entails:**

**“Also, for the sake of completeness, I suggest adding the word "out-of-sample performances" every time you use CV because it must be clear to reader that all the metrics are computed in a training-test framework to assess predictive capacities of model (and not in-sample fitting);”**

Line 225-230. - 20-fold cross-validation means the training set is divided into 20 parts. The global model was trained with 482 observations. In each iteration, only 24 observations are replaced. It is fine to do 20-fold cross-validation. But for the small size of samples, it is not a proper choice.

**To clarify, we used repeated random sampling validation rather than traditional 20-fold cross-validation. In our approach, we sampled 25% of the observations as test sets in each of the 20 iterations, with the remaining data used for training. This method involved repeated sampling with different random states to ensure robust evaluation. We recognize the limitations of using cross-validation with smaller datasets and appreciate your input on this matter. The relevant part in the main text is adjusted to this.**

Line 235-240 - This part belongs to the discussion section.

**Inserted this part under section “Accounting for spatial groups” in the discussion section**

Figure 6. - Would it be possible to plot also the distribution of mobile predictions from Kerckhoffs? Another crucial aspect of the air pollution map is the spatial variations. I expect to see the different levels of variations captured by global and local models as well as fixed-site vs mobile measurements.

**A new table (8) showing the residuals per model, by comparing it with the open NO2 dataset (Kerckhoffs), is added to the main text. The spatial residuals per global and local model can be found in supplementary figure 22 and figure 23 respectively.**

Line 290.
- UK = universal kriging? Please use the full name in the bracket.
**Adjusted.**
- What is the linear model in the table? Lasso? Ridge? Why are the algorithms used for global and local models not aligned?

**The performances of the algorithms used for global models, perform substantially poorer on the local dataset, also for the Lasso and Ridge algorithms. See the leave-one-out cross validation results (also included in supplementary, table 4; a reference is made in the main. text):**

Line 325-330.
- It is great to involve external datasets for cross-checking.

**Also added a table (8) with model residuals (comparing with NO2 mobile map of Kerckhoffs)**

- What is the true reason for filtering out outliers? Enhancing the low correlation is not a proper reason. Rephrase, please.
- Use the past tense. This is something you have done.

**Adjusted. To improve the clarity of the correlations between the models and the open NO2 dataset, we addressed some extreme prediction values. These outliers were removed to prevent them from skewing the analysis and to provide a more accurate representation of the correlations. This is mentioned in the main text too.**

Line 335. - Another reason for kriging is its stationary assumption.

**Added**

Line 360-375. - If the temporal analysis is not performed, please put it into the limitations or future work section.

**Shortened the discussion section. The feedback to which this apply, is removed to make the paper more compact.**

Figure 7. - Can you specify which models are global and which are local models directly in the figure? To increase the readability.

**Added in figure 7.**

Line 380-405. - Another nice paper I found that discusses also the difference between the global and local models is “Integrating large-scale stationary and local mobile measurements to estimate hyperlocal long-term air pollution using transfer learning methods”. They found also a significant improvement in modeling performance in urban background areas when involving global knowledge. This would be a good citation.

**Thank you for this suggestion. I put it in the discussion (section “Global and local predictions”).**

Line 445-450. - Need to also mention the missing meteorological information.

**Shortened the discussion section. The feedback to which this apply, is removed to make the paper more compact.**