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.

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.

Line 100-105.

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

Figure 6.

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

Minor:

Abstract:

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

Line 75-80.

-Please clarify the spatiotemporal resolution of the models.

- Add the number of stations of the two datasets.

Line 180-185.

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

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)?

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.

Line 235-240

- This part belongs to 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.

Line 290.

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

Line 325-330.

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

Line 335.

- Another reason for kriging is its stationary assumption.

Line 360-375.

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

Figure 7.

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

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.

Line 445-450.

- Need to also mention the missing meteorological information.