Major revision

Monday 25 march 2024

Foeke Boersma & Meng Lu

We appreciate your time and effort into the feedback given. There are a lot of good suggestions which helped us improve the quality of the content. Our replies to your feedback is expressed in bold font.

Feedback – reviewer 1

General comments:

Currently, many studies are using spatially sparse fixed-site measurements to map air pollution on a large scale, ignoring the local spatial heterogeneities such as the intra-city variations. This article evaluated the performance of various algorithms across different scales and validated the accuracy separately in subsets categorized by road density and population density. They found that the model performance varied significantly at different spatial locations. The pattern was found to be different in “global” and local models. The comparison between “global” and local models in terms of intra-city distribution patterns is valuable. However, in its present form, I cannot recommend the article for publication. With substantial revision and restructuring, this article could be a useful addition to the existing literature.

The writing needs further improvement. The current version is not easy to read. First, this is too long. I appreciate the solid work of the authors. But please simplify the main text and consider moving some descriptions/figures to the Appendix. Keep only the core story in the main text and make sure the primary findings and the most important messages stand out. Second, consider restructuring the method/data and the result section. Third, the caption of figures and tables needs more details, including the unit of NO_2. Fourth, Clarify definitions like “Far from road” vs “Rural”. Last, please pay attention to the tense usage.

Specific comments:

1. The mobile measurements from Kerckhoffs et. al., 2019 were measured on the road. How can they validate the accuracy for the “far from road” group? Did you perform any adjustments?

Reference to Kerckhoffs is relevant as the study area entails the area of Amsterdam as well (i.e. similar area of interest)

2. Table 1 describes the predictor features. Why not include land use proportions? Land Use Regression models are efficient and well-accepted methods in air pollution modeling.

Thank you for your comprehensive and useful feedback and suggestions. Considering your suggestion about including land use proportions in our predictor features, we believe that we already incorporated the land use proportions via various buffer analyses for variables including industry, building density, and road types. In case we

misinterpreted this suggestion, please let us know so that we potentially can make the necessary changes.

3. Figure 4. It would be better to plot the map of differences between the model tested and the benchmark (i.e., NO2 estimations from Kerckhoffs et. al., 2019). I would be curious about the difference in spatial distributions between the "global" and local models.
4. A restructuring of the data/method section is recommended. Begin with the introduction of the data source, ensuring clarity on the source of the population information and road class when discussing spatial groups. Consider adding a table summarizing model input/algorithms for ease of understanding. Move some algorithm introductions to the appendix.

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.

5. Please explain why 20-fold cross-validation?

We select a 20-fold cross-validation to encourage stable estimate; added this sentence in main text.

Technical corrections:

I have listed some specific points. But not limited to them.

Abstract:

The abstract attempts to encompass numerous findings but allocates insufficient space to elucidate the methodology and experimental setting. A substantial rephrasing of the abstract is needed.

Line 1-5, toing and froing, can be simplified.

Line 6, please provide more details about the meaning of "spatial heterogeneity" in this context.

Added: "in which characteristics and/or phenomena may change over space"

Line 9-10 what is the local and global model? Define first, before using it.

Elaboration given:

"Local models are based on the Amsterdam area and perform predictions on the same area. Global models are based on observations throughout Germany and The Netherlands while predictions apply to several smaller areas of interest in Germany and The Netherlands"

Methodology:

Line 100-105, not clear. How do you divide the area? Purpose? What is the time frame of these national measurements? Frequency of measuring? Any preprocessing? More details are needed here. How do you define the less densely populated area? What is the source of the population density data?

Line 121, “rural”= “Far from roads”? Please keep the terminology consistent. Changed to “far from roads”

Changed to “far from roads”

Line 123, the label of models should be provided as the legend in the figure instead of in the caption.

Line 130-135, unit of NO_2 is missing. This paragraph is not informative. The values can be integrated into the figure 1.

Removed alinea

Line 145, More details about kriging and accuracy are needed.

Added a supplementary, equations and supplementary, parameters containing more details. Reference in main text is added too.

Line 160-165, is the traffic volume used as the annual average? “The traffic volume is expressed in average hourly traffic” > “The traffic volume is expressed in average hourly traffic, measured over the year 2017.” Table 1. it would help readers to understand the data distribution by adding columns such as numbers and some statistics like mean, median, quantiles etc.

Removed table 1, however added a table with descriptives of predictors that are used to classify spatial groups (Table 1. Descriptive statistics for each relevant variable in the determination of spatial groups for the local- and global datasets)

Line 168, the section title should begin with a capital letter, and further refinement is necessary in terms of formatting.

Line 190, not clear. Please do not refer to the citation but to the dataset you have described in section 2.1. Removed this line, also with the intention to limit the main text.

Removed this line, also with the intention to limit the main text.

Line 195, rephrase please instead of a direct quote. Removed quote, also with the intention to limit the main text.

Removed quote, also with the intention to limit the main text.

Line 196, details of the tuning strategy are missing.

Added a supplementary, equations and supplementary, parameters containing more details. Reference in main text is added too.

Result and discussion:

Line 465, how do you compare the influence of predictors between cities? The feature importance is a relative value. The magnitude is not meaningful when compared to the other models.

Actually by looking at the prediction patters. For Hamburg, high clusters of NO2 seem to coincide more with traffic related variables and less with population related variables; for Utrecht, high clusters of NO2 seem to coincide more in the city center itself (higher building density; population) than traffic related variables (high NO2 clusters along highways are less visible here).

Line 515, which is opposite to the common knowledge (see Hoek et. al., 2008). Can you explain why non-linear model predictions were smoother?

Based on this, changed:

“The spatial prediction patterns show that non-linear methods generally predict more smoothly than linear methods. Additionally, clusters of predicted air pollution differ within and between cities.” → Lastly, non-linear prediction patterns seem to be less prone to overfitting compared to linear methods, and different modeling techniques lead to different NO${_2}$ clusters in the prediction map. (conclusion)

“The spatial prediction patterns show that non-linear methods generally predict more smoothly than linear methods.” → The spatial prediction patterns show that non-linear methods generally are less prone to overfitting than linear methods. (abstract)

“Generally, the linear spatial predictions are more discrete, compared to the non-linear techniques. To elaborate, values in the linear predictions are more extreme (i.e. below 15 or above 50) and relatively low and high values tend to be closer to each other, in contrast to non-linear prediction patterns, where the predicted NO2 patterns tend 305 to be more smooth” → Generally, the linear spatial predictions seem to be more prone to overfitting as these prediction maps are characterized by a higher share of extreme values compared to non-linear techniques (i.e. below 15μg/m3 or above 50μg/m3). (section 3.1.1)

Feedback – reviewer 2

Introduction In this paper, the authors aim at discussing the role of spatial heterogeneity in spatio-temporal prediction of ground-level air quality (i.e., airborne pollutant concentrations) using both a global and a local approach. The authors refer to a global dataset consisting of all the ground station measurements in Germany and the Netherlands, and to a local dataset comprising only the ground monitoring station in the Amsterdam area. The authors attempt to assess the performance of several algorithms across different spatial scales (global and local) and validate the predictive accuracy when ignoring and when considering local spatial characteristics (i.e., density and population density). The main findings state that that the model performance strongly depends on the considered spatial scale and on the considered spatial locations.

The paper addresses the issue of spatial prediction of air quality in a very broad way and tests several interesting dimensions. However, the work done and the methodology are not rigorous and have several critical points. In particular, several methodological inaccuracies, poor analytical rigor and unclear (if not unwarranted) choices emerged during the reading. Therefore, I suggest that the paper should not be accepted in its present form and should be subject to major revisions (especially in methodology).

General comments:

Hereafter, I state my major concerns that need to be addressed and clarified.

• The overall readability of the text is very poor:
o 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. A reference to the equations and technical aspects of the considered models is now available in supplementary, equations.

o names and acronyms are inserted into the text without appropriate discussion and description;

lightgbm, xgboost, and lasso are now first written out before referring to the acronyms/names.

o many sentences need to be rewritten as they are unclear;

check
o 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.

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.

• Methodology:

o 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;

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.

o Machine learning models (e.g., random forests, xgboost and lightgbm) require great caution and understanding before their use. They are (partially) black-box models with attributes devoted to prediction rather than interpretation of phenomena (while they greatly improve predictions, they also make the results lose interpretive meaning and risk becoming tools that cannot be used by policy makers or practitioners);

Thank you for the comment. We agree that tree-based machine learning models require great caution and understanding before their use. However, despite that classical statistical methods such as standard linear regression (linear regression without penalty) have clear interpretations on the parameters, a correct interpretation also depend on the model assumptions. That is, if the true model is highly non-linear and a linear model is used, the standard linear regression can also lose interpretive meaning and risk becoming tools that are not usable. Models such as ensemble trees, besides their potential predictive power, can be interpreted using for example marginals and permutations, also, the uncertainties can be assessed. We admit that many desirable properties they don't have, and that is a main reason that in our study, we compare them with statistical models such as Lasso, standard linear regression, and Kriging.

o The expression "linear models" denotes the class of models that are linear in their parameters. It is a very large family. Personally, I struggled to understand which linear models you considered: linear regression? at what scale (original or logarithmic)? ridge and LASSO are linear, but differ significantly from pure OLS because of the penalty;

Thank you for your question. By linear model we mean the model with a linear relationship between predictors and response,

Y= Xß

We see Lasso and ridge as a general form of linear regression models compared to "pure" linear regression (regression without regularization) due to their regularization terms.

We clarified this in the revised manuscript. Rewritten to:

"The key variables highlighted by the random forest model are chosen as predictors in Multiple Linear Regression (MLS). MLS, a statistical method employing multiple explanatory variables to forecast the response variable, operates within a linear framework, where the relationship between predictors and response follows the form Y = Xß. However, linear regression techniques can be characterized by complexity and/or overfitting of the model. In this context, Least Absolute Shrinkage and Selection Operator (LASSO) and RIDGE regression emerge as broader forms of linear regression models, incorporating regularization terms, unlike "pure" linear regression lacking such regularization." (~151 – 157)

o Transformations and distribution of data: the text does not mention the problem of positive skewness (typical of the airborne pollutant concentrations world) (Mudelsee). Clearly, the prediction changes considerably if transformations (e.g., logarithm) are applied to adjust for skewness or if general models with non-Gaussian distributions (e.g., GLMs and GAMs) are used. Where do you stand with respect to this problem?

Thank you for the comment. We agree that the air pollution concentration is also not normally distributed in our study and this is indeed a concern. Our study uses the results from Lu, 2023, who used the same dataset as our global dataset. In their study, careful considerations have been given to different transformations, likelihood functions, and loss functions to address the issue of non-Gaussian distribution, with a detailed discussion of the results. It was found that using a transformation, likelihood function, and loss function that matches with the more-likely distribution (the Gamma distribution) does not improve the modelling results but worsened the prediction errors and the uncertainty quantification. In future study we aim at uncovering the reason for these model behaviours.

We add this issue in the revised manuscript in discussion.

Added:

Complementary, airborne pollutant concentrations are often positively skewed. To adjust for positive skewness, transformations can be applied but also
cause prediction changes which currently are not revised in our research. Simultaneously, Lu et al. (2023) examine several techniques such as transformations, likelihood functions, and loss functions to address the issue of non-Gaussian distributions. Thereby, they observed that using a transformation, likelihood function, and loss function that matches with the more-likely distribution (i.e. Gamma) does not improve the modeling results but worsened the prediction errors and the uncertainty quantification (Lu et al, 2023). (~428-434)

o There are dozens of linear models and spatio-temporal mixed-effects models in the literature that provide 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];

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." (~380)

o The use of cross-validation is relevant for assessing the predictive ability of models. However, remember that random K-fold, as well as LOOCV, are studied for the crosssectional world. In a spatiotemporal context, they require ad-hoc adjustments that preserve the correlation structure in time and space. In this regard, see the work of Meyer-Pebesma [3-8] on the role of spatial CV and how it is the ideal substitute for random K-fold in the context you study.

Thank you for this comment. We agree with your comment that the cross-validation methods is highly relevant for assessing the predictive ability of models. We are aware of the literature and critiques regarding spatial cross validation, as well as various cross validation methods. However, we considered spatial cross validation methods not suitable for our study. The reason is well explained in two discussions regarding spatial cross validation methods, Wadoux (2021) and Lu (2023). We agree with the arguments in these two papers consider randomly bootstrapped cross validation suitable to the accuracy assessment of our study.

We add this discussion in the revised manuscript.

Added: "Wadoux et al. (2021) argue that standard cross-validation (i.e. ignoring autocorrelation) results in smaller bias than spatial cross-validation.400
Moreover, they state that spatial cross-validation methods should not be used for map assessment as they have no theoretical underpinning, while standard cross-validation is applicable and is sufficient in clustered data scenario's (Wadoux et al., 2021; Lu et al, 2023)"

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);

Added the words "out of sample performances" before "CV".

o Feature selection involves an overwhelming number of alternative techniques: the authors use Shapley values, variables importance, penalty (lasso and ridge), bust subset, etc. Ideally, only one method should be chosen to select relevant covariates so that model results are comparable and not data-dependent. Similarly, it is somewhat problematic to have two or more CV schemes (20-fold and LOO) that prevent proper comparison of the models;

The application of shapely values is used for all global models whereby comparisons between global models are made, not taking into account local models as the approach for the local models is indeed different. This can be attributed to the following. Due to the poor performances by the random forest over all the local station measurements (supplementary, figure 10a, 10b, and 10c), and per spatial group (supplementary, table 2), the best model is not selected by the random forest algorithm and cross-validated Shapley approach. Rather, the best subset regression is used for variable selection.

o When reading, I had the feeling (probably wrong) that there was a misunderstanding of the statistical tools used. For example, box plots do not assess the "variance" but the "variability" of a phenomenon and use the median (not the mean) as a reference point, as it is robust to the presence of outliers (frequent in air quality).

Thank you, we changed the terminology to variability and median

Specific comments and technical corrections:

- Abstract, line 2: “model and predict air pollution over space and time”

No temporal focus

- Section 2.1, page 5, row 143: the authors state that “We used the precipitation from weather stations (National Centers for Environmental Information, 2017)”. Why not directly using Copernicus ECMWF ERA-5 data, which naturally cover the whole Europe with a fine scale (compared to the study area)? Since you used spatial interpolation (ordinary kriging), how do you account for the interpolation uncertainty generated by this approach? How does it reflect on the following stages?

That is a difficult and good question. ECMWF ERA-5 provides a precipitation at 31 km grid resolution. There are also many other products that model precipitation. We firstly need to know if the precipitation significantly affects the air pollution, that is the reason that we chose to use a reliable source of data and simple interpolation method. It is known that spatial interpolation method such as kriging provide reasonable interpolation for precipitation in Europe (E. Lupikasza 2006).

In our study we did not find precipitation to be significant or important. If precipitation contribute significantly to our statistical models, an interesting next-step is indeed comparing different precipitation products.

It is also a very good question regarding considering or incorporating the uncertainty of the predictors in the model. It is difficult to account for the uncertainty of each predictors or data source directly, for example by inputting distributions for each data point into the model. However, a probabilistic model could quantify the uncertainty separately from data and model (Hüllermeier and Wägeman 2021; Kendall and Gal 2017). This however deviate the goal of this study.

E. Lupikasza, Interpolation methods for precipitation fields in Europe, Geophysical Research Abstracts, Vol. 8, 06493, 2006

Hüllermeier, Eyke, and Willem Waegeman. "Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods." ***Machine learning*** **110.3 (2021): 457-506.**

Kendall, Alex, and Yarin Gal. "What uncertainties do we need in bayesian deep learning for computer vision?." Advances in neural information processing systems 30 (2017).

• Section 2.1, formula after row 165: please, explicitly define the symbols/quantities used in the formula. In its current form, it's not easy to understand how the traffic is computed; **Replaced to supplementary; added in main text: "The formula for calculating average hourly traffic can be found in supplementary, equations." (~137).**

• Section 2.2

o Row 171: The average median of what? which values are you considering to rank the features? (later on we discover that you take the median of the rankings… but here it is not clear) **Removed: "The average median of"**

o Rows 175-177: Being the first time you cite LASSO, lightgbm and xgboost models, I suggest using the extended names followed by the acronyms. Also, add some theoretical references on the models (e.g., papers explaining the full methodology); **Full methodology can be better used in supplementary, I assume, as paper (*main text*) is of considerable size already. In main text added: "The equations for the ensemble trees can be found in supplementary, equations." (~149)**

o Row 178: for an extensive comparison in assessing the spatio-temporal prediction accuracy of tree-based methods, linear mixed models and geostatistical mixed models I suggest the papers from the Fassò research group [9-12]; **Again, no focus on the temporal in this paper.**

• Section 2.3

o Actually, the considered models are only described in words but it is difficult to compare it from the analytical perspective. I suggest adding a synoptic table which synthesize the characteristics of the considered models. For instance, the table could state if a model explicitly considers (or not) spatial, temporal or spatio-temporal components (e.g., spatial random effects), if a model is penalized or not (e.g., LASSO), if a model includes covariates or not (e.g., ordinary kriging); **Added table 3 that takes into account model complexity and the consideration of the spatial component**

o Section 2.3.1: can you consider including the recent works on spatial random forests, which extend classical RF to a spatial prediction context [13]? As the aim of the paper is to assess the spatial prediction accuracy of models, this new class could improve a lot your findings;

Interesting read, I integrated some theory of that paper in the discussion: "Moreover, Patelli et al. (2023) identify three main categories in which random forests can be linked to spatial data, being pre-, in-, and post-processing. While random forest performance is linked with spatial groups in our study (which can arguably be linked to a form of post-processing), there is potential in better integrating spatial data in ensemble tree based models such as random forests, to potentially increase predictive performance (Patelli et al., 2023)" (~398). A further integration of the theories discussed (spatial autocorrelation in random forest modeling) in Patelli into this paper, unfortunately is out of scope. Furthermore, a lot of different approaches to linking spatial data with random forests are discussed, however, most if not all approaches[1] are in it's infancy and not widely adopted measures.

[1] RF with SI: Random Forest with Spatial Information; RF with FFS: Random Forest with Forward Feature Selection; RF-RK: Random Forest Residual Kriging; RF-sGs: Random Forest sequential Gaussian simulation; RF-RK with SI: Random Forest Residual Kriging with Spatial Information; RF-RK with SB: Random Forest Residual Kriging with Spatial Bootstrap; RF-GLS-RK: Random Forest based on GLS Residual Kriging.

o Section 2.3.2, row 214 (α): Please, explicitly define the parameter alpha. Also, if alpha refers to the elastic net mixing parameter, with $0 \leq \alpha \leq 1$, then you are considering the elastic net penalization, which is a combination of LASSO and ridge, and not exactly LASSO or ridge;

Thank you for you acknowledging this. The text now specifically mentions that the alpha refers to the separate LASSO and RIDGE models.

- Section 2.4

o Row 224: "N describes a set of n features" is not clear. which is the difference between N and n? What do you mean by payout? Is it the prediction with the N features? Is S the cardinality of the subset of N? **Changed in supplementary, equations**

o Rows 228-229: please, consider rephrasing the whole sentence: the current sentence seems to state that in general/typically Shapley values are embedded in the two alternative CV approaches. However, this seems to be one of your proposals;

We do not mention mean approach anymore to avoid confusion; figure of mean cv is still apparent in supplementary material.

o Rows 239-242: Please, consider rephrasing the whole sentence as currently it is confused. My interpretation of Figure 2 is that a sensible/remarkable prediction accuracy improvement is obtained when considering at least 12 predictors. However, the improvement is marginal considering more than 12 covariates (the curves become flat);

Added:

A remarkable prediction accuracy improvement is obtained when considering at least 12 predictors. However, the improvement is marginal considering more than 12 covariates as the curve flattens. (~192)

o Section 2.4.2 (best subset regression): usually, best subset is used in a linear regression framework. Still, it is not clear to me if you are considering its application in linear or non-linear (in this case, which model?) models. Also, best subset regression is typically affected by computational inefficiency as it requires the computation of 2^k-1 models (where k=30 in the number of covariates). Do you have any insights about the computational burden of this step?

Added: "Rather, the best subset regression is used for variable selection for the local models" (~197)

Computationally unfeasible to continue with a higher k number; moreover global models indicate that the prediction accuracy of a model seems to stagnate way before 30, taking this as an inspiration to set a limit.

o Section 2.4.2 (linear models): still, it is not clear to me which class of mixed-effects models are you considering. Please, can you state (in Appendix or Supplementary Materials) the exact formulae and parameter specifications (i.e., which is the structure of the random effects? Are they i.i.d. sequence of Gaussian RVs or are spatio-temporally structured?). Also, later on you state that the "... linear models (i.e., LASSO and ridge) ...": why not considering a multiple linear regression without penalization? This last model should be directly comparable with penalized approaches.

Equation for mixed effects model can now be found in supplementary, equations; parameter specifications in supplementary, parameters. For the local model, a multiple linear regression without penalization is indeed considered.

- Figure 3 (caption): I would say that the upper and lower whiskers provide information about the overall variability of the estimates rather than "variance". Box-whisker plots are typically computed using the IRQ-rule, that is, the whiskers are +-1.5 x IQR (interquartile range, i.e., X_0.75 - X_0.25). Also, box-plots typically use the median as central value. Is the orange line the median? If so, why do you talk about "mean statistic" in the first paragraph of Section 3.1.1? In air quality statistics there is a huge difference among robust (median) and nonrobust (mean) methods for assessing the centrality of air quality distributions.

Thank you, we changed the terminology to variability and median

- Section 3.1.1

o row 286: what do you mean by "uncommon"?

Changed to "may be present" (~285)

o right before Table 2: Spatial characteristics is a fundamental feature in air quality statistical modeling. Indeed, local air quality is substantially affected by local weather and environmental conditions. However, why did you not include such variable in the features selection stage? You should be sure about the effective predictive capacity of such variables before including it "a priori". Also, I suppose you used the information through a set of dummy variables (I guess 2 vars). Is that correct? Which one did you choose as reference category? Otherwise, did you used separated/independent models by category (i.e., you estimated all the previous models only for Urban and then for low pop and then for far from road)? In the latter case, you should compare the results with the full dataset very carefully as in the submodels you are ignoring a large part of the information contained in the full data;

Local weather and environmental conditions (e.g. wind, temperature, precipitation) are considered in the feature selection stage. I separated the models by spatial group (urban, lowpop, far from road) on which table 2 (now 3) is based. The table shows some NO2 descriptives per spatial group and already unravels some significant differences in NO2 variability.

o row 302: please, clearly state the definition of "more discrete" outcomes or models;

We removed this already, based on 1st reviewer

o rows 307 on: whenever you cite a specific place (e.g., Harleem), please make sure that the area is recognizable on the maps. Where is Harleem in Figures 4 and 5? The same comment holds for all the other cities/locations. **Added a spatial reference in supplementary figure (figure 10d - spatial references Amsterdam area)**

- Figures 4 and 5: they compare different models for different locations. Can you justify this choice? It seems an unfair comparison: to understand the effects of models one should compare different models at the same locations. The comparison you propose is meaningful only if you are sure that, independently on the local conditions, the predictions are comparable (thus there is no spatial effect) and the only relevant factor is the model's definition;

Not entirely true I would say, figure 4 projects the predictions of the global models on the same spatial extent. Figure 5, a comparison between a linear (RIDGE) and non-linear (random forest) is made for every spatial extent, being Bayreuth, Hamburg, Utrecht (in supplementary, again each global model has a prediction for each spatial extent, making comparisons between global models for a spatial extent possible).

Moreover, based on figure 5, you still can say something about the influence of every predictor on the prediction patterns (no2) for every spatial extent.

- Figure 6: still, if you use box-plots, then the central value you are comparing is the median. Also, it is not clear to me what you are representing on the box-plots. Are they the distribution of the estimated NO2 concentrations at every point (if so, how many points did you interpolate?) in a specific area (Figures 4 and 5) or are they temporal predictions at some locations (in this case, which locations?) or are the spatio-temporal predictions? Also, where are the results associated with linear mixed models?

I removed this figure out of the main text so, NA

- 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;

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."**

- Table 4: Where are the machine/statistical learning results (i.e., lightgbm, xgboost, random forest)? What is "linear model" and which is its relationship with ridge and LASSO? Why do you compare a different set of models? As for the mapping, if different models are used to compare local and global modeling, the comparison will be biased and unfair;

Using different models is not worrisome perse, but using different data is hence we divide between global and local and compare the models within the groups. Table 4 only applies to the local models (as it is part of the "local model" section); since the machine/statistical learning results are based on the global data, the machine/statistical learning are irrelevant for table 4.

- Figure 9: are Kerckhoffs's data used to train the models? Why not using the actual NO2 observations (used as response variable of the models) as benchmark? Also, I would plot the original NO2 concentrations used as response variables in the models. They are the actual benchmark. **No, our data originates from the municipality of Amsterdam (open source) while Kerckhoffs et al (2019) use other sources.**

Actions by author(s)

Added:

- Table 1. Descriptive statistics for each relevant variable in the determination of spatial groups for the local- and global datasets
- Added table 2 and 3 about model specifications.
- “The formula for calculating average hourly traffic can be found in supplementary, equations.” (~137).
- “The equations for the ensemble trees can be found in supplementary, equations.” (~149)
- “The key variables highlighted by the random forest model are chosen as predictors in Multiple Linear Regression (MLS). MLS, a statistical method employing multiple explanatory variables to forecast the response variable, operates within a linear framework, where the relationship between predictors and response follows the form Y = Xß. However, linear regression techniques can be characterized by complexity and/or overfitting of the model. In this context, Least Absolute Shrinkage and Selection Operator (LASSO) and RIDGE regression emerge as broader forms of linear regression models, incorporating regularization terms, unlike "pure" linear regression lacking such regularization.” (~151 – 157)
- “The relevant equations can be found in supplementary, equations.” (~172)
- A remarkable prediction accuracy improvement is obtained when considering at least 12 predictors. However, the improvement is marginal considering more than 12 covariates as the curve flattens. (~192)
- Added: “Rather, the best subset regression is used for variable selection for the local models” (~197)
- ““Moreover, Patelli et al. (2023) identify three main categories in which random forests can be linked to spatial data, being pre-, in-, and post-processing. While random
  forest performance is linked with spatial groups in our study (which can arguably be

linked to a form of post-processing), there is potential in better integrating spatial data in ensemble tree based models such as random forests, to potentially increase predictive performance (Patelli et al., 2023)" (~398).

- 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 dimensionality." (~380)
- Added: "Wadoux et al. (2021) argue that standard cross-validation (i.e. ignoring autocorrelation) results in smaller bias than spatial cross-validation.400
  Moreover, they state that spatial cross-validation methods should not be used for map assessment as they have no theoretical underpinning, while standard cross-validation is applicable and is sufficient in clustered data scenario's (Wadoux et al., 2021; Lu et al., 2023)"
- "Complementary, airborne pollutant concentrations are often positively skewed. To adjust for positive skewness, transformations can be applied but also cause prediction changes which currently are not revised in our research. Simultaneously, Lu et al. (2023) examine several techniques such as transformations, likelihood functions, and loss functions to address the issue of non-Gaussion distributions. Thereby, they observed that using a transformation, likelihood function, and loss function that matches with the more-likely distribution (i.e. Gamma) does not improve the modeling results but worsened the prediction errors and the uncertainty quantification (Lu et al, 2023)." (~428-434)
- Supplementary equations – equations relating to several elements that are discussed in the main text: traffic volume; ensemble trees; lasso- and ridge regression; kriging methods; mixed effects model; feature selection
- Supplementary parameters – mainly supportive to methodology section
- Supplementary, spatial reference (e.g. Haarlem)

Key removals

- Equation(s) (1)(2)
- Table 1 (data -descriptives)
- Figure 6 (Distribution predicted NO2 (µg/m3) per model and per location) (now part of supplementary figures)
- Figure 8 (Distribution predicted NO2 (µg/m3) per model (local) with outliers correction. LR = linear regression, LRsp = linear regression accounting for spatial groups, MEM = mixed- effects model, UK = universal kriging, UKsp = universal kriging accounting for spatial groups, OK = ordinary kriging) (now part of supplementary figures)