Public justification (visible to the public if the article is accepted and published):
Dear authors

Thank you for revising the manuscript. Based on the latest reviews, a further revision is necessary. Please take this opportunity to clarify what the statistics in Table 1 summarise and the notation of the predictors in Table 2. In this context, please also provide the resolutions of all data sets used.

Thanks for your suggestions. We added two tables to address this: table 2 and table 3 that show the most important predictors with their descriptive statistics.

Review 1

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.

While we acknowledge that the "global" dataset includes data from only two neighboring countries with similar characteristics, our choice of terminology was intended to distinguish between the larger-scale, cross-border dataset (encompassing cross countries) and the smaller, localized dataset. In this context, "global" is used relatively, to denote a broader scope compared to the "local" dataset.

Changing the terminology to "country" versus "multi-country"/"cross-country" might introduce further ambiguity, as it would still not fully capture the comparative distinction we aim to convey. To address the concern raised, we have clarified the relative nature of these terms in the text to ensure readers understand our intended framing (section 2.1).

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.

NO2 is well-documented as a traffic-related pollutant, with concentrations often strongly influenced by proximity to major roads and traffic density. Therefore, incorporating distance to roads as a feature in the spatial group definition aligns with the pollutant's source characteristics and its spatial distribution patterns. While traditional classifications of urban, suburban, and rural areas often rely solely on population data, the inclusion of road-related variables provides additional relevant distinction in the context of air pollution studies (Chen et al., 2019. Lu et al., 2023).

Chen, J., de Hoogh, K., Gulliver, J., Hoffmann, B., Hertel, O., Ketzel, M., ... & Hoek, G. (2019). A comparison of linear regression, regularization, and machine learning algorithms to develop Europe-wide spatial models of fine particles and nitrogen dioxide. *Environment international*, *130*, 104934.

Lu, M., Cavieres, J., Moraga, P. (2023). A Comparison of Spatial and Nonspatial Methods in Statistical Modeling of NO. *Geographical Analysis*, *55*, 703-727.

3. Line 111-103: Discuss this more, as it seems like these two statements are contradictory.

Thank you, we have rewritten this part, now focusing on how to mitigate bias in the statistical learning methods. The (local) spatial groups are converged in sample size via this threshold adjustment with a slightly higher share of the urban group, aiming to compensate for the relatively high heterogeneity in this group.

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.

We appreciate the reviewer's concern regarding the use of 50,000 estimators in the boosting algorithms. While this number may appear excessive compared to traditional practices, it is important to note that boosting algorithms, such as gradient boosting, are designed to iteratively optimize residuals at each step. Boosting inherently uses the residuals from all observations to build the next tree, and if the gradient no longer descends (i.e., it reaches a minimum), the predictions stabilize, preventing further changes regardless of the number of estimators. This behavior is a fundamental difference from very large individual decision trees, which can overfit by continually splitting into smaller segments using local optimizers. Thus, while increasing the number of estimators significantly may extend computation time, it does not inherently increase the risk of overfitting. Instead, it provides the model with the flexibility to fully converge on the training data (Vezhnevets & Barinova, 2007). See also:

https://datascience.stackexchange.com/questions/11272/is-boosting-resistant-to-overfitting-for-both-number-of-iterations-and-number-of

https://tomatofox.wordpress.com/2021/01/30/why-a-very-large-number-of-trees-wont-overfit-boosting/

Vezhnevets, A., & Barinova, O. (2007). Avoiding boosting overfitting by removing confusing samples. In *Machine Learning: ECML 2007: 18th European Conference on Machine Learning, Warsaw, Poland, September 17-21, 2007. Proceedings 18* (pp. 430-441). Springer Berlin Heidelberg.

5. Line 138, 141: Specific references to the supplementary material is missing. e.g. In which table are these results presented?

We now added "section parameters".

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.

Thank you for this observation. We changed table numbers in supplementary and references in main accordingly.

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.

Thank you for your feedback regarding the need for further elaboration in Sections 2.2.2 and 2.2.3. We agree that more detail on the methodology for the mixed-effects model and kriging would be beneficial to clarify the approach. We don't want the main to extent with too much information to guard simplicity and compactness – therefore we mainly extended only section 2.2.3.

In Section 2.2.2 on Multiple Linear Regression, we utilize regularization techniques such as LASSO and Ridge regression to prevent overfitting. These methods enable the selection of the most important predictors, while optimizing for model performance through tuning.

For Section 2.2.3 concerning the Mixed-Effects Model and Kriging, we have provided a more detailed explanation of the role of each method in our analysis

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.

Added in main. Every fold is visible in the supplementary material (Figure 9).

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.

For tree-based models like RandomForestRegressor, normalization is typically not required for shaply value calculation because they do not rely on distance metrics and coefficients of a linear regression model. This also becomes evident as one of the most important features, identified through the 10-fold repeated random sampling validation, contains very small values (the feature "trop mean filt"), see also table 2 in

main. See also:

Mohammadi, A., Karimzadeh, S., Banimahd, S. A., Ozsarac, V., & Lourenço, P. B. (2023). The potential of region-specific machine-learning-based ground motion models: application to Turkey. Soil Dynamics and Earthquake Engineering, 172, 108008.

Junda, E., Málaga-Chuquitaype, C., & Chawgien, K. (2023). Interpretable machine learning models for the estimation of seismic drifts in CLT buildings. Journal of Building Engineering, 70, 106365.

And following complementary article:

https://shap.readthedocs.io/en/latest/example_notebooks/overviews/An%20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html

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.

The process of feature selection is guided by the out-of-sample performance in a 10-fold repeated random sampling validation, where Shapley values are calculated in each iteration of the random forest models. The median of every feature is used to determine the order from least influential to most influential feature on NO2 predictions. The specific features are now mentioned in section 2.3. Moreover, a table of global (2) and local (3) predictors (with descriptive statistics) are apparent in main.

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.

Thank you for your comment. We acknowledge that the temporal resolution of this benchmark data differs from the coarser temporal scales used in our models. The Kerckhoffs et al. (2019) data represents measurements over specific, limited time periods, while our models address predictions over broader temporal spans. Despite this temporal inconsistency, the detailed spatial granularity of the Kerckhoffs et al. map provides valuable insights and remains an appropriate standard for assessing spatial prediction quality. We added this acknowledgement in the revised version (section 2.4).

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?

Twenty-fold validation was performed by repeatedly splitting the dataset into training (75%) and testing (25%) subsets, using 20 different random states to ensure diverse splits (we now mention this division in section 2.4 and 3.1). Each fold maintained the same test size to ensure consistency, and performance metrics (R^2, MAE, RMSE) were evaluated on the test sets for each model. This approach provided robust estimates of model performance while accounting for variability due to data splitting.

2. A graph of ground truth vs predictions would be beneficial here to identify edge cases at which the algorithms do not perform well.

To provide additional insights, we have added supplementary 26 and 27.

3. Line 216: It's not clear what the spatial resolution of the predictions is nor how these maps were created.

Thank you for this observation, we can indeed clarify the resolution of the prediction area by mentioning it in the methodology (before, it was only apparent in the figure description). The resolution is 100m. TIF files are converted to 100m grid cells for the different regions of Amsterdam, Bayreuth, Hamburg, and Utrecht. Since we have the most influential predictor information (for global models, see table 2 and 4 on which predictors this are; for local models, table 3 and 4) available at 100m for the extent of above mentioned regions, we can use this information to predict the NO2 values for the respective 100m grids, based on the trained local and global models. We added this elaboration in the main (section 2.4).

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.

See supplementary 26 and 27.

5. Line 256: When was leave-one-out cross-validation used before this? Discuss how this was implemented.

Leave-one-out cross validation is opted because of the limited number of observations. Applying the 75/25 train/test methodology is valid for the global dataset due to a larger number of observations but the local dataset contains to few samples to obtain stable testing results. We have added this now in the methodology of main (section 2.4).

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.

That is a good point. This is particularly a limitation for the global dataset. We updated the discussion of the main and acknowledge this by stating that an imbalance between relatively few samples and high heterogeneity cause poor performance. The urban group is more adequately represented in the local dataset

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.

You make a good point that this number of models makes the comparison harder to follow. Therefore, we removed the LightGBM analysis from main and only added this in the supplementary with references to it in main.

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?

The poor performance of these points could be attributed to several factors. One possible reason is out-of-range predictor values; regions with unusual predictor combinations may not be well-represented in the training data, leading to unreliable predictions. Below is a copy of the summary. The count of NO2 values below zero and 85 or above (i.e. classified as outliers) are shown, as well as the number of "not a number"-values. The results per model are visible.

Omission Summary Details

Total number of samples: 10746

Variable-specific omission counts:

Model: random forest (global)

- Not Meeting Criteria (<0 or >=85): 0

- NAs: 0

Model: XGBoost (global)

- Not Meeting Criteria (<0 or >=85): 0

- NAs: 0

Model: LASSO (global)

- Not Meeting Criteria (<0 or >=85): 0

- NAs: 0

Model: Ridge (global)

- Not Meeting Criteria (<0 or >=85): 0

- NAs: 0

Variable: Linear (local)

- Not Meeting Criteria (<0 or >=85): 0

- NAs: 0

Model: Linear separating for spatial groups (local)

- Not Meeting Criteria (<0 or >=85): 30

- NAs: 0

Model: Mixed-effects model (local)

- Not Meeting Criteria (<0 or >=85): 0

- NAs: 0

Model: universal kriging (local)

- Not Meeting Criteria (<0 or >=85): 0

- NAs: 0

Model: universal kriging separating for spatial groups (local)

- Not Meeting Criteria (<0 or >=85): 30

- NAs: 0

Model: ordinary kriging (local)

- Not Meeting Criteria (<0 or >=85): 0

- NAs: 66

Discussion

1. Line 323 “While a well-trained model...” Not clear what this sentence conveys to the reader.

In the discussion of main, we have changed it to: “argue that the growing popularity of global models, due to their ability to capture both linear and non-linear relationships, may lead to misinformation. Although global models can make accurate predictions in regions where the predictor variables are well-represented in the training data, their performance may degrade in areas with predictor values that deviate significantly from the training range, highlighting the risk of spatial bias in predictions.”

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

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.

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.

In main, changed to:

“In the local dataset, the threshold for defining "urban" areas was adjusted from the upper 75% quartile (0.75) to the median (0.5). This adjustment was necessary due to the limited sample size, which required a broader definition to ensure sufficient data coverage for urban areas. However, this change also resulted in a less stringent definition of "urban," potentially including areas with lower population densities. While this adjustment expands the number of training samples available for the most heterogeneous group (urban), it introduces a limitation by diluting the urban group

and affecting the comparability of results. This trade-off underscores the challenges of balancing data representation with statistical robustness in spatial analyses."

Conclusions

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

We have changed this to "investigate" as recommended.

Review 2

Thank you for the comprehensive revisions. This version effectively highlights varying performances across different spatial groups, which may inspire a critical and interesting reconsideration of current validation methodologies in air pollution modeling. I particularly like the methods section, as it provides details on the hyperparameters tuning. I have no major comments, only a few minor suggestions.

- Line 14:
The term "overfitting" is not correct. Please consider replacing it.

We have changed this term. Now the relevant section states: "The spatial prediction patterns of global models show that non-linear methods generally are less sensitive to extreme values compared to linear methods."

- Lines 16-17:
If I understand correctly, this work did not build models specific to spatial groups. Could you clarify the phrase, "but modeling spatial groups does," as its reference is unclear?

We have clarified the phrase, changing it to "Using the local dataset of our study, explicitly accounting for spatial autocorrelation in the universal and ordinary kriging models does not improve accuracy; however, analyzing prediction performance across spatial groups provides valuable insights."

- Line 329:
It is unclear why the statement suggests that nonlinear models outperform linear models in rural areas due to the homogeneous distribution of air pollution levels in these regions. My understanding is that linear models are typically more suitable for fitting homogeneous distributions.

It might be more suitable when there is no enough data, as they sufficiently fit the data and provide clear interpretation. But when data is sufficient and the covariates are expressive to the response, the nonlinear model may provide an equivalent or even better fit if the relationship deviate from linear.

- Lines 331-332:
The reasoning provided here seems unsupported. I do not see how this result substantiates the statement made.

We have removed that part.

- Lines 347-350:
These lines lack informativeness and could be deleted.

Thanks for your observation on this, we removed that part.

- Line 385:
The sentence here appears disconnected from the preceding context. Consider deleting it or adding a transition, like "Moreover."

Removed.