Thank you for the latest revision. The manuscript has progressed well, but there are still a few issues that need to be addressed. In addition to the reviewer's comments, please consider the following:

- The captions of the new Tables 2 and 3 are certainly helpful in understanding Table 1, but please clarify already in Table 1 that the table provides statistics on station characteristics and that the distances in the "Variable" column represent buffer radii.

We have included your suggestion in the table title.

- https://doi.org/10.5281/zenodo.7948161 seems to have been superseded by https://zenodo.org/records/8219003.

There was a misunderstanding about the Git repository snapshot: please include a zip archive of the Git repository (e.g. from "Download ZIP" on GitHub) that represents the version used for the manuscript in the Zenodo repository (and then update the Zenodo link in the Code and Data Availability section accordingly).

Thank you for your clarification. We updated the link to the zenodo archives regarding the main scripts and data <100MB: (https://doi.org/10.5281/zenodo.15193954) and data repository for files >100MB (https://doi.org/10.5281/zenodo.15194548)

Regarding the new paragraph on the Influence of Cross-Validation Techniques:

Page 26, line 396: "90/10 train-test split" should more clearly read "random 90/10 train-test split"

Changed according to your suggestion.

Line 403ff: The term "bootstrapping" often implies sampling with replacement, but apparently here the data were randomly split 90/10, which means sampling without replacement. Therefore, the term Monte Carlo cross-validation used by the reviewer seems more appropriate, and I suggest that the terminology be adjusted accordingly.

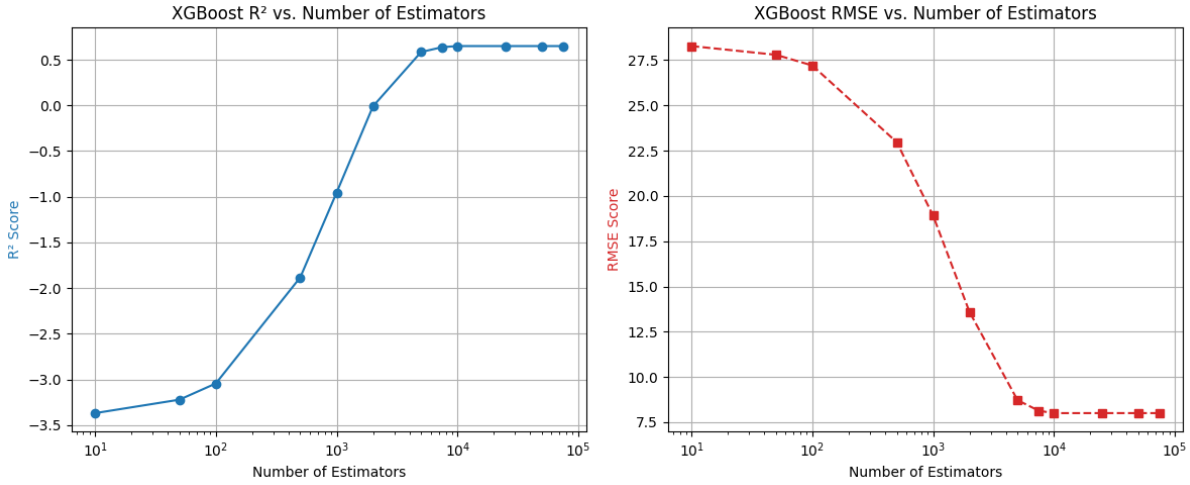
Changed according to your suggestion.

Review 1

I would like to thank the authors for the revised manuscript and for their detailed replies to the comments they received. The revised manuscript addresses most of my previous concerns. There are however a couple of points that I believe were not addressed adequately before publication:

1. While it's true that gradient boosting algorithms optimize residuals iteratively, having too many estimators can lead to overfitting even after apparent convergence. The argument that "predictions stabilize" ignores that small changes in later trees can still accumulate to overfit the training data and it's the reason that XGBoost includes regularization terms, as unlimited (or very large number of) boosting rounds can lead to overfitting. The authors should justify the inclusion of such a high number of estimators and provide evidence (i.e. Validation curves). The reasoning that further trees that necessary don't affect the model is not valid, as later trees can still make small adjustments that collectively overfit. The authors can provide some empirical justification for this choice, like a curve of model performance with models trained on an increasing number of estimators until 50000. This can demonstrate whether 50,000 estimators were indeed necessary for convergence or not. The cited paper (Vezhnevets & Barinova, 2007) doesn't directly address modern gradient boosting implementation best practices.

We have investigated the number of estimators further via a validation curve, as suggested:



With a number of 10,000 estimators, the performance of the xgboost model seems to stagnate. Having 50,000 estimators is therefore unnecessary. In the new version, the xgboost model contains 10,000 estimators.

2. The 20-fold cross-validation methodology followed does not adhere to the standard k-fold cross validation methodology, where the data is divided into k non-overlapping folds, where each data point appears exactly once in the test set. The methodology used in the manuscript adheres more to the Monte-Carlo Cross-Validation method and not k-fold cross-validation. The random split employed can lead to biased performance estimates, especially with small datasets (some points may be used multiple times while others might not be used at all). I recommend to use fewer folds (5-fold cross validation is generally a good balance between bias and variance and make sure each point is used once.

You are right that our explanation can be adjusted to fit the situation better. We abstain from using "fold" and prefer words such as "times" and "iterations" to describe the situation. We changed the naming to Monte Carlo Cross-Validation (CV) and stick to this methodology, rather than using the x-fold divide. The Monte Carlo CV aims at reducing the bias because of the random drawing of training points. We have reduced the number of testing samples for evaluating global model performances, meaning that the random train/test split is now 90/10 instead of 75/25. Moreover, we chose to use Monte Carlo CV because:

- it provides a robust measure of uncertainty for relatively small datasets.

- it works well when data is limited, as it generates multiple datasets through resampling.

In the discussion, we now also discuss the importance of choosing different CV approaches can make a difference in the results.

3. The manuscript would benefit from clearer methodology descriptions, particularly regarding the mixed-effects and kriging models. The authors should also provide more detailed information about data resolution and how predictions were generated at 100m resolution. The limitation of having the most heterogeneous group (urban) being the least represented in terms of data points should be more thoroughly discussed.

Since the previous iteration, we elaborated more on the theories of mixed-effects and kriging models (section 2.2.3). We have expanded the discussion of both the mixed-effects and kriging

models to provide more detailed explanations of the theoretical underpinnings. Specifically, we have clarified the role of fixed and random effects in the mixed-effects models and how they are used to account for spatial variation. In the case of kriging, we have elaborated on both ordinary and universal kriging methods used for local modeling, including how they account for spatial dependencies. At the same time, we explain the conversion of predictor information to a 100m by 100m grid so that we can make predictions via local and global models for the same resolution. Although we did our best to make an objective representation of the land use categories and for comparison between methods, the lack of observations of urban areas may be a cause of the unsatisfactory model performance. Note that the final prediction maps are unaffected by how we separate different land categories and can provide us additional perspectives in model performance, as well as the degree of details that we could expect. Also, we acknowledge the limitations of having the most heterogeneous group (urban) being the least represented in terms of global data points and/or represented by few local data points in the methodology (119-124), results (306-309), and discussion(390-405) sections.

We also agree that we may elaborate further on the points mentioned by you, but at the same time we have to keep our paper compact where possible as the research already has several extensive analyses.