We would like to thank the two reviewers for their time and useful feedback. In this author response, we reply to all comments and suggestions and highlight the changes we plan to implement. Our response is organised in a point-by-point fashion, where our reply follows each comment in italics. We used quotation marks for literal text mentions to our manuscript. Line numbers refer to the preprint with date 24th of January 2024.

Anonymous Referee 1

This manuscript takes random forests as an example to analyze spatial agents such as coordinates and Euclidean distance fields in environmental modeling, which has positive value for spatial analysis based on machine learning models. However, there are some significant shortcomings in the work of this manuscript.

(1) Like other models, random forests require a set of influencing or predictive factors. Therefore, the proxy of environmental factors in spatial analysis models is not a special case of random forest models. Therefore, it is recommended that the author provide additional information on this point.

We agree with the reviewer on the fact that the addition of spatial proxies is not a special case of a random forest, and in our manuscript we have included this information in different parts of the text.

In the introduction (lines 32-36): "The lack of consideration of space in ML models has motivated researchers to find ways to account for spatial autocorrelation to improve model performance. One straightforward approach is to add "spatial proxies" as predictors to the ML model without any modification of the algorithm. We define spatial proxies as a set of spatially-indexed variables with long or infinite autocorrelation ranges that are not causally related to the response variable".

In the methods section (Figure 1 and lines 96-99): "For each set of samples, we extracted the corresponding values of the response and predictors, deleted duplicate observations (i.e. two or more points intersecting with the same cell), and fitted a baseline model, which used predictors according to the corresponding scenario (Table 1). We also fitted coordinates, EDF, and RFsp models (see introduction for details) which included the predictors in the baseline model plus the spatial proxies".

To improve the clarity of this point, we will also include this information in the abstract by modifying the first sentence to "Spatial proxies such as coordinates and Euclidean distance fields are often added as predictors in random forest models **without any modification of the algorithm**".

(2) The use of coordinates and Euclidean distance fields as spatial factor proxies is undoubtedly due to the influence of these spatial factors on the target, or the need to use spatial regions to reflect certain undiscovered factors. This is determined by the specific work, and such a spatial agent is undoubtedly reasonable. Even if the accuracy obtained in some models may not appear to have improved numerically. And this important aspect was not taken into account in this manuscript.

As the reviewer, we think that one of the uses of spatial proxies is to account for residual autocorrelation prompted by the unavailability of relevant yet undiscovered or unmeasured predictors. In fact, we decided to use the word "proxy" precisely for this reason, i.e. a predictor that can be used as a substitute for these unknown or unmeasured variables. The use of proxies to account for unavailable features was also the motivation behind the "partial scenario" in the simulation study, where only a subset of the predictors was available for modelling; as well as in the ”naive models” in the study cases (lines 156-157):

”Our motivation for the naive scenario was to examine whether spatial proxies could help explain residual spatial autocorrelation due to missing predictors and therefore be used in predictor scarcity settings.”

That said, we do think that this link should be made stronger throughout the text so we plan to implement the following changes. First, we will clarify the rationale behind the word ”proxy” in the introduction as explained in this response. Second, we will add a new paragraph in the introduction that will discuss the possible sources of residual autocorrelation, including missing predictors and/or autocorrelated errors. Third, we will motivate the ”partial scenario” and ”naive models” by explicitly mentioning that the goal of these analyses is to examine the ability of coordinate and distance fields to act as proxies of the missing predictors.

(3) It is meaningless to evaluate the superiority or inferiority of a certain agent solely based on the accuracy of the final results, without considering specific issues. In summary, it is recommended to reject the manuscript.

As the reviewer, we also believe that accuracy is not the only way to evaluate the usefulness of a model, although in many cases where knowledge discovery is not a priority it may well be. In our discussion, we acknowledge this limitation (lines 309-314):

”Second, our analysis was based on the adequacy of spatial proxies from a prediction accuracy point of view. When using RF for knowledge discovery, variables with long or infinite autocorrelation ranges such as spatial proxies have been identified to be beyond the prediction horizon (Behrens & Viscarra Rossel 2020, Wadoux et al. 2020, Fourcade et al. 2018) and variable importance statistics in models including them should be interpreted with extreme caution (Meyer et al. 2019).”

However, when the aim is prediction as in our study, we think that accuracy metrics are indeed meaningful to evaluate the relevance of the predictors. For that purpose, we considered several specific factors including different sample distributions, autocorrelation ranges and scenarios to analyse in which cases the proxy predictors are relevant or not.

Dr. Carsten F. Dormann

This study compares different approaches to address spatial autocorrelation in random forest analyses. Using simulated data, and two case studies, the authors assess prediction error and variable importance for differently clustered spatial data.

The study finds that clustering of spatial data had a substantial effect on RMSE, and that different spatial random forest versions differed less than that effect.

There are a few points I do, and a few I do not like about this study. On the plus side, I think the comparison of the RF-approaches is comprehensive and reflects nicely what people have been doing in the past. The evaluation against simulated data is how it should be (see caveat below), and I find the attempts to interpret performance using AOA nice and useful. The case studies illustrate the application case well, and also the problems, particularly using lat/lon as predictor.

My main criticisms are these:

1. The goal of the study does not become clear in the introduction and is confounded throughout the papers. To me, the structure should relate to the three "scenarios" one could have in mind for this study: interpolation, extrapolation and effect estimation (predictive inference). These targets are very different and need to be assessed differently, too. For example, regression kriging is an interpolation method (by design), while identifying importance is effect estimation. Extrapolation (to regions beyond the training data) is explicitly most often the target in the simulations presented here, but not always. I find the results hence sometimes confounding the different issues and hence have trouble interpreting them.

This also relates to the CV strategy. For interpolation error, random CV is fine, for extrapolation it is not. Thus, if the authors find a difference between randomCV and kNNDM-CV, this may or may not be relevant, depending on the goal of the study.

I think this problem pertains particularly to the introduction and the results and will not require much work to address.

We have reviewed the introduction and confirmed that the structure from paragraph 4 onwards can be improved to better characterise the factors that determine the adequacy of spatial proxies, as well as incorporate the ideas pointed out by the reviewer. In the next version of the manuscript, we plan to restructure the second part of the introduction into three different paragraphs according to the following factors relevant for the suitability of spatial proxies:

- Study goal: interpolation vs. extrapolation vs. predictive inference (see major comment 4).
- Residual spatial autocorrelation: strength of the residual spatial autocorrelation, autocorrelation due to missing covariates vs. autocorrelated error (see major comment 2).
- Sampling pattern: regular vs. random vs. clustered samples.

These factors will also be listed in the study objectives at the end of the introduction so that they can be easily identified by the reader. In addition to these, we will also devote a separate paragraph on validation methods to empirically determine whether adding proxies is recommended (i.e. model selection), which will include probability test samples (see major comment 4) as well as random k-fold and kNNDM CV. Regarding the results section, we will also structure the simulation findings into three different blocks according to the three factors listed above.

2. Spatial autocorrelation is entirely related to environmental variables, when in the real world it is also related to mass effects (dispersal, diffusion, contagion): the error is spatially autocorrelated, too. In my understanding, the authors did not address that. That is problematic, as Table 1's second row is thus a model WITHOUT spatial autocorrelation in the residuals, as all predictors are present to correct for it. This is, from a statistical point of view, a no-problem data set. While that does in no way invalidate the simulations, it must be clearly communicated that this is NOT a situation one would even consider using spatial representation for: no residual SA, no problem.

Spatial error in the residuals has been simulated in previous such studies (since Dormann et al. 2007 Ecography), and is a bit annoying to fine-tune; it can be done, though, and I think it should be done (see also simData here: https://github.com/biometry/FReibier/tree/master).

On the back of such simulation, a GLS fitted to the correct model would give the best possible reference analysis; anything better than that would be a biased assessment of error.

As the reviewer pointed out, we did not consider the case of autocorrelated errors in our simulation study, where all the spatial structure in the outcome generation was introduced via covariates. To address this point, we propose to add an additional scenario ”autocorrelated error” where the response generation equation includes an autocorrelated error term, for which the RF-GLS approach will probably also result to be the best approach.

Regarding the ”complete” scenario (Table 1 second row) where indeed no spatial representation is needed, we discussed how spatial proxies are unnecessary in this case (lines 265-270):

”However, in the complete scenario where no residual autocorrelation was expected, we hypothesise that the similar or sometimes worse performance is due to adding an irrelevant set of predictors that are noise to the model. Unlike RK, where spatial autocorrelation is modelled in the residuals and in its absence would result in a pure nugget effect, i.e. a flat variogram leading to an ordinary least squares estimation (Hengl 2007), in a ML model the irrelevant proxies are still included in the trend model. Even though RF is fairly robust to the addition of irrelevant predictors (Kuhn & Johnson 2019), a decrease in performance was sometimes observed.”

In the revised manuscript, we will also mention that no space representation is needed in this case in the methods section when the scenarios are introduced.

3. Minimum RMSE according to simulations should be indicated in Fig. 3. Since the authors use the standard normal as error distribution, the best possible RMSE should also be 1. On average, it is closer to 2, but for the "complete, range 40" it looks as if it was below 1. That would be, well, surprising and important for interpretation: a fit into the spatial noise.

We agree that considering the minimum possible RMSE would help to interpret results. In the revised version of the manuscript, we will add it to figure 3 by adding a dashed line. We checked our results and the minimum RMSE across all simulations is above 1 (1.04), which will become clear once we add the minimum possible RMSE horizontal line.

4. While I read about and like the kNNDM, I still prefer a truly independent test data set. Since the authors have invented the data, they could simply extend the area by doubling it to one side, and use that second half for validation in the sense of a true extrapolation. My feeling is that kNNDM will work well if the range of data is much larger than the spatial autocorrelation, but not if SA is *large relative to the spatial extent. An extent of 100x100 is "only" 2.5 times larger than the range of 40. Thus, the sampled data points will fall within the SA-range virtually always (Fig. 1.2). I am not convinced that this is an independent-enough test case.*

We agree with the reviewer on that a probability test sample is always preferable to any type of cross-validation, which is something we discussed in our other works (Milà et al 2022, Linnenbrink et al 2023) but not here. In the future version of the manuscript, we will include this information in the introduction. Furthermore, we will include random probability test samples as an additional validation method in our simulations (together with the already-included random k-fold and kNNDM CV).

Regarding extrapolation, we agree on the fact that the way we designed our simulation study does not allow to isolate extrapolation effects and explore their impact on the CV results. While in some cases we will indeed have some extrapolation when predicting (e.g. with very clustered samples), there will always be interpolation as well. Following the reviewer's suggestion, we propose to modify the dimensions of the simulation area following a similar approach to our previous work (Milà et al 2022). Briefly, we will simulate in a 300x100 grid (rather than a 100x100 grid like we are doing at the moment) where samples will be taken in the 100x100 left window while the 100x100 right window will be used to evaluate extrapolation.

5. More as a suggestion: The problem of using spatial coordinates or proxies is that they replace the causal predictors in a random forest due to collinearity. As a consequence, the importances are wrongly estimated. One can, for the simulated X, compute how well each predictor can be represented by the specific spatial predictors used. If X1 can be predicted by lat/lon or the EDFs or distances with an r of 0.8 (or so), then clearly they will compete for explanation and substantially bias importance estimates. That is the reason why the ME-approach in spdep adds the PCNMs only to explain the RESIDUALs of the model, after fitting the non-space-variables X1-X6.

So, I would be interested in seeing how well space can replace actual predictors. Either by reporting such "predictability of predictors by space" in the appendix, or by having another model entirely without X1-X6 in the comparison.

This would also tie in nicely with the difference between inter/extrapolation: space-only should work fine for inter-, but fail for extrapolation.

Following the reviewer's recommendation, we propose to add an additional "only proxy" scenario to those listed in Table 1 where only spatial proxies will be used to predict the outcome and for which extrapolation performance should be low. Beyond this, however, we do not plan to further investigate the impact of spatial proxies on predictive inference in this specific study, as this would require a different simulation study design targeting the ability of the models to estimate the true variable importance (lines 309-314):

"Second, our analysis was based on the adequacy of spatial proxies from a prediction accuracy point of view. When using RF for knowledge discovery, variables with long or infinite autocorrelation ranges such as spatial proxies have been identified to be beyond the prediction horizon (Behrens & Viscarra Rossel 2020, Wadoux et al. 2020, Fourcade et al. 2018) and variable importance statistics in models including them should be interpreted with extreme caution (Meyer et al. 2019)."

Overall, I think this is a nice paper almost as it is and with a little bit more integration of WHY *we would expect which approach to work better and a clearer structuring of the purposes of the analysis it will be just fine. IMHO it could be a greater paper, if the authors would allow for spatial autocorrelation in the error term and try to get at the bottom of WHEN space affects inference (i.e. here: importance) of predictors (that is, investigate the effect of collinearity on the intrapolation, extrapolation, variable importance).*

Thank you for the acknowledgement. As a wrap-up of the major comments, we list again the action points to address the reviewer's major comments:

- Change the structure of the introduction and results sections based on these three factors: study goal, origin and strength of the spatial autocorrelation, and sampling pattern.
- Add probability test samples as a validation method.
- Extend the area of the simulation study to 300x100 to evaluate extrapolation.
- Add the simulation "autocorrelated error" scenario to consider an autocorrelated error term.
- Add simulation "only proxy" scenario where only spatial proxy predictors are used in the models.

Minor points: L56: Also cite other people's work here, much earlier, e.g. Le Rest et al. 2014 and whatever else we cited in Roberts et al. (2017 Ecography) on that topic.

Since cross-validation is not the main topic of the manuscript, we kept a short reference list and did not include all references related to the topic. Nonetheless, we agree that giving more background could be useful to the non-familiar reader. We will extend the list of references for cross-validation with the works mentioned by the reviewer as well as the earlier literature regarding block-based and buffer-based spatial cross-validation.

L57: I find the restriction to RF too narrow. This is a logical and fundamental problem, not one specific to RF. Ploton et al. (2020) showed it for random forest, Kattenborn et al. (2022) for CNNs. It is the same problem of extrapolation in space with poor design for the CV.

Our manuscript focused on RF because it is 1) the algorithm with which spatial proxies have been mostly used, and 2) one of the most widely used models in the geosciences (lines 28-31):

"One of the most popular ML algorithms in the geospatial community is Random Forest (RF), a decision tree ensemble (Breiman 2001) that has shown good performance across many applications (Wylie et al. 2019) and centred the attention of many methodological studies (e.g. Meyer & Pebesma (2021), Hengl et al. (2018), Sekulić et al. (2020), Georganos et al. (2021), Saha et al. (2023))."

In our discussion, we acknowledged the fact that although our analysis was restricted to RF, it will possibly also apply to other algorithms as well (lines 307-309):

"First, we focused on RF regression and, while we think that our results are likely to extend to other ML algorithms, the extrapolation behaviour and sensitivity to irrelevant predictors differs by algorithm and might limit the ability to generalize our results."

That said, we agree that our work would benefit from more detail regarding the generalization of the results to other algorithms. To address this, we will include the references for other algorithms the reviewer suggested.

L73: "scenarios" are what I called "target", "goal" or "purpose": Make clear what the goals are in the intro!

Please see response to major comment 1 to check the new structure we propose to clarify this point.

L178: Why would anybody use randomForest and not ranger? Much faster and hence less energy consumption.

We used the package randomForest rather than ranger because, when running our simulations in an HPC environment, we faced threading issues when using the ranger package. While using randomForest does indeed affect energy consumption, as a different random forest implementation will not impact results. In the next round of simulations we will try to solve these issues and use ranger instead.

What is the point of Fig. 7? I can see neither RMSEs or biases or anything, so why look at these maps? Also, we are typically more impressed by high-resolution maps, even if they are completely wrong; map visualisation is thus either uninformative and misleading in many cases.

What is the point of Fig. 8, apart from the funny lines in "A Coordinates"?

We think that Figures 7 (temperature) and 8 (air pollution) are relevant since they allow us to assess 1) the differences in predicted surfaces between baseline and proxy models, and 2) whether naive models with spatial proxies can approximate the spatial patterns of complete baseline models. While in Figure 7 we showed that the inclusion of proxies in naive models resulted in a predicted surface that approximated the spatial patterns of complete models successfully, we find that this is not the case for the air pollution case study where they resulted in artifacts (Figure 8). In the revised version of the manuscript, we will further highlight these points in the text.

Table 2 and 3: Where are the standard errors on these estimates? (Yes, I understood that some of them are a bit a pain to compute for one of the models. Still, without an estimate of the error, how can the reader interpret a value of "0.92" vs "0.87"? Might well be the same value if SD=0.4.)

We agree with the reviewer on the importance of showing the standard deviation of the statistics to be able to ascertain whether the differences are relevant. In the future version of the manuscript, we will compute these for random k-fold cross-validation. For kNNDM, however, it will not be possible as statistics are computed, similarly to a LOO CV, by stacking all out-of-sample predicted and observed values (see lines 109-111 of the manuscript) and computing a single statistic. This is due to potential unequal sample sizes across folds, as well as how the folds are created according to the distribution of all samples simultaneously (Linnenbrink et al. 2023).

I missed the discussion of some existing approaches to manage space into ML.

Hajjem, A., Bellavance, F., & Larocque, D. (2011). Mixed effects regression trees for clustered data. Statistics & Probability Letters, 81(4), 451–459. https://doi.org/10.1016/j.spl.2010.12.003

Hajjem, A., Bellavance, F., & Larocque, D. (2014). Mixed-effects random forest for clustered data. Journal of Statistical Computation and Simulation, 84(6), 1313–1328. https://doi.org/

10.1080/00949655.2012.741599

Li, L., Girguis, M., Lurmann, F., Wu, J., Urman, R., Rappaport, E., Ritz, B., Franklin, M., Breton, C., Gilliland, F., & Habre, R. (2019). Cluster-based bagging of constrained mixed-effects models for high spatiotemporal resolution nitrogen oxides prediction over large regions. Environment International, 128, 310–323. https://doi.org/10.1016/j.envint.2019.04.057

Li, L., Lurmann, F., Habre, R., Urman, R., Rappaport, E., Ritz, B., Chen, J.-C., Gilliland, F. D., & Wu, J. (2017). Constrained mixed-effect models with ensemble learning for prediction of nitrogen oxides concentrations at high spatiotemporal resolution. Environmental Science & Technology, 51(17), 9920–9929. https://doi.org/10.1021/acs.est.7b01864

Zhan, Y., Luo, Y., Deng, X., Chen, H., Grieneisen, M. L., Shen, X., Zhu, L., & Zhang, M. (2017). Spatiotemporal prediction of continuous daily PM2.5 concentrations across China using a spatially explicit machine learning algorithm. Atmospheric Environment, 155, 129–139. https://doi.org/10.1016/j.atmosenv.2017.02.023

Kattenborn, T., Schiefer, F., Frey, J., Feilhauer, H., Mahecha, M. D., & Dormann, C. F. (2022). Spatially autocorrelated training and validation samples inflate performance assessment of convolutional neural networks. ISPRS Open Journal of Photogrammetry and Remote Sensing, 5, 100018. https://doi.org/10.1016/j.ophoto.2022.100018

Le Rest, K., Pinaud, D., Monestiez, P., Chadoeuf, J., & Bretagnolle, V. (2014). Spatial leave-one-out cross-validation for variable selection in the presence of spatial autocorrelation. Global Ecology and Biogeography, 23, 811–820. https://doi.org/10.1111/geb.12161

Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., Rossi, V., Dormann, C., Cornu, G., Viennois, G., Bayol, N., Lyapustin, A., Gourlet-Fleury, S., & Pélissier, R. (2020). Spatial validation reveals poor predictive performance of large-scale ecological mapping models. Nature Communications, 11(1), Article 1. https://doi.org/10.1038/s41467-020-18321-y

Thank you very much for providing these references; we will include them in the relevant paragraphs of the text.

References

Behrens, T. & Viscarra Rossel, R. A. (2020), 'On the interpretability of predictors in spatial data science: The information horizon', *Scientific Reports* **10**(1), 16737.

Breiman, L. (2001), 'Random forests', *Machine learning* **45**(1), 5–32.

Fourcade, Y., Besnard, A. G. & Secondi, J. (2018), 'Paintings predict the distribution of species, or the challenge of selecting environmental predictors and evaluation statistics', *Global Ecology and Biogeography* **27**(2), 245–256.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/geb.12684*

Georganos, S., Grippa, T., Gadiaga, A. N., Linard, C., Lennert, M., Vanhuysse, S., Mboga, N., Wolff, E. & Kalogirou, S. (2021), 'Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling', *Geocarto International* **36**(2), 121–136.
URL: *https://doi.org/10.1080/10106049.2019.1595177*

Hengl, T. (2007), 'A practical guide to geostatistical mapping of environmental variables.', *Office for Official Publications of the European Communities* .
URL: *https://publications.jrc.ec.europa.eu/repository/handle/JRC38153*

Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B. & Gräler, B. (2018), 'Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables', *PeerJ* **6**, e5518.

Kuhn, M. & Johnson, K. (2019), *Feature engineering and selection: A practical approach for predictive models*, Chapman and Hall/CRC.

Linnenbrink, J., Milà, C., Ludwig, M. & Meyer, H. (2023), 'knndm: k-fold nearest neighbour distance matching cross-validation for map accuracy estimation', *EGUsphere* **2023**, 1–16.
URL: *https://egusphere.copernicus.org/preprints/2023/egusphere-2023-1308/*

Meyer, H. & Pebesma, E. (2021), 'Predicting into unknown space? estimating the area of applicability of spatial prediction models', *Methods in Ecology and Evolution* **12**(9), 1620–1633.

Meyer, H., Reudenbach, C., Wöllauer, S. & Nauss, T. (2019), 'Importance of spatial predictor variable selection in machine learning applications – moving from data reproduction to spatial prediction', *Ecological Modelling* **411**, 108815.
URL: *https://www.sciencedirect.com/science/article/pii/S0304380019303230*

Saha, A., Basu, S. & Datta, A. (2023), 'Random forests for spatially dependent data', *Journal of the American Statistical Association* **118**(541), 665–683.
URL: *https://doi.org/10.1080/01621459.2021.1950003*

Sekulić, A., Kilibarda, M., Heuvelink, G. B., Nikolić, M. & Bajat, B. (2020), 'Random forest spatial interpolation', *Remote Sensing* **12**(10).
URL: *https://www.mdpi.com/2072-4292/12/10/1687*

Wadoux, A. M. J.-C., Samuel-Rosa, A., Poggio, L. & Mulder, V. L. (2020), 'A note on knowledge discovery and machine learning in digital soil mapping', *European Journal of Soil Science* **71**(2), 133–136.
URL: *https://bsssjournals.onlinelibrary.wiley.com/doi/abs/10.1111/ejss.12909*

Wylie, B. K., Pastick, N. J., Picotte, J. J. & Deering, C. A. (2019), 'Geospatial data mining for digital raster mapping', *GIScience & Remote Sensing* **56**(3), 406–429.
URL: *https://www.tandfonline.com/doi/abs/10.1080/15481603.2018.1517445*