We thank the reviewers for the time taken to review our work and for their valuable comments and suggestions. In these final author comments, we will address the reviewer and community comments and point out the changes we have implemented in the revised version of our manuscript. Our responses are organised in a point-by-point fashion following each comment in italics.

## Reviewer 1

*The manuscript presents a much-needed methodology for cross-validation of spatial data. In my opinion, the strongest point is the use of the W statistic to identify the best CV split. However, there are a few points which I feel should be addressed in the discussion.*

*The proposed methodology using clustering algorithms seems valid, but how can we know if it provides the best possible result? An algorithm that optimizes the W statistic directly as a function of the CV fold indices would be more desirable, instead of relying on the clustering algorithm´s internal metric as a proxy. As a suggestion for future work, I recommend using a genetic algorithm to assign CV indices to the data points directly.*

We would like to thank professor Gonçalves for his suggestion. Fold assignment via clustering is an intuitive and straightforward approach, however, we agree that using an advanced optimization procedure such as a genetic algorithm might be worth testing. In the revised version of the manuscript we discuss this point as future work, which we will try to take onboard in future versions of the algorithm.

*The W statistic explained 60% of the variability in map accuracy, but would this be consistent across different datasets? At least one more case study would be needed to verify this.*

We agree that a second simulation is helpful to assess if this relationship holds in a different setting. Therefore, we include an additional simulation using the Above-Ground Biomass example described in de Bruin et al. (2022) as supplementary material in the revised version of the manuscript.

## Reviewer 2

*The study proposes a novel cross-validation method for spatial data that aims to deliver more representative measurements of spatial map accuracy than commonly-used methods. This is a relevant concern for GMD readers with the rise in use of machine learning methods for geoscientific modelling. Issues with model evaluation in the spatial setting have been identified in a number of recent studies. The paper is well-written and contributes a practical solution for a common issue.*

*In my opinion, the most exciting/innovative idea in this work is the concept of defining the evaluation method based on the desired data for which he model is intended to return predictions. This would require researchers to more carefully define the purpose of their models before and during the model creation process, which should be common practice. In reality, this is often not done, or done in a 'standard' way which doesn't accurately reflect the intended use of the model.*

*The method presented in this paper is a very practical solution to this, where the desired target dataset is an input of the evaluation algorithm and therefore researchers are required to clearly consider and define it. I think this is a significant contribution to model development methodology and should be more clearly emphasised in the manuscript. The possibilities, benefits and disadvantages of this concept could also be discussed - for example, when models are used in production, the prediction area is a moving target; would that require continual re-evaluation?*

We thank reviewer 2 for their comments. As discussed in our previous work (Meyer & Pebesma 2022, Milà et al. 2022), we agree with the reviewer on that defining the objective of the prediction is a key step to define an appropriate map accuracy estimation method. In the kNNDM manuscript, we explain this idea in the third paragraph of the introduction and, in the revised version of the manuscript, we also mention it in the concluding paragraph of the discussion.

If the prediction area differs when the model is used for production, the CV estimate might not be a suitable proxy for map accuracy anymore. This would require other testing strategies. Re-evaluation using kNNDM might be one option. We addded a sentence in the discussion section on that issue.

*The paper suggests that kNNDM is, essentially, a computationally-cheaper alternative to the previously-published method by the authors, NNDM LOO. In the article, the only limitation of leave-one-out CV methods described is that of computational time. However, to my knowledge, even if computation is not considered, LOO CV methods may not be the optimum method due to higher variation in the resulting models (due to the bias-variance tradeoff). Could this explain why kNNDM 10-fold seems to perform better in the case of strong clusters (Figure 5)? For me, this would be more convincing than the computation speedup comparison, which is relatively trivial given that LOO CV is the most extreme version of k-fold CV.*

We would like to thank the reviewer for pointing out the bias-variance trade-off in LOO vs. k-fold CV. While in the strong clustered simulations NNDM k-fold CV provides slightly more accurate RMSE (mean (SD) -0.008 (0.043) for kNNDM vs -0.015 (0.038) for NNDM LOO) and MAE estimates (mean (SD) -0.003 (0.031) for kNNDM vs -0.013 (0.026) for NNDM LOO), the difference between the two methods is small and cannot be detected for the $R^2$. Moreover, dispersion estimates for NNDM LOO CV were generally smaller as indicated in the parentheses above. We think that although the two algorithms will tend to provide similar results, differences are expected because the way to match the distributions is different, i.e. a buffering approach is used for the LOO CV while clustering is used for kNNDM. In addition to these, as the reviewer points out, the size of the hold-out data may have an impact on the results as well, with some studies suggesting the aforementioned bias-variance trade-off (Kohavi et al. 1995, Hastie et al. 2009) while others argue this will actually depend on the modelling algorithm and its stability (Zhang & Yang 2015).

In the revised version of the manuscript, we now discuss how LOO NNDM CV and kNNDM can result in different estimates due to the different methodologies as well as the size of the holdout data in the first paragraph of the discussion.

*Following on from this, it seems likely that the value of k would impact the results. Use of 10 folds is very common; is there theoretical justification for this? It would be useful to see some comparisons of the results with multiple values of k.*

We chose 10 folds since, as the reviewer points out, it is a frequent choice amongst ML practitioners. However, we agree that $k$ might influence results as it has important implications for Nearest Neighbour Distances (NNDs), e.g. lower numbers of $k$ can better address severe clustering in the data (at the cost of resulting in potentially more biased performance estimates). To address this, in our updated manuscript we have added a new experiment where we investigate the influence of different numbers of $k$ on the performance and quality of the match of kNNDM.

*In Figure 1, it is shown that the W statistic will also be larger if training points are regularly distributed, as well as when clustered. Does this mean that the null hypothesis might be rejected for regularly distributed datapoints? Does this explain why NNDM LOO performed better for regularly* *distributed data (Figure 5)?*

Regarding regularly-distributed samples in Figure 1, the $W$ statistic between $\hat{G}_{ij}(r)$ and $\hat{G}_j(r)$ for regular samples will indeed be larger since distances between training points will be longer than during prediction (i.e. $\hat{G}_j(r) \leq \hat{G}_{ij}(r)$). This, however, does not impact our results, since the Kolmogorov-Schirnov test we perform as a first step of the algorithm is one-sided and will only perform clustering if the null hypothesis $H_0 : \hat{G}_j(r) \leq \hat{G}_{ij}(r)$ is rejected in favour of the alternative hypothesis $H_1 : \hat{G}_j(r) > \hat{G}_{ij}(r)$, which will not occur for regular samples. As a result, the algorithm will return a random k-fold CV instead.

We think that the worse performance of NNDM k-fold CV compared to NNDM LOO CV for regular samples in Figure 5 is due to the fact that in absence of clustering, NNDM LOO CV generalises to LOO CV while NNDM k-fold CV generalises to random k-fold CV. In a random k-fold CV, neighbouring points can still be in the same fold due to a random chance that will increase with smaller number of folds $k$, thus causing slightly longer NND during CV. In contrast, in a NNDM LOO CV all points except the one being validated will be included in the model leading to the smallest possible W.

*Minor comment: I assume the hyperparameters of the models are not tuned as it is not mentioned, but this could be stated explicitly.*

We confirm we did not tune the model hyperparameters in order to save computation time in the already computationally-demanding simulations, given that our objective was not to optimize the performance of the models. We will add the respective information in the updated version of the manuscript.

*Finally, I would recommend testing the method on at least one additional dataset, as the results presumably depend on the spatial autocorrelation present in the dataset used.*

We agree with the reviewer that testing the model on a different dataset is needed. We added a second simulation in the revised version of the manuscript where we test the kNNDM method using the Above-Ground Biomass example presented in de Bruin et al. (2022).

## Community comments

*Just a quick hint that https://doi.org/10.1016/j.jag.2023.103364 was just published - may or may not be relevant for your discussion.*

We thank Dr. Nils Tjaden for providing the reference, which we have included in the revised version of the manuscript.

# References

de Bruin, S., Brus, D. J., Heuvelink, G. B., van Ebbenhorst Tengbergen, T. & Wadoux, A. M.-C. (2022), 'Dealing with clustered samples for assessing map accuracy by cross-validation', *Ecological Informatics* **69**, 101665.
**URL:** *https://linkinghub.elsevier.com/retrieve/pii/S1574954122001145*

Hastie, T., Tibshirani, S. & Friedman, H. (2009), *The Elements of Statistical Learning. Data Mining, Inference, and Prediction.*, Springer Science  Business Media.

Kohavi, R. et al. (1995), A study of cross-validation and bootstrap for accuracy estimation and model selection, *in* 'Ijcai', Vol. 14, Montreal, Canada, pp. 1137–1145.

Meyer, H. & Pebesma, E. (2022), 'Machine learning-based global maps of ecological variables and the challenge of assessing them', *Nature Communications* **13**(1), 2208.

Milà, C., Mateu, J., Pebesma, E. & Meyer, H. (2022), 'Nearest neighbour distance matching leave-one-out cross-validation for map validation', *Methods in Ecology and Evolution* **13**(6), 1304–1316.
**URL:** *https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13851*

Zhang, Y. & Yang, Y. (2015), 'Cross-validation for selecting a model selection procedure', *Journal of Econometrics* **187**(1), 95–112.