Dear Reviewer,

We sincerely thank you for your detailed and constructive feedback. We have carefully considered each of your comments and will revise the manuscript accordingly. We are pleased that you recognize the value of our work and are confident that the revisions made in response to your suggestions will further strengthen the manuscript. Please find our detailed responses to each of your points below.

Reviewer comments	**Response**
1. Introduction: The introduction is well written and effectively prepares the reader for the paper. However, the authors largely restrict their literature review to soil erosion modelling. While this is understandable to a certain degree, the claimed novelty of the paper lies in applying "new" methods such as CNNs and multi-layer neural networks. These models, however, are not particularly novel in this context, as CNNs have been applied to soil prediction tasks at least since 2019 (e.g., Padarian et al., 2019). The study would offer stronger novelty by considering more recently proposed methods from the broader ML literature (for instance, the high-quality TabArena benchmark by Erickson et al., 2025, which compares state-of-the-art tabular learners). Several of these modern methods have already been successfully tested in soil science, and established approaches such as CatBoost have been available for even longer. I understand that it is not feasible to cover every recent method, but the current comparison does feel somewhat outdated for a paper that aims to emphasize on machine learning aspects. Padarian, J., Minasny, B., & McBratney, A. B. (2019). Using deep learning for digital soil mapping. Soil, 5(1), 79-89. Erickson, N., Purucker, L., Tschalzev, A., Holzmüller, D., Desai, P. M., Salinas, D., & Hutter, F. (2025). Tabarena: A living benchmark for machine learning on tabular data. arXiv preprint arXiv:2506.16791.	We thank the reviewer for the positive feedback on the introduction and for the valuable comments. We agree that CNNs have already been applied in soil science for various purposes, particularly for modelling soil properties such as soil organic carbon. However, the cited studies and similar works do not apply complex neural network architectures, such as CNNs, for quantifying continuous soil erosion rates. Erosion rates are not a soil property but a function of various natural and management factors including soil properties, erosive rainfall, topography, management, etc. The novelty of our study therefore lies in the application of complex neural networks to model patterns of continuous soil erosion rates at the field-to-landscape scale. The focus of this study is to explore and compare neural networks with a benchmark method (Random Forest) in this context. To our knowledge, no previous study has done this and used CNNs to predict continuous soil erosion rates at this spatial scale. Nonetheless, we acknowledge that additional, recently proposed machine-learning could also provide valuable insights and should be considered for future research and we will add these points to the discussion.
33: The use of the term AI does not seem appropriate in this context and comes across more as a buzzword. Since the paper exclusively discusses machine learning methods (e.g., L. 67), I suggest using machine learning consistently instead of AI.	Thank you for bringing this to our attention. We agree that consistent terminology is important and will rephrase L33 accordingly.

2. Methodology: I have several concerns about the hyperparameters and the validation used in this study. Other comments are of minor nature: Hyperparameters: It remains unclear how the authors tuned their models. From the description (L. 178–179), it appears that hyperparameters were adjusted directly on the validation folds of the 5-fold CV. This approach introduces data leakage, as the same data are effectively used both for model selection and for performance estimation, which reduces the penalty for overfitting. Proper hyperparameter optimisation requires a nested cross-validation scheme, where the data are split into three parts: a training set for fitting the model, an inner validation set for selecting hyperparameters, and an outer test set (or fold) for obtaining a performance estimate. I looked into the provided code but could not find any script related to hyperparameter tuning. Instead, in the models script I found only fixed parameter settings. This is problematic, as optimal hyperparameters should be determined separately for each training fold within the cross-validation. Without such a procedure, the reported results may not reflect the best achievable model performance and risk being biased by arbitrary parameter choices. Lastly, the search space for the hyperparameters was not given. This is extremely important for a fair model comparison, if a poorly tuned RF is compared to a well-tuned NN, the comparison would not be fair. There is a lot of studies on how this can induce bias in benchmarking (e.g., Nießl et al. 2022). Nießl, C., Herrmann, M., Wiedemann, C., Casalicchio, G., & Boulesteix, A. L. (2022). Over-optimism in benchmark studies and the multiplicity of design and analysis options when interpreting their results. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 12(2), e1441.	We apologize for our lack of sufficient documentation regarding the hyperparameter tuning. The tuning of hyperparameters was conducted separately from the main training and validation procedure, using a grid search performed prior to the main cross-validation runs (Raschka, 2020; Yu and Zhu, 2020). We agree that it should be described more thoroughly to avoid confusion, and we will revise the description in the manuscript accordingly. In addition, we will include the hyperparameter tuning scripts and the respective search space in the referenced repository. Raschka S. (2020): Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning, https://arxiv.org/abs/1811.12808 Yu, T. and Zhu, H. (2020): Hyper-parameter optimization: A review of algorithms and applications, arXiv preprint arXiv:2003.05689, https://doi.org/10.48550/arXiv.2003.05689
Do I understand correctly that this figure shows the "ground-truth" soil erosion dataset, and that these data are available in raster format, i.e., the true (or approximate true) erosion values are known across the	The raster data displayed in Fig. 1 represent spatially continuous mapped erosion patterns rather than single point measurements based on erosion pins (as in Gholami et al., 2021). The dataset is based on empirical long-term

entire study area? If so, I find this somewhat questionable, since such complete "ground truth" presumably relies on interpolation or modelling itself, and may therefore not represent true independent measurements. More importantly, it is unclear why additional modelling is applied, given that each cross-validation repetition already uses 80% of the study area for training. In digital soil mapping, modelling is typically motivated by sparse point observations, where the objective is to generate high-resolution maps from limited data. In contrast, this study seems to assume ground-truth values for every raster cell, a setup that almost inevitably leads to overly optimistic performance estimates with poor generalization value. Would a strategy such as "leave-one-validation-site-out" not provide a more realistic evaluation of model performance? I may be missing a domain-specific aspect of soil erosion mapping, but from a classical digital soil mapping perspective this design appears problematic. For example in Gholami et al. (2021), which is also cited in this paper, they used some point data and they have specified validation points. I am missing something like this in this study. To me, this makes much more sense but I do not see this in Fig. 1. Gholami, V., Sahour, H., & Amri, M. A. H. (2021). Soil erosion modeling using erosion pins and artificial neural networks. Catena, 196, 104902.	soil erosion monitoring data obtained in surveys, which were subsequently aggregated to a raster format to enable spatial analysis (see Steinhoff-Knopp & Burkhard, 2018). It is not directly derived from interpolation or modelling, based on single points. The aim of our study was to assess how well different machine-learning models can reproduce these observed erosion patterns and loss rates at the field-to-landscape scale and detect underlying relationships. We thank the reviewer for pointing out the potential value of a "leave-one-area-out" validation approach. In fact, we applied this approach during our study, and the results also show that the CNN achieves the best predictive performance among the tested models. However, this approach also has its own limitations given the available data and was not the primary focus of our analysis. Nevertheless, we agree that it adds further validity to our results and provides insight into the models' generalizability. We will therefore include the corresponding "leave-one-area-out" results in the revised manuscript. Steinhoff-Knopp, B. and Burkhard, B. (2018): Soil erosion by water in Northern Germany: long-term monitoring results from Lower Saxony,450 Catena, 165, 299–309, https://doi.org/10.1016/j.catena.2018.02.017
Minor Comments	
104: I may be wrong, but the overall study areas cover a few hundred ha, but the grid of the original R-factor was 1 km x 1 km. Even if resampled (how?), is this not too broad for the study area context. Maybe a reference which refers to this procedure could be useful?	The R factor indeed shows only regional variation on a 1 km x 1 km resolution and does not differ in a relevant manner within individual study areas. But it differs between the study areas which are situated in different regions of lower Saxony resulting in different R factors. We agree that further details are needed to describe the different predictor variables and will add a table to the appendix with comprehensive information on each variable, including the R factor.
123: It would be more precise to write "a random subset of the feature [or variables]". Using a subset of data (i.e., training data) is also possible as a hyperparameter but not by definition a classical parameter in Random Forest.	Thank you for pointing this out. We agree and will change the phrasing accordingly.

2.3.4 It is not clear from the section but implied. Did the authors use a “2D” CNN, with what Y x Y raster cell?	A 2D CNN was used with 7 by 7 pixels. We will add more details to the description in the manuscript to make this clearer.
Figure 4: Why do the ECDF curves of the models appear so smooth? I would expect them, similar to the mapped erosion rate, to be step functions. This suggests that the ECDFs may have been constructed differently for the models and for the mapped erosion rate. Could the authors please clarify how these curves were generated?	The ECDFs were generated directly from all continuous erosion values by sorting the data and plotting the cumulative proportion of values ≤ x using a step function (matplotlib.pyplot.step). No interpolation or smoothing was applied. The smoother visual appearance of the model ECDFs results from the continuous and smoother nature of the model estimates.
Figure 5: The unit is missing. It is not simply [%], but rather increase of MSE in %. While this may be clear from the context, the figure should explicitly state the correct unit.	Thank you for highlighting this. The figure represents the relative permutation importance [%], which is based on each variable’s normalized contribution to the total increase in MSE. We will adapt the manuscript and figure label to make this clearer.