Response to Reviewers

Dear Reviewers,

We sincerely thank both Reviewers again for their constructive feedback. We have carefully considered each of the comments, implemented all suggested changes and revised the manuscript accordingly. We are confident that the new revisions made in response to your suggestions strengthened the manuscript further and made it suitable for publication. Please find our detailed responses to each of your points below.

Response to Anonymous Referee #1

Reviewer comments	Changes	Response
L 9: was	L 9	Thank you for catching this. We've fixed it accordingly.
L 72: unit for RMSE is missing	L 12	We added units.
L 13: unit missing	L 13	We added units.
L 22: and pedological factors! You should not forget soil in a soil journal.	L 23	Yes, the sentence has been revised to include "pedological factors"
L 86: Is 'increasing' the appropriate word? Increasing relative to what and over which time?	L 87	We changed the phrasing.
L 88: followed in importance or in rotation? This will also make clear what is meant by 'primary'.	L 89 f.	We changed the phrasing to make this clearer.
L 93: Is 'include' the correct word? Are there more than two? Better 'are'.	L 94	Thank you for the suggestion. We changed it to "are".
L 98: There is no variable in the data set which reflects snowmelt. I do not criticise this, but it may justify a remark in the discussion. Even with a relatively large dataset collected for an exceptionally well-known area (20 years of research), important information may still be missing. It remains always speculative whether we regard a data set as complete. We can hardly afford the same effort everywhere. Insufficient information will always be our fate. Although this is speculative, I would expect that ML handles missing information better than physical models. Physically based models must fail if only one key variable is missing or incorrect.	L102 f. L 327 - 332	Thank you for pointing this out. Precipitation effects during the winter dormant phase are cumulatively mapped after the snowmelt in early spring before the first agricultural measurements take place in the field. Snowmelt as well as single rainfall events during winter time (November to Feburary/March) should thus be cumulatively included in the data set, but cannot be separated. We added an explanatory phrase in the methodology section. Nonetheless, we agree that not all processes/factors were (or can be) included. We have added a corresponding paragraph to the Discussion section addressing this.
L 98: All organisations of standards require that units are printed upright while variables are tilted. Please follow this rule and please be consistent (upright and tilted units alternate).	e.g. L 99	We changed this and units are now consistently typeset in accordance with conventions.
L 98: Your definition of an erosive event cannot be found in DIN. It is acceptable to use your own definition, provided it is clearly defined. Is it correct that the rainfall amount was irrelevant for your definition? You may have missed many events.	L 99 ff.	Thanks for pointing this out. Our definition, however, for an erosive rainfall as basis for the whole erosion mapping programme can be found in Schwertmann et al. (1987) (the Bavarian/German adaption of the USLE) and the German DIN standards for the USLE valid during the monitoring period 2000 to 2020. We also fixed the rainfall amount to reflect the correct threshold used during the survey period (10 mm (the standard implemented by Schwertmann et al. (1987) and in the first two DIN versions) instead of the 12.7 mm threshold implemented in the DIN after 2022). Furthermore, we now clarified the definition further and added the respective DINs.

		Events below these thresholds were, by definition, not classified as erosive, but of course, it cannot be ruled out that single precipitation events occurring close to each other may have been missed.
Table 1: Wrong units for R and K. What unit should 'M' be?	Table 1	Thank you for pointing this out. We fixed the formatting.
L 111: 'resampled' is not clear. Was there any modification involved, or did you use the same value for each of your pixels within one radar pixel?	L 116 f.	This could indeed be more precise. We directly assigned the values within the 1 km pixels to the pixels without any modification. We specified the text accordingly.
L 115: This sentence should appear much earlier. Otherwise, the reader may be puzzled by some descriptions and may not realise that additional information is available..	L 112	We moved the sentence to the beginning of the variable description. The corresponding information can also be found in the Caption of Table 1 to avoid confusion for the reader.
L 170: continuously?		Continuous here refers to the continuous (i.e., non-discrete) model output.
L 209 -211 and Table 1: RMSE and MAE require units. Your statements that one method performs better than the other remains unjustified because you provide no statistical test that the values differ. I would suggest calculating the 95% confidence interval. This will also show whether three decimals are justified. I doubt this. Presumably, not more than one decimal is justified.	Table 2, L 217 -230	We implemented changes as suggested, added units, 95 % confidence intervals and changed the number of decimals accordingly.
Table 3: Is the number of decimals justified?	Table 3	We changed the number of decimals to two.
Figure 2: This is a nice map, and I appreciate it. However, I hardly see any correspondence with Fig. 1 in erosion severity (particularly panel b). I do not criticise this, but it is not reflected in the text. The text only reports that very high values are underrepresented, but this is not sufficient. The patterns, at least in panel b, but to a lesser degree also in other panels, are very different. I think that this is really interesting. In your previous version, you criticised the USLE. If the USLE fails, we may attribute this to the limitations and constraints of the USLE. In this case, ML was free of all these limitations and still had problems reproducing the pattern even with good and abundant input data. To me, it appears that we have to live with considerable uncertainty, and this is not a critique but apparently a fact that we cannot overcome (at least presently).	L 231 ff.	We added further statements regarding the spatial distribution and patterns of the predicted soil erosion. While we agree that several erosion features were not correctly reproduced, we clarify that some spatial patterns are visible, show partial correspondence with the mapped data and are not random. We therefore expanded the manuscript to more clearly indicate where the models partially succeeded and where erosion features were underestimated or not reproduced (e.g., in Barum). We agree that there are still many uncertainties remaining.
L 228: The relative error can still be high (and likely is). Perhaps, the relative error of the lowest classes is even larger than that of the high classes. It would be good to differentiate RMSE and MAE in Table 2 according to class (similar to Table 3). This would allow us to judge the relative error depending on soil loss.	Table 2	Thank you for this comment. We revised Table 2 as suggested and included RMSE and MAE differentiated by soil-erosion severity. This further highlights where the RF and neural network models differ (slightly) and where they perform similarly.
Chapter 3. 3: Really nice and insightful. The large influence of aspect is strange to me. I would not have expected this. Would you? Is there a physical reason? Snowmelt?		Please note that the results of the permutation importance analysis changed partially following the implementation of the stricter nested hyperparameter tuning procedure (see response to Reviewer #2). Aspect does no longer have an high permutation importance, but Machining direction vs.

		aspect is still important. Nevertheless, a potentially high importance of aspect could be explained by the fact that the models do not evaluate predictors in isolation but captures nonlinear interactions among input variables. Although aspect alone may have limited predictive value, its combination with machining direction is highly informative (as shown also in the new results) as a large share of linear erosion features (rills) was monitored in tramlines in direction of slope.
L 243: There was no test to prove this statement	L 253 - 258	The revised models show a difference, with regard to the F1 score. We restructured the paragraph, rephrased the statement and put it into the respective context to reflect the new results.
L 249: was	L 256 f.	Changed.
L 254: I disagree with your interpretation that this is a failure of the models. This result was caused by your decision to use log values, which gives large weight to the low erosion rates. This effect is well known because it also appears in other statistical contexts, such as regression analysis. Fitting a power function to untransformed data yields a different equation than calculating a regression with log-transformed data. Log-transformation (and other transformations) leads to a distortion of the variances. There is a wealth of statistical papers on this effect.	L 263 f.	The log-transformation was applied to stabilize variance and to reduce the influence of extreme erosion values during training. This approach stabilises training in machine learning approaches. We do not claim that the failure to predict high values is a failure of the models, but a limitation in the data (only 3.9 % of the pixels show loss rates higher than 5 t ha-1 yr-1). This is indeed amplified partly by the log transformation. We added a sentence to the discussion to make this clearer.
L 255: More data with high soil losses cannot help. Suppose you include a field from Sumatra with significant losses in your dataset. How could this improve the pattern matching within your fields?	L 314 f.	A larger representation of high-erosion events within the same landscape may help reduce class imbalances. In theory, this would improve the models ability to learn relationships between high-erosion events and predictor combinations that are currently underrepresented in the training data. However, whether this is indeed the case requires further testing and we have therefore adjusted the phrasing accordingly.
L 269-273: This was not tested. Nothing can be said.	L 280 - 283	The paragraph was changed to better reflect the current results.
L 288-290: Strange and interesting. Field observations and aerial photos consistently demonstrate a significant influence of these parameters. The calculation of these parameters strongly depends on the scale. May it be that your data were obtained on a wrong scale (either too coarse or too detailed) to capture these influences?	L 320 f.	We agree, that the low importance does not necessarily imply that they have no physical relevance for soil erosion by water. The models may rely on other models to capture the underlying relationships, relevant processing may be missing or curvature related variables may indeed influence soil erosion by water at a different scale. This requires further testing and we added this point to our discussion.
L 303-304: No; see remarks above	L 314 - 318	We changed the phrasing. See answer above.
Table A1: Slope and other parameters: Only the number of grid cells is reported, but not their size. The size determines on which scale a certain parameter is determined and which features can be detected. This may explain the low importance of some topographic attributes.	Table A1	Grid cell size: Cell size has been added to the table caption (5 m x 5 m). Aspect: Thank you for pointing this out. We clarified this in the Table. LS factor: We revised accordingly.

Aspect 360, 780: How is zero defined? Given the large importance of these variables, a more detailed description is required. LS factor: I found the description in the reply letter, which is not available to the readers, clearer (pixel-based LS factors using the Desmet and Govers, 1996, method, which includes field boundaries) K factor: either the numbers or the units are wrong (or even both) C factor: Schwertmann is a bit outdated regarding the C factor due to climate change (different growing periods and seasonal distribution of rain erodibility). R factor: Either the numbers or the units are wrong (or even both). I am not aware that R can be taken from Winterrath et al. (2018).		C factor: The C-factor was parameterized following the detailed method in Schwertmann et al. 1987 (using soil loss rations and relative erosivity index) with locally updated information on the seasonal distribution of rainfall erosivity and crop development. Information was taken from our monitoring and the RADKLIM-data. We incorporated this information in the Table. K factor and R factor: Thank you for catching this error. The originally reported R and K values were indeed expressed in a German unit convention. Both factors have now been converted to the commonly used units: R: MJ mm ha^{-1} h^{-1} yr^{-1} K: t h MJ^{-1} mm^{-1} The conversion was purely multiplicative (R ×10, K ÷10). This would not affect the results, as the model was trained using standardized input variables (feature scaling). Nonetheless, the converted factors were used in the new implemented version of the models (see response to Reviewer #2). Winterrath et al. (2018) refers to the RADKLIM-data. We added a reference to the R-factor calculation.

Response to Anonymous Referee #2

Reviewer comments	Changes	Response
Based on the code and the accompanying description (please correct me if I have misinterpreted your implementation), the current workflow appears to have a data leakage because the hyperparameter selection is not strictly separated from the final model evaluation. Specifically, it seems that the same held-out data (or area) are being used both to (1) choose the hyperparameter configuration and (2) report the test performance. This would constitute a leakage: selecting the hyperparameter setting that performs best on the test set effectively "overfits" to the same test set and leads to optimistically biased performance estimates. I consider this a major methodological flaw, as the performance of the CNN may be inflated to a (potentially substantial) degree, and makes the benchmark with Random Forest "unfair". To avoid this, hyperparameter tuning should be performed within a nested cross-validation (or an equivalent three-way split strategy). Concretely, for each outer (leave-one-area-out) split used for performance estimation, there must be an inner split performed only on the outer training portion to select hyperparameters; the outer held-out area must remain completely untouched until final evaluation.	L 195 – 200 Results, Discussion and Code	Thank you for this comment. Based on your suggestion, we implemented a stricter hyperparameter tuning process. A nested cross-validation approach is now applied, where the left-out area is used only for the final prediction and validation. The left-out area remains completely untouched throughout hyperparameter tuning and model training. Instead, the six remaining areas are used exclusively for hyperparameter selection within an inner leave-one-area-out procedure. In this inner loop, five areas are used for training and one area for validation, iteratively. To keep the required computational resources of this workflow at a reasonable level, random search was used within the inner cross-validation loop. Information on the implemented approach has also been added to the main manuscript and all relevant scripts have been updated. While the main performance metrics somewhat remained similar, all results have been replaced with those obtained from the revised nested hyperparameter tuning workflow. This ensures that model selection and performance evaluation are strictly separated and that no data leakage occurs.
Given that erosion rate is a continuous target variable, I consider it unusual to use the F1 score as a relative error measure rather than reporting R^2. This may be established conventions from soil erosion literature, but the choice could be very shortly justified.	L 173 - 176	Soil erosion is often classified into erosion severity classes (Borelli et al. 2018, Steinhoff-Knopp & Burkhard 2018). The F1 score complements the regression metrics to evaluate the performance of the models within different severity classes. We added a brief justification to the main manuscript.
Out-of-distribution evaluation (as induced by leave-one-area-out) can involve substantial predictive bias. It would be informative to include Mean Error (ME) as an additional metric to quantify bias, alongside the existing performance measures, without any need for much further interpretation.	Fig. 3, Fig. C5 – C8	The mean error of each model is now reported alongside the difference between predictions and mapped erosion rates (see e.g. Figure 3).
I still find the term "Permutation importance [%]" misleading or at least unusual, as the % can be related to the (absolute) permutation-based increase. However, the figure caption (Figure 4) provides the correct explanation, so this can be deemed acceptable.	Fig. 4	We do not want the axis label to be misleading and therefore changed it accordingly.