Supplement of Biogeosciences, 22, 3769–3784, 2025
https://doi.org/10.5194/bg-22-3769-2025-supplement

Biogeosciences Open Access EGU

Supplement of

Improved understanding of nitrate trends, eutrophication indicators, and risk areas using machine learning

Deep S. Banerjee and Jozef Skákala

Correspondence to: Deep S. Banerjee (dba@pml.ac.uk)

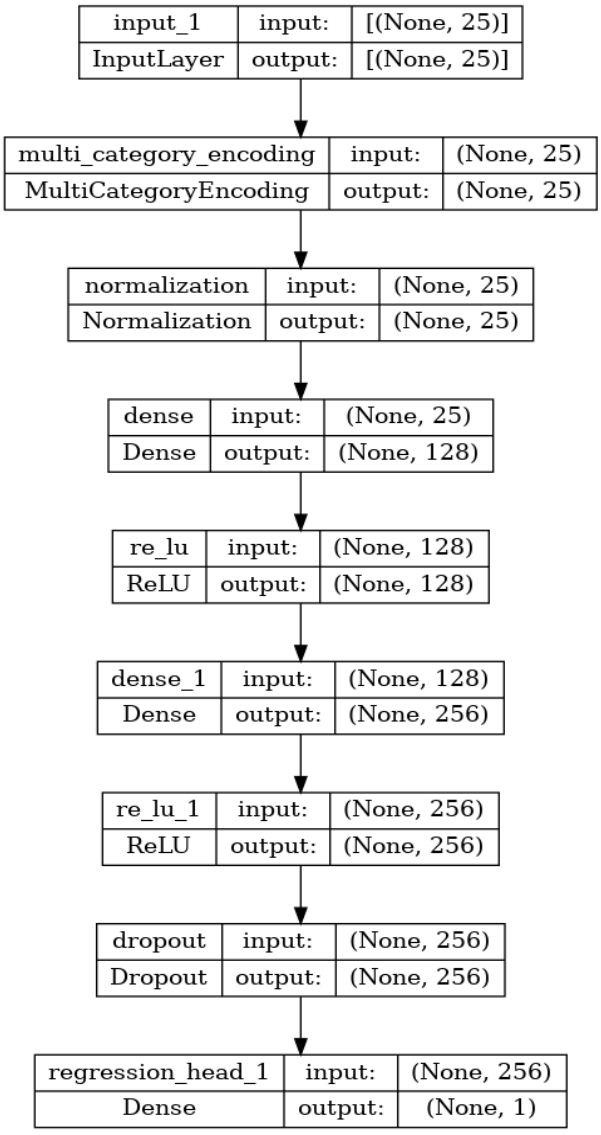

Figure S1. The architecture of the NN model used in this study.

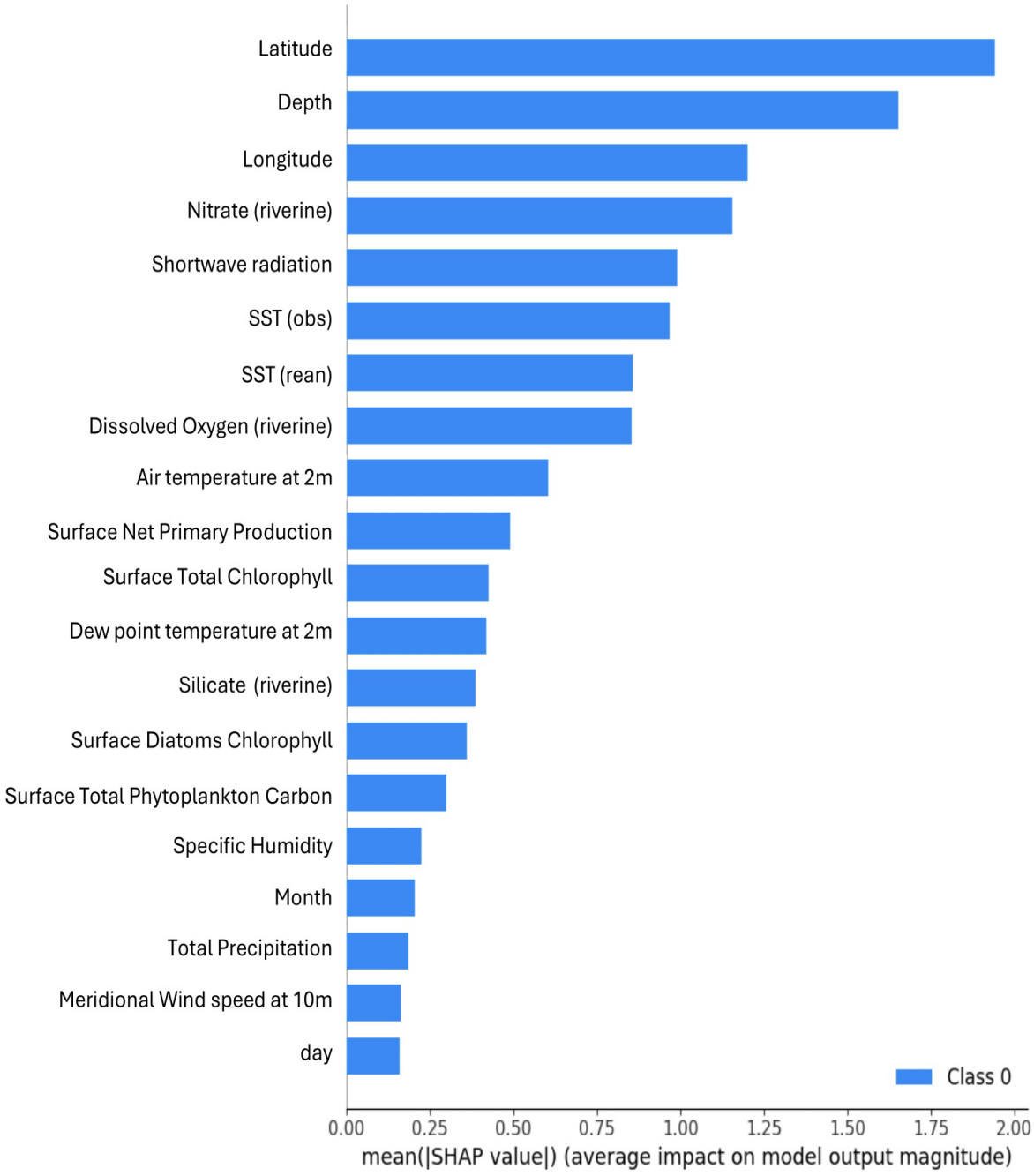

Figure S2. Evaluation of input feature importance using SHAP analysis. The features are ordered according to their importance, with 20 most important features shown.

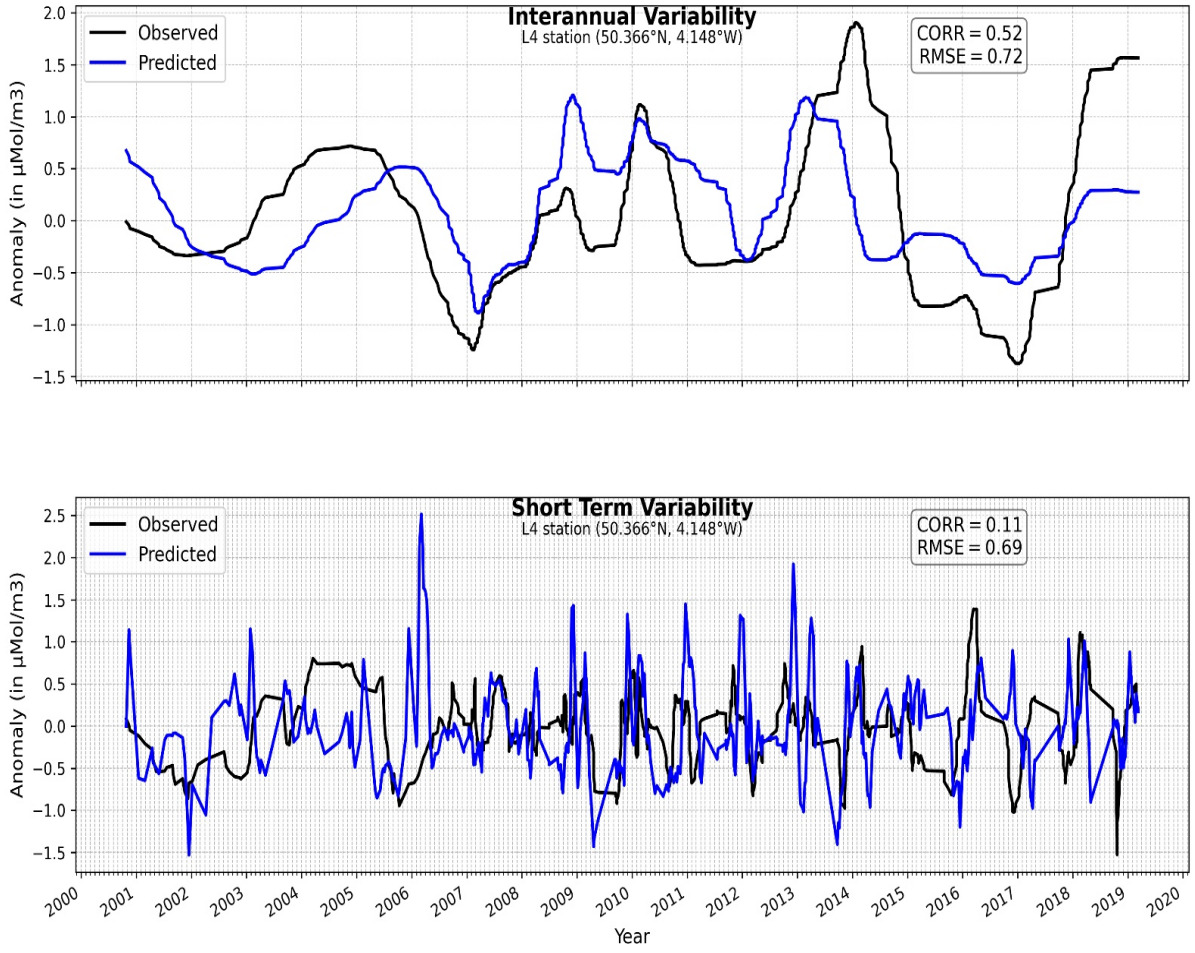

Figure S3. The comparison of interannual series (upper plot) and sub-monthly series (bottom plot) for surface nitrate at L4 station, between the NN-prediction and the L4 observations. The time-series were obtained by (i) subtracting seasonal climatology (smoothed on daily time-scale) from nitrate time-series and then (ii) either low-pass filtering the data (inter-annual time-series), or high-pass filtering the data (sub-monthly time-series), both on a 100 day scale.

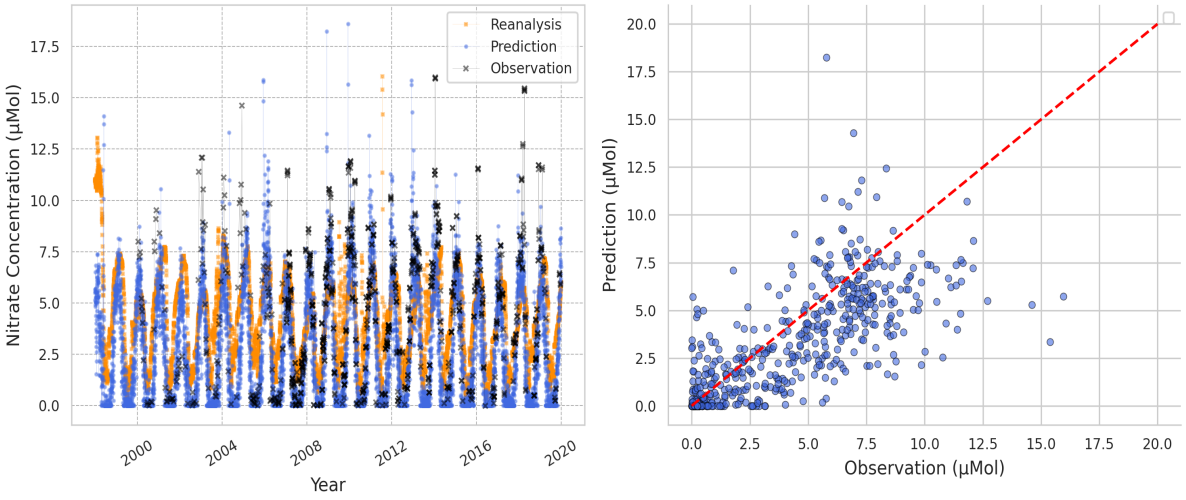

Figure S4. The comparison of the surface nitrate NN model prediction with the Copernicus reanalysis and the observations at the L4 station. The NN model prediction is from the NN-generated 1998-2020 gridded nitrate product. Both NN-model prediction and reanalysis are structured on the same spatial grid and were taken from the nearest model grid point to L4.

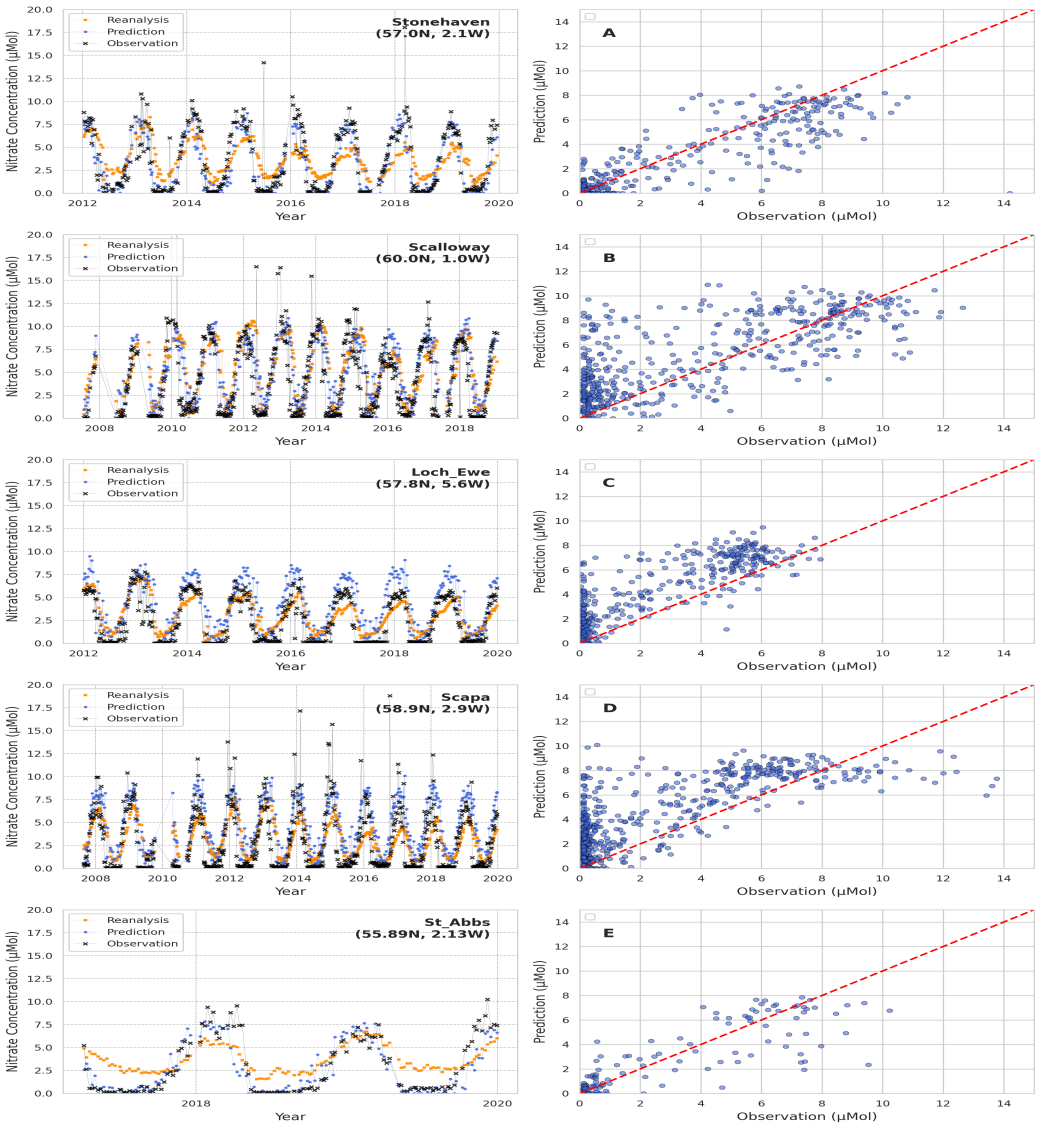

Figure S5. The comparison of the surface nitrate NN model prediction with the Copernicus reanalysis and the observations at five Scottish coastal stations. The NN model prediction is from the NN-generated 1998-2020 gridded nitrate product. Both NN-model prediction and reanalysis are structured on the same spatial grid and were taken from the nearest model grid points to the different stations.

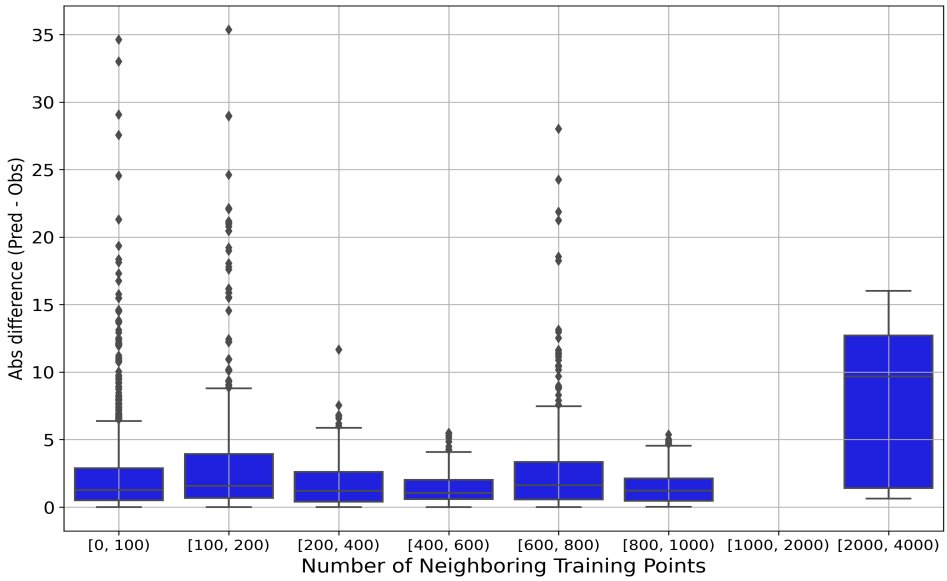

Figure S6. The RMSE skill score (μmol/m^3) of the NN-model calculated as a distance from the training data. The score is calculated as follows: (i) for each ICES test data-point we determine the number of training data that fall into the 50 km neighborhood of the test data point. The spatial separation of that test data-point from the training data can be then defined as inversely proportional to the number of training data that fall into the 50 km neighborhood of that test data-point. (ii) The test data are then binned into categories based on their spatial separation from the training data. The x-axis shows the bins, with their defining feature: the [X,Y] bin includes test data that have between X and Y training data in their 50 km spatial neighborhood. The y-axis then shows how well the NN model predicts the test data from the specific bin, i.e. it shows the distribution of absolute values of difference between the NN-model prediction and the ICES test data. The black horizontal lines show median values, the blue bars around those lines show two quantiles around the median, the error-bars show the interval with 95$\%$ of the values and the remaining extreme values are shown as separate dots.