AGILE: GIScience Series, 3, 66, 2022. https://doi.org/10.5194/agile-giss-3-66-2022
Proceedings of the 25th AGILE Conference on Geographic Information Science, 2022.
Editors: E. Parseliunas, A. Mansourian, P. Partsinevelos, and J. Suziedelyte-Visockiene.
This contribution underwent peer review based on a full paper submission.

ML-based water quality modeling at national level in a data-scarce region

Holger Virro[1], Alexander Kmoch[1], Marko Vainu[2], and Evelyn Uuemaa[1]

[1]Department of Geography, Institute of Ecology and Earth Sciences, University of Tartu, Tartu, Estonia
[2]Institute of Ecology, Tallinn University, Tallinn, Estonia

Correspondence: holger.virro@ut.ee

Abstract. Water quality (WQ) modeling can be used for gaining insight into WQ issues in order to implement effective mitigation efforts. Process-based nutrient models are very complex, requiring a lot of input parameters and computationally expensive calibration. Recently, ML approaches have shown to achieve an accuracy comparable to the process-based models and even outperform them when describing nonlinear relationships. We used observations from 242 Estonian catchments, amounting to 469 yearly total nitrogen (TN) and 470 total phosphorus (TP) measurements covering the period 2016–2020 to train random forest (RF) models for predicting annual N and P concentrations. We used a total of 82 predictor variables, including land use and land cover (LULC), soil, climate and topography parameters and applied a feature selection strategy to reduce the number of dependent features in the models. The SHAP method was used for deriving the most relevant predictors. The performance of our models is comparable to previous process-based models used in the Baltic region. However, as input data used in our models is easier to obtain, the models offer superior applicability in areas, where data availability is insufficient for process-based approaches.

Keywords. water quality, interpretable machine learning, random forest

1 Introduction

WQ modeling plays an important role in better understanding the magnitude and impact of WQ issues and in providing evidence for policy-making and implementing water management plans (Tang et al., 2019). Thus far, a common approach to WQ modeling has been the use of process-based models such as Soil and Water Assessment Tool (SWAT) (Arnold et al., 1998; Malagó et al., 2017; Me et al., 2015) or HYPE (Arheimer et al., 2012; Lindström et al., 2010), which have been widely used at the catchment level. However, process-based modeling is data- and calculation-intensive (Clark et al., 2017), and requires data that is not often easily available (e.g. soil bulk density, soil organic carbon content) or difficult to measure across larger areas.

In a thorough review article Tiyasha et al. (2020) investigated studies from the period 2000–2020 and found there has been an exponential growth in adopting ML-based methods for WQ modeling purposes. Neural network (NN) models have shown to be good at detecting the non-linear relationships in hydrological processes, resulting in accurate predictions of flow, sediment and nutrient concentrations (Kuo et al., 2004; Sarkar and Pandey, 2015; Singh et al., 2009). However, in order to achieve high accuracy in an NN model, calibration in the form of hyperparameter tuning needs to be implemented. For regression tasks, tree-based ML methods can often provide a robust alternative to deep learning techniques (Ho et al., 2019; Visser et al., 2022). Although many of the RF models have been applied only on the catchment scale, recent studies have shown that RF can also perform well in the case of large-scale WQ datasets. In particular, N and P models using RF (Marzadri et al., 2021; Sheikholeslami and Hall, 2022; Shen et al., 2020) have been applied successfully on the subcontinental and global level, thus producing valuable benchmarks about the scalability of RF models.

The aim of the study was to model annual total nitrogen (TN) and total phosphorus (TP) concentrations at national level using an ML approach. We used WQ data originating from the Environmental Monitoring Database KESE (Estonian Environment Agency, 2021) to train RF models for nutrient concentration prediction in 242 catchments across Estonia. A total of 82 environmental variables were used as predictors in the models. In order to yield the best results, a feature selection strategy along with hyperparameter optimization was performed when building the models. The models are applicable for predicting nutrient loads on an annual level, e.g. for the purpose of reporting national level WQ statistics in regional projects, such as HELCOM

in the Baltic Sea region (HELCOM, 2009). The results showed that this relatively basic RF modeling approach can have a performance similar to process-based models. Moreover, these models are easier to reuse and apply on a larger scale, since the required inputs can be derived from freely available datasets (e.g. satellite imagery).

2 Methods

2.1 WQ observations

Estonian WQ data was obtained from the KESE environment monitoring system website maintained by the Environment Agency (Estonian Environment Agency, 2021). The yearly mean values were calculated for a particular site if it had observations available in at least four distinct months within a year. The summary statistics of the aggregated yearly mean values are given in Table 1.

2.2 Catchments of WQ sites

For each of the 242 sites, catchments were delineated from the 5 m resolution LiDAR digital elevation model (DEM) provided by the Estonian Land Board (Estonian Land Board, 2020a) (Fig. 1). Hydrological conditioning involving burning in culverts, bridges, rivers, streams and larger ditches, as well as sink filling was applied to the DEM in order for the flow direction to better represent reality. The catchment delineation workflow was performed using the ArcHydro toolbox in ArcGIS (Esri, 2020). The size of the catchments ranged from 0.9 to 8,513.9 km^2.

2.3 WQ predictors

A total of 82 variables were used as predictors in the model (Table 2). The source data included layers in raster (GeoTIFF), vector (SHP and GPKG) and NetCDF formats. With the exception of the hydrology and agriculture parameters, all other predictor datasets were converted into raster layers with 5 m resolution in alignment with the DEM. The *rasterstats* Python package was used to calculate zonal statistics of the predictors for each catchment. Fertilizer deposition (manure_dep_n and manure_dep_p), the climate variables and the three forest disturbance derivations were the only predictors available on an annual basis in the corresponding source datasets and, thus were calculated for each year in the study period as spatio-temporal input data. All other predictors were considered to stay static throughout the study period.

2.4 WQ modeling using RF

2.4.1 Feature selection

In general, RF is considered to be comparatively resistant to collinear features. However, reducing the number of features can provide further support, e.g. trying to limit the amount of predictors to reduce data requirements in general, pre-processing of predictor data when reusing the model in the future or applying to other regions, or where data is not yet available in an ML ready form. In our case, the main purpose of feature reduction was the potential reuse of the model, i.e. as many parameters are not updated enough or are available only tentatively then it is preferable to use the ones that are the most easily obtained.

In order to reduce the number of collinear features, we employed a feature selection strategy shown in Fig. 2. The strategy for reducing the number of features was as follows:

- Each predictor was assigned a subcode given in Table 2. In general, predictors having multiple statistical derivations, but originating from the same source data had the same subcode.
- Pairwise correlations between all features were calculated along with correlations between features and the target value (WQ concentration)
- Feature pairs with correlation values above a certain correlation threshold (Table 3) were then extracted and sorted based on the strength of the correlation
- For each pair, the feature with a lower correlation with the target was determined
- While iterating over the feature pairs, the feature with the lower target correlation was removed from the set as long as at least one feature from the corresponding group remained in the set. For example, slope_mean, slope_min and slope_max could all be removed, provided that slope_std remained in the feature set due to being less collinear with the other features and, thus, having a lower placement in the correlation order.

As a result of the feature selection procedure, four feature sets were generated for TN and TP (Table 3). Each set corresponded to a Pearson correlation coefficient (0.9, 0.8, 0.7 or 0.6) used as a threshold to determine whether the feature selection procedure was applied to a feature pair.

2.4.2 Model building workflow

In order to investigate the predictive capabilities of the selected environmental variables, RF regression models for both nutrients were built using the Scikit-learn (Pedregosa et al., 2011) ML package in Python. A separate model was built for each of the feature sets derived from the feature selection procedure using the RandomForestRegressor class from Scikit-learn. The workflow used for developing the models is given in Fig. 3.

Hyperparameter optimization was carried out by using the RandomizedSearchCV algorithm (Bergstra and Bengio, 2012) from Scikit-learn's *model_selection* module using cross-validation with $k = 5$. The set of hyperparameters

Table 1. Statistics for WQ observations at the yearly level.

Parameter	Observations	Sites	Minimum	Mean	Maximum	Median	Standard deviation
			mg L^{-1}	mg L^{-1}	mg L^{-1}	mg L^{-1}	mg L^{-1}
TN	469	242	0.478	2.662	11.933	2.09	1.952
TP	470	242	0.008	0.051	0.27	0.044	0.032

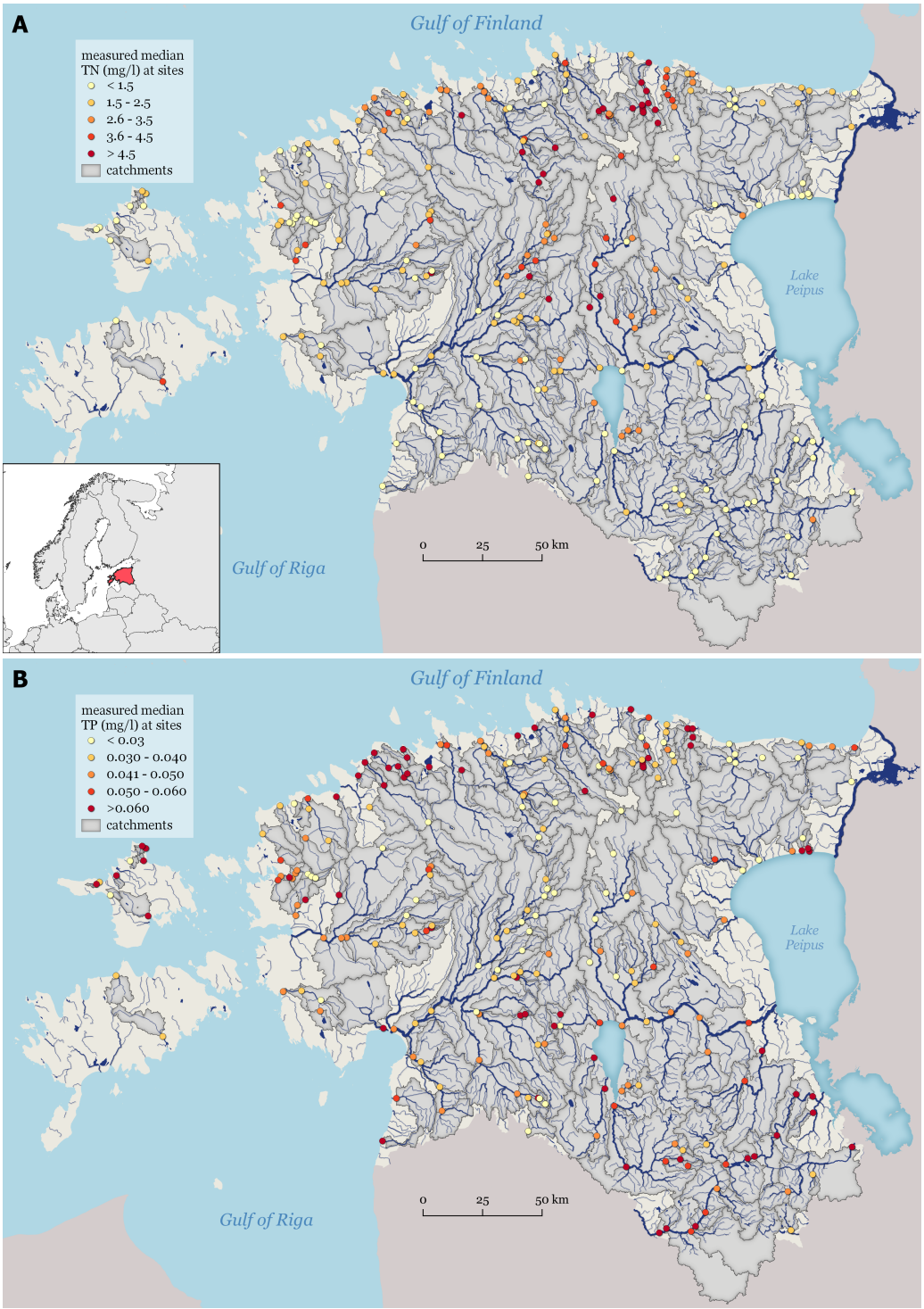

Figure 1. Median TN (A) and TP (B) concentration in observation sites 2016–2020.

Table 2. List of WQ predictor variables.

Category	Code	Subcode	Description	Unit	Source
topography	dem_min, dem_max, dem_mean, dem_std	dem	Elevation	m	(Estonian Land Board, 2020a)
topography	tri_min, tri_max, tri_mean, tri_std	tri	Terrain ruggedness index (TRI)		(Estonian Land Board, 2020a)
topography	twi_min, twi_max, twi_mean, twi_std	twi	Topographic wetness index (TWI)		(Estonian Land Board, 2020a)
topography	flowlength_min, flowlength_max, flowlength_mean, flowlength_std	flowlength	Flow length in the catchment		(Estonian Land Board, 2020a)
topography	slope_min, slope_max, slope_mean, slope_std	slope	Slope		(Estonian Land Board, 2020a)
soil	awc1_min, awc1_max, awc1_mean, awc1_std	awc1	Water holding capacity (first layer)	mm H_2O mm^{-1} soil	(Kmoch et al., 2021)
soil	bd1_min, bd1_max, bd1_mean, bd1_std	bd1	Bulk density (first layer)	g cm^{-3}	(Kmoch et al., 2021)
soil	clay1_min, clay1_max, clay1_mean, clay1_std	clay1	Clay content (first layer)	% mass of fine earth fraction	(Kmoch et al., 2021)
soil	k1_min, k1_max, k1_mean, k1_std	k1	Hydraulic conductivity (first layer)	mm h^{-1}	(Kmoch et al., 2021)
soil	rock1_min, rock1_max, rock1_mean, rock1_std	rock1	Rock content (first layer)	% mass of fine earth fraction	(Kmoch et al., 2021)
soil	sand1_min, sand1_max, sand1_mean, sand1_std	sand1	Sand content (first layer)	% mass of fine earth fraction	(Kmoch et al., 2021)
soil	silt1_min, silt1_max, silt1_mean, silt1_std	silt1	Silt content (first layer)	% mass of fine earth fraction	(Kmoch et al., 2021)
soil	soc1_min, soc1_max, soc1_mean, soc1_std	soc1	Soil organic carbon (SOC) content (first layer)	% of soil weight	(Kmoch et al., 2021)
LULC	arable_prop	arable	Proportion of arable land	% of catchment area	(Estonian Land Board, 2020b)
LULC	forest_prop	forest	Proportion of forest	% of catchment area	(Estonian Land Board, 2020b)
LULC	grassland_prop	grassland	Proportion of grassland	% of catchment area	(Estonian Land Board, 2020b)
LULC	other_prop	other	Proportion of other LULC	% of catchment area	(Estonian Land Board, 2020b)
LULC	urban_prop	urban	Proportion of urban land	% of catchment area	(Estonian Land Board, 2020b)
LULC	water_prop	water	Proportion of water	% of catchment area	(Estonian Land Board, 2020b)
LULC	wetland_prop	wetland	Proportion of wetland	% of catchment area	(Estonian Land Board, 2020b)
LULC	arable_prop_buff_100, arable_prop_buff_500, arable_prop_buff_1000	arable	Proportion of arable land within 100/500/1000 m stream buffer	% of stream buffer	(Estonian Land Board, 2020b)
LULC	forest_disturb_prop_buff_100, forest_disturb_prop_buff_500, forest_disturb_prop_buff_1000	forest_disturb	Proportion of disturbed forest area within 100/500/1000 m stream buffer	% of stream buffer	(Senf, 2021)
LULC	rip_veg_nat_prop, rip_veg_drain_prop	rip_veg	Total area of riparian vegetation buffer around natural streams/drainage ditches divided by catchment area	% of catchment area	(Uuemaa et al., 2021)
hydrology	area	area	Area of the catchment	m^{-2}	
hydrology	stream_density	stream_density	Stream density	m m^{-2}	(Estonian Land Board, 2020b)
hydrology	pol_sen_drain_m_prop, pol_sen_drain_h_prop, pol_sen_drain_vh_prop	pol_sen	Proportion of riparian buffer moderately/highly/very highly sensitive to pollution around drainage ditches/natural streams	% of riparian buffer	Uuemaa et al. (2021)
agriculture	livestock_density	livestock_density	Density of livestock	livestock units per ha	PRIA
agriculture	manure_dep_n, manure_dep_p	manure	Mean deposition of nitrogen/phosphorus in manure	kg ha^{-1}	(Statistics Estonia, 2021)
climate	precip_mean	precip	Mean annual total precipitation	mm	(Muñoz Sabater, 2021)
climate	snow_depth_mean	snow_depth	Mean annual snow depth	cm	(Muñoz Sabater, 2021)
climate	temp_max, temp_mean, temp_min	temp	Maximum/mean/minimum annual temperature	C°	(Muñoz Sabater, 2021)
geology	limestone_prop	limestone	Proportion of catchment located on limestone	% of catchment area	(Estonian Land Board, 2020b)

Table 3. Feature sets generated during the feature selection procedure.

Threshold	TN feature set	Number of features	TP feature set	Number of features
0.9	TN_FEAT_V1	62	TP_FEAT_V1	64
0.8	TN_FEAT_V2	55	TP_FEAT_V2	56
0.7	TN_FEAT_V3	47	TP_FEAT_V3	47
0.6	TN_FEAT_V4	38	TP_FEAT_V4	40

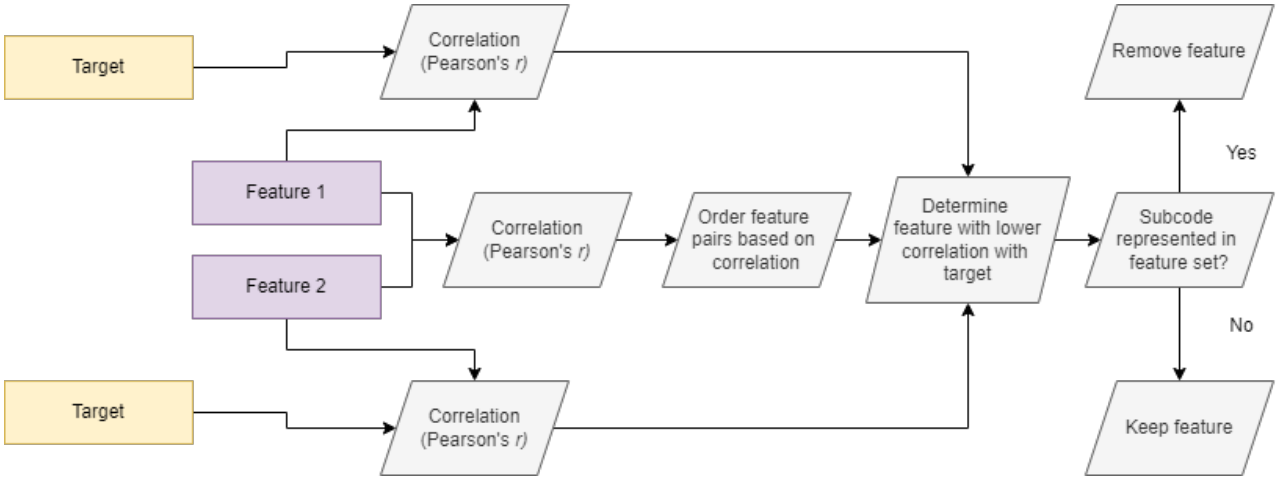

Figure 2. Workflow used for the feature selection strategy. Here, *target* refers to WQ concentration, while *feature* refers to predictor variable. *Subcode* refers to the grouping given in Table 2.

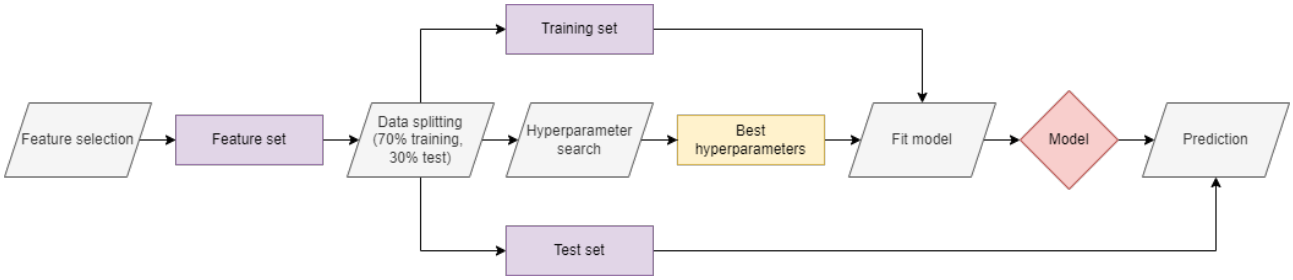

Figure 3. Workflow used for building the RF model for each feature set.

evaluated with RandomizedSearchCV is given in Table 4. The procedure was repeated for each feature set, since the best combination of hyperparameters can depend on the available features. The algorithm randomly selected different hyperparameter combinations and scored each of the iterations based on the mean squared error (MSE). Finally, the hyperparameter set with the smallest MSE was flagged as the best for a particular model configuration and used for fitting the model.

2.4.3 Model evaluation

Three accuracy indicators were calculated for the different model configurations of both nutrients:

- **r2_train:** the coefficient of determination (R^2) calculated on training data
- **r2_test:** R^2 calculated on test data
- **mape_train:** mean absolute percentage error (MAPE) calculated on training data
- **mape_test:** MAPE calculated on test data

For MAPE calculated on the test set scores below 20% show good prediction accuracy, while scores in the 20–50% range are considered reasonable. Out of the four model versions, the best model for both nutrients was determined based on r2_test, while also trying to minimize the number of features used, i.e. if multiple models had a similar accuracy then the model with the least amount of features was chosen. In addition to aforementioned performance indicators, feature importances were derived for the best TN and TP models. As a default option, the feature importances in RF are based on Gini importance (or mean decrease impurity), which is considered to be biased towards features with high cardinality (Grömping, 2009). To achieve a less biased result, the SHapley Additive exPlanations (SHAP) explainable AI (XAI) method from the corresponding Python package was implemented to detect the most important features. The method uses Shapley values from game theory to estimate how each feature contributes to the prediction (Lundberg et al., 2020).

For each catchment, the ratio between the observed and predicted values was also calculated. This ratio indicates whether the model under (values greater than 1.0) or overestimated (lower than 1.0) the nutrient concentration of a particular sample. In order to explore the spatial variability in the model performance, the ratios were plotted on the map.

Table 4. Hyperparameters used in RandomizedSearchCV.

Parameter	Values
n_estimators	10, 20, 30, 40, 50, 60, 70, 80, 90, 100
max_depth	10, 20, 30, 40, 50, 60, 70, 80, 90, 100, None
min_samples_split	2, 5, 10
min_samples_leaf	1, 2, 4
max_features	auto, sqrt, log2
bootstrap	True, False
oob_score	TRUE

2.5 Data and software availability

The scripts used for data processing and building the models are available on Zenodo at https://doi.org/10.5281/zenodo.5910319 (Virro et al., 2022). The models along with their corresponding input data and the results of the modeling are available at https://doi.org/10.5281/zenodo.6325311 (Virro and Kmoch, 2022).

3 Results

3.1 Performance of the models corresponding to each feature set

The results of TN models using the four feature sets are given in Table 5 and the results of TP models in Table 6. The difference between r2_train and r2_test was significantly smaller in case of the TN models compared to TP. Therefore, the TN models are less likely to overfit. Since the accuracy of the models corresponding to the four feature sets was similar for both nutrients, the best model was determined by trying to minimize the size of the feature set. Thus, TN_MODEL_V4 (38 features) and TP_MODEL_V4 (40) were deemed as the best, because they used less than half of the original predictors (82).

3.2 Feature importances

In order to detect the most important factors contributing to the nutrient concentrations, feature importances based on SHAP values were calculated for the best models. The impact on the the prediction is given in the units of the target, i.e. mg L^{-1}. Due to differences in their corresponding concentration values (Table 1), the SHAP values of TN are several magnitudes higher than those of TP.

The most important features along with the direction of the impact a given feature has on the prediction are given in Fig. 4. It can be seen that arable_prop and rock1_mean had a positive correlation with TN concentration, while k1_mean and sand1_mean a negative one (Fig. 4A). As with the latter two features, TN was also higher in catchments were forest_prop was low. In the case of TP, high limestone_prop resulted in a lower TP concentration, while both higher grassland_prop and urban_prop had a positive correlation with the target, meaning that higher urban and grassland proportion in the catchment results in higher TP values (Fig. 4B). A negative correlation was detected for dem_min and a positive one for stream_density.

3.3 Spatial distribution of modeling results

The spatial distribution of modeling results is given in Fig. 5. In the case of both models the ratios in around half the catchments were in the range 0.91–1.11, meaning that the prediction was either over or underestimated by only 10%. The TN model resulted in 114 catchments with ratios in this range, while the total range of the values was 0.27–2.17 (Fig. 5A). In general, the greatest discrepancies between the observed and predicted TN values were more common in smaller catchments with 11 catchments having a ratio below 0.5 and five a ratio above 1.5.

The ratios of the less accurate TP model showed more variability with ratios ranging from 0.33 to 3.33 (Fig. 5B). Still, there were 119 catchments in the range 0.91–1.11. The greatest over (four catchments) and underestimations (five catchments) were present in a few small catchments on the northern coast and the islands.

4 Discussion

4.1 Performance of the models

The RF models built in this study were able to reach an accuracy comparable to process-based models used in similar studies previously. Model uncertainty expressed here through the observed to predicted ratio was within the range previously shown by the Balt-HYPE model (Arheimer et al., 2012) and the R^2 values were either higher or similar to catchments tested with HYPE in similar environmental conditions (Lindström et al., 2010). The difference in the accuracy of the best TN ($R^2 = 0.83$) and TP ($R^2 = 0.52$) models matches the observations described in previous nutrient modeling efforts. Both process-based (Hollaway et al., 2018; Malagó et al., 2017; Me et al., 2015) and ML (Álvarez-Cabria et al., 2016; Shen et al., 2020) modeling studies have shown that the predictive power of TN models is usually greater than TP models.

Although catchment area was not among the most important features in either model, a relationship between model

Table 5. Results of the four model versions used for TN prediction along with hyperparameters derived from the RandomizedSearchCV algorithm.

Attribute	TN_MODEL_V1	TN_MODEL_V2	TN_MODEL_V3	TN_MODEL_V4
n_features	62	55	47	38
test_size	0.3	0.3	0.3	0.3
n_samples_train	328	328	328	328
n_samples_test	141	141	141	141
r2_train	0.877	0.937	0.917	0.928
r2_test	0.791	0.835	0.821	0.833
mape_train	0.21	0.16	0.17	0.17
mape_test	0.31	0.30	0.30	0.29
max_depth	90	30	30	60
max_features	sqrt	sqrt	sqrt	sqrt
min_samples_leaf	4	2	2	2
min_samples_split	5	2	2	5
n_estimators	30	30	30	60
bootstrap	True	True	True	True

Table 6. Results of the four model versions used for TP prediction along with hyperparameters derived from the RandomizedSearchCV algorithm.

Attribute	TP_MODEL_V1	TP_MODEL_V2	TP_MODEL_V3	TP_MODEL_V4
n_features	64	56	47	40
test_size	0.3	0.3	0.3	0.3
n_samples_train	329	329	329	329
n_samples_test	141	141	141	141
r2_train	0.916	0.856	0.94	0.926
r2_test	0.496	0.52	0.484	0.517
mape_train	0.17	0.17	0.13	0.15
mape_test	0.27	0.26	0.27	0.26
max_depth	10	60	NaN	10
max_features	log2	sqrt	log2	log2
min_samples_leaf	1	2	1	1
min_samples_split	2	5	2	2
n_estimators	60	60	60	60
bootstrap	True	True	True	True

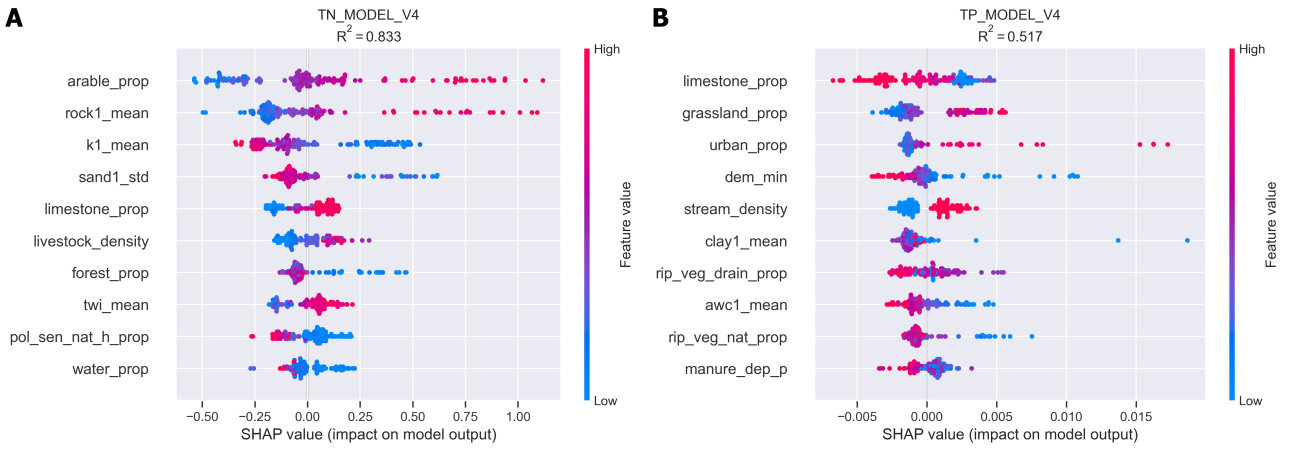

Figure 4. SHAP summary plots of the best models for TN (A) and TP (B). Here, each sample is colored by its corresponding feature value with higher values shown in pink and lower values in blue.

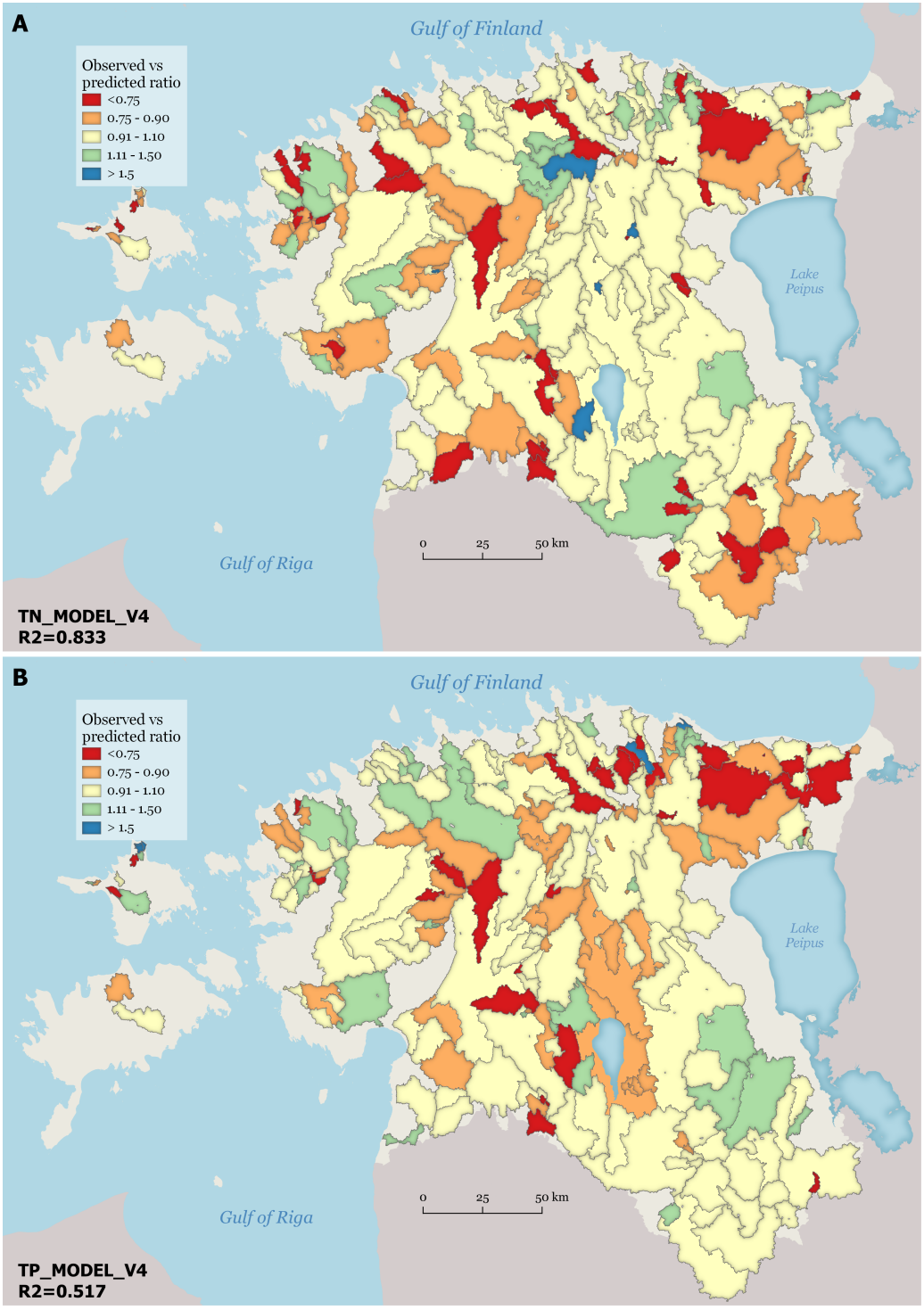

Figure 5. Spatial distribution of the ratio between observed and predicted values in catchments for the best TN (A) and TP (B) model. Overestimated concentrations are indicated by ratios lower than 1.0, while ratios greater than 1.0 show where the model underestimated the nutrient concentrations.

performance and catchment size was detected with smaller catchments being more likely to be either under or overestimated (Fig. 5). In general, larger catchments are more diverse in LULC with forests and wetlands acting as buffers between fields and streams. Therefore, more nutrients are adsorbed during transport, which results in more stable concentrations in streams (Lintern et al., 2018). Small catchments are more likely to be more uniform in LULC and soil (Bartley et al., 2012), which can make them less resilient when dealing with the effects of nutrient runoff, resulting in greater fluctuations in concentration (Bhattacharjee et al., 2021; Smith et al., 2005).

4.2 Most important features

The feature selection strategy worked well and the best performing models used the smallest corresponding feature sets in the case of both TN and TP. Our study showed that large numbers of features (predictors) are not always necessary to achieve good accuracy with ML models and rather relevant features are more important to achieve good accuracy.

From the list of the most important features in the TN model (Fig. 4A), occurrence of arable_prop and livestock_density were expected as they have been commonly described as some of the most influential predictors due to increasing fertilizer and manure deposition in the catchment (He et al., 2011; Hooda et al., 2000; Liu et al., 2020). Likewise, the role of forests in retaining N explains the negative correlation between forest_prop and TN concentration (Moreno-Mateos et al., 2008; Peterjohn and Correll, 1984).

One of the most important predictors for TP was found to be limestone_prop, which had a negative correlation with TP (Fig. 4B), meaning that TP losses were smaller from areas where limestone bedrock was dominating. This can be explained by the well-known fact that neutral or even higher pH values are optimal for the uptake of phosphates by plants, which in turn reduces TP losses (Barrow, 2017). Wastewater treatment plants and sewage systems are a major source of P from urban areas (Edwards and Withers, 2008; Lintern et al., 2018; Yang and Toor, 2018), which explains why urban_prop was the strongest predictor for TP related to LULC. The occurrence of dem_min is likely related to urban_prop as some of the biggest urban areas in Estonia are located in low lying areas on the coast. Nutrients reach the waterbodies faster if the stream network is dense, which explains the higher TP values from catchments with higher stream_density (Ebeling et al., 2021; Gentry et al., 2007).

5 Conclusions and outlook

Compared to process-based models, the RF models offer superior scalability and reusability. Discrete and expensive to measure input data (e.g. soil bulk density) is required in order to parameterize and validate process-based models successfully (Clark et al., 2017; Yilmaz et al., 2008). Such models can only be used in areas with sufficient and freely available input data, which limits their applicability for regional level modeling. On the other hand, input data for ML is easier to obtain as there are abundant open data sources available (e.g. LULC from satellite imagery) for extracting predictors. Thus, our models are applicable in regions, where data availability is insufficient for process-based solutions. Additionally, the use of XAI techniques such as SHAP enable to extract insight (e.g. feature importance) needed for understanding the specific forces affecting WQ.

Some of the uncertainty in our RF models is due to gaps in the KESE (Estonian Environment Agency, 2021) WQ time series. In most of the catchments the annual mean concentrations were calculated based on a limited number of monthly samples. Great gains in model performance could be made with improving sampling consistency. The improvement and harmonization of national hydrological datasets is one of the focus areas of the European Union Water Framework Directive (Brack et al., 2017). We hope, that this will increase the international applicability of ML models in the near future.

Our models are currently able to model only at the yearly level and cannot be used for predicting time series. However, there is potential for a finer temporal scale, provided there is sufficient input data. In addition to aforementioned improvements in WQ data, predictors would have to be adjusted accordingly. Although many of the predictors used in the models can be considered as static (e.g. elevation, soil), modeling on the seasonal or monthly level means that certain predictors (e.g. precipitation, temperature) must be calculated at corresponding intervals as well.

6 Acknowledgments

This research was funded by Mobilitas+ program grant no. MOBERC34, Marie Skłodowska-Curie Actions individual fellowship under the Horizon 2020 Programme grant agreement number 795625, NUTIKAS program and European Regional Development Fund (EcolChange Centre of Excellence). Holger Virro is also thankful for technical support from the the High Performance Computing Center of the University of Tartu.

References

Álvarez-Cabria, M., Barquín, J., and Peñas, F. J.: Modelling the spatial and seasonal variability of water quality for entire river networks: Relationships with natural and anthropogenic factors, Science of the Total Environment, 545, 152–162, https://doi.org/10.1016/j.scitotenv.2015.12.109, 2016.

Arheimer, B., Dahné, J., Donnelly, C., Lindström, G., and Strömqvist, J.: Water and nutrient simulations using the HYPE model for Sweden vs. the Baltic Sea basin–influence of input-

data quality and scale, Hydrology research, 43, 315–329, https://doi.org/10.2166/nh.2012.010, 2012.

Arnold, J. G., Srinivasan, R., Muttiah, R. S., and Williams, J. R.: Large area hydrologic modeling and assessment part I: model development, JAWRA Journal of the American Water Resources Association, 34, 73–89, https://doi.org/10.1111/j.1752-1688.1998.tb05961.x, 1998.

Barrow, N.: The effects of pH on phosphate uptake from the soil, Plant and soil, 410, 401–410, https://doi.org/10.1007/s11104-016-3008-9, 2017.

Bartley, R., Speirs, W. J., Ellis, T. W., and Waters, D. K.: A review of sediment and nutrient concentration data from Australia for use in catchment water quality models, Marine pollution bulletin, 65, 101–116, https://doi.org/10.1016/j.marpolbul.2011.08.009, 2012.

Bergstra, J. and Bengio, Y.: Random search for hyper-parameter optimization., Journal of machine learning research, 13, 2012.

Bhattacharjee, J., Marttila, H., Launiainen, S., Lepistö, A., and Kløve, B.: Combined use of satellite image analysis, land-use statistics, and land-use-specific export coefficients to predict nutrients in drained peatland catchment, Science of The Total Environment, 779, 146 419, https://doi.org/10.1016/j.scitotenv.2021.146419, 2021.

Brack, W., Dulio, V., Ågerstrand, M., Allan, I., Altenburger, R., Brinkmann, M., Bunke, D., Burgess, R. M., Cousins, I., Escher, B. I., Hernández, F. J., Hewitt, L. M., Hilscherová, K., Hollender, J., Hollert, H., Kase, R., Klauera, B., Lindim, C., López Herráez, D., Miège, C., Munthe, J., O'Toole, S., Posthuma, L., Rüdels, H., Schäfert, R. B., Sengl, M., Smedes, F., van de Meentv, D., van den Brink, P. J., van Gils, J., van Wezel, A. P., Vethaak, A. D., Vermeirssen, E., von der Ohe, P. C., and Vrana, B.: Towards the review of the European Union Water Framework Directive: recommendations for more efficient assessment and management of chemical contamination in European surface water resources, Science of the Total Environment, 576, 720–737, https://doi.org/10.1016/j.scitotenv.2016.10.104, 2017.

Clark, M. P., Bierkens, M. F., Samaniego, L., Woods, R. A., Uijlenhoet, R., Bennett, K. E., Pauwels, V., Cai, X., Wood, A. W., and Peters-Lidard, C. D.: The evolution of process-based hydrologic models: historical challenges and the collective quest for physical realism, Hydrology and Earth System Sciences, 21, 3427–3440, https://doi.org/10.5194/hess-21-3427-2017, 2017.

Ebeling, P., Kumar, R., Weber, M., Knoll, L., Fleckenstein, J. H., and Musolff, A.: Archetypes and Controls of Riverine Nutrient Export Across German Catchments, Water Resources Research, 57, e2020WR028 134, https://doi.org/10.1029/2020WR028134, 2021.

Edwards, A. and Withers, P.: Transport and delivery of suspended solids, nitrogen and phosphorus from various sources to freshwaters in the UK, Journal of Hydrology, 350, 144–153, https://doi.org/10.1016/j.jhydrol.2007.10.053, 2008.

Esri: Arc Hydro GIS for Water Resources, [software], https://www.esri.com/en-us/industries/water-resources/arc-hydro, last accessed Feb 18, 2022, 2020.

Estonian Environment Agency: Keskkonnaseire infosüsteem KESE, [dataset], https://kese.envir.ee/kese/welcome.action, last accessed Mar 6, 2022, 2021.

Estonian Land Board: Elevation Data, [dataset], https://geoportaal.maaamet.ee/eng/Spatial-Data/Elevation-Data-p308.html, last accessed Mar 6, 2022, 2020a.

Estonian Land Board: Estonian Topographic Database, [dataset], https://geoportaal.maaamet.ee/eng/Spatial-Data/Estonian-Topographic-Database-p305.html, last accessed Mar 6, 2022, 2020b.

Gentry, L., David, M., Royer, T., Mitchell, C., and Starks, K.: Phosphorus transport pathways to streams in tile-drained agricultural watersheds, Journal of Environmental Quality, 36, 408–415, https://doi.org/10.2134/jeq2006.0098, 2007.

Grömping, U.: Variable importance assessment in regression: linear regression versus random forest, The American Statistician, 63, 308–319, https://doi.org/10.1198/tast.2009.08199, 2009.

He, B., Kanae, S., Oki, T., Hirabayashi, Y., Yamashiki, Y., and Takara, K.: Assessment of global nitrogen pollution in rivers using an integrated biogeochemical modeling framework, Water research, 45, 2573–2586, https://doi.org/10.1016/j.watres.2011.02.011, 2011.

HELCOM: Eutrophication in the Baltic Sea: An Integrated Thematic Assessment of the Effects of Nutrient Enrichment in the Baltic Sea Region. Executive Summary, Helsinki Commission. Baltic Marine Environment Protection Commission, https://doi.org/10.13140/RG.2.1.2758.0564, 2009.

Ho, J. Y., Afan, H. A., El-Shafie, A. H., Koting, S. B., Mohd, N. S., Jaafar, W. Z. B., Sai, H. L., Malek, M. A., Ahmed, A. N., Mohtar, W. H. M. W., Elshorbagy, A., and El-Shafie, A.: Towards a time and cost effective approach to water quality index class prediction, Journal of Hydrology, 575, 148–165, https://doi.org/10.1016/j.jhydrol.2019.05.016, 2019.

Hollaway, M., Beven, K., Benskin, C., Collins, A., Evans, R., Falloon, P., Forber, K., Hiscock, K., Kahana, R., Macleod, C., Ockenden, M., Villamizar, M., Wearing, C., Withers, P., Zhou, J., Barber, N., and Haygarth, P.: The challenges of modelling phosphorus in a headwater catchment: Applying a 'limits of acceptability' uncertainty framework to a water quality model, Journal of Hydrology, 558, 607–624, https://doi.org/10.1016/j.jhydrol.2018.01.063, 2018.

Hooda, P. S., Edwards, A. C., Anderson, H. A., and Miller, A.: A review of water quality concerns in livestock farming areas, Science of the total environment, 250, 143–167, https://doi.org/10.1016/S0048-9697(00)00373-9, 2000.

Kmoch, A., Kanal, A., Astover, A., Kull, A., Virro, H., Helm, A., Pärtel, M., Ostonen, I., and Uuemaa, E.: EstSoil-EH: a high-resolution eco-hydrological modelling parameters dataset for Estonia, Earth System Science Data, 13, 83–97, https://doi.org/10.5194/essd-13-83-2021, 2021.

Kuo, Y.-M., Liu, C.-W., and Lin, K.-H.: Evaluation of the ability of an artificial neural network model to assess the variation of groundwater quality in an area of blackfoot disease in Taiwan, Water research, 38, 148–158, https://doi.org/10.1016/j.watres.2003.09.026, 2004.

Lindström, G., Pers, C., Rosberg, J., Strömqvist, J., and Arheimer, B.: Development and testing of the HYPE (Hydrological Predictions for the Environment) water quality model for different spatial scales, Hydrology research, 41, 295–319, https://doi.org/10.2166/nh.2010.007, 2010.

Lintern, A., Webb, J., Ryu, D., Liu, S., Waters, D., Leahy, P., Bende-Michl, U., and Western, A.: What are the key catchment characteristics affecting spatial differences in riverine water quality?, Water Resources Research, 54, 7252–7272, https://doi.org/10.1002/wat2.1260, 2018.

Liu, X., Wang, Y., Li, Y., Wang, M., Liu, J., Yin, L., Zuo, S., and Wu, J.: Riverine nitrogen export and its natural and anthropogenic determinants in a subtropical agricultural catchment, Agriculture, Ecosystems & Environment, 301, 107 021, https://doi.org/10.1016/j.agee.2020.107021, 2020.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I.: From local explanations to global understanding with explainable AI for trees, Nature Machine Intelligence, 2, 2522–5839, https://doi.org/10.1038/s42256-019-0138-9, 2020.

Malagó, A., Bouraoui, F., Vigiak, O., Grizzetti, B., and Pastori, M.: Modelling water and nutrient fluxes in the Danube River Basin with SWAT, Science of the Total Environment, 603, 196–218, https://doi.org/10.1016/j.scitotenv.2017.05.242, 2017.

Marzadri, A., Amatulli, G., Tonina, D., Bellin, A., Shen, L. Q., Allen, G. H., and Raymond, P. A.: Global riverine nitrous oxide emissions: The role of small streams and large rivers, Science of The Total Environment, 776, 145 148, https://doi.org/10.1016/j.scitotenv.2021.145148, 2021.

Me, W., Abell, J., and Hamilton, D.: Effects of hydrologic conditions on SWAT model performance and parameter sensitivity for a small, mixed land use catchment in New Zealand, Hydrology and Earth System Sciences, 19, 4127–4147, https://doi.org/10.5194/hess-19-4127-2015, 2015.

Moreno-Mateos, D., Mander, Ü., Comín, F. A., Pedrocchi, C., and Uuemaa, E.: Relationships between landscape pattern, wetland characteristics, and water quality in agricultural catchments, Journal of environmental quality, 37, 2170–2180, https://doi.org/10.2134/jeq2007.0591, 2008.

Muñoz Sabater, J.: ERA5-Land hourly data from 1950 to 1980, [dataset], https://doi.org/10.24381/cds.e2161bac, last accessed Mar 6, 2022, 2021.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É.: Scikit-learn: Machine learning in Python, the Journal of machine Learning research, 12, 2825–2830, 2011.

Peterjohn, W. T. and Correll, D. L.: Nutrient dynamics in an agricultural watershed: observations on the role of a riparian forest, Ecology, 65, 1466–1475, https://doi.org/10.2307/1939127, 1984.

Sarkar, A. and Pandey, P.: River water quality modelling using artificial neural network technique, Aquatic procedia, 4, 1070–1077, https://doi.org/10.1016/j.aqpro.2015.02.135, 2015.

Senf, C.: European forest disturbance map, [dataset], https://doi.org/10.5281/zenodo.4746129, last accessed Mar 7, 2022, 2021.

Sheikholeslami, R. and Hall, J. W.: A global assessment of nitrogen concentrations using spatiotemporal random forests, Hydrology and Earth System Sciences Discussions, pp. 1–30, https://doi.org/10.5194/hess-2021-618, 2022.

Shen, L. Q., Amatulli, G., Sethi, T., Raymond, P., and Domisch, S.: Estimating nitrogen and phosphorus concentrations in streams and rivers, within a machine learning framework, Scientific data, 7, 1–11, https://doi.org/10.1038/s41597-020-0478-7, 2020.

Singh, K. P., Basant, A., Malik, A., and Jain, G.: Artificial neural network modeling of the river water quality—a case study, Ecological modelling, 220, 888–895, https://doi.org/10.1016/j.ecolmodel.2009.01.004, 2009.

Smith, S., Swaney, D., Buddemeier, R., Scarsbrook, M., Weatherhead, M., Humborg, C., Eriksson, H., and Hannerz, F.: River nutrient loads and catchment size, Biogeochemistry, 75, 83–107, https://doi.org/10.1007/s10533-004-6320-z, 2005.

Statistics Estonia: PM0646: Nitrogen, phosphorus and potassium in livestock manure by county, [dataset], https://andmed.stat.ee/en/stat/majandus__pellumajandus__pellumajandussaaduste-tootmine__taimekasvatussaaduste-tootmine/PM0646, last accessed Mar 6, 2022, 2021.

Tang, T., Strokal, M., van Vliet, M. T., Seuntjens, P., Burek, P., Kroeze, C., Langan, S., and Wada, Y.: Bridging global, basin and local-scale water quality modeling towards enhancing water quality management worldwide, Current opinion in environmental sustainability, 36, 39–48, https://doi.org/10.1016/j.cosust.2018.10.004, 2019.

Tiyasha, T., Tung, T. M., and Yaseen, Z. M.: A survey on river water quality modelling using artificial intelligence models: 2000–2020, Journal of Hydrology, 585, 124 670, https://doi.org/10.1016/j.jhydrol.2020.124670, 2020.

Uuemaa, E., Kull, A., Mõisja, K., Nurm, H.-I., and Kmoch, A.: Dimensioning of riparian buffer zones in agricultural catchments at national level, in: EGU General Assembly Conference Abstracts, pp. EGU21–12 634, https://doi.org/10.5194/egusphere-egu21-12634, 2021.

Virro, H. and Kmoch, A.: Code supplement for the Estonian water quality modeling study, [software], https://doi.org/10.5281/zenodo.5910320, last accessed Mar 6, 2022, 2022.

Virro, H., Kmoch, A., Vainu, M., and Uuemaa, E.: Data for water quality modeling using ML in Estonia, [dataset], https://doi.org/10.5281/zenodo.6325312, last accessed Mar 6, 2022, 2022.

Visser, H., Evers, N., Bontsema, A., Rost, J., de Niet, A., Vethman, P., Mylius, S., van der Linden, A., van den Roovaart, J., van Gaalen, F., Knoben, R., and de Lange, H. J.: What drives the ecological quality of surface waters? A review of 11 predictive modeling tools, Water Research, 208, 117 851, https://doi.org/10.1016/j.watres.2021.117851, 2022.

Yang, Y.-Y. and Toor, G. S.: Stormwater runoff driven phosphorus transport in an urban residential catchment: Implications for protecting water quality in urban watersheds, Scientific reports, 8, 1–10, https://doi.org/10.1038/s41598-018-29857-x, 2018.

Yilmaz, K. K., Gupta, H. V., and Wagener, T.: A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model, Water Resources Research, 44, https://doi.org/10.1029/2007WR006716, 2008.