DISSERTATIONES
GEOGRAPHICAE
UNIVERSITATIS
TARTUENSIS
86

HOLGER VIRRO

Geospatial data harmonization and machine learning for large-scale water quality modelling

DISSERTATIONES GEOGRAPHICAE UNIVERSITATIS TARTUENSIS
86

DISSERTATIONES GEOGRAPHICAE UNIVERSITATIS TARTUENSIS
86

HOLGER VIRRO

Geospatial data harmonization and machine learning for large-scale water quality modelling

UNIVERSITY OF TARTU
Press
1632

Department of Geography, Institute of Ecology and Earth Sciences, Faculty of Science and Technology, University of Tartu, Estonia

Dissertation has been accepted for the commencement of the degree of *Doctor philosophiae* in Geoinformatics at the University of Tartu on August 25, 2022 by the Scientific Council of the Institute of Ecology and Earth Sciences, University of Tartu.

Supervisors: Associate Prof. Evelyn Uuemaa
Institute of Ecology and Earth Sciences
University of Tartu
Estonia

Dr Alexander Kmoch
Institute of Ecology and Earth Sciences
University of Tartu
Estonia

Opponent: Assistant Prof. Victor Francisco Rodriguez Galiano
Department of Physical Geography and Geographical Regional Analysis
University of Seville
Spain

Commencement: Senate Hall, University of Tartu, Ülikooli 18, Tartu, on December 8, 2022 at 14:15.

Publication of this dissertation is granted by the Institute of Ecology and Earth Sciences, University of Tartu.

ISSN 1406-1295 (print)
ISBN 978-9916-27-033-2 (print)
ISSN 2806-2302 (pdf)
ISBN 978-9916-27-034-9 (pdf)

University of Tartu Press
www.tyk.ee

TABLE OF CONTENTS

LIST OF ORIGINAL PUBLICATIONS

This thesis is based on the following publications, which are referred to in the text by Roman numerals:

I **Virro, H.**, Kmoch, A., Amatulli, G., Shen, L., Uuemaa, E. (2021) GRQA: Global River Water Quality Archive. *Earth System Science Data, 13*(12), 5483–5507. DOI: 10.5194/essd-13-5483-2021

II Kmoch, A., Kanal, A., Astover, A., Kull, A., **Virro, H.**, Helm, A., Pärtel, M., Ostonen, I., Uuemaa, E. (2021) EstSoil-EH: a high-resolution eco-hydrological modelling parameters dataset for Estonia. *Earth System Science Data, 13*(1), 83–97. DOI: 10.5194/essd-13-83-2021

III **Virro, H.**, Kmoch, A., Vainu, M., Uuemaa, E. (2022) Random forest-based modeling of stream nutrients at national level in a data-scarce region. *Science of The Total Environment*, 156613.
DOI: 10.1016/j.scitotenv.2022.156613

Author's contribution to the articles denoted by: '*' a minor contribution, '**' a moderate contribution, '***' a major contribution.

	Articles		
	I	II	III
Original idea	***	*	***
Study design	***	**	***
Data processing and analysis	***	**	***
Interpretation of the results	***	**	***
Writing the manuscript	***	*	***

LIST OF ABBREVIATIONS

CSV	Comma-Separated Values
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DEM	Digital Elevation Model
HYPE	Hydrological Predictions for the Environment
IQR	InterQuartile Range
MAPE	Mean Absolute Percentage Error
ML	Machine Learning
nMAD	normalized Median Absolute Deviation
R^2	coefficient of determination
RF	Random Forest
RMSE	Root Mean Squared Error
SD	Standard Deviation
SDG	Sustainable Development Goal
SHAP	SHapley Additive exPlanations
SOC	Soil Organic Carbon
SWAT	Soil and Water Assessment Tool
TN	Total Nitrogen
TP	Total Phosphorus
WQ	Water Quality
XAI	eXplainable Artificial Intelligence

1. INTRODUCTION

1.1. Global water quality issues

The quality of freshwater—one of the world's most important natural resources serving humans and ecosystems alike—has continued to deteriorate globally during the 21st century, hindering global efforts to improve water security (Gain et al. 2016; Downing et al. 2021). According to the 2021 report about the progress regarding UN Sustainable Development Goal (SDG) 6, the world is still far from ensuring access to clean water for all due to anthropological pressures fueled by population growth and intensifying agriculture (Mueller et al. 2012; Grizzetti et al. 2017; UN-Water 2021). Water quality (WQ) in rivers and streams is affected by various anthropogenic point source and non-point source pollutants. Point source pollution, e.g. wastewater from treatment plants is generally easier to detect and manage, since measurements can be made directly near the source (Ahearn et al. 2005). Most of the WQ degradation, however, is driven by non-point source or diffuse pollutants without a fixed outlet, such as agricultural production or urban runoff (Ouyang et al. 2019). Nutrient runoff from fertilizers used in agriculture is the leading cause of eutrophication in rivers (Bouwman et al. 2017; Vilmin et al. 2018; Ibáñez & Peñuelas 2019), while impervious surfaces and artificial drainage accelerate the transport of pollutants in urban areas (McGrane 2016; Y.-Y. Yang & Toor 2018; Newhart et al. 2019). The aforementioned pressures are magnified by the effects of climate change (Desmit et al. 2018; Sinha et al. 2019). Increasing temperatures intensify eutrophication (Whitehead et al. 2009; Giri 2021), higher amounts of rainfall and more intensive rainfall events speed up pollutant transport (Wagena & Easton 2018) and droughts can lead to an accumulation of nutrients during low streamflow conditions (van Vliet et al. 2013; L. Li et al. 2022). In order to better understand the issues affecting global WQ, scalable methods for quantifying the complex interactions between WQ and catchment characteristics are needed. Here, WQ modeling can provide support for decision-making and the development of effective mitigation measures and water management plans (Sivapalan & Blöschl 2017; Tang et al. 2019).

1.2. Environmental modeling

When modeling environmental phenomena, the value of a variable (the target) at a certain location can be predicted if the factors affecting it (predictors) at that location are known. Environmental parameters that are mainly dependent on local conditions, can be modeled based on independent variables in the proximity of the measurement site. For example, soil properties (e.g. soil organic carbon) in a particular plot of land are expected to be strongly linked to other soil properties (e.g. soil porosity and bulk density) in the direct vicinity of that plot (Suuster et al. 2011; Heung et al. 2014). In WQ modeling, certain indicators (e.g. dissolved oxygen) can be also be modeled based on known conditions (water temperature, pH etc.) of the site (Olyaie et al. 2017; Zhi et al. 2021). However, many parameters,

e.g. the concentration of nutrients—nitrogen (N) and phosphorus (P)—in water, are highly dependent on the combined environmental conditions (e.g. soil, land use and climate) of the contributing upstream area, meaning that data for the whole catchment is needed for accurate predictions at a specific site (Ahearn et al. 2005; Hasani Sangani et al. 2015).

1.3. Process-based water quality modeling

The traditional approach for modeling WQ parameters has been the use of process-based models, such as the Soil and Water Assessment Tool (SWAT) (Arnold et al. 1998; Malagó et al. 2017) or HYPE (Arheimer et al. 2012). In the process-based approach, the model is based on complex equations simulating the environmental processes affecting the target variable (Cuddington et al. 2013). In order to run a process-based model like SWAT, all parameters needed to simulate the processes have to be either derived from input data or replaced with default values (physical constants or coefficients), when corresponding data is missing. Furthermore, not all physical processes are necessarily fully understood and encoded in these models. These factors can result in significant uncertainty when applying a process-based modelling approach across large areas (Yilmaz et al. 2008; Clark et al. 2017). Therefore, the process-based approach is most effective in catchments, where the discrete input data (e.g. soil bulk density) needed for parameterizing and validating the model has good spatial coverage. The requirements for input data combined with the computational demands needed for calibrating the model in many catchments limits the scalability of process-based models. Instead, data-driven machine learning (ML) methods are increasingly used in large-scale (e.g. national, global) hydrological studies as they are able to capture interactions between variables even when the specific set of input parameters needed to run process-based models is not available (Tahmasebi et al. 2020). Rather than trying to simulate real-world processes mathematically, ML methods can infer the non-linear relationships between variables from a large number of data points fed into the model (Solomatine et al. 2009; Amato et al. 2020; Giri 2021). Compared to process-based models, ML models also do not have such tightly constrained requirements for the set of predictor variables used as input. This flexibility also allows to include additional and potentially informative predictors (e.g. stream density) in WQ ML models, which can not be used in process-based models without altering the model structure.

1.4. Machine learning in hydrology

In hydrological modeling, some of the earliest uses of ML were deep learning methods (Tiyasha et al. 2020), such as neural networks, which have shown to achieve results comparable to process-based models in predicting nutrient concentrations, sediments and oxygen content in rivers (Singh et al. 2009; Sarkar & Pandey 2015; García-Alba et al. 2019; Zhou 2020). Like process-based approaches, extensive computational resources are needed for training deep learning models (LeCun

et al. 2015; Schmidhuber 2015). In addition, deep learning currently only offers so-called black-box solutions, meaning that interpreting how the model reaches its prediction is not straightforward (Montavon et al. 2018; Reichstein et al. 2019). Those shortcomings have limited the adoption of such models in environmental sciences (Karpatne et al. 2018; C. Shen 2018; T. Xu & Liang 2021). There is a need for white-box or explainable artificial intelligence (XAI) solutions (Adadi & Berrada 2018; Dikshit & Pradhan 2021) in order to give reliable support to decision-making (e.g. pollution mitigation measures). As an alternative, more robust, tree-based ML techniques like random forest (RF) have shown to achieve a comparable performance to both process-based and deep learning models when applied to large environmental datasets (Hengl et al. 2018; Fox et al. 2020; Yaseen 2021). RF is considered to be resistant to overfitting, insensitive to noisy input data and skewed distributions (Breiman 2001; Caruana & Niculescu-Mizil 2006), which is why it is suitable for environmental modeling, where the distribution of samples is often sparse and uneven (J. Li et al. 2011; Tyralis et al. 2019). RF models have also been used successfully for predicting different WQ indicators, such as nutrient concentrations (Álvarez-Cabria et al. 2016; R. Wang et al. 2021), sediments (Al-Mukhtar 2019) and oxygen content (Kim et al. 2020). Moreover, the availability of large-scale geospatial datasets that can be used for extracting WQ predictors has continued to improve (Butler 2014; Hengl et al. 2017; McCabe et al. 2017; L. Chen & Wang 2018), which has allowed RF to be applied in the first continental and global WQ studies (L.Q. Shen et al. 2020; Toming et al. 2020; Sheikholeslami & Hall 2022).

1.5. Spatial representativeness of machine learning input data

Despite the aforementioned advancements, conducting large-scale WQ modeling studies has still been hindered by the lack of global WQ datasets with a sufficient spatiotemporal coverage (Pellerin et al. 2016; Blöschl et al. 2019). Although large-scale models cannot be used to study one specific catchment, they allow detecting general trends and relationships in WQ not possible with catchment-scale models (Gupta et al. 2014). In order to ensure predictive power over large areas, the data used for training a large-scale ML model has to be spatially representative, i.e. encompass the heterogeneity of the environmental conditions in different regions. If the predictors of a target variable do not capture a diverse set of conditions then the models might achieve a good accuracy when testing on a slice of the source data, but will ultimately fail when applied to other regions or the whole extent of the target domain (Gupta et al. 2014; Elmes et al. 2020). In the case of WQ, this means that the training data has to contain samples from a diverse (e.g. land use configurations, topography, soil types) selection of catchments (O'Hare et al. 2020; de Almeida et al. 2022). Unfortunately, monitoring networks have often been designed around logistical or economic considerations which leads to biased sampling that does not fully represent the heterogeneity of the environmental variables affecting WQ (R.O. Strobl & Robillard 2008). Geospatial sampling bias caused by suboptimal sampling practices is a common issue in hydrology (Wadoux

et al. 2019). For example, a single river can have good coverage in sampling in order to monitor the transport of pollutants from point sources, whereas many rivers with less anthropogenic influence are not sampled at all (Oudin et al. 2010; O'Hare et al. 2020). Like most environmental variables, WQ has a strong temporal variability in addition to the spatial variability, which means that the ability of long-term monitoring data to accurately describe changes in WQ over time is also dependent on the continuity of the sampling time series.

1.6. Data harmonization for large-scale water quality modeling

Taking into account the training data requirements for large-scale WQ models, there is a need for global WQ datasets that represent a variety of environments, while also having good temporal coverage. The publication of open national (Read et al. 2017) and global (Hartmann et al. 2019; International Centre for Water Resources and Global Change 2020) datasets has alleviated the problem somewhat, but issues regarding data scarcity in many regions remain (K. Chen et al. 2020; Giri 2021). However, the spatiotemporal coverage can be improved by combining and harmonizing the already existing large-scale datasets. Thus far, a major issue inhibiting the creation of multi-source datasets has been the lack of guidelines to follow when publishing open WQ data. This has resulted in inconsistencies in parameter naming conventions, units and other attributes, which makes merging data from different sources not straightforward (McMillan et al. 2012; Crochemore et al. 2020). Here, it would help to include statistical insights into the data (e.g. outliers, histograms, time series continuity) along with the proper documentation regarding units of measurement, sampling methods and supplementary details about the sampling location (name, catchment size etc.) when publishing the data (Plana et al. 2019; Peng et al. 2022). These additions would also allow to make reliable assessments about the validity and suitability of the data for ML purposes (Gudmundsson et al. 2018; Addor et al. 2020).

1.7. Improving understanding of water quality through machine learning

Although improving the spatiotemporal coverage and representativeness of WQ data makes it feasible to create large-scale WQ ML models, the models will likely not have uniform accuracy across space (Wadoux et al. 2019; Brus 2021). Still, assessing the performance of the ML models spatially can help identify, where and why predictions were less accurate. For example, catchments, where the model either significantly overestimated or underestimated the target value might have common characteristics (e.g. dominated by certain land use or soil type), so adding samples from similar areas to training data could improve the generalization capabilities of the model. In addition to giving insights into the representativeness of the training data, combining ML with explainable AI (XAI) techniques such as SHAP (Lundberg et al. 2020) can help to further increase

transparency of the model and identify the most relevant predictor variables—or features in ML terminology—influencing the variability of WQ. Although an ML method like RF is considered to be relatively insensitive to noisy features, reducing the amount of features can result in an improved performance and a more robust model (J. Li et al. 2011; Yaseen 2021). Additionally, feature reduction can improve the reusability of the model as a smaller feature set also reduces the amount on input data needed in future applications (Tahmasebi et al. 2020; Yaseen 2021). Therefore, investigating the effects of feature reduction on ML performance can help assess the applicability of the model in large-scale studies.

1.8. Research questions

The overall aim of the thesis is to improve and harmonize datasets for WQ modeling purposes and create an ML framework for suitable large-scale WQ modeling.

The following research questions were investigated:

- What are the challenges regarding harmonizing large-scale datasets for WQ modeling?
- How does the spatial representativeness of training data affect the modeling results?
- Can an ML-based WQ model be used on a national scale and what are the potential advantages ML has over process-based models?
- Does feature reduction improve the modeling performance of WQ ML models?

The challenges regarding harmonizing large-scale WQ datasets were investigated in **Article I** and **Article II** as part of the compilation of the Estonian soil database EstSoil-EH and Global River Water Quality Archive (GRQA). EstSoil-EH is a numerical soil database intended for ecohydrological modeling that is derived from the National Soil Map of Estonia. The compilation process involved converting text-based soil properties of over 750,000 soil units into a machine-readable format with more than 30 actionable numerical and categorical variables, modeling additional variables based on the physical soil attributes and incorporating supplementary attributes (e.g. topography) from other data sources. GRQA is a new WQ dataset improving the spatiotemporal coverage of open global WQ data, which is built from five existing datasets and contains 42 parameters and over 17 million measurements around the globe covering the 1898–2020 time period. The harmonization involved unifying the attributes, flagging outliers and detecting duplicate observations in areas of spatial overlap. In addition, supplementary time series statistics were calculated to assess the temporal coverage of WQ observations.

The effects of the spatial representativeness of training data on modeling performance were studied in **Article II** and **Article III**. In **Article II**, we used soil properties from EstSoil-EH to predict soil organic carbon (SOC) content and investigated the relationships between prediction accuracy and the land forms, where the SOC samples originated from. Spatial assessment in **Article III** involved

calculating the observed to predicted ratios for each catchment as proxy measures for residuals. The ratios were plotted on a map to visually identify potential spatial patterns in accuracy. Additionally, the relationship between prediction accuracy and catchment areas was studied.

Article III focused on building a scalable RF model for predicting nutrient concentrations in Estonian rivers. The study extended the modeling workflow developed for predicting SOC in **Article II** by adding a feature selection procedure focused on optimizing the number of relevant features by reducing collinearity between features based on correlation. Although used here in the context of WQ, the procedure can be applied for many other targets, including soil properties. In addition, the SHAP method was implemented to investigate the most important features and their effects on nutrients in Estonian rivers. The combination of feature selection and importance extraction helps reduce the amount of preprocessing needed for the predictors in future studies, thus improving the reusability and scalability of the models.

2. MATERIAL AND METHODS

2.1. Harmonization of large-scale datasets for water quality modeling

2.1.1. The EstSoil-EH dataset

Along land use and land cover (LULC), climate and topography, the hydraulic properties of soil are also some of the most important catchment characteristics affecting WQ as they control precipitation infiltration and surface runoff potential (X. Yang & Jin 2010; Ross et al. 2018). Good quality soil data is an important input for WQ modeling and soil texture (Sandström et al. 2020; Guo et al. 2021), bulk density (Liu et al. 2020), hydraulic conductivity (Mittelstet et al. 2019) and soil organic carbon (SOC) (Fabre et al. 2019) have all been linked to riverine WQ. Currently available global soil datasets such as the Harmonized World Soil Database (HWSD) v1.2 (Fischer et al. 2008) or SoilGrids250m (Hengl et al. 2017)—the latter of which is notably produced using a random forest ML —could be used for extracting soil properties as predictors in large-scale WQ models (Abbaspour et al. 2015). Nevertheless, finer spatial resolution would allow for more accurate predictions when conducting national-scale studies. Furthermore, some key WQ predictors (e.g. hydraulic conductivity) are not available in aforementioned soil datasets.

The EstSoil-EH dataset (**Article II**) is a numerical soil database intended to support national-scale ecohydrological ML modeling. As a basis, EstSoil-EH built upon the existing highly detailed digitized National Soil Map of Estonia (1 : 10,000) that had been created based on Soviet-era field mapping, where 75% of mapped units are smaller than 4 ha. The creation of EstSoil-EH involved extensive standardization in order to convert the descriptive text-based attributes in the old soil map into a machine-readable format as well as extracting supplementary variables from other datasets (**Article II**, Figure 1). The overview of the resulting attributes is along with the sources is given in Section 3.1.1.

2.1.2. Compilation of GRQA

Source datasets of GRQA. The source datasets used for compiling GRQA are given in Table 1. The GLObal RIver Chemistry database (GLORICH) (Hartmann et al. 2019) and Global Freshwater Quality Database GEMStat (International Centre for Water Resources and Global Change 2020) were the two existing global datasets, while the European Environment Agency's (EEA) Waterbase dataset (European Environment Agency 2020) had a semicontinental (Europe) spatial coverage. Additionally, data from the Canadian Environmental Sustainability Indicators program (CESI) (Environment and Climate Change Canada 2020) and the Water Quality Portal (WQP) (United States Geological Survey 2020) had a national extent. Within GRQA, over half of the total number of observations originated from WQP, which also had the longest time series and had data for 37 out of the 42 WQ parameters (**Article I**, Table 7) presented in GRQA.

Table 1: Source datasets used for compiling GRQA with their total number of observations, parameters and timeframe length in GRQA. All datasets were retrieved on November 16, 2020. Source: **Article I**, Table 1.

Dataset	Name	Data provider	Observations *n*	Timeframe	Parameters (source/ **GRQA**) *n*/***n***
CESI	Water quality in Canadian rivers	Environment Canada	30,457	2002–2018	8/**42**
GEMStat	Global Freshwater Quality Database	International Centre for Water Resources and Global Change	2,094,598	1950–2020	32/**42**
GLORICH	GLObal RIver Chemistry database	Institute of Geology of the University of Hamburg	3,231,797	1942–2011	26/**42**
Waterbase	Waterbase - Water Quality	European Environment Agency	306,332	2008–2018	15/**42**
WQP	USGS Water Quality Portal	Environmental Protection Agency	8,689,335	1898–2020	37/**42**

Metadata harmonization. GRQA consists of 42 parameters, focusing on some of the most frequently used WQ indicators (e.g. water temperature, oxygen, phosphorus, nitrogen and carbon compounds). The full list of GRQA parameters along with their descriptive statistics is given in **Article I**, Table 7. Ambiguities in parameter naming conventions, units and chemical forms are a common problem when combining WQ datasets (McMillan et al. 2012; Sprague et al. 2017) and standardization is often needed before merging. Here, the first step was mapping the different parameter names and codes present in source data to the common denominator presented in GRQA. Examples of some of the differences in parameter codes and names along with standardized versions are given in Table 2. Lookup tables created for mapping the parameters and used in subsequent preprocessing steps are provided in GRQA (files with naming pattern **_code_map.csv*).

With the exception of temperature (°C), pH and dissolved oxygen (%), parameters in GRQA are given in mg L^{-1}. However, before harmonization there were many cases, where the same parameter had been measured in different units depending on the source. In some cases, the units varied by magnitude (e.g. mg L^{-1}

Table 2: Examples of differences in parameter naming conventions in GRQA source datasets based on Nitrate Nitrogen (NO3N) and Total Suspended Solids (TSS).

Parameter code	Source code	Parameter name	Source name	Source
NO3N	NITRATE	Nitrate Nitrogen	NITRATE	CESI
NO3N	NO3N	Nitrate Nitrogen	Nitrate	GEMStat
NO3N	NO3	Nitrate Nitrogen	Nitrate concentration, dissolved	GLORICH
NO3N	CAS_14797-55-8	Nitrate Nitrogen	Nitrate	Waterbase
NO3N	00618	Nitrate Nitrogen	Nitrate, water, filtered, milligrams per liter as nitrogen	WQP
NO3N	00620	Nitrate Nitrogen	Nitrate, water, unfiltered, milligrams per liter as nitrogen	WQP
TSS	TSS	Total Suspended Solids	Total Suspended Solids	GEMStat
TSS	SPM	Total Suspended Solids	Suspended matter concentration	GLORICH
TSS	EEA_31-02-7	Total Suspended Solids	Total suspended solids	Waterbase
TSS	00400	Total Suspended Solids	Suspended solids, water, unfiltered, milligrams per liter	WQP

vs μL^{-1}), while others were reported in different chemical forms. For example, depending on the measurement method nitrate (NO_3) can be reported as NO_3 or its nitrogen form (NO_3-N), which also affects the corresponding observation values. In order to harmonize units in the aforementioned cases, conversion constants had to be used.

The formula for conversion constants was

$$x_2 = \frac{x_1 \times M_{x_2}}{n \times M_{x_1}} \tag{2.1}$$

where x_1 and x_2 are observation values before and after conversion, M is the corresponding molar mass and n the magnitude difference between source and converted unit. Some examples of unit conversion are given in Table 3. The full list of all unit conversion procedures is given in the **Article I**, Table A1.

Table 3: Examples of unit conversion from the chemical form in source data to the GRQA version. x_1 and x_2 are observation values before and after conversion, respectively. Source: **Article I**, Table 4.

Parameter code	Form	Source form	Unit	Source unit	x_1	M_{x_2}	n	M_{x_1}	x_2
TAN	N	NH3	mg/l	mg/l	0.106	14.007	1	17.031	0.087
NO2N	N	NO2	mg/l	mg/l NO2	0.024	14.007	1	46.005	0.007
NO3N	N	NO3	mg/l	μmol/l	210.268	14.007	1000	62.004	0.048
NH4N	N	NH4	mg/l	mg/l	0.063	14.007	1	18.039	0.049

Some instances of duplicate site IDs were present in GLORICH and Waterbase, which would result in duplicate time series when merging the data. Site ID duplicates could indicate that were have been small shifts in the site location or that the site had been closed and reinstated at some point. The duplicate site pairs are provided in corresponding files in GRQA (e.g. *GLORICH_dup_sites*) and can be used to determine, which time series is more valid. In the case of WQP and CESI, coordinates had to be converted from the North American Datum of 1983 (NAD83) to World Geodetic System 1984 (WGS84). Where possible, source metadata about the quality of the observation was taken into account in order to remove implausible (e.g. negative) or poor quality (e.g. from unreliable source) observations and values outside detection limits are reported as such in GRQA (column *detection_limit_flag*). In addition, information about whether the sample had been filtered or not was retained as filtration can affect observation values in some parameters (Sprague et al. 2017). Observation dates were converted into a common format (%Y-%m-%d) and observations having invalid dates detected using the *datetime* Python package were removed. Where provided, measurement methods were also retained along with information about the catchment characteristics (size, drainage region) of the observation site.

Outlier flagging. The distribution of WQ observation values is often illustrated by a right skewed histogram, meaning that the peak is left of center due to a small number of very high values (McMillan et al. 2012; Sprague et al. 2017). Although some of these extreme outliers can be caused by data entry errors or faulty equipment, distinguishing errors based on peaks in a histogram might not be straightforward. For example, fertilizer spills or sudden increases in runoff due to floods can lead to short-term spikes in the amount of nutrients in river water (Hughes et al. 2016), resulting in outliers that should not be removed from data as they are valuable training data for ML models (Boldetti et al. 2010). For this reason, none of the outliers in GRQA were removed and were instead flagged based on the interquartile range (IQR) test (Figure 1), where IQR is defined as the difference between the third ($Q3$) and first ($Q1$) quartile. For a visual overview of distributions, histograms along with box and whisker plots were also produced for every parameter and are included in the GRQA data repository.

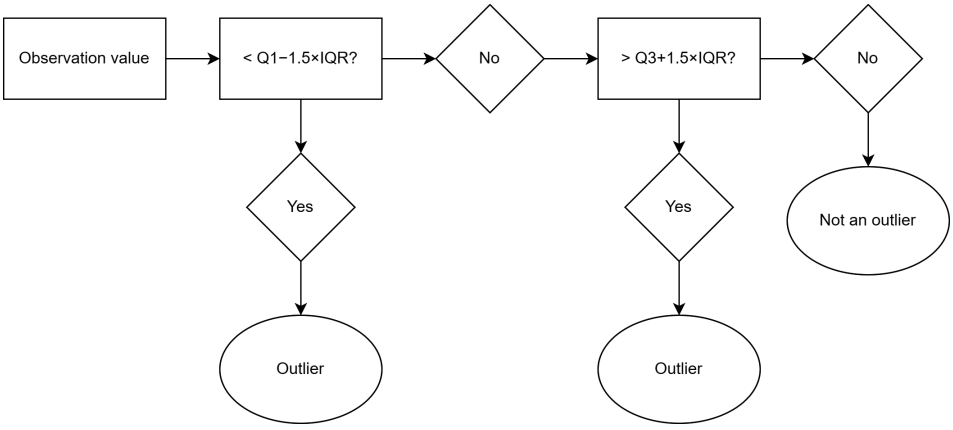

Figure 1: Workflow used for flagging outliers in GRQA based on interquartile range (*IQR*) test ($IQR = Q3 - Q1$).

Duplicate observation detection. Since GEMStat and GLORICH both had partial spatial overlap with the other source datasets, it is possible that the same site location could appear in multiple datasets under a different name or code. For example, WQP is also one of the source datasets of GLORICH, so it is likely that at least some of the observation time series would end up being duplicated after merging, which would result in falsely overrepresented areas. Here, rather than comparing site IDs, the duplicates had to be identified by spatial proximity and time series similarity.

The duplicate detection workflow is shown in Figure 2. First, site locations were clustered using the DBSCAN (density-based spatial clustering of applications with noise) algorithm (X. Xu et al. 1998) from the Scikit-learn Python library (Pedregosa et al. 2011) and setting a search radius of 1 km, which is an accuracy of around two decimal points in latitude/longitude degrees. DBSCAN was chosen as it has been previously applied to big data (Parimala et al. 2011; Khan et al. 2014) and can be run without setting the number of clusters beforehand (Birant & Kut 2007).

After the clusters consisting of potential duplicates had been created, the corresponding time series were compared based on root-mean-square error (RMSE). Only observations made on matching dates were used for calculating the RMSE and only pairs where RMSE was equal to zero were considered as potential duplicates. Finally, the duplicates were exported into separate CSV files (e.g. *TP_dup_obs.csv*) along with relevant metadata to help the user decide whether the sites can be considered duplicate. A high number of matching dates with the same observation value would indicate a higher likelihood of duplication.

Time series availability and continuity. The discontinuity of time series has been another issue previously reported when dealing with large-scale WQ data (Read et al. 2017; L.Q. Shen et al. 2020). As temporal coverage can affect the applicability of different modeling methods, monthly availability and monthly continuity statistics appropriated from Crochemore et al. (2020) were calculated

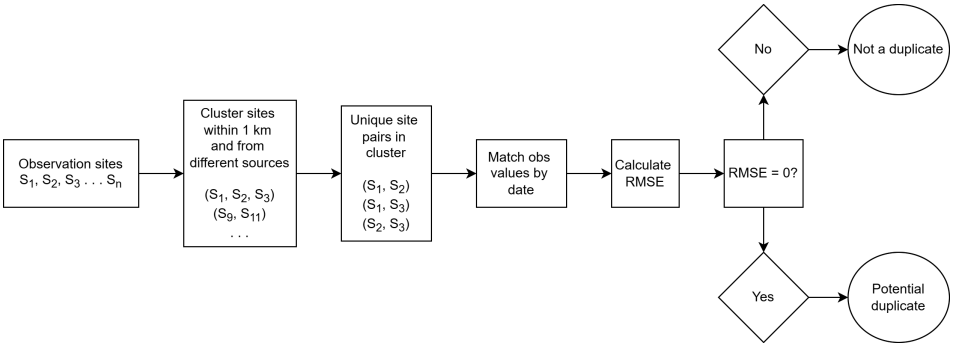

Figure 2: Workflow used for identifying potential duplicate time series in GRQA. The DBSCAN clustering algorithm was used for grouping nearby observation sites, while the similarity of the time series was determined based on RMSE.

for each site and parameter combination, to illustrate the suspected granularity of the time series. Monthly availability of observation data was defined as the ratio between number of months with at least one observation and the total number of months between the first and last observation. A ratio of 1.0 would mean that there was at least one observation in every month of the time series. Monthly continuity was calculated as the ratio between the longest period of consecutive months with any measurements and the length of time series in months. Here, a ratio of 1.0 would mean that there were no months without observations and the time series is continuous on a monthly level. The resulting characteristics were added as columns in the output files and were plotted on maps to identify areas with better or worse temporal coverage.

2.2. National-scale random forest-based modeling framework

2.2.1. Soil organic carbon prediction using random forest

Being part of the carbon cycle, SOC (% soil weight) is an important indicator of soil health, which is why it is important to map and monitor SOC under future climate and land use changes (Yigini & Panagos 2016; Vitharana et al. 2017). Prior to EstSoil-EH, SOC had not been mapped at national level in Estonia, which is why it was chosen as the initial target for building the ML framework that would form the basis of the national-scale WQ model. Training data consisted of SOC field measurements from soil units (Figure 3) located in peatlands (175 samples), arable fields (8,964), forests (100) and grasslands (446), amounting to 3,373 distinct point locations (Suuster et al. 2011; Kriiska et al. 2019). In the case of arable fields and peatlands many SOC samples had been collected from the same soil unit, meaning that the values had to be averaged over the whole unit to reduce bias in training data. As a result, the total number of distinct soil units used for training was reduced to 397.

Figure 3: Distinct soil unit polygons including all soil sampling locations for the SOC machine-learning training sample. Source: **Article II**, Figure 2.

The national-scale modeling framework for both SOC and WQ was developed using random forest (RF), which is an ensemble learning method based on combining a large number of regression trees (Breiman 2001; Loh 2011). RF was preferred over deep learning methods, since it has showed to yield reasonable accuracy when applied to big environmental datasets even when there is considerable noise (e.g. outliers, large number of redundant predictors) in training data (Fox et al. 2020; L.Q. Shen et al. 2020; Visser et al. 2022). The individual trees in RF have low correlation due to bootstrap aggregating, also known as bagging. During bootstrapping, new versions of training data are created by drawing random samples from the original training set for constructing the trees in RF (Bühlmann & Yu 2002; Prasad et al. 2006). Sampling is done with replacement, i.e. some of the samples in the bootstrap datasets are duplicated to match the size of the original training set. Random sampling in the bootstrapping phase leads to low bias and low variance in the overall RF model as predictions of the individual (low correlation, but high variance) trees are averaged in the aggregation phase of bagging, resulting in a model more resistant to overfitting (J. Li et al. 2011; Tyralis et al. 2019).

The RandomForestRegressor function from the Scikit-kearn Python package (Pedregosa et al. 2011) was used both for implementing RF for SOC **Article II** and WQ prediction **Article III**. For each of the soil units, soil attributes from the soil map were extracted as predictors. The data was randomly split into training (60%) and test (40%) sets and the model was validated by predicting SOC based on the

predictors in the test set. The resulting RF model was then used for predicting SOC into all soil units of Estonia. Modeling performance was evaluated based on the coefficient of determination (R^2) score as well as feature importances extracted using the built-in Gini importance (or mean decrease impurity) method available for RandomForestRegressor. The method calculates the importance of a feature based on the number of splits the feature was involved proportionally to the number of samples it splits, i.e. a feature that is part of many decisions in the trees making up the RF receives a high importance score (Menze et al. 2009). As additional measures of accuracy, RMSE and the normalized Median Absolute Deviation (nMAD) along with descriptive statistics of prediction errors were calculated per land form to examine whether the characteristics of soil units had an effect on performance.

2.2.2. Estonian water quality data

The aim of **Article III** was to model total nitrogen (TN) and total phosphorus (TP) concentrations in Estonian rivers during the period 2016–2020. Although two source datasets (Waterbase and GEMStat) are updated with new data originating from European country level databases on an irregular basis, the updates depend on the voluntary incentive of the corresponding national institutions. Therefore, WQ data from some of the smaller countries is often lacking and this was also the case with Estonia, where the nutrient sampling time series were insufficient for modeling. Instead, Estonian WQ data had to be obtained from the Estonian environment monitoring system (Keskkonnaseire Infosüsteem KESE) website maintained by the Environment Agency (Estonian Environment Agency 2021). Yearly mean concentrations were calculated for the 242 sites extracted from KESE (Figure 4), since sampling had been too sporadic for a more fine grained aggregation level (**Article III**, Figure 1). For each of the 242 sites, catchments were delineated from the 5 m resolution LiDAR DEM provided by the Estonian Land Board (Estonian Land Board 2020). The observation site locations and their corresponding catchment boundaries are presented along with median TN and TP concentrations in Figure 5.

2.2.3. Water quality predictor variables

The set of predictors used for WQ prediction in **Article III** consisted of 82 variables in total. The selection of predictors was based on what has been successfully used for TN and TP prediction in recent studies as well as the availability of the required source datasets in Estonia (**Article III**, Table S1). All predictors were aggregated based on the boundaries of the catchments delineated for each observation site location. Soil and topographic attributes from EstSoil-EH formed a significant part of the overall predictor set in the national-scale WQ model. Other than soil, topography (Ebeling et al. 2021; Lei et al. 2021) and LULC (Steidl et al. 2019; Dong et al. 2021; Song et al. 2021), which are generally considered as key predictors for WQ, variables describing the hydrology, climate, agriculture (fertilizer deposition and livestock) and geology of the catchments were included

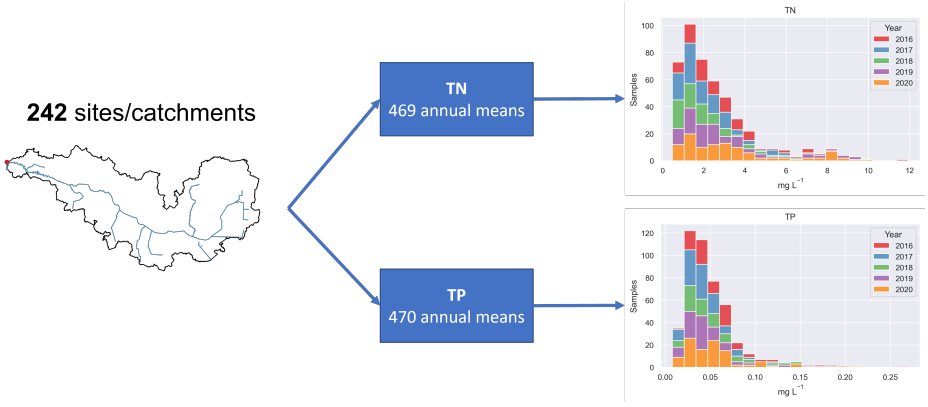

Figure 4: Distributions of yearly mean TN and TP concentrations in the 242 observation sites extracted from KESE.

in the predictor set (Table 4). In addition to a unique code, each predictor also was assigned a subcode, which was used during the feature selection procedure (Section 2.2.4).

Table 4: List of WQ predictor variables used in the national-scale RF model. Source: **Article III**, Table A1.

Category	Subcode	Description
topography	dem	Elevation
topography	tri	Terrain ruggedness index (TRI)
topography	twi	Topographic wetness index (TWI)
topography	flowlength	Flow length in the catchment
topography	slope	Slope
soil	awc1	Water holding capacity (first layer)
soil	bd1	Bulk density (first layer)
soil	clay1	Clay content (first layer)
soil	k1	Hydraulic conductivity (first layer)
soil	rock1	Rock content (first layer)
soil	sand1	Sand content (first layer)
soil	silt1	Silt content (first layer)
soil	soc1	Soil organic carbon (SOC) content (first layer)
LULC	arable	Proportion of arable land
LULC	forest	Proportion of forest
LULC	grassland	Proportion of grassland
LULC	other	Proportion of other LULC

Table 4 Continued.

Category	Subcode	Description
LULC	urban	Proportion of urban land
LULC	water	Proportion of water
LULC	wetland	Proportion of wetland
LULC	arable	Proportion of arable land within 100/500/1000 m stream buffer
LULC	forest_disturb	Proportion of disturbed forest area within 100/500/1000 m stream buffer
LULC	rip_veg	Total area of riparian vegetation buffer around natural streams/drainage ditches divided by catchment area
hydrology	area	Area of the catchment
hydrology	stream_density	Stream density
hydrology	pol_sen	Proportion of riparian buffer moderately/highly/very highly sensitive to pollution around drainage ditches/natural streams
agriculture	livestock_density	Density of livestock
agriculture	manure	Mean deposition of nitrogen/phosphorus in manure
climate	precip	Mean annual total precipitation
climate	snow_depth	Mean annual snow depth
climate	temp	Maximum/mean/minimum annual temperature
geology	limestone	Proportion of catchment located on limestone

2.2.4. National-scale water quality modeling

Feature selection workflow. As mentioned in Section 2.2.1, the modeling framework used for predicting TN and TP concentrations in **Article III** was an extension of the workflow used for SOC prediction. Experiences and lessons learnt from developing the SOC model helped reiterate the workflow with the intention to create models that are the most optimal for large-scale WQ prediction in terms of the amount of features needed to achieve an accuracy comparable to similar process-based alternatives. Separate models were developed for the two nutrients, both of which used the same predictor set described in Section 2.2.3. As both TN and TP concentrations showed a right skewed distribution (Figure 4) and the time series were fragmented due to infrequent sampling (**Article III**, Figure 1), the robustness of RF makes it a suitable method for WQ prediction in Estonia.

Compared to the SOC model, a major difference in this RF workflow (Figure 6) was the addition of a feature selection procedure given in Figure 7. In ML terminology, predictor variables are also called features, i.e. measurable properties of the target variable (Chandrashekar & Sahin 2014). Feature selection is a method for deriving a subset of features that reduces the complexity of the model, while maximizing the performance and explainability (Guyon & Elisseeff 2003; J. Li et al. 2017). Additionally, it can eliminate redundant features (those that have similar

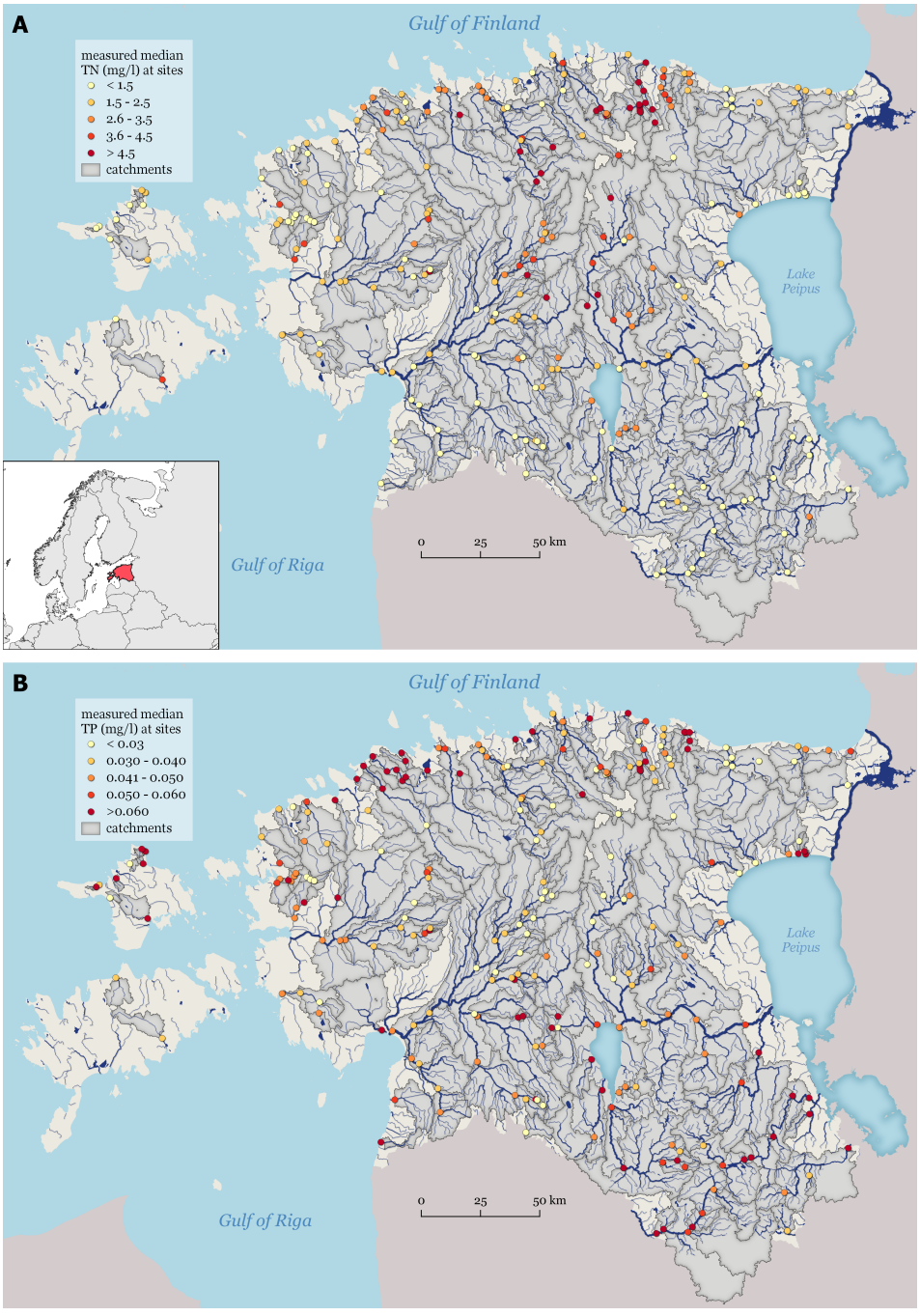

Figure 5: Median TN (A) and TP (B) concentration in observation sites 2016–2020. Source: **Article III**, Figure 3.

effects) by looking at correlations (Hall 1999). Although RF is not too sensitive to noisy features, reducing the number of features can result in an improved performance and a more robust model (J. Li et al. 2011; Yaseen 2021).

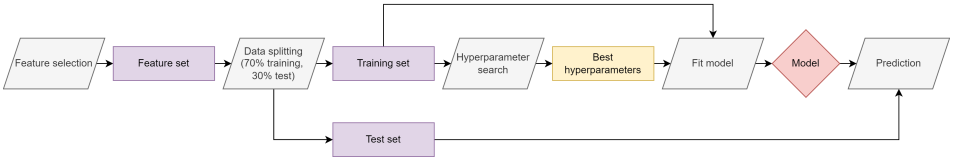

Figure 6: Workflow used for building the RF model for each feature set. Source: **Article III**, Figure 4.

First, pairwise correlations between all features along with correlations between features and the target (TN or TP concentration) were calculated. After that, feature pairs having correlation values above a set threshold were extracted. Finally, the feature having a lower correlation with the target was removed from each pair, provided that at least one feature from the corresponding subcode group (Table 4) remained in the set. This meant that features dem_max, dem_min and dem_std could be removed as long as dem_mean was retained due to being less collinear with the other features. Four different correlation thresholds were used for extracting the feature pairs, each resulting in a separate feature set (**Article III**, Table 3) and a corresponding RF model.

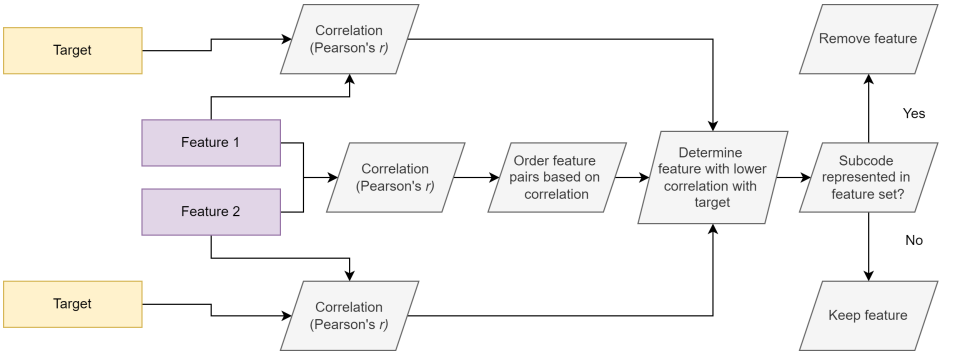

Figure 7: Workflow used for the feature selection strategy. Source: **Article III**, Figure 5.

Model evaluation. As with the SOC model, R^2 scores calculated for both the training (r2_train) and test (r2_test) sets were used to determine the accuracy of each model and feature set combination. In addition, mean absolute percentage error (MAPE) was calculated accordingly. 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.

Feature importances. Additionally, feature importances were derived from each model. However, instead of using the default Gini importance applied in the SOC model, the SHapley Additive exPlanations (SHAP) explainable AI (XAI) method from the corresponding Python package was implemented here. While Gini importance is faster to calculate, the method tends to prefer features with

high cardinality (i.e. having many distinct numerical values) and neglect feature correlations (C. Strobl et al. 2007; Grömping 2009). The SHAP method uses Shapley values from game theory to estimate how each feature contributes to the prediction and allows to investigate each sample's contribution to the prediction (Lundberg et al. 2020), which can be useful for detecting outliers that cause uncertainty in predictions. SHAP values show by how much (mg L^{-1} in the case of nutrients) and in which direction (increase or decrease) a prediction made based on the particular value of a feature differs from the general mean target value. Features accounting for more variance in predictions end up being more important. Due to being based on the number of splits per feature, Gini importance can only be used with tree-based models. SHAP, on the other hand, is model-agnostic, i.e. it can be applied to both tree-based models (D. Wang et al. 2022) and deep learning models (García & Aznarte 2020), which makes it useful for studies comparing different ML techniques.

Spatial assessment. The ratio between the observed and predicted values was also calculated for each catchment in order to assess spatial differences in model accuracy. This ratio indicates whether the model under (values greater than 1.0) or overestimated (lower than 1.0) the concentrations in a particular catchment. Here, catchments with lower accuracy could help identify potentially underrepresented regions or catchment types. In general, smaller catchments tend to be more uniform in environmental conditions (e.g. dominated by the same LULC class), which makes them less resistant to fluctuations in nutrient concentration (Lintern et al. 2018; Bhattacharjee et al. 2021). Moreover, as smaller catchments are often underrepresented in national monitoring networks, they can be the cause of uncertainty in model accuracy (Harmel et al. 2006). Hence, the relationship between aforementioned ratios and the size of the catchments was also investigated using scatter plots.

3. RESULTS

3.1. Harmonization of large-scale datasets

3.1.1. Overview of EstSoil-EH

EstSoil-EH is a numerical soil database intended to support national-scale ecohydrological modeling. The structure of the EstSoil-EH attributes is given in Table 5. The dataset consists of over 750,000 soil units that originate from the National Soil Map of Estonia. Each unit has attributes that describe its physical soil properties, which we derived by converting the text-based descriptive attributes of the National Soil Map into actionable numerical and categorical machine-readable values. The conversion was the most integral part during the creation of EstSoil-EH, since prior to numerical values the properties could not be used for modeling purposes. The physical soil properties were used as input for calculating additional attributes related to chemical properties (SOC) and soil hydrology, such as bulk density and soil hydraulic conductivity. The soil properties are accompanied by information about the topography, land use and drainage, which were derived from corresponding national datasets.

For each of the soil unit polygons in EstSoil-EH the following attributes were derived:

- The main soil properties (i.e. soil type, texture classes, fine earth fractions) were obtained from the National Soil Map using a conversion procedure described in **Article II**, Section 2.2.
- Topographic variables slope, topographic wetness index (TWI), terrain ruggedness index (TRI) and slope length and steepness (LS) factor derived from the Estonian 5 m digital elevation model (DEM).
- The LULC of each unit was derived from the Estonian Topographic Database (ETAK).
- Drainage information was extracted from the official register of drainage systems by the Agricultural Board of Ministry of Rural Affairs of Estonia.
- Soil organic carbon (SOC) was predicted using a random forest-based ML model described in Section 2.2.1.
- Bulk density (BD), which affects nitrification and mineralization processes in soil (Liu et al. 2020), was calculated based on the modeled SOC values.
- Hydraulic conductivity was calculated based on the EstSoil-EH fine earth fractions using the Rosetta3 software (Zhang & Schaap 2017).
- Available water capacity (AWC) was calculated from the EU-SoilHydroGrids dataset (Tóth et al. 2017).

Table 5: Description of variables and parameters available in the EstSoil-EH dataset. Source: **Article II**, Table 3.

Name of variable per mapped soil unit	Description
Name of variable per mapped soil unit	Description
est_soiltype	Estonian soil type
wrb_code	FAO WRB soil reference group (first and second level)
wrb_main	FAO WRB main soil reference group (first level)
est_txcode	reconstructed error-free interpretation of Estonian texture encoding description
nlayers	number of recognized layers/horizons
zmx	depth in millimetres: max depth of the sample-analysed soil profile in the mapped soil unit
z1-4	depth of each layer in millimetres (referring to bottom number) if nlayers indicates defined
est_txt1-4	Estonian texture class per layer number
lxtype1-4	USDA texture class
est_crs1-4	Estonian coarse fragment type
sand1-4	percentage mass of sand in fine earth fraction
silt1-4	percentage mass of silt in fine earth fraction
clay1-4	percentage mass of clay in fine earth fraction
rock1-4	volumetric content in percentage
soc1-4	SOC content percentage of soil weight
bd1-4	bulk density (g/cm^3)
k1-4	saturated hydraulic conductivity (mm/hr)
awc1-4	mm H_2O per millimetre of soil
slp_mean	mean slope (degrees), from DEM (also median and SD)
twi_mean	mean terrain wetness index (also median and SD)
ls_mean	LS factor (also median and SD)
tri_mean	terrain roughness index (also median and SD)
area_drain	area (m^2) per unit under a (e.g. tile) drainage regimen
drain_pct	percentage of the area of the soil unit under drainage
area_arable	area (m2) of LULC arable (six additional LULC types)
arable_pct	percentage of area that is LULC arable (six additional LULC types)
geometry	polygon, EPSG:3301 Estonian National Grid

3.1.2. Overview of GRQA

The structure of the GRQA is shown in **Article I**, Figure 1, while the full list of parameters and their descriptive statistics is given in **Article I**, Table 7. For each parameter, a separate CSV file is provided with observations and corresponding harmonized observation metadata. In addition, a data catalog (*GRQA_data_catalog.pdf*) with maps showing the spatiotemporal coverage and graphs describing the distribution of all 42 parameters is included in the repository. The structure of observation files is given in Table 6. Harmonized parameter codes, names, forms and units are reported along with the versions originally in source data, so that users could validate the conversion results. Where possible,

information about the site's upstream basin (area, drainage region) was retained as most tools for delineating catchments require the user to set a window size to be used as search radius during the delineation (Jasiewicz & Metz 2011).

Table 6: Summary table of output water quality observation file attributes. Source: **Article I**, Table 6.

Attribute name	Description	Data type
obs_id	Unique observation ID generated by hashing	string
lat_wgs84	Observation site latitude in WGS84	float
lon_wgs84	Observation site longitude in WGS84	float
obs_date	Observation date in the %Y-%m-%d format	string
obs_time	Observation time in the %H:%M:%S format	string
obs_time_zone	Observation time zone code	string
site_id	Observation site ID	string
site_name	Observation site name	string
site_country	Observation site country	string
upstream_basin_area	Site upstream basin area	string
upstream_basin_area_unit	Site upstream basin area unit	string
drainage_region_name	Drainage region where site is located in	string
param_code	Parameter code in GRQA	string
source_param_code	Parameter code in source dataset	string
param_name	Parameter name in GRQA	string
source_param_name	Parameter name in source dataset	string
obs_value	Observation value in GRQA	float
source_obs_value	Observation value in source dataset	float
detection_limit_flag	Whether a value was flagged as below (<) or above (>) detection limit in source data	string
param_form	Parameter chemical form in GRQA	string
source_param_form	Parameter chemical form in source dataset	string
unit	Parameter unit in GRQA	string
source_unit	Parameter unit in source dataset	string
filtration	Sample filtration information	string
source	Source dataset name	string
obs_percentile	Percentile of the observation value	float
obs_iqr_outlier	Flag to mark whether observation value is an outlier according to the interquartile range test	string
site_ts_availability	Monthly availability of the time series per site	float
site_ts_continuity	Monthly continuity of the time series per site	float
*_meta_*	Other observation metadata with a reference to the corresponding source column (e.g., GEMSTAT_meta_Method Description)	string
...	...	

As expected, many of the parameters are characterized by right skewed distributions (**Article I**, Figure 4), since outliers were flagged, rather than omitted (Section 2.1.2). Regarding outliers, no clear relationship between the parameter groups and the percentage of observations flagged as outliers was found (**Article I**, Table 7). However, the average percentage of outliers per parameter was 9.1%. Accord-

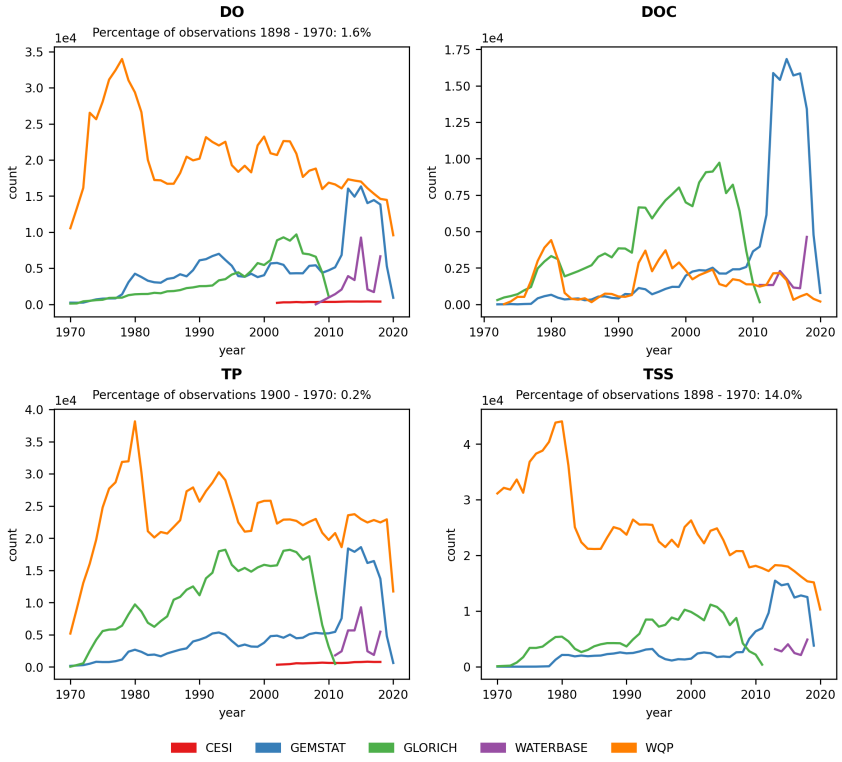

Figure 8: Temporal distribution of observations for dissolved oxygen (DO), dissolved organic carbon (DOC), total phosphorus (TP) and total suspended solids (TSSs) for the period 1970–2020. Source: **Article I**, Figure 6.

ing to the output files of the duplicate detection procedure (**_dup_obs.csv*), the parameters having the most duplicate observations were dissolved oxygen (DO), dissolved oxygen saturation (DOSAT) and temperature, which can be expected as these parameters are usually measured at higher intervals than others.

Temporal distribution of the observations of some example parameters—DO, dissolved organic carbon (DOC), TP and TSS— is shown in Figure 8. It can be seen that combining multiple sources can help extend the time series, e.g. the large amount of recent data from GEMStat can significantly improve the overall length of DOC time series. Although the spatial coverage of WQ observations improved in general (**Article I**, Figure 2), certain areas still have significantly better coverage (e.g. North America, most of Europe, Brazil), while empty regions remain in much of the developing world. In particular, the arid regions (North Africa, Central Asia) and tropical forests (Central Africa, Indonesia) are still underrepresented in GRQA.

Monthly continuity (Figure 9) plots used for illustrating the fragmented nature of time series indicated that observations originating from national datasets (CESI and WQP) had slightly better temporal coverage than others. Still, no clear spatial

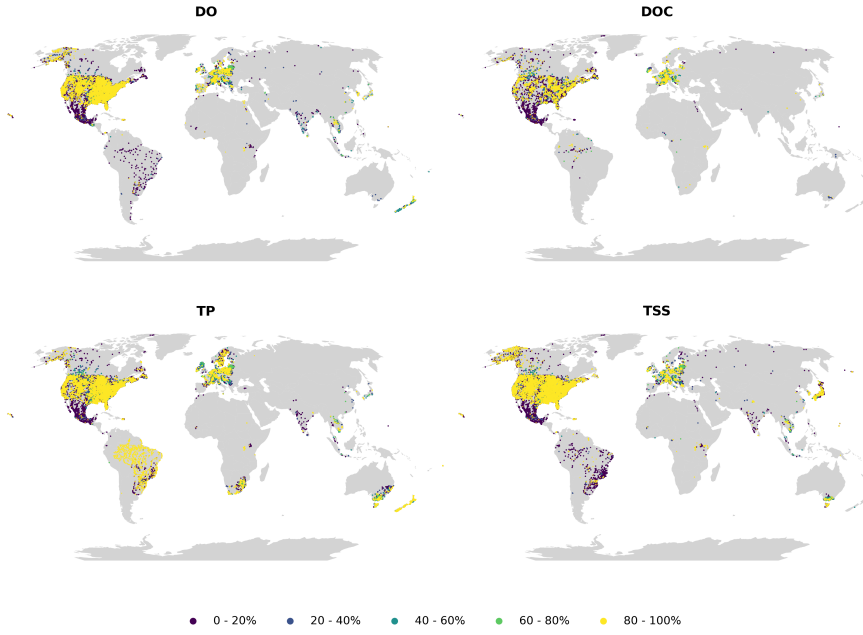

Figure 9: Monthly continuity for dissolved oxygen (DO), dissolved organic carbon (DOC), total phosphorus (TP) and total suspended solids (TSS). Source: **Article I**, Figure 6.

pattern could be identified. It has to be noted that due to how the statistic was calculated, short time series were more likely to have better continuity. For example, although TP time series were very short in Brazil, the continuity was among the highest.

3.2. National-scale random forest-based modeling framework

3.2.1. Soil organic carbon prediction

The results of the SOC prediction are discussed in **Article II**, Section 3.2. The RF-based model achieved an R^2 score of 0.69 (69%). More importantly, out of all soil attributes, clay content in the first soil layer (CLAY1) was found to be by far the most important based on the Gini importance scores. CLAY1 accounted for around 65% (importance score 0.65) of the overall prediction, while the next best features only had a marginal effect on the results, meaning that a reasonably accurate prediction could be made only using clay content as a predictor. The overwhelming importance of clay is consistent with previous findings as the clay fraction has an important role in stabilizing and retaining organic carbon within soil. Furthermore, a high clay-fraction was assigned for peat soils, in order to emulate hydraulic conditions in eco-hydrological models (Zhong et al. 2018; Prout et al. 2021).

The relationship between model accuracy and the land forms are shown in Table 7. Both nMAD and RMSE were larger in the case of soil units in grasslands and forests, which also had greater standard deviation (SD) than peatlands and arable fields. Considering that the original number of samples from grasslands (446) and forests (100) differed significantly, it is unlikely that the amount of training samples per class was the reason for the lower accuracy as the second best nMAD and RMSE values were detected in the case of peatlands, which had 175 samples. Arable fields, which had the most original samples (8,964), showed the lowest nMAD values. However, in many peatlands and arable fields the sampling points were located in the same polygon, so the value had to be averaged to reduce bias. Therefore, those classes were represented by more generalized values in training data, which could be one reason, why those classes were predicted better.

Table 7: Statistical description of SOC prediction error per land form. Source: **Article II**, Table 4.

Land form	Count	Min	Mean	Median	Max	SD	nMAD	RMSE
Wetland	150	−5.22	1.74	1.71	8.09	2.73	2.15	3.23
Arable	6675	−21.2	−1.54	−1	6.82	1.78	1.12	2.35
Forest	1299	−24.56	−2.08	−1.52	25.65	4.46	2.79	4.92
Grassland	74	−8.47	1.06	0.52	11.78	4.28	3.21	4.38

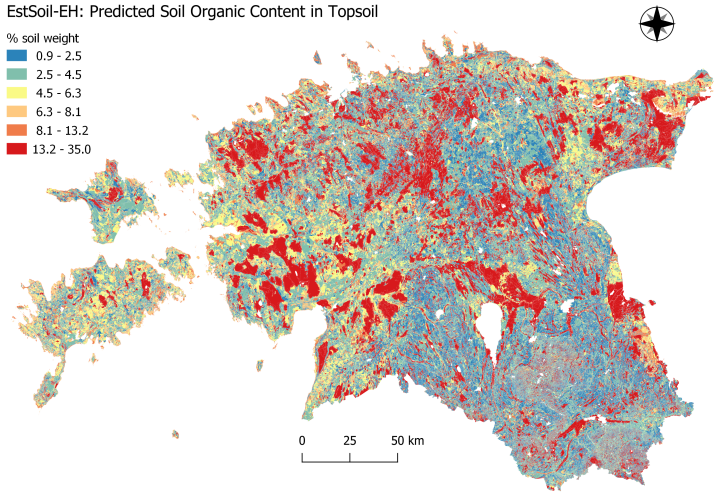

Figure 10: Predicted soil organic carbon (SOC) of the top soil layer. Source: **Article II**, Figure 5.

3.2.2. National-scale water quality modeling

Effects of feature reduction on machine learning performance. As explained in Section 2.2.4, each feature set derived from the feature selection procedure had a separate RF model built. The performance indicators for the four TN models are given in Table 8. Differences between the r2_train and r2_test scores indicate that TN models not as prone to overfitting as the TP models (Table 9). MAPE values were within the reasonable range in all cases. The best model was determined based on r2_test score, while minimizing the size of the feature set. As TN_MODEL_V4 managed to reach an accuracy of 0.83 (83%) with less than half (38) of the original features (82), it was deemed as the best.

Table 8: Performance indicators of the four model versions used for TN prediction. Source: **Article III**, Table 5.

Attribute	TN_MODEL_V1	TN_MODEL_V2	TN_MODEL_V3	TN_MODEL_V4
n_features	62	55	47	38
r2_train	0.88	0.94	0.92	0.93
r2_test	0.79	0.83	0.82	0.83
mape_train	0.21	0.16	0.17	0.17
mape_test	0.31	0.30	0.30	0.29

Regardless of the size of the feature set, all TP models showed significantly lower r2_test scores than the TN models (Table 9). TP models were also much more prone to overfitting, illustrated by the high r2_train scores. Still, MAPE scores were similar to the ones reported for TN. Based on the aforementioned criteria, TP_MODEL_V4 was selected as the best TP model, again showing that an optimal model could be built with only half of the features.

Table 9: Performance indicators of the four model versions used for TP prediction. Source: **Article III**, Table 7.

Attribute	TP_MODEL_V1	TP_MODEL_V2	TP_MODEL_V3	TP_MODEL_V4
n_features	64	56	47	40
r2_train	0.92	0.86	0.94	0.93
r2_test	0.50	0.52	0.48	0.52
mape_train	0.17	0.17	0.13	0.15
mape_test	0.27	0.26	0.27	0.26

Feature importances. Feature importances of the best nutrient models were calculated based on SHAP values. Unlike Gini importance scores, SHAP importances are given in the units of the corresponding feature (mg L^{-1}). In addition, SHAP values also indicate the direction of the contribution, i.e. which values of the features increased or decreased the target value. For TN, the most important features were related to arable land proportion (arable_prop) and soil rock content (rock1_mean), which both had a positive relationship with TN concentration, while

and hydraulic conductivity (k1_mean) and forest proportion (forest_prop) showed a negative correlation. In the TP model, high limestone content resulted in lower TP, while both grassland and urban proportion increased the concentration.

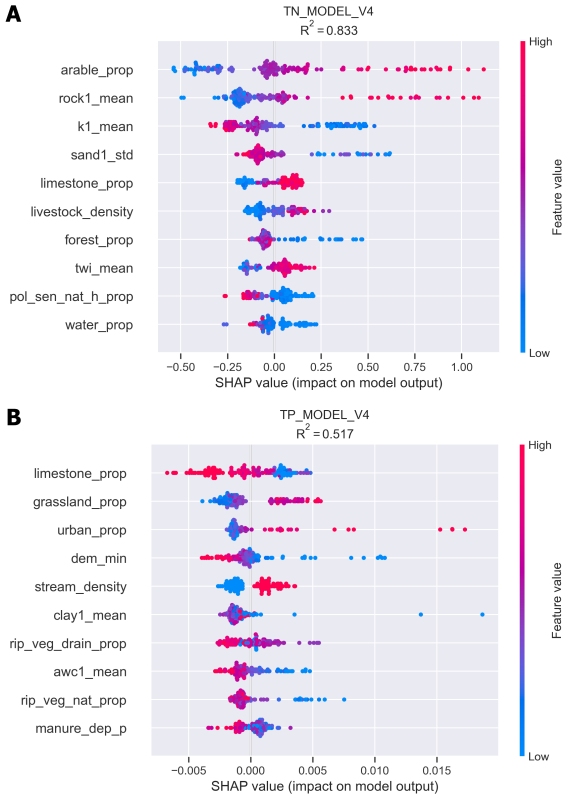

Figure 11: SHAP summary plots of the best models for TN (A) and TP (B). Here, each sample is colored by its corresponding feature value with higher values shown in pink and lower values in blue. Source: **Article III**, Figure 7.

Spatial assessment. For both TN and TP models, the observed to predicted ratios in around half of the overall catchments (242) were between 0.91 and 1.11, indicating that the models over or underestimated by around only 10% (Figure 12). Despite the lack of a discernible pattern in modeling performance, predictions seemed to be more accurate in the eastern (TN) and southeastern (TP) parts of mainland Estonia. Both models had a cluster of overestimations in the northeastern oil shale mining region, while scattered similarly overestimated areas where also present in the western half of the country.

The relationship between the previously calculated observed to predicted ratios and catchment sizes was also investigated (Fig. 13). In general, predicted values of samples from larger catchments match those of observed values reasonably well. On the other hand, smaller catchments showed significantly greater variance

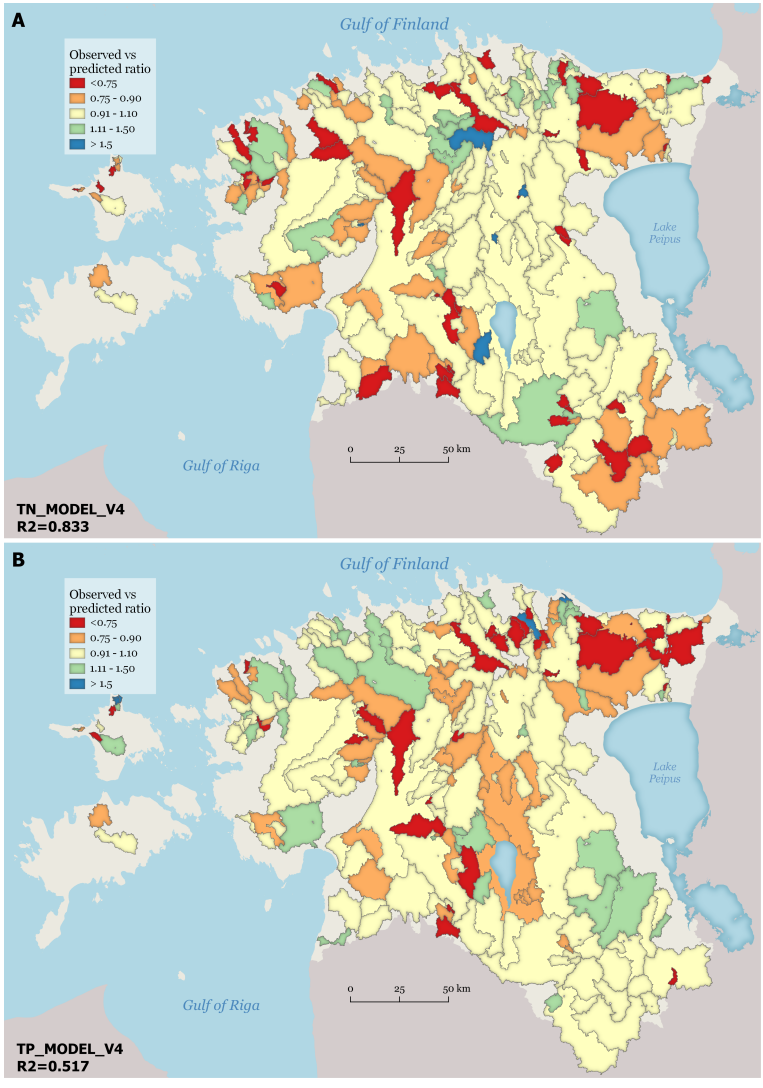

Figure 12: Spatial distribution of the ratio between observed and predicted values in catchments for the best TN (A) and TP (B) model. Overestimated concentrations are indicated by ratios lower than 1.0, while ratios greater than 1.0 show where the model underestimated the nutrient concentrations. Source: **Article III**, Figure 10.

in prediction accuracy and the majority of the biggest over (ratio below 0.5) and underestimations (ratio above 1.5) for both TN and TP were detected among the smallest catchments in the overall dataset.

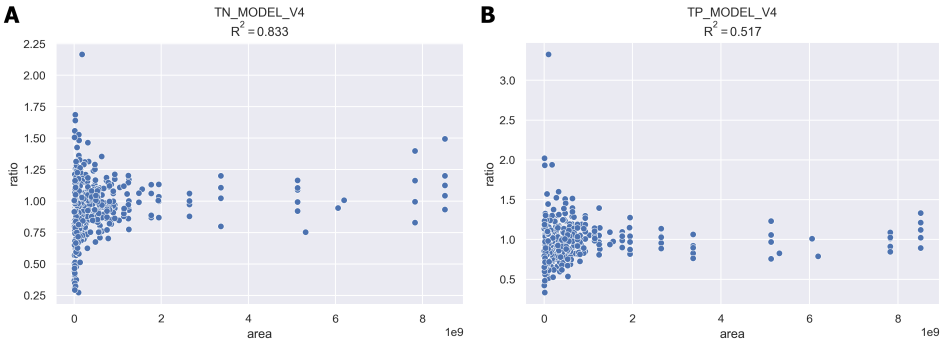

Figure 13: Relationship between catchment size and ratio of observed and predicted value for the best TN (A) and TP (B) model. Source: **Article III**, Figure 11.

4. DISCUSSION

4.1. Challenges regarding the harmonization of datasets for water quality modeling

Prior to the publication of the EstSoil-EH (**Article II**) there was no national-scale soil dataset in Estonia that could be used for modeling. Although the previously digitized Estonian Soil Map did have geometries for over 750,000 soil units, the soil properties of those units were mostly text-based and descriptive. Manual interpretation was needed to derive information about soil properties from the specialized soil types, which is only feasible for small-scale (e.g. single field) studies. Thus, it was not possible to use them as predictors in ML models and a major effort was made to convert the formerly descriptive soil properties into numerical machine-readable values (**Article II**, Section 2.2). The effort needed to translate the text-based properties into numerical ones highlights the importance of considering machine-readability and modern modeling requirements when creating national-scale datasets. Although different countries often have differences in soil classification schemes, good quality metadata published together with a well-documented reproducible workflow makes comparing and combining data originating from different countries much easier and allows to conduct cross-border studies. Nevertheless, there remains some uncertainty in some of the EstSoil-EH attributes, which has to be considered when using the data in models. As both SOC and hydraulic conductivity are parameters obtained through ML, using them as inputs in other ML models can result in additional uncertainty in predictions (Meyer et al. 2019). Finally, because soil properties in EstSoil-EH are attached to polygons as discrete values and we did not use them to predict a continuous surface, it cannot be directly compared to existing large-scale soil datasets (Hengl et al. 2017; Tóth et al. 2017) for validation purposes.

As in the case of GRQA source datasets, the diversity and ambiguity in parameter naming conventions, units and the chemical forms of parameters is one of the key problems when compiling multi-source WQ data (Sprague et al. 2017; Crochemore et al. 2020). For example, misinterpreting the chemical form or unit of a parameter can lead to an increase in outliers as mixing units varying in magnitude and forms with different molar masses can result in a falsely skewed distribution. In fact, since publishing **Article I** the GRQA data repository has been updated as the assumptions originally made regarding the N parameters with missing chemical form information in GLORICH had been incorrect, which resulted in inaccurate conversion constants that had to be fixed (Virro et al. 2021).

Different methods can also result in differences in observation values, which is why keeping corresponding metadata can help validate the results. However, there was too much variety in the method descriptions included in GRQA to classify them reasonably, so methods from different sources had to be included as separate columns. Although harmonizing should aim to reduce noise in metadata by restructuring and unifying the source attributes, retaining original metadata as supplementary material would help users in detecting such errors and, thus aid

data providers in actively improving the quality of the data. Another attribute included in GRQA is the upstream catchment area of the observation site, which unfortunately was only available in GLORICH. In order to relate WQ to conditions in the catchment, the boundary of the catchment is needed. As this boundary is not always available, it has to be delineated using corresponding geospatial software (Jasiewicz & Metz 2011). Knowing the estimated area beforehand would lessen ambiguity in the delineation process. Annotating data that is either incomplete or suspect is another aspect of good metadata governance (Gudivada et al. 2017). Good examples are the internal quality control measures used in GEMStat and Waterbase, which could help users eliminate some of the outliers during preprocessing. The adoption of such common machine-readable metadata quality measures would also help evaluate the usability and applicability of big WQ datasets for modeling purposes (Hutton et al. 2016; Stagge et al. 2019; Peng et al. 2022).

In many WQ parameters, observation values are characterized by skewed distributions due to outliers. Although some outliers can be either measurement or data entry errors, others are the result of fluctuations caused by pollution events. Thus, caution should be taken before removing outliers. For this reason, the outliers identified by the IQR test were flagged, rather than removed in GRQA to avoid discarding valuable information prematurely. The length and continuity of WQ time series are important indicators for determining, whether the data can be used as input for modeling. ML models cannot learn the interannual and seasonal variability of WQ from time series that are very fragmented. Therefore, we calculated time series statistics (availability and continuity) that can be used to assess the temporal quality of the WQ parameters.

The spatiotemporal analysis of GRQA parameters showed that although certain gains were made in the spatial and temporal coverage of global WQ data, both would benefit from significant improvements in the future. The importance of open WQ data has been emphasized as one of the keys to implement more efficient water management programs and pollution mitigation strategies (Brack et al. 2017; UN-Water 2021). However, the spatial distribution of WQ sites (**Article 1**, Figure 2) indicates that large areas with insufficient coverage remain in Africa and Asia, while others (Europe, North America) are characterized by dense monitoring networks. Similarly, the time series of WQ observations remain to be very discontinuous in many GRQA sites. In some of the sparsely covered regions, helping local institutions develop their own data management strategy by providing an example workflow when designing WQ pipelines, such as the schema recently proposed by Plana et al. (2019), could potentially lead to filling in some of the undersampled zones. Still, as it is up to the local institutions to adhere to an open data policy, gains in spatial coverage will likely be slow in the near future, despite increased collaboration efforts (Blöschl et al. 2019; Tang et al. 2019).

4.2. Effects of spatial representativeness on machine learning performance

4.2.1. Spatial assessment of soil organic carbon prediction

Although no discernible spatial pattern was detected in the accuracy of the RF models applied in this study, both SOC and nutrient models were still affected by the representativeness of the training data. In the case of SOC, the model seemed to perform better when predicting SOC in soil units located in peatlands and arable fields (Table 7), which both had instances, where soil units had multiple samples within one polygon, so the measured SOC values where averaged before training. This subdued the effect of potential outliers and as the averaged value is closer to the overall mean SOC content, the model would be able to replicate it better. Larger nMAD values were present in soil units originating from grasslands and forests. For the former, the issue could have simply been a lack of samples, as only a fraction of the reduced training data remained from grassland plots. Thus, the model did not manage to learn the variability of SOC in grasslands properly. On the other hand, forests represented a large part of the available training data. Here, the evaluation of representativeness is somewhat constrained by the generalization level of the land form types. Different types of forests are much more likely to appear on certain soil types, e.g. the pine-dominated coniferous forests are often growing on less fertile sandy soils. Having a more detailed classification scheme would have helped further narrow down the specific forest types, where the model was less accurate to decide whether undersampling could be one of the causes. In addition, the method that had been used for taking some of the SOC samples in forests did not remove the SOC found in litter from the measurement, which resulted in very high SOC content in some forest samples. The removal of litter would have decreased the number of outliers and, therefore, likely improved SOC predictions in forests. In addition, of the main 120–130 Estonian soil groups only about 70 were represented in the sampled training data, and different grassland and forest areas could be considered more diverse in general. Thus, the overall soil diversity was not captured well enough under the current sampling schemes.

With ML techniques becoming more common for soil mapping purposes (Grimm et al. 2008; Hengl et al. 2017), more attention is also paid to optimizing sampling in order to create representative training datasets (Wadoux et al. 2019). Investigating modeling performance in the aforementioned manner can provide valuable input for building a training set that better describes the full heterogeneity of interactions between soil and other environmental variables. In the case of SOC, Wadoux et al. (2019) showed that an optimal sampling design would take the spatial variability of predictor variables into account when planning sampling locations. Still, it has to be noted that different target variables depend on different predictors and a sampling design that is fine-tuned for SOC might not be optimal for predicting other soil properties (Brus 2021).

4.2.2. Spatial assessment of water quality prediction

As with the SOC model, neither of the RF models applied for nutrient concentration prediction showed clear spatial differences in performance, although the models were slightly more accurate in the southeastern part of Estonia (Figure 12). However, a clear relationship between catchment size and model performance existed with smaller catchments having more over and underestimations (Figure 13). Larger catchments tend to be more resistant to fluctuations in nutrient concentration, since they usually are more diverse in LULC, containing forests and wetlands, which adsorb some of the runoff before it reaches streams (Lintern et al. 2018). The more diverse environmental conditions in bigger catchments help even out fluctuations in WQ and reduce the amount of outliers in data. Smaller catchments, on the other hand, are less resilient as they are more likely to be dominated by a single LULC class and have more uniform soil and climate conditions, meaning that single pollution events have a greater effect on WQ (Bartley et al. 2012; Bhattacharjee et al. 2021). As a result, there tend to be more outliers, which are much more difficult for ML models to predict compared to the more generalized values of bigger catchments. Moreover, smaller catchments are generally underrepresented in monitoring networks, so the variability in their WQ time series might be missing from training data altogether (O'Hare et al. 2020).

However, the relationship between WQ in smaller headwater and larger downstream catchments can be nonlinear, i.e. nutrient concentrations may both increase and decrease with stream order as they are more dependent on the location of pollution sources in relation to the river, rather than river size (Oudin et al. 2010; Abbott et al. 2018). Therefore, rather than simply focusing on the size of catchments, the characteristics of the catchments with the largest over and underestimations could yield valuable input for designing a training dataset that better represents the relationships between WQ and the predictors (O'Hare et al. 2020). For example, clustering catchments based on the most important features (e.g. LULC classes) similarly to the procedure used for soil units can reveal potentially underrepresented feature configurations. Additionally, using SHAP to investigate the internal variability of features in relation to predictions can help detect potential outliers having a disproportionate impact on predictions. Because SHAP also shows the direction of the impact that features have on predictions and can be used to inspect predictions separately for each sample, it is more informative than Gini importance. Statistical tests (e.g. linear models) can also be used to determine, which variables are more unstable, i.e. need more focus when designing optimal sampling strategies (Levine et al. 2014).

4.3. Suitability of machine learning for national-scale water quality modeling

The performance of the RF models used for nutrient modeling showed good alignment with both process-based and ML models used in similar studies. The R^2 scores presented here are comparable to those previously shown by the HYPE

process-based models used in the Baltic region (Lindström et al. 2010; Arheimer et al. 2012). As expected, the best TN model achieved a significantly higher accuracy compared to the TP model. This has been the case with both process-based (Me et al. 2015; Malagó et al. 2017; Hollaway et al. 2018) and ML (Álvarez-Cabria et al. 2016; L.Q. Shen et al. 2020) nutrient models due to the processes affecting nitrogen being better understood, which makes selecting relevant predictors more straightforward. The most important features detected by SHAP for TN (Figure 11) were also consistent with what has been described in nutrient modeling studies with arable land, hydraulic properties of soil and livestock density being the main environmental variables affecting the amount of nitrogen in rivers (Hooda et al. 2000; He et al. 2011; Liu et al. 2020). Similarly, limestone (Barrow 2017), urban land (Lintern et al. 2018; Y.-Y. Yang & Toor 2018) and stream density (Ebeling et al. 2021) have shown to be effective predictors in TP models.

Taking into account the comparable accuracy to process-based models, the RF models used in this study have the advantage of being more applicable for large-scale studies. Process-based models are often limited to catchment level or at most regional scale studies, since the discrete input data required to parameterize and run the models lacks good spatial coverage and the computational resources are significant when applying them across many catchments (Yilmaz et al. 2008; Clark et al. 2017). RF models, on the other hand, have no fixed requirements for input data and are more flexible regarding the amount and relevance of the variables used as predictors (Tahmasebi et al. 2020; Giri 2021). Thus, RF models are more applicable in regions, where the specific input data needed for the process-based approach is lacking, but other data sources suitable for extracting predictors (e.g. LULC from satellite imagery) are abundant (McCabe et al. 2017; L. Chen & Wang 2018). Furthermore, the RF models used here for predicting annual TN and TP concentrations can be easily combined with streamflow data, which would allow to model annual N an P loads, providing valuable input to water management programs like HELCOM (HELCOM 2009), which monitors the eutrophication of Baltic Sea among other WQ indicators.

Still, the RF models showcased here could be developed further to increase their performance. It is likely that some loss in accuracy was due to the discontinuity of the time series used as training data (Estonian Environment Agency 2021). This meant that the annual TN and TP concentrations were calculated based on a very limited number of monthly samples in many sites, resulting in an incomplete representation of the interannual and seasonal variability of nutrient concentrations. Although the current models are built for predicting annual averages, improving the temporal consistency in sampling could potentially allow for seasonal or even monthly predictions. Finally, in order to apply these models across different countries, changes would have to be made in the source data used for extracting predictors. Most of the predictors used here were derived from national datasets, which might not have high resolution equivalents in other countries. Instead, existing continental or global datasets could be used to extract catchment characteristics to ensure better scalability, e.g. Corine Land Cover (CLC) for LULC (European Environment Agency 2018) or SoilGrids250m (Hengl

et al. 2017) for soil. However, considering that large-scale datasets are likely to have different classification levels for variables like soil properties or are products of ML themselves, additional uncertainty could be introduced when replacing input data (Meyer et al. 2019).

4.4. Effects of feature reduction on machine learning performance

The number of input features in an ML model can often be large and contain features that are actually not relevant, especially when the method used is relatively robust and resilient towards noisy features (e.g. RF). However, when dealing with big datasets, having too many redundant features can result in longer training times and demands on computation, limiting the wider applicability of the model, so reducing the number of features is often advisable. Using a more relevant feature set not only decreases the complexity of the model, but also increases its reusability, since the amount of preprocessing needed for predictors in the future is reduced (J. Li et al. 2011; Yaseen 2021).

The feature selection strategy implemented as part of the development of nutrient models (Figure 7) revealed that only half of the original features were required to reach a reasonable accuracy. For both TN (Table 8) and TP (Table 9), the best models were using the feature sets generated by eliminating many of the collinear features. It is likely that the number of features could be further reduced, as SHAP analysis showed that most features had only a marginal impact on the predictions. When interpreting the results of both feature selection and SHAP it has to be noted that seemingly redundant features detected here might be relevant when applying the models in other areas. For example, neither slope nor the climate variables seemed to have any impact on modeling results, although both are generally considered to significantly affect the transport and variability of nutrients in rivers (Mittelstet et al. 2019; Dong et al. 2021; Ebeling et al. 2021; Guo et al. 2021). Thus, as each study area is likely to have a uniquely optimal feature set, premature feature reduction should be avoided.

5. CONCLUSIONS

Large-scale modeling could provide valuable insight into global WQ issues. WQ modeling depends on input datasets with good spatial coverage, which are often not available, so they have to created by combining existing data sources. The majority of the challenges encountered during the creation of the two large-scale datasets in this study were related to metadata. Considerable effort was needed to convert the text-based soil properties from the existing National Soil Map of Estonia into the numerical database EstSoil-EH. The compilation of EstSoil-EH showed that although there might be valuable data sources available, their applicability is limited if the data is not machine-readable and needs significant domain knowledge to interpret. Thus, the numerical soil properties and supplementary attributes (topography, LULC, drainage) in EstSoil-EH significantly improved the usability of Estonian soil data for modeling purposes. Most of the issues regarding harmonizing existing WQ datasets for GRQA were related to ambiguities in parameter metadata. Differences in parameter naming conventions, codes, units and chemical forms have to be considered and attributes have to be properly mapped before combining observations from multiple sources. It is also advisable to retain metadata about the aforementioned attributes, since that allows users to detect potential conversion errors. Additionally, information about outliers and time series statistics can be useful for assessing the suitability of the data for modeling purposes. Although GRQA extended the spatiotemporal coverage of global WQ data, the adoption of common metadata guidelines and open data policies continue to be the key elements in advancing the coverage and quality of WQ data in the near future and, therefore, in helping develop more efficient water management programs worldwide.

In order to increase the generalization capabilities of training data, the ability to capture the variability of predictor variables has to be considered when designing monitoring networks. The spatial representativeness of training data had effects on modeling performance in both SOC and nutrient concentration prediction. For example, the assessment of prediction accuracy per land form showed that the predicted values were closer to the observed ones in wetlands and arable fields, which had very different sampling sizes within training data. Therefore, in addition to the amount of samples the variability within the predictor variables can affect modeling results and sampling networks should be designed with that in mind. Although no discernible spatial pattern existed in the accuracy of TN and TP models, a relationship between catchment area and prediction accuracy was identified, showing that concentrations in smaller catchments were more likely to be over or underestimated. Nutrient fluxes tend to have a greater effect in smaller catchments, resulting in higher variability in proportion to the catchment size and constituting land forms, which ML models are unable to capture. In larger catchments these single or extreme events are averaged out and the ML model can deal with them better. Therefore, additional sampling in small catchments would help to increase the variability captured by training data and, thus improve the predictive power of the ML model.

The process-based models traditionally used for WQ prediction have limited scalability, because they require discrete input data for parameterization that is not available in many areas. ML models, on the other hand, do not have such tightly constrained requirements for the variables used as predictors. Both nutrient concentration models developed in this study achieved an accuracy comparable to previous process-based models applied in the region as well as similar ML models used elsewhere. However, our developed ML models are more applicable in large-scale studies, since they lack the strict requirements for input data that process-based models have. Therefore, the RF models can be used for predicting annual mean nutrient concentrations in areas, where input data requirements for process-based approaches cannot be reasonably satisfied, but source data for extracting predictor variables (e.g. remote sensing products) is abundant. In addition, the models enable estimating nutrient losses at national and even regional level and they enable to capture the spatial variability of nutrient runoff better than existing process-based solutions. The SHAP method applied for extracting feature importances from the RF models has the benefit of showing both the size and direction of the impact each feature has on a particular prediction. Therefore, it is more informative than the Gini importance method used as default in RF.

Feature reduction is a method for reducing dimensionality in data by removing redundant features, which have a minimal effect on predictions. Although RF can generally handle noisy high-dimensional data well, the removal of redundant features makes the model less complex, which reduces training times. Furthermore, less preprocessing is required with a smaller feature set, which increases the applicability and reusability of the model in the future. The feature selection procedure implemented in the nutrient models revealed that the accuracy of the TN and TP models stayed stable during the step-by-step feature reduction. Reasonable performance was achieved with feature sets that were less than half the size of the original number of features considered (82). The benefit of the simple feature selection strategy used in our study over other well-known dimensionality reduction methods like principal component analysis is that it does not create any new features and feature values are kept as they appear originally. However, redundant features detected for one dataset might not match those in another one, so the feature selection procedure should be reapplied when the modeling framework is used in another area that differs in terms of geographic criteria.

REFERENCES

Abbaspour, K.C., Rouholahnejad, E., Vaghefi, S., Srinivasan, R., Yang, H., Kløve, B. A continental-scale hydrology and water quality model for Europe: Calibration and uncertainty of a high-resolution large-scale SWAT model. *Journal of Hydrology*, *524*, 733–752.

Abbott, B.W., Gruau, G., Zarnetske, J.P., Moatar, F., Barbe, L., Thomas, Z., Fovet, O., Kolbe, T., Gu, S., Pierson-Wickmann, A.-C., Davy, P., Pinay, G. Unexpected spatial stability of water chemistry in headwater stream networks. *Ecology letters*, *21*(2), 296–308.

Adadi, A., Berrada, M. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE access*, *6*, 52138–52160.

Addor, N., Do, H.X., Alvarez-Garreton, C., Coxon, G., Fowler, K., Mendoza, P.A. Large-sample hydrology: Recent progress, guidelines for new datasets and grand challenges. *Hydrological Sciences Journal*, *65*(5), 712–725.

Ahearn, D.S., Sheibley, R.W., Dahlgren, R.A., Anderson, M., Johnson, J., Tate, K.W. Land use and land cover influence on water quality in the last free-flowing river draining the western sierra nevada, california. *Journal of hydrology*, *313*(3-4), 234–247.

Al-Mukhtar, M. Random forest, support vector machine, and neural networks to modelling suspended sediment in tigris river-baghdad. *Environmental Monitoring and Assessment*, *191*(11), 1–12.

Álvarez-Cabria, M., Barquín, J., 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.

Amato, F., Guignard, F., Robert, S., Kanevski, M. A novel framework for spatio-temporal prediction of environmental data using deep learning. *Scientific reports*, *10*(1), 1–11.

Arheimer, B., Dahné, J., Donnelly, C., Lindström, G., 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*(4), 315–329.

Arnold, J.G., Srinivasan, R., Muttiah, R.S., Williams, J.R. Large area hydrologic modeling and assessment part i: Model development. *JAWRA Journal of the American Water Resources Association*, *34*(1), 73–89.

Barrow, N. The effects of ph on phosphate uptake from the soil. *Plant and soil*, *410*(1), 401–410.

Bartley, R., Speirs, W.J., Ellis, T.W., Waters, D.K. A review of sediment and nutrient concentration data from australia for use in catchment water quality models. *Marine pollution bulletin*, *65*(4-9), 101–116.

Bhattacharjee, J., Marttila, H., Launiainen, S., Lepistö, A., 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*, 146419.

Birant, D., Kut, A. St-dbscan: An algorithm for clustering spatial–temporal data. *Data & knowledge engineering*, *60*(1), 208–221.

Blöschl, G., Bierkens, M.F., Chambel, A., Cudennec, C., Destouni, G., Fiori, A., Kirchner, J.W., McDonnell, J.J., Savenije, H.H., Sivapalan, M., et al. Twenty-three unsolved problems in hydrology (uph)–a community perspective. *Hydrological Sciences Journal*, *64*(10), 1141–1158.

Boldetti, G., Riffard, M., Andréassian, V., Oudin, L. Data-set cleansing practices and hydrological regionalization: Is there any valuable information among outliers? *Hydrological Sciences Journal–Journal des Sciences Hydrologiques*, *55*(6), 941–951.

Bouwman, A., Beusen, A., Lassaletta, L., Van Apeldoorn, D., Van Grinsven, H., Zhang, J., van Ittersum, M. Lessons from temporal and spatial patterns in global use of n and p fertilizer on cropland. *Scientific reports*, *7*(1), 1–11.

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., ... 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.

Breiman, L. Random forests. *Machine learning*, *45*(1), 5–32.

Brus, D.J. Statistical approaches for spatial sample survey: Persistent misconceptions and new developments. *European Journal of Soil Science*, *72*(2), 686–703.

Bühlmann, P., Yu, B. Analyzing bagging. *The annals of Statistics*, *30*(4), 927–961.

Butler, D. Earth observation enters next phase. *Nature*, *508*(7495), 160–161.

Caruana, R., Niculescu-Mizil, A. An Empirical Comparison of Supervised Learning Algorithms. *Proceedings of the 23rd International Conference on Machine Learning*, 161–168.

Chandrashekar, G., Sahin, F. A survey on feature selection methods. *Computers & Electrical Engineering*, *40*(1), 16–28.

Chen, K., Chen, H., Zhou, C., Huang, Y., Qi, X., Shen, R., Liu, F., Zuo, M., Zou, X., Wang, J., Zhang, Y., Chen, D., Chen, X., Deng, Y., Ren, H. Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. *Water research*, *171*, 115454.

Chen, L., Wang, L. Recent advance in earth observation big data for hydrology. *Big Earth Data*, *2*(1), 86–107.

Clark, M.P., Bierkens, M.F., Samaniego, L., Woods, R.A., Uijlenhoet, R., Bennett, K.E., Pauwels, V., Cai, X., Wood, A.W., 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*(7), 3427–3440.

Crochemore, L., Isberg, K., Pimentel, R., Pineda, L., Hasan, A., Arheimer, B. Lessons learnt from checking the quality of openly accessible river flow data worldwide. *Hydrological Sciences Journal*, *65*(5), 699–711.

Cuddington, K., Fortin, M.-J., Gerber, L., Hastings, A., Liebhold, A., O'Connor, M., Ray, C. Process-based models are required to manage ecological systems in a changing world. *Ecosphere*, *4*(2), 1–12.

de Almeida, R.G.B., Lamparelli, M.C., Dodds, W.K., Cunha, D.G.F. Spatial optimization of the water quality monitoring network in são paulo state (brazil) to improve sampling efficiency and reduce bias in a developing sub-tropical region. *Environmental Science and Pollution Research*, *29*(8), 11374–11392.

Desmit, X., Thieu, V., Billen, G., Campuzano, F., Dulière, V., Garnier, J., Lassaletta, L., Ménesguen, A., Neves, R., Pinto, L., Silvestre, M., Sobrinho, J., Lacroix, G. Reducing marine eutrophication may require a paradigmatic change. *Science of the Total Environment*, *635*, 1444–1466.

Dikshit, A., Pradhan, B. Interpretable and explainable ai (xai) model for spatial drought prediction. *Science of the Total Environment*, *801*, 149797.

Dong, B., Qin, T., Wang, Y., Zhao, Y., Liu, S., Feng, J., Li, C., Zhang, X. Spatiotemporal variation of nitrogen and phosphorus and its main influencing factors in huangshui river basin. *Environmental monitoring and assessment*, *193*(5), 1–22.

Downing, J.A., Polasky, S., Olmstead, S.M., Newbold, S.C. Protecting local water quality has global benefits. *Nature Communications*, *12*(1), 1–6.

Ebeling, P., Kumar, R., Weber, M., Knoll, L., Fleckenstein, J.H., Musolff, A. Archetypes and controls of riverine nutrient export across german catchments. *Water Resources Research*, *57*(4), e2020WR028134.

Elmes, A., Alemohammad, H., Avery, R., Caylor, K., Eastman, J.R., Fishgold, L., Friedl, M.A., Jain, M., Kohli, D., Laso Bayas, J.C., Lunga, D., McCarty, J.L., Pontius Jr., R.G., Reinmann, A.B., Rogan, J., Song, L., Stoynova, H., Ye, S., Yi, Z.-F., Estes, L. Accounting for training data error in machine learning applied to earth observations. *Remote Sensing*, *12*(6), 1034.

Environment and Climate Change Canada *Water quality in canadian rivers* [Accessed on November 16]. Retrieved November 16, 2020, from https://open.canada.ca/data/en/dataset/55cc50dc-feb3-46d1-b40f-09254f3c00c5

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

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

European Environment Agency Corine Land Cover (CLC) 2018 [last accessed May 3, 2022]. https://land.copernicus.eu/pan-european/corine-land-cover/clc2018

European Environment Agency *Waterbase - water quality icm* [Accessed on November 16]. Retrieved November 16, 2020, from https://www.eea.europa.eu/data-and-maps/data/waterbase-water-quality-icm

Fabre, C., Sauvage, S., Tananaev, N., Espitalier Noël, G., Teisserenc, R., Probst, J., Sánchez Pérez, J. Assessment of sediment and organic carbon exports into the arctic ocean: The case of the yenisei river basin. *Water research*, *158*, 118–135.

Fischer, G., Nachtergaele, F., Prieler, S., Van Velthuizen, H., Verelst, L., Wiberg, D. Global agro-ecological zones assessment for agriculture (gaez 2008). *IIASA, Laxenburg, Austria and FAO, Rome, Italy*, *10*.

Fox, E.W., Ver Hoef, J.M., Olsen, A.R. Comparing spatial regression to random forests for large environmental data sets. *PloS one*, *15*(3), e0229509.

Gain, A.K., Giupponi, C., Wada, Y. Measuring global water security towards sustainable development goals. *Environmental Research Letters*, *11*(12), 124015.

García, M.V., Aznarte, J.L. Shapley additive explanations for no2 forecasting. *Ecological Informatics*, *56*, 101039.

García-Alba, J., Bárcena, J.F., Ugarteburu, C., Garcıa, A. Artificial neural networks as emulators of process-based models to analyse bathing water quality in estuaries. *Water research*, *150*, 283–295.

Giri, S. Water quality prospective in twenty first century: Status of water quality in major river basins, contemporary strategies and impediments: A review. *Environmental Pollution*, *271*, 116332.

Grimm, R., Behrens, T., Märker, M., Elsenbeer, H. Soil organic carbon concentrations and stocks on barro colorado island—digital soil mapping using random forests analysis. *Geoderma*, *146*(1-2), 102–113.

Grizzetti, B., Pistocchi, A., Liquete, C., Udias, A., Bouraoui, F., Van De Bund, W. Human pressures and ecological status of european rivers. *Scientific reports*, *7*(1), 1–11.

Grömping, U. Variable importance assessment in regression: Linear regression versus random forest. *The American Statistician*, *63*(4), 308–319.

Gudivada, V., Apon, A., Ding, J. Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations. *International Journal on Advances in Software*, *10*(1), 1–20.

Gudmundsson, L., Do, H.X., Leonard, M., Westra, S. The global streamflow indices and metadata archive (gsim)–part 2: Quality control, time-series indices and homogeneity assessment. *Earth System Science Data*, *10*(2), 787–804.

Guo, D., Liu, S., Singh, D., Western, A.W. Predicting quantiles of water quality from catchment characteristics. *Hydrological Processes*, *35*(1), e13996.

Gupta, H., Perrin, C., Blöschl, G., Montanari, A., Kumar, R., Clark, M., Andréassian, V. Large-sample hydrology: A need to balance depth with breadth. *Hydrology and Earth System Sciences*, *18*(2), 463–477.

Guyon, I., Elisseeff, A. An introduction to variable and feature selection. *Journal of machine learning research*, *3*(Mar), 1157–1182.

Hall, M.A. Correlation-based feature selection for machine learning.

Harmel, R., Cooper, R., Slade, R., Haney, R., Arnold, J. Cumulative uncertainty in measured streamflow and water quality data for small watersheds. *Transactions of the ASABE*, *49*(3), 689–701.

Hartmann, J., Lauerwald, R., Moosdorf, N. Glorich-global river chemistry database. *PANGAEA https://doi. org/10.1594/PANGAEA*, *902360*.

Hasani Sangani, M., Jabbarian Amiri, B., Alizadeh Shabani, A., Sakieh, Y., Ashrafi, S. Modeling relationships between catchment attributes and river water quality in southern catchments of the caspian sea. *Environmental Science and Pollution Research*, *22*(7), 4985–5002.

He, B., Kanae, S., Oki, T., Hirabayashi, Y., Yamashiki, Y., Takara, K. Assessment of global nitrogen pollution in rivers using an integrated biogeochemical modeling framework. *Water research*, *45*(8), 2573–2586.

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.

Hengl, T., Mendes de Jesus, J., Heuvelink, G.B., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M.N., Geng, X., Bauer-Marschallinger, B., Guevara, M.A., Vargas, R., MacMillan, R.A., Batjes, N.H., Leenaars, J.G., Ribeiro, E., Wheeler, I., Mantel, S., Kempen, B. Soilgrids250m: Global gridded soil information based on machine learning. *PLoS one*, *12*(2), e0169748.

Hengl, T., Nussbaum, M., Wright, M.N., Heuvelink, G.B., Gräler, B. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ*, *6*, e5518.

Heung, B., Bulmer, C.E., Schmidt, M.G. Predictive soil parent material mapping at a regional-scale: A random forest approach. *Geoderma*, *214*, 141–154.

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., 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.

Hooda, P.S., Edwards, A.C., Anderson, H.A., Miller, A. A review of water quality concerns in livestock farming areas. *Science of the total environment*, *250*(1-3), 143–167.

Hughes, A.O., Tanner, C.C., McKergow, L.A., Sukias, J.P. Unrestricted dairy cattle grazing of a pastoral headwater wetland and its effect on water quality. *Agricultural Water Management*, *165*, 72–81.

Hutton, C., Wagener, T., Freer, J., Han, D., Duffy, C., Arheimer, B. Most computational hydrology is not reproducible, so is it really science? *Water Resources Research*, *52*(10), 7548–7555.

Ibáñez, C., Peñuelas, J. Changing nutrients, changing rivers. *Science*, *365*(6454), 637–638.

International Centre for Water Resources and Global Change *Global water quality database gemstat* [Accessed on November 16]. Retrieved November 16, 2020, from https : //gemstat.org/data/data-portal/

Jasiewicz, J., Metz, M. A new grass gis toolkit for hortonian analysis of drainage networks. *Computers & Geosciences*, *37*(8), 1162–1173.

Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H.A., Kumar, V. Machine learning for the geosciences: Challenges and opportunities. *IEEE Transactions on Knowledge and Data Engineering*, *31*(8), 1544–1554.

Khan, K., Rehman, S.U., Aziz, K., Fong, S., Sarasvady, S. Dbscan: Past, present and future. *The fifth international conference on the applications of digital information and web technologies (ICADIWT 2014)*, 232–238.

Kim, S., Alizamir, M., Zounemat-Kermani, M., Kisi, O., Singh, V.P. Assessing the biochemical oxygen demand using neural networks and ensemble tree approaches in south korea. *Journal of Environmental Management*, *270*, 110834.

Kriiska, K., Frey, J., Asi, E., Kabral, N., Uri, V., Aosaar, J., Varik, M., Napa, Ü., Apuhtin, V., Timmusk, T., Ostonen, I. Variation in annual carbon fluxes affecting the soc pool in hemiboreal coniferous forests in estonia. *Forest Ecology and Management*, *433*, 419–430.

LeCun, Y., Bengio, Y., Hinton, G. Deep learning. *Nature*, *521*(7553), 436–444.

Lei, C., Wagner, P.D., Fohrer, N. Effects of land cover, topography, and soil on stream water quality at multiple spatial and seasonal scales in a german lowland catchment. *Ecological Indicators*, *120*, 106940.

Levine, C.R., Yanai, R.D., Lampman, G.G., Burns, D.A., Driscoll, C.T., Lawrence, G.B., Lynch, J.A., Schoch, N. Evaluating the efficiency of environmental monitoring programs. *Ecological Indicators*, *39*, 94–101.

Li, J., Heap, A.D., Potter, A., Daniell, J.J. Application of machine learning methods to spatial interpolation of environmental variables. *Environmental Modelling & Software*, *26*(12), 1647–1659.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R.P., Tang, J., Liu, H. Feature selection: A data perspective. *ACM computing surveys (CSUR)*, *50*(6), 1–45.

Li, L., Stewart, B., Zhi, W., Sadayappan, K., Ramesh, S., Kerins, D., Sterle, G., Harpold, A., Perdrial, J. Climate controls on river chemistry. *Earth's Future*, *10*(6), e2021EF002603.

Lindström, G., Pers, C., Rosberg, J., Strömqvist, J., Arheimer, B. Development and testing of the hype (hydrological predictions for the environment) water quality model for different spatial scales. *Hydrology research*, *41*(3-4), 295–319.

Lintern, A., Webb, J., Ryu, D., Liu, S., Waters, D., Leahy, P., Bende-Michl, U., Western, A. What are the key catchment characteristics affecting spatial differences in riverine water quality? *Water Resources Research*, *54*(10), 7252–7272.

Liu, X., Wang, Y., Li, Y., Wang, M., Liu, J., Yin, L., Zuo, S., Wu, J. Riverine nitrogen export and its natural and anthropogenic determinants in a subtropical agricultural catchment. *Agriculture, Ecosystems & Environment*, *301*, 107021.

Loh, W.-Y. Classification and regression trees. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, *1*(1), 14–23.

Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.-I. From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, *2*(1), 2522–5839.

Malagó, A., Bouraoui, F., Vigiak, O., Grizzetti, B., Pastori, M. Modelling water and nutrient fluxes in the danube river basin with swat. *Science of the Total Environment*, *603*, 196–218.

McCabe, M.F., Rodell, M., Alsdorf, D.E., Miralles, D.G., Uijlenhoet, R., Wagner, W., Lucieer, A., Houborg, R., Verhoest, N.E., Franz, T.E., Shi, J., Gao, H., Wood, E.F. The future of earth observation in hydrology. *Hydrology and earth system sciences*, *21*(7), 3879–3914.

McGrane, S.J. Impacts of urbanisation on hydrological and water quality dynamics, and urban water management: A review. *Hydrological Sciences Journal*, *61*(13), 2295–2311.

McMillan, H., Krueger, T., Freer, J. Benchmarking observational uncertainties for hydrology: Rainfall, river discharge and water quality. *Hydrological Processes*, *26*(26), 4078–4111.

Me, W., Abell, J., 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*(10), 4127–4147.

Menze, B.H., Kelm, B.M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., Hamprecht, F.A. A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC bioinformatics*, *10*(1), 1–16.

Meyer, H., Reudenbach, C., Wöllauer, S., Nauss, T. Importance of spatial predictor variable selection in machine learning applications–moving from data reproduction to spatial prediction. *Ecological Modelling*, *411*, 108815.

Mittelstet, A.R., Gilmore, T.E., Messer, T., Rudnick, D.R., Heatherly, T. Evaluation of selected watershed characteristics to identify best management practices to reduce nebraskan nitrate loads from nebraska to the mississippi/atchafalaya river basin. *Agriculture, Ecosystems & Environment*, *277*, 1–10.

Montavon, G., Samek, W., Müller, K.-R. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, *73*, 1–15.

Mueller, N.D., Gerber, J.S., Johnston, M., Ray, D.K., Ramankutty, N., Foley, J.A. Closing yield gaps through nutrient and water management. *Nature*, *490*(7419), 254–257.

Newhart, K.B., Holloway, R.W., Hering, A.S., Cath, T.Y. Data-driven performance analyses of wastewater treatment plants: A review. *Water research*, *157*, 498–513.

O'Hare, M.T., Gunn, I.D., Critchlow-Watton, N., Guthrie, R., Taylor, C., Chapman, D.S. Fewer sites but better data? optimising the representativeness and statistical power of a national monitoring network. *Ecological Indicators*, *114*, 106321.

Olyaie, E., Abyaneh, H.Z., Mehr, A.D. A comparative analysis among computational intelligence techniques for dissolved oxygen prediction in delaware river. *Geoscience Frontiers*, *8*(3), 517–527.

Oudin, L., Kay, A., Andréassian, V., Perrin, C. Are seemingly physically similar catchments truly hydrologically similar? *Water Resources Research*, *46*(11).

Ouyang, W., Hao, X., Wang, L., Xu, Y., Tysklind, M., Gao, X., Lin, C. Watershed diffuse pollution dynamics and response to land development assessment with riverine sediments. *Science of The Total Environment*, *659*, 283–292.

Parimala, M., Lopez, D., Senthilkumar, N. A survey on density based clustering algorithms for mining large spatial databases. *International Journal of Advanced Science and Technology*, *31*(1), 59–66.

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., Duchesnay, É. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, *12*, 2825–2830.

Pellerin, B.A., Stauffer, B.A., Young, D.A., Sullivan, D.J., Bricker, S.B., Walbridge, M.R., Clyde Jr, G.A., Shaw, D.M. Emerging tools for continuous nutrient monitoring networks: Sensors advancing science and water resources protection. *JAWRA Journal of the American Water Resources Association*, *52*(4), 993–1008.

Peng, G., Lacagnina, C., Downs, R.R., Ganske, A., Ramapriyan, H., Ivánová, I., Wyborn, L., Jones, D., Bastin, L., Shie, C.L., Moroni, D.F. Global community guidelines for documenting, sharing, and reusing quality information of individual digital datasets. *Data Science Journal*.

Plana, Q., Alferes, J., Fuks, K., Kraft, T., Maruéjouls, T., Torfs, E., Vanrolleghem, P.A. Towards a water quality database for raw and validated data with emphasis on structured metadata. *Water Quality Research Journal*, *54*(1), 1–9.

Prasad, A.M., Iverson, L.R., Liaw, A. Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. *Ecosystems*, *9*(2), 181–199.

Prout, J.M., Shepherd, K.D., McGrath, S.P., Kirk, G.J., Haefele, S.M. What is a good level of soil organic matter? an index based on organic carbon to clay ratio. *European Journal of Soil Science*, *72*(6), 2493–2503.

Read, E.K., Carr, L., De Cicco, L., Dugan, H.A., Hanson, P.C., Hart, J.A., Kreft, J., Read, J.S., Winslow, L.A. Water quality data for national-scale aquatic research: The water quality portal. *Water Resources Research*, *53*(2), 1735–1745.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., Prabhat Deep learning and process understanding for data-driven earth system science. *Nature*, *566*(7743), 195–204.

Ross, C.W., Prihodko, L., Anchang, J., Kumar, S., Ji, W., Hanan, N.P. Hysogs250m, global gridded hydrologic soil groups for curve-number-based runoff modeling. *Scientific data*, *5*(1), 1–9.

Sandström, S., Futter, M.N., Kyllmar, K., Bishop, K., O'Connell, D.W., Djodjic, F. Particulate phosphorus and suspended solids losses from small agricultural catchments: Links to stream and catchment characteristics. *Science of the Total Environment*, *711*, 134616.

Sarkar, A., Pandey, P. River water quality modelling using artificial neural network technique. *Aquatic procedia*, *4*, 1070–1077.

Schmidhuber, J. Deep learning in neural networks: An overview. *Neural networks*, *61*, 85–117.

Sheikholeslami, R., Hall, J.W. A global assessment of nitrogen concentrations using spatiotemporal random forests. *Hydrology and Earth System Sciences Discussions*, 1–30.

Shen, C. A transdisciplinary review of deep learning research and its relevance for water resources scientists. *Water Resources Research*, *54*(11), 8558–8593.

Shen, L.Q., Amatulli, G., Sethi, T., Raymond, P., Domisch, S. Estimating nitrogen and phosphorus concentrations in streams and rivers, within a machine learning framework. *Scientific data*, *7*(1), 1–11.

Singh, K.P., Basant, A., Malik, A., Jain, G. Artificial neural network modeling of the river water quality—a case study. *Ecological modelling*, *220*(6), 888–895.

Sinha, E., Michalak, A., Calvin, K.V., Lawrence, P.J. Societal decisions about climate mitigation will have dramatic impacts on eutrophication in the 21st century. *Nature communications*, *10*(1), 1–11.

Sivapalan, M., Blöschl, G. The growth of hydrological understanding: Technologies, ideas, and societal needs shape the field. *Water Resources Research*, *53*(10), 8137–8146.

Solomatine, D., See, L.M., Abrahart, R. Data-driven modelling: Concepts, approaches and experiences. *Practical hydroinformatics*, 17–30.

Song, M., Jiang, Y., Liu, Q., Tian, Y., Liu, Y., Xu, X., Kang, M. Catchment versus riparian buffers: Which land use spatial scales have the greatest ability to explain water quality changes in a typical temperate watershed? *Water*, *13*(13), 1758.

Sprague, L.A., Oelsner, G.P., Argue, D.M. Challenges with secondary use of multi-source water-quality data in the united states. *Water research*, *110*, 252–261.

Stagge, J.H., Rosenberg, D.E., Abdallah, A.M., Akbar, H., Attallah, N.A., James, R. Assessing data availability and research reproducibility in hydrology and water resources. *Scientific data*, *6*, 190030.

Steidl, J., Kalettka, T., Bauwe, A. Nitrogen retention efficiency of a surface-flow constructed wetland receiving tile drainage water: A case study from north-eastern germany. *Agriculture, Ecosystems & Environment*, *283*, 106577.

Strobl, C., Boulesteix, A.-L., Zeileis, A., Hothorn, T. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, *8*(1), 1–21.

Strobl, R.O., Robillard, P.D. Network design for water quality monitoring of surface freshwaters: A review. *Journal of environmental management*, *87*(4), 639–648.

Suuster, E., Ritz, C., Roostalu, H., Reintam, E., Kõlli, R., Astover, A. Soil bulk density pedotransfer functions of the humus horizon in arable soils. *Geoderma*, *163*(1-2), 74–82.

Tahmasebi, P., Kamrava, S., Bai, T., Sahimi, M. Machine learning in geo-and environmental sciences: From small to large scale. *Advances in Water Resources*, *142*, 103619.

Tang, T., Strokal, M., van Vliet, M.T., Seuntjens, P., Burek, P., Kroeze, C., Langan, S., 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.

Tiyasha, T., Tung, T.M., Yaseen, Z.M. A survey on river water quality modelling using artificial intelligence models: 2000–2020. *Journal of Hydrology*, *585*, 124670.

Toming, K., Kotta, J., Uuemaa, E., Sobek, S., Kutser, T., Tranvik, L.J. Predicting lake dissolved organic carbon at a global scale. *Scientific reports*, *10*(1), 1–8.

Tóth, B., Weynants, M., Pásztor, L., Hengl, T. 3d soil hydraulic database of europe at 250 m resolution. *Hydrological Processes*, *31*(14), 2662–2666.

Tyralis, H., Papacharalampous, G., Langousis, A. A brief review of random forests for water scientists and practitioners and their recent history in water resources. *Water*, *11*(5), 910.

United States Geological Survey *Water quality portal* [Accessed on November 16]. Retrieved November 16, 2020, from https://www.waterqualitydata.us/portal/

UN-Water Summary progress update 2021: Sdg 6 — water and sanitation for all.

van Vliet, M.T., Franssen, W.H., Yearsley, J.R., Ludwig, F., Haddeland, I., Lettenmaier, D.P., Kabat, P. Global river discharge and water temperature under climate change. *Global Environmental Change*, *23*(2), 450–464.

Vilmin, L., Mogollón, J.M., Beusen, A.H., Bouwman, A.F. Forms and subannual variability of nitrogen and phosphorus loading to global river networks over the 20th century. *Global and Planetary Change*, *163*, 67–85.

Virro, H., Amatulli, G., Kmoch, A., Shen, L.Q., Uuemaa, E. Grqa: Global river water quality archive. *Earth System Science Data*, *13*(12), 5483–5507.

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., de Lange, H.J. What drives the ecological quality of surface waters? a review of 11 predictive modeling tools. *Water Research*, *208*, 117851.

Vitharana, U., Mishra, U., Jastrow, J., Matamala, R., Fan, Z. Observational needs for estimating alaskan soil carbon stocks under current and future climate. *Journal of Geophysical Research: Biogeosciences*, *122*(2), 415–429.

Wadoux, A.M.-C., Brus, D.J., Heuvelink, G.B. Sampling design optimization for soil mapping with random forest. *Geoderma*, *355*, 113913.

Wagena, M.B., Easton, Z.M. Agricultural conservation practices can help mitigate the impact of climate change. *Science of The Total Environment*, *635*, 132–143.

Wang, D., Thunéll, S., Lindberg, U., Jiang, L., Trygg, J., Tysklind, M. Towards better process management in wastewater treatment plants: Process analytics based on shap values for tree-based machine learning methods. *Journal of Environmental Management*, *301*, 113941.

Wang, R., Kim, J.-H., Li, M.-H. Predicting stream water quality under different urban development pattern scenarios with an interpretable machine learning approach. *Science of The Total Environment*, *761*, 144057.

Whitehead, P.G., Wilby, R.L., Battarbee, R.W., Kernan, M., Wade, A.J. A review of the potential impacts of climate change on surface water quality. *Hydrological sciences journal*, *54*(1), 101–123.

Xu, T., Liang, F. Machine learning for hydrologic sciences: An introductory overview. *Wiley Interdisciplinary Reviews: Water*, *8*(5), e1533.

Xu, X., Ester, M., Kriegel, H.-P., Sander, J. A distribution-based clustering algorithm for mining in large spatial databases. *Proceedings 14th International Conference on Data Engineering*, 324–331.

Yang, X., Jin, W. Gis-based spatial regression and prediction of water quality in river networks: A case study in iowa. *Journal of Environmental Management*, *91*(10), 1943–1951.

Yang, Y.-Y., 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), 1–10.

Yaseen, Z.M. An insight into machine learning models era in simulating soil, water bodies and adsorption heavy metals: Review, challenges and solutions. *Chemosphere*, *277*, 130126.

Yigini, Y., Panagos, P. Assessment of soil organic carbon stocks under future climate and land cover changes in europe. *Science of the Total Environment*, *557*, 838–850.

Yilmaz, K.K., Gupta, H.V., Wagener, T. A process-based diagnostic approach to model evaluation: Application to the nws distributed hydrologic model. *Water Resources Research*, *44*(9).

Zhang, Y., Schaap, M.G. Weighted recalibration of the rosetta pedotransfer model with improved estimates of hydraulic parameter distributions and summary statistics (rosetta3). *Journal of Hydrology*, *547*, 39–53.

Zhi, W., Feng, D., Tsai, W.-P., Sterle, G., Harpold, A., Shen, C., Li, L. From hydrometeorology to river water quality: Can a deep learning model predict dissolved oxygen at the continental scale? *Environmental Science & Technology*, *55*(4), 2357–2368.

Zhong, Z., Chen, Z., Xu, Y., Ren, C., Yang, G., Han, X., Ren, G., Feng, Y. Relationship between soil organic carbon stocks and clay content under different climatic conditions in central china. *Forests*, *9*(10), 598.

Zhou, Y. Real-time probabilistic forecasting of river water quality under data missing situation: Deep learning plus post-processing techniques. *Journal of Hydrology*, *589*, 125164.

SUMMARY IN ESTONIAN

Ruumiandmete harmoniseerimine ja masinõpe veekvaliteedi modelleerimiseks

Traditsiooniliselt on veekvaliteedi modelleerimiseks kasutatud nn protsessipõhiseid mudeleid, mis üritavad veekvaliteedi ja seda mõjutavate keskkonnategurite (mullastik, maakasutus, kliima jne) vahelisi interaktsioone looduses simuleerida matemaatiliste võrranditega. Protsessipõhiste mudelite rakendamise eelduseks on spetsiifiliste sisendandmete (nt mulla veejuhtivus) olemasolu. Kuna antud andmete kättesaadavus on sageli piiratud, on raskendatud ka selliste mudelite kasutamine veekvaliteedi modelleerimises geograafiliselt suurematel aladel. Siinkohal on tänapäeval alternatiiviks masinõppel (ingl *machine learning*) põhinevad mudelid, mis on sisendandmete osas paindlikud ja on võimelised tuvastama eelpool kirjeldatud interaktsioone valemeid defineerimata. Sellegipoolest on masinõppel põhinevate globaalse ja regionaalse ulatusega veekvaliteedi mudelite arendamist takistanud veekvaliteedi näitajate puudulik ruumiline katvus ja kvaliteet.

Doktoritöö eesmärk oli parandada ja harmoniseerida veekvaliteedi modelleerimise andmestikke ning arendada välja masinõppe raamistik, mida saaks kasutada riigiüleseks veekvaliteedi modelleerimiseks. Püstitati järgmised uurimisküsimused:

- Milliste kitsaskohtadega tuleb arvestada veekvaliteedi modelleerimiseks vajalike suurte andmestike harmoniseerimisel?
- Kuidas mõjutab masinõppe mudelite täpsust treeningandmete ruumiline esindatus?
- Kas masinõppel põhinevat raamistikku on võimalik kasutada riigiüleseks veekvaliteedi modelleerimiseks?
- Kuidas mõjutab masinõppe mudelite täpsust prognoosimiseks kasutatavate tunnuste hulk?

Töö käigus valmis kaks veekvaliteedi modelleerimist toetavat andmestikku. Esimene neist on mullastikuandmebaas EstSoil-EH, mis on üle-eestilise katvusega andmestik, kus on üle 750 000 mullaüksuse koos erinevate atribuutidega, mis kirjeldavad mulla füüsilisi ja hüdroloogilisi omadusi. Mullaandmed on veekvaliteedi mudelite oluliseks sisendiks, kuna mulla omadused mõjutavad pinnavee äravoolu ja seeläbi toitainete transporti veekogudesse. Eestis aga puudus modelleerimiseks sobiv mullaandmestik ja seega seati eesmärk vastava masinloetava andmebaasi loomine olemasoleva digitaalse Eesti mullastikukaardi põhjal. EstSoil-EH koostamisel läks põhirõhk mullastikukaardi tekstipõhiste atribuutide masinloetavaks (numbrilisteks atribuutideks) teisendamisele. Nende põhjal tuletati seejärel mulla veejuhtivus ja lasuvustihedus. Andmestikule lisati täiendavatest allikatest mullaüksuste topograafiat, maakasutust ja kuivendustaset kirjeldavad atribuudid.

Globaalse veekvaliteedi andmete katvuse parandamiseks koostati viie olemasoleva suuremõõtmelise andmestiku põhjal uus andmebaas Global River Water Quality Archive ehk GRQA, mis koosneb 42 olulise veekvaliteedi näitaja (hapniku-

sisaldus, setted, toitained jne) mõõteandmetest. Lähteandmestike harmoniseerimiseks tuli standardiseerida parameetrite nimed ja koodid ning teisendada ühikud, mis võisid sama parameetri puhul sõltuvalt allikast erineda. Kvartiilidevahelise vahemiku (ingl *interquartile*) testi alusel tuvastati võõrväärtused (ingl *outlier*) ehk keskmisest oluliselt erinevad mõõteandmed, mis märgistati vastavalt. Selleks, et masinõppe mudel õpiks ära veekvaliteedi sesoonsed kõikumised, peaks mõõteandmete aegread olema võimalikult pidevad ja seetõttu viidi läbi ka aegridade täielikkuse analüüs. Algandmed kattusid osaliselt ruumiliselt ning üks ja sama mõõtekoht võis seega esineda mitmes andmestikus, mistõttu viidi läbi ka aegridade korrelatsioonanalüüs, et tuvastada võimalikud mõõtekohtade duplikaadid.

EstSoil-EH oli oluliseks sisendiks töö raames loodud masinõppemudelitele. Esmalt kasutati selle mullaatribuute üle-eestiliseks mulla orgaanilise süsiniku sisalduse modelleerimiseks. Mudelis rakendati otsustusmetsa (ingl *random forest*) masinõppe meetodit, mis on leidnud laialdast kasutust erinevate keskkonnategurite modelleerimisel, kuna on võimeline saavutama häid tulemusi ka juhul, kui prognoosimiseks kasutatavad mõõteandmed on mürarikkad ja ei vasta normaaljaotusele. Mudeli treenimiseks kasutati olemasolevaid mullasüsiniku proovide andmeid (kokku 3373 proovi), mis pärinesid metsadest, põldudelt, märgaladelt ja rohumaadelt. Mullasüsiniku mudeli täpsust hinnati determinatsioonikordaja (R^2), ruutkeskmise vea (RMSE) ja normaliseeritud mediaanabsoluuthälve (nMAD) kaudu. Viimase abil selgitati välja prognooside täpsuse sõltuvus eelpool mainitud maakasutustüüpidest.

Mullasüsiniku mudeli loomise käigus õpitu põhjal arendati välja raamistik üle-eestiliseks veekvaliteedi modelleerimiseks, keskendudes toitainete—üldlämmastiku ja üldfosfori—kontsentratsioonile jõgedes. Mõlema veekvaliteedi näitaja jaoks loodi omaette mudel, mille treenimiseks kasutati kokku 242 mõõtekoha aasta keskmisi toitainesisalduse andmeid, mis on kogutud aastatel 2016–2020 ja pärinevad Keskkonnaagentuuri keskkonnaseire infosüsteemist (KESE). Mudelites kasutati kokku 82 erinevat tunnust ehk mõõdetavat väärtust, mille alusel toitainete kontsentratsioone prognoositi. Tunnused hõlmasid keskkonnategureid, mis jõevee toitainete sisaldust kõige enam mõjutavad, sh maakasutus, kliima, topograafia ja mulla omadused, millest viimased saadi EstSoil-EH andmestikust. Tunnuste numbrilised väärtused keskmistati iga mõõtekoha jaoks vastavalt valgla piirile. Sarnaselt mullasüsiniku mudelile põhinesid ka veekvaliteedi mudelid otsustusmetsal, kuid mudeleid täiendati nn tunnuste vähendamise (ingl *feature reduction*) protseduuriga. Selle eesmärk oli vähendada omavahel tugevalt korreleeruvate tunnuste hulka mudelis, mis omakorda vähendab mudeli keerukust ja parandab kirjeldusvõimet. Protseduuri tulemusena saadi kummagi veekvaliteedi näitaja puhul neli tunnuste kogumit ja iga kogumi jaoks ehitati eraldi otsustusmetsa mudel. Nende hulgast valiti parimaks mudelid, mis saavutasid optimaalse tasakaalu täpsuse (R^2) ja tunnuste arvu vahel. Seejärel kasutati SHAP (SHapley Additive exPlanations) meetodit, et tuvastada toitainete kontsentratsiooniprognoosi kõige enam mõjutanud tunnused. Lõpuks uuriti parimate mudelite näitel modelleerimise täpsuse ruumilisi erinevusi, arvutades selleks iga valgla jaoks mõõdetud ja prognoositud toitainete sisalduse vahe, mida kujutati seejärel kaardil.

Veekvaliteedi modelleerimist toetavate andmete harmoniseerimisel oli peamine kitsaskoht ebatäpsused metaandmetes. EstSoil-EH puhul oli põhiraskuseks eelkõige tekstipõhiste lähteandmete masinloetavaks muutmine. Kuigi olemasolevad atribuutandmed (mullatüüp, lõimis jms) olid määratud kõrge täpsusastmega, ei sobinud need numbriliste väärtuste puudumise tõttu modelleerimiseks. Seega parandas EstSoil-EH oluliselt Eesti mullastikuandmete kasutatavust modelleerimisel. GRQA koostamist raskendasid peamiselt ebakõlad lähteandmete dokumentatsioonis. Andmestike kombineerimiseks tuli kaardistada erinevused parameetrite nimedes, ühikutes ja keemilistes vormides, et tagada korrektsed teisendused. Seejuures oli oluline säilitada vastavad atribuudid ka nende algkujul, mis võimaldab kasutajatel teisenduse tulemusi valideerida. Täiendavad aegridade täielikkuse ja võõrväärtuste näitajad võimaldavad hinnata andmete sobivust masinõppeks. Kuigi GRQA parandas veekvaliteedi andmete globaalset katvust, on suur osa Aasiast ja Aafrikast siiani hõredalt kaetud. Ruumilist katvust aitaks lähitulevikus suurendada standardsete metaandmete kasutuselevõtt riiklikes institutsioonides, mis võimaldaks andmeid hõlpsamalt ühildada.

Masinõppe mudelite treenimiseks kasutatud mõõteandmete ruumiline esindatus mõjutas nii mulla orgaanilise süsiniku kui toitainete kontsentratsioonide prognoosi täpsust. Mullasüsiniku modelleerimine oli täpsem märgaladel ja põldudel asuvate mullaüksuste puhul, kuigi treeningandmetes olid nende maakasutustüüpide vahekorrad väga erinevad. Seega mõjutab prognoosi edukust lisaks proovide arvule ka nende esindatus ehk kui suure osa tunnuste varieeruvusest kindlast asukohast võetud proov suudab ära kirjeldada. Kuigi üldlämmastiku ja üldfosfori mudelite täpsuses ei tuvastatud selget ruumilist mustrit, olid prognoosid täpsemad suuremates valglates. Toitainete kontsentratsioonide kõikumised avaldavad rohkem mõju väiksemates valglates, põhjustades mõõteandmetes varieeruvust, mida masinõppe mudelid prognoosida ei suuda. Seevastu suuremates valglates aitavad mitmekesisemad keskkonnatingimused järske kontsentratsioonide kõikumisi tasandada, mistõttu mudelitel on kontsentratsioonide dünaamikat lihtsam jäljendada. Võimalik, et mõõtmiste arvu suurendamine väiksemates valglates aitaks seega parandada treeningandmete kirjeldusvõimet.

Nii töös loodud üldlämmastiku kui üldfosfori masinõppe mudelite täpsus on võrreldav Baltimaades ja Skandinaavias varem rakendatud protsessipõhiste mudelitega. Sellegipoolest on loodud masinõppe mudelitel protsessipõhiste ees selged eelised laiaulatuslikes uuringutes. Erinevalt protsessipõhistest mudelitest ei ole masinõppe mudelitel fikseeritud nõudmisi sisendparameetrite osas, mistõttu saab neid rakendada ka piirkondades, kus protsessipõhiste mudelite parameetrite vajalikud lähteandmed puuduvad. Siinkohal kõige olulisemate tunnuste leidmiseks kasutatud SHAP meetod võimaldab tuvastada iga tunnuse mõju suuruse ja suuna (suurendab või vähendab) prognoositud väärtusele. Seega suurendab otsustusmetsa kombineerimine SHAP meetodiga mudeli läbipaistvust. Veekvaliteedi mudelites rakendatud tunnuste vähendamise protseduur näitas, et rahuldava täpsuse saavutamiseks piisas vaid vähem kui pooltest tunnustest. Kuigi otsustusmets ei ole üldjuhul mürarikaste tunnuste suhtes tundlik, aitab ebaoluliste tunnuste eemaldamine vähendada mudeli keerukust. Lisaks tähendab väiksem tunnuste hulk, et tunnuste

lähteandmete töötlemine on ressursisäästlikum. Sellegipoolest tuleb arvestada, et ühes piirkonnas ebaoluliseks peetud tunnused võivad olla prognoosile suurema mõjuga mõnes muus piirkonnas, mistõttu tuleks tunnuste vähendamist rakendada igas piirkonnas eraldi, et vältida tunnuste ennatlikku eemaldamist.

ACKNOWLEDGEMENTS

Above all, I am thankful for the guidance and support of my supervisors **Dr Evelyn Uuemaa** and **Dr Alexander Kmoch** throughout my PhD studies. I would not have managed to get this far without them keeping me on track when I started to veer off course. It remains a mystery to me, where they get the energy to put so much effort into their students, while at the same time teaching what seem to be a countless number of courses and being involved in a wide array of projects in academia and beyond, not to mention that Evelyn also has a whole department to run. On top of that, they have managed to have time to organize all these extracurricular activities ranging from hiking to the many visits to Pahad Poisid for the members of the Landscape Geoinformatics Lab. Prior to starting the PhD, I really did not anticipate that I would spend so much time outside the office together with my supervisors and colleagues (and also enjoy it).

Speaking of which, all the time spent together with the other current and former PhD students and postdocs in the Chair of Geoinformatics, especially **Bruno Montibeller**, **Isaac Buo**, **Iuliia Burdun**, **Oleksandr Karasov**, **Desalew Moges** and **Oleksandr Matsibora**, has turned these former colleagues into good friends of mine. The adventures we had in Estonia and elsewhere really were the highlights of my doctoral studies.

I would also like to thank **Prof Ülo Mander** and **Dr Jüri Roosaare**, who are both partially responsible of me signing up for the PhD in the first place. I was writing my MSc thesis under Jüri's supervision, when he suggested to contact some new German postdoc called Alex, who was looking for a PhD student to join him and Evelyn in their new project, which was supposed to focus on "big data". I had no intention to continue in academia after my MSc studies, so in the next day I headed to the department to meet Alex and politely decline the offer. However, I was stopped in the hallway by Ülo, who managed to drag me to visit Alex and Evelyn and at least hear out the offer. I have to say the conversation the four of us had is kind of a blur by now, so perhaps it was too traumatic to remember, but all I know is that by the end of it I had agreed to try this PhD thing out and see where it goes. And then the PhD thing went on for the next four years.

Many thanks also to **Dr Giuseppe Amatulli** for the guidance during my time abroad at Yale University and for sharing his knowledge and clever tricks related to large-scale open-source geocomputation. Those became very handy when there was a need to use parallel computing to process large datasets for my papers. Our conversations also gave good insight into the business side of academia as well as the academic lifestyle in the United States.

I also thank **Tauri Tampuu** for sharing his LaTeX template with me, so that I would not have to write this thesis in MS Word.

Last, but not least, I thank **my family** for their support over all these years and for keeping discussions related to academic studies to a minimum with me, so that I would not forget that there is life outside of the PhD as well.

PUBLICATIONS

CURRICULUM VITAE

Name Holger Virro
Date of birth 14.07.1993
E-mail holger.virro@ut.ee

Education

2018–2022	University of Tartu, PhD in Geoinformatics
2016–2018	University of Tartu, MSc in Geoinformatics and Cartography (cum laude)
2013–2016	University of Tartu, BSc in Geography (cum laude)
2009–2012	Tartu Tamme Gymnasium (gold medal)
2000–2009	Tartu Tamme Gymnasium

Institutions and positions

2018	Regio Ltd, GIS Specialist
2017	The City Council of Tartu, GIS Specialist

Scholarships and awards

2020	Hydroinformatics Innovation Fellowship, Consortium of Universities for the Advancement of Hydrologic Science (CUAHSI), USA
2019–2020	Dora Plus PhD student mobility scholarship
2019	ESTGIS Scholarship, Estonian Geoinformatics Society (ESTGIS)
2018–2022	Smart specialisation scholarship, Archimedes Foundation
2018	Esri Young Scholars Award, Esri, USA
2017	Raefond Scholarship, University of Tartu Foundation

Supplementary education

2019–2020	Visiting Assistant in Research, Yale School of the Environment, Yale University, New Haven, Connecticut, USA
2019	Summer school *GeoComputation using free and open source software*, University of Basilicata, Matera, Italy

Supervised master's thesis

Ivan Vasilyev, Master's Degree, 2021, (sup) Alexander Kmoch; Holger Virro, Integrating environmental datasets into a Data Cube using Discrete Global Grid Systems, University of Tartu, Faculty of Science and Technology, Institute of Ecology and Earth Sciences

Publications

Virro, Holger; Kmoch, Alexander; Vainu, Marko; Uuemaa, Evelyn (2022). Random forest-based modeling of stream nutrients at national level in a data-scarce region. *The Science of The Total Environment*, 840, 156613. DOI: 10.1016/j.scitotenv.2022.156613.

Kmoch, Alexander; Vasilyev, Ivan; Virro, Holger; Uuemaa, Evelyn (2022). Area and shape distortions in open-source discrete global grid systems. *Big Earth Data*, 1–20. DOI: 10.1080/20964471.2022.2094926.

Kmoch, Alexander; Kanal, Arno; Astover, Alar; Kull, Ain; Virro, Holger; Helm, Aveliina; Pärtel, Meelis; Ostonen, Ivika; Uuemaa, Evelyn (2021). EstSoil-EH: a high-resolution eco-hydrological modelling parameters dataset for Estonia. *Earth System Science Data*, 13 (1), 83–97. DOI: 10.5194/essd-13-83-2021.

Virro, Holger; Amatulli, Giuseppe; Kmoch, Alexander; Shen, Longzhu; Uuemaa, Evelyn (2021). GRQA: Global River Water Quality Archive. *Earth System Science Data*, 13 (12), 5483–5507. DOI: 10.5194/essd-13-5483-2021.

Montibeller, Bruno; Kmoch, Alexander; Virro, Holger; Mander, Ülo; Uuemaa, Evelyn (2020). Increasing fragmentation of forest cover in Brazil's Legal Amazon from 2001 to 2017. *Scientific Reports*, 10 (1). DOI: 10.1038/s41598-020-62591-x.

ELULOOKIRJELDUS

Nimi Holger Virro
Sünniaeg 14.07.1993
E-mail holger.virro@ut.ee

Haridustee

2018–2022	Tartu Ülikool, doktoriõpe, geoinformaatika
2016–2018	Tartu Ülikool, magistriõpe, geoinformaatika ja kartograafia (cum laude)
2013–2016	Tartu Ülikool, bakalaureuseõpe, geograafia (cum laude)
2009–2012	Tartu Tamme Gümnaasium (kuldmedal)
2000–2009	Tartu Tamme Gümnaasium

Töökohad ja ametid

2018	AS Regio, geoinformaatik
2017	Tartu Linnavalitsus, geoinformaatik

Stipendiumid

2020	Hüdroinformaatika innovatsiooni stipendium (Hydroinformatics Innovation Fellowship), Consortium of Universities for the Advancement of Hydrologic Science (CUAHSI), USA
2019–2020	Dora Pluss doktorantide õpirände stipendium
2019	ESTGISi stipendium, Eesti Geoinformaatika Selts (ESTGIS)
2018–2022	Nutika spetsialiseerumise doktorandistipendium, Sihtasutus Archimedes
2018	Esri Noore Teadlase stipendium (Esri Young Scholars Award), Esri, USA
2017	Tartu Raefondi stipendium, Tartu Ülikooli Sihtasutus

Täiendõpe

2019–2020	Külalisdoktorant, Yale School of the Environment, Yale'i Ülikool, New Haven, Connecticut, USA
2019	Suveülikool *GeoComputation using free and open source software*, Basilicata Ülikool, Matera, Itaalia

Juhendatud magistritöö

Ivan Vasilyev, magistrikraad, 2021, (juh) Alexander Kmoch; Holger Virro, Integrating environmental datasets into a Data Cube using Discrete Global Grid Systems (Keskkonnaandmestike integreerimine andmekuupi jagatud globaalsete võrkude süsteemide (DGGS) abil), Tartu Ülikool, Loodus- ja täppisteaduste valdkond, ökoloogia ja maateaduste instituut

Publikatsioonid

Virro, Holger; Kmoch, Alexander; Vainu, Marko; Uuemaa, Evelyn (2022). Random forest-based modeling of stream nutrients at national level in a data-scarce region. *The Science of The Total Environment*, 840, 156613. DOI: 10.1016/j.scitotenv.2022.156613.

Kmoch, Alexander; Vasilyev, Ivan; Virro, Holger; Uuemaa, Evelyn (2022). Area and shape distortions in open-source discrete global grid systems. *Big Earth Data*, 1–20. DOI: 10.1080/20964471.2022.2094926.

Kmoch, Alexander; Kanal, Arno; Astover, Alar; Kull, Ain; Virro, Holger; Helm, Aveliina; Pärtel, Meelis; Ostonen, Ivika; Uuemaa, Evelyn (2021). EstSoil-EH: a high-resolution eco-hydrological modelling parameters dataset for Estonia. *Earth System Science Data*, 13 (1), 83–97. DOI: 10.5194/essd-13-83-2021.

Virro, Holger; Amatulli, Giuseppe; Kmoch, Alexander; Shen, Longzhu; Uuemaa, Evelyn (2021). GRQA: Global River Water Quality Archive. *Earth System Science Data*, 13 (12), 5483–5507. DOI: 10.5194/essd-13-5483-2021.

Montibeller, Bruno; Kmoch, Alexander; Virro, Holger; Mander, Ülo; Uuemaa, Evelyn (2020). Increasing fragmentation of forest cover in Brazil's Legal Amazon from 2001 to 2017. *Scientific Reports*, 10 (1). DOI: 10.1038/s41598-020-62591-x.

DISSERTATIONES GEOGRAPHICAE UNIVERSITATIS TARTUENSIS

1. **Вийви Руссак.** Солнечная радиация в Тыравере. Тарту, 1991.
2. **Urmas Peterson.** Studies on Reflectance Factor Dynamics of Forest Communities in Estonia. Tartu, 1993.
3. **Ülo Suursaar.** Soome lahe avaosa ja Eesti rannikumere vee kvaliteedi analüüs. Tartu, 1993.
4. **Kiira Aaviksoo.** Application of Markov Models in Investigation of Vegetation and Land Use Dynamics in Estonian Mire Landscapes. Tartu, 1993.
5. **Kjell Weppling.** On the assessment of feasible liming strategies for acid sulphate waters in Finland. Tartu, 1997.
6. **Hannes Palang.** Landscape changes in Estonia: the past and the future. Tartu, 1998.
7. **Eiki Berg.** Estonia's northeastern periphery in politics: socio-economic and ethnic dimensions. Tartu, 1999.
8. **Valdo Kuusemets.** Nitrogen and phosphorus transformation in riparian buffer zones of agricultural landscapes in Estonia. Tartu, 1999.
9. **Kalev Sepp.** The methodology and applications of agricultural landscape monitoring in Estonia. Tartu, 1999.
10. **Rein Ahas.** Spatial and temporal variability of phenological phases in Estonia. Tartu, 1999.
11. **Эркки Таммиксаар.** Географические аспекты творчества Карла Бэра в 1830–1840 гг. Тарту, 2000.
12. **Garri Raagmaa.** Regional identity and public leaders in regional economic development. Tartu, 2000.
13. **Tiit Tammaru.** Linnastumine ja linnade kasv Eestis nõukogude aastatel. Tartu, 2001.
14. **Tõnu Mauring.** Wastewater treatment weltlands in Estonia: efficiency and landscape analysis. Tartu, 2001.
15. **Ain Kull.** Impact of weather and climatic fluctuations on nutrient flows in rural catchments. Tartu, 2001.
16. **Robert Szava-Kovats.** Assessment of stream sediment contamination by median sum of weighted residuals regression. Tartu, 2001.
17. **Heno Sarv.** Indigenous Europeans east of Moscow. Population and Migration Patterns of the Largest Finno-Ugrian Peoples in Russia from the 18th to the 20th Centuries. Tartu, 2002.
18. **Mart Külvik.** Ecological networks in Estonia — concepts and applications. Tartu, 2002.
19. **Arvo Järvet.** Influence of hydrological factors and human impact on the ecological state of shallow Lake Võrtsjärv in Estonia. Tartu, 2004.
20. **Katrin Pajuste.** Deposition and transformation of air pollutants in coniferous forests. Tartu, 2004.

21. **Helen Sooväli.** *Saaremaa waltz*. Landscape imagery of Saaremaa Island in the 20th century. Tartu, 2004.
22. **Antti Roose.** Optimisation of environmental monitoring network by integrated modelling strategy with geographic information system — an Estonian case. Tartu, 2005.
23. **Anto Aasa.** Changes in phenological time series in Estonia and Central and Eastern Europe 1951–1998. Relationships with air temperature and atmospheric circulation. Tartu, 2005.
24. **Anneli Palo.** Relationships between landscape factors and vegetation site types: case study from Saare county, Estonia. Tartu, 2005.
25. **Mait Sepp.** Influence of atmospheric circulation on environmental variables in Estonia. Tartu, 2005.
26. **Helen Alumäe.** Landscape preferences of local people: considerations for landscape planning in rural areas of Estonia. Tartu, 2006.
27. **Aarne Luud.** Evaluation of moose habitats and forest reclamation in Estonian oil shale mining areas. Tartu, 2006.
28. **Taavi Pae.** Formation of cultural traits in Estonia resulting from historical administrative division. Tartu, 2006.
29. **Anneli Kährik.** Socio-spatial residential segregation in post-socialist cities: the case of Tallinn, Estonia. Tartu, 2006.
30. **Dago Antov.** Road user perception towards road safety in Estonia. Tartu, 2006.
31. **Üllas Ehrlich.** Ecological economics as a tool for resource based nature conservation management in Estonia. Tartu, 2007.
32. **Evelyn Uuemaa.** Indicatory value of landscape metrics for river water quality and landscape pattern. Tartu, 2007.
33. **Raivo Aunap.** The applicability of gis data in detecting and representing changes in landscape: three case studies in Estonia. Tartu, 2007.
34. **Kai Treier.** Trends of air pollutants in precipitation at Estonian monitoring stations. Tartu, 2008.
35. **Kadri Leetmaa.** Residential suburbanisation in the Tallinn metropolitan area. Tartu, 2008.
36. **Mare Remm.** Geographic aspects of enterobiasis in Estonia. Tartu, 2009.
37. **Alar Teemusk.** Temperature and water regime, and runoff water quality of planted roofs. Tartu, 2009.
38. **Kai Kimmel.** Ecosystem services of Estonian wetlands. Tartu, 2009.
39. **Merje Lesta.** Evaluation of regulation functions of rural landscapes for the optimal siting of treatment wetlands and mitigation of greenhouse gas emissions. Tartu, 2009.
40. **Siiri Silm.** The seasonality of social phenomena in Estonia: the location of the population, alcohol consumption and births. Tartu, 2009.
41. **Ene Indermitte.** Exposure to fluorides in drinking water and dental fluorosis risk among the population of Estonia. Tartu, 2010.
42. **Kaido Soosaar.** Greenhouse gas fluxes in rural landscapes of Estonia. Tartu, 2010.

43. **Jaan Pärn.** Landscape factors in material transport from rural catchments in Estonia. Tartu, 2010.
44. **Triin Saue.** Simulated potato crop yield as an indicator of climate variability in Estonia. Tartu, 2011.
45. **Katrin Rosenvald.** Factors affecting EcM roots and rhizosphere in silver birch stands. Tartu, 2011.
46. **Ülle Marksoo.** Long-term unemployment and its regional disparities in Estonia. Tartu, 2011, 163 p.
47. **Hando Hain.** The role of voluntary certification in promoting sustainable natural resource use in transitional economies. Tartu, 2012, 180 p.
48. **Jüri-Ott Salm.** Emission of greenhouse gases CO_2, CH_4, and N_2O from Estonian transitional fens and ombrotrophic bogs: the impact of different land-use practices. Tartu, 2012, 125 p.
49. **Valentina Sagris.** Land Parcel Identification System conceptual model: development of geoinfo community conceptual model. Tartu, 2013, 161 p.
50. **Kristina Sohar.** Oak dendrochronology and climatic signal in Finland and the Baltic States. Tartu, 2013, 129 p.
51. **Riho Marja.** The relationships between farmland birds, land use and landscape structure in Northern Europe. Tartu, 2013, 134 p.
52. **Olle Järv.** Mobile phone based data in human travel behaviour studies: New insights from a longitudinal perspective. Tartu, 2013, 168 p.
53. **Sven-Erik Enno.** Thunderstorm and lightning climatology in the Baltic countries and in northern Europe. Tartu, 2014, 142 p.
54. **Kaupo Mändla.** Southern cyclones in northern Europe and their influence on climate variability. Tartu, 2014, 142 p.
55. **Riina Vaht.** The impact of oil shale mine water on hydrological pathways and regime in northeast Estonia. Tartu, 2014, 111 p.
56. **Jaanus Veemaa.** Reconsidering geography and power: policy ensembles, spatial knowledge, and the quest for consistent imagination. Tartu, 2014, 163 p.
57. **Kristi Anniste.** East-West migration in Europe: The case of Estonia after regaining independence. Tartu, 2014, 151 p.
58. **Piret Pungas-Kohv.** Between maintaining and sustaining heritage in landscape: The examples of Estonian mires and village swings. Tartu, 2015, 210 p.
59. **Mart Reimann.** Formation and assessment of landscape recreational values. Tartu, 2015, 127 p.
60. **Järvi Järveoja.** Fluxes of the greenhouse gases CO_2, CH_4 and N_2O from abandoned peat extraction areas: Impact of bioenergy crop cultivation and peatland restoration. Tartu, 2015, 171 p.
61. **Raili Torga.** The effects of elevated humidity, extreme weather conditions and clear-cut on greenhouse gas emissions in fast growing deciduous forests. Tartu, 2016, 128 p.
62. **Mari Nuga.** Soviet-era summerhouses On homes and planning in post-socialist suburbia. Tartu, 2016, 179 p.

63. **Age Poom.** Spatial aspects of the environmental load of consumption and mobility. Tartu, 2017, 141 p.
64. **Merle Muru.** GIS-based palaeogeographical reconstructions of the Baltic Sea shores in Estonia and adjoining areas during the Stone Age. Tartu, 2017, 132 p.
65. **Ülle Napa.** Heavy metals in Estonian coniferous forests. Tartu, 2017, 129 p.
66. **Liisi Jakobson.** Mutual effects of wind speed, air temperature and sea ice concentration in the Arctic and their teleconnections with climate variability in the eastern Baltic Sea region. Tartu, 2018, 118 p.
67. **Tanel Tamm.** Use of local statistics in remote sensing of grasslands and forests. Tartu, 2018, 106 p.
68. **Enel Pungas.** Differences in Migration Intentions by Ethnicity and Education: The Case of Estonia. Tartu, 2018, 142 p.
69. **Kadi Mägi.** Ethnic residential segregation and integration of the Russian-speaking population in Estonia. Tartu, 2018, 173 p.
70. **Kiira Mõisja.** Thematic accuracy and completeness of topographic maps. Tartu, 2018, 112 p.
71. **Kristiina Kukk.** Understanding the vicious circle of segregation: The role of leisure time activities. Tartu, 2019, 143 p.
72. **Kaie Kriiska.** Variation in annual carbon fluxes affecting the soil organic carbon pool and the dynamics of decomposition in hemiboreal coniferous forests. Tartu, 2019, 146 p.
73. **Pille Metspalu.** The changing role of the planner. Implications of creative pragmatism in Estonian spatial planning. Tartu, 2019, 128 p.
74. **Janika Raun.** Mobile positioning data for tourism destination studies and statistics. Tartu, 2020, 153 p.
75. **Birgit Viru.** Snow cover dynamics and its impact on greenhouse gas fluxes in drained peatlands in Estonia. Tartu, 2020, 123 p.
76. **Iuliia Burdun.** Improving groundwater table monitoring for Northern Hemisphere peatlands using optical and thermal satellite data. Tartu, 2020, 162 p.
77. **Ingmar Pastak.** Gentrification and displacement of long-term residents in post-industrial neighbourhoods of Tallinn. Tartu, 2021, 141 p.
78. **Veronika Mooses.** Towards a more comprehensive understanding of ethnic segregation: activity space and the vicious circle of segregation. Tartu, 2021, 161 p.
79. **Johanna Pirrus.** Contemporary Urban Policies and Planning Measures in Socialist-Era Large Housing Estates. Tartu, 2021, 142 p.
80. **Gert Veber.** Greenhouse gas fluxes in natural and drained peatlands: spatial and temporal dynamics. Tartu, 2021, 210 p.
81. **Anniki Puura.** Relationships between personal social networks and spatial mobility with mobile phone data. Tartu, 2021, 144 p.
82. **Alisa Krasnova.** Greenhouse gas fluxes in hemiboreal forest ecosystems. Tartu, 2022, 185 p.

83. **Tauri Tampuu.** Synthetic Aperture Radar Interferometry as a tool for monitoring the dynamics of peatland surface. Tartu, 2022, 166 p.
84. **Najmeh Mozaffaree Pour.** Urban Expansion in Estonia: Monitoring, Analysis, and Modeling. Tartu, 2022, 169 p.
85. **Bruno Montibeller**. Evaluating human-induced forest degradation in different biomes using spatial analysis of satellite-derived data. Tartu, 2022, 112 p.