Science of the Total Environment 851 (2022) 158090

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Predicting water quality from geospatial lake, catchment, and buffer zone characteristics in temperate lowland lakes

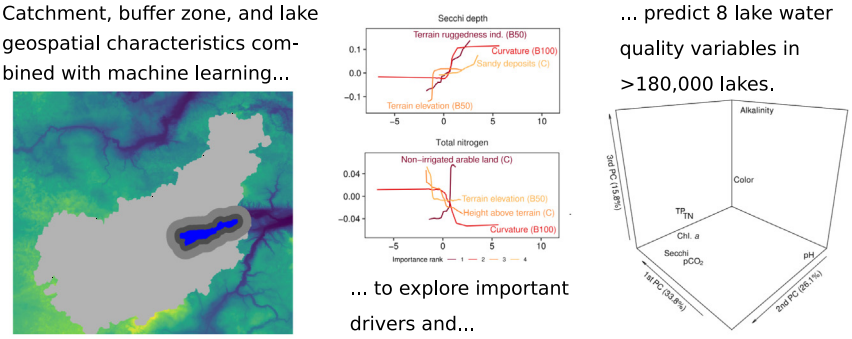
Kenneth Thorø Martinsen *, Kaj Sand-Jensen

Freshwater Biological Laboratory, Department of Biology, University of Copenhagen, Universitetsparken 4, 3rd floor, 2100 Copenhagen, Denmark

HIGHLIGHTS

- Lake water quality can be predicted with machine learning and geospatial predictors.
- Buffer zone geomorphology metrics are predictors of eutrophication related variables.
- Landscape history and catchment soil type are influential on alkalinity and pH.
- Lake surface area is a master variable with strong influence on pH, color, and pCO_2.
- National upscaling can be improved using water quality estimates for >180.000 lakes.

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: José Virgílio Cruz

Keywords:
Machine learning
Predictive modeling
Watershed
Nutrients
Geomorphology
Carbon dioxide

ABSTRACT

Lakes provide essential ecosystem services and strongly influence landscape nutrient and carbon cycling. Therefore, monitoring water quality is essential for the management of element transport, biodiversity, and public goods in lakes. We investigated the ability of machine learning models to predict eight important water quality variables (alkalinity, pH, total phosphorus, total nitrogen, chlorophyll *a*, Secchi depth, color, and pCO_2) using monitoring data from 924 to 1054 lakes. The geospatial predictor variables comprise a wide range of potential drivers at the lake, buffer zone, and catchment level. We compared the performance of nine predictive models of varying complexity for each of the eight water quality variables. The best models (Random Forest and Support Vector Machine in six and two cases, respectively) generally performed well on the test set (R^2 = 0.28–0.60). Models were then used to predict water quality for all 180,377 mapped Danish lakes. Additionally, we trained models to predict each water quality variable by using the predictions we had generated for the remaining seven variables. This improved model performance (R^2 = 0.45–0.78). Overall, the uncovered relationships were in line with the findings of previous studies, e.g., total nitrogen was positively related to catchment agriculture and chlorophyll *a,* Secchi depth, and alkalinity were influenced by soil type and landscape history. Remarkably, buffer zone geomorphology (curvature, ruggedness, and elevation) had a strong influence on nutrients, chlorophyll *a*, and Secchi depth, e.g., curvature was positively related to nutrients and chlorophyll *a* and negatively to Secchi depth. Lake area was a strong predictor of multiple variables, especially its relationship with pH (positive), pCO_2 (negative), and color (negative). Our analysis shows that the combination of machine learning methods and geospatial data can be used to predict lake water quality and improve national upscaling of predictions related to nutrient and carbon cycling.

* Corresponding author.
E-mail address: kenneth.martinsen@bio.ku.dk (K.T. Martinsen).

http://dx.doi.org/10.1016/j.scitotenv.2022.158090
Received 31 March 2022; Received in revised form 11 July 2022; Accepted 13 August 2022
Available online 17 August 2022

1. Introduction

Lakes and wetlands provide a diverse range of essential ecosystem services, e.g., biodiversity, carbon sequestration, food production, water supply, and recreation (Janssen et al., 2021; Peterson et al., 2003). Lakes only cover approximately 3 % of the global surface area (Pekel et al., 2016), nevertheless, lakes – and small lakes in particular – have disproportionately high greenhouse gas (GHG) emissions of CO_2 and CH_4 (Holgerson and Raymond, 2016) and support high biodiversity (Biggs et al., 2017). Freshwater habitats face multiple threats due to human activity, including reclamation, habitat degradation, and eutrophication (Moreno-Mateos et al., 2012; Riis and Sand-Jensen, 2001). Accurate predictions of essential water quality variables in lakes across large scales and understanding the relationships with important drivers can pave the way for better management strategies (Read et al., 2015). Furthermore, the ability to make local, lake-level predictions may reduce the uncertainty associated with the upscaling of GHG emissions (Martinsen et al., 2020a) and other important processes. Here, we used country-level data to investigate the ability of machine learning models to predict eight essential water quality variables – alkalinity, pH, total phosphorus (TP), total nitrogen (TN), chlorophyll *a*, Secchi depth, color, and the partial pressure of carbon dioxide (pCO_2) – from readily available geospatial data.

Several water chemical variables can be used to evaluate the state and quality of freshwater ecosystems (Bhateria and Jain, 2016; Kalff, 2002). These variables are routinely measured in monitoring programs in many countries to determine the ecological quality and function of lakes. This is especially the case for the major nutrients, nitrogen and phosphorus (Stanley et al., 2019), which are closely connected to biodiversity (Jeppesen et al., 2000), phytoplankton development (Kalff and Knoechel, 1978), GHG-emission (Beaulieu et al., 2019; Huttunen et al., 2001), and water clarity (Jeppesen et al., 2000). The nutrient state influences lake primary production and, in turn, the variation in pH and pools of inorganic carbon species, e.g., pCO_2 and alkalinity (Kragh and Sand-Jensen, 2018; Trolle et al., 2012). The relationship between nutrients and production is modulated by the amount of available light (Krause-Jensen and Sand-Jensen, 1998), suspended particles, and colored dissolved organic compounds (Kirk, 1994). Thus, an understanding of lake water quality is obtainable from water samples and measurements of a restricted set of lake variables. However, this approach does not scale well to larger regions and unvisited lakes. High degrees of temporal and spatial variation intensify the challenge of predicting water quality variables in lakes.

Lakes are influenced by their surroundings, both the immediate surroundings (buffer zone) and the topographical area (catchment area, watershed, basin, upslope area, etc.; the term catchment is used onwards) that delivers water to the lake (Jeppesen et al., 1999; Staehr et al., 2012). On its way to the lake as either surface water or groundwater, water chemistry is influenced by land use, geology, soil type, natural vegetation, and human activity (Marx et al., 2017; Rapp et al., 1985; Read et al., 2015). The strength of the relationship between lake chemistry and catchment conditions is modulated by the speed at which water travels through the landscape (Fraterrigo and Downing, 2008; Smits et al., 2017) and the influence of internal lake processes. The presence of streams may reinforce the relationship between catchment and lake chemistry by offering efficient transportation (Abell et al., 2011; Nielsen et al., 2012). The shape of the catchment landscape can be described using geomorphological variables that can be computed from a digital elevation model (DEM; Hengl and Reuter, 2008) and is often used to predict quantities such as water occurrence (Lidberg et al., 2020) and soil composition (Hengl et al., 2014).

Several attempts have been made to quantify the relationship between lake water quality and spatial catchment or buffer zone characteristics (Nielsen et al., 2012; Toming et al., 2020). In-lake concentrations of the major nutrients, nitrogen and phosphorus, increase with agriculture intensity in buffer zones and catchments (Arbuckle and Downing, 2001; Taranu and Gregory-Eaves, 2008). The ability to predict a major limiting nutrient such as phosphorus prompts an expectation that closely related water quality measures such as Secchi depth and chlorophyll *a* levels are predictable from similar sets of variables. Carbonate or silicate mineral levels in the catchment are, as expected, strong predictors of the level of chemical weathering products in streams and downstream lakes (Marx et al., 2017). Similarly, pCO_2 in streams is predictable from catchment characteristics (Lauerwald et al., 2015; Martinsen et al., 2020a). However, parallel relationships for inorganic carbon and nutrients in streams might not be readily transferable to lakes, as the pronounced increase in residence time and, consequently, higher influence of in-lake metabolism, may modify the relationships (Hotchkiss et al., 2015; Martinsen et al., 2020b). Potentially, some of this variation can be accounted for by including variables related to lake bathymetry, which in turn may be related to the slope and shape of the landscape in the buffer zone surrounding the lake (Messager et al., 2016). Despite the expectation that a lake's water quality and its buffer zone or catchment are closely connected, the relationships are not simple and their effects might be outweighed by the pronounced spatial and temporal variation. Furthermore, non-linear relationships and interactions among predictor variables may challenge linear approaches to assessing lake water quality.

Applying methods and models from the field of machine learning could alleviate some of these problems (Olden et al., 2008). These models are trained to minimize prediction error on new observations, i.e., observations not used for training the model, with the goal of maximizing the ability of the models to generalize. They are well-suited for tasks with many observations and predictor variables and can handle complex relationships and inter-correlation (James et al., 2013). The models can be viewed as flexible, functional approximators that capture the relationships between the response and predictor variables (Breiman, 2001). This approach contrasts with traditional statistical modeling, e.g., the family of linear models, where the functional relationships are specified upfront and performance is assessed with the data that had been used to fit the model.

It could be argued that the complexity of some machine learning models may result in 'black-box' models. However, several techniques exist to identify influential variables and assess their relationship with the response variable, making machine learning models a tool for both improving generalization and uncovering important drivers (Molnar, 2019). Therefore, machine learning appears to be a tractable option for improving our ability to predict lake water quality across a wide range of different lakes (Read et al., 2015). Predictions of water quality for a national collection of lakes may both yield estimates for unvisited lakes and act as a reference for ongoing monitoring. That is, the predictions can be used as readily available predictor variables in future modeling efforts of water quality, large-scale nutrient and carbon cycling, and biodiversity (Amatulli et al., 2020; Domisch et al., 2015; Shen et al., 2020).

In this study, we set out to test how well the levels of eight water quality variables can be predicted using geospatial predictors and machine learning models. This is done using lake monitoring data from up to 1054 Danish lakes, collected during the last 20 years, as the response variables and a large collection of readily available predictor variables at the lake, buffer zone, and catchment scale. Specifically, we hypothesize that: 1) the predictability of water quality variables can be improved by including a wider range of predictor variables; 2) complex relationships between response and predictor variables can be modeled using flexible machine learning models; and 3) similar or closely connected water quality variables share important drivers. The overall aim is to produce a country-level dataset with predictions of the eight water quality variables for all 180,378 Danish lakes.

2. Data and methods

2.1. Study region

We used water quality data collected as part of the Danish national monitoring program to train machine learning models to use predictor variables at the lake, buffer zone, and catchment level, and used these models to make predictions for Denmark's 180,378 mapped lakes. Despite Denmark's small area (approx. 43,000 km^2), the lakes span a large gradient in water chemistry due to the variable influence of the last glacial period

(Weichselian glaciation). At their maximum extent, glaciers covered only parts of Denmark (Fig. 1), resulting in large differences in catchment geology between the eastern (glacier-influenced) and western parts of the country. This makes Denmark a suitable study area for investigating the influence of catchment geology and land use on lake water quality and it is representative of the lowland North-temperate regions. The mean annual air temperature is 8.1 °C and annual precipitation is 704 mm (Fick and Hijmans, 2017).

2.2. Water quality response variables

2.2.1. Data selection

We used publicly available data from the national surface water monitoring program (MFVM and DCE, 2021) to calculate annual averages of eight key water quality variables: Alkalinity, pH, total phosphorus and nitrogen, chlorophyll *a*, Secchi depth, color, and pCO_2. These eight variables were selected because they are important for the ecological quality (nutrients, chlorophyll *a*, Secchi depth, and color), involved in carbon cycling (pCO_2, alkalinity, and pH) or predictors of biodiversity (several). Furthermore, measurements of these variables were available for a wide range of lakes with good seasonal coverage. Color is affected by the quantity of humic compounds or the 'colored' fraction of the dissolved organic matter pool. The water quality variables were measured using standard methods, and pCO_2 was calculated from alkalinity, pH, and water temperature using the *seacarb* R-package (Gattuso et al., 2021). The investigated lakes generally have high carbonate alkalinity (Table 1), which reduces the potential influence of organic alkalinity on the estimation of pCO_2 (Abril et al., 2015; Liu et al., 2020). To further reduce the potential bias, and similar to other studies (Hastie et al., 2017), we excluded observations with pH below 5.4. However, the data generally vary in both temporal and spatial sampling intensity, and thus required some preprocessing to calculate a robust annual average for each lake and response variable.

2.2.2. Interpolation and annual averages

We used data from a 20-year period (2000–2019). We used only surface water samples (during water column stratification) or mixed samples (no stratification). We calculated the monthly median for each lake and month, excluding lakes with less than four monthly values. To interpolate values for missing months, we fitted generalized additive models (GAM) for each water quality variable. The term 'month' was modeled as a cyclic cubic spline to make December and January line up, with 'lake' as a random effect, using the *mgcv* R-package (Wood, 2017, 2011). For the GAM analysis, the response variables were $\log_{10}(x + 1)$ transformed. The model fits a global shape of the intra-annual variation with a different intercept for each lake. All models showed a good fit, with R^2 values ranging from 0.56 to 0.96. The fitted GAM were then used to fill the gaps in the annual time series. Finally, we calculated the annual average for each response variable resulting in data from 924 to 1054 lakes (1054 unique lakes; Table 1).

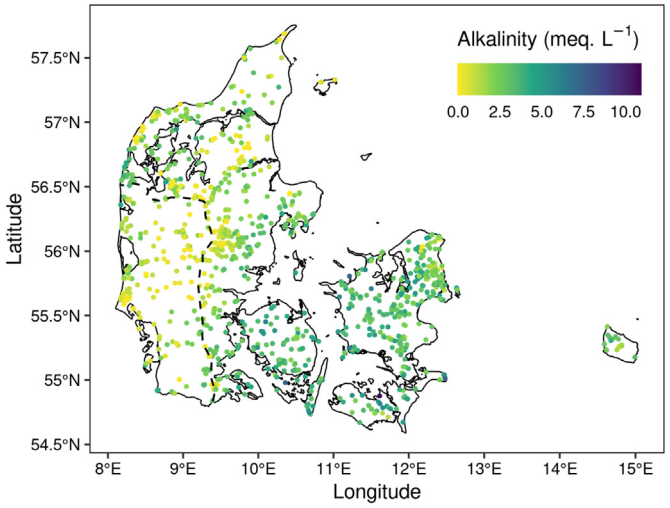


Fig. 1. Map of Denmark with the location of the 1.054 lakes included in the analysis. Lakes are colored by alkalinity and the stippled line identifies the maximum extent of the Weichselian glaciation. Areas to the south-west of the stationary line remained ice-free.

Table 1
Summary statistics of the eight water quality response variables part of the study.

Variable	Unit	Min.	Q25	Median	Mean	Q75	Max.	N
Alkalinity	meq. L^{-1}	0.0	1.1	2.3	2.3	3.3	10.9	999
Chlorophyll *a*	µg L^{-1}	1.0	12.3	27.9	44.6	52.7	1001.5	1046
Color	mg Pt L^{-1}	1.2	22.3	40.8	67.4	73.8	736.3	924
pCO_2	µatm	58	892	1472	2289	2844	34,185	951
pH	pH	3.5	7.5	8.0	7.7	8.3	9.3	1049
Secchi depth	M	0.2	0.7	1.0	1.4	1.6	9.5	1054
Total nitrogen	mg L^{-1}	0.3	1.3	1.8	2.3	2.6	156.8	1052
Total phosphorus	mg L^{-1}	0.0	0.0	0.1	0.3	0.2	56.5	1050

2.3. Catchment delineation

We delineated the topographical catchment of 180,377 lakes (one of the country's 180,378 mapped lakes could not be delineated), including catchments of the 1054 lakes whose water quality data were used in this study. Lake polygons and streamlines for Danish lakes are publicly available (SDFE, 2021). We used a publicly available high-resolution digital elevation model (DEM; 1.6 m resolution) based on LiDAR data from a national survey conducted in 2007, with some hydrological corrections added subsequently (SDFE, 2021).

2.3.1. DEM preprocessing

To ease the computations, we first delineated drainage sub-basins that were then used as units for further hydrological processing. Sub-basins were delineated based on an aggregated (10 m resolution) version of the high-resolution DEM. The DEM was hydrologically corrected by 'breaching', which carves through obstacles to enable flow routing. This contrasts with 'filling', which raises the elevation within depressions that would otherwise impede flow routing (Lindsay and Creed, 2005). Following this, sub-basins were labeled using the algorithm described in Barnes et al. (2014a). The breaching and labeling algorithms are part of the the *RichDEM* Python/C++ package (Barnes, 2016). The delineated sub-basins were then merged iteratively with smaller neighboring sub-basins using GRASS GIS (Neteler and Mitasova, 2013) and further enlarged using a spatial buffer zone of 1000 m to avoid edge artifacts. This resulted in 114 sub-basins and 164 islands for further hydrological processing.

2.3.2. Lake catchments

We split the high-resolution DEM based on the delineated sub-basins and performed hydrological corrections by breaching as described above, flat resolution (Barnes et al., 2014b), and, finally, assignment of flow directions (deterministic-eight method; O'Callaghan and Mark (1984)). These methods are also part of the *RichDEM* library. The topographical catchment for each lake was delineated by identifying all grid cells that contribute flow to the lake, using the *NumPy* and *Numba* Python libraries (Harris et al., 2020; Lam et al., 2015). Finally, catchment polygons were simplified, retaining approximately 10 % of their constituting points to speed up the extraction of summary statistics. This geospatial analysis also relied on other software libraries such as *GDAL* (GDAL/OGR contributors, 2021) and the *sf* (Pebesma, 2018), *exactextractr* (Baston, 2021), and *raster* (Hijmans, 2021) R-packages.

2.4. Predictor variables

We extracted data for 132 predictor variables at the lake (L), buffer zone (B), and catchment (C) levels from a wide range of geospatial data sources (Table S1). The lake level includes attributes based on the lake polygons, e.g., area and shoreline length. The catchment level includes summary statistics of land use (Corine Land Cover: Bossard et al., 2000), geology (Pedersen et al., 2011), geomorphology (DEM and digital surface model derivatives, e.g., slope, aspect, and curvature (Horn, 1981; Zevenbergen and Thorne, 1987), and climate (Fick and Hijmans, 2017). For each lake, we also calculated summary statistics of geomorphological variables within buffer zones of different distances (50, 100, and 250 m). We expect that catchment characteristics are useful for predicting lake water quality, though this relationship might be influenced by, in this case, unknown lake bathymetric attributes such as volume. For this reason, we have included buffer-zone geomorphological variables, which can be expected to correlate with lake bathymetry (Messager et al., 2016).

2.5. Predictive modeling

We used nine machine learning models (Table S2) to predict each of the eight water quality variables from the lake, buffer zone, and catchment characteristics. The models vary in complexity and their ability to approximate the functional relationship between the water quality variables and geospatial characteristics. Importantly, we assessed each model's predictive performance on unseen data using the R-squared (R^2), root-mean-squared-error (RMSE), and mean-absolute-error (MAE) metrics. Therefore, the data were split randomly into training (80 %) and test (20 %) sets, where the training set was used for model selection and training, and the test set was reserved for the final assessment.

2.5.1. Preprocessing

We applied some preprocessing to both the response and predictor variables. This is necessary to reduce distributional skewness and differences in units that could affect model performance. Seven water quality response variables were $\log_{10}(x)$ transformed; the eighth, alkalinity, was $\log_{10}(x + 1)$ transformed. For the predictor variables, we used median imputation to fill a few (three observations in the training data) missing values, removed near-zero variance variables, square-root transformed the catchment land use and geology proportions, applied Yeo-Johnson transformation (Yeo and Johnson, 2000), standardized variables to unit standard deviation, and finally removed predictor variables iteratively until the Spearman rank correlation coefficient of all variable pairs were below 0.7. This left 50 out of the 132 candidate predictor variables (Table S1) for further analysis. The *recipes* R-package was used for preprocessing (Kuhn and Wickham, 2021).

2.5.2. Model selection and training

We compared the predictive performance of nine machine learning models: featureless (AVG), linear model (LM), k-nearest neighbor (NN: Beygelzimer et al., 2019), regression tree (TREE: Therneau and Atkinson, 2019), partial least square regression (PLSR: Liland et al., 2021), elastic net (ELAST: Friedman et al., 2010), neural network (NNET: Venables and Ripley, 2002), support vector machine (SVM: Meyer et al., 2021), and random forest (RF: Wright and Ziegler, 2017). For each response variable, we compared the performance of the nine models using 5-fold cross-validation repeated 5 times (outer loop) to select the best model for further analysis. Many of the models depend on hyperparameter tuning for optimal performance (Table S2). A tuning search was performed for each cross-validation split using 4-fold cross-validation (inner loop) and a 30-iteration random search, and the best set of hyperparameters was then used to fit the model. Following model selection, the best-performing model was trained on the entire training set using the same hyperparameter tuning procedure as for the model selection, except that the number of random search iterations was increased to 100. The final models were then evaluated on the test set and used to make predictions for all lakes in Denmark.

2.5.3. Model interpretation

To identify the important predictor variables and the functional relationships between response and predictor variables, we computed the permutation variable importance (normalized to 0–1 range) and accumulated local effects (ALE) using the *iml* R-package (Molnar et al., 2018). Furthermore, we visualized the relationship between the permutation variable importance scores for each response variable using principal components analysis (PCA). Machine learning models were trained and evaluated using the *mlr* R-package (Bischl et al., 2016).

2.5.4. Using predictions as predictor variables

Thus far, the eight water quality variables have been treated separately. However, the water quality predictions, now available for all Danish lakes, provide a straightforward way to upscale related lake variables. To provide an example of such usage and a performance estimate, we used 5-fold cross-validation to evaluate the performance of an RF model that incorporated all observations of the eight water quality variables, along with the predicted values of the remaining seven. The same tuning procedure as in the model selection analysis was used.

2.6. Data availability

All analysis was performed in R version 4.1 (R Core Team, 2021) or Python version 3.8 (Van Rossum and Drake, 2011). All raw data are publicly available from the sources cited in the main text. All scripts used for the analysis and the resulting data products (e.g., lake catchments, predictions, and models) are available from an online repository (http://doi.org/10.17894/ucph.a344db4b-3d71-4a48-8293-a17b4ccf0e9d).

3. Results

3.1. Danish lakes and catchments

180,378 Danish lakes have been mapped, covering a total area of 713 km^2 (1.66 % of the national area). Small lakes <1 ha make up 97.5 % of the total (Fig. 2 A). Mean and median lake surface areas are 3951 m^2 and 495 m^2 respectively. The catchments of Danish lakes cover a non-overlapping area of 31,811 km^2 (73.8 % of the national area) with a mean and median area of 1.1 km^2 and 0.016 km^2, respectively (Fig. 2 B). Thus, the catchment area is generally much larger than the lake surface area, with a mean and median lake to catchment ratio of 0.47 and 0.04, respectively (Fig. 2 C).

3.2. Water quality variables

Data for eight water quality variables for 924–1054 Danish lakes cover a wide range of conditions in terms of nutrients, inorganic carbon, and light attenuation (Table 1). These lowland lakes are generally nutrient-rich because of the high population density and intense agricultural land use in the country. However, a sampling bias in the lakes that are included in the analysis results in medium- and large-size lakes being overrepresented (Fig. 2 A). Further, there is a pronounced variation in lake alkalinity from the western to eastern parts of Denmark, a consequence of the last glacial period, which left the southwestern part free of ice cover and subject to sand deposition by rivers (Fig. 1). The eight water quality variables are highly inter-correlated (Table S3), as expected.

3.3. Predicting lake water quality

3.3.1. Model selection

For each of the eight response variables, we compared the performance of nine predictive models ranging from very simple (AVG and LM) to more complex and often better-performing models. For six response variables, the best-performing model was RF, with SVM having the best performance for the remaining two variables (pH and color; Fig. S1). Generally, all models performed well on the test set, explaining 28–60 % of the variation (R^2; Table 2).

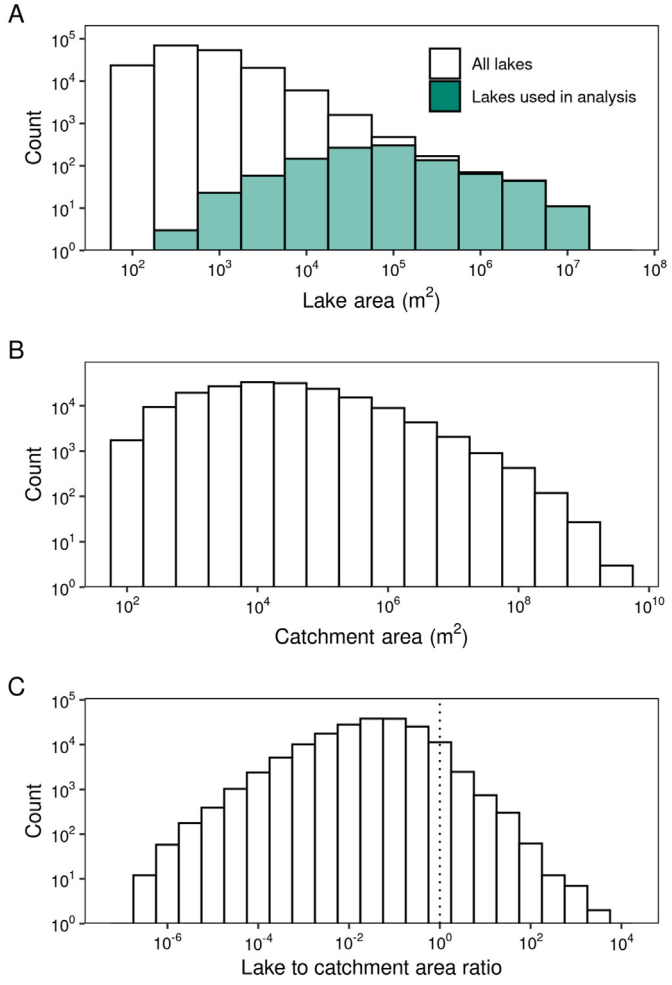


Fig. 2. Density distributions of A) surface area, B) catchment area, and C) the ratio of the lake to the catchment area for 180,377 Danish lakes. In A, the green bars show the distribution of the 1054 lakes that were part of the analysis. One observation is removed in panel C to improve the visualization.

3.3.2. Water quality predictions for Danish lakes

We used the trained models to make predictions for 180,377 Danish lakes (Fig. 3). This is straightforward because the predictor variables are readily available from geospatial databases. The density distributions of the predictor for some response variables differ from those of the observations used to train the models, which is likely a consequence of the bias in sampling of a higher proportion of medium- to large lakes, relative to small lakes (Fig. 2 A). This is the case for pCO_2, pH, color, and Secchi depth, for which the predicted national average (Fig. 3) is much different from the observations (Table 1) because the many small lakes are predicted to have higher pCO_2 and color along with lower pH and Secchi depth. The difference is less pronounced for TP, TN, alkalinity, and chlorophyll *a*, which are not affected by lake size to the same degree.

Table 2
Predictive performance on the test set of the best performing machine learning models as root-mean-squared-error (RMSE), mean-absolute-error (MAE), and variance explained (R^2) for eight water quality variables (log_{10}-transformed).

Variable	RMSE	MAE	R^2	N
Alkalinity	0.142	0.104	0.60	195
Chlorophyll *a*	0.396	0.319	0.28	211
Color	0.285	0.204	0.55	179
pCO_2	0.288	0.209	0.36	187
pH	0.036	0.024	0.50	210
Secchi depth	0.227	0.167	0.38	211
Total nitrogen	0.207	0.160	0.33	211
Total phosphorus	0.431	0.329	0.38	211

3.3.3. Using water quality predictions to train new models

In addition to generating knowledge that is immediately useful in nature management, model predictions can be used in future national modeling efforts. To demonstrate this, we computed cross-validated model performance using observations of each of the eight water quality variables as the response variable, with the predictions of the remaining seven variables as the predictors (Table 3). For all variables, the mean performance was superior to that of models that only use geospatial variables at the lake, buffer zone, and catchment level. This is a result of the interdependence between the water quality variables, which often are highly correlated (Table S3). Moreover, it indicates that the trained predictive models must capture relevant relationships for each of the eight water quality variables and not only consider them individually. This in turn improves the performance when the predictions are subsequently used as part of new modeling efforts.

3.4. Drivers of lake water quality

Fifty predictor variables were used to train the predictive models. These variables directly or indirectly cover a range of effects and interactions at the lake, buffer zone, and catchment level. The influence of each predictor variable on each of the eight response variables generally differs widely, but some commonalities are present for similar variables such as nutrients (Fig. 4). Generally, the incorporation of several predictor variables enhanced the performance of the predictive models. The exception is the pCO_2 levels, which was reliably and consistently predicted on the basis of a single predictor variable: lake area. Among the top ten predictors for the water quality variables, the major categories of predictor variables are present, i.e. geomorphology (curvature and elevation), land use (non-irrigated arable land and coniferous forest), geology (clayey till and sandy deposits), and lake metrics (area, distance to the coastline, distance to main stationary line of ice cover during the last glaciation), and the three spatial levels (lake, buffer zone, and catchment).

3.4.1. Permutation variable importance

In general, many of the predictor variables expected a priori to be important are present among the most important variables (Fig. 4). This is the case for lake area and many of the catchment soil and land use variables, which previous studies suggested would be promising predictor variables. Perhaps more surprising is the apparent importance of geomorphological variables at both the buffer and catchment levels. These variables may influence lake water quality through both direct (e.g., soil erosion) and indirect pathways (e.g., buffer zone geomorphology as related to lake bathymetry).

For alkalinity and pH, clayey till, the distance to the main stationary line, and coniferous forest for pH, are important predictors, emphasizing the influence of the last glaciation on inorganic carbon chemistry in Danish lakes (Fig. 1). Lake area, catchment forest cover, and presence of freshwater deposits were important predictors of color. For response variables directly related to eutrophication (TP, TN, Secchi depth, and chlorophyll *a*), buffer zone curvature and elevation, unexpectedly, were among the most important predictors; for Secchi depth, ruggedness also was a surprisingly important predictor. Less surprisingly, agriculture (non-irrigated arable land), ice cover during the last ice age, catchment geology (clayey till and sandy deposits), and landscape position (distance to the coastline) were all important predictors (Fig. 4).

3.4.2. Response for the most important predictors

The average influence of each predictor variable along a range of values was examined with ALE. From the response curves of the most important variables (Fig. 5), it is evident that the trained predictive models can capture a range of relationships, with the RF appearing more step-like because it is an ensemble of regression trees, in contrast to the more linear

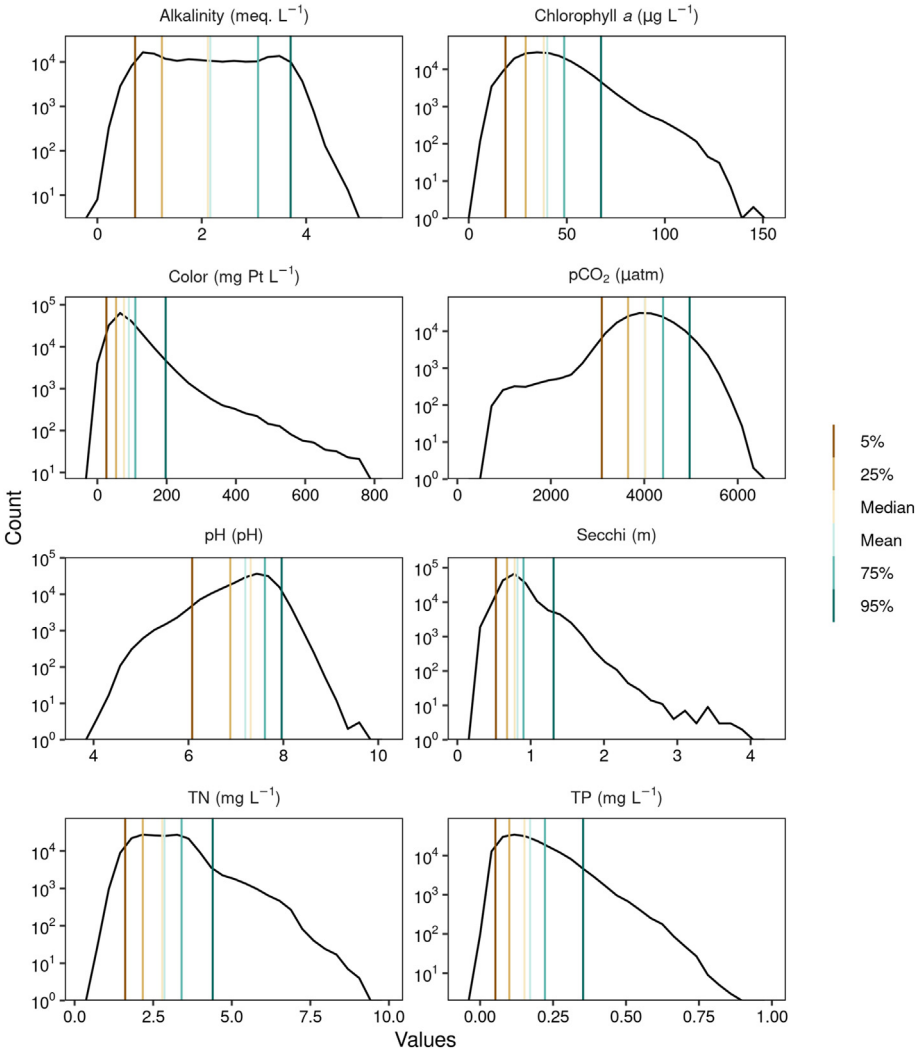


Fig. 3. Density distributions of eight water quality variables in 180,377 Danish lakes predicted using the best performing machine learning models. Vertical lines show distributional summary statistics. A total of 37 observations were removed to improve the visualization of distributions. Ordinate axes are in log-units.

relationship with SVM (color and pH). For many of the predictor variables, the relationships, either positive or negative, are similar to a priori expectations. Alkalinity increases with the proportion of clayey till in the catchment and the distance to the main stationary line and, thus coverage by glaciers during the last glaciation (Fig. 1). Color decreases with lake area and increases with the proportion of mixed forest in the catchment. For pCO_2, the dominant relationship is the negative relationship with lake area, and consequently, its relationships are positive with color and negative with pH. For the eutrophication-related variables, the influence of buffer zone geomorphology (e.g., curvature and elevation, and ruggedness for Secchi depth) is perhaps the most striking, with a positive relationship with TP, TN, and chlorophyll *a*.

Table 3
5-fold cross-validation performance of an RF model trained using observations of each of the eight water quality variables (log_{10}-transformed) as the response variable and the predicted values of the remaining seven as the predictor variables. The root-mean-squared-error (RMSE), mean-absolute-error (MAE), and variance explained (R^2) are reported as mean (±SD).

Variable	RMSE	MAE	R^2
Alkalinity	0.132 (±0.008)	0.099 (±0.006)	0.66 (±0.055)
Chlorophyll *a*	0.313 (±0.034)	0.242 (±0.026)	0.544 (±0.069)
Color	0.262 (±0.011)	0.196 (±0.008)	0.62 (±0.071)
pCO_2	0.272 (±0.019)	0.201 (±0.009)	0.449 (±0.058)
pH	0.028 (±0.003)	0.017 (±0.002)	0.778 (±0.031)
Secchi depth	0.182 (±0.012)	0.141 (±0.008)	0.61 (±0.026)
Total nitrogen	0.182 (±0.015)	0.134 (±0.009)	0.472 (±0.07)
Total phosphorus	0.348 (±0.015)	0.26 (±0.01)	0.574 (±0.07)

3.4.3. Similarities in important drivers of water quality

As highlighted above, there are similarities in the permutation variable importance among the eight response variables. This is not surprising since there are significant correlations among the eight variables (Table S3). The similarities can be visualized using PCA (Fig. 6), showing that the eutrophication-related variables generally group together, near Secchi depth and pCO_2, but distanced from pH, alkalinity, and color.

4. Discussion

4.1. Predictability of water quality variables

From a large candidate set of 132 predictor variables, we used 50 to train predictive models for eight water quality variables. We selected readily available predictor variables, so that predictions could be made for all current and potential future lakes. Some variables belong to well-defined categories, e.g., land use, geology, or geomorphology, while others place the lake in the landscape and account for spatial autocorrelation, e.g., distance to the coastline, distance to the main stationary line of the

Fig. 4. The permutation variable importance (normalized to 0–1 range with 1 and 0 being the most and least important respectively) of 50 predictor variables calculated using the best performing machine learning model for each of the eight water quality variables. Variables are determined at the levels of the lake (L), catchment (C), and buffer zones (B; e.g., B100 is a 100 m wide buffer zone surrounding the lake).

long-gone glaciers (Hengl et al., 2018). The predictive performance ranged from moderate to strong, both regarding the accuracy as RMSE or MAE and regarding the proportion of explained variance as R^2. Thus, this study demonstrates the ability of machine learning models to yield good predictions of several lake water quality variables from readily available data on lake, buffer zone, and catchment geospatial variables (Fig. S2). This result is a significant improvement over previous approaches. Many existing models use geospatial variables to predict water quality by predicting the concentrations of nutrients like TN and TP (Stanley et al., 2019), but few have attempted to generate predictions for pCO_2, pH, and alkalinity. This motivated us to include a broader range of important lake variables and evaluate the relationships among them.

Multiple studies have used landscape characteristics to predict lake water quality, but their accuracy and the proportion of variance explained has generally been low (Gémesi et al., 2011; Nobre et al., 2020). Comparing the performance of various models may be difficult because they use different predictor variables and because some studies evaluate their model's performance on the basis of data used to train the model, while others use an independent test set. For TN and TP, the R^2 values reported here are similar to previous studies of Denmark (Nielsen et al., 2012) and climatically similar Estonia (Sepp et al., 2022). For pCO_2, other studies have achieved similar R^2 values using mainly lake variables instead of the more indirect landscape drivers (Humborg et al., 2010; Lapierre et al., 2017). In general, the framework presented here yielded good performance compared to existing studies, despite not using any measured lake chemical variables as predictors. This is a great step forward because the data can be upscaled efficiently, whereas models that are dependent on data from individual lakes cannot.

The fact that data from catchments and buffer zones are good predictors of lake chemistry is highly promising, though not a surprise (Arbuckle and Downing, 2001; Gémesi et al., 2011; Nobre et al., 2020). Previous studies have noted that the degree of lake-catchment interaction depends on the

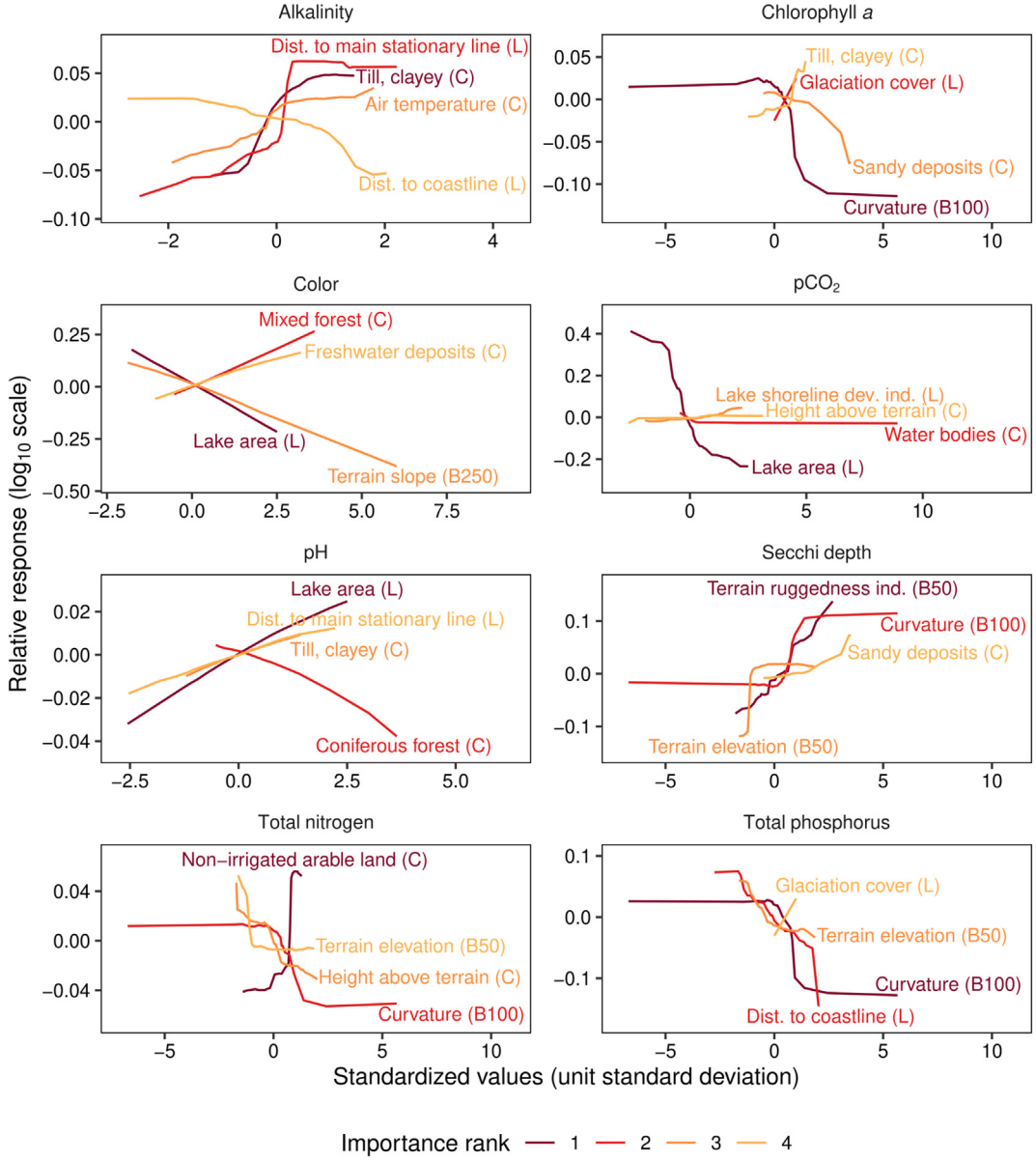


Fig. 5. Influence of the four most important predictor variables assessed using Accumulated Local Effects (ALE) for each of the eight water quality response variables. Lines are colored according to their ranked importance, with 1 being the most important.

fluvial connectivity of the catchment, i.e., how efficiently water and elements are transported within and to the lake system (Fraterrigo and Downing, 2008; Read et al., 2015). In an effort to accommodate this, we included variables describing lake–stream connectivity but these were found not to be of great importance. However, the distance to the coastline and terrain elevation were important for several water quality variables and should capture some aspects of the lakes' position in the fluvial network (Olden et al., 2001). This could be developed further by applying network analysis and thereby emplace the lake in the fluvial network (Jones, 2010).

While the topographical catchment units are straightforward to delineate, they are also prone to error (Oksanen and Sarjakoski, 2005). Errors may be due to inaccurate catchment delineations or water transport that does not follow the terrain surface (Lindsay, 2016). This may especially be the case in flat regions like Denmark, which has a high degree of subsurface water transportation via, e.g., sewerage, drainage, or groundwater. The use of a very high-resolution DEM and the applied preprocessing methods should alleviate some of the issues associated with flow routing in flat landscapes (Lidberg et al., 2017).

4.2. Drivers of lake water quality

We expected catchment land use, geology, and soil composition to be influential predictor variables of water quality. For several of the response variables, the predictive models also uncovered such relationships but additionally found geomorphology at the buffer and catchment level to be of particular importance. Especially for eutrophication-related variables TN, TP, Secchi depth, and chlorophyll *a,* the influence was apparent. Generally, the water-quality variables were highly inter-correlated (Table S3) and in line with results presented in numerous studies. One exception is pCO_2, that has been considered to a lesser extent in this context, and shows the strongest correlations to color and pH, but exhibited no apparent relationship to TN and TP. The similarities among important predictor variables between the eight water quality variables were examines using PCA. The water quality variables directly related to eutrophication, TN, TP, and chlorophyll *a*, group together, with pCO_2 and Secchi depth slightly distanced. This is expected, as pCO_2 and Secchi may not be directly influenced by the drivers of eutrophication but are affected by attributes such

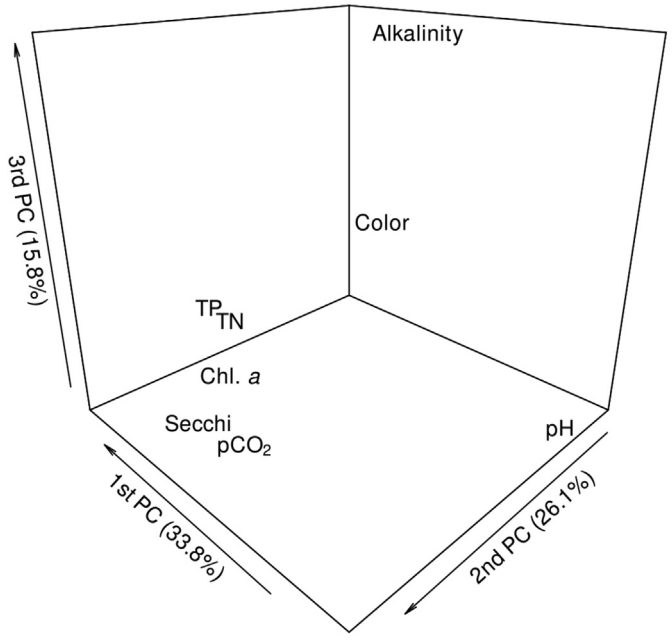


Fig. 6. Principal components analysis of the eight response variables based on the permutation variable importance calculated for 50 predictor variables. The water quality variables are shown in a three-dimensional space using the first three principal components (PC) with the variance explained by the respective component shown in parenthesis.

as lake area. Further distanced from these are alkalinity, pH, and color, which differ to a much larger degree in their drivers.

For alkalinity, the strong impact of the last glaciation was pronounced, and as expected shared influential predictors with pH (Rebsdorf et al., 1991), which was positively influenced by lake area and negatively by coniferous forest cover. Other studies have also found an influence of coniferous forest cover on lake pH (D'Arcy and Carignan, 1997). The positive influence of surface area on pH, and the inverse for pCO_2, is likely related to the higher gas transfer velocity (Holgerson et al., 2017), the decreasing influence of terrestrial organic matter, and CO_2 supersaturated groundwater inputs relative to lake volume associated with increasing lake area (Martinsen et al., 2020b). The overshadowing influence of lake area on pCO_2 is somewhat surprising, given that landscape processes and geomorphology are important predictors of pCO_2 in streams (Martinsen et al., 2020a; Rocher-Ros et al., 2019). It is likely that the increased residence time and seasonal stratification-mixing dynamics (Weyhenmeyer et al., 2012), even in the shallow Danish lakes, overrides these effects on surface water pCO_2. For color, the negative influence of lake area and positive influence of forest cover is in line with previous studies in Danish lakes and the higher input of terrestrial humic organic material relative to lake water volume (Sand-Jensen and Staehr, 2009, 2007). We also observed a negative relationship between color and terrain slope, which has been reported by other studies as well (Rasmussen et al., 1989). The success of terrain slope as a predictor of terrestrially influenced carbon pools, e.g., color, DOC (D'Arcy and Carignan, 1997), and pCO_2 (Martinsen et al., 2020a), is likely due to the influence of terrain slope on soil thickness and organic matter accumulation. Jankowski et al. (2014) found that the quality of organic matter in streams, expressed as the temperature sensitivity of respiration increased as catchment slopes became steeper, while the quantity of organic matter decreased. This may explain some of the observed variations in the pools of dissolved carbon, as differences in substrate, and the degree to which they have been exposed to microbial processing is influenced by the transit time within the catchment. A steep catchment terrain also impacts pCO_2 directly, through higher gas exchange velocities (Wallin et al., 2011), and the relative role of photochemical degradation of colored dissolved organic matter during transit (Köhler et al., 2002).

Geomorphological variables may influence lake nutrient variables directly or indirectly. For example, soil erosion, which is directly influenced by landscape geomorphology, land use, and soil type is known to be an important delivery mechanism for phosphorus (Laubel et al., 2003). Erosion of soil that contains particle-bound phosphorus is driven by physical forcing through precipitation and wind, and is particularly important in regions with exposed soils and steep slopes (Grant et al., 1996; Kronvang et al., 2007). Our finding regarding the influence of geomorphology on chlorophyll *a* may be a result of its impact on TP, as these two variables are highly correlated (Table S3). Furthermore, phosphorus is often the limiting nutrient for phytoplankton development in lakes and thus an important predictor of chlorophyll *a* and potentially Secchi depth (Jackson et al., 2007; Jeppesen et al., 2000). The high importance of catchment agriculture cover for TN is expected, as most nitrogen input to Danish streams comes from agriculture-dominated catchments, whereas phosphorus inputs are generally dominated by point sources (Jeppesen et al., 1999) and are very sensitive to soil erosion in steep terrain. Indirect effects of geomorphology on lake nutrients are also likely, because buffer zone topography is a good predictor of lake bathymetry (Messager et al., 2016; Sobek, 2011) and in turn TP, TN, chlorophyll *a*, and Secchi depth (Fee et al., 1996; Qin et al., 2020).

The lake water quality predictions produced in this study can also be used to investigate the traditional water quality relationships of interest, e.g., the relationships between chlorophyll *a* and TP and TN (Kalff, 2002). Our predictions cover the nationwide range of environmental characteristics and lake types which should provide an improved view of such relationships. The slope between chlorophyll *a* and TN appears similar to those found in earlier more restricted studies but appears to be lower for TP (Fig. S3). This could be due to the presence of inflection points and non-linearities (Quinlan et al., 2021) which is more likely when considering a much wider gradient in lake sizes and nutrient conditions in the nationwide data. This could be further promoted by the inclusion of small lakes which tend to have higher dissolved organic carbon and color levels (as also shown here) which may restrict phytoplankton biomass and primary production (Sand-Jensen and Staehr, 2009, 2007). Additionally, the relationships are modified by water depth (e.g., euphotic depth: mixed depth) in a complex manner as light, vertical mixing, and wind-induced particle resuspension may reach shallow bottoms (Krause-Jensen and Sand-Jensen, 1998; Yuan and Jones, 2020).

4.3. Predictive modeling and large-scale investigations

Machine learning models, as opposed to traditional linear models, can have several advantages. Their predictive performance can often be superior because they are sufficiently flexible to capture complex, non-linear relationships and interdependencies that are present in monitoring data (Read et al., 2015), and they scale well to many more observations (Olden et al., 2008). Furthermore, the researchers do not define a priori the functional relationship between the response and predictor variables, a task that is more straightforward for experimental data but challenging for a diverse range of predictors presented here (Breiman, 2001). A model that can approximate the functional relationships may even reveal new or surprising relationships (Read et al., 2015). However, the flexibility of the models may also result in poor performance when extrapolating outside the range of predictor variables or new spatio-temporal domains, and this possibility should be taken into account when evaluating model performance in an extrapolation setting (Meyer et al., 2018; Meyer and Pebesma, 2021). Consequently, while we are confident that equivalent datasets can be assembled for other countries and regions, especially when considering the availability of readily available data on elevation, land use, soil composition, etc. with near-global coverage (Amatulli et al., 2020; Hengl et al., 2014), the model performances reported here is strictly valid only for Denmark, the range of environmental conditions is so extensive that the models should be applicable in many regions with a similar climate, geology, and soil conditions.

In the present analysis, we only considered the spatial variation, that is, we used multi-year aggregations of lake water quality, assuming that the

inter-annual variation is low. We believe this is a reasonable assumption for the 2000–2019 period in a Danish context, where lake water quality improved markedly with the implementations of the water action plans in 1987, but has since slowed down (Kronvang et al., 2005). Including temporal variation in a predictive model could be advantageous for some use-cases. Indeed, GAM models that interpolated monthly values captured seasonal variation very well, suggesting that a spatio-temporal predictive (forecasting) model also could perform well. Approaches that combine machine learning models and process knowledge could result in further improvements (Hanson et al., 2020).

As a secondary step in our analysis, we show that the water quality predictions can be used subsequently to train even better-performing models. This highlights an advantage of considering multiple response variables simultaneously and suggests that making predictions for all available lakes, and not only for the lakes used for modeling, could be a suitable direction for future studies. This is commonly done for variables that are continuous in space, e.g., climate or soil type (Amatulli et al., 2020; Fick and Hijmans, 2017; Hengl et al., 2014), and less so for discrete units such as lakes. Adding additional water chemical response variables using the existing monitoring data to create a collection of predictions for each lake could provide exciting stepping stones for future upscaling exercises. However, monitoring data, as in Denmark, may contain biases. In our study as well as others (Stanley et al., 2019; Wagner et al., 2008), small lakes were undersampled, leading to biases in sample statistics that potentially influenced model performance. The lack of studies of small lakes may result in significant differences between sampled observations and large-scale predictions of water quality variables where lake size is influential, e.g., pH, pCO_2, and color in this study. For pCO_2, we found an approximately two-fold difference between the mean computed for observations from monitoring data and the national predicted values, and, despite the higher gas transfer velocity in larger lakes, this may have significant implications for carbon emission estimated over large geographical areas (Holgerson and Raymond, 2016; Martinsen et al., 2020b).

4.4. Conclusions

A significant proportion of the variation in water quality between lakes can be explained using geospatial predictor variables related to land use, geology, geomorphology, and lake morphometry. Some of the relationships have previously been described, however, we present a surprisingly strong influence of buffer zone geomorphology (curvature, elevation, ruggedness, slope, etc.) on lake water quality. Eutrophication related variables share many important predictor variables but some variables (e.g. pH, alkalinity, and color) are ultimately driven by other processes, which appear in our analysis when considering similarity between water quality variables based on the contribution of important predictor variables. Furthermore, we show that relationships between water quality and predictor variables can be approximated using flexible machine learning models and, subsequently, the most important variables and their relationships can be identified. These models automate an otherwise error-prone step of the modeling process, thereby improving predictive performance. The water quality predictions can be included as predictor variables in new models and, if available for entire regions, form the basis for improved large-scale estimates of nutrient and carbon cycling by moving beyond estimates based on discrete size classes. Additionally, the predictive models can be used to inform approaches to nature management by providing water quality estimates for lakes, both existing and contemplated, with varying lake, buffer zone, and catchment characteristics. Future studies should address how the temporal dimension can be incorporated and the potential influence of biases associated with data from regional monitoring programs.

CRediT authorship contribution statement

Kenneth Thorø Martinsen: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. **Kaj Sand-Jensen:** Conceptualization, Methodology, Funding acquisition, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

KTM and KSJ were supported by a grant from the Independent Research Fund Denmark (0217-00112B) to the project: “Supporting climate and biodiversity by rewetting low-lying areas”. We thank David Stuligross for linguistic corrections.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2022.158090.

References

Abell, J.M., Özkundakci, D., Hamilton, D.P., Miller, S.D., 2011. Relationships between land use and nitrogen and phosphorus in New Zealand lakes. Mar. Freshw. Res. 62, 162–175. https://doi.org/10.1071/MF10180.

Abril, G., Bouillon, S., Darchambeau, F., Teodoru, C.R., Marwick, T.R., Tamooh, F., Ochieng Omengo, F., Geeraert, N., Deirmendjian, L., Polsenaere, P., Borges, A.V., 2015. Technical note: large overestimation of pCO_2 calculated from pH and alkalinity in acidic, organic-rich freshwaters. Biogeosciences 12, 67–78. https://doi.org/10.5194/bg-12-67-2015.

Amatulli, G., McInerney, D., Sethi, T., Strobl, P., Domisch, S., 2020. Geomorpho90m, empirical evaluation and accuracy assessment of global high-resolution geomorphometric layers. Sci. Data 7, 162. https://doi.org/10.1038/s41597-020-0479-6.

Arbuckle, K.E., Downing, J.A., 2001. The influence of watershed land use on lake N: P in a predominantly agricultural landscape. Limnol. Oceanogr. 46, 970–975. https://doi.org/10.4319/lo.2001.46.4.0970.

Barnes, R., 2016. RichDEM: terrain analysis software. http://github.com/r-barnes/richdem.

Barnes, R., Lehman, C., Mulla, D., 2014a. Priority-flood: an optimal depression-filling and watershed-labeling algorithm for digital elevation models. Comput. Geosci. 62, 117–127. https://doi.org/10.1016/j.cageo.2013.04.024.

Barnes, R., Lehman, C., Mulla, D., 2014b. An efficient assignment of drainage direction over flat surfaces in raster digital elevation models. Comput. Geosci. 62, 128–135. https://doi.org/10.1016/j.cageo.2013.01.009.

Baston, D., 2021. exactextractr: fast extraction from raster datasets using polygons. https://CRAN.R-project.org/package=exactextractr.

Beaulieu, J.J., DelSontro, T., Downing, J.A., 2019. Eutrophication will increase methane emissions from lakes and impoundments during the 21st century. Nat. Commun. 10, 1–5. https://doi.org/10.1038/s41467-019-09100-5.

Beygelzimer, A., Kakadet, S., Langford, J., Arya, S., Mount, D., Li, S., 2019. FNN: fast nearest neighbor search algorithms and applications. https://CRAN.R-project.org/package=FNN.

Bhateria, R., Jain, D., 2016. Water quality assessment of lake water: a review. Sustain. Water Resour. Manag. 2, 161–173. https://doi.org/10.1007/s40899-015-0014-7.

Biggs, J., von Fumetti, S., Kelly-Quinn, M., 2017. The importance of small waterbodies for biodiversity and ecosystem services: implications for policy makers. Hydrobiologia 793, 3–39. https://doi.org/10.1007/s10750-016-3007-0.

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., Casalicchio, G., Jones, Z.M., 2016. mlr: machine learning in R. J. Mach. Learn. Res. 17, 1–5. https://doi.org/10.5555/2946645.3053452.

Bossard, M., Feranec, J., Otahel, J., 2000. CORINE Land Cover Technical Guide: Addendum 2000. European Environment Agency, Copenhagen.

Breiman, L., 2001. Statistical modeling: the two cultures (with comments and a rejoinder by the author). Stat. Sci. 16, 199–231. https://doi.org/10.1214/ss/1009213726.

MFVM, D.C.E.collab, 2021. Surface water database. (in Danish). Accessed October 2021Ministry of Environment and Food of Denmark and Danish Centre for Environment and Energy. https://odaforalle.au.dk/.

D'Arcy, P., Carignan, R., 1997. Influence of Catchment Topography on Water Chemistry in Southeastern Québec Shield Lakes. 54, p. 13. https://doi.org/10.1139/f97-129.

Domisch, S., Amatulli, G., Jetz, W., 2015. Near-global freshwater-specific environmental variables for biodiversity analyses in 1 km resolution. Sci.Data 2, 1–13. https://doi.org/10.1038/sdata.2015.73.

Fee, E.J., Hecky, R.E., Kasian, S.E.M., Cruikshank, D.R., 1996. Effects of lake size, water clarity, and climatic variability on mixing depths in Canadian shield lakes. Limnol. Oceanogr. 41, 912–920. https://doi.org/10.4319/lo.1996.41.5.0912.

Fick, S.E., Hijmans, R.J., 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. Int. J. Climatol. 37, 4302–4315. https://doi.org/10.1002/joc.5086.

Fraterrigo, J.M., Downing, J.A., 2008. The influence of land use on lake nutrients varies with watershed transport capacity. Ecosystems 11, 1021–1034. https://doi.org/10.1007/s10021-008-9176-6.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33, 1. https://doi.org/10.18637/jss.v033.i01.
Gattuso, J.-P., Epitalon, J.-M., Lavigne, H., Orr, J., 2021. seacarb: seawater carbonate chemistry. https://CRAN.R-project.org/package=seacarb.
GDAL/OGR contributors, 2021. GDAL Geospatial Data Abstraction Software Library. Open Source Geospatial Foundation.
Gémesi, Z., Downing, J.A., Cruse, R.M., Anderson, P.F., 2011. Effects of watershed configuration and composition on downstream lake water quality. J. Environ. Qual. 40, 517–527. https://doi.org/10.2134/jeq2010.0133.
Grant, R., Laubel, A., Kronvang, B., Andersen, H.E., Svendsen, L.M., Fuglsang, A., 1996. Loss of dissolved and particulate phosphorus from arable catchments by subsurface drainage. Water Res. 30, 2633–2642. https://doi.org/10.1016/S0043-1354(96)00164-9.
Hanson, P.C., Stillman, A.B., Jia, X., Karpatne, A., Dugan, H.A., Carey, C.C., Stachelek, J., Ward, N.K., Zhang, Y., Read, J.S., Kumar, V., 2020. Predicting lake surface water phosphorus dynamics using process-guided machine learning. Ecol. Model. 430, 109136. https://doi.org/10.1016/j.ecolmodel.2020.109136.
Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., 2020. Array programming with NumPy. Nature 585, 357–362. https://doi.org/10.1038/s41586-020-2649-2.
Hastie, A., Lauerwald, R., Weyhenmeyer, G., Sobek, S., Verpoorter, C., Regnier, P., 2017. CO_2 evasion from boreal lakes: revised estimate, drivers of spatial variability, and future projections. Glob. Chang. Biol. 24, 711–728. https://doi.org/10.1111/gcb.13902.
Hengl, T., Reuter, H.I., 2008. Geomorphometry: concepts, software, applications. Developments in Soil Science. Elsevier, Amsterdam, Netherlands.
Hengl, T., de Jesus, J.M., MacMillan, R.A., Batjes, N.H., Heuvelink, G.B.M., Ribeiro, E., Samuel-Rosa, A., Kempen, B., Leenaars, J.G.B., Walsh, M.G., Gonzalez, M.R., 2014. SoilGrids1km — global soil information based on automated mapping. PLOS ONE 9, e105992. https://doi.org/10.1371/journal.pone.0105992.
Hengl, T., Nussbaum, M., Wright, M.N., Heuvelink, G.B.M., Gräler, B., 2018. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ 6, e5518. https://doi.org/10.7717/peerj.5518.
Hijmans, R.J., 2021. raster: geographic data analysis and modeling. https://CRAN.R-project.org/package=raster.
Holgerson, M.A., Raymond, P.A., 2016. Large contribution to inland water CO_2 and CH_4 emissions from very small ponds. Nat. Geosci. 9, 222–226. https://doi.org/10.1038/ngeo2654.
Holgerson, M.A., Farr, E.R., Raymond, P.A., 2017. Gas transfer velocities in small forested ponds. J. Geophys. Res. Biogeosci. 122, 1011–1021. https://doi.org/10.1002/2016JG003734.
Horn, B.K., 1981. Hill shading and the reflectance map. Proc. IEEE 69, 14–47. https://doi.org/10.1109/PROC.1981.11918.
Hotchkiss, E.R., Hall Jr., R.O., Sponseller, R.A., Butman, D., Klaminder, J., Laudon, H., Rosvall, M., Karlsson, J., 2015. Sources of and processes controlling CO_2 emissions change with the size of streams and rivers. Nat. Geosci. 8, 696. https://doi.org/10.1038/ngeo2507.
Humborg, C., Mörth, C.-M., Sundbom, M., Borg, H., Blenckner, T., Giesler, R., Ittekkot, V., 2010. CO_2 supersaturation along the aquatic conduit in Swedish watersheds as constrained by terrestrial respiration, aquatic respiration and weathering. Glob. Chang. Biol. 16, 1966–1978. https://doi.org/10.1111/j.1365-2486.2009.02092.x.
Huttunen, J.T., Lappalainen, K.M., Saarijärvi, E., Väisänen, T., Martikainen, P.J., 2001. A novel sediment gas sampler and a subsurface gas collector used for measurement of the ebullition of methane and carbon dioxide from a eutrophied lake. Sci. Total Environ. 266, 153–158. https://doi.org/10.1016/S0048-9697(00)00749-X.
Jackson, L.J., Lauridsen, T.L., Søndergaard, M., Jeppesen, E., 2007. A comparison of shallow Danish and Canadian lakes and implications of climate change. Freshw. Biol. 52, 1782–1792. https://doi.org/10.1111/j.1365-2427.2007.01809.x.
James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning. Springer, New York https://doi.org/10.1007/978-1-4614-7138-7.
Jankowski, K., Schindler, D.E., Lisi, P.J., 2014. Temperature sensitivity of community respiration rates in streams is associated with watershed geomorphic features. Ecology 95, 2707–2714. https://doi.org/10.1890/14-0608.1.
Janssen, A.B.G., Hilt, S., Kosten, S., de Klein, J.J.M., Paerl, H.W., Van de Waal, D.B., 2021. Shifting states, shifting services: linking regime shifts to changes in ecosystem services of shallow lakes. Freshw. Biol. 66, 1–12. https://doi.org/10.1111/fwb.13582.
Jeppesen, E., Søndergaard, M., Kronvang, B., Jensen, J.P., Svendsen, L.M., Lauridsen, T.L., 1999. Lake and catchment management in Denmark. In: Harper, D.M., Brierley, B., Ferguson, A.J.D., Phillips, G. (Eds.), The Ecological Bases for Lake and Reservoir Management. Springer, Dordrecht, Netherlands, pp. 419–432.
Jeppesen, E., Jensen, J.P., Sondergaard, M., Lauridsen, T., Landkildehus, F., 2000. Trophic structure, species richness and biodiversity in Danish lakes: changes along a phosphorus gradient. Freshw. Biol. 45, 201–218. https://doi.org/10.1046/j.1365-2427.2000.00675.x.
Jones, N.E., 2010. Incorporating lakes within the river discontinuum: longitudinal changes in ecological characteristics in stream–lake networks. Can. J. Fish. Aquat. Sci. 67, 1350–1362. https://doi.org/10.1139/F10-069.
Kalff, J., 2002. Limnology: Inland Water Ecosystems. Prentice Hall, New Jersey.
Kalff, J., Knoechel, R., 1978. Phytoplankton and their dynamics in oligotrophic and eutrophic lakes. Annu. Rev. Ecol. Syst. 9, 475–495. https://doi.org/10.1146/annurev.es.09.110178.002355.
Kirk, J.T., 1994. Light and Photosynthesis in Aquatic Ecosystems. Cambridge University Press, Cambridge.
Köhler, S., Buffam, I., Jonsson, A., Bishop, K., 2002. Photochemical and microbial processing of stream and soil water dissolved organic matter in a boreal forested catchment in northern Sweden. Aquat. Sci. 64, 269–281. https://doi.org/10.1007/s00027-002-8071-z.
Kragh, T., Sand-Jensen, K., 2018. Carbon limitation of lake productivity. Proc. R. Soc. B Biol. Sci. 285, 20181415. https://doi.org/10.1098/rspb.2018.1415.
Krause-Jensen, D., Sand-Jensen, K., 1998. Light attenuation and photosynthesis of aquatic plant communities. Limnol. Oceanogr. 43, 396–407. https://doi.org/10.4319/lo.1998.43.3.0396.
Kronvang, B., Jeppesen, E., Conley, D.J., Søndergaard, M., Larsen, S.E., Ovesen, N.B., Carstensen, J., 2005. Nutrient pressures and ecological responses to nutrient loading reductions in Danish streams, lakes and coastal waters. J. Hydrol. Nutr. Mobil. River Basins Eur. Perspect. 304, 274–288. https://doi.org/10.1016/j.jhydrol.2004.07.035.
Kronvang, B., Vagstad, N., Behrendt, H., Bøgestrand, J., Larsen, S.E., 2007. Phosphorus losses at the catchment scale within Europe: an overview. Soil Use Manag. 23, 104–116. https://doi.org/10.1111/j.1475-2743.2007.00113.x.
Kuhn, M., Wickham, H., 2021. recipes: preprocessing tools to create design matrices. https://CRAN.R-project.org/package=recipes.
Lam, S.K., Pitrou, A., Seibert, S., 2015. Numba: a llvm-based python jit compiler. Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC, pp. 1–6 https://doi.org/10.1145/2833157.2833162.
Lapierre, J.-F., Seekell, D.A., Filstrup, C.T., Collins, S.M., Emi Fergus, C., Soranno, P.A., Cheruvelil, K.S., 2017. Continental-scale variation in controls of summer CO_2 in United States lakes. J. Geophys. Res. Biogeosci. 122, 875–885. https://doi.org/10.1002/2016JG003525.
Laubel, A., Kronvang, B., Hald, A.B., Jensen, C., 2003. Hydromorphological and biological factors influencing sediment and phosphorus loss via bank erosion in small lowland rural streams in Denmark. Hydrol. Process. 17, 3443–3463. https://doi.org/10.1002/hyp.1302.
Lauerwald, R., Laruelle, G.G., Hartmann, J., Ciais, P., Regnier, P.A.G., 2015. Spatial patterns in CO_2 evasion from the global river network. Glob. Biogeochem. Cycles 29, 534–554. https://doi.org/10.1002/2014gb004941.
Lidberg, W., Nilsson, M., Lundmark, T., Ågren, A.M., 2017. Evaluating preprocessing methods of digital elevation models for hydrological modelling. Hydrol. Process. 31, 4660–4668. https://doi.org/10.1002/hyp.11385.
Lidberg, W., Nilsson, M., Ågren, A., 2020. Using machine learning to generate high-resolution wet area maps for planning forest management: a study in a boreal forest landscape. Ambio 49, 475–486. https://doi.org/10.1007/s13280-019-01196-9.
Liland, K.H., Mevik, B.-H., Wehrens, R., 2021. pls: partial least squares and principal component regression. https://CRAN.R-project.org/package=pls.
Lindsay, J.B., 2016. Efficient hybrid breaching-filling sink removal methods for flow path enforcement in digital elevation models. Hydrol. Process. 30, 846–857. https://doi.org/10.1002/hyp.10648.
Lindsay, J.B., Creed, I.F., 2005. Removal of artifact depressions from digital elevation models: towards a minimum impact approach. Hydrol. Process. 19, 3113–3126. https://doi.org/10.1002/hyp.5835.
Liu, S., Butman, D.E., Raymond, P.A., 2020. Evaluating CO_2 calculation error from organic alkalinity and pH measurement error in low ionic strength freshwaters. Limnol. Oceanogr. Methods 18, 606–622. https://doi.org/10.1002/lom3.10388.
Martinsen, K.T., Kragh, T., Sand-Jensen, K., 2020a. Carbon dioxide partial pressure and emission throughout the Scandinavian stream network. Glob. Biogeochem. Cycles 34, e2020GB006703. https://doi.org/10.1029/2020GB006703.
Martinsen, K.T., Kragh, T., Sand-Jensen, K., 2020b. Carbon dioxide efflux and ecosystem metabolism of small forest lakes. Aquat. Sci. 82, 9. https://doi.org/10.1007/s00027-019-0682-8.
Marx, A., Dusek, J., Jankovec, J., Sanda, M., Vogel, T., van Geldern, R., Hartmann, J., Barth, J.A.C., 2017. A review of CO_2 and associated carbon dynamics in headwater streams: a global perspective. Rev. Geophys. 55, 560–585. https://doi.org/10.1002/2016rg000547.
Messager, M.L., Lehner, B., Grill, G., Nedeva, I., Schmitt, O., 2016. Estimating the volume and age of water stored in global lakes using a geo-statistical approach. Nat. Commun. 7, 13603. https://doi.org/10.1038/ncomms13603.
Meyer, H., Pebesma, E., 2021. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. Methods Ecol. Evol. 12, 1620–1633. https://doi.org/10.1111/2041-210X.13650.
Meyer, H., Reudenbach, C., Hengl, T., Katurji, M., Nauss, T., 2018. Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. Environ. Model Softw. 101, 1–9. https://doi.org/10.1016/j.envsoft.2017.12.001.
Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., 2021. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (formerly: E1071). TU Wien.
Molnar, C., 2019. Interpretable Machine Learning. A Guide for Making Black Box Models Explainable.
Molnar, C., Bischl, B., Casalicchio, G., 2018. iml: an r package for interpretable machine learning. J.Open Source Softw. 3, 786. https://doi.org/10.21105/joss.00786.
Moreno-Mateos, D., Power, M.E., Comín, F.A., Yockteng, R., 2012. Structural and functional loss in restored wetland ecosystems. PLoS Biol. https://doi.org/10.1371/journal.pbio.1001247.
Neteler, M., Mitasova, H., 2013. Open Source GIS: A GRASS GIS Approach. Springer, New York.
Nielsen, A., Trolle, D., Søndergaard, M., Lauridsen, T.L., Bjerring, R., Olesen, J.E., Jeppesen, E., 2012. Watershed land use effects on lake water quality in Denmark. Ecol. Appl. 22, 1187–1200. https://doi.org/10.1890/11-1831.1.
Nobre, R.L.G., Caliman, A., Cabral, C.R., Araújo, F.de C., Guérin, J., Dantas, F.da C.C., Quesado, L.B., Venticinque, E.M., Guariento, R.D., Amado, A.M., Kelly, P., Vanni, M.J., Carneiro, L.S., 2020. Precipitation, landscape properties and land use interactively affect water quality of tropical freshwaters. Sci. Total Environ. 716, 137044. https://doi.org/10.1016/j.scitotenv.2020.137044.
O'Callaghan, J.F., Mark, D.M., 1984. The extraction of drainage networks from digital elevation data. Comput. Vision Graph. Image Process. 28, 323–344. https://doi.org/10.1016/S0734-189X(84)80011-0.
Oksanen, J., Sarjakoski, T., 2005. Error propagation analysis of DEM-based drainage basin delineation. Int. J. Remote Sens. 26, 3085–3102. https://doi.org/10.1080/01431160500057947.
Olden, J.D., Jackson, D.A., Peres-Neto, P.R., 2001. Spatial isolation and fish communities in drainage lakes. Oecologia 127, 572–585. https://doi.org/10.1007/s004420000620.

Olden, J.D., Lawler, J.J., Poff, N.L., 2008. Machine learning methods without tears: a primer for ecologists. Q. Rev. Biol. 83, 171–193. https://doi.org/10.1086/587826.
Pebesma, E., 2018. Simple features for R: standardized support for spatial vector data. R J. 10, 439–446. https://doi.org/10.32614/RJ-2018-009.
Pedersen, S.A.S., Hermansen, B., Nathan, C., Tougaard, L., 2011. Geological Survey of Denmark and Greenland (GEUS) Report. (in Danish).
Pekel, J.-F., Cottam, A., Gorelick, N., Belward, A.S., 2016. High-resolution mapping of global surface water and its long-term changes. Nature 540, 418–422. https://doi.org/10.1038/nature20584.
Peterson, G.D., Beard Jr., T.D., Beisner, B.E., Bennett, E.M., Carpenter, S.R., Cumming, G.S., Dent, C.L., Havlicek, T.D., 2003. Assessing future ecosystem services: a case study of the northern highlands lake district, Wisconsin. Conserv. Ecol. 7. https://doi.org/10.5751/ES-00557-070301.
Qin, B., Zhou, J., Elser, J.J., Gardner, W.S., Deng, J., Brookes, J.D., 2020. Water depth underpins the relative roles and fates of nitrogen and phosphorus in lakes. Environ. Sci. Technol. 54, 3191–3198. https://doi.org/10.1021/acs.est.9b05858.
Quinlan, R., Filazzola, A., Mahdiyan, O., Shuvo, A., Blagrave, K., Ewins, C., Moslenko, L., Gray, D.K., O'Reilly, C.M., Sharma, S., 2021. Relationships of total phosphorus and chlorophyll in lakes worldwide. Limnol. Oceanogr. 66, 392–404. https://doi.org/10.1002/lno.11611.
R Core Team, 2021. R: A Language and Environment for Statistical Computing Vienna, Austria.
Rapp Jr., G., Allert, J.D., Liukkonen, B.W., Ilse, J.A., Loucks, O.L., Glass, G.E., 1985. Acid deposition and watershed characteristics in relation to lake chemistry in northeastern Minnesota. Environ. Int. 11, 425–440. https://doi.org/10.1016/0160-4120(85)90226-0.
Rasmussen, J.B., Godbout, L., Schallenberg, M., 1989. The humic content of lake water and its relationship to watershed and lake morphometry. Limnol. Oceanogr. 34, 1336–1343. https://doi.org/10.4319/LO.1989.34.7.1336.
Read, E.K., Patil, V.P., Oliver, S.K., Hetherington, A.L., Brentrup, J.A., Zwart, J.A., Winters, K.M., Corman, J.R., Nodine, E.R., Woolway, R.I., Dugan, H.A., Jaimes, A., Santoso, A.B., Hong, G.S., Winslow, L.A., Hanson, P.C., Weathers, K.C., 2015. The importance of lake-specific characteristics for water quality across the continental United States. Ecol. Appl. 25, 943–955. https://doi.org/10.1890/14-0935.1.
Rebsdorf, A., Thyssen, N., Erlandsen, M., 1991. Regional and temporal variation in pH, alkalinity and carbon dioxide in Danish streams, related to soil type and land use. Freshw. Biol. 25, 419–435. https://doi.org/10.1111/j.1365-2427.1991.tb01386.x.
Riis, T., Sand-Jensen, K., 2001. Historical changes in species composition and richness accompanying perturbation and eutrophication of Danish lowland streams over 100 years. Freshw. Biol. 46, 269–280. https://doi.org/10.1046/j.1365-2427.2001.00656.x.
Rocher-Ros, G., Sponseller, R.A., Lidberg, W., Mörth, C.-M., Giesler, R., 2019. Landscape process domains drive patterns of CO_2 evasion from river networks. Limnol. Oceanogr. Lett. 4, 87–95. https://doi.org/10.1002/lol2.10108.
Sand-Jensen, K., Staehr, P.A., 2007. Scaling of pelagic metabolism to size, trophy and forest cover in small Danish lakes. Ecosystems 10, 128–142. https://doi.org/10.1007/s10021-006-9001-z.
Sand-Jensen, K., Staehr, P.A., 2009. Net heterotrophy in small Danish lakes: a widespread feature over gradients in trophic status and land cover. Ecosystems 12, 336–348. https://doi.org/10.1007/s10021-008-9226-0.
SDFE, 2021. Danish map supply. Accessed October 2021SDFE (Agency for Datasupply and Efficiency). https://download.kortforsyningen.dk/.
Sepp, M., Kõiv, T., Nõges, P., Nõges, T., Newell, S.E., McCarthy, M.J., 2022. Catchment soil characteristics predict organic carbon, nitrogen, and phosphorus levels in temperate lakes. Freshw .Sci. 41, 1–17. https://doi.org/10.1086/717954.
Shen, L.Q., Amatulli, G., Sethi, T., Raymond, P., Domisch, S., 2020. Estimating nitrogen and phosphorus concentrations in streams and rivers, within a machine learning framework. Sci. Data 7, 161. https://doi.org/10.1038/s41597-020-0478-7.
Smits, A.P., Schindler, D.E., Holtgrieve, G.W., Jankowski, K.J., French, D.W., 2017. Watershed geomorphology interacts with precipitation to influence the magnitude and source of CO_2 emissions from Alaskan streams. J. Geophys. Res. Biogeosci. 122, 1903–1921. https://doi.org/10.1002/2017jg003792.
Sobek, S., 2011. Predicting the depth and volume of lakes from map-derived parameters. Inland Waters 1, 177–184. https://doi.org/10.5268/IW-1.3.426.
Staehr, P.A., Baastrup-Spohr, L., Sand-Jensen, K., Stedmon, C., 2012. Lake metabolism scales with lake morphometry and catchment conditions. Aquat. Sci. 74, 155–169. https://doi.org/10.1007/s00027-011-0207-6.
Stanley, E.H., Collins, S.M., Lottig, N.R., Oliver, S.K., Webster, K.E., Cheruvelil, K.S., Soranno, P.A., 2019. Biases in lake water quality sampling and implications for macroscale research. Limnol. Oceanogr. 64, 1572–1585. https://doi.org/10.1002/lno.11136.
Taranu, Z.E., Gregory-Eaves, I., 2008. Quantifying relationships among phosphorus, agriculture, and lake depth at an inter-regional scale. Ecosystems 11, 715–725. https://doi.org/10.1007/s10021-008-9153-0.
Therneau, T., Atkinson, B., 2019. rpart: recursive partitioning and regression trees. https://CRAN.R-project.org/package=rpart.
Toming, K., Kotta, J., Uuemaa, E., Sobek, S., Kutser, T., Tranvik, L.J., 2020. Predicting lake dissolved organic carbon at a global scale. Sci. Rep. 10, 8471. https://doi.org/10.1038/s41598-020-65010-3.
Trolle, D., Staehr, P.A., Davidson, T.A., Bjerring, R., Lauridsen, T.L., Søndergaard, M., Jeppesen, E., 2012. Seasonal dynamics of CO_2 flux across the surface of shallow temperate lakes. Ecosystems 15, 336–347. https://doi.org/10.1007/s10021-011-9513-z.
Van Rossum, G., Drake, F.L., 2011. The Python Language Reference Manual. Network Theory Ltd., Massachusetts.
Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics With s. 4th ed. Springer, New York.
Wagner, T., Soranno, P.A., Cheruvelil, K.S., Renwick, W.H., Webster, K.E., Vaux, P., Abbitt, R.J., 2008. Quantifying sample biases of inland lake sampling programs in relation to lake surface area and land use/cover. Environ. Monit. Assess. 141, 131–147. https://doi.org/10.1007/s10661-007-9883-z.
Wallin, M.B., Öquist, M.G., Buffam, I., Billett, M.F., Nisell, J., Bishop, K.H., 2011. Spatiotemporal variability of the gas transfer coefficient (KCO_2) in boreal streams: Implications for large scale estimates of CO_2 evasion. Glob. Biogeochem. Cycles 25. https://doi.org/10.1029/2010gb003975.
Weyhenmeyer, G.A., Kortelainen, P., Sobek, S., Müller, R., Rantakari, M., 2012. Carbon dioxide in boreal surface waters: a comparison of lakes and streams. Ecosystems 15, 1295–1307. https://doi.org/10.1007/s10021-012-9585-4.
Wood, S.N., 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. J. R. Stat. Soc. Ser. B 73, 3–36. https://doi.org/10.1111/j.1467-9868.2010.00749.x.
Wood, S.N., 2017. Generalized Additive Models: An Introduction With R. CRC Press, Florida.
Wright, M.N., Ziegler, A., 2017. ranger: a fast implementation of random forests for high dimensional data in C++ and R. J. Stat. Softw. 77, 1–17. https://doi.org/10.18637/jss.v077.i01.
Yeo, I.-K., Johnson, R.A., 2000. A new family of power transformations to improve normality or symmetry. Biometrika 87, 954–959. https://doi.org/10.1093/biomet/87.4.954.
Yuan, L.L., Jones, J.R., 2020. Rethinking phosphorus–chlorophyll relationships in lakes. Limnol. Oceanogr. 65, 1847–1857. https://doi.org/10.1002/lno.11422.
Zevenbergen, L.W., Thorne, C.R., 1987. Quantitative analysis of land surface topography. Earth Surf. Process. Landf. 12, 47–56. https://doi.org/10.1002/esp.3290120107.