sustainability

MDPI

Article

Watershed-Based Evaluation of Automatic Sensor Data: Water Quality and Hydroclimatic Relationships

Jacopo Cantoni *, Zahra Kalantari and Georgia Destouni

Department of Physical Geography and Bolin Center for Climate Research, Stockholm University, SE-106 91 Stockholm, Sweden; zahra.kalantari@natgeo.su.se (Z.K.); georgia.destouni@natgeo.su.se (G.D.)
* Correspondence: jacopo.cantoni@natgeo.su.se

Received: 1 November 2019; Accepted: 31 December 2019; Published: 3 January 2020

Abstract: Water is a fundamental resource and, as such, the object of multiple environmental policies requiring systematic monitoring of its quality as a main management component. Automatic sensors, allowing for continuous monitoring of various water quality variables at high temporal resolution, offer new opportunities for enhancement of essential water quality data. This study investigates the potential of sensor-measured data to improve understanding and management of water quality at watershed level. Self-organizing data maps, non-linear canonical correlation analysis, and linear regressions are used to assess the relationships between multiple water quality and hydroclimatic variables for the case study of Lake Mälaren in Sweden, and its total catchment and various watersheds. The results indicate water discharge from dominant watersheds into a lake, and lake water temperature as possible proxies for some key water quality variables in the lake, such as blue-green algae; the latter is, in turn, identified as a potential good proxy for lake concentration of total nitrogen. The relationships between water discharges into the lake and lake water quality dynamics identify the dominant contributing watersheds for different water quality variables. Seasonality also plays an important role in determining some possible proxy relationships and their usefulness for different parts of the year.

Keywords: water quality; water discharge; hydroclimate; data mining; automatic sensor; monitoring; watershed; lake Mälaren; Stockholm region

1. Introduction

Availability of clean water is critical for life and human societies. Systematic monitoring of water availability and quality is a key management requirement, for example, in global sustainable development goal (SDG) 6 of the United Nations [1] and the European Water Framework Directive (WFD) [2]. Nevertheless, water quality monitoring is still largely lacking, years after WFD implementation and even in countries at the forefront of environmental management, such as Sweden [3]. A general need for improved and more efficient water monitoring is recognized [4], and the rapidly growing field of information and communication technology [5] can play a key role in meeting this challenge.

For example, new opportunities for meeting water monitoring needs may be provided by the Internet of Things (IoT). The IoT concept is about connecting identifiable devices that can observe the world and interact with other devices, with information and communication networks that can support and inform decision-making [5]. This study tests the potential usefulness of relatively cheap automatic sensors for continuous water quality monitoring, which can, e.g., be part of an IoT concept for water resources management. In testing, we considered the example case of the whole regional catchment and (sub)watersheds within it that feed water and waterborne tracers, nutrients, and

possible pollutants into Lake Mälaren (Figure 1), the third-largest lake in Sweden and the main water supply for 1.7 million people living in and around the Swedish capital Stockholm [6].

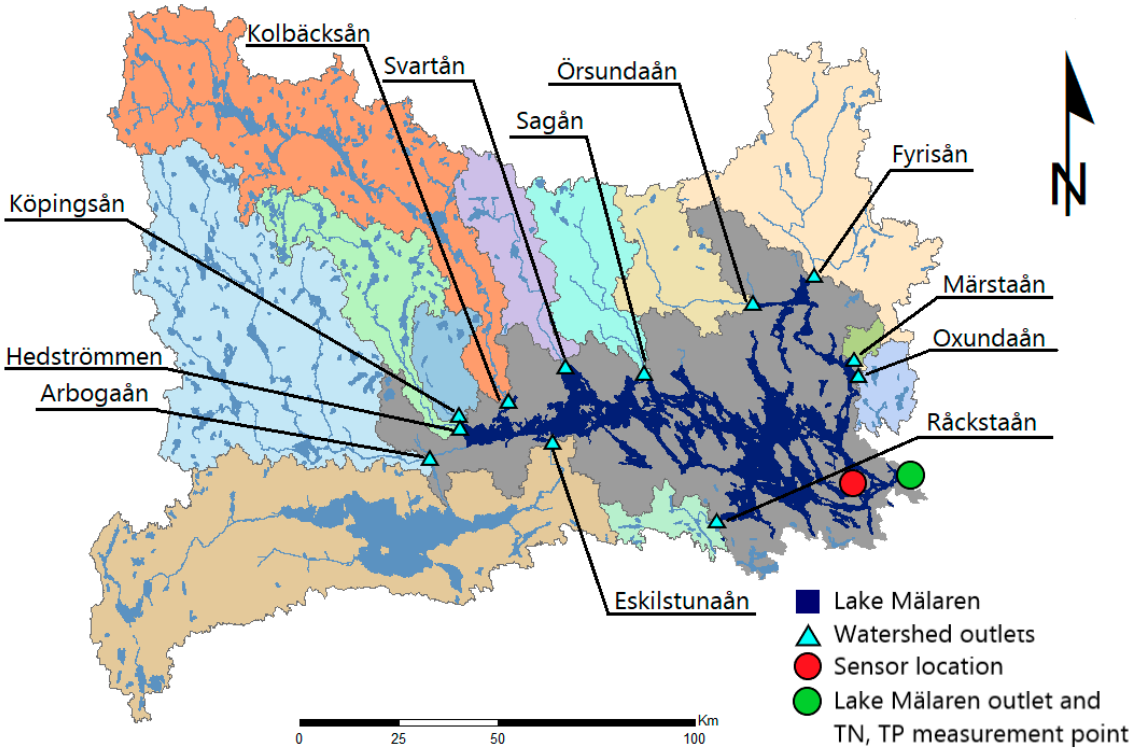

Figure 1. Map of Lake Mälaren showing the location of the automatic sensor (red dot) and the total catchment area of the lake. The total catchment is divided into unmonitored parts (gray) and monitored (sub)watersheds (colored fields with names), defined by the corresponding watershed outlets (light blue triangles). The Lake Mälaren outlet to the Baltic Sea (green dot) is also the monitoring point for total nitrogen (TN) and total phosphorus (TP) data.

To address this aim, we analyzed available data from an automatic water quality sensor installed in Lake Mälaren (red circle, Figure 1) and measuring a range of variables with high time resolution (minute data), but also with some hiatuses, over the year between October 2016 and October 2017. The sensor-measured water quality variables are: water temperature (T_w, °C), electrical conductivity (EC, uS/cm), pH, oxidation reduction potential (ORP, mV), dissolved oxygen (ODO, %), turbidity (FNU), fluorescent dissolved organic matter (fDOM, RFU), chlorophyll (Chl, ug/L), and blue-green algae (BGA, ug/L). In addition, using measured T_w and EC values, water salinity was calculated in terms of the concentration of total dissolved solids (TDS, mg/L). In order to explore the relationships between these sensor variables and a basic set of independently measured (i) hydroclimatic variables, and (ii) additional (nutrient-related) water quality variables of main water resource management importance in the region, we also examined data on (i) water discharge (Q, m^3/s) into Lake Mälaren from its whole catchment and individual watersheds within it (Figure 1), air temperature (T_a, °C) and precipitation (P, mm/day) over the catchment and individual watersheds, and (ii) concentrations of waterborne total nitrogen (TN, ug/L) and total phosphorus (TP, ug/L) measured at the Lake Mälaren outlet (green circle, Figure 1).

The relationships between all these variables are investigated in order to assess whether, and to what degree, variables of high environmental relevance, e.g., for WFD-required ecological status assessment (such as Chl, BGA, TN, TP), relate to spatiotemporal variations in other water quality and/or hydroclimatic variables that are monitored more simply, cheaply, or over longer time periods. If strong relationships can be found, they can provide an important novel way of using already available

long-term data or intensively monitored new data as proxies for more complex water quality variables. New data on these variables from automatic sensors can further feed into cloud communication and Artificial Intelligence interpretation algorithms for direct, effective guidance of water resources management. Moreover, spatial assessment of strong relationships between hydroclimatic drivers in contributing watersheds and water quality variables in receiving lakes can identify critical watershed conditions and contributions for resulting lake water quality, which can also guide adaptation for projected future hydroclimatic scenarios. To determine these spatiotemporal relationships, a set of statistical tools is used in this study. Beyond simple linear regressions, more advanced tools include Non-Linear Canonical Correlation Analysis (NLCCA), a generalization of Canonical Correlation Analysis (CCA) [7], which is used to analyze the complexity of interrelations among sets of variables, and Self-organizing Maps (SOM) [8], which are used for clustering temporal data relations over the study period and measurement hiatuses within it.

2. Materials and Methods

2.1. Study Area

Lake Mälaren covers an area of 1074 km^2 [9] and its total hydrological catchment has an area of 22,640 km^2 [10], with a total population of 2,074,844 [11]. The lake has 12 main tributaries, the monitored watersheds of which are shown in color in Figure 1. The sum of discharge from these watersheds accounts for 84% of total inflow to the lake, while the remaining inflow comes from unmonitored parts (gray in Figure 1) within the lake's total regional catchment [12]. Table 1 lists some main characteristics of the monitored watersheds.

Table 1. Characteristics of the monitored watersheds within the total Lake Mälaren catchment. Temporal average values for each watershed are over the study period, with available data of about a year (352 days from October 2016 to October 2017). The top row shows corresponding spatial sums or area-weighted average values over all monitored watersheds (sums for area, discharge (Q), population; area-weighted averages for air temperature (T_a), precipitation (P), population density).

	Area [a] (km^2)	T_a [b] (C°)	P [b] (mm/year)	Q [c] (m^3/s)	Total Population [d/a] (persons)	Population Density [d/a] (persons/km^2)
Total sum or average	17,189	6.5	608	68.0	890,316	52
Arbogaån	3490	6.2	730	21.6	54,933	16
Eskilstunaån	4179	7.1	584	9.2	293,099	70
Fyrisån	2003	6.6	620	6.0	195,440	98
Hedstrommen	1048	6.2	694	5.2	7987	8
Kolbäcksån	3170	5.6	657	14.3	79,532	25
Köpingsån	377	6.8	584	1.3	21,204	56
Märstaån	79	7.0	584	0.4	31,913	406
Orsundaån	735	6.7	584	2.2	13,860	19
Oxundaån	272	7.3	584	1.0	112,732	415
Råckstaån	262	7.0	548	0.5	5133	20
Sagaån	856	6.6	548	3.8	27,502	32
Svartån	720	6.5	584	2.4	46,981	65

[a] [10] [b] [13] [c] [14] [d] [11].

2.2. Data

The time period with available sensor data is limited by the sensor operation period, which was 26 October 2016 to 12 October 2017. The water quality data obtained from the automatic sensor include the variables outlined in Section 1 of this paper: T_w, EC, pH, ORP (measured with a platinum electrode), ODO, FNU, fDOM, Chl. and BGA. The last three variables are measured based on the

fluorescent response of specific fractions of organic elements present in the water, with UV light used for DOM, an in vivo fluorescence technique used for BGA, and fluorescence at wavelength 435–470 nm used for Chl; further specific information for each variable measurement can be found at https://www.ysi.com/parameters. In addition, using the T_w and EC data, TDS concentration is calculated as (Equation (1) from [15]; Equation (2) from [16]):

$$EC_{25} = \frac{EC}{1 + [0.025(T_w - 25)]} \quad (1)$$

$$TDS = 0.55EC_{25} \quad (2)$$

where EC_{25} is normal EC at a temperature of 25 °C.

The temporal resolution of the available sensor data is at the scale of minutes. However, the hydroclimatic variables have only daily resolution. To be directly comparable, the finer-resolution sensor datasets are aggregated to daily data for each variable. Moreover, an important issue with the sensor data is that measurements were interrupted, due to technical sensor failure, on four occasions during 2017: 5–7 February, 22 April–11 May, 5–7 August, and 9–11 October. The three shorter hiatuses are excluded from the study, while the handling of the longer hiatus (22 April–11 May) is explained further in Section 2.3.1. Some of the water quality time series (Turbidity, Chl, BGA, Cond, and TDS) show apparent data discontinuities in the trajectories from before to after the longer hiatus. Further discussion on these discontinuities can be found in Section 3.1.

Plots of the available data time series of the sensor-based water quality variables are shown in Supplementary Material (SM), Figures S1–S12. Further descriptions of technicalities of the collection process for quality data are also provided in SM.

The independently measured hydroclimatic data represent conditions in the monitored watersheds within the total Lake Mälaren catchment, and are taken from the Swedish Meteorological and Hydrological Institute (SMHI), as referred to in Table 1. The P and T_a values are based on raster data provided by SMHI [13] that have been cropped and spatially averaged over the shape of each watershed. The Q data are provided by SMHI as daily time series for each watershed (http://vattenwebb.smhi.se/), based on combined measurements and model interpretations by the SMHI model Hydrological Predictions for the Environment (HYPE) [17]. Time series of T_a and Q over the study period and the total Lake Mälaren catchment are shown in Figure 2a,b, and the corresponding P time series is shown in Figure S13 in SM.

Furthermore, independently measured TN and TP concentrations close to the Lake Mälaren outlet to the Baltic Sea (green circle, Figure 1) are considered as reported from the official Swedish national water quality database by the Swedish University of Agricultural Sciences (SLU) (http://miljodata.slu.se/mvm/). The reported data are monthly, meaning that only 11 monthly data points are available for direct comparison with the sensor data. Time series of the TN and TP concentrations over the sensor data period are shown in Figures S13 and S14 in SM.

This study does not consider time lags between the data, because all chemical variables are measured at the same location by the same sensor, except TN and TP that are measured a few kilometers downstream. Furthermore, there is no particular need to consider time lags between the water quality variables and the water discharges into the lake (that integrate and represent the most downstream outcome of the hydroclimatic processes in the contributing watersheds), as the data analysis shows high correlation results even without any time lag consideration (see result Sections 3.2 and 3.3).

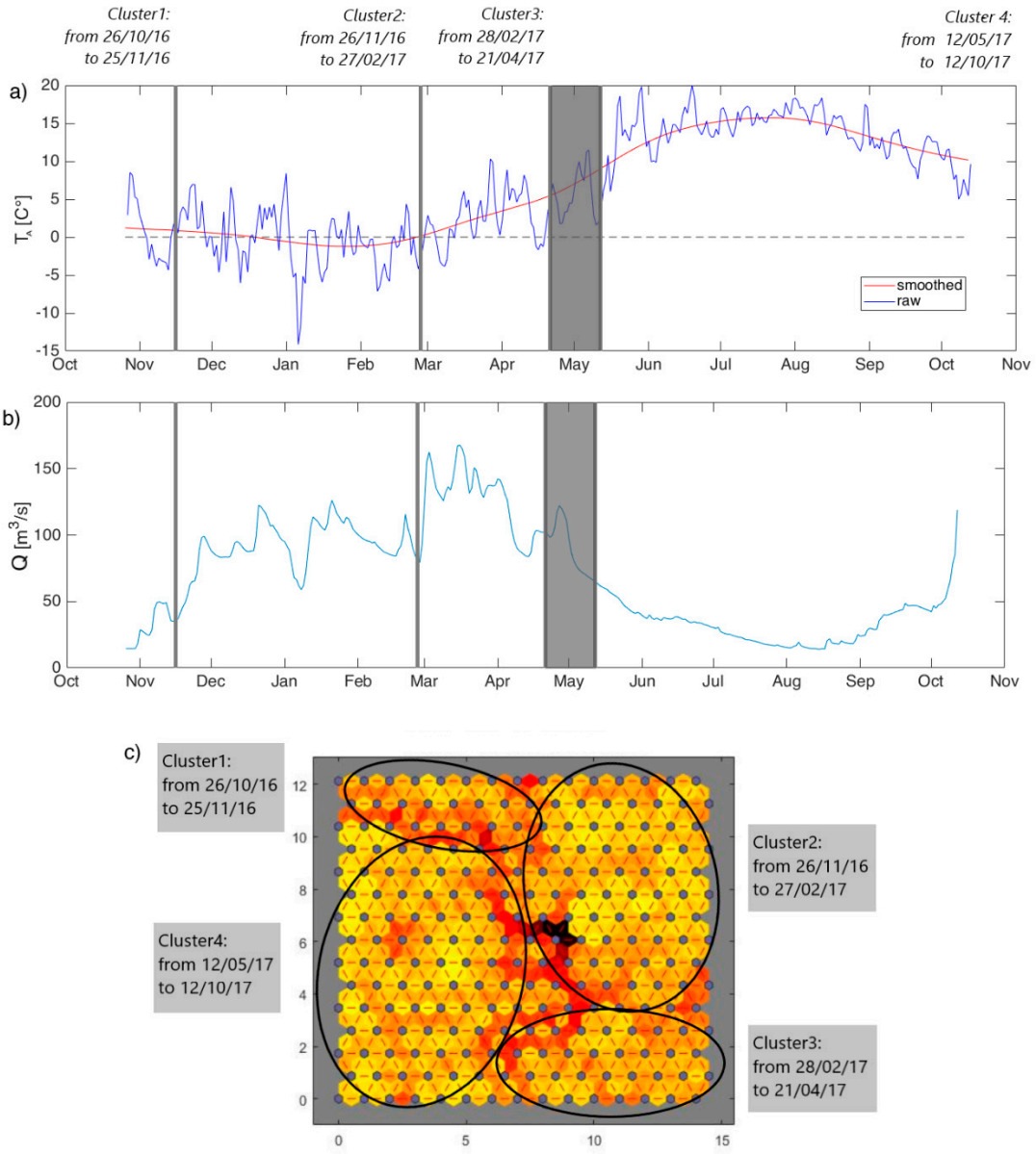

Figure 2. Time series of (**a**) mean air temperature (T_a) over the total Lake Mälaren catchment (blue line; smoothed time series shown by red line), and (**b**) total water discharge from all monitored watersheds within the Lake Mälaren catchment. The gray fields in (**a**,**b**) show the times of data hiatuses within the total sensor operation period October 2016–October 2017. (**c**) Two-dimensional representation of data distances in the self-organizing data-space map (SOM), leading to identification of four main temporal data clusters (Cluster 1: from 26 October to 25 November 2016, Cluster 2: from 26 November 2016 to 27 February 2017, Cluster 3: from 28 February to 21 April 2017, Cluster 4: from 12 May to 8 October 2017).

To summarize, the structure of the dataset used in this study comprises three different components:

- 352 × 10 (day × water quality parameter) data matrix of water quality measured with the automatic sensor
- 352 × 12 × 3 (day × watershed × hydroclimatic variable) data matrix of discharge, air temperature, and precipitation divided for each specific watershed
- 11 × 2 (day × nutrient) matrix of available data on TN and TP during the operation of the automatic sensor.

The number of days in the total time period of study is 352 with the hiatuses included and 323 without the hiatuses.

2.3. Methods

2.3.1. Data Clustering and Self-Organizing Map (SOM)

Visual inspection of the sensor data time series (Figures S1–S10 in SM) indicates the presence of possible temporal data clusters in different periods of the year, with notable differences between data values before and after the long sensor hiatus (22 April–11 May 2017). Based on that indication, a data clustering exploration step was applied using the SOM concept [18]. SOM algorithms have various applications, but are commonly used for identification of data clustering [19–21]. The underlying principle of SOM algorithms has been explained as: "Every input data item shall select the model that matches best with the input item, and this model, as well as a subset of its spatial neighbors in the grid, shall be modified for better matching" [22]. In the present case, we use the SOM concept to investigate if data value discontinuities depend on (and thus are explainable by and cluster according to) actual condition shifts from before to after the data hiatuses, or are indicative of potential measurement errors. To investigate the potential data self-organization in condition-consistent clusters, we use a map of 15 by 15 possible models (also referred to as nodes) and take as input to this model (node) network a series of 323 vectors (sensor operation days, excluding hiatuses). Each vector contains eight values of the sensor-measured water quality variables (excluding EC and TDS, as explained below), one value representing time, and 12 discharge values from the different watersheds. The variable values in each vector are normalized to be in the range 0–1, in order to make them directly comparable with each other in spite of their different magnitude ranges and units.

The EC and directly related TDS values are excluded from the clustering analysis because they exhibit both a particularly major drop and a highly different variability pattern from before to after the main April-to-May hiatus. More specifically, while the EC value exhibits an almost stationary behavior in the period October–March, it has much higher variability over the rest of the monitoring period. Even though EC values still remain within the freshwater range (50–1500 uS/cm) [23], this changed variability behavior may be indicative of potential measurement error; as such, EC and the derived TDS time series are excluded from the clustering analysis to avoid confounding the results for the potentially more reliable data for the other water quality variables. The Q values are included in the clustering analysis, as representative of the hydroclimatic conditions in the watersheds and because preliminary regression analysis indicated discharge as the hydroclimatic variable most closely related to several of the sensor-measured water quality variables.

For the data clustering analysis, the most important SOM result is a distance matrix (illustrated in Figure 2c), which visualizes in a two-dimensional space the Euclidean distance between various nodes of the SOM. This distance matrix shows clouds of relatively well-clustered (similar normalized) data values, and boundary regions between these where the data density (degree of similarity) is lower [8]. Note that small differences may occur in this network structure if the method is repeated due to an element of randomness intrinsic to the method itself.

2.3.2. Non-Linear Canonical Correlation Analysis (NLCCA)

NLCCA is a generalization of the underlying linear CCA method of multivariate analysis of interrelationships between two sets of variables, with each set aggregated into a canonical variable such that correlation can be analyzed between two canonical variables [7]. The limitation of linearity in CCA can be overcome by use of a neural network in creation of the canonical variables in NLCCA [24]. This study uses the Hsien version of NLCCA, with a code available under GNU General Public License for the Matlab environment [25]. The method allows for selection between the biweight midcorrelation and the Pearson correlation as methods of analysis of the relationship between the canonical variables. The biweight midcorrelation method is used in this analysis, based on its greater robustness [26].

The NLCCA method is further used to explore the potential relationships of each hydroclimatic variable for each watershed with the set of different water quality variables. The NLCCA analysis is then iterated three times, such that the coupling of the water quality dataset is at each time explored against a different hydroclimatic variable, i.e., the air temperature T_a, the discharge Q, or the precipitation P of each watershed. In each such iteration, the NLCCA is performed with a number of neuron layers varying from 1 to 4. Increasing the numbers of layers increases the NLCCA capacity to handle non-linearity, with use of just one single layer being equivalent to the linear CCA method [24].

2.3.3. Linear Regressions

Application of the NLCCA method in identification of hydroclimatic variable correlations with the sensor-measured set of water quality variables showed that linear correlations are mostly as good as or better than the studied non-linear correlations. Based on this result, the study focuses on quantifying the degree of linear correlation between pairs of variables. The linear correlation between each pair of variables is evaluated through a series of linear regressions, each of which relating the time series of: (a) a water quality variable and a watershed discharge; (b) two different water quality variables. The coefficient of determination is calculated for each linear regression as a key parameter expressing the degree of linear correlation. For each variable pair, statistical significance is tested for the obtained slope of the associated regression line by checking that zero slope is not included in the 95% confidence interval of slope values [27,28].

The variables considered and coupled in this regression analysis include the 10 sensor-measured water quality parameters (i.e., including also EC and TDS), the discharge Q from the 12 watersheds, and their total Q into Lake Mälaren. In addition, the degree of linear regression between the TN and TP data, measured at the Lake Mälaren outlet (green circle, Figure 1), is checked against all sensor-measured water quality variables and total Q. As mentioned in Section 2.2, the TN and TP concentrations are measured once per month, independently from the automatic sensor and its hiatuses. Therefore, it is possible to use more TN, TP, and Q data for statistical analysis, so that the linear regressions between each of the nutrients and Q are studied over a longer period (from 1 January 2014 to 26 February 2018).

3. Results and Discussion

3.1. SOM Application and Data Clustering

The SOM algorithm identifies four main data clusters (Figure 2c), each representing somewhat different data conditions than the others, including to contain data only from a unique associated time interval within the studied year. This temporal clustering suggests a seasonality effect on data conditions, which can at least to some degree explain data discontinuities from before to after a hiatus and is not a forced or predetermined result, as time is one of the data values included in each daily vector considered in the cluster analysis.

Figure 2c shows the resulting SOM distance mapping of the study data, with lighter and darker map areas representing shorter and longer Euclidean distances in the SOM space, respectively. The clearest boundary visible in the map is that starting in the top left corner and extending to the center of the map at the bottom (Figure 2c). On the left side is a relatively homogeneous area identified as a single main cluster (Cluster 4, from 12 May to 8 October 2017). Another main homogeneous map area stretches over the top three quarters of the right side of the map (identifying data Cluster 2: from 26 November 2016 to 27 February 2017). Two more area divisions in the top left and bottom right parts of the map identify two more data clusters (Cluster 1: from 26 October to 25 November 2016, and Cluster 3: from 28 February to 21 April 2017).

The visual indication of data clustering before (Clusters 1–3) and after (Cluster 4) the longest sensor hiatus (22 April–11 May 2017) is thus also supported by the SOM tool application, which, in addition, gives a further division of the data before that hiatus into Clusters 1–3. This result is relevant to solve an important question with regard to some of the water quality parameters. However, the period of the

year in which the hiatus occurs presents a challenge since, as shown in Figure 2a, the hiatus lies in the transition between the colder period of the year and the warmer period. In conjunction, as observed in Section 2.2, some of the water quality time series show strong discontinuities around the hiatus. It is difficult to define whether these discontinuities are due to sensor bias, or are a seasonal fluctuation due to the change from a colder to a warmer season. The identification by the SOM algorithm of a well-defined cluster boundary in association with the longer hiatus suggests that these discontinuities are more likely to be caused by seasonal effects, rather than by sensor bias. This is because this finding applies for all water quality indices, i.e., those with trajectory discontinuity and those without, together with the discharges not affected by the hiatus.

A comparison of Figure 2a,b shows how this time-based clustering relates to the temporal behavior of the two hydroclimatic variables air temperature T_a and water discharge Q. Specifically, average T_a values are distinctly higher, while average Q values are lower, in the data for Cluster 4 (after the longest hiatus) than for Clusters 1–3 (before that hiatus), with no such clear relation evident for precipitation P; see Figure S11 in SM). Cluster 2 roughly overlaps with a period of average T_a decrease to below 0 °C, while Clusters 1 and 3 are low-temperature periods but still remaining on average above 0 °C (Figure 2a). In terms of discharge, Clusters 1–3 represent times of relatively high and/or increasing Q values, while Cluster 4 is a period of low average Q (Figure 2b).

The data cluster results thus reflect seasonal hydroclimatic variations in the Lake Mälaren catchment, which have been reported to affect water quality [29–31]. Some studies propose use of time-dependent model parameters to analyze relationships between dissolved oxygen and water flows [32]. In this study, the SOM algorithm application led to consistent time division into statistically homogeneous seasonal periods, with possible different relationships among water quality variables and between these variables and the hydroclimatic variables.

The fine temporal resolution of available sensor data is important for facilitating clear identification of data clustering. In the present case study, this fine time resolution allows for 352-day division of the study period, providing a sufficient number of data for statistical analysis. A major limitation of the study is that the availability of data is limited to a single year. Further fine-resolution monitoring and research is thus needed to explore and test the seasonality effects and variable relationships observed in this study.

3.2. NLCCA Application

Figure 3 shows the results of the NLCCA application for the main (longest) period with continuous data, Clusters 2 and 4, expressed as the biweight midcorrelation between the studied canonical variables for different analysis modes. The mode numbering represents the number of layers used in the neural network model, with higher number indicating higher non-linearity in variable correlations. While it is difficult to make an exact physical interpretation of resulting correlation values, the highest values indicate the overall strongest water quality relationships with the studied hydroclimatic variables. Overall, for the main data Clusters 2 and 4 (and for the relatively minor Clusters 1 and 3, not shown), the results show that the discharge (Q) dataset has the strongest relationships with and across the water quality data (Figure 3). Air temperature, T_a, exhibits the next strongest relationships, while the precipitation (P) dataset often suffers from overfitting issues, even with just two neuron layers and especially for the warm Cluster 4 period. For both clusters, allowing the NLCCA method to account for higher degree of non-linearity does not increase the degree of correlation. Based on these findings, the further analysis of relationships between the Q dataset and water quality data focuses on the linear regression approach.

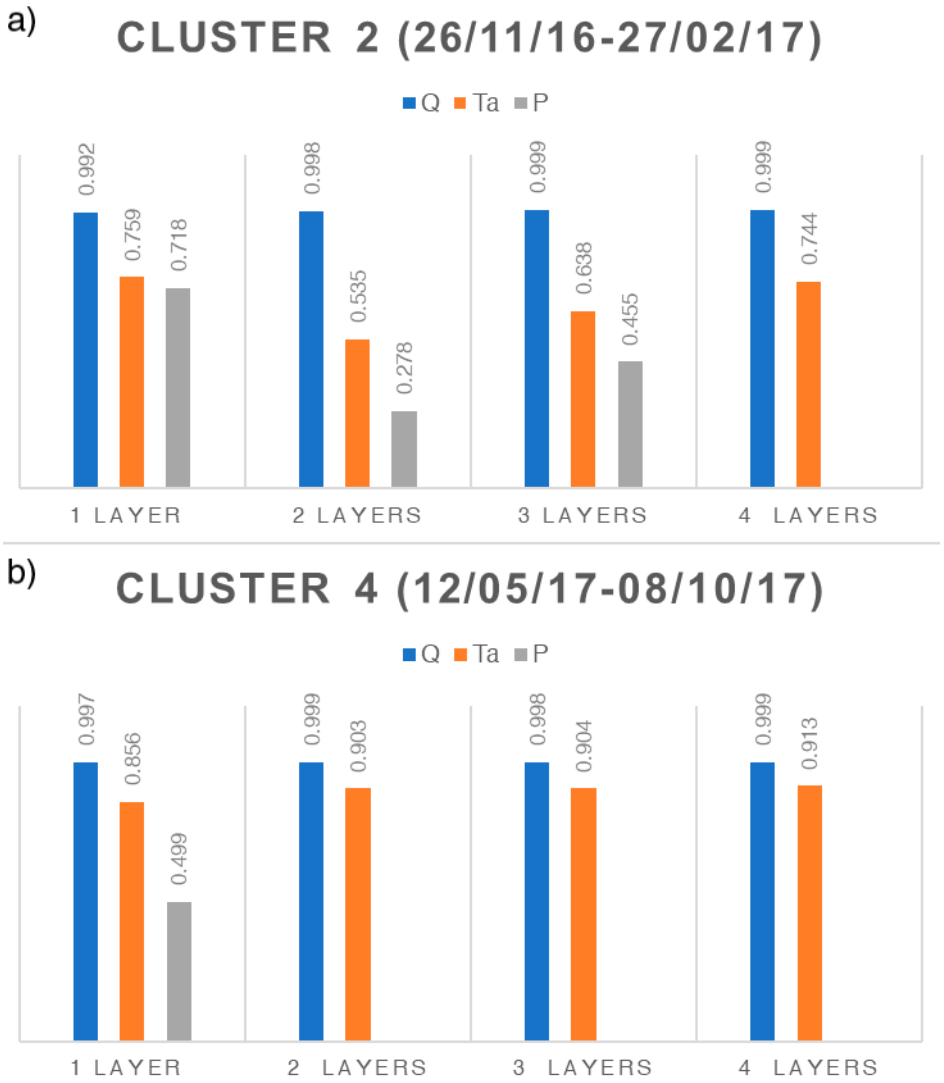

Figure 3. Canonical correlations between water quality indices and the studied hydroclimatic variables discharge (Q), air temperature (T_a), and precipitation (P) for (**a**) data Cluster 2 and (**b**) data Cluster 4. Higher model number implies higher number of layers in the neural network model and higher degree of non-linearity in variable correlations (missing values for P imply failed test on overfitting).

In general, the NLCCA application reveals a dominant role of Q as an overall water quality driver in comparison with the other studied hydroclimatic variables. It also reveals important differences between the Q relationship and the considerably weaker P relationship with water quality. Relationships between water discharge and various water quality variables have been studied previously [32–34]. In the more detailed linear regression analysis, the present study goes further in showing which specific water quality variables are pair-wise most strongly related with each other and with total and specific watershed Q around Lake Mälaren, as well as how seasonality affects these relationships.

3.3. Linear Regressions

Figures 4–6 show the coefficient of determination (R^2) for different pairs of the studied variables. Slope and intercept values for the linear regression equations between Q (from different watersheds) and each sensor-measured water quality variable, and between each possible pair of the latter, are also shown in Figures S16 and S18, respectively (see SM).

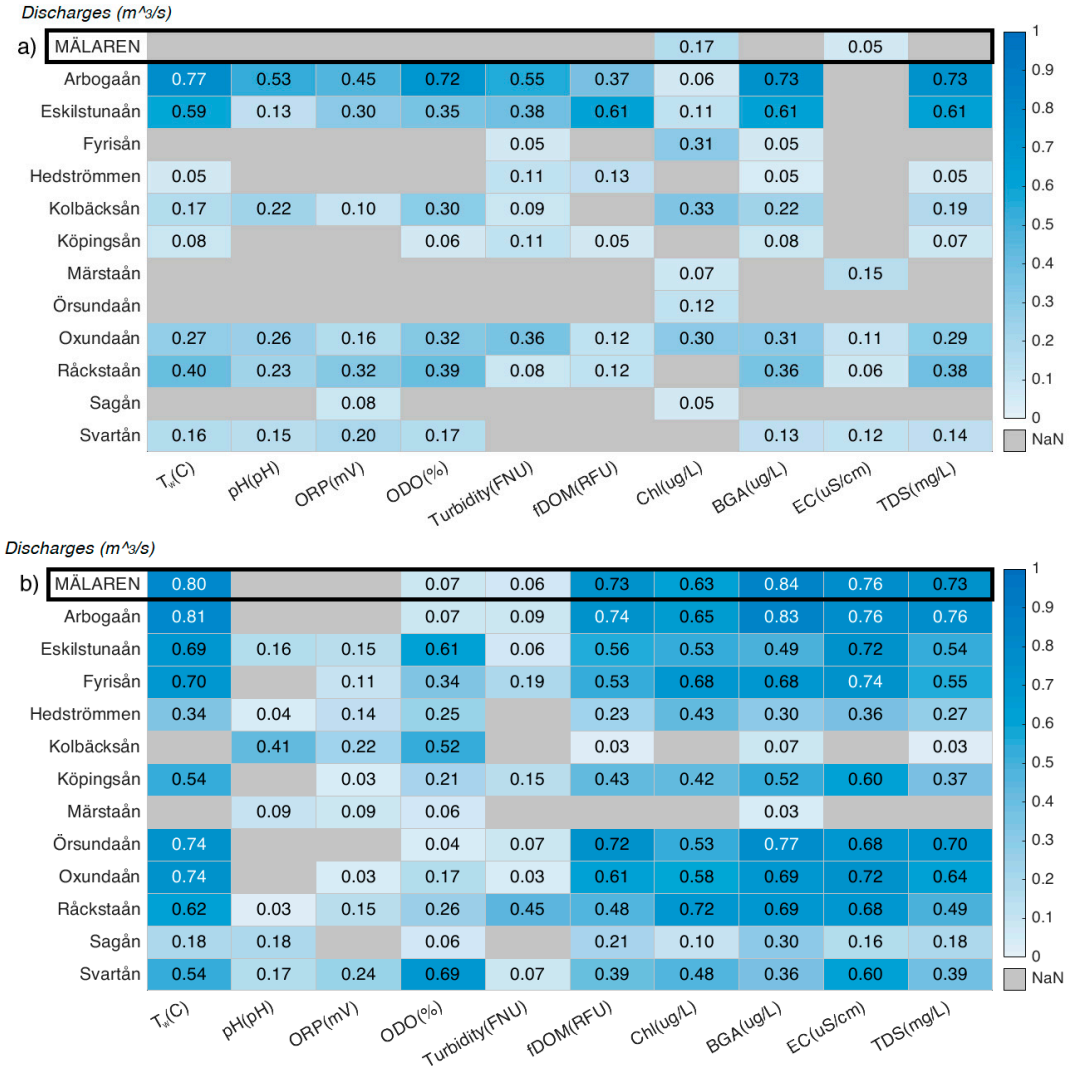

Figure 4. Coefficient of determination (R^2) values for linear regression lines between the data for total water discharge (Q) (top row), and Q from each individual watershed (rows below the top), into Lake Mälaren, and the different water quality data. (**a**) Data Cluster 2 (26 November 2016–27 February 2017). (**b**) Data cluster 4 (12 May–8 October 2017). Gray cells represent no significant regression line slope (0.05 significance level).

For Q from different watersheds and in total, the linear regression analysis reveals strong seasonal effects. Overall, correlations between Q and various water quality variables are higher in the warm Cluster 4 period (Figure 4b) than the cold Cluster 2 period (Figure 4a). This seasonal difference is apparent for total Q, and for Q from some individual watersheds. Among the watersheds, Arbogaån (with the highest individual Q) and Eskilstunaån (with the largest catchment area, the largest population in the catchment, in combination with the third largest Q) have the overall highest R^2 values. In the warm Cluster 4 period (Figure 4b), the total Q from all watersheds into Lake Mälaren exhibits mostly similar or somewhat lower correlations compared with Q from the most dominant watershed, Arbogaån. The correlations of Q from Arbogaån remain high ($R^2 \geq 0.7$) for the water quality variables T_w, BGA, and TDS in the cold Cluster 2 as in the warm Cluster 4 period; for ODO, the Arbogaån Q correlation is even much higher in Cluster 2 ($R^2 = 0.72$) than Cluster 4 ($R^2 = 0.07$). For the warm Cluster 4 period, correlations of total Q and Q from Arbogaån are also high ($R^2 \geq 0.6$) for fDOM, Chl, and EC.

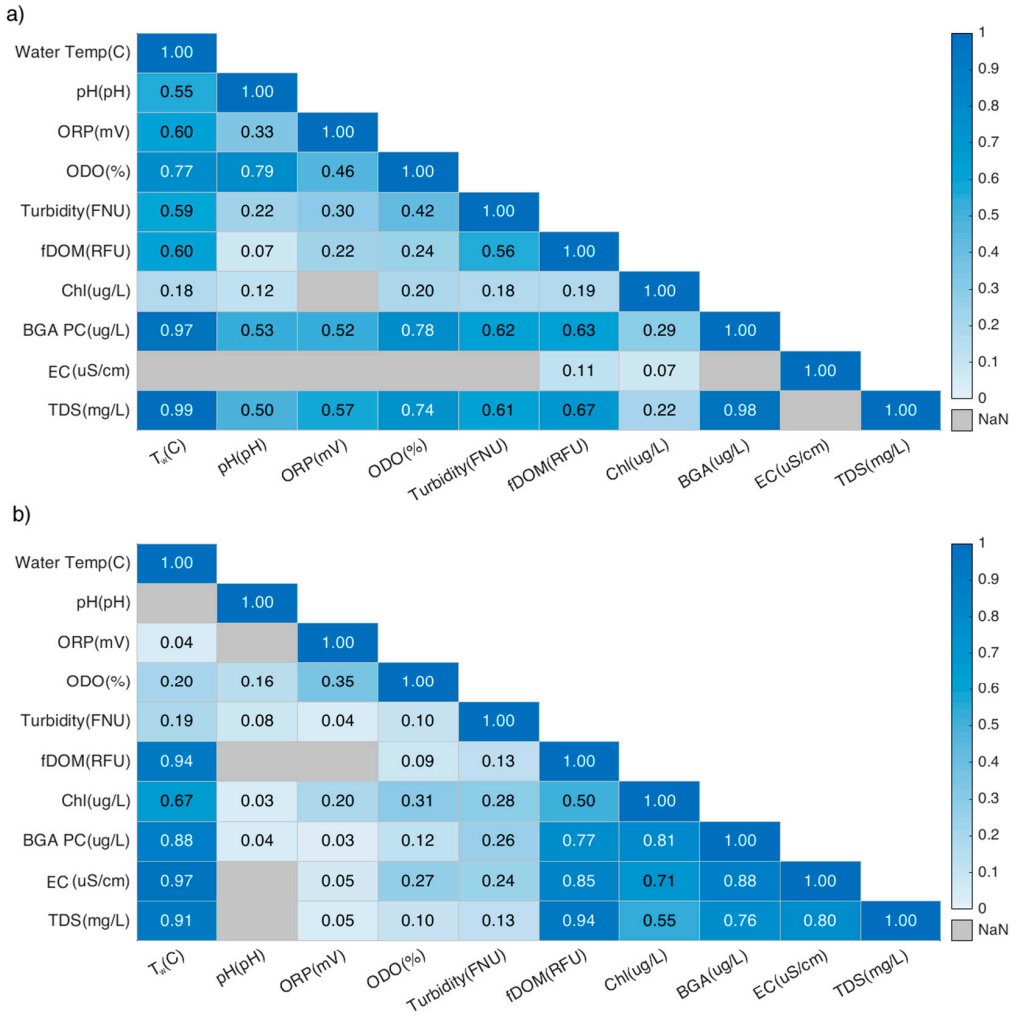

Figure 5. Coefficient of determination (R^2) values for cross-correlations between pairs of sensor-measured water quality data for (**a**) Cluster 2 (26 November 2016–27 February 2017) and (**b**) Cluster 4 (12 May–8 October 2017). Gray cells represent no significant regression line slope (0.05 significance level).

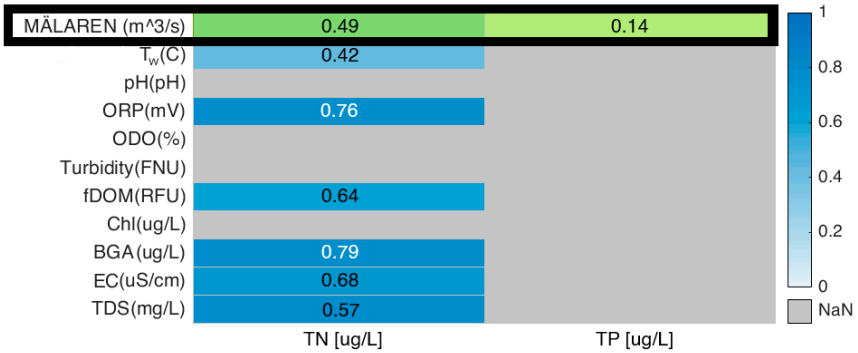

Figure 6. Coefficient of determination (R^2) for linear regressions between independently measured data for total nitrogen (TN) or total phosphorus (TP) concentration and data for total discharge (Q) (first row; for the extended study period, 1 January 2014–26 February 2018) or sensor-measured water quality data (for the basic study period 26 October 2016–8 October 2017). For regressions based on data over the basic study period, there is no significant correlation between TN or TP concentration and Q. For each study period, the regression analysis considers data for all days with available TN and TP measurements, without clustering.

Seasonal effects are also apparent in the regressions focusing on cross-correlations between water quality variables (Figure 5). Again, correlations are mostly higher for the warm Cluster 4 period than the cold Cluster 2 period. However, water temperature, T_w, exhibits relatively high correlation levels ($R^2 > 0.5$) also in the Cluster 2 period, with most other water quality variables except for Chl and EC. The pairs of water quality variables with the highest cross-correlations ($R^2 > 0.7$) are ODO–pH, BGA–ODO, ODO–TDS, and BGA–TDS in the cold Cluster 2 period, and nearly all pair combinations among fDOM, Chl, BGA, EC, and TDS (with the few exceptions still having $R^2 > 0.5$), and between all of these and T_w, in the warm Cluster 4 period. Overall, the water quality variables that exhibit relatively low linear correlation with (some dominant discharge) Q tend to be even less well correlated with the other water quality variables, especially in the warm Cluster 4 period.

The relatively readily and cheaply sensor-measured water quality data may or may not also correlate with TN and TP concentrations. On the one hand, loads of TN and TP are known to drive eutrophication and algae blooms [35–37], and are key variables for assessing hydro-ecological conditions and status [3,38]. On the other hand, nutrient load is the product of concentration and discharge, Q, with Q found to dominate nutrient load variability, while concentration levels are much less variable over time [3,39]. As such, nutrient concentrations may not correlate well with Q or with other water quality variables that are highly correlated with Q. The measured TP and TN concentration data are monthly values, meaning that there are only 11 data points available over the present study period. For these 11 data points, no significant linear correlation is evident between TP or TN and total Q, as a relevant discharge measure for the Lake Mälaren outlet. As the nutrient data are not measured with the automatic sensor, their correlation with Q is also tested over the longer time series from 1 January 2014 to 26 February 2018, with the results listed in the first row of Figure 6. While some considerable correlation is found between the TN concentration and total Q ($R^2 = 0.49$), the corresponding correlation for the TP concentration is much lower ($R^2 = 0.14$). Furthermore, the TN concentration exhibits high correlation with some of the sensor-measured water quality variables, including ORP and BGA (at $R^2 > 0.7$) and fDOM, EC, and TDS (at $R^2 > 0.5$), whereas no correlation is found between the TP concentration and the sensor-measured variables (Figure 6, with corresponding additional regression line parameters (slopes, intercepts) listed in Figure S18 in SM).

The capability of modern automatic sensors of recording concurrent time series of multiple water quality variables at high temporal resolution has led to meaningful seasonality analysis and results, even though the temporal extent of the available data time series is limited to only one year. The richness of the temporally high-resolved datasets still facilitates linear correlation analysis of a large number of daily data points (92 for cluster 2 and 147 for cluster 4), which are, in turn, based on an even larger number of finer-resolved observation data (SM Section 1). Although seasonality effects could only be studied over a single year, the findings are clear in that potentially useful proxy relationships, e.g., between water discharges and some water quality variables, or among some of the latter, cannot be expected to remain constant over any year, but vary seasonally over each year. More research, with multivariable datasets extending over longer time scales than just one year, is needed to further assess the details of seasonality effects, as well as the relevance and usefulness of potential proxy relationships for multivariable water quality evolution.

These regression results indicate a potential for possible use of some measured sensor data, such as ORP or BGA, as potential proxies for other water quality conditions, such as the monthly TN concentrations. With ORP and BGA data not exhibiting strong correlation with each other, especially under warm Cluster 4 conditions (Figure 5b), these would then be potential independent proxies. However, more investigations are needed to test this potential further, beyond the limited data availability and coarse temporal resolution of nutrient concentration data in the present study.

4. Conclusions

This study shows promising results for potential use of water discharge, Q, from dominant watersheds into a lake, and lake water temperature, T_w, as possible proxies for some key variables of

lake water quality, such as blue-green algae, BGA. The latter is, in turn, indicated as a possible good proxy for TN concentration in the lake. In general, the relationships between water discharges into the lake and lake water quality dynamics can identify the dominant contributing watersheds for different water quality variables. Furthermore, seasonality plays an important role in determining some other possible proxy relationships and their usefulness for different parts of the year (e.g., EC for fDOM, Chl, BGA in the warm season).

These findings call for further investigation and testing of their robustness, generality, and transferability, based on longer-term and more consistently time-resolved datasets for different study locations and scales around the world. If similar or additional proxies are found to be useful in new extensive studies, the results can be widely useful for: (a) expanding water quality monitoring based on proxy measurements by automatic sensors; (b) constructing better models and interpretation algorithms, e.g., in an IoT concept, based on data for a range of variables and accounting for important seasonality effects to obtain better water quality results; (c) identifying and directing effective pollution mitigation measures to dominant watershed sources and drivers of water quality changes; and, (d) assessing and deciding on adaptation and mitigation measures for future water quality based on projected hydroclimate and discharge scenarios.

Supplementary Materials: The following are available online at http://www.mdpi.com/2071-1050/12/1/396/s1. This supplemental material provides a description of the automatic sensor and the methodology used for water quality data collection, and a set of additional graphs and figures about the water quality data from the automatic sensor and other related data: Figures S1–S12, time series of sensor-measured water quality variables; Figure S13, Time series of independently measured catchment-average precipitation (P); Figures S14 and S15, Time series of independently measured nutrient concentrations; Figure S16, Slope and intercept values of linear regression lines between water discharge Q and water quality variables; Figure S17, Slope and intercept values of linear regression lines water quality variables; Figure S18, Slope and intercept values of linear regression lines between water discharge Q and nutrient concentrations.

Author Contributions: Conceptualization: J.C., Z.K. and G.D.; Data curation: J.C.; Formal analysis: J.C.; Supervision: Z.K. G.D.; Writing—original draft: J.C.; Writing—review & editing, Z.K. and G.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Swedish Innovation Agency (Vinnova), grant number 2018-01509.

Acknowledgments: We thank Stockholm Vatten och Avfall for access to the sensor-measured water quality data, SMHI for the hydroclimatic data, and SLU for the nutrient concentration data. We also thank William W. Hsieh for the GNU general public license to the Matlab code used to perform the NLCCA data analysis.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. UN. *General Assembly Transforming Our World: The 2030 Agenda for Sustainable Development*; United Nations: New York, NY, USA, 2015.
2. Council, E. Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for Community action in the field of water policy. *Off. J. L* **2000**, *327*, 1–73.
3. Destouni, G.; Fischer, I.; Prieto, C. Water quality and ecosystem management: Data-driven reality check of effects in streams and lakes. *Water Resour. Res.* **2017**, *53*, 6395–6406. [CrossRef]
4. Carvalho, L.; Mackay, E.B.; Cardoso, A.C.; Baattrup-Pedersen, A.; Birk, S.; Blackstock, K.L.; Borics, G.; Borja, A.; Feld, C.K.; Ferreira, M.T.; et al. Protecting and restoring Europe's waters: An analysis of the future development needs of the Water Framework Directive. *Sci. Total Environ.* **2019**, *658*, 1228–1238. [CrossRef]
5. Tsai, C.W.; Lai, C.F.; Chiang, M.C.; Yang, L.T. Data Mining for Internet of Things: A Survey. *IEEE Commun. Surv. Tutor.* **2014**, *16*, 77–97. [CrossRef]
6. Darracq, A.; Greffe, F.; Hannerz, F.; Destouni, G.; Cvetkovic, V. Nutrient transport scenarios in a changing Stockholm and Mälaren valley region, Sweden. *Water Sci. Technol.* **2005**, *51*, 31–38. [CrossRef] [PubMed]
7. Härdle, W.K.; Simar, L. *Applied Multivariate Statistical Analysis*; Springer: Berlin/Heidelberg, Germany, 2015; ISBN 978-3-662-45171-7.
8. Vesanto, J.; Sulkava, M. Distance Matrix Based Clustering of the Self-Organizing Map. In Proceedings of the ICANN, Madrid, Spain, 28–30 August 2002.

9. SMHI Fakta om Mälaren|SMHI. Available online: https://www.smhi.se/kunskapsbanken/hydrologi/fakta-om-malaren-1.5089 (accessed on 14 March 2019).
10. Sveriges Meteorologiska Och Hydrologiska Institut (SMHI) Delavrinningsområden SVAR_2016_3. 2016. Available online: https://www.smhi.se/data/hydrologi/sjoar-och-vattendrag/ladda-ner-data-fran-svenskt-vattenarkiv-1.20127 (accessed on October 2018).
11. Statistiska Centralbyrån SCB: Befolkning Vektor. 2018. Available online: https://www.zeus.slu.se (accessed on March 2019).
12. Ledesma, J.L.J. Dynamics of Color and Organic Carbon within the Mälaren Catchment. Available online: https://stud.epsilon.slu.se/2659/ (accessed on 18 April 2018).
13. Sveriges Meteorologiska Och Hydrologiska Institut (SMHI) SMHI Data. 2018. Available online: https://www.smhi.se/data (accessed on October 2018).
14. Sveriges Meteorologiska Och Hydrologiska Institut (SMHI) Modelldata Per Område. 2018. Available online: https://vattenwebb.smhi.se/modelarea/ (accessed on October 2018).
15. Hayashi, M. Temperature-Electrical Conductivity Relation of Water for Environmental Monitoring and Geophysical Data Inversion. *Environ. Monit. Assess.* **2004**, *96*, 119–128. [CrossRef] [PubMed]
16. Atekwana, E.A.; Atekwana, E.A.; Rowe, R.S.; Werkema, D.D.; Legall, F.D. The relationship of total dissolved solids measurements to bulk electrical conductivity in an aquifer contaminated with hydrocarbon. *J. Appl. Geophys.* **2004**, *56*, 281–294. [CrossRef]
17. 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. *Hydrol. Res.* **2010**, *41*, 295–319. [CrossRef]
18. Kohonen, T. *Self-Organizing Maps*; Springer: Berlin, Germany, 2001; ISBN 978-3-642-56927-2.
19. Kalteh, A.M.; Hjorth, P.; Berndtsson, R. Review of the self-organizing map (SOM) approach in water resources: Analysis, modelling and application. *Environ. Model. Softw.* **2008**, *23*, 835–845. [CrossRef]
20. Lu, R.-S.; Lo, S.-L. Diagnosing reservoir water quality using self-organizing maps and fuzzy theory. *Water Res.* **2002**, *36*, 2265–2274. [CrossRef]
21. Nguyen, T.T.; Kawamura, A.; Tong, T.N.; Nakagawa, N.; Amaguchi, H.; Gilbuena, R. Clustering spatio–seasonal hydrogeochemical data using self-organizing maps for groundwater quality assessment in the Red River Delta, Vietnam. *J. Hydrol.* **2015**, *522*, 661–673. [CrossRef]
22. Kohonen, T. Essentials of the self-organizing map. *Neural Netw.* **2013**, *37*, 52–65. [CrossRef] [PubMed]
23. Behar, S. *Testing the Waters: Chemical and Physical Vital Signs of a River*; River Watch Network; United States: New York, NY, USA, 1997; ISBN 978-0-7872-3492-8.
24. Hsieh, W.W. Nonlinear canonical correlation analysis by neural networks. *Neural Netw.* **2000**, *13*, 1095–1105. [CrossRef]
25. NeuMATS Neuralnets for Multivariate and Time Series Analysis (NeuMATSA). Available online: http://www.ocgy.ubc.ca/projects/clim.pred/download.html (accessed on October 2019).
26. Cannon, A.J.; Hsieh, W.W. Robust nonlinear canonical correlation analysis: Application to seasonal climate forecasting. *Nonlinear Process. Geophys.* **2008**, *15*, 221–232. [CrossRef]
27. McKillup, S.; Dyar, M.D. *Geostatistics Explained: An Introductory Guide for Earth Scientists*; Cambridge University Press: Cambridge, UK; New York, NY, USA, 2010; ISBN 978-0-521-76322-6.
28. Harris, R.; Jarvis, C. *Statistics in Geography and Environmental Science*; Pearson Education Limited: Harlow, Essex, UK; New York, NY, USA, 2011; ISBN 978-0-13-178933-3.
29. Li, L.; He, Z.; Li, Z.; Zhang, S.; Li, S.; Wan, Y.; Stoffella, P.J. Spatial and temporal variation of nitrogen concentration and speciation in runoff and storm water in the Indian River watershed, South Florida. *Environ. Sci. Pollut. Res.* **2016**, *23*, 19561–19569. [CrossRef]
30. Lintern, A.; Webb, J.A.; Ryu, D.; Liu, S.; Bende-Michl, U.; Waters, D.; Leahy, P.; Wilson, P.; Western, A.W. Key factors influencing differences in stream water quality across space. *Wiley Interdiscip. Rev. Water* **2018**, *5*, e1260. [CrossRef]
31. Singh, S.; Dash, P.; Sankar, M.S.; Silwal, S.; Lu, Y.; Shang, P.; Moorhead, R.J. Hydrological and Biogeochemical Controls of Seasonality in Dissolved Organic Matter Delivery to a Blackwater Estuary. *Estuar. Coasts* **2019**, *42*, 439–454. [CrossRef]
32. Monteiro, M.; Costa, M. A Time Series Model Comparison for Monitoring and Forecasting Water Quality Variables. *Hydrology* **2018**, *5*, 37. [CrossRef]

33. Johnson, T.; Butcher, J.; Deb, D.; Faizullabhoy, M.; Hummel, P.; Kittle, J.; McGinnis, S.; Mearns, L.O.; Nover, D.; Parker, A.; et al. Modeling Streamflow and Water Quality Sensitivity to Climate Change and Urban Development in 20 U.S. Watersheds. *JAWRA J. Am. Water Resour. Assoc.* **2015**, *51*, 1321–1341. [CrossRef]
34. Hanslík, E.; Marešová, D.; Juranová, E.; Vlnas, R. Dependence of Selected Water Quality Parameters on Flow Rates at River Sites in the Czech Republic. *Water Environ. Syst.* **2016**, *4*, 127–140.
35. Paerl, H.W.; Fulton, R.S.; Moisander, P.H.; Dyble, J. Harmful freshwater algal blooms, with an emphasis on cyanobacteria. *Sci. World J.* **2001**, *1*, 76–113. [CrossRef] [PubMed]
36. Castillo, C.R.; Sarmento, H.; Álvarez-Salgado, X.A.; Gasol, J.M.; Marraséa, C. Production of chromophoric dissolved organic matter by marine phytoplankton. *Limnol. Oceanogr.* **2010**, *55*, 446–454. [CrossRef]
37. Mao, X.-F.; Wang, C.; Wei, X.; Chen, Q.; Liu, G. The Distribution of Chlorophyll-a and Its' Correlation with Related Indicators in the Ulansuhai Lake, China. *J. Environ. Account. Manag.* **2014**, *2*, 123–131. [CrossRef]
38. Bond, N.R.; Kennard, M.J. Prediction of Hydrologic Characteristics for Ungauged Catchments to Support Hydroecological Modeling. *Water Resour. Res.* **2017**, *53*, 8781–8794. [CrossRef]
39. Destouni, G.; Jarsjö, J. Zones of untreatable water pollution call for better appreciation of mitigation limits and opportunities. *Wiley Interdiscip. Rev. Water* **2018**, *5*, e1312. [CrossRef]