water

MDPI

Article

Investigating Water Quality Data Using Principal Component Analysis and Granger Causality

Maryam Zavareh [1,*], Viviana Maggioni [1] and Vadim Sokolov [2]

[1] Department of Civil, Environmental and Infrastructure Engineering, George Mason University, Fairfax, VA 22030, USA; vmaggion@gmu.edu
[2] Department of System Engineering and Operational Research, George Mason University, Fairfax, VA 22030, USA; vsokolov@gmu.edu
* Correspondence: mzavareh@masonlive.gmu.edu

Abstract: This work investigates the inter-relationships among stream water quality indicators, hydro-climatic variables (e.g., precipitation, river discharge), and land characteristics (e.g., soil type, land use), which is crucial to developing effective methods for water quality protection. The potential of using statistical tools, such as Principal Component (PC) and Granger causality analyses, for this purpose is assessed across 10 watersheds in the Eastern United States. The PC analysis shows consistency across the ten locations, with most of the variation explained by the first two PCs, except for the least developed watershed that presents three PCs. Results show that stronger Granger causality relationships and correlation coefficients are identified when considering a lag of one day, compared to longer lags. This is mainly due to the watersheds' limited size and, thus, their fast hydrological response. The strongest Granger causalities are observed when water temperature and dissolved oxygen concentration are considered as the effect of the other variables, which corroborates the importance of these two water properties. This work also demonstrates how watershed size and land use can impact causalities between hydrometeorological variables and water quality, thus, highlighting how complex these relationships are even in a region characterized by overall similar climatology.

Keywords: principal component analysis; Granger causality; watershed; water quality; urbanization; land cover

Citation: Zavareh, M.; Maggioni, V.; Sokolov, V. Investigating Water Quality Data Using Principal Component Analysis and Granger Causality. *Water* **2021**, *13*, 343. https://doi.org/10.3390/w13030343

Received: 2 January 2021
Accepted: 22 January 2021
Published: 30 January 2021

1. Introduction

Water quality information is essential to protect lives and manage water resources effectively [1]. This requires state-of-the-art collection procedures [2] and in-depth analysis skills [3,4]. However, data scarcity in Earth science is a well-known problem because of the high cost of monitoring systems and low reliability of the measurements. Across the United States (U.S.), the U.S. Geological Survey's (USGS) National Water Information System (NWIS) is in charge of acquiring, processing, and storing water quality data [5]. Analyzing and interpreting these data is complicated by several factors. First, water quality data are complex in nature, as multiple water quality indicators are commonly combined to holistically characterize the condition of streams, lakes, and groundwater (including pH, dissolved oxygen level, nutrients' concentration, among others). Second, water quality changes both in time—and is often affected by seasonality—and space, influenced by topography, urban effluents, farm waste, fertilizer runoffs, and industrial waste discharge. Third, different water bodies may be affected by different environmental issues. For instance, the concentration of nutrients such as nitrates and phosphates may be more problematic in lakes during summertime, because of the risk eutrophication. On the other hand, conductivity may be an indicator for control in rivers where fishery is a predominant activity, since conductivity is an indirect measure of the saltiness of the water and freshwater may not tolerate large increases in saltiness.

All the challenges highlighted above may translate into inadequate representations of complex environmental systems [6] and poor management decisions [7]. Nevertheless, exploring state-of-the-art analytical tools may make up for the lack and complexity of such data. For instance, water quality datasets can be investigated through multivariate analyses, a set of techniques to analyze data that comprise more than one variable, which are usually correlated among themselves [8].

One of the most common multivariate methods to extract fundamental information from a dataset and reduce its initial size is the Principal Component Analysis [8]. PCA has been used in the past to evaluate water quality data and reduce the number of variables without losing any information in the system. For instance, Ou et al. [9] applied PCA to reduce the number of water quality data collected at 22 monitoring stations in a study area in northeast Florida. Their analysis showed that they could reduce the number of stations to 19, based on 3 years of data. Gangopadhyay et al. [10] used PCA to identify monitoring wells that are important for predicting the dynamic variation in potentiometric head in Bangkok, Thailand. Additionally, Khound and Bhattacharyya [11] conducted a multivariate analysis to determine the source and extend of water pollution in the North Brahmaputra Plain in India, using 50 groundwater and 35 surface water samples to analyze 15 chemical parameters from 2009 to 2011. PCA identified the main source of variation in groundwater chemistry and the geogenic and anthropogenic influences on water quality. In a similar study, Rao et al. [12] coupled PCA with ionic spatial distribution and entropy water quality index techniques to evaluate groundwater quality in the Wanaparthy District in Telangana State, India and identify the wells that need treatment prior to use. Matitatos et al. [13], Kamtchueng et al. [14], and Kumar et al. [15] also adopted PCA to evaluate surface and groundwater quality and identify the possible stressors on water systems. In Europe, Iticescu et al. [16] quantified water quality in the lower Danube region by using multivariate techniques and a water quality index. They studied 18 physiochemical water quality parameters through PCA and response surface methods and were able to reduce the number of variables, assess the correlation between variables, and verify the existence of a seasonal gradient in the dataset. Additionally, Villegas et al. [17] identified the main groundwater hydro chemical patterns, specifically variation in water quality in recharge and discharge in the northwestern part of Antioquia in Colombia, using PCA.

Another powerful multivariate analysis technique to investigate causal information and hidden relationships among variables is Granger causality, one of the most common data-driven approaches to explore cause-and-effect relationships [18,19] evaluated the causality between daily and monthly water temperatures in the Notec river in Poland and proved that the forecasting river water temperature was more accurate if air temperature of the previous day was considered. Val et al. [20] studied the impact of anthropogenic and natural changes in the Ebro river basin in Spain using Granger causality. Furthermore, Sun et al. [21] coupled the feature selection method with Granger causality to overcome the two-dementia feature selection problem (i.e., selecting feature and their window size of effective lagged values). Zelaya et al. [22] combined Granger causality with vector auto-regressive models to assess the relationship between richness and geochemistry in three designated wells in Oak Ridge, TN. In another study, Salvucci et al. [23] investigated the Granger causality of soil moisture on precipitation in 18 stations in Illinois and concluded that soil moisture was not linked to precipitation alone.

This work builds upon these past studies and combines PCA analysis with Granger causality to investigate stream water quality data collected across multiple watersheds in Virginia, Maryland, and the District of Colombia from January 2010 to May 2019. The PCA analysis evaluates and potentially reduces the number of water quality indicators, while the Granger causality reveals cause-and-effect relationships between water quality indicators and select environmental variables (e.g., precipitation, streamflow). Such relationships are interpreted based on the watershed characteristics, including land cover, soil type, and the presence of farm animals. The study area and datasets are presented in the next

section. Section 3 presents and discusses the methodology, while results and conclusion are discussed in Sections 4 and 5, respectively.

2. Study Area and Dataset

This study investigates ten watersheds across the District of Columbia, Maryland, and Virginia (also known as the DMV region), one of the most rapidly growing urban coastal areas in the United States [24]. Its population growth is still on the rise and is predicted by the Northern Virginia Regional Commission to continue to at least 2040 [25]. Its proximity to the coast makes the region vulnerable to hydro-meteorological hazards exacerbated by sea level rise and increased urban development. Population growth and multiple climatic stressors call for comprehensive plans that increase the community's resilience to ensure clean water to the entire population. Moreover, excessive nutrient loads in the Chesapeake Bay area and its tidal tributaries often resulted in eutrophication in the past years [26]. The recovery is slow and the Bay is listed as an impaired water body according to the Clean Water Act [27]. As a result, monitoring water quality in the area has been of interest to many researchers and engineers [28].

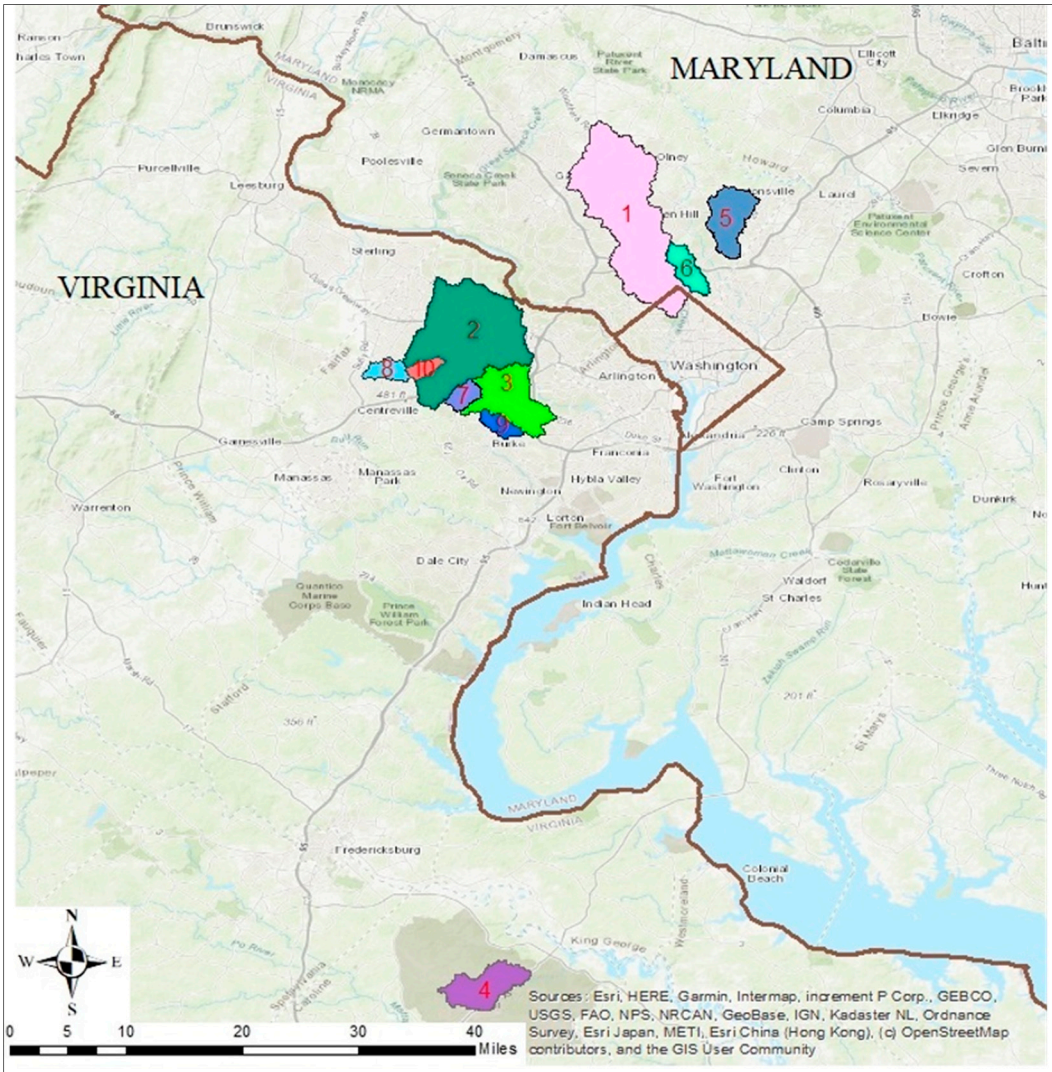
Figure 1 displays the location of the 10 watersheds located across the DMV selected for this study. Moreover, a USGS station is located at the discharge point (or outlet) of each watershed. Watersheds and their associated USGS IDs (Hydrological unit code) are listed based on their size from large to small in the legend of Figure 1. In the discussion of results, the first three watersheds are considered large, watersheds 4, 5, and 6 medium, and the last four are classified as small. Watershed areas vary from 7 to 168 km^2 with an average of 50 km^2 and a standard deviation of 56.8 km^2.

Watershed information, streamflow discharge, and water quality data are collected from USGS [5] during the period January 2010–May 2019. However, 9 years of hydrological readings are not available for every single watershed. All data used in this study along with their sources and description are listed in Table 1. These include watershed characteristics like area, land cover, and soil type. Precipitation data are obtained from the North America Land Data Assimilation Systemb (NLDAS) [28]. The four water quality indicators are chosen based on their availability across the 10 watersheds and include water temperature (WT), dissolved oxygen (DO), turbidity (Tu), and specific conductivity (K). DO concentration is a measure of how much oxygen is dissolved in water, turbidity is a measure of the relative clarity of water, and specific conductivity is the ability of a solution to conduct electricity.

Table 1. Watershed data, sources, and units.

Variable	Source	Description and Unit
Date	USGS	2010 to 2019
Land Use	NLCD [1] 2011	Classified as land, open water, developed, barren, forest, shrubland, herbaceous, planted/cultivated, wetlands
Soil type	USDA (gSSURGO 2016) [2]	Type A: High Infiltration and A/D—High/Very Slow Infiltration Type B Moderate Infiltration and B/D—Medium/Very Slow Infiltration Type C: Slow Infiltration and C/D—Medium/Very Slow Infiltration Type D: Very Slow Infiltration
Discharge	USGS	Average daily discharge from a watershed at exit point in cubic feet per second
Air Temperature (T)	USGS	Average daily air temperature in degree Celsius
Water Temperature (WT)	USGS	Average daily water temperature in degree Celsius
Precipitation	NLDAS [3]-2	Average daily precipitation
Specific Conductivity (K)	USGS	Average daily specific conductivity in microsiemens per centimeter at 25 degrees Celsius
Dissolved Oxygen (DO)	USGS	Average daily Dissolved oxygen concentration in milligram per liter
Turbidity (Tu)	USGS	Average daily turbidity in Nephelometric Unit (NTU)

Note: [1] National Land Cover Database; [2] United State Department of Agriculture; [3] United State Department of Agriculture.



ID	USGS ID	Area (km²)	Description
1	USGS 01648010	169	ROCK CREEK AT JOYCE RD WASHINGTON, DC
2	USGS 01646000	149	Difficult Run-Great Falls, VA
3	USGS 01654000	62	ACCOTINK CREEK NEAR ANNANDALE, VA
4	USGS 0166818623	37	MILL CREEK ABV PEUMANSEND CREEK NR PORT ROYAL, VA
5	USGS 01649190	34	PAINT BRANCH NEAR COLLEGE PARK, MD
6	USGS 01650800	17	SLIGO CREEK NEAR TAKOMA PARK, MD
7	USGS 0165389205	10	ACCOTINK CREEK NEAR RANGER ROAD AT FAIRFAX, VA
8	USGS 01656903	10	FLATLICK BRANCH ABOVE FROG BRANCH AT CHANTILLY, VA
9	USGS 01654500	10	LONG BRANCH NEAR ANNANDALE, VA
10	USGS 01645762	7	Difficult Run-Vienna, VA

Figure 1. Location, area, and extension of the 10 watersheds selected for the study in the DMV region.

Table 2 presents additional information for each of the 10 watersheds, including land cover and soil type. Watershed 6 is the most developed (87% of the total area), while watershed 4 is only developed by ~8%. Watersheds 1, 4, 5, and 6 are mainly characterized by soil type B, which has moderate infiltration. On the other hand, watersheds 2, 3, and all

the small watersheds (7–10) show a prevalence of soil type C with slow infiltration. Information regarding the land use and soil type can be particularly useful when interpreting relationships among water quality indicators and environmental characteristics, as both land use and soil type impact infiltration rates, and, as a consequence, both streamflow and stormwater runoff and the contaminants they carry (thus affecting water quality).

Table 2. Land use and soil type characteristics of each watershed in percent values.

	1	2	3	4	5	6	7	8	9	10
Open Water	0.3	0.4	0	0.1	0	0	0	0.1	0	0
Developed	69.2	53.5	74.2	7.9	61.1	87.8	85.4	86.0	70.6	44
Barren	0.2	0	0	0	0	0	0	0	0	0.2
Forest	20.9	38.9	22.8	77.6	29.1	11.7	14.3	11.6	27.1	51
Shrubland	0.9	0.9	0.2	5.2	0.9	0.3	0.2	0.4	0.1	0.7
Herbaceous	0.1	0.1	0	0.3	0.2	0	0	0.5	0	0.1
Planted-Cultivated	6.5	1.8	0.2	1.9	5.9	0	0	0.4	0	0
Wetlands	1.8	4.3	2.7	6.8	2.7	0.1	0.1	0.9	2.3	3.8
Soil Type A	0.7	2.9	1.2	0	1.0	0	0	0	0.7	4.0
Soil Type B	73.6	29.9	18.1	99.8	76.2	81.2	6.0	4.3	20.5	29.4
Soil Type C	16.0	66.7	80.7	0.2	14.5	11.1	93.6	89.7	78.9	66.5
Soil Type D	9.8	0.5	0.1	0	8.3	7.7	0.3	6.0	0	0

3. Methodology

The flowchart in Figure 2 maps out the methodological process followed in this work. A set of pre-processing steps is required before applying the PCA and Granger causality analysis to the collected data, as described in the next sub-section.

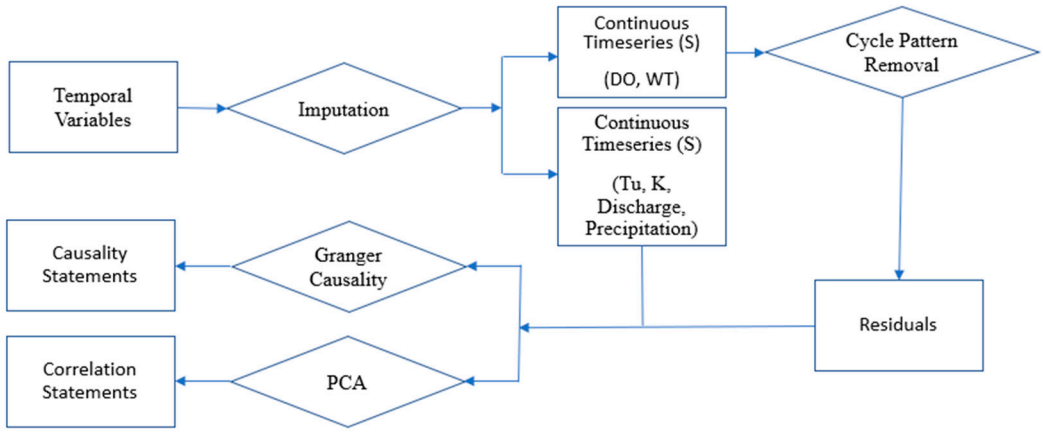


Figure 2. Methodology flowchart. The rectangular shapes represent data and the diamond shapes represent processes.

3.1. Data Pre-Processing

Before applying the PCA and Granger causality analysis to the collected data, there are a few pre-processing steps that need to be considered. First off, missing observations in time series are very common and many methods are available to overcome this issue [29]. Many time series analysis techniques assume there is no gap in data frame, and the analysis of incomplete time series may result in biased results. Thus, in this work, the missing values are imputed with the median of the nearest values to obtain continuous time series.

Secondly, time series of environmental variables often show complex cyclic patterns. Some of the data considered here exhibit daily and seasonal patterns. For instance, as shown in Figure 3 for the Difficult Run watershed near Vienna, VA during a 6-year period (June 2011–August 2017), both water temperature and DO present a strong seasonal pattern.

Moreover, such variables present a diurnal cycle (temperature is higher during the day and lower at night). On the other hand, other variables, like Tu and K, are more closely related to streamflow (which is driven by precipitation events) and do not present any strong temporal cycle in their timeseries (Figure 3).

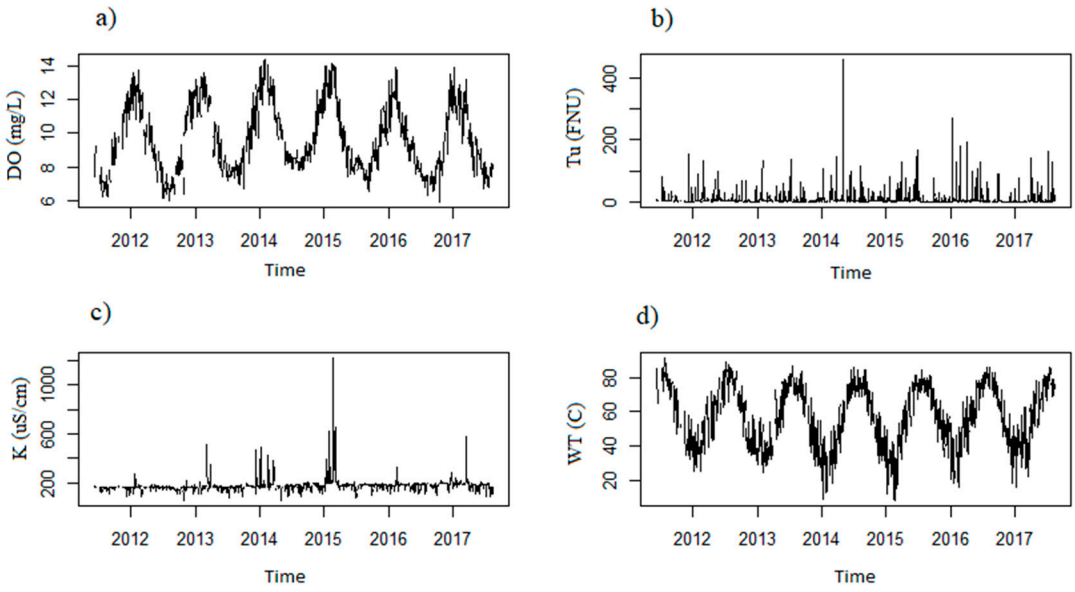


Figure 3. Time series of (**a**) dissolved oxygen (DO), (**b**) turbidity (Tu), (**c**) specific conductivity (K), and (**d**) water temperature (WT) at the outlet of the Difficult Run watershed during the period 2012–2017.

In order to identify relationships among variables that are independent of any diurnal or seasonal patterns, it is recommended to remove any cyclic behavior from the original data. Thus, we fit temperature and DO timeseries with a Fourier transformation, which is subsequently subtracted by the initial time series. A Fourier series $f(x)$ is defined according to Tolstov [30] as:

$$f(x) = \frac{1}{2}a_0 + \sum_{n=1}^{\infty} a_n \cos\left(\frac{n\pi x}{p}\right) + \sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi x}{p}\right)$$

where:

$$a_0 = \frac{1}{p}\int_{-p}^{p} f(x)dx$$

$$a_n = \frac{1}{p}\int_{-p}^{p} f(x)\cos\left(\frac{n\pi x}{p}\right)dx$$

$$b_n = \frac{1}{p}\int_{-p}^{p} f(x)\cos\left(\frac{n\pi x}{p}\right)dx$$

where $p > 0$ is a fixed value, $f(x)$ is a periodic function with period $2p$, defined within $(-p, p)$, and n is the number of cycles. Residuals at each time t (R_t) are calculated as follows:

$$R_t = S_t - F_t$$

where S_t is an observation at time t and F_t is the fitted Fourier series.

As a third pre-processing step, stationarity needs to be tested, as Granger causality adopted in this study assumes stationary time series. A time series is stationary if its statistical properties (e.g., mean, variance, and autocorrelation) do not change over time. One common test to assess stationarity is the Augmented Dickey Fuller (ADF) test, which

is the extended version of the Dickey Fuller (DF) [31] test and allows for higher order autoregressive processes:

$$\Delta y_t = d_0 + \theta y_{t-1} + \sum_{i=2}^{k} \varphi \Delta y_{t-1+i} + \varepsilon_t$$

where φ_i is the coefficient of trend with $\varphi_i = -\sum_{j=1}^{k} b_j$, d is a constant, k is the lag in the autoregressive process, and ε is the magnitude of the random error, i.e., the white noise [32]. θ is a constant and if it is equal to zero, then the time series is not stationary. If the ADF test is not cleared, methods to remove non-stationary trends need to be considered.

3.2. PCA

PCA is a data reduction technique that transforms a dataset into a new set of variables, the principal components (PCOMs), which are a linear combination of the original variables. The main goal of PCA is to maintain the original variation of the data [33], while creating an uncorrelated dataset. It also reveals patterns that might not be apparent using common analysis and graphic techniques. As a first step, the covariance matrix is calculated. If X is the original dataset in a matrix format, with m rows (which account for different measurements of a specific attribute) and n columns (which represent the attributes), then the covariance matrix C_x is:

$$C_x = \frac{1}{n} X X^T$$

where X^T is transpose matrix of X.

Next, eigenvalues and eigenvectors are computed. The eigenvector $\vec{v}$ is defined as:

$$C_x \vec{v} = \lambda \vec{v}$$

where λ is a scalar value, i.e., the eigenvalue. The following equations show steps to solve for eigenvalue and eigenvector:

$$C_x \vec{v} - \lambda \vec{v} = 0$$
$$\vec{v}(C_x - \lambda I) = 0$$
$$Det(C_x - \lambda I) = 0$$

where I is the identity matrix of the same dimension as C_x. As a result, each eigenvector is produced by each λ times $\vec{v}$ which is called principal component. The number of principal components is equal to the dimension of the dataset, however, PCA loads the maximum possible information in the first component, the maximum remaining information in the second component, and so on. The number of PCs is usually based on the number of eigenvalues greater than 1 [34]. The ratio between the eigenvalue of a component and the sum of the eigenvalues shows the percent of variance of in the original dataset represented by that component.

3.3. Granger Causality

The notion of Granger causality was introduced by Granger [35] and soon found application in many fields (e.g., economics) because of its simplicity and robustness [21]. For this study, we adopt the first-order Granger causality test, which investigates the linear causal interaction between time series of data. The causal relation exists if the following two conditions are fulfilled: (i) the cause precedes the effect; and (ii) the cause contains information about the effect that is not available in other variables. As mentioned above, one main assumption to test Granger causality is the stationarity of the time series. The

bi-variate Granger causality between two stationary time series (X and Y) is formulated as follows:

$$Y_t = \sum_{j=1}^{m} a_t Y_{t-j} + \sum_{j=1}^{m} b_j X_{t-j} + \varepsilon_t$$

where a and b are coefficient ($b \neq 0$) and ε is white noise. In this case, variable X Granger causes variable Y.

It is important to mention that Granger causality measures precedence and information content, rather than the effect or the result. Granger causality tries to answer the question of how much of the current variable can be explained by the past values of a different values and whether adding lagged values can improve such explanation [36].

4. Results

4.1. Pre-Processing

After imputing all the hydrometeorological variables (i.e., the four water quality indicators, air temperature, precipitation, and discharge), the diurnal and seasonal patterns in water temperature and DO were removed by fitting a Fourier series and calculating the residuals in each watershed. Figure 4 shows time series DO and water temperature in watershed 10 as an example. The DO and water temperature readings started from June of 2011 to August of 2017, which results in six cold weather seasons and almost eight hot weather seasons. Such effects together with any diurnal cycle were removed in the residuals, which are used in any further analysis. Besides DO and water temperature, no other variable presented a strong cyclic pattern and therefore did not undergo the de-cycling process.

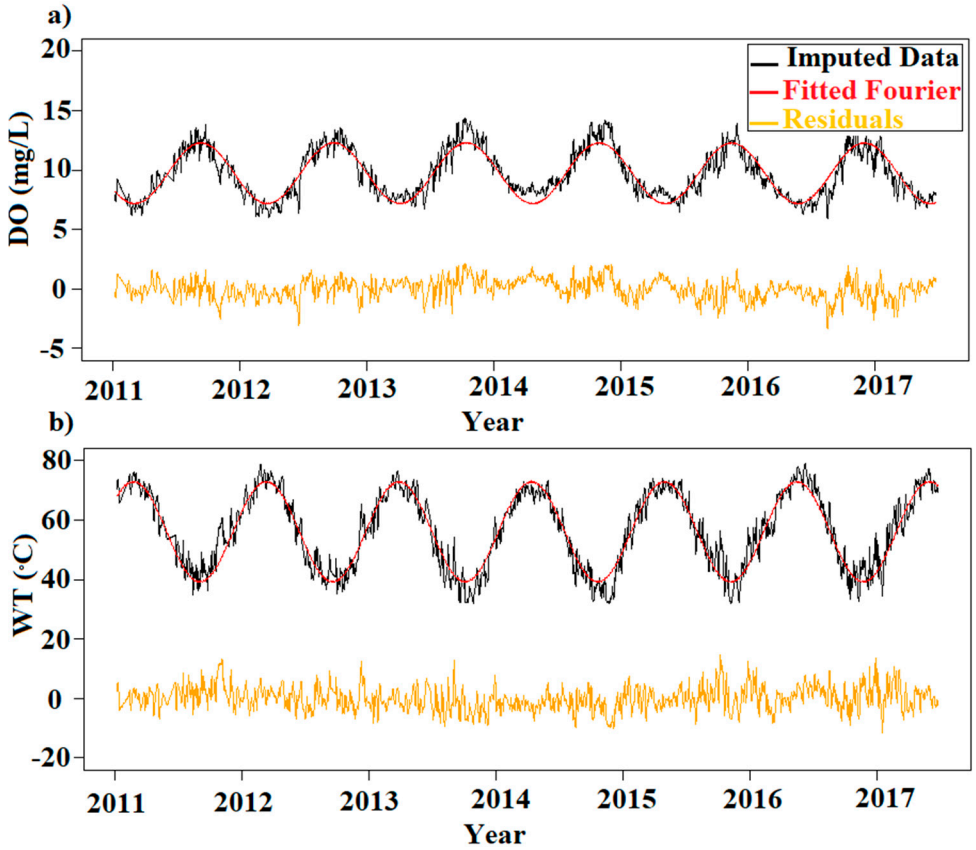


Figure 4. Time series, fitted Fourier series, and residuals (i.e., time series from which cyclic patterns are removed) for (**a**) DO and (**b**) water temperature.

Next, the ADF test is applied to imputed time series of precipitation, discharge, turbidity, specific conductivity, DO, air temperature, and water temperature to check if such datasets are stationary. Results show that residuals of DO and water temperature are stationary at a 95% confidence level. Precipitation, discharge, turbidity, specific conductivity, and air temperature also pass the ADF test, at an even higher (99%) confidence level.

Before investigating the PCA and Granger causality, some preliminary analyses are performed. Specifically, residuals of water temperature are plotted against residuals of DO for three different lags (1 day, 2 days, and 3 days). For the sake of brevity, one large (1), one medium (4), and one small watershed (7) are shown in Figure 5. In other words, for lag 1, we assess the relationship between the water temperature of yesterday (or *t-1*) and the DO of today (at time *t*). Similarly, for lags 2 and 3, the temperature recorded at *t-2* and *t-3* is linked to the DO observed at time *t*. As expected, DO and water temperature are negatively correlated: when water temperature rises, the DO concentration drops. What is interesting though is that such a relationship is even stronger in large and medium watersheds relative to smaller watersheds (Figure 5). This may be due to the lower number of readings in the small watersheds where data were mainly collected during the cold months. As shown in Figure 5, the correlation decreases when the lag increases. That is, the relationship between today's DO and yesterday's temperature is stronger than the one with the temperature measured 2 or more days ago.

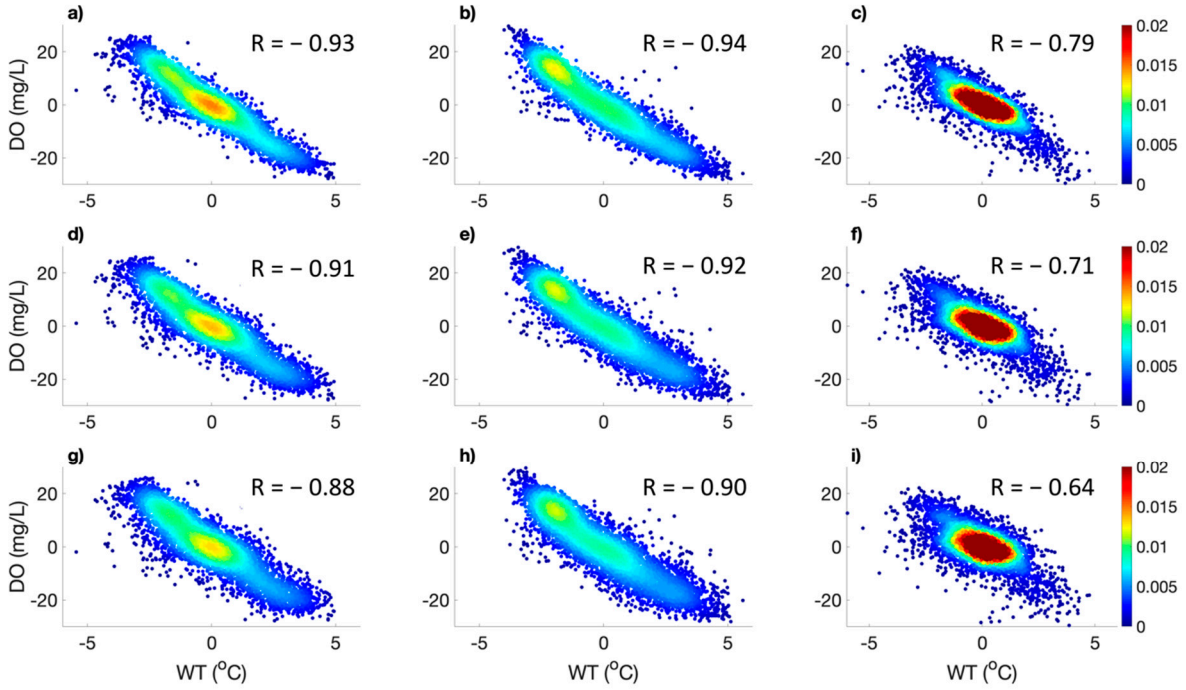


Figure 5. Density scatterplots of de-seasonalized DO against water temperature for a large watershed (**a**,**d**,**g**); a medium watershed (**b**,**e**,**h**); and a small watershed (**c**,**f**,**i**) and for three different lags: 1 day (**a**–**c**), 2 days (**d**–**f**), and 3 days (**g**–**i**). Correlation coefficients for each case are also shown.

4.2. PCA

The PCA analysis is performed on hydrometeorological variables (precipitation, discharge, T, WT, DO, Tu, K) in every watershed. In most watersheds, the first two principal components explain most of the variance, except for watershed #4 that presents three principal components (Figure 6). Interestingly, this watershed is the least developed among all the ones considered in the study with almost 78% of its entire area covered by forest. Hydrological processes in a more natural environment are highly non-linear, because the relationship between precipitation, discharge, and water quality is complicated by the pres-

ence of dense vegetation that intercepts rainfall, slows down infiltration, and withholds a portion of runoff and the pollutants it carries. Furthermore, most watersheds show similar behavior, with between 30% and 50% of the variance explained by the first component and between 20% and 30% explained by the second component. As a result, PCOM 1 and 2 cumulatively explain more than 60% of the variation in every watershed.

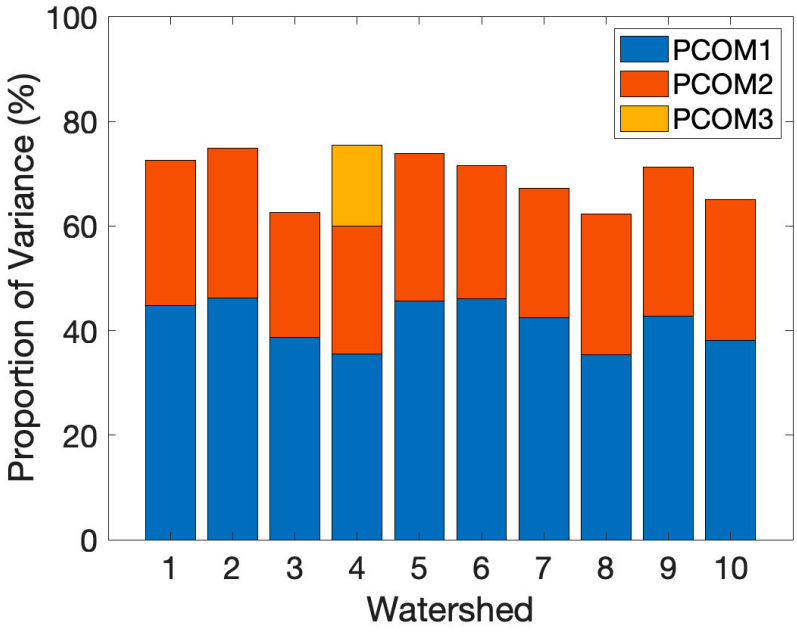


Figure 6. Proportion of variance explained by each principal component for the 10 watersheds.

Table 3 shows how individual hydrometeorological variables contribute to each principal component. If the loading associated with one variable is positive and the loading associated with another variable is negative, then those two variables are negatively correlated. Vice versa, if the loadings associated with two variables have the same sign, they are positively correlated. For instance, the loadings associated with WT and DO always show opposite signs, confirming the well-known negative correlation between the two variables. Stronger loadings correspond to variables that contribute more to the variation in the original dataset. DO and discharge usually are associated with strong loadings (greater than 0.5) in PCOM 1 (except for watersheds 3 and 7), showing that they carry fundamental information that cannot be disregarded in most of the watersheds. In PCOM 2, Tu, K, and precipitation appear to be the most prominent variables (with larger loadings), thus, identifying a set of variables that are also significant sources of variability in the dataset.

The PCA results indicate that in small watersheds the most relevant variables (i.e., the ones that explain most of the variance) are Tu, discharge, DO, T, and WT. Small watersheds are also the most urbanized watersheds with a low soil infiltration (i.e., soil type C). As a result, a higher volume of water travels in shorter amount of time in these watersheds. On the other hand, in medium watersheds, most of the variation is explained by Tu, precipitation, DO, T, and WT. Medium watersheds are the least urbanized watersheds in comparison to the others with higher infiltration rates (i.e., soil type B) and the largest forest land cover. This corresponds to lower volumes of runoff and higher chance of infiltration in comparison to the small watersheds. In summary, discharge plays a more important role in small urbanized watersheds, whereas precipitation plays that role in medium watersheds. Similarly, in large watersheds, most of the variability in data is explained by Tu, discharge, T, and WT. In terms of land cover, large watersheds are similar to the small watersheds, i.e., highly urbanized with lower percentage of forest. As a result, discharge plays a more important role than precipitation, which once again is in line with the fact that the more urbanized areas tend to have higher volume of runoff in a shorter period of time.

Table 3. Principal component loadings.

Watershed	Principal Component	Tu	Precipitation	Discharge	K	DO	T	WT
1	PCOM 1	0.61	0.46	0.64	−0.31	−0.55	0.55	0.55
	PCOM 2	0.17	−0.5	0.14	−0.8	0.12	−0.14	−0.13
2	PCOM 1	0.14	0.12	0.64	−0.32	−0.54	0.53	0.53
	PCOM 2	0.62	0.41	0.25	−0.38	−0.15	0.18	−0.11
3	PCOM 1	0.33	0.1	0.29	−0.25	−0.46	0.52	0.5
	PCOM 2	0.58	0.34	0.6	0.12	0.21	−0.24	−0.29
4	PCOM 1	0.14	0.15	0.65	−0.44	−0.54	0.56	0.59
	PCOM 2	0.43	0.4	0.11	0.69	0.19	−0.22	−0.24
	PCOM 3	0.61	−0.88	0.14	−0.24	−0.21	0.12	0.76
5	PCOM 1	0.65	0.36	0.67	−0.32	−0.55	0.54	0.55
	PCOM 2	0.27	−0.82	0.2	−0.44	−0.16	0.17	0.19
6	PCOM 1	0.65	0.38	0.65	−0.35	−0.54	0.54	0.54
	PCOM 2	0.29	−0.92	0.26	0.93	−0.17	0.21	0.22
7	PCOM 1	0.18	0.29	0.15	−0.21	−0.54	0.55	0.55
	PCOM 2	0.65	−0.73	0.65	−0.62	0.12	−0.14	−0.16
8	PCOM 1	−0.14	−0.39	−0.65	0.23	0.51	−0.57	−0.58
	PCOM 2	−0.62	−0.44	0.18	0.8	−0.13	0.52	0.13
9	PCOM 1	0.14	0.32	0.66	−0.22	−0.55	0.55	0.56
	PCOM 2	0.65	−0.53	0.25	−0.81	0.12	−0.53	−0.12
10	PCOM 1	0.65	0.31	0.66	−0.22	−0.55	0.55	0.58
	PCOM 2	0.25	−0.68	0.27	−0.17	−0.68	−0.64	−0.8

4.3. Granger Causality

Granger causality assesses whether one variable at time *t*—*lag* causes another variable at time *t*. In our analysis, we considered three different lags (i.e., 1 day, 2 days, and 3 days). This choice was dictated by the fact that most of the watersheds have a fast hydrological response, due to the fact that they are limited in size, highly developed, and characterized by low to very low soil infiltration rates. Thus, going beyond a 3-day lag would not be recommended. The Granger causality test is performed on each of the four water quality indicators (WT, DO, Tu, K), which are considered the effects, and all the hydrometeorological variables (WT, DO, Tu, K, T, precipitation, and discharge) considered as possible causes. The null hypothesis is defined as follows: there is no Granger causality between the cause and effect. Thus, lower *p*-values correspond to stronger causality and vice versa.

Figure 7 shows *p*-values for the Granger causality test, when WT is considered as the effect. The Granger causality relationship is strong in any lag for all the variables. Water temperature is known to be an important physical property and any change in the other variables can impact it. When discharge is the cause, the variability around median *p*-value is larger in lag 1 and gets smaller when moving to lag 3, showing that such relationship is not as strong in all the watersheds Nevertheless, this uncertainty is not due to uncertainty in the precipitation relationship, which shows very low *p*-values. Therefore, this may be due to the fact that the WT is highly dependent on the air temperature and precipitation rather than discharge.

When DO is considered as the effect, as shown in Figure 8, the Granger causality relationship is very similar to the previous case, where WT is Granger caused by the rest of the variables. This means that the streamflow and environmental conditions of the previous days impact the amount of dissolved oxygen in the water today. However, higher variability is observed around K, which can be explained by the fact that the ability of water to conduct an electrical current does not impact the amount of dissolved oxygen directly.

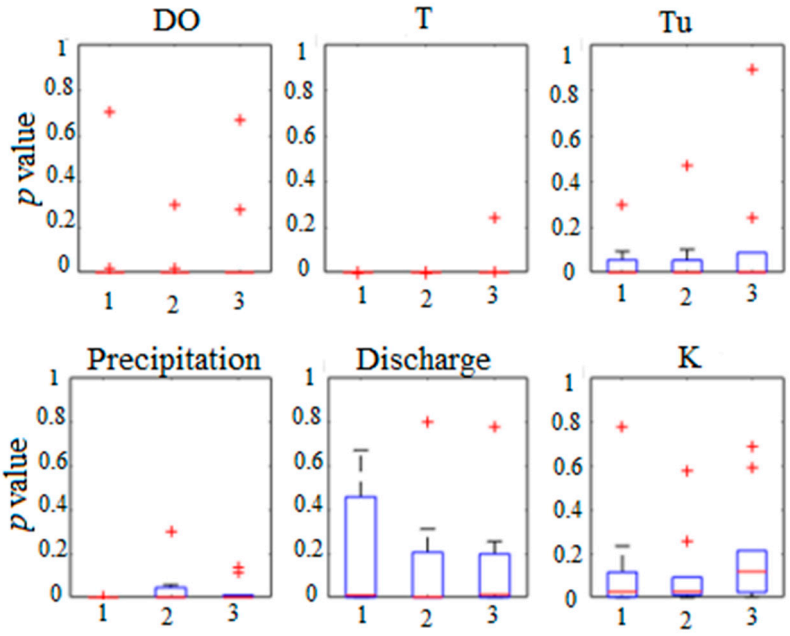


Figure 7. Box plots of *p*-values resulting from the Granger causality test when WT is Granger caused by DO, T, Tu, precipitation, discharge, and K based on a lag time of 1, 2, and 3 days and performed for each watershed. The 25th and 75th percentiles are shown by the blue boxes, while the central red line shows the median. Whiskers extended to minimum and maximum do not consider the outliers, shown with red crosses.

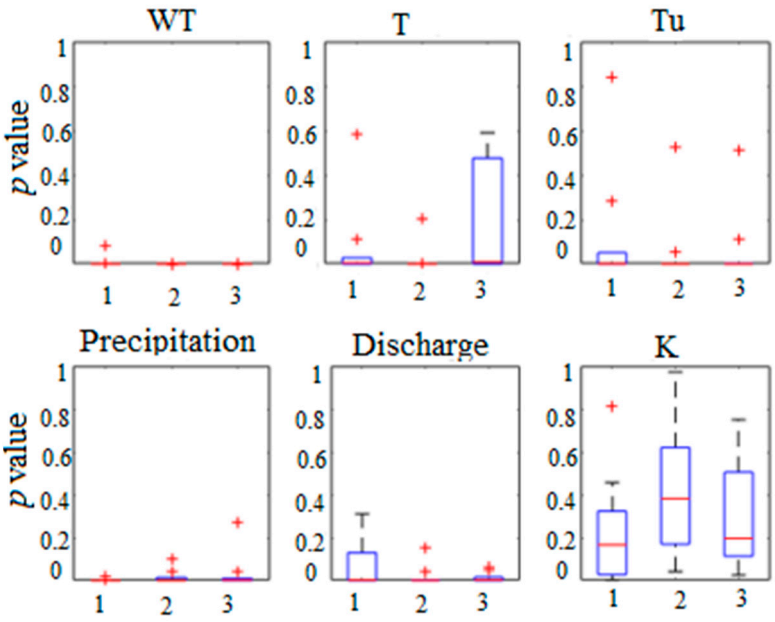


Figure 8. Same as in Figure 7, but for DO as an effect in the Granger causality relationship.

When Tu is the effect (Figure 9), the Granger causality relationships are strong (i.e., low *p*-values) when WT and DO are the causes in all lags. This means that the values of WT and DO measured 1, 2, and 3 days ago influence the number of suspended particles in the water today. Granger causality relationship between T, precipitation, discharge, and K is strong at lag 1, but weaker for longer lags, which is expected since the retention time is short in the studied watersheds. In addition, the variability around the median *p*-value decreases in WT and DO when the lag increases, but the opposite happens for precipitation

and K. This means that past values of WT and DO (values from 2 and 3 days ago) impact the Tu of today, which is nonetheless not impacted by the precipitation and K of 2 or 3 days ago. This may be due to the fact that the variation in WT and DO values is not as significant as the variation in precipitation and K.

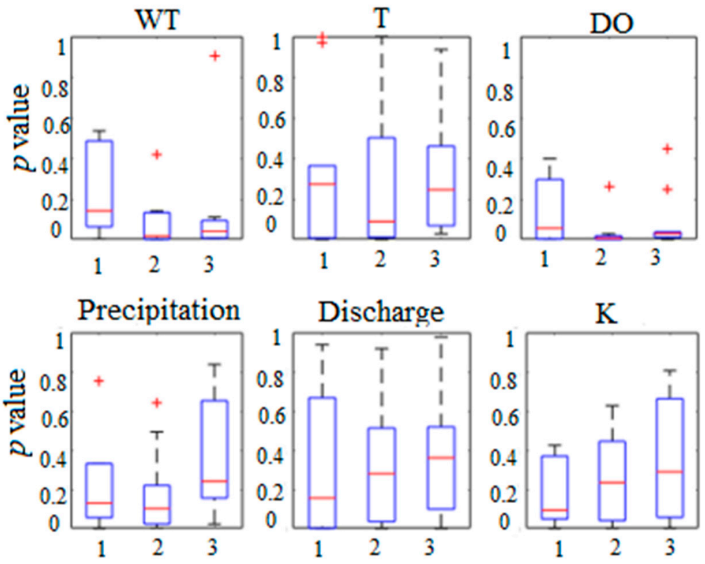


Figure 9. Same as in Figure 7, but for Tu as an effect in the Granger causality relationship.

When K is Granger caused by the rest of the variables, as shown in Figure 10, the relationship is very strong when WT and DO are the cause. This can be easily explained, since WT influences the ability of water to conduct electrical current and DO is highly related to WT, as discussed above. The Granger relationships are more uncertain when the other variables are considered as the causes, as shown by the wider box plots. This demonstrates that the ability of water to conduct electrical current is not as closely linked to variables like precipitation, air temperature (that is more variable than water temperature), discharge, and K at least at lags between 1 and 3 days. Nevertheless, investigating lags shorter than 1 day may identify stronger causality among these variables in the watersheds analyzed here.

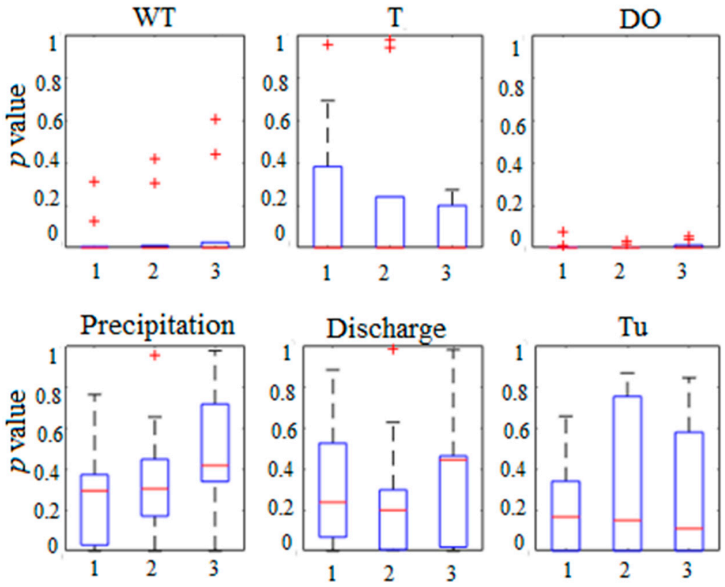


Figure 10. Same as in Figure 7, but for K as an effect in the Granger causality relationship.

After analyzing all four scenarios, it is concluded that the strongest relationships (i.e., smallest *p*-values) are observed for lag 1. Based on these results, lag 1 relationships are further investigated as a function of the watershed size and urbanized area. For the sake of brevity, we only show the Granger causality relationship when DO is caused by the other six variables based on different watershed sizes (Figure 11). Relationships are strong when WT and precipitation are the cause across all watersheds, regardless of their size, which may be in part due to the similar environmental and climatic characteristics of the watersheds. The Granger causality is also strong when T, discharge, and Tu are the cause, but in this case the *p*-value depends on the watershed size. As a result of lower retention time in smaller watersheds, the causality relationship between Tu and DO weakens as the watershed size decreases. The *p*-value is higher when DO is Granger caused by K in comparison to when other variables are the cause.

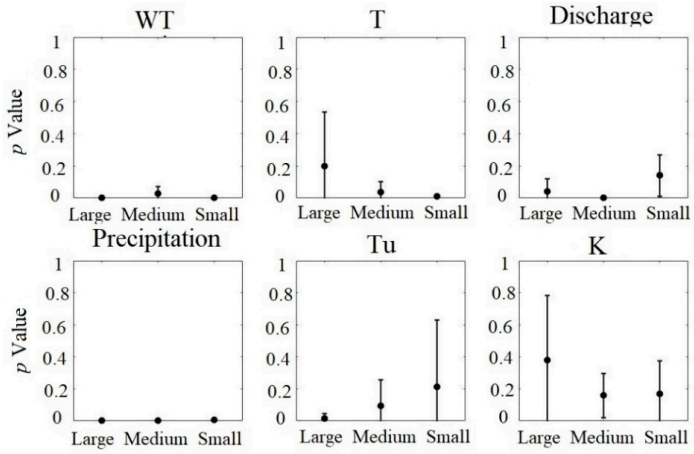


Figure 11. Box plots of *p*-values resulting from the Granger causality test based on watershed size (small, medium, and large) when DO is Granger caused by WT, T, discharge, precipitation, Tu, and K. Dots show the median and the vertical lines indicate +/− standard deviation.

Figure 12 displays Granger causality test results when DO is caused by the other six variables based on the level of urbanization in the watersheds. Watersheds 1, 2, 4, 5, and 10 show a level of urbanization lower than 70%, whereas the others are considered highly urbanized. Watersheds characterized by less urban area show stronger Granger causality relationship when WT, discharge, precipitation, and Tu are the cause of a change in DO. When soil infiltration is slow (which is commonly the case in more developed regions), the retention time is possibly shorter than 1 day, which is why such relationships are weaker.

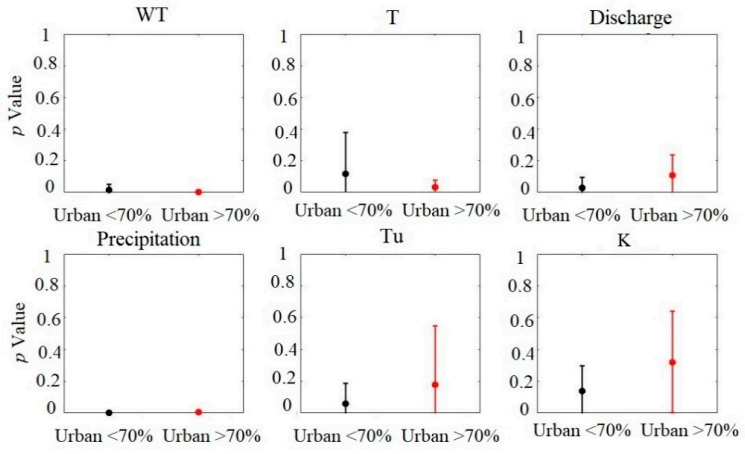


Figure 12. Same as in Figure 11, but for different urbanization percentages (<70% and >70%).

5. Conclusions

This study proposes a set of analytical tools to better understand the relationship between a suite of water quality indicators (WT, DO, Tu, and K) and select hydroclimatic variables (precipitation, discharge, and air temperature) in a series of watersheds in the DMV region in the Eastern United States. Stream water quality in this region is of particular interest because of two main reasons: (i) it is one of the most rapidly growing urban coastal areas in the United States and (ii) its tributaries feed into the Chesapeake Bay, listed as an impaired water body according to the Clean Water Act. Thus, investigating how water quality relates to the local hydrometeorology and urbanization is crucial to developing effective and sustainable methods for water quality protection.

An extensive pre-processing is applied to the collected data, as a first step to fill in any missing observations in the time series and to remove any cyclic pattern and ensure stationarity, before applying the PCA and testing for Granger causality. Results from the PCA show that most of the variation in each watershed can be explained by only considering the first two principal components, except for watershed 4, which is the least developed watershed (almost entirely covered with forest) among all watersheds. This may be due to the highly non-linear hydrological processes ongoing in a more natural environment, where the relationship between precipitation, discharge, and water quality is complicated by the presence of dense vegetation that intercepts rainfall, slows down infiltration, and withholds a portion of stormwater runoff and the pollutants it carries. The selection of important indicators is based on the absolute value of loading. Higher loadings point to a strong relationship between an indicator and specific component. The most important variables identified by the PCA in all watersheds are turbidity, water temperature, and air temperature. Furthermore, while discharge is an important variable in highly urbanized watersheds, precipitation is more fundamental in less urbanized watersheds with higher soil infiltration rates and therefore lower volumes of runoff.

Results from the Granger causality analysis show how different lag times (1, 2, and 3 days) affect the causality relation between hydroclimatic variables and water quality indicators. In general, lag 1 shows more and stronger Granger causality relationships in comparison to lags 2 and 3. This is due to the limited size of the basins (characterized by hydrological responses faster than 2 days), but also to the fact that most watersheds are highly developed and therefore characterized by low soil infiltration rates (which corresponds to relatively short retention times). The strongest Granger causalities are observed when water temperature and dissolved oxygen concentration (which are highly correlated) are considered as the effect of the hydrometeorological variables and other water quality indicators, which corroborates the importance of these two water properties, since any change in the other variables can impact them. When dissolved oxygen concentration is caused by water temperature and precipitation, the watershed size does not play a role in the Granger causality relationships. In contrast, the watershed size changes the strength of such relationships when air temperature, discharge, and turbidity are the cause when dissolved oxygen is the effect. Moreover, urbanization triggers weaker Granger causality when discharge, conductivity, and turbidity are the cause and dissolved oxygen is the effect.

This work explored how PCA and Granger causality analysis can inform relationships between water quality, hydroclimatic variables, and watershed characteristics. A main conclusion is that even within watersheds characterized by similar climate, land use distribution, and size, such relationships vary largely. Thus, if a predictive model were to be built, such information should be carefully considered and predictions like watershed size and urban area should be included. For instance, machine learning algorithms for estimating and predicting water quality variables, based on the set of hydrometeorological variables and watershed information identified here, could be developed and adopted for the assessment of such relationships during extreme weather events, when collecting in-situ data is the most difficult, but also crucial.

Future studies should extend the proposed framework to watersheds characterized by different climatology and size in other regions of the world, to verify the impact of different climates on the identified relationships between hydrometeorology and water quality. Future work could also expand our analyses to more water quality indicators (e.g., pH, nitrate and phosphorous concentrations). Finer temporal resolutions should also be considered to investigate hydrological responses that are shorter than 1 day. Furthermore, a wider set of water quality indicators (including for instance nitrate concentration) should be investigated to generalize the results of this work and make the proposed analyses more useful for areas affected by different types of pollution.

Author Contributions: All three authors conceptualized the study and developed the methodology. M.Z. is responsible for the data analysis, data curation, and the writing of the original draft. V.M. and V.S. were involved in reviewing and editing the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not Applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: "USGS Water-Quality Data for the Nation." https://waterdata.usgs.gov/nwis/qw (accessed 6 January 2020).

Acknowledgments: Water quality data are provided by the U.S. Geological Survey. The authors would like to thank Ishrat Jahan Dollan for providing Figure 1.

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

References

1. Surface Water Use, the USGS Water Science School. Available online: https://water.usgs.gov/edu/wusw.html (accessed on 9 May 2018).
2. Wagner, R.J.; Boulger, R.W., Jr.; Oblinger, C.J.; Smith, B.A. *Guidelines and Standard Procedures for Continuous Water-Quality Monitors: Station Operation, Record Computation, and Data Reporting*; USGS Numbered Series 1-D3; 2006. Available online: http://pubs.er.usgs.gov/publication/tm1D3 (accessed on 1 June 2018).
3. Hounslow, A. *Water Quality Data: Analysis and Interpretation*; Lewis Publishers: Boca Raton, FL, USA, 1995.
4. Zavareh, M.; Maggioni, V. Application of Rough Set Theory to Water Quality Analysis: A Case Study. *Data* **2018**, *3*, 50. [CrossRef]
5. USGS Water-Quality Data for the Nation. Available online: https://waterdata.usgs.gov/nwis/qw (accessed on 6 January 2020).
6. Gorgoglione, A.; Castro, A.; Chreties, C.; Etcheverry, L. Overcoming Data Scarcity in Earth Science. *Data* **2020**, *5*, 5. [CrossRef]
7. Longqin, X.; Shuangyin, L. Study of short-term water quality prediction model based on wavelet neural network. *Math. Comput. Model.* **2013**, *58*, 807–813.
8. Grimnes, S.; Martinsen, Ø.G. Chapter 9—Data and Models. In *Bioimpedance and Bioelectricity Basics*, 3rd ed.; Grimnes, S., Martinsen, Ø.G., Eds.; Academic Press: Oxford, UK, 2015; pp. 329–404.
9. Ou, M.; Liang, S.; Zhang, R.; Xiong, Q. Evaluation of water quality for the Beilun Gulf and Zhenzhu Bay by principal component analysis. In Proceedings of the 2017 International Conference on Advanced Mechatronic Systems (ICAMechS), Xiamen, China, 6–9 December 2017; pp. 324–328. [CrossRef]
10. Subhrendu, G.; Gupta, A.D.; Nachabe, M.H. Evaluation of Ground Water Monitoring Network by Principal Component Analysis. *Groundwater* **2001**, *39*, 181–191. [CrossRef]
11. Khound, N.J.; Bhattacharyya, K.G. Assessment of Water Quality in and around Jia-Bharali River Basin, North Brahmaputra Plain, India, Using Multivariate Statistical Technique. *Appl. Water Sci.* **2018**, *8*, 221. [CrossRef]
12. Rao, N.S.; Sunitha, B.; Adimalla, N.; Chaudhary, M. Quality Criteria for Groundwater Use from a Rural Part of Wanaparthy District, Telangana State, India, through Ionic Spatial Distribution (ISD), Entropy Water Quality Index (EWQI) and Principal Component Analysis (PCA). *Environ. Geochem. Health* **2019**. [CrossRef]
13. Matiatos, I. Apostolos Alexopoulos, and Athanasios Godelitsas. Multivariate Statistical Analysis of the Hydrogeochemical and Isotopic Composition of the Groundwater Resources in Northeastern Peloponnesus (Greece). *Sci. Total Environ.* **2014**, *476*, 577–590. [CrossRef]
14. Kamtchueng, B.T.; Fantong, W.Y.; Wirmvem, M.J.; Tiodjio, R.E.; Takounjou, A.F.; Ngoupayou, J.R.; Kusakabe, M.; Zhang, J.; Ohba, T.; Tanyileke, G.; et al. Hydrogeochemistry and Quality of Surface Water and Groundwater in the Vicinity of Lake Monoun, West Cameroon: Approach from Multivariate Statistical Analysis and Stable Isotopic Characterization. *Environ. Monit. Assess.* **2016**, *188*, 524. [CrossRef]

15. Krishna, K.S.; Babu, S.H.; Rao, P.E.; Selvakumar, S.; Thivya, C.; Muralidharan, S.; Jeyabal, G. Evaluation of Water Quality and Hydrogeochemistry of Surface and Groundwater, Tiruvallur District, Tamil Nadu, India. *Appl. Water Sci.* **2017**, *7*, 2533–2544. [CrossRef]
16. ticescu, C.; Georgescu, L.P.; Murariu, G.; Topa, C.; Timofti, M.; Pintilie, V.; Arseni, M.; Timofti, M. Lower Danube Water Quality Quantified through WQI and Multivariate Analysis. *Water* **2019**, *11*, 1305. [CrossRef]
17. Villegas, P.; Paredes, V.; Betancur, T.; Ribeiro, L. Assessing the hydrochemistry of the Urabá Aquifer, Colombia by principal component analysis. *J. Geochem. Explor.* **2013**, *134*, 120–129. [CrossRef]
18. Granger, C.W.J. Testing for causality: A personal viewpoint. *J. Econ. Dyn. Control* **1980**, *2*, 329–352. [CrossRef]
19. Graf, R. Analysis of granger causality between daily and monthly temperatures of water and air, as illustrated with the example of noteć river. *Acta Sci. Pol. Form. Circumiectus* **2018**, *17*, 101–117. [CrossRef]
20. Val, J.; Chinarro, D.; Pino, M.R.; Navarro, E. Global change impacts on river ecosystems: A high-resolution watershed study of Ebro river metabolism. *Sci. Total Environ.* **2016**, *569–570*, 774–783. [CrossRef] [PubMed]
21. Sun, Y.; Li, J.; Liu, J.; Chow, C.; Sun, B.; Wang, R. Using causal discovery for feature selection in multivariate numerical time series. *Mach. Learn.* **2015**, *101*, 377–395. [CrossRef]
22. Zelaya, A.J.; Parker, A.E.; Bailey, K.L.; Zhang, P.; Van Nostrand, J.; Ning, D.; Elias, D.A.; Zhou, J.; Hazen, T.C.; Arkin, A.P.; et al. High Spatiotemporal Variability of Bacterial Diversity over Short Time Scales with Unique Hydrochemical Associations within a Shallow Aquifer. *Water Res.* **2019**, *164*, 114917. [CrossRef] [PubMed]
23. Salvucci, G.D.; Saleem, J.A.; Kaufmann, R. Investigating Soil Moisture Feedbacks on Precipitation with Tests of Granger Causality. Adv. *Water Resour.* **2002**, *25*, 1305–1312. [CrossRef]
24. Kneebone, E. The Changing Geography of US Poverty. *Brookings* **2017**. Available online: https://www.brookings.edu/testimonies/the-changing-geography-of-us-poverty/ (accessed on 29 January 2020).
25. Resiliency Planning | Northern Virginia Regional Commission—Website. Available online: https://www.novaregion.org/1354/Resiliency-Planning (accessed on 28 January 2020).
26. Zhang, Q.; Murphy, R.R.; Tian, R.; Forsyth, M.K.; Trentacoste, E.M.; Keisman, J.; Tango, P.J. Chesapeake Bay's Water Quality Condition Has Been Recovering: Insights from a Multimetric Indicator Assessment of Thirty Years of Tidal Monitoring Data. *Sci. Total Environ.* **2018**, *637*, 1617–1625. [CrossRef]
27. Zhang, Q.; Hirsch, R.M.; Ball, W.P. Long-Term Changes in Sediment and Nutrient Delivery from Conowingo Dam to Chesapeake Bay: Effects of Reservoir Sedimentation. *Environ. Sci. Technol.* **2016**, *50*, 1877–1886. [CrossRef]
28. Earthdata. Available online: https://earthdata.nasa.gov// (accessed on 26 January 2021).
29. Lepot, M.; Aubin, J.-B.; Clemens, F.H.L.R. Interpolation in Time Series: An Introductive Overview of Existing Methods, Their Performance Criteria and Uncertainty Assessment. *Water* **2017**, *9*, 796. [CrossRef]
30. Tolstov, G.P. *Fourier Series*; Courier Corporation: Mineola, NY, USA, 2012.
31. Cromwell, J.B. (Ed.) *Multivariate Tests for Time Series Models*; Sage Publications: Thousand Oaks, CA, USA, 1994.
32. Nugroho, A.; Simanjuntak, B.H. ARMA (Autoregressive Moving Average) Model for Prediction of Rainfall in Regency of Semarang—Central Java—Republic of Indonesia. *Int. J. Comput. Sci. Issues (IJCSI)* **2014**, *11*, 27–32.
33. Everitt, B.; Hothorn, T. Principal Components Analysis. In *An Introduction to Applied Multivariate Analysis with R*; Everitt, B., Hothorn, T., Eds.; Springer: New York, NY, USA, 2011; pp. 61–103.
34. Jolliffe, I. Principal Component Analysis. In *Wiley StatsRef: Statistics Reference Online*; American Cancer Society: Hoboken, NJ, USA, 2014.
35. Granger, C.W.J. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica* **1969**, *37*, 424–438. [CrossRef]
36. Kirchgässner, G.; Wolters, J.; Hassler, U. *Introduction to Modern Time Series Analysis*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 2013.