sustainability

MDPI

Article

Evaluation of Watershed Water Quality Management According to Flow Conditions through Factor Analysis and Naïve Bayes Classifier

Woo Suk Jung [1] and Young Do Kim [2,*]

[1] Nakdong River Support Team, Presidential Water Commission, Changwon-si 51439, Republic of Korea
[2] Department of Civil & Environmental Engineering, Myongji University, Yongin 17058, Republic of Korea
* Correspondence: ydkim@mju.ac.kr; Tel.: +82-31-330-6407

Abstract: Previous studies on water quality assessment for watershed management have predominantly focused on specific seasonal or annual average values, rather than considering water quality variations based on flow fluctuations. It is crucial to identify the water quality characteristics within a watershed by incorporating flow conditions to establish a customized watershed management approach over different time periods. In this study, a vulnerability analysis was conducted to attain the target water quality (TWQ) in 22 watersheds within the Nakdong River system in South Korea. Additionally, factor analysis (FA) was employed to analyze the characteristics of water quality fluctuations in relation to flow conditions. The FA results categorized the pollution source characteristics of the 22 watersheds into various types, indicating the need for specific pollution source management strategies. These findings enabled an initial decision-making process regarding which water pollution sources to prioritize based on flow conditions. Moreover, detailed analyses of pollution sources were performed for watersheds, where achieving TWQ was challenging. Subsequently, a data-based prediction model was developed using the naïve Bayes classification model to determine the likelihood of achieving TWQ. As a result, this study proposes a technique for water quality management in watersheds by introducing a water quality excess probability model, which employs data-based analysis instead of traditional numerical modeling for watershed water quality assessment and proactive prediction. The study discusses the potential of various data-based tools to reduce development and analysis time, providing a powerful alternative to physical-based models that require extensive input data and are time-consuming. To advance future studies, the establishment of comprehensive water environment big data, improvement of real-time monitoring systems within watersheds, and advancements in spatial and temporal observation technologies are emphasized as essential for the development of an advanced watershed management system.

Keywords: TMDL; target water quality (TWQ); watershed management; water quality assessment; factor analysis; naïve Bayes classifier

Citation: Jung, W.S.; Kim, Y.D. Evaluation of Watershed Water Quality Management According to Flow Conditions through Factor Analysis and Naïve Bayes Classifier. *Sustainability* **2023**, *15*, 10038. https://doi.org/10.3390/su151310038

Academic Editors: Venkata Krishna Kumar Upadhyayula and Carina da Conceição Mendes de Almeida

Received: 30 March 2023
Revised: 2 June 2023
Accepted: 14 June 2023
Published: 25 June 2023

1. Introduction

For water quality management, the assessment of the status of contaminations in river systems and the identification of problems within the watershed based on management goals are important. Appropriate countermeasures must be established to address these issues. A representative water quality measure is the total maximum daily load (TMDL) system, which allocates and manages the acceptable pollutant load from the set target water quality (TWQ) limit for each watershed. The acceptable pollutant load is determined using water quality modeling to ensure that the TWQ can be met under baseline flow conditions [1]. The evaluation of model calibration and validation uncertainties for TMDL reliability is limited. This is due to the complexity of water quality modeling, the availability and quality of data, and the absence of standardized guidelines [2]. Furthermore, this

system manages pollutant loads in each watershed to maintain TWQs at a specific standard flow rate; however, there are limitations to managing water quality in a watershed with different TMDL conditions than the standard flow [3]. Undoubtedly, water quality management of a watershed can be carried out in a scientific and systematic manner. However, it is problematic to establish a water quality management plan solely through a top-down uniform approach without taking into account the specific characteristics of the watershed in detail.

Nakdong River, the study area, is a watershed where TMDL has been implemented since 2004 [4]. To achieve TWQ, various pollution source management plans are being established, and watershed water quality management is being undertaken continuously through step-by-step TMDLs. For sustainable water quality management, an important evaluation factor is to analyze whether the assessed watershed water quality has achieved and maintained the TWQ [5]. In the Nakdong River TMDL, TWQ is evaluated as the average transformed water quality measurement for one year, without considering the reference flow. This method does not adequately reflect seasonal variability in water quality and flow. Although it is convenient for management to achieve TWQ, it is insufficient for the healthy recovery of water bodies as well as identifying and addressing fundamental watershed management problems [6]. In addition, the Nakdong River watershed has severe seasonal flow deviations, which causes difficulties in water quality management [7].

In the United States and Europe, the load duration curve (LDC) method is used to identify water pollution problems and to effectively evaluate TMDLs by considering the magnitude of water quality changes or pollutant loads under full river flow conditions [8]. The Nakdong River watershed was also studied to evaluate TWQs and review their adequacy using LDC [9]. The LDC analysis can be utilized as an analytical tool to identify water body damage and improve water quality problems by focusing on months and seasons. However, it has limited application when flow is not the dominant factor [3]. In addition, TWQ is limited to biological oxygen demand (BOD), an organic matter indicator [10], as well as total phosphorus (TP), a nutrient indicator [11]. Therefore, it is necessary to characterize the variation of TWQ owing to internal and external factors. Recently, various approaches, such as numerical modeling methods and data-driven analysis, have been used to assess watershed management water quality. In [12], seven methods were used to evaluate water quality impairment. In particular, the comprehensive water quality identification index was proposed as an effective method to evaluate water quality impaired by multiple pollutants. Ref. [13] analyzed the correlation between water quality variables and proposed an alternative method for water quality assessment through a hierarchical clustering analysis based on Mahalanobis distance analysis. In [14–17], multivariate statistical techniques were utilized to evaluate the spatial and seasonal variation of surface river water quality data. In [18], the usefulness of numerical modeling as a decision-making tool for water quality assessment was described and applied to predict future conditions after water quality assessment and implementation of remediation measures. However, numerical models have limitations in reproducing nonlinear water quality variations due to rainfall and various external watershed factors [19]. Most of the prior studies on water quality assessment for watershed management were conducted on a partial basis using specific seasonal or annual average water quality values, instead of assessing water quality according to flow fluctuations. It is crucial to identify water quality characteristics within a watershed according to flow conditions, and to set a customized direction for watershed management by period.

Therefore, this study analyzed the vulnerability of 22 watersheds in the Nakdong River system in South Korea to achieve TWQ. In addition, it examined the characteristics of water quality fluctuations according to flow conditions using factor analysis (FA). Using the FA results, the watershed pollution source characteristics were classified. In addition, the watersheds with difficulty in achieving TWQ were analyzed in detail through a factor analysis network graph (FANG). Finally, a data-driven prediction model was developed using a naïve Bayes classification model to predict whether TWQ is achieved or not.

2. Research Background

2.1. Descriptions of Study Area

The Nakdong River system consists of a total of 22 watersheds and eight multifunctional weirs, which are major hydraulic structures in the mainstream (Figure 1). The numbering of watersheds indicated in the Figure 1 has been accurately labeled with the corresponding watershed names in Table 1. To manage water quality effectively, target water quality (TWQ) standards are established within the framework of the national water management master plan, which is formulated by the government. These TWQ standards serve as a guideline for pollution source management activities implemented within the watershed. The Nakdong River Watershed Water Environment Management Plan is an overall plan that is the basis for subwatershed-specific water environment management plans in the Nakdong River watershed. The plan expands and develops water quality and water ecosystem conservation measures. It analyzes water quality standard achievements through sample analysis of representative points in the watershed, and it also conducts overall water environment management. Further, the watershed has undergone environmental water changes owing to the installation of multifunctional weirs due to the Major Rivers Project in 2012. This makes it an important watershed to evaluate the water quality management due to environmental changes [20]. The Major Rivers Project in 2012, also known as the Four Major Rivers Restoration Project, refers to a large-scale initiative undertaken in South Korea for the comprehensive restoration and management of four major rivers: the Han River, Nakdong River, Geum River, and Yeongsan River.

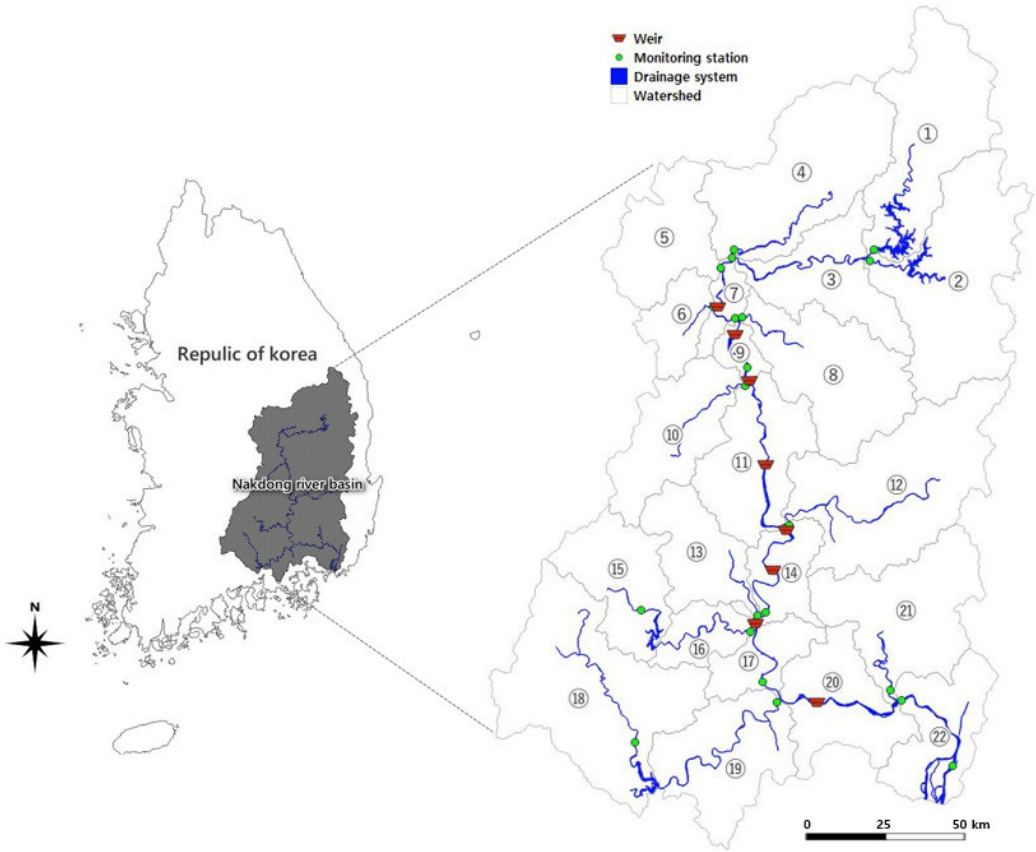

Figure 1. Nakdong River watershed status map.

Table 1. Water quality status of 22 watersheds in the Nakdong River system.

Watershed	Standard			Water Quality					
	Grade	BOD (mg/L)	T-P (mg/L)	BOD (mg/L)			T-P (mg/L)		
				'16	'17	'18	'16	'17	'18
① Andong Dam	Ia	1.0	0.020	0.9	0.9	0.9	0.021	0.023	0.027
② Imha Dam	Ib	2.0	0.040	1.4	1.1	1.0	0.018	0.015	0.017
③ Andong Dam downstream	Ia	1.0	0.020	1.2	1.3	1.3	0.017	0.039	0.023
④ Naeseongcheon	Ia	1.0	0.020	0.9	0.7	0.9	0.045	0.038	0.050
⑤ Yeong River	Ia	1.0	0.020	1.3	1.1	1.3	0.020	0.024	0.027
⑥ Byeongseongcheon	Ib	2.0	0.040	1.6	1.4	1.7	0.048	0.053	0.064
⑦ NakdongSangju	Ia	1.0	0.020	2.0	1.7	1.9	0.031	0.033	0.041
⑧ Wicheon	Ib	2.0	0.040	2.1	2.0	2.3	0.035	0.030	0.039
⑨ Gumi Wier	Ia	1.0	0.020	2.0	2.1	2.1	0.027	0.026	0.032
⑩ Gamcheon	Ia	1.0	0.020	1.5	1.1	1.1	0.055	0.038	0.050
⑪ Gangjunggoryeong Weir	Ib	2.0	0.040	2.5	2.4	2.4	0.034	0.030	0.040
⑫ Geumho River	II	3.0	0.100	3.7	3.8	2.9	0.085	0.070	0.060
⑬ Hoecheon	Ib	2.0	0.040	1.3	1.4	1.4	0.033	0.028	0.031
⑭ HapcheonChangnyeong Weir	II	3.0	0.100	2.4	2.2	2.8	0.053	0.039	0.052
⑮ Hapcheon Dam	Ib	2.0	0.040	1.3	1.4	1.3	0.032	0.033	0.035
⑯ Hwang River	Ia	1.0	0.020	0.9	0.6	0.7	0.029	0.020	0.023
⑰ NakdongChangnyeong	Ib	2.0	0.040	2.2	2.4	2.4	0.045	0.034	0.042
⑱ Namgang Dam	Ib	2.0	0.040	1.2	1.2	1.1	0.029	0.027	0.027
⑲ Namgang	Ib	2.0	0.040	2.2	2.6	2.2	0.040	0.042	0.036
⑳ NakdongMiryang	Ib	2.0	0.040	2.0	2.1	2.2	0.052	0.042	0.061
㉑ Miryang River	Ib	2.0	0.040	1.9	2.9	2.0	0.038	0.044	0.031
㉒ Nakdong River estuary	Ib	2.0	0.040	1.9	2.1	2.0	0.050	0.039	0.064

The main status and water quality of the 22 watersheds of the Nakdong River system are shown in Table 1 below. The water environment monitoring network data provided by the Water Environment Information System (http://water.nier.go.kr (accessed on 1 June 2019)) were used. The water quality target grade is set based on BOD and T-P. Currently, the conditions for grade Ia are less than 1 mg/L and 0.020 mg/L of BOD and T-P, respectively. Currently, eight watersheds are classified as Ia. They are relatively more prevalent in the upper watersheds. The conditions for grade Ib are less than 2 mg/L and 0.040 mg/L of BOD and T-P, respectively, and 12 watersheds are classified as Ib. These are mostly distributed in the middle and lower watersheds. The conditions for grade II are less than 3 mg/L and 0.100 mg/L of BOD and T-P, respectively, and two watersheds are classified as grade II.

2.2. Factor Analysis

Factor analysis (FA) is a technique that analyzes the interrelationships between variables using covariance and correlation between multiple variables [21]. In addition, it is based on the results, identifies the correlation and structure between questions and variables, and represents the information of multiple variables by grouping them into a small number of factors. By conducting an FA, information on multiple variables is condensed into a few key underlying factors. This makes the information more understandable and easier to analyze further [19]. The purpose of FA is to organize variables in groups by determining how they relate to each other. Among the studies for water quality management through FA, the authors in [22] inferred that the parameters responsible for groundwater chemistry through FA were due to mineral weathering of the parent rock, dissolution of chloride salts, excessive use of chemical-based fertilizers, and anthropogenic activities. Of note, [23] analyzed the main factors and empirical orthogonal function (EOF-patterns) which cause water quality fluctuations. In addition, the Nakdong River monitoring network was evaluated spatially and seasonally based on the contribution of each factor through exploratory factor analysis (EFA) and empirical orthogonal

functions. Through [24], the main analyzed parameters related to pollution were found to be nutrient factor, dissolved solids factor, and sediment factor. In addition, household, industrial, and agricultural activities all contribute to pollution sources in the study area. The studies listed above investigated the contribution and characteristics of water pollution through FAs. In contrast to previous studies, this study analyzed the characteristics of water pollution by flow rate and conducted a detailed analysis of the causes. In addition, this study aims to suggest the direction of pollution source management for achieving TWQ by watershed.

2.3. Naïve Bayes Classification

The naïve Bayes classifier, a traditional classifier for classification, is not part of the artificial neural network algorithm. However, it is a mainstream algorithm in machine learning and is known to perform well in classification. It is data-driven, not model-driven, based on conditional probabilities which do not make assumptions about the data [25,26]. The naïve Bayes classifier is a probability classifier type which uses Bayes' theorem, a probability theory that assumes independence between characteristics, and calculates the relationship between prior and posterior probabilities using conditional probabilities [27,28].

$$p(C_k|X) = \frac{p(X|C_k)p(C_k)}{p(X)} \quad (1)$$

The probability that new data falls into each classification is equal to the right-hand side of the equation. Without the assumption of independence between the independent variables in the data, the right-hand side of the equation requires complex computations to account for the effect of each independent variable on each other. However, with the independence assumption, it can be implemented as a multiplicative operation on the probability of each independent variable, as shown in the equation, and can be calculated simply. In the naïve Bayes (NB) model, the parameters are conditionally independent, and therefore it is simple to manipulate data (add, delete, and change) within the network [29]. Existing water quality studies using NB are characterized by the training data for classifying pollution sources, as well as predicting that water quality grades are all limited to water quality variables [30–32]. Therefore, this study aimed to develop a model to classify TWQ grades using flow and drought conditions that could directly affect water quality. It is believed that these meteorological variables can be implemented in a sophisticated model that considers the nonlinear relationship to water quality variation characteristics.

3. Study Methods

The Nakdong River watershed implements water quality management through the division of low and normal seasons as part of the TMDL system. In this study, the flow data from all time periods was sorted in descending order based on the results of the Hydrological Simulation Program—FORTRAN (HSPF) model conducted in the Phase IV TMDL Management Criteria Setting Study (I) for the Nakdong River. This enabled the development of a flow duration curve. The water quality standard excess rates were analyzed in 22 watersheds by dividing the range into four flow sections instead of five flow sections on the flow duration curve presented in EPA (2007) [8]. The flood season was excluded from the analysis due to the limited occurrence of events and insufficient water quality measurement data. A flow duration curve was calculated through the HSPF watershed model, and FA and naïve Bayes classification were applied to classify low- and high-flow periods. Data preprocessing was conducted as well. The FA results were visually represented and analyzed using a network graph to understand the water quality variation characteristics of each factor. Moreover, the water quality variation characteristics of each watershed were examined, and the main influencing factors on water quality based on flow duration were identified. Additionally, the water quality variation characteristics of each watershed were classified according to the pollution source associated with each flow duration. The data used in this study were collected and analyzed from 2006 to 2018.

A water quality excess probability model was developed and validated using a naïve Bayes classifier for each of the 22 watersheds in the Nakdong River system. The train data consisted of 60% of the total data, and 40% was utilized as validation data. The predictor variables in this study were the achievement of water quality control items, specifically BOD and T-P. The condition variables considered were flow duration conditions, including four periods: base flow, low flow, average flow, and flood flow. Additionally, we incorporated seven hydrological drought conditions based on the standardized precipitation index (SPI), twelve conditions corresponding to each month of the year, and conditions based on compliance with water quality standards. No existing studies apply the naïve Bayes classifier to predict TWQ grade by applying flow conditions and drought conditions as well as water quality variables. Therefore, this study proposes a technique for water quality management in a watershed by developing a water quality excess probability model for watershed water quality evaluation, as well as the preemptive prediction through data-based analysis instead of conventional numerical modeling. The overall flow and key summary of the research are described in Figure 2.

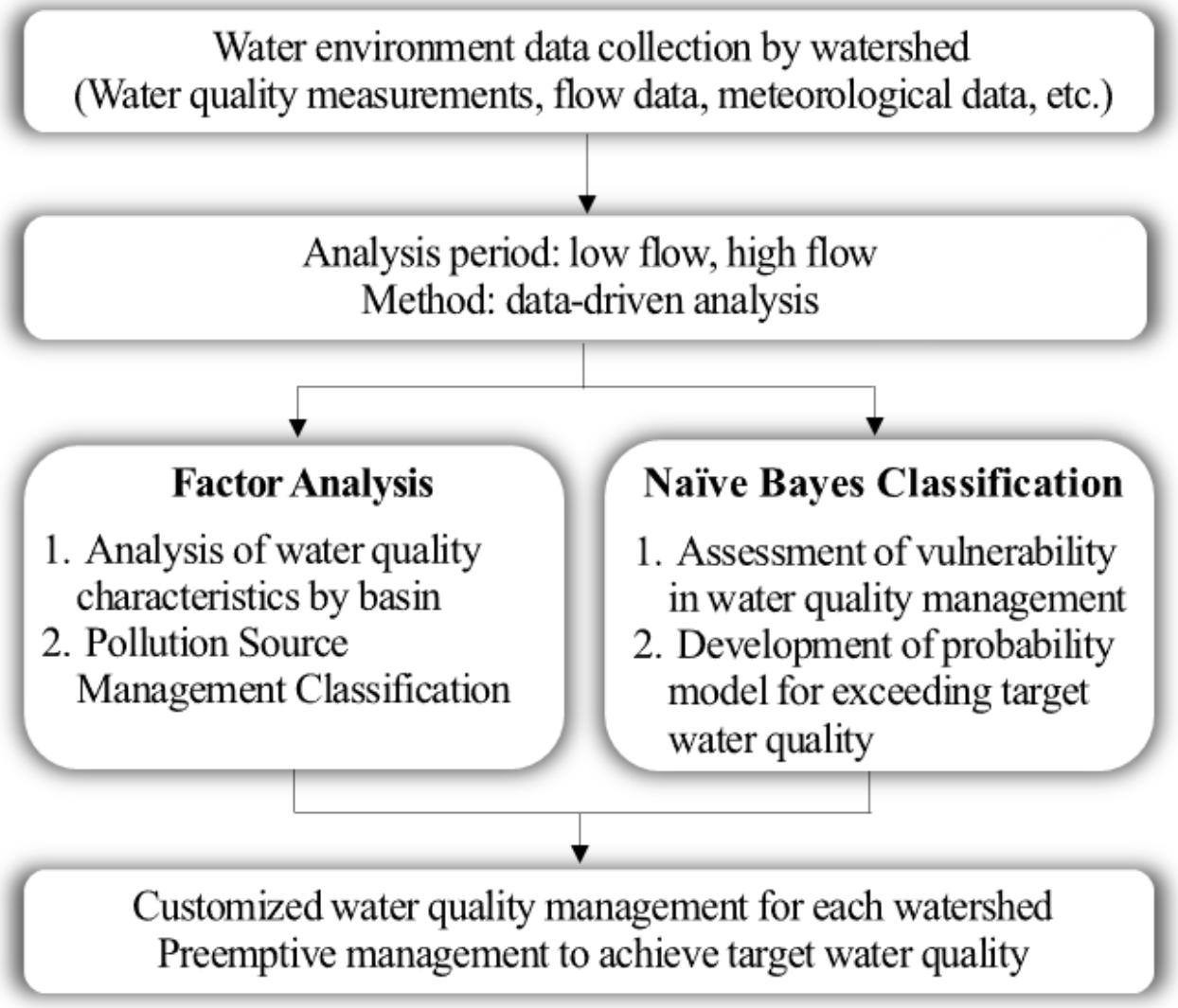

Figure 2. Research flow chart and method of deriving results.

4. Results

4.1. Evaluation of Water Quality Vulnerability

The ratio of attainment to non-attainment values of water quality standards for the watersheds were divided into four grades, and utilized as vulnerability assessment met-

rics. Figures 3 and 4 show the vulnerability of achieving TWQ for each watershed through heat mapping.

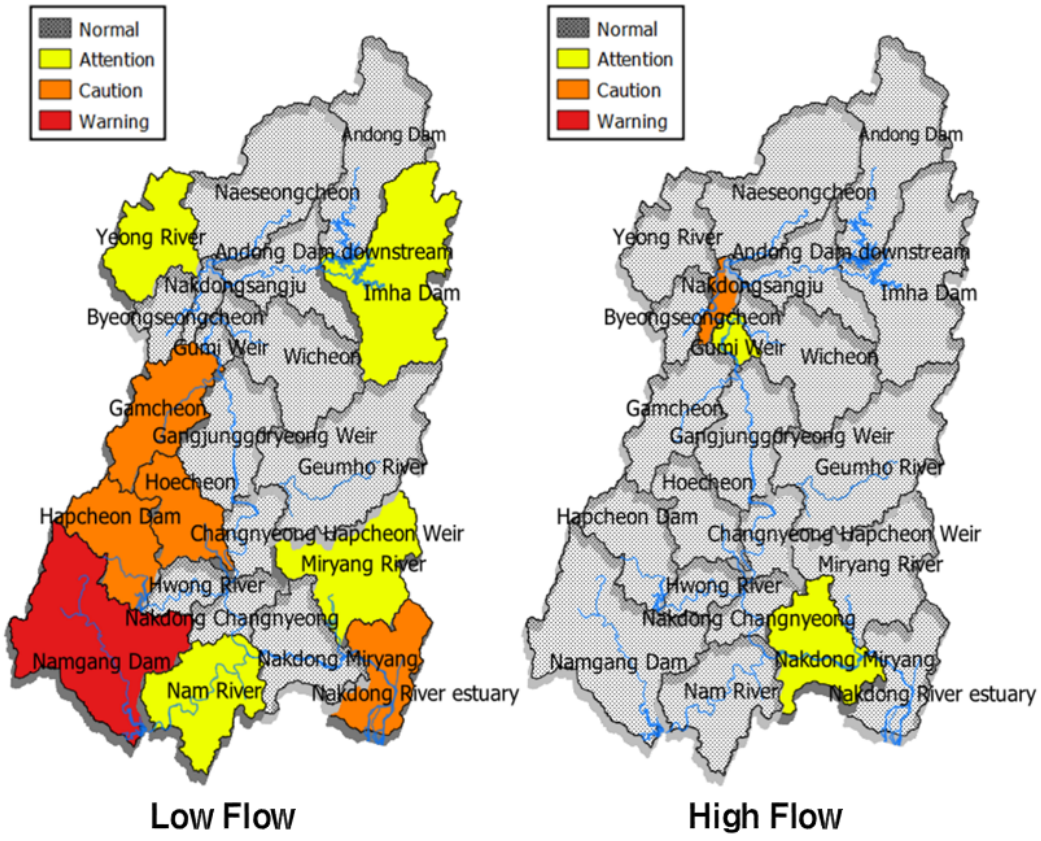

Figure 3. Assessment of vulnerability in achieving target water quality according to flow conditions (BOD).

The BOD was found to have a high probability of exceeding the target standard under low flow (approximately 60–100%) conditions. Gamcheon, Hoecheon, Hapcheon Dam, and Nakdong River estuary watersheds were found to be particularly vulnerable to BOD water quality control with a probability of exceeding the target standard by more than 1.4 times. Under high flow (approximately 10–60%) conditions, Nakdong Sangju, Gumi Weir, and Nakdong Milyang middle watersheds were found to be vulnerable to BOD with a probability of exceeding the target standard by 1.2 times or more.

In contrast to the BOD results, T-P showed a higher probability of exceeding the target standard under high flow (approximately 10–60%) conditions. Eight watersheds, including Imha Dam, Naeseong Stream, Nakdong Sangju, Gumi Weir, Gamcheon, Gangjeong Goryeong Weir, Hapcheon Changnyeong Weir, and Hwanggang Middle Watershed, were found to be particularly vulnerable to T-P with a probability of exceeding the target standard by 1.4 times or more. Under low-flow (approximately 60–100%) conditions, Hapcheon Dam and Namgang Dam middle watershed were found to be vulnerable to T-P water quality management with a probability of exceeding the target standard by 1.2 times. Even though watershed management is currently being implemented through various water management policies, the target standard for middle watershed management in the Water Environment Management Plan was simplified. Therefore, it is necessary to set custom target standards considering watershed conditions and characteristics, and in addition, the target standards for each middle watershed required an increase.

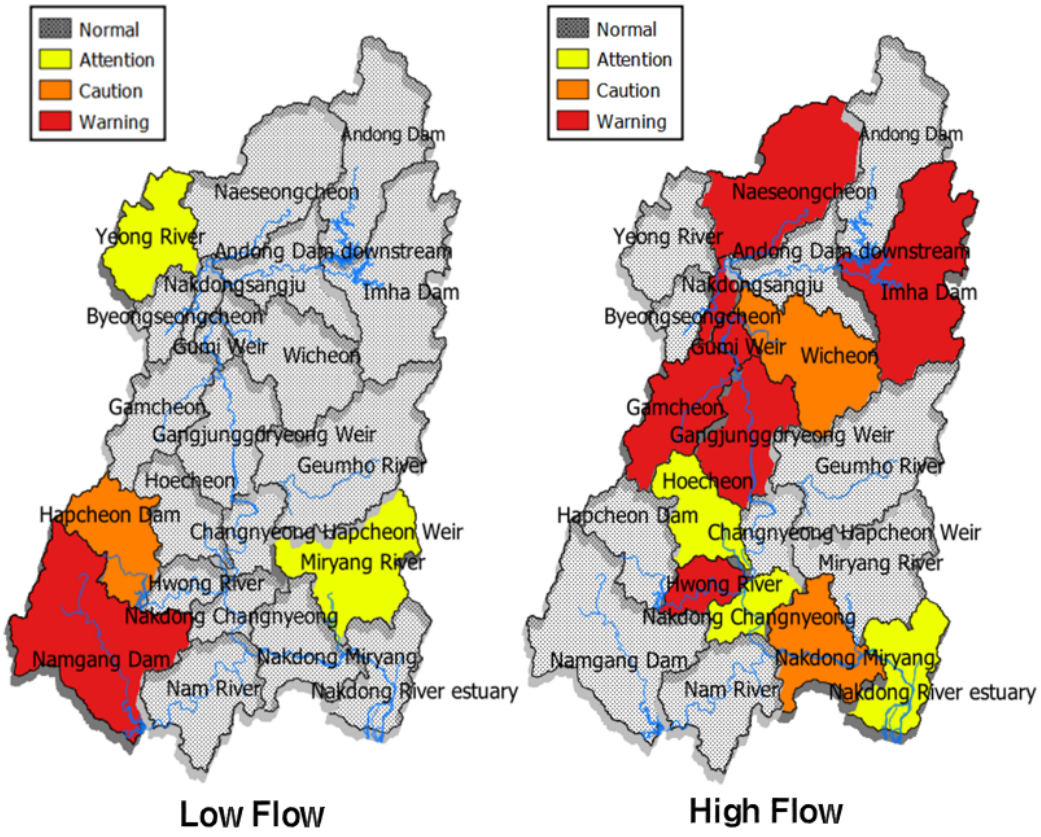

Figure 4. Assessment of vulnerability in achieving target water quality according to flow conditions (T-P).

4.2. Detailed Analysis of Vulnerable Watersheds

Hapcheon Dam and Namgang Dam watersheds were selected as watersheds vulnerable to water quality management during low flow periods. Therefore, they were categorized into priority management watersheds according to low flow, and FA was conducted. Based on the FA results, the following networking visualization analysis was conducted. In Figures 5 and 6, the circle plot layout node represents the classified factors, while the square plot layout node represents the variables. Correlations between the factors are color-coded. Green and magenta indicate a positive and negative correlation, respectively. The brightness of the color displays the scale of the correlation. The higher the correlation between variables per node, the closer they are distributed to each other [33].

The Hapcheon Dam watershed shows different water quality variation characteristics during low and high flows. The peculiarity is that the phosphorus-related water quality items clustered in factor 2 are negatively correlated with flow during low flow periods. This contrasts with the phenomenon that phosphorus is generally introduced by nonpoint pollution sources due to rainfall runoff.

In the Namgang Dam watershed, we found that the variation of water quality characteristics during low and high flows are only marginally different. At low flows, phosphorous-based water quality items and total coliform counts are grouped in the same factor. This contrasts with the influx of livestock manure, a nonpoint source of pollution that is generally [34] attributed to rainfall runoff. It is believed that runoff from these nonpoint sources also occurs at low flows.

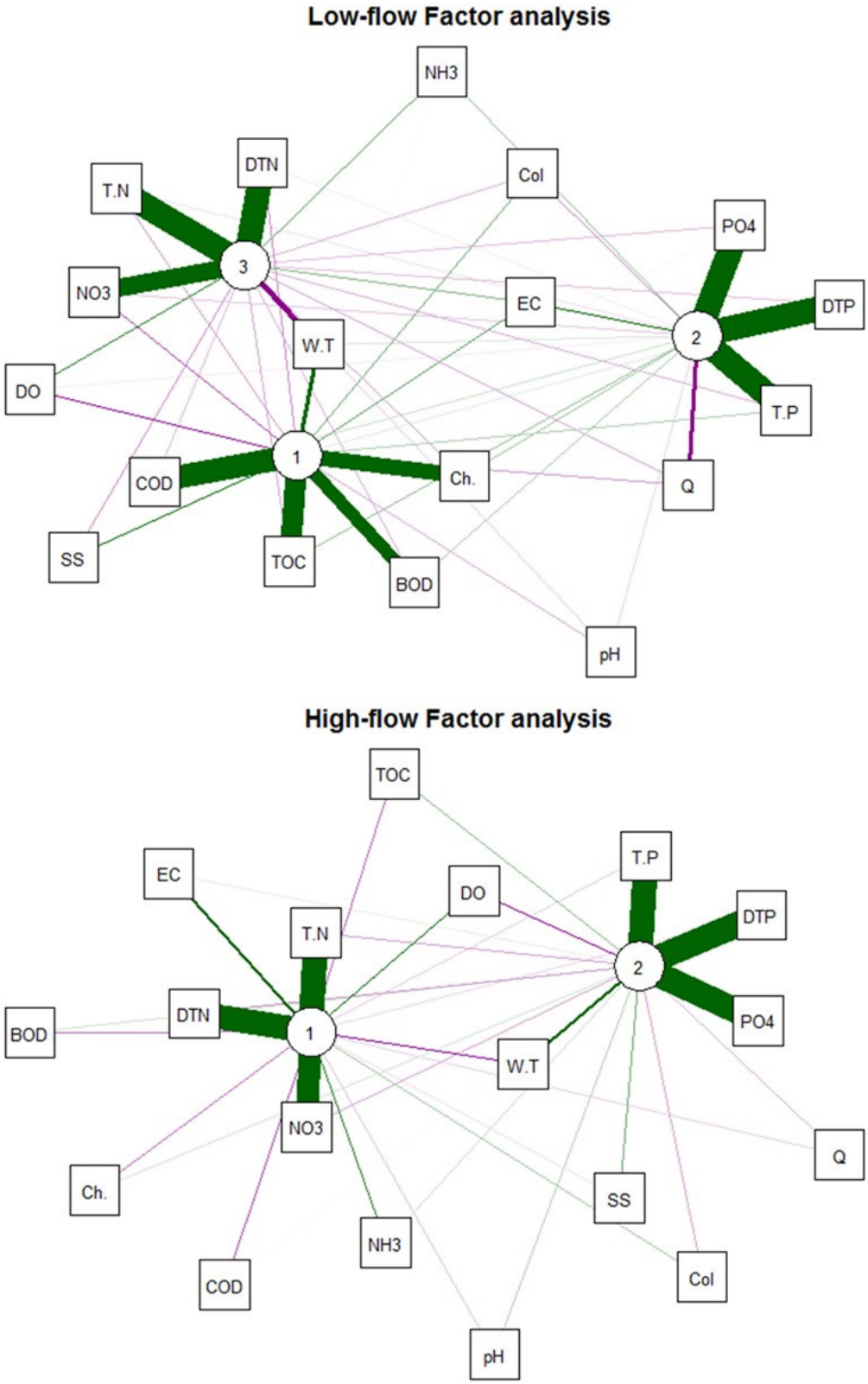

Figure 5. Hacheon Dam factor analysis network graph. W.T: water temperature, DTN: dissolved total nitrogen, DTP: dissolved total phosphorus, Col: total coliform count, Q: Flow rate, Ch.: chlorophyll a, NO3: nitrate nitrogen NH3: ammonia nitrogen, PO4: phosphate phosphorus, EC: electrical conductivity, DO: dissolved oxygen.

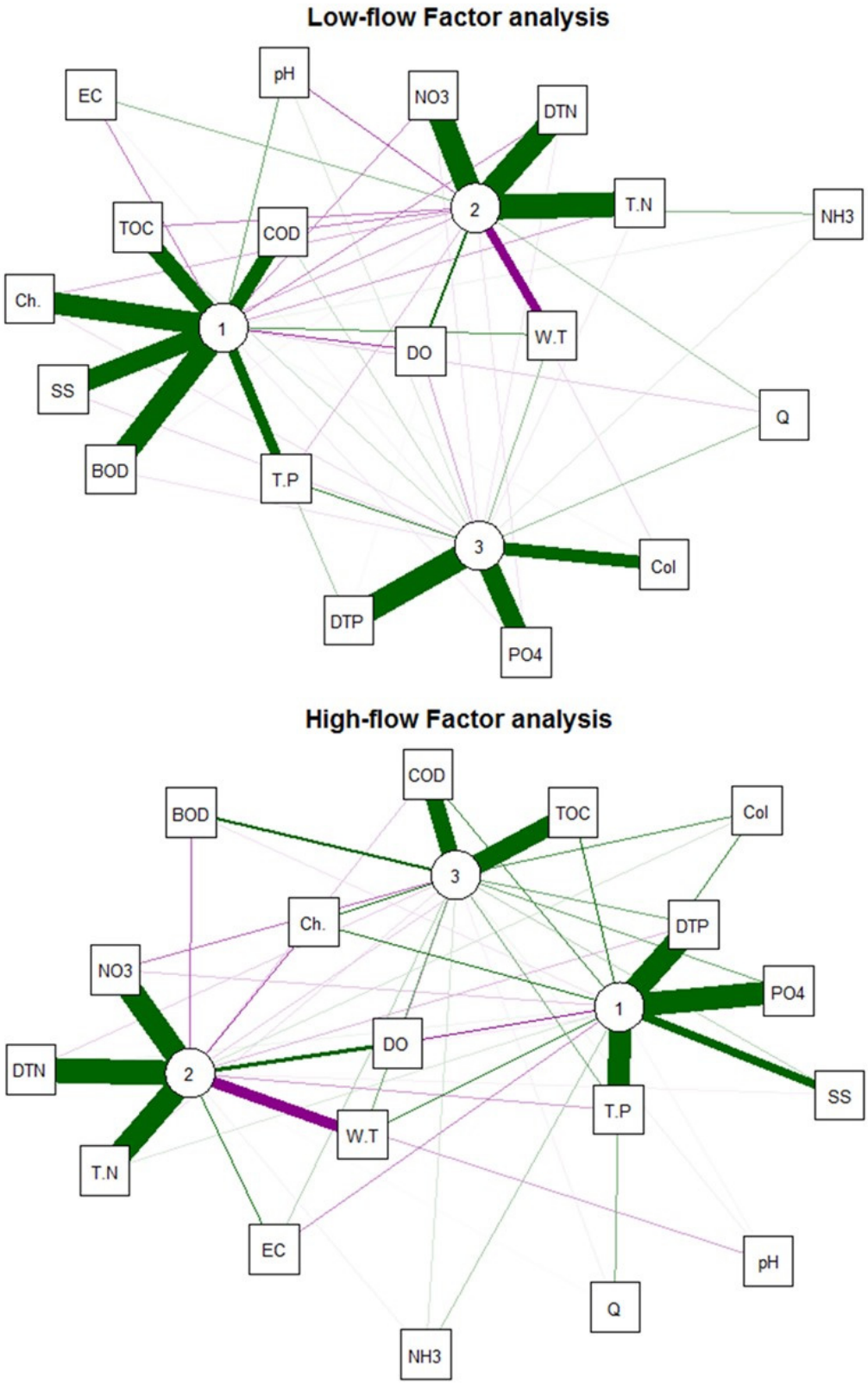

Figure 6. Namgang Dam factor analysis network graph. W.T: water temperature, DTN: dissolved total nitrogen, DTP: dissolved total phosphorus, Col: total coliform count, Q: Flow rate, Ch.: chlorophyll a, NO3: nitrate nitrogen NH3: ammonia nitrogen, PO4: phosphate phosphorus, EC: electrical conductivity, DO: dissolved oxygen.

Figures 7 and 8 below display the results of analyzing the pollutant load of two watersheds. The Hapcheon Dam watershed was characterized by nutrients (nitrogen and phosphorus) and organic matter. Furthermore, the proportion of livestock farming in the pollutant load was very high (BOD: 56%, T-P: 51%). Phosphorous-based water quality items were negatively correlated with flow. This contrasts with the influx of phosphorus from nonpoint pollution sources during high flows. This suggests that certain pollution sources are entering the watershed during low flows. In particular, the T-P discharge concentration of Geochang Wastewater Treatment Plant, a major sewage treatment facility, was 0.169 mg/L, which is much higher than the TWQ of 0.04 mg/L. Thus, it is necessary to strengthen the effluent concentration of sewage treatment plants during low-flow periods.

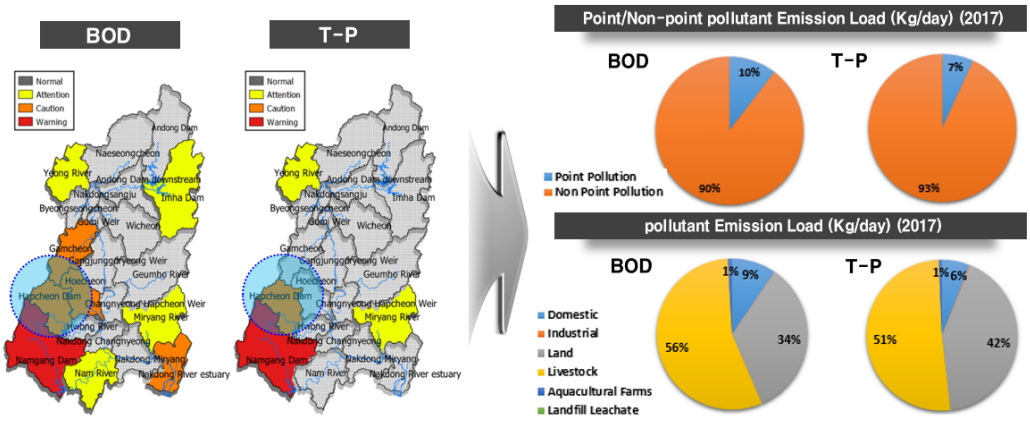

Figure 7. Hapcheon Dam pollution source status analysis.

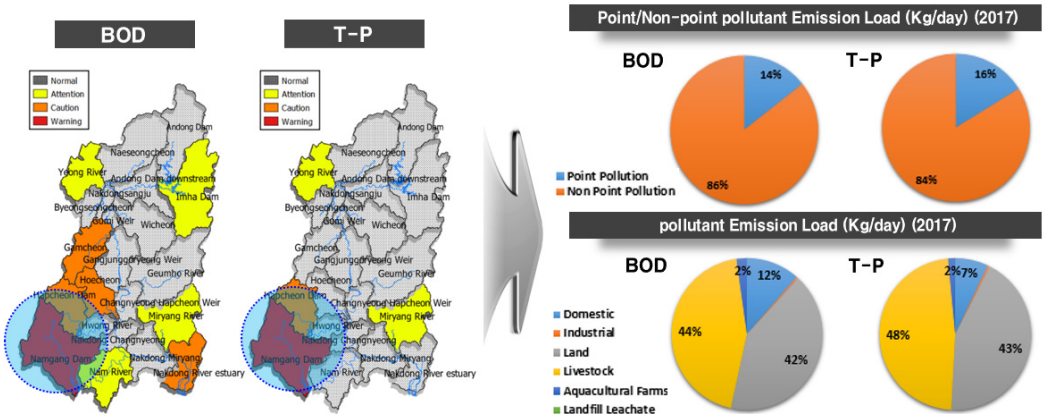

Figure 8. Namgang Dam pollution source status analysis.

Namgang Dam watershed also has a very high proportion of livestock farming in the pollutant load (BOD: 44%, T-P: 48%). The FA results showed that phosphorus and total coliform bacteria belong to the same factor. This suggests that runoff from livestock pollution sources occurs at low flows. In addition, with nine public watershed treatment plants (16,600 tons/day in total) and 145 small-scale public watershed treatment plants (11,562 tons/day in total), it should be necessary to control the discharge amount of treatment facilities during low flow periods in addition to managing livestock pollution sources.

The FA can be used to infer the watershed water pollution characteristics. Thus, the water quality characteristics of a watershed can be identified efficiently and easily through utilizing FA in a watershed. Before conducting detailed pollution source analysis, it is effective for water quality management to identify flow water quality characteristics within a watershed, even if only using water quality data.

4.3. Classification of Water Quality Characteristics

Pollution source characteristics were classified based on the FA results for each watershed. The water quality variables which were characterized by pollution source characteristics such as organic matter, nutrients, rainfall runoff, sewage, and livestock wastewater are shown in the Table 2. Water quality variables were grouped by each factor using the results of [23]. The variation characteristics of water pollution sources for each factor were compared and summarized. In addition, the pollution source characteristics under low and high flow conditions were analyzed for each watershed.

Table 2. Relevant source of pollution for water quality parameters.

W.Q. Parameter	Relevant Source of Pollution
pH	organic matter, biochemical reactions of pollutant, atmospheric inputs, chemical contaminants (industrial wastewater)
BOD	organic matter, sewage
COD	organic matter, sewage
SS	organic matter, algal or plankton blooms, aquatic plants, stormwater run-off
TN, TP	stormwater run-off, fertilizers, domestic wastewater
NH3-N	municipal and agricultural wastewater
NO3-N	biological treatment plants, eutrophication, algal blooms
DTN, DTP	underground water seeping, sewage
W.T.	inflow of tributaries or discharge from industrial or wastewater treatment plant, change of seasons
EC	industrial wastewater, eutrophication, algal blooms, salinity
FC, TC	livestock wastewater, sewage
Chl-a	algal or plankton blooms, aquatic plants

Quantitative pollution source classification was performed using the factor matrix values of the watershed, and the FA results displayed various pollution source characteristics in the analyzed results. Two water quality variables, BOD and Chl-a, were used to classify water pollution sources according to biodegradable organic matter. Nutrients were classified into N-based and P-based items to characterize water pollution sources. The chemical oxygen demand (COD) and total organic carbon (TOC) were used to characterize the water pollution sources based on poorly degradable organic matter. The classification criteria are similar to the eigenvector of the principal component analysis (PCA). The average value of each water quality variable was calculated using the factor loading value, which indicates the influence of each common factor on the measured variable. Subsequently, the pollution source status of each watershed was evaluated for relative flow condition vulnerabilities. Figure 9 displays the average water quality factor loading value of the variables of 22 watersheds classified by pollution source, divided by quantile, and compared by heat mapping. It was possible to distinguish between watersheds with various pollution source characteristics and watersheds which required specific pollution source management. Through this result, it was possible to make a primary decision on which water pollution source to manage according to flow conditions.

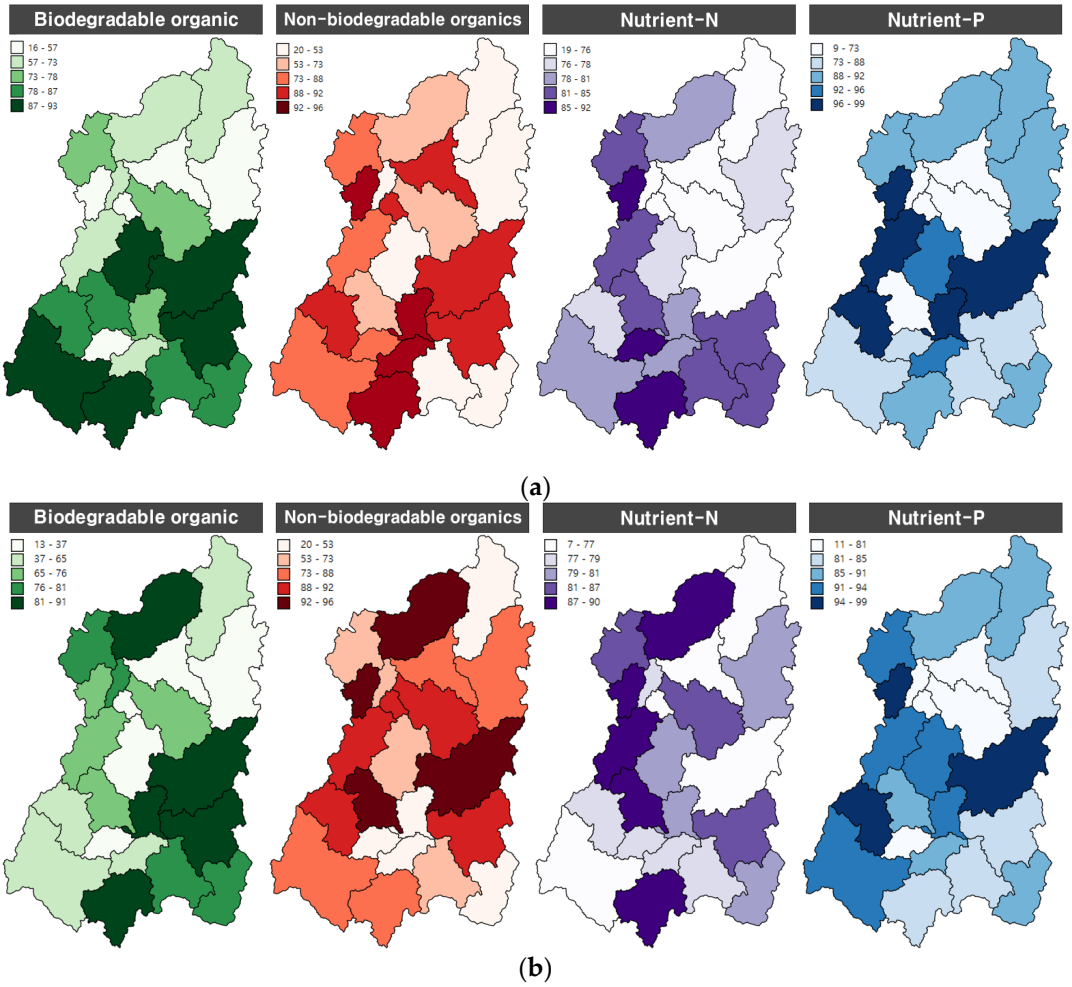

Figure 9. Quantitative assessment heat map of contaminants by basin. They should be listed as: (**a**) low flow; (**b**) high flow.

4.4. Water Quality Excess Probability Model

The training and validation results of the water quality excess probability model are presented in Table 3. In the training phase, the average BOD for the 22 watersheds was 82.39%, while in the validation phase, it was 72.67%. Similarly, the average T-P in the training and validation phases was 76.46% and 80.60% respectively. These results provide an overview of the performance of the model in predicting water quality exceedances for the studied watersheds. The model accuracy varied for each watershed. Generally, it is not common to use the naïve Bayes classifier for water quality prediction. However, the naïve Bayes classifier assists in decision-making for rapid water quality management as it can provide a final decision based on probability distribution, i.e., known uncertainty [19]. In addition, from the results, the nature of the naïve Bayes classification model is considered to be based on probability according to various conditions. Therefore, it can be supplemented with a precise model through data accumulation in the future. As a data-driven model, it is expected to contribute to rapid decision-making for TWQ management in watersheds by reducing the extensive input data and time required for conventional physical-based models for watershed water quality management.

Table 3. Naïve Bayes classifier accuracy results.

	Accuracy (%)			
	BOD		T-P	
	Training	Validation	Training	Validation
Andong Dam	83.91	80.17	81.03	70.69
Imha Dam	93.05	90.40	97.33	93.60
Andong Dam downstream	73.08	65.71	78.85	68.57
Naeseongcheon	74.59	73.39	86.49	82.26
Yeong River	80.27	71.43	76.87	72.45
Byeongseongcheon	90.21	87.50	76.92	73.96
NakdongSangju	96.34	87.50	73.17	75.00
Wicheon	70.95	72.00	74.32	74.00
Gumi Wier	98.04	97.00	84.31	70.59
Gamcheon	73.33	72.22	97.04	96.67
Gangjunggoryeong Weir	75.26	68.18	85.57	84.85
Geumho River	80.54	70.16	80.00	71.77
Hoecheon	84.87	81.37	86.18	82.35
HapcheonChangnyeong Weir	82.99	82.83	89.80	88.89
Hapcheon Dam	92.57	84.85	69.59	70.71
Hwang River	79.23	80.68	81.54	79.55
NakdongChangnyeong	74.36	71.15	73.72	71.15
Namgang Dam	86.58	89.00	93.29	84.00
Namgang	74.73	72.58	71.51	72.58
NakdongMiryang	82.89	61.54	68.42	67.31
Miryang River	80.67	70.00	77.33	73.00
Nakdong River estuary	84.13	65.12	69.84	58.14
Average	82.39	72.67	80.60	76.46

5. Discussion

The varying accuracies of the model across different watersheds indicate that the performance of the model can vary depending on the characteristics and conditions of each watershed. This suggests that various factors such as geographical, climatic, and geological characteristics of the watershed, types and concentrations of pollutants, and the state of water quality management systems can influence the prediction of water quality exceedances. For example, one watershed may experience frequent water quality exceedances at higher levels due to high concentrations of pollutants from industrial areas. On the other hand, another watershed may have fewer occurrences of water quality exceedances due to relatively lower concentrations of pollutants from agricultural activities. These differences can lead to variations in the predictive accuracy of the model for each watershed. Furthermore, there may be other conditions that influence the prediction of water quality exceedances. Factors such as rainfall, inflow and outflow rates, and land use patterns within the watershed can also play important roles in predicting water quality exceedances. Therefore, it is important to consider these diverse conditions in order to improve and refine the model. These findings can help water quality managers gain a better understanding of the water quality status in specific watersheds and develop appropriate management and improvement strategies. Additionally, by incorporating and addressing these various factors, the model can be further enhanced to achieve more accurate predictions in the future.

The data-driven model, as opposed to the physics-based model, can capture nonlinear relationships through powerful data modeling techniques. While physics-based models rely on the understanding of physical principles and processes, they may struggle to accurately capture complex nonlinear relationships. In contrast, data-driven models leverage various input data and statistical analyses to learn and predict nonlinear relationships. These data models excel in capturing nonlinear factors and intricate interactions, thereby enhancing the accuracy of predictions. Data models utilize statistical techniques, machine learning algorithms, artificial neural networks, and other approaches to learn the relation-

ship between input data and the desired output (in this case, water quality prediction). By doing so, data models identify patterns and features in the input data and generate predictions considering nonlinear relationships. Consequently, data models offer more flexibility in considering diverse variables and conditions, leading to improved prediction accuracy compared to physics-based models. The application of data models holds significant implications in the field of environmental water management. By learning and predicting nonlinear relationships, data models can effectively consider multiple variables and complex interactions, thereby aiding in accurate water quality prediction, monitoring, and the development of water quality management systems. Therefore, the data mining techniques presented in this study can serve as valuable resources for the development of data-driven models that capture nonlinear relationships and contribute to the advancement of water quality management technologies.

6. Conclusions

In this study, FA was used to classify customized water pollution source management indicators for each watershed according to low and high flows. In addition, FANGs were used to visualize the water quality variation characteristics during low and high flows per watershed unit. Pollution source classification was performed using the factor matrix values of FA results. Watersheds exhibiting diverse pollution source characteristics were classified into specific pollution source management categories. To effectively manage water quality in these watersheds, a naïve Bayes classification model was developed to predict the attainment of TWQ. Various conditions, such as flow conditions, month, drought index, and compliance with water quality standards were considered as predictors to determine if the water quality standards have been exceeded. This predictive model provides valuable insights for proactive water quality management strategies in the watershed, aiding in the identification and implementation of appropriate pollution control measures. However, since it is based on probabilities according to the various conditions, it could be improved to a precise model through data accumulation in the future. Physical-based models, currently used as decision-making tools for policy water quality management, require large amounts of input data and are time-consuming. However, the existing development time can be shortened, and in addition. The analysis time can be shortened using various data-based tools applied in this study. It can be used as a powerful data model through nonlinear relationship learning compared to physical models. Therefore, it is necessary to build vast water environment big data, with the improvement of real-time monitoring systems in watersheds, and the development of spatial and temporal observation technologies. For future studies, the data mining techniques analyzed in this study can be applied as water quality management technologies in watersheds to establish an advanced watershed management system.

Author Contributions: Conceptualization, W.S.J. and Y.D.K.; methodology, W.S.J.; formal analysis, W.S.J. and Y.D.K.; writing—original draft preparation W.S.J.; writing—review and editing, W.S.J. and Y.D.K.; supervision, Y.D.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Korea Environment Industry & Technology Institute (KEITI) through Aquatic Ecosystem Conservation Research Program (or Project), funded by Korea Ministry of Environment (MOE) (2022003050007).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Hwang, H.S.; Park, J.H.; Kim, Y.S.; Rheu, D.H.; Choi, Y.J.; Lee, S.J. Research on How to Set 3rd Phase Target Water Quality on the Boundary between Metropolitan Cities/Dos Specified in Nakdong River Basin. *J. Korean Soc. Water Environ.* **2017**, *33*, 70–77.
2. Ahmadisharaf, E.; Camacho, R.A.; Zhang, H.X.; Hantush, M.M.; Mohamoud, Y.M. Calibration and validation of watershed models and advances in uncertainty analysis in TMDL studies. *J. Hydrol. Eng.* **2019**, *24*, 03119001. [CrossRef]
3. Lee, S.; Kang, J.; Park, H.; Kang, J.; Kim, S.; Kim, J.P.; Kim, G. A Study on the Variation of Water Quality and the Evaluation of Target Water Quality Using LDC in Major Tributaries of Nakdong River Basin. *J. Korean Soc. Water Environ.* **2020**, *36*, 521–534.
4. Cho, H.K.; Kwak, E.T.; Kim, S.M. A Study on the Spatial Variation of Target Water Quality and Excess Rate at 41 Stations in Nakdong River Basin after the Total Maximum Daily Loads. *J. Korean Soc. Agric. Eng.* **2020**, *62*, 97–109.
5. Kim, S.; Kim, H.; Han, M. Assessment of Total Maximum Daily Loads (TMDLs) for water quality management in the Nakdong River basin, South Korea. *Environ. Monit. Assess.* **2018**, *190*, 399.
6. uration Curve (LDC) to Evaluate the Achievement Rate of Target Water Quality in the Nakdong River Unit Watersheds. *J. Environ. Sci. Int.* **2017**, *26*, 433–445. [CrossRef]
7. Choi, J.Y.; Kim, J.W.; Kim, M.H. Evaluation of water quality changes in the Nakdong River watershed under climate change scenarios. *Water* **2019**, *11*, 849.
8. Environmental Protection Agency. *An Approach for Using Load Duration Curves in the Development of TMDLs*; Environmental Protection Agency: Washington, DC, USA, 2007.
9. Kim, M.H.; Yoon, C.G. Development of an evaluation method for target water quality (TWQ) in watersheds using a data-based classification model. *Environ. Earth Sci.* **2016**, *75*, 146.
10. Boopathy, R. Factors limiting bioremediation technologies. *Bioresour. Technol.* **1999**, *74*, 63–67. [CrossRef]
11. Hecky, R.E.; Kilham, P. Nutrient limitation of phytoplankton in freshwater and marine environments: A review of recent evidence on the effects of enrichment. *Limnol. Oceanogr.* **1988**, *33*, 796–822. [CrossRef]
12. Ji, X.; Dahlgren, R.A.; Zhang, M. Comparison of seven water quality assessment methods for the characterization and management of highly impaired river systems. *Environ. Monit. Assess.* **2016**, *188*, 15. [CrossRef] [PubMed]
13. Du, X.; Shao, F.; Wu, S.; Zhang, H.; Xu, S. Water quality assessment with hierarchical cluster analysis based on Mahalanobis distance. *Environ. Monit. Assess.* **2017**, *189*, 335. [CrossRef]
14. Barakat, A.; El Baghdadi, M.; Rais, J.; Aghezzaf, B.; Slassi, M. Assessment of spatial and seasonal water quality variation of Oum Er Rbia River (Morocco) using multivariate statistical techniques. *Int. Soil Water Conserv. Res.* **2016**, *4*, 284–292. [CrossRef]
15. Kükrer, S.; Mutlu, E. Assessment of surface water quality using water quality index and multivariate statistical analyses in Saraydüzü Dam Lake, Turkey. *Environ. Monit. Assess.* **2019**, *191*, 71. [CrossRef] [PubMed]
16. Muangthong, S.; Shrestha, S. Assessment of surface water quality using multivariate statistical techniques: Case study of the Nampong River and Songkhram River, Thailand. *Environ. Monit. Assess.* **2015**, *187*, 548. [CrossRef]
17. Rakotondrabe, F.; Ngoupayou, J.R.N.; Mfonka, Z.; Rasolomanana, E.H.; Abolo, A.J.N.; Ako, A.A. Water quality assessment in the Bétaré-Oya gold mining area (East-Cameroon): Multivariate statistical analysis approach. *Sci. Total. Environ.* **2018**, *610*, 831–844. [CrossRef]
18. Menendez, A.N.; Badano, N.D.; Lopolito, M.F.; Re, M. Water quality assessment for a coastal zone through numerical modeling. *J. Appl. Water Eng. Res.* **2013**, *1*, 8–16. [CrossRef]
19. Jung, W.S.; Kim, S.E.; Kim, Y.D. Prediction of Surface Water Quality by Artificial Neural Network Model Using Probabilistic Weather Forecasting. *Water* **2021**, *13*, 2392. [CrossRef]
20. Hair, J.F.; Black, W.C.; Babin, B.J.; Anderson, R.E.; Tatham, R.L. *Multivariate Data Analysis*, 8th ed.; Cengage Learning: Boston, MA, USA, 2019.
21. Jung, W.S.; Kim, Y.D. Effect of abrupt topographical characteristic change on water quality in a river. *KSCE J. Civ. Eng.* **2019**, *23*, 3250–3263. [CrossRef]
22. Kale, A.; Bandela, N.; Kulkarni, J.; Raut, K. Factor analysis and spatial distribution of water quality parameters of Aurangabad District, India. *Groundw. Sustain. Dev.* **2020**, *10*, 100345. [CrossRef]
23. Kim, S.E.; Seo, I.W.; Choi, S.Y. Assessment of water quality variation of a monitoring network using exploratory factor analysis and empirical orthogonal function. *Environ. Model. Softw.* **2017**, *94*, 21–35. [CrossRef]
24. Loi, J.X.; Chua, A.S.M.; Rabuni, M.F.; Tan, C.K.; Lai, S.H.; Takemura, Y.; Syutsubo, K. Water quality assessment and pollution threat to safe water supply for three river basins in Malaysia. *Sci. Total. Environ.* **2022**, *832*, 155067. [CrossRef] [PubMed]
25. Rish, I. An Empirical Study of the Naïve Bayes Classifier. In Proceedings of the IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, Seattle, WA, USA, 4–6 August 2001; Volume 41, pp. 412–417.
26. Zhang, H. The optimality of Naïve Bayes. In Proceedings of the 17th International Florida Artificial Intelligence Research Society Conference (FLAIRS), Miami Beach, FL, USA, 12–14 May 2004; Volume 2, pp. 562–567.
27. Zhang, H.; Batuwita, R. Optimality of naive Bayes for text classification under linguistic perspective. *IEEE Trans. Knowl. Data Eng.* **2006**, *18*, 525–535.
28. Wang, H.; Fan, W. Effective measures to cope with attribute incompleteness in naïve Bayes classifiers. *Data Knowl. Eng.* **2005**, *55*, 191–210.
29. Ilić, M.; Srdjević, Z.; Srdjević, B. Water quality prediction based on Naïve Bayes algorithm. *Water Sci. Technol.* **2022**, *85*, 1027–1039. [CrossRef]

30. Sakizadeh, M. Assessment the performance of classification methods in water quality studies, A case study in Karaj River. *Environ. Monit. Assess.* **2015**, *187*, 573. [CrossRef] [PubMed]
31. Babbar, R.; Babbar, S. Predicting river water quality index using data mining techniques. *Environ. Earth Sci.* **2017**, *76*, 504. [CrossRef]
32. Kyungpook National University. *A Study on the Establishment of Management Standards of Total Maximum Daily Loads for the Four Major River Basins*; National Institute of Environmental Research: Incheon, Republic of Korea, 2016.
33. R for Earth-System Science. Available online: https://pjbartlein.github.io/REarthSysSci/PCA.html (accessed on 1 June 2019).
34. Carpenter, S.R.; Caraco, N.F.; Correll, D.L.; Howarth, R.W.; Sharpley, A.N.; Smith, V.H. Nonpoint pollution of surface waters with phosphorus and nitrogen. *Ecol. Appl.* **1998**, *8*, 559–568. [CrossRef]