water

MDPI

Article

Integrated Quality Control Process for Hydrological Database: A Case Study of Daecheong Dam Basin in South Korea

Gimoon Jeong [1], Do-Guen Yoo [2], Tae-Woong Kim [3], Jin-Young Lee [4], Joon-Woo Noh [5] and Doosun Kang [1,*]

[1] Department of Civil Engineering, Kyung Hee University, 1732 Deogyeong-daero, Giheung-gu, Yongin 17104, Korea; gimoon1118@gmail.com
[2] Department of Civil Engineering, The University of Suwon, Hwaseong 18323, Korea; dgyoo411@suwon.ac.kr
[3] Department of Civil and Environmental Engineering, Hanyang University, Ansan 15588, Korea; twkim72@hanyang.ac.kr
[4] Rural Research Institute, Korea Rural Community Corporation (KRC), Sejong 30130, Korea; hydrojy@gmail.com
[5] K-Water Research Institute, K-Water (Korea Water Resources Corporation), Daejeon 34045, Korea; jnoh@kwater.or.kr
* Correspondence: doosunkang@khu.ac.kr

Abstract: In our intelligent society, water resources are being managed using vast amounts of hydrological data collected through telemetric devices. Recently, advanced data quality control technologies for data refinement based on hydrological observation history, such as big data and artificial intelligence, have been studied. However, these are impractical due to insufficient verification and implementation periods. In this study, a process to accurately identify missing and false-reading data was developed to efficiently validate hydrological data by combining various conventional validation methods. Here, false-reading data were reclassified into suspected and confirmed groups by combining the results of individual validation methods. Furthermore, an integrated quality control process that links data validation and reconstruction was developed. In particular, an iterative quality control feedback process was proposed to achieve highly reliable data quality, which was applied to precipitation and water level stations in the Daecheong Dam Basin, South Korea. The case study revealed that the proposed approach can improve the quality control procedure of hydrological database and possibly be implemented in practice.

Keywords: data reconstruction; data validation; hydrological data; quality control; smart water management

Citation: Jeong, G.; Yoo, D.-G.; Kim, T.-W.; Lee, J.-Y.; Noh, J.-W.; Kang, D. Integrated Quality Control Process for Hydrological Database: A Case Study of Daecheong Dam Basin in South Korea. *Water* **2021**, *13*, 2820. https://doi.org/10.3390/w13202820

Academic Editor: Aizhong Ye

Received: 19 August 2021
Accepted: 5 October 2021
Published: 11 October 2021

1. Introduction

In the water resources field, forecasting, planning, and management technologies have been actively developed based on vast data collected using intelligent technologies and stored in various databases. Hydrological data, such as precipitation and river water levels, are valuable and essential for flood and drought analysis. However, databases affect the reliability of analysis on water resources because they include various types of missing and false-reading data owing to errors in measurement equipment or data processing. To prevent this problem, the accuracy of measurement equipment has been improved, and quality management systems for the collected data have been developed. Data quality control is the process of validating missing data or false-reading data in an abnormal range for a collected dataset and reconstructing them into normal-range data.

As for research on such quality control, early researchers, such as Bedient and Cressman [1], Shuman [2], and Haug [3], have recognized the reliability problems with such data and attempted to manage weather observation data using computer systems. In particular, for research in the field of weather observation data, including temperature, precipitation, water level, and wind speed, the overall data quality control process has been

identified through technical reports published by the World Meteorological Organization (WMO) [4–7].

As for research on the validation of false-reading data, mathematical theories considering the inconsistency and nonhomogeneity of observation data were first adopted in the 20th century [8]. Various methods have been proposed, including a method that uses spatial consistency [9] and a study for establishing standardized rules, such as the high-low range and change limit [10]. Abbott [5] and Aguilar et al. [6] systematically established and shared methodologies and standards for data validation, leading to studies that reflect regional characteristics. In this instance, most data validation algorithms exhibit different levels of applicability depending on spatiotemporal factors; thus, the automation of the data validation field has been actively researched to derive and apply the most suitable validation technique. There exist some representative cases [11–13] establishing automation platforms that consider database management structures, such as on/offline and real-time/postanalysis quality control.

For many false-reading data validation methods, expected observation values are predicted first considering the continuity and consistency of data, and these are compared with the actual observed values. In other words, the expected data range is produced along with data validation, which naturally leads to research on the reconstruction of missing and false-reading data. March [14] first divided the data reconstruction field into two sets, one including data obtained from the same observation station as the error data and one including data obtained from other observation stations, and presented simple interpolation methods. Linacre [15] and Acock and Pachepsky [16] presented initial reconstruction methods using statistical values, such as mean and standard deviation, or nonlinear regression. These methods are widely used owing to their ease of application and constitute representative reconstruction techniques for databases obtained from the same observation station. In contrast, Willmott et al. [17], Xia et al. [18], and Teegavarapu and Chandramouli [19] introduced well-known spatial interpolation methods, such as arithmetic averaging and inverse distance weighing method. Subsequently, spatiotemporal interpolation [20,21] and machine learning-based reconstruction techniques [22–24], such as artificial neural networks (ANNs), have been developed to improve the reliability of reconstructed data.

HYMOS [25] and AQUARIUS Time Series [26] are representative software that provides quality control functions for hydrological databases using such data validation and reconstruction techniques. They include functions for determining and reconstructing outliers using statistical comparative analysis of time-series data and homogeneity analysis of spatial data. Despite the efforts mentioned above, however, there are still concerns over validating the false-reading data in the collected dataset and reconstructing them into user-defined values in terms of data reliability. Therefore, in the field of data quality control based on practical systems, decision making by data managers still represents the largest proportion. This indicates that quality control procedures must be developed to support data managers for effective decision making.

However, previous studies on data quality control often revealed the limited practical implementation. Improved validation algorithms such as machine learning techniques have been developed, but the algorithms still include drawbacks and have limited practical implementation due to their complexity and lack of theoretical background. Here, we propose an approach that utilizes and integrates well-known conventional algorithms for easy implementation in the field. The developed process reduces unnecessary data processing by classifying false data into suspected false and confirmed false datasets based on the combined results of the validation methods. In addition, an iterative feedback analysis to further improve the quality control procedure of hydrological databases is proposed.

In this study, for the effective quality control of precipitation and water level data, the observation data were classified into missing, suspected false-reading, confirmed false-reading, and normal data using various validation algorithms (validation process). In this instance, depending on the individual validation method, the data determined to

be false-readings were divided into suspected and confirmed false-reading data based on the combination of each validation type to support a more precise identification of false-readings. Next, the developed quality control process suggested reconstructed values through various reconstruction methods to replace missing and false-reading data and support data reconstruction of managers (reconstruction process). Finally, an integrated quality control process was proposed to provide the best data availability through an iterative feedback process considering the influence of repeated data validation and reconstruction. In this case study, the developed quality control process was applied to the hydrological data collected in the Daecheong Dam basin, South Korea, and the suitability of the process for each procedure was analyzed in detail.

The remainder of this paper is organized as follows. In Section 2.1, the overall structure and procedure of the proposed approaches are introduced. In Section 2.2, the missing and false-reading data validation methods are summarized. In Section 2.3, the reconstruction methods of error data are summarized, then the iterative quality control process is summarized in Section 2.4. The status of precipitation and water level stations in the target basin for the case study is summarized in Section 3.1. The results of applying the quality control process to the target basin are analyzed in Sections 3.2 and 3.3. In Sections 4 and 5, the discussions and conclusions are presented, respectively.

2. Methodology

2.1. Overview

Hydrological data quality control techniques in South Korea, which are the basis of this study, have been presented by major institutions such as the Korea Water Resources Corporation (K-water), Korea Institute of Civil Engineering and Building Technology (KICT), and Korea Meteorological Administration (KMA). They have presented historical or statistical normal observation ranges based on the climate environment of South Korea and validated outliers through temporal or spatial relationships based on a theoretical approach. For data reconstruction, they also use representative practical techniques, such as the mean value reconstruction, linear programming, exponential function method, and reciprocal distance squared (RDS) method, for overall quality control. In the quality control process developed in this study, the above data quality control methods were mainly divided into data validation and reconstruction steps, as shown in Figure 1.

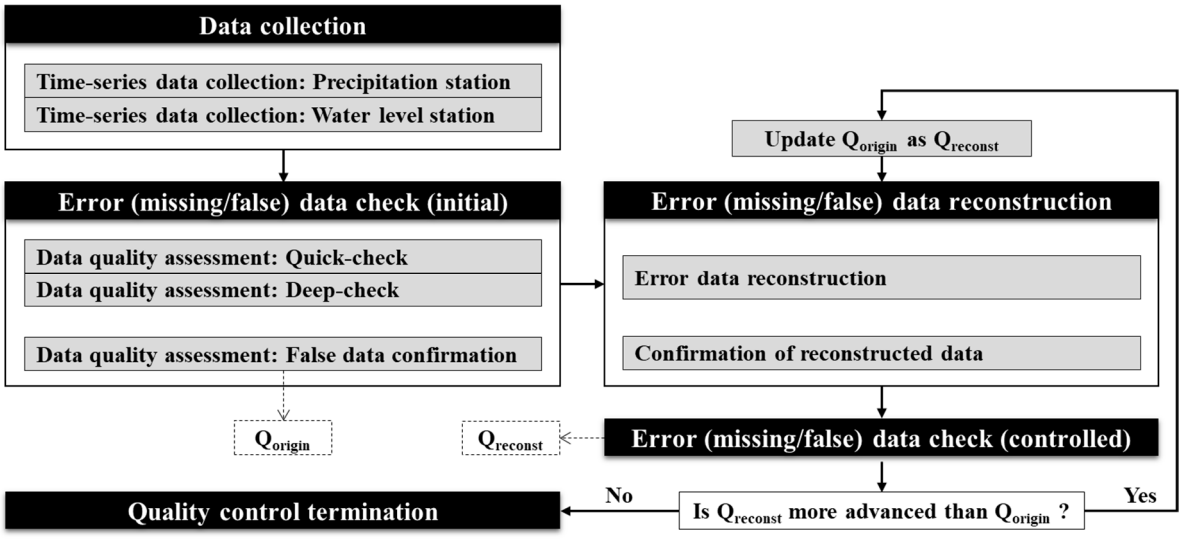

Figure 1. Scheme of proposed quality control process.

In the data validation step of this study, the observed data were validated through six types (missing, physical, duration, trend, statistical, and spatial false) of validation algorithms. As for the major characteristics of the developed process, the six algorithms can

be divided into quick-check and deep-check algorithms and applied gradually according to the intensity of the data quality control. In the data reconstruction step, the data classified as missing or false-reading data in the previous validation step were reconstructed into normal range through five types (selected-time, linear, spline, autoregressive integrated moving average (ARIMA), and spatial) of reconstruction algorithms. Different reconstructed values are proposed by each algorithm, and the final reconstructed values are determined by the data managers.

During the validation and reconstruction of error data, referring to preceding and subsequent time-series data or data from nearby observation stations is essential, and reliability may decrease if abnormal data is referred to. Therefore, data quality can be further improved if the already modified data are used again for quality control. This study presents an iterative validation and reconstruction process, in which quality control was repeatedly applied until the data quality was no longer improved.

In the case study, the developed integrated quality control process was applied to 32 precipitation stations and 24 water level stations located in the Daecheong Dam basin with a data collection period of more than five years. The databases contained precipitation and water level data with a 10-min interval.

2.2. Missing and False-Reading Data Validation

In this study, six (missing, physical, duration, trend, statistical, and spatial) validation methods were applied to validate the quality of observation data. In this instance, relatively simple validation methods may require relatively short-term data, but more rigorous validation methods require long-term data or data from nearby stations. The proposed quality control process comprises quick-check algorithms, which are based on missing, physical, duration, and trend methods, and deep-check algorithms for precise quality control based on statistical and spatial methods.

For the precipitation and water level data used in this study, the overall validation procedure can be applied in the same manner. However, the detailed application process and reference value may vary depending on the characteristics of the data. Figure 2 shows a conceptual diagram of the applied six validation methods. The details of validation methods are explained in the following sections.

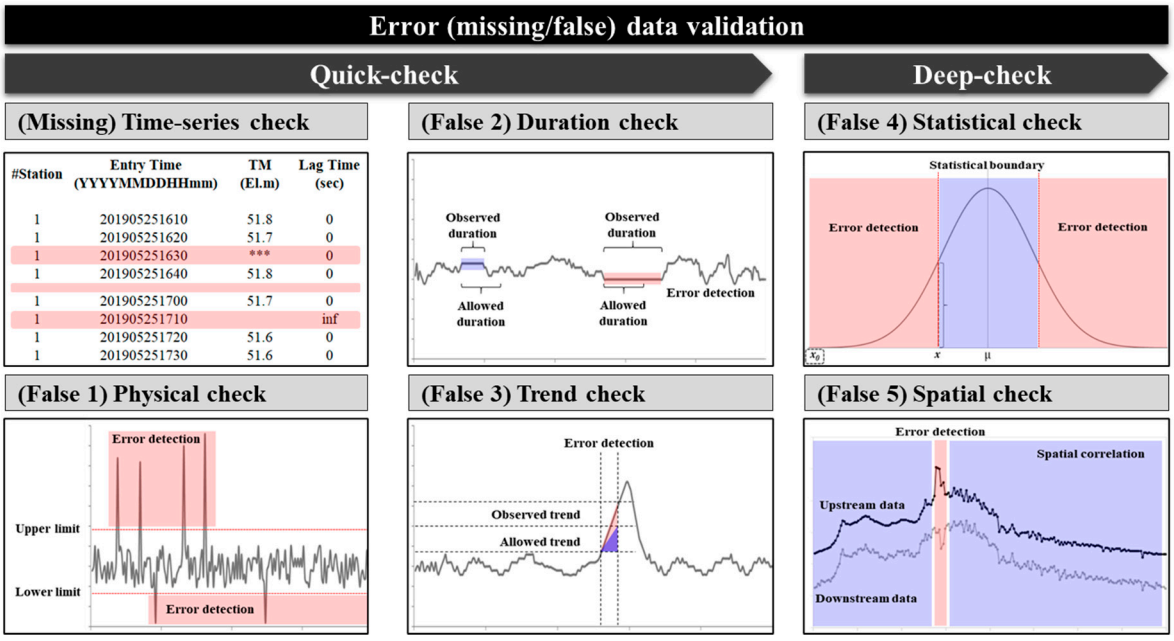

Figure 2. Conceptual diagram of the applied six data validation algorithms.

2.2.1. Data Validation Algorithms: Quick-Check

The quick-check algorithms are simple validation methods that can be applied in real-time or at a quasi-real-time because relatively small amount of data is required for validation. The inspection methods of the quick-check items presented here, and the application standards are as follows:

First, missing data represents the data not secured in time-series databases because measuring instruments failed or errors occurred in the data transmission and storage steps. Such missing data can be specifically divided into the following cases in which: (1) the corresponding time series have not been recorded, (2) data of the corresponding time series have been recorded as null, and (3) data of the corresponding time series have been recorded in a non-numeric form. In this study, missing data were validated for the databases constructed in the form of a CSV spread sheet using a time-series check algorithm that identifies cases corresponding to cases 1–3.

Second, false-reading data are data recorded differently from the actual values or data that are corrupted due to errors in the transmission and storage steps, even though numeric data were recorded in the database. Among the validation methods that correspond to the quick-check items, the physical check algorithm is practically difficult to implement considering the climate and locational characteristics of the area where the target station is located. In other words, cases that exceed the upper or lower boundaries of the physical observation range can be classified. In this study, false-reading data were validated, as shown in Equation (1). In this instance, the applied physical boundaries for precipitation and water level data are classified as shown in Table 1.

$$X_t = \begin{cases} FALSE & X_t < B_{P,\,min} \\ TRUE & B_{P,\,min} \le X_t \le B_{P,\,max} \\ FALSE & X_t > B_{P,\,max} \end{cases}, \quad (1)$$

where X_t is t-th observation data, $B_{P,min}$ is the minimum physical boundary of the data, and $B_{P,max}$ is the maximum physical boundary of data.

Table 1. Validation boundaries applied for physical check algorithm.

Data	Boundary	Value	Note
Precipitation	Lower $(B_{P,min})$	0 (mm per 10 min)	- Nonprecipitation
	Upper $(B_{P,max})$	100 (mm per 10 min)	- South Korea guideline (KMA) [27]
Water level	Lower $(B_{P,min})$	Lowest observation (m)	- South Korea guideline (KICT) [28]
	Upper $(B_{P,max})$	Highest observation (m)	

KMA [27] presented 100 mm per 10 min and 300 mm per hour as the maximum physical observation limits of precipitation data in Korea. Therefore, in this study, the maximum observation limit of 100 mm per 10 min and minimum of 0 mm per 10 min (nonprecipitation) were applied for the physical check of precipitation data. Meanwhile, KICT [28] suggested a method of applying the past maximum and minimum observation histories as the maximum and minimum physical observation limits of water level data. Therefore, in this study, the past maximum and minimum observation values of the target stations were collected and applied as observation limits of the station for the physical check of water level data.

Among the validation methods corresponding to the quick-check items, the duration check algorithm inspects whether the same observation value is recorded during an abnormally long time period. Therefore, false-reading data can be validated according to the duration boundaries for precipitation and water level data, as summarized in Table 2.

Table 2. Validation boundaries applied for duration check algorithm.

Data	Boundary	Value	Note
Precipitation	Upper ($B_{D,max}$)	60 (min)	- Except nonprecipitation period
Water level	Upper ($B_{D,max}$)	1440 (min)	- South Korea guideline (MLTM) [29]

Regarding the duration check of precipitation data, there is no specific duration limit for identical observations in Korea. In this study, when the same 10-minute precipitation data were repeated for more than one hour (i.e., more than six consecutive identical data), except for the nonprecipitation period, they were validated as false-reading data. In the case of water level data, the ministry of land, transport, and maritime affairs (MLTM) [29] in Korea classifies the same water level data repeated for more than 24 h as false-reading data. This guideline was adopted in this study, and 24 h was set as the duration limit of the 10-minute water level data (i.e., more than 144 consecutive identical data).

Among the validation methods that correspond to the quick-check items, the trend check algorithm inspects whether the data within a specific period abnormally fluctuates. The application of the increase and decrease limits may vary depending on the data. In this study, false-readings of precipitation and water level data were validated as shown in Equations (2a) and (2b), and the applied increase and decrease limits for the data are summarized in Table 3.

Table 3. Validation boundaries applied for trend check algorithm.

Data	Boundary	Value	Note
Precipitation	Upper ($B_{T,\,max}$)	$\overline{X} + 6\sigma$	- Shulski et al. [30] - Except change between nonprecipitation
Water level	Lower ($B_{T,min}$)	$\begin{cases} -1 \times \frac{dX}{dt}, if\ \frac{dX}{dt} > 0 \\ 3 \times \frac{dX}{dt}, if\ \frac{dX}{dt} < 0 \end{cases}$	- South Korea guideline (KICT) [28]
	Upper ($B_{T,\,max}$)	$\begin{cases} 3 \times \frac{dX}{dt}, if\ \frac{dX}{dt} > 0 \\ -1 \times \frac{dX}{dt}, if\ \frac{dX}{dt} < 0 \end{cases}$	

$$X_t = \begin{cases} TRUE & X_t - X_{t-1} \le B_{T,\,max} \\ FALSE & X_t - X_{t-1} > B_{T,\,max} \end{cases} \text{ for precipitation,} \tag{2a}$$

$$X_t = \begin{cases} FALSE & X_t - X_{t-1} < B_{T,\,min} \\ TRUE & B_{T,\,min} \le X_t - X_{t-1} \le B_{T,\,max} \\ FALSE & X_t - X_{t-1} > B_{T,\,max} \end{cases} \text{ for water level,} \tag{2b}$$

where $B_{T,\,min}$ is the decrease limit of observation data within unit time-step and $B_{T,\,max}$ is the increase limit of observation data within unit time-step.

Shulski et al. [30] suggested the fluctuation range of precipitation data within a unit time-step through a statistical approach. In this study, the increase limit of precipitation data was applied using the mean ($\overline{X}$) and standard deviation (σ) of precipitation variation for the same month of the previous year, as shown in Table 3. Meanwhile, KICT [28] suggested the water level fluctuation boundary depending on the slope ($\frac{dX}{dt}$) of the water level change during the last two hours: that is, if water level increased (i.e., $\frac{dX}{dt} > 0$), a range of -1 to 3 times the slope ($\frac{dX}{dt}$) is applied, while if it decreased (i.e., $\frac{dX}{dt} < 0$), a range of 3 to -1 times the slope ($\frac{dX}{dt}$) is applied as given in Table 3. This guideline was adopted in this study, and the increase and decrease limits of water level data were applied for cases in which the average slope over the last two hours was not zero, as shown in Table 3.

2.2.2. Data Validation Algorithms: Deep-Check

Deep-check algorithms for data validation are precision validation methods that can be applied at a non-real-time because they require a relatively large amount of reference data for validation. The inspection methods of the deep-check algorithms and the detailed application standards are as follows:

The statistical check algorithm inspects the data based on the statistical tolerance calculated using historical data, with the data period used to calculate the appropriate observation range being applied differently depending on the characteristics of hydrological data. In this study, data false-readings were validated, as shown in Equation (3) and the applied statistical boundaries for the data are summarized in Table 4. In the case of statistical inspection, statistical differences depending on the observation season of the data are reflected, and a larger amount of the collected historical data is more favorable for quality control. In this study, it was found that the applicability of the corresponding validation method can be applicable to a data period of at least five years.

$$X_t = \begin{cases} FALSE & X_t < B_{ST,\,min} \\ TRUE & B_{ST,\,min} \le X_t \le B_{ST,\,max} \\ FALSE & X_t > B_{ST,\,max} \end{cases}, \quad (3)$$

where $B_{ST,\,min}$ and $B_{ST,\,max}$ are the minimum and maximum statistical boundary of the data, respectively.

Table 4. Validation boundaries applied for statistical check algorithm.

Data	Boundary	Value	Note
Precipitation	Lower $(B_{ST,min})$	$\overline{X} - 3\sigma$	- Hubbard et al. [31]
	Upper $(B_{ST,max})$	$\overline{X} + 3\sigma$	- Except nonprecipitation data
Water level	Lower $(B_{ST,min})$	$\overline{X} - 3\sigma$	- Hubbard et al. [31]
	Upper $(B_{ST,max})$	$\overline{X} + 3\sigma$	- Separate flood and nonflood season

Hubbard et al. [31] suggested statistical boundaries considering the mean $(\overline{X})$ and standard deviation (σ) ranges of past precipitation observation data. In this study, nonprecipitation data were excluded, and statistical boundaries were calculated based on the precipitation data for the same month of all available years, as shown in Table 4. The same method was also applied for the water level data, but statistical boundaries were calculated by dividing the data corresponding to the flood season (June to September) and nonflood season (October to May) of all available years.

The spatial check algorithm inspects false-reading data through spatial boundaries, considering the consistency between the observation data near the target observation station. In this study, data false-readings were validated, as shown in Equation (4) and the applied statistical boundaries for the data are summarized in Table 5.

$$X_t = \begin{cases} FALSE & X_t < B_{SP,\,min} \\ TRUE & B_{SP,\,min} \le X_t \le B_{SP,\,max} \\ FALSE & X_t > B_{SP,\,max} \end{cases}, \quad (4)$$

where $B_{SP,\,min}$ and $B_{SP,\,max}$ is the minimum and maximum spatial boundary of the data, respectively.

Table 5. Validation boundaries applied for spatial check algorithm.

Data	Boundary	Value	Note
Precipitation	Lower $(B_{SP,min})$ Upper $(B_{SP,max})$	$0.2 \times X_{t,j}$ $2.8 \times X_{t,j}$	- South Korea guideline (K-water) [32]
Water level	Lower $(B_{SP,min})$ Upper $(B_{ST,max})$	$\overline{X} - 3\sigma$ $\overline{X} + 3\sigma$	- Hubbard et al. [31]

K-water [32] in South Korea suggested a range corresponding to 0.2–2.8 times the predicted value through the RDS method as the spatial boundary of precipitation data. Here, the RDS method is a reconstruction method based on the concept that the spatial correlation of precipitation is inversely proportional to the physical distance, and the precipitation value reconstructed using the RDS method can be calculated based on data from two or more observation stations located near the target station, as shown in Equation (5).

$$X_{t,j} = \frac{\sum_{i=1}^{m} \frac{X_{t,i}}{d_i^2}}{\sum_{i=1}^{m} \frac{1}{d_i^2}}, \quad (5)$$

where $X_{t,j}$ is the precipitation value of the target station (j) at time t reconstructed using the RDS method, $X_{t,i}$ is the observation data of the i-th station near the target station at time t, d_i is the distance between the target station (j) and the nearby i-th station, and m is the number of nearby stations used for the RDS method.

In the case of water level data, validation methods considering the spatial association between the upstream and downstream stations of a river have been presented through various studies. In this study, the method of Hubbard et al. [31], which was used as a statistical validation method, was utilized, and spatial boundaries were calculated based on the mean $(\overline{X})$ and standard deviation (σ) of water level differences between upstream and downstream stations during the past week, as shown in Table 5.

2.2.3. Distinction between Suspected False-Reading and Confirmed False Data

It is likely that a normal data is classified as false data after conducting the aforementioned validation checks. In particular, the false-reading rate of the database increases as more types of validation methods are applied. To improve the reliability of the quality control procedure, efforts to accurately identify false-reading data are required, while various validation methods are applied. In this study, the data determined to be false-readings by each validation method were first classified as suspected false data, then a procedure for determining the data as confirmed false data according to the type of validation method was proposed. In other words, in the proposed quality control procedure, the entire dataset was classified into normal, missing, suspected false, and confirmed false data. Error data, such as missing and confirmed false data, were then reconstructed, and the suspected false data were treated in the same manner as normal data but classified into a separate category.

False data were confirmed by considering the characteristics of each validation method, as shown in Figure 3. First, continuity, which is the most important observational characteristic of hydrological data, was considered in this study; thus, the data classified as false-reading by trend-check among the quick-check methods were immediately classified as confirmed false data. Next, the data classified as false-reading by physical check and duration check among the quick-check methods were first tagged as suspected false data and were further determined as confirmed false data depending on the deep-check results. In other words, among the data classified as suspected false data by quick-check algorithms, those that were also determined to be suspected false data by the deep-check algorithms were finally classified as confirmed false data.

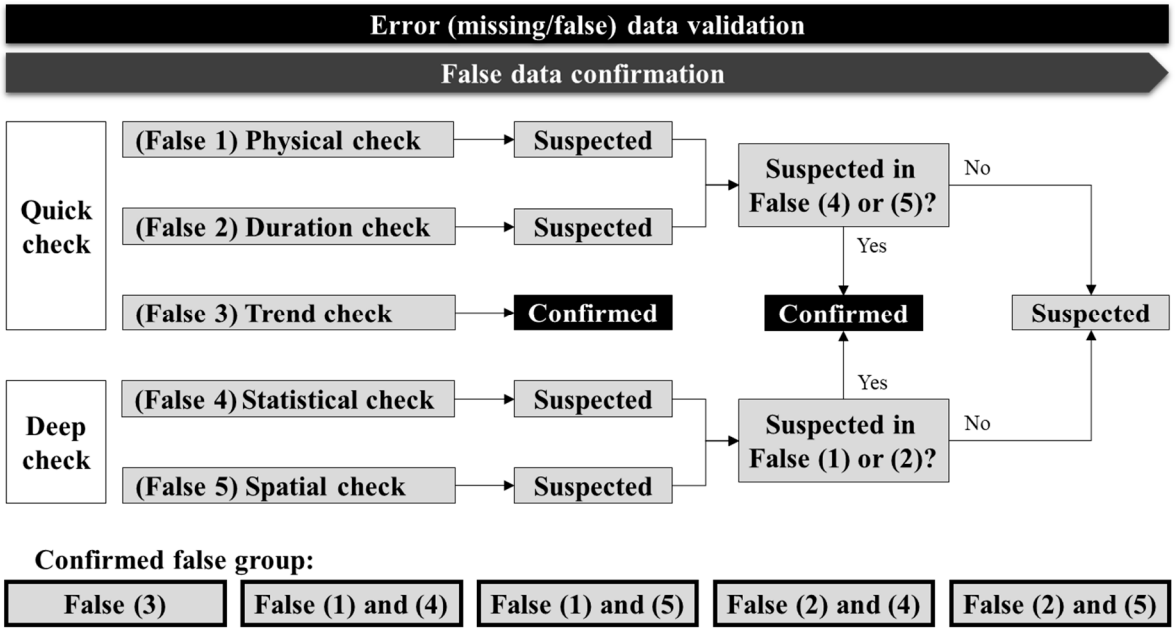

Figure 3. Process for suspected and confirmed false data categorization.

2.3. Missing and False-Reading Data Reconstruction

The proposed quality control process includes a procedure for reconstructing the data classified as missing or confirmed false-reading using five types (selected-time, linear, spline, ARIMA, and spatial) of reconstruction methods. In this instance, all the reconstruction methods, except for the spatial reconstruction method, use only data from the target station, and can be applied regardless of the data type (precipitation and water level) in the same manner. In case of the spatial reconstruction method, however, completely different methods are applied depending on the data type. In other words, the reconstruction methods used in this study can be summarized as six methods in five types, as shown in Figure 4, and the details of each reconstruction method are summarized as follows.

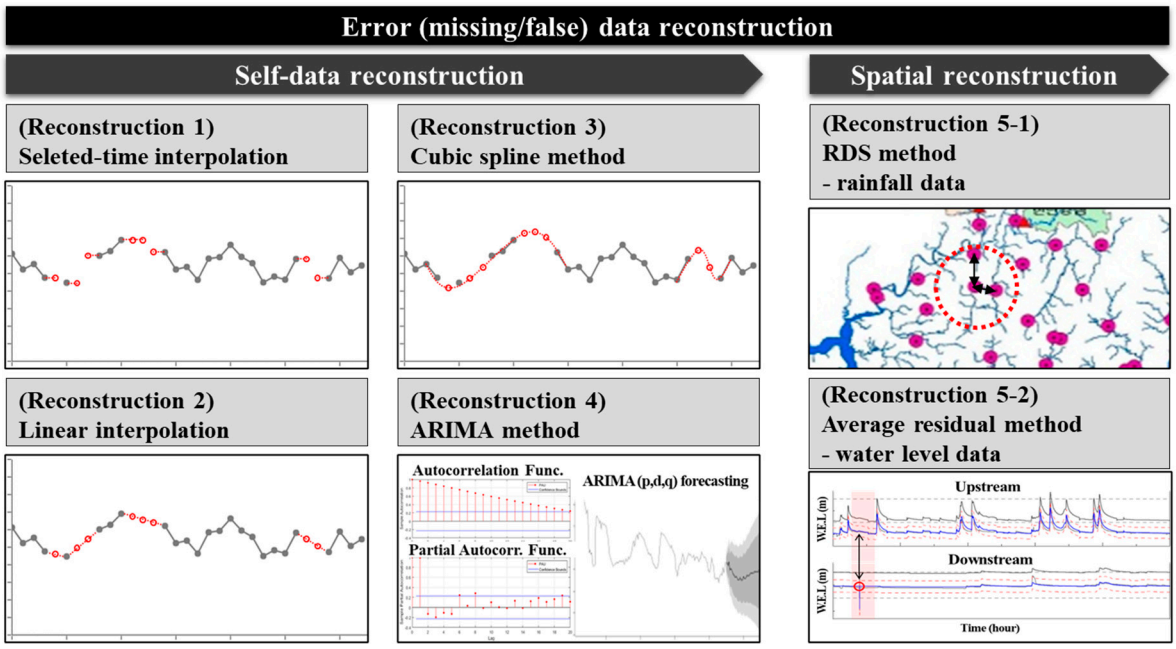

Figure 4. Conceptual diagram of the applied six data reconstruction methods.

The selected-time interpolation, which is the simplest method, replaces the error data with preceding or subsequent data directly. In general, alternative data for selected-time interpolation can be manually selected by the data manager. However, in this study, the nearest data, which are classified as normal, from the issued error data were selected for automated data reconstruction.

The linear interpolation method reconstructs data at a specific time under the assumption that data linearly increases or decreases, as shown in Equation (6). This method is easy to apply and can reflect the continuity of data in a relatively short time window. That is, the corresponding error data are reconstructed by linearly connecting the data before and after the error data. When missing or false-reading data occurred continuously during a certain period, continuous reconstruction was performed based on the preceding and subsequent data of the corresponding period so that the entire target period could have a linear relationship.

$$X_t = X_{t_0} + \frac{X_{t_1} - X_{t_0}}{t_1 - t_0}(t - t_0), \tag{6}$$

where t is the observation time point for the missing and false-reading data to be reconstructed; t_0 and t_1 are the time points immediately before and after the error observation, respectively; and X_{t_0} and X_{t_1} are the observed values at time t_0 and t_1, respectively.

The spline method, similar to the linear interpolation, is a reconstruction method based on the assumption that the observed tendency of the data follows a mathematical pattern. Although only instantaneous continuity with preceding and subsequent single data was considered in the linear interpolation method, a longer continuity pattern was considered in the spline method by referring to the data obtained one day before and after the error observation. When the error data were observed continuously, the spline method could reconstruct the entire period simultaneously.

The ARIMA reconstruction method predicts and reconstructs the data of the target time series. The ARIMA model considers cointegration in addition to the ARMA model, which combines the AR (autoregressive) model that uses the past data pattern and the MA (moving-average) model that utilizes the average of the past data. This can be expressed as an ARIMA (p, d, q) model, where p, d, and q denote the order of the AR model, degree of differencing for securing data stationarity, and order of the MA model, respectively. In this study, the reconstructed values were predicted by referring to the data obtained one week before the error observation. Here, the ARIMA parameters can be directly input by the user or derived by the ARIMA module for a specific observation period.

Finally, in the spatial reconstruction, different methods were applied to the precipitation and water level data. For the precipitation data reconstruction, the calculated RDS value obtained from Equation (5) was applied. For the water level reconstruction, the error data were reconstructed by adding the average difference in water levels between the target and nearby stations estimated from the data obtained during the previous week, as shown in Equation (7).

$$X_{t,j} = X_{t,i} + \frac{\sum_{k=1}^{n}\left(X_{t-k,j} - X_{t-k,i}\right)}{n}, \tag{7}$$

where $X_{t,j}$ is the reconstructed water level of the target station (j) at time t, $X_{t,i}$ is the observed water level of the nearby station (i) at time t, $X_{t-k,j}$ is the observed value of the target station (j) at time $t-k$, $X_{t-k,i}$ is the observed value of the nearby station (i) at time $t-k$, and n is the maximum period of the past data referenced for data reconstruction (here, one week).

2.4. Iterative Quality Control Feedback Process

Most validation and reconstruction methods require the process of referring to the data of the preceding and subsequent periods from the error observation. Thus, the data validation and reconstruction results may vary as the overall data pool is improved owing

to the quality control process. For example, identification of the false-reading data may vary as reconstruction data is newly included, and the proposed reconstructed value may still fall within the range of the false value. In this study, such data dependency of quality control was considered, and an iterative feedback process was proposed. In the proposed feedback process, quality control was applied iteratively until the error rate is no longer improved by comparing the data availabilities before and after the quality control process as shown in Figure 5. In specific, the error data were identified and reconstructed by data validation and reconstruction methods (Step 1 and 2, respectively). As the data manager obtained five sets of reconstructed values, determination of the reconstruction method was required. Here, an iterative quality control process is used to apply data validation methods to reconstructed datasets and reevaluate their quality (Step 3). Based on this, the data manager could compare and determine the best reconstruction method for each iteration of quality control (Step 4). However, the reconstructed database can be further improved owing to the improved data pool. The iterative quality control process compares the data quality before and after reconstruction, and then repeats Steps 2–4 until no more improvement is identified.

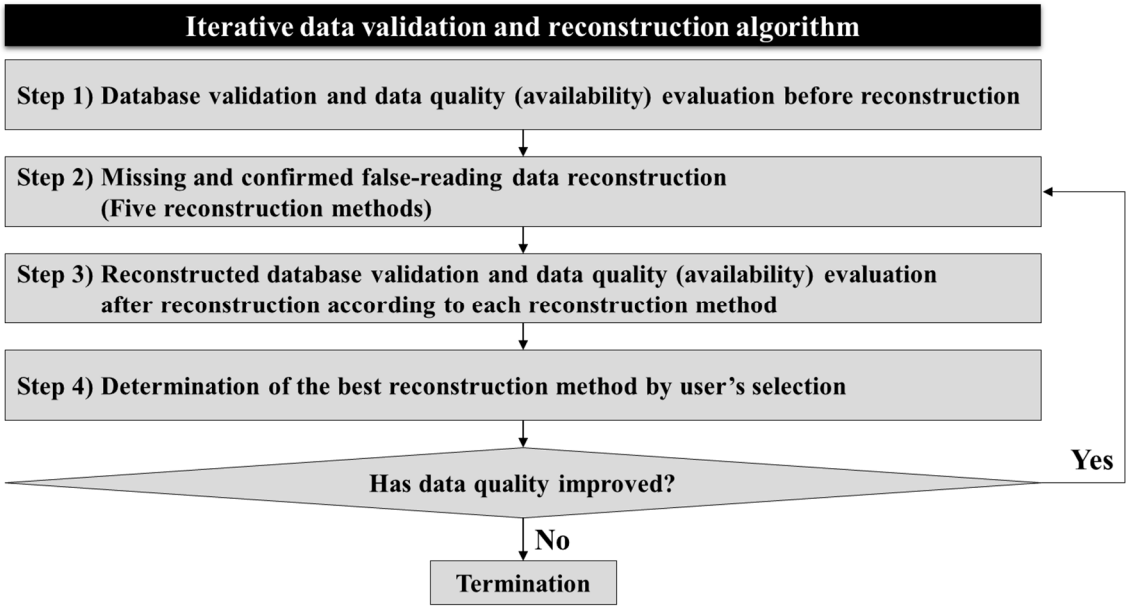

Figure 5. Process for iterative quality control algorithm.

For illustration of the proposed iterative quality control process, Figure 6a shows the time series, including the false-reading data in the validation step, and Figure 6b shows the proposed reconstructed values by each reconstruction method. In this step, the data manager can choose the most suitable values among the suggested reconstructed data, as shown in Figure 6c. After fixing the error data, as shown in Figure 6d, the reconstructed data were validated again to examine whether the reconstructed data falls within the acceptable range. As seen in the figure, one value is still identified as an error data, then one more iteration of data reconstruction and validation is conducted. The proposed iterative feedback process enhances the data availability by repeating quality control process until the validated error rate is no longer improved.

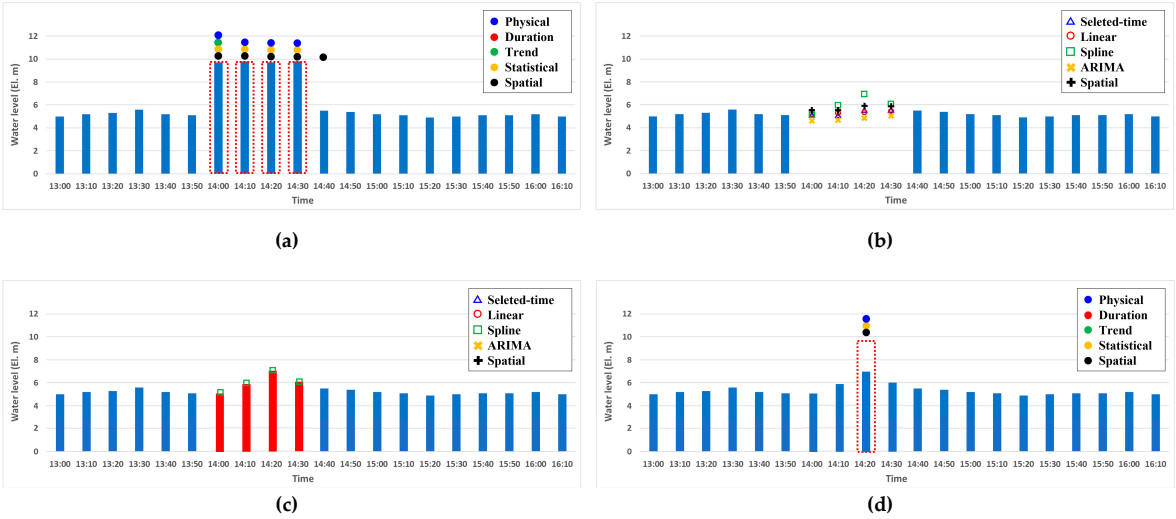

Figure 6. Illustration of iterative quality control process. (**a**) Data validation of original dataset; (**b**) Reconstructed values by five reconstruction methods; (**c**) Data reconstruction chosen by data manager; (**d**) Data validation of reconstructed dataset.

3. Application Study

3.1. Study Area

The proposed quality control process was applied to the Daecheong Dam Basin in the Geum River, South Korea, which contains 32 precipitation stations and 24 water level stations, as shown in Figure 7. The target examination period was one year, from 00:10 on 1 January 2016, to 24:00 on 31 December 2016; a total of 52,704 data points was collected at 10-min intervals.

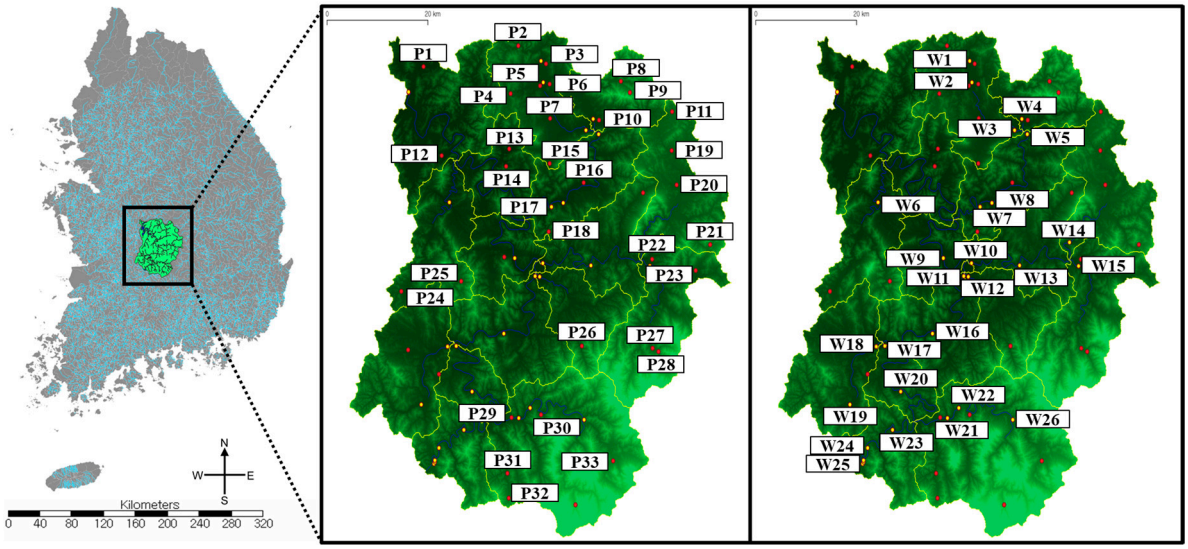

Figure 7. Study area (Daecheong Dam basin, South Korea): precipitation stations (**left**) and water level stations (**right**).

3.2. Application Results of Validation Process

3.2.1. Validation of Precipitation Data

First, a time-series check was applied to the 32 precipitation stations for validating missing data. The precipitation stations can be largely divided into first group for 19 precipitation stations representing a number of missing data points and second group for 13 stations showing few missing data points. In the first group, the P-23 station showed the largest number of missing data (1209 data points), and 560 missing data points were identified at each of 16 precipitation stations (P-2, 4, 5, 7, 8, 10, 11, 13, 15, 16, 18, 19, 20, 21,

26, and 30) during the same time periods. In other words, the corresponding missing type was understood to be missing data generated from database systems, such as transmission or processing, rather than equipment errors. Moreover, 842 missing data points occurred at the P-6 precipitation station, 209 missing data points at the P-28 precipitation station. On the contrary, only two missing data points were identified for the remaining 13 precipitation stations. The data missing rate in the overall precipitation databases was found to be 0.67% on average for 32 stations, with the P-23 station exhibiting the highest rate (2.29%).

Next, quick-check (physical, duration, and trend) and deep-check (statistical and spatial) were applied to the precipitation database for the validation of false-reading data. The result of false-reading validation for each check exhibited patterns, as shown in Figure 8. As single false data can be classified as false-reading multiple times by the various check methods, the total number of stacked bars in Figure 8 does not represent the total amount of suspected false-readings. For example, the total number of datasets classified as false-readings for the P-1 station was approximately 80, but the actual number of suspected false datasets could be 80 or less because of the overlapped counts by multiple checks.

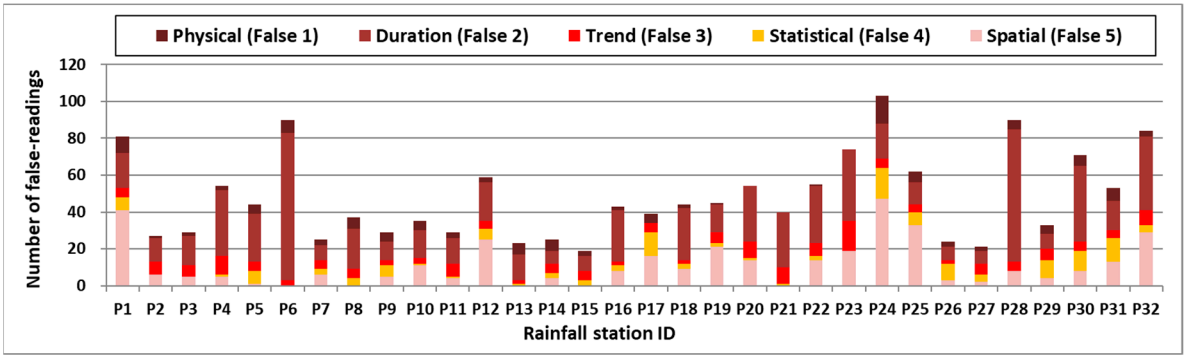

Figure 8. False data validation results of 32 precipitation stations.

When precipitation data were validated by individual false type, the data identified as false-reading by duration check (False 2) or spatial check (False 5) represented the highest proportion overall. Most of them were eventually found to correspond to suspected false data, as they were classified as normal data in the results of other check algorithms. In contrast, the data validated as false by physical check (False 1) exhibited a high level of precipitation that exceeded 100 mm per 10 min. As most of them were also classified as false-reading in the statistical check (False 4), it was found that the case of determining the confirmed false-reading data based on the combination of False 1 and False 4 types represented the highest proportion. In the case of the data immediately classified as confirmed false by the trend check (False 3), relatively high values were observed between preceding and subsequent low values in most cases. Moreover, most of them were in agreement with the results of the physical check (False 1), indicating the high probability of the actual false data.

When suspected/confirmed false data were classified based on the above precipitation data validation results, approximately 0.08% of the data from 32 stations were initially classified as suspected false data on average. However, only 0.01% of the data (i.e., 15% of the suspected false data) were finally classified as confirmed false data, and the remaining 0.07% were still regarded as the suspected false data. The final error rate of the precipitation data in 32 stations under first application of the validation process was found to be approximately 0.68% (i.e., missing data 0.67% and confirmed false-reading data 0.01% combined) on average, with the P-23 station exhibiting the highest rate of 2.32%. Table 6 summarizes the data error rates for each station.

Table 6. Data availability analysis in 32 precipitation stations.

Precipitation Station	Missing Rate (%)	Confirmed False Rate (%)	Suspected False Rate (%)	Total Error Rate (%)	Normal Data Availability (%)
P-1	0.004	0.013	0.114	0.017	99.98
P-2	1.063	0.013	0.036	1.076	98.92
P-3	0.004	0.011	0.044	0.015	99.98
P-4	1.063	0.021	0.080	1.083	98.92
P-5	1.063	0.015	0.053	1.078	98.92
P-6	1.598	0.006	0.161	1.603	98.40
P-7	1.063	0.015	0.027	1.078	98.92
P-8	1.063	0.015	0.046	1.078	98.92
P-9	0.004	0.009	0.032	0.013	99.99
P-10	1.063	0.008	0.055	1.070	98.93
P-11	1.063	0.013	0.038	1.076	98.92
P-12	0.004	0.011	0.091	0.015	99.98
P-13	1.063	0.004	0.036	1.066	98.93
P-14	0.004	0.009	0.030	0.013	99.99
P-15	1.063	0.015	0.015	1.078	98.92
P-16	1.063	0.006	0.070	1.068	98.93
P-17	0.004	0.009	0.046	0.013	99.99
P-18	1.063	0.006	0.072	1.068	98.93
P-19	1.063	0.011	0.068	1.074	98.93
P-20	1.063	0.017	0.085	1.080	98.92
P-21	1.063	0.017	0.059	1.080	98.92
P-22	0.004	0.013	0.089	0.017	99.98
P-23	2.294	0.030	0.108	2.324	97.68
P-24	0.004	0.019	0.142	0.023	99.98
P-25	0.004	0.009	0.093	0.013	99.99
P-26	1.063	0.006	0.030	1.068	98.93
P-27	0.004	0.011	0.027	0.015	99.98
P-28	0.397	0.011	0.152	0.408	99.59
P-29	0.004	0.011	0.038	0.015	99.98
P-30	1.063	0.015	0.101	1.078	98.92
P-31	0.004	0.009	0.072	0.013	99.99
P-32	0.004	0.015	0.133	0.019	99.98

3.2.2. Validation of Water Level Data

In the time-series check (i.e., validation of missing data) for the 24 water level stations, the stations were divided into two groups with large and small amounts of missing data points, respectively. In the first group of 19 water level stations, W-18 station exhibited the largest number of missing data points (386 data points), and 352 missing data points were identified at each of 18 water level stations (W-1, 2, 3, 4, 5, 6, 8, 10, 11, 12, 13, 14, 16, 17, 20, 21, 22, and 24) during the same missing time periods. However, only one missing data point was identified at each of the four water level stations (W-7, 9, 15, 19). The W-23 water level station exhibited no missing data, as all of the time-series data were recorded as numeric data. The data missing rate in the overall water level databases was found to be 0.53% on average for 24 stations, with the W-18 station exhibiting the highest rate of 0.73%.

Next, the validation results of false-reading data based on the quick-check and deep-check algorithms are shown in Figure 9. As shown in the bar chart, the false-reading data by the duration check (False 2) and statistical check (False 4) represented the highest proportion. In most cases, false-reading data were suspected by the duration check (False 2) and primarily corresponded to the case where low water levels were maintained in long periods during the dry season. Except for W-1, 3, and 18 stations, most stations exhibited a statistically normal range, thereby remaining in the suspected false data group. For the W-1 and W-3 stations, however, a large amount of data was classified as confirmed false data because statistically low water level values were maintained during a certain data period. The number of data classified as false-reading by the physical check (False 1) and the trend check (False 3) was found to be very small, and most of them were in agreement with results of statistical (False 4) and spatial checks (False 5), indicating high probability of the actual false data.

W1 and W3 show relatively high confirmed false-reading rate compared to those of other stations, which were identified by combining False 2 (duration check) and False 4 (statistical check). It is observed that several stations (W-4, 7, 8, 10, 12, 14, 15, 17, and 20) show high suspected false rate (higher than 30%) by False 2 (Duration check), but they are eventually classified as "suspected" because most of them are normal in other validation checks. These results indicate that the proposed validation approach (i.e., separation of suspected and confirmed false data) can be practically implemented.

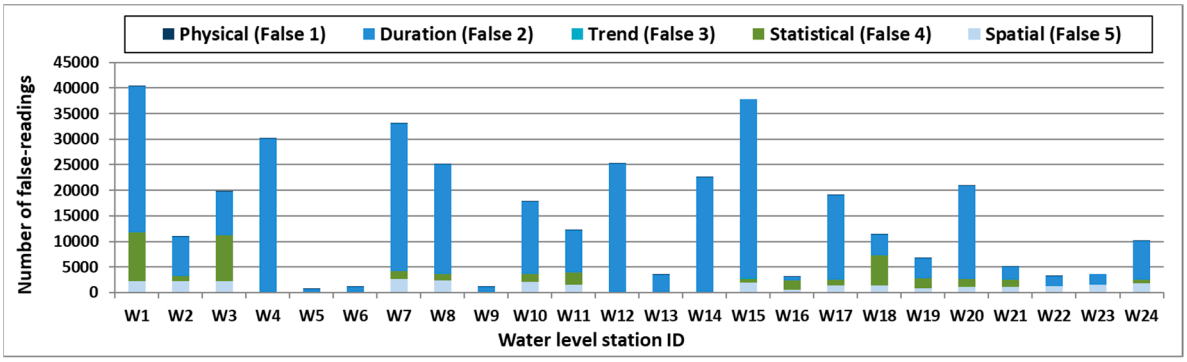

Figure 9. False data validation results of 24 water level stations.

From the suspected/confirmed false data classification based on the above validation results, approximately 27% of the data from 24 stations were initially classified as suspected false data on average. Finally, only 1.13% of the data (i.e., 4% of the suspected false data) was classified as confirmed false data, and the remaining 26% were finally determined to be suspected false data. The final error rate of the water level data in the 24 stations was found to be approximately 1.66% (i.e., missing data 0.53% and confirmed false-reading data 1.13% combined) on average, with the W-1 station exhibit the highest rate of 11.55%. Table 7 summarizes the data error rates for each station.

Table 7. Data availability analysis in 24 water level stations.

Water Level Station	Missing Rate (%)	Confirmed False Rate (%)	Suspected False Rate (%)	Total Error Rate (%)	Normal Data Availability (%)
W-1	0.668	10.886	54.116	11.553	88.45
W-2	0.668	0.670	18.785	1.338	98.66
W-3	0.668	7.876	20.376	8.544	91.46
W-4	0.668	0.013	57.094	0.681	99.32
W-5	0.668	0.006	1.345	0.674	99.33
W-6	0.668	0.000	2.171	0.668	99.33
W-7	0.002	2.736	56.858	2.738	97.26
W-8	0.668	0.865	45.402	1.533	98.47
W-9	0.002	0.017	1.985	0.019	99.98
W-10	0.668	0.609	32.404	1.277	98.72
W-11	0.668	0.393	21.894	1.061	98.94
W-12	0.668	0.006	47.808	0.674	99.33
W-13	0.668	0.008	6.666	0.675	99.32
W-14	0.668	0.013	42.776	0.681	99.32
W-15	0.002	1.617	68.267	1.619	98.38
W-16	0.668	0.083	5.561	0.751	99.25
W-17	0.668	0.412	35.146	1.080	98.92
W-18	0.732	0.121	20.752	0.854	99.15
W-19	0.002	0.072	12.274	0.074	99.93
W-20	0.668	0.290	38.757	0.958	99.04
W-21	0.668	0.180	8.819	0.848	99.15
W-22	0.668	0.063	5.897	0.731	99.27
W-23	0.000	0.053	6.749	0.053	99.95
W-24	0.668	0.131	18.401	0.799	99.20

3.3. Application Results of Reconstruction Process

3.3.1. Reconstruction of Precipitation Data

Table 8 shows the error rates after applying the five reconstruction methods for the missing and false-reading precipitation data classified above. The error rates were calculated individually by applying each reconstruction method to compare the efficiency of each reconstruction method. In Table 8, the marked values indicate the minimum error rate after reconstruction, which shows the best error improvement at each station. For 32 precipitation stations, the selected time, linear, spline, ARIMA, and spatial methods were selected as the optimal reconstruction method for 6, 19, 12, 9, and 11 times, respectively. Thus, in general, the linear interpolation method was found to be the most effective for reconstructing the precipitation error data in this case study.

Table 8. Error rate comparison of five reconstruction methods in 32 precipitation stations.

Precipitation Station	Error Rate of Original Data (%)	Error Rate after Single Iteration of Reconstruction (%)				
		Selected-Time	Linear	Spline	ARIMA	Spatial
P-1	0.017	**0.011**	**0.011**	**0.011**	**0.011**	**0.011**
P-2	1.076	0.011	0.011	0.015	0.013	**0.008**
P-3	0.015	0.002	**0.000**	**0.000**	0.002	**0.000**
P-4	1.083	**0.008**	0.011	0.013	0.011	0.009
P-5	1.078	0.011	**0.009**	0.013	0.011	0.025
P-6	1.603	0.017	0.015	0.015	0.017	**0.013**
P-7	1.078	0.013	0.021	0.021	**0.009**	0.017
P-8	1.078	0.017	**0.011**	0.013	0.017	0.019
P-9	0.013	0.015	**0.011**	**0.011**	0.013	0.015
P-10	1.070	0.015	**0.011**	0.015	0.015	0.017
P-11	1.076	0.006	0.011	0.011	**0.004**	0.006
P-12	0.015	0.019	**0.017**	**0.017**	0.019	0.021
P-13	1.066	0.013	0.015	0.015	0.030	**0.011**
P-14	0.013	0.013	0.009	0.011	0.011	**0.008**
P-15	1.078	0.013	0.021	0.021	**0.009**	0.017
P-16	1.068	0.015	0.015	0.015	0.021	**0.004**
P-17	0.013	0.008	**0.006**	**0.006**	**0.006**	0.008
P-18	1.068	**0.015**	**0.015**	**0.015**	0.021	0.021
P-19	1.074	0.021	**0.013**	0.015	0.017	0.021
P-20	1.080	0.015	**0.011**	**0.011**	0.034	0.019
P-21	1.080	0.015	**0.011**	**0.011**	0.034	0.019
P-22	0.017	0.009	0.017	0.019	**0.008**	**0.008**
P-23	2.324	0.028	0.034	0.028	0.027	**0.023**
P-24	0.023	0.009	0.009	0.011	**0.008**	**0.008**
P-25	0.013	**0.009**	**0.009**	**0.009**	**0.009**	0.011
P-26	1.068	**0.019**	**0.019**	**0.019**	0.021	0.027
P-27	0.015	0.004	**0.000**	**0.000**	0.004	0.006
P-28	0.408	0.019	0.017	0.017	0.017	**0.013**
P-29	0.015	0.013	**0.009**	0.011	0.011	0.013
P-30	1.078	0.019	**0.015**	0.021	0.047	0.019
P-31	0.013	**0.011**	**0.011**	**0.011**	**0.011**	0.017
P-32	0.019	0.011	**0.009**	0.011	0.013	0.017

Note: Marked value denotes the lowest error rate at each station after single iteration of reconstruction.

In the proposed quality control process, the iterative feedback process can be applied, as described in Section 2.4. For demonstration of the proposed iterative feedback process, a representative station was selected from the 32 precipitation stations, and iterative reconstruction and validation were simulated. The P-30 was selected as a representative station because it included all types of missing and false-reading data.

Figure 10 shows the error rate improvement through the iterative application of the quality control process for the P-30 station. The original data validation results revealed an error rate of 1.078%, which was improved to 0.015–0.047% after first application of the five reconstruction methods. After iteration 1, the linear interpolation method exhibited the best reconstruction results; thus, the database reconstructed by linear interpolation with 0.015% of the error rate was selected for the next iteration. After iteration 2, the linear interpolation, spline, and ARIMA reconstruction methods showed the best reconstruction results (i.e., an error rate of 0.010%). When the same data validation error rate was obtained by multiple reconstruction methods, the data manager can choose the favorable reconstruction method. In this case study, the reconstructed data by the cubic spline method was selected for the next iteration. Finally, after iteration 3, no more missing and false-reading data existed in the database of P-30 station by both of linear interpolation and cubic spline methods. The data manager can choose the final results either by the linear interpolation or cubic spline methods in this simulation.

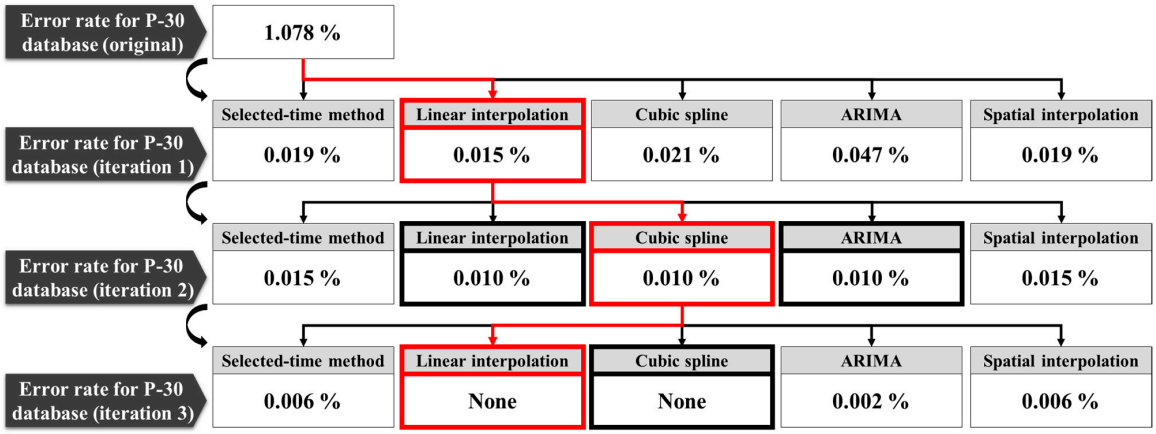

Figure 10. Improvement of data quality through the iterative quality control process for P-30 precipitation data.

3.3.2. Reconstruction of Water Level Data

Table 9 shows the error rates after applying the reconstruction methods for missing and false-reading water level data. For 24 water level stations, the selected time, linear, spline, ARIMA, and spatial methods were selected as the best method for three, six, two, fifteen, and one times, respectively. Thus, in this case study, the ARIMA reconstruction method was found to be the most effective for reconstructing the error data and attributed to the water level data being highly related to the past water level state.

Table 9. Error rate comparison of five reconstruction methods in 24 water level stations.

Water Level Station	Error Rate of Original Data (%)	Error Rate after Single Iteration of Reconstruction (%)				
		Selected-Time	Linear	Spline	ARIMA	Spatial
W-1	11.553	11.627	4.690	3.152	**0.049**	1.376
W-2	1.338	0.729	0.617	0.379	**0.013**	0.888
W-3	8.544	8.184	2.577	**0.156**	0.645	1.328
W-4	0.681	0.008	**0.004**	0.021	0.009	0.254
W-5	0.674	**0.004**	0.006	0.034	0.009	0.247
W-6	0.668	0.006	**0.000**	0.008	0.009	0.241
W-7	2.738	2.457	2.131	1.856	**0.588**	1.560
W-8	1.533	0.780	0.729	0.757	0.137	**0.700**
W-9	0.019	0.011	**0.004**	0.008	0.011	0.019
W-10	1.277	0.527	0.254	0.326	**0.011**	0.125
W-11	1.061	0.250	0.250	0.429	**0.004**	0.332
W-12	0.674	**0.006**	**0.006**	0.011	**0.006**	0.222
W-13	0.675	0.006	**0.000**	0.019	0.006	0.249
W-14	0.681	**0.008**	**0.008**	0.013	0.009	0.254
W-15	1.619	1.620	1.285	1.104	**0.087**	1.249
W-16	0.751	0.047	0.030	0.030	**0.004**	0.013
W-17	1.080	0.575	0.577	0.416	**0.013**	0.768
W-18	0.854	0.104	0.102	0.078	**0.004**	0.787
W-19	0.074	0.013	0.013	0.049	**0.000**	0.017
W-20	0.958	0.374	0.334	0.296	**0.004**	0.677
W-21	0.848	0.154	0.156	0.176	**0.009**	0.679
W-22	0.731	0.055	0.044	0.032	**0.011**	0.717
W-23	0.053	0.044	0.042	**0.011**	0.013	0.042
W-24	0.799	0.321	0.300	0.116	**0.025**	0.721

Note: Marked value denotes the lowest error rate at each station after single iteration of reconstruction.

For demonstration of the iterative reconstruction simulation, the W-10 station was selected as a representative station among the 24 water level stations. Figure 11 shows the error rate improvement through the iterative application of the quality control process for W-10. The raw data validation results showed an error rate of 1.277%, which was improved to 0.011–0.527% after first application of the five reconstruction methods. After iteration 1, the ARIMA method exhibited the best reconstruction results, then corresponding reconstructed database was selected for next iteration. After iteration 2, the database reconstructed by the cubic spline showed the best data quality and was thus selected for the next step. Finally, after iteration 3, no missing and false-reading data existed in the database reconstructed by both ARIMA and spatial interpolation methods. The data

manager can choose the final results either by the ARIMA or spatial interpolation methods in this simulation.

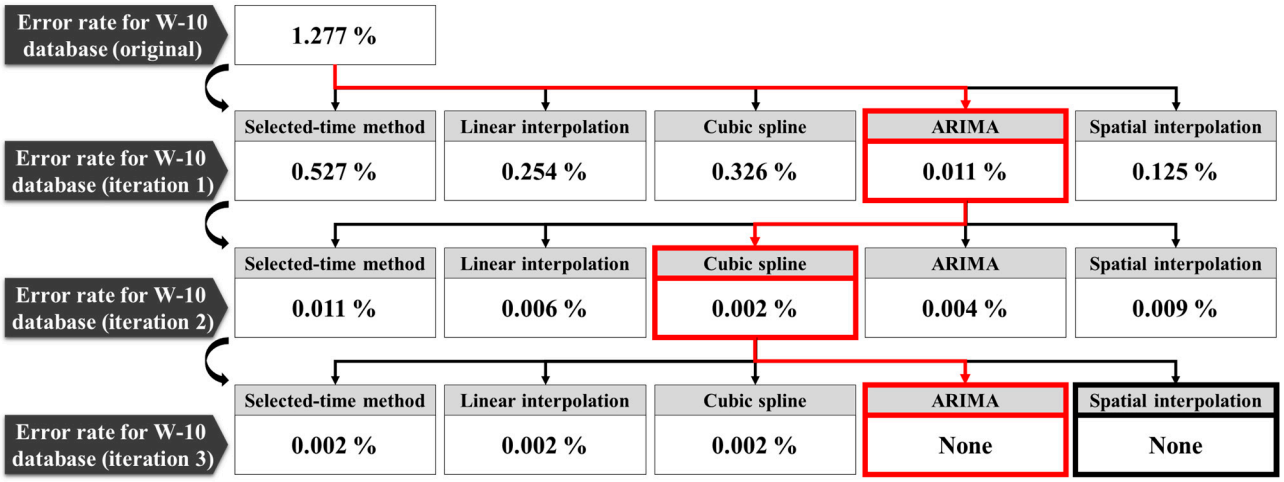

Figure 11. Improvement of data quality through the iterative quality control process for W-10 water level data.

4. Discussion

The developed quality control process reduces unnecessary data processing by classifying false data into suspected false and confirmed false datasets based on the combined results of the validation methods. In particular, the proposed reconstruction process was designed to support data managers more reasonably by providing various types of reconstructed values. In addition, the false data quality can be enhanced through the proposed iterative feedback process of validating and reconstructing the error data. As for related research, validation standards for some validation methods will be improved, and various validation and reconstruction methods can be added and combined in the future to further improve the quality control of hydrological databases.

When combining the validation results and classifying the false data, universal classification standards are required for practical implementation. In the proposed iterative process, the best reconstruction values in each iteration are selected based on the error rate. That is, the algorithm that results in the lowest error rate is prioritized and selected for reconstruction. This may cause inconsistency in the reconstructed data set and further study is required to improve the proposed methodology.

5. Conclusions

In this study, an integrated quality control process capable of validating and reconstructing the missing and false-reading precipitation and water level data were developed. The validation procedure was composed of a time-series check algorithm for validating missing data and five types of algorithms for validating false-reading data, which was composed of physical, duration, trend, statistical, and spatial checks, which were categorized into the quick-check and deep-check algorithms. For error data reconstruction, five reconstruction methods such as selected-time method, linear interpolation, cubic spline, ARIMA, and spatial interpolation were applied. The developed validation and reconstruction process was applied to 32 precipitation stations and 24 water level stations in the Daecheong Dam basin, South Korea. The application results were obtained by securing observation data at a 10-minute interval for at least five years and applying the developed quality control process to the data for one year in 2016. The analysis results are summarized as follows:

(1) The precipitation data validation results revealed that the 32 stations in the target basin had an average of 0.68% error data (missing and confirmed false-reading data combined).

(2) When precipitation data were reconstructed, it was found that the error rates of the 32 stations was improved through the five reconstruction methods, and the linear interpolation method exhibited the largest overall error rate improvement.
(3) For demonstration purposes, the precipitation database from the P-30 station was continuously reconstructed through the feedback procedure. The most effective reconstruction method was selected by evaluating the error reduction rate and user's selection, and all missing and false-reading data were completely reconstructed after three iterations.
(4) The water level data validation results showed that the 24 stations in the target basin had an average of 1.66% error data (missing and confirmed false-reading data combined).
(5) When the water level data were reconstructed, the ARIMA method exhibited the largest error rate improvement owing to the hydrological characteristics of water level with high autocorrelation.
(6) When the water level database of the W-10 station was continuously reconstructed through the feedback procedure for demonstration, all missing and false-reading data were completely reconstructed after three iterations.

This study proposed an approach that utilizes and integrates the well-known conventional algorithms for easy implementation in the field. The developed process reduces unnecessary data processing by classifying false data into suspected false and confirmed false datasets based on the combined results of validation methods. In addition, an iterative feedback analysis is proposed to further improve and finally correct the error data.

Most reconstruction methods showed a promising performance, and it is expected that the proposed multiple reconstructed values and iterative feedback process will present a more reasonable basis for developing a data quality control tool to aid data managers. As for related research, real-time automatic validation and reconstruction techniques could be applied to develop a practical quality control model. This will make a wide contribution from basic database management to data-driven diverse research in the field of water resources engineering.

Author Contributions: Conceptualization, D.-G.Y., T.-W.K., and D.K.; methodology, D.-G.Y., T.-W.K., and D.K.; software, G.J. and J.-Y.L.; validation, D.-G.Y., T.-W.K., and D.K.; formal analysis, J.-W.N. and D.-G.Y.; writing—original draft preparation, G.J. and J.-Y.L.; writing—review and editing, D.-G.Y., T.-W.K., and D.K.; supervision, J.-W.N. and T.-W.K.; funding acquisition, D.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by (1) Korea Ministry of Environment (MOE) as the "Graduate School specializing in Climate Change" and (2) National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2020R1A2C2009517) (MSIT: Ministry of Science and ICT).

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available because of privacy concerns.

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

References

1. Bedient, H.A.; Cressman, G.P. An experiment in automatic data processing. *Mon. Weather Rev.* **1957**, *85*, 333–340. [CrossRef]
2. Shuman, F.G. Numerical methods in weather prediction: II. Smoothing and filtering. *Mon. Weather Rev.* **1957**, *85*, 357–361. [CrossRef]
3. Haug, O. *A Method for Numerical Weather Map Analysis*; Scientific Report, No. 5; Norske Meteorologiske Institutt: Oslo, Norway, 1959; p. 10.
4. Filippov, V.V. *Quality Control Procedures for Meteorological Data*; Secretariat of the World Meteorological Organization: Geneva, Switzerland, 1968; p. 38.
5. Abbott, P.F. *Guidelines on the Quality Control of Surface Climatological Data*; WMO/TD: Geneva, Switzerland, 1986; p. 111.
6. Aguilar, E.; Auer, I.; Brunet, M.; Peterson, T.C.; Wieringa, J. *Guidance on Metadata and Homogenization*; WMO/TD: Geneva, Switzerland, 2003; pp. 1–53.
7. Zahumenský, I. *Guidelines on Quality Control Procedures for Data from Automatic Weather Stations*; World Meteorological Organization: Geneva, Switzerland, 2004; pp. 1–10.

8. Yevjevich, V.M.; Jeng, R.I.S. *Properties of Non-Homogeneous Hydrologic Time Series*; Colorado State University: Fort Collins, CO, USA, 1969.
9. Peterson, T.C.; Vose, R.; Schmoyer, R.; Razuvaëv, V. Global Historical Climatology Network (GHCN) quality control of monthly temperature data. *Int. J. Climatol. A J. R. Meteorol. Soc.* **1998**, *18*, 1169–1179. [CrossRef]
10. Meek, D.W.; Hatfield, J.L. Data quality checking for single station meteorological databases. *Agric. For. Meteorol.* **1994**, *69*, 85–109. [CrossRef]
11. Upton, G.J.G.; Rahimi, A.R. On-line detection of errors in tipping-bucket raingauges. *J. Hydrol.* **2003**, *278*, 197–212. [CrossRef]
12. Tollerud, E.; Collander, R.; Lin, Y.; Loughe, A. On the Performance, Impact, Liabilities of Automated Precipitation Gage Screening Algorithms. In Proceedings of the 21st Conference on Weather Analysis and Forecasting/17th Conference on Numerical Weather Prediction, Washington, DC, USA, 4 August 2005.
13. Qi, Y.; Martinaitis, S.; Zhang, J.; Cocks, S. A real-time automated quality control of hourly rain gauge data based on multiple sensors in MRMS system. *J. Hydrometeorol.* **2016**, *17*, 1675–1691. [CrossRef]
14. March, W.J. *Compendium of Lecture Notes in Climatology for Class IV Meteorological Personnel*; World Meteorological Organization: Geneva, Switzerland, 1973.
15. Linacre, E. *Climate Data and Resources: A Reference and Guide*; Routledge: Oxfordshire, UK, 1992.
16. Acock, M.C.; Pachepsky, Y.A. Estimating missing weather data for agricultural simulations using group method of data handling. *J. Appl. Meteorol.* **2000**, *39*, 1176–1184. [CrossRef]
17. Willmott, C.J.; Robeson, S.M.; Feddema, J.J. Estimating continental and terrestrial precipitation averages from rain-gauge networks. *Int. J. Climatol.* **1994**, *14*, 403–414. [CrossRef]
18. Xia, Y.; Fabian, P.; Stohl, A.; Winterhalter, M. Forest climatology: Estimation of missing values for Bavaria, Germany. *Agric. For. Meteorol.* **1999**, *96*, 131–144. [CrossRef]
19. Teegavarapu, R.S.; Chandramouli, V. Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. *J. Hydrol.* **2005**, *312*, 191–206. [CrossRef]
20. Franklin, M.; Kotamarthi, V.R.; Stein, M.L.; Cook, D.R. Generating Data Ensembles over a Model Grid from Sparse Climate Point Measurements. In *Journal of Physics: Conference Series*; IOP Publishing: Washington, DC, USA, 2008.
21. Gentilucci, M.; Barbieri, M.; Burt, P.; D'Aprile, F. Preliminary data validation and reconstruction of temperature and precipitation in Central Italy. *Geosciences* **2008**, *8*, 202. [CrossRef]
22. Kuligowski, R.J.; Barros, A.P. Using artificial neural networks to estimate missing precipitation data 1. *J. Am. Water Resour. Assoc.* **1998**, *34*, 1437–1447. [CrossRef]
23. Coulibaly, P.; Evora, N.D. Comparison of neural network methods for infilling missing daily weather records. *J. Hydrol.* **2007**, *341*, 27–41. [CrossRef]
24. Kim, J.W.; Pachepsky, Y.A. Reconstructing missing daily precipitation data using regression trees and artificial neural networks for SWAT streamflow simulation. *J. Hydrol.* **2010**, *394*, 305–314. [CrossRef]
25. Crebas, J.I. HYMOS: A database management and processing system for hydrometeorological data. In Proceedings of the First International Conference on Hydroinformatics, Delft, The Netherlands, 19–23 September 1994.
26. AQUATIC Informatics. AQUARIUS Time-Series Developer Guide: Field Data Plug-In Framework; 2017. Available online: https://usermanual.wiki/Pdf/AQUARIUSDeveloperGuideFieldDataPluginFramework.291959327/view (accessed on 19 August 2021).
27. KMA. *The Guidelines for Integrated Data Quality Management of National Climate Data*; KMA (Korea Meteorological Administration): Seoul, Korea, 2016.
28. Kim, H.S.; Kim, C.S. Application of the Quality Control System for Hydrological Data. In Proceedings of the Korean Society of Civil Engineers 2005 Conference, Jeju, Korea, 20–21 October 2005.
29. MLTM. *2010–2019 Master Plan for Hydrological*; MLTM (Ministry of Land, Transport and Maritime Affairs): Sejong, Korea, 2008.
30. Shulski, M.D.; You, J.; Krieger, J.R.; Baule, W.; Zhang, J.; Zhang, X.; Horowitz, W. Quality assessment of meteorological data for the Beaufort and Chukchi Sea coastal region using automated routines. *Arctic* **2014**, *67*, 104–112. [CrossRef]
31. Hubbard, K.G.; Goddard, S.; Sorensen, W.D.; Wells, N.; Osugi, T.T. Performance of quality assurance procedures for an applied climate information system. *J. Atmos. Ocean. Technol.* **2005**, *22*, 105–112. [CrossRef]
32. K-water. *Development of Quality Control Algorithm for Standard Database of Water Information*; Interim Report of Korea Water Resources Corporation: Daejeon, Korea, 2019.