

Article

A Novel Online Hydrological Data Quality Control Approach Based on Adaptive Differential Evolution

Qun Zhao ¹, Shicheng Cui ^{1,*}, Yuelong Zhu ², Rui Li ¹ and Xudong Zhou ^{3,4}¹ School of Computer Engineering, Nanjing Institute of Technology, Nanjing 211167, China; qunzhao@njit.edu.cn (Q.Z.); lirui_92@njit.edu.cn (R.L.)² College of Computer and Information, Hohai University, Nanjing 211100, China; ylzhu@hhu.edu.cn³ Institute of Hydraulics and Ocean Engineering, Ningbo University, Ningbo 315211, China; zhouxudong@nbu.edu.cn⁴ School of Civil and Environmental Engineering, Ningbo University, Ningbo 315211, China

* Correspondence: shicheng.cui@njit.edu.cn; Tel.: +86-13851822286

Abstract: The quality of hydrological data has a significant impact on hydrological models, where stable and anomaly-free hydrological time series typically yield more valuable patterns. In this paper, we conduct data analysis and propose an online hydrological data quality control method based on an adaptive differential evolution algorithm according to the characteristics of hydrological data. Taking into account the characteristics of continuity, periodicity, and seasonality, we develop a Periodic Temporal Long Short-Term Memory (PT-LSTM) predictive control model. Building upon the real-time nature of the data, we apply the Adaptive Differential Evolution algorithm to optimize PT-LSTM, creating an Online Composite Predictive Control Model (OCPT-LSTM) that provides confidence intervals and recommended values for control and replacement. The experimental results demonstrate that the proposed data quality control method effectively manages data quality; detects data anomalies; provides suggested values; reduces reliance on manual intervention; provides a solid data foundation for hydrological data analysis work; and helps hydrological personnel in water resource scheduling, flood control, and other related tasks. Meanwhile, the proposed method can also be applied to the analysis of time series data in other industries.



Citation: Zhao, Q.; Cui, S.; Zhu, Y.; Li, R.; Zhou, X. A Novel Online Hydrological Data Quality Control Approach Based on Adaptive Differential Evolution. *Mathematics* **2024**, *12*, 1821. <https://doi.org/10.3390/math12121821>

Academic Editors: Zhihua Zhang and António Lopes

Received: 26 April 2024

Revised: 19 May 2024

Accepted: 7 June 2024

Published: 12 June 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: hydrological data quality; data analysis; predictive control model; differential evolution algorithm; online optimization

MSC: 68T09

1. Introduction

The continuous development of the Information Age has led to the establishment and improvement of an increasing number of intelligent hydrological monitoring stations. The coverage of hydrological data has become more comprehensive, with massive numbers of historical and real-time hydrological data being collected and stored in databases. These hydrological data have a large quantity and diverse types. Extracting potential patterns from massive hydrological data and conducting data analysis of hydrological phenomena holds significant importance for hydrological work.

Hydrological data mining models have certain requirements for data quality. Only high-quality data can provide valuable information and knowledge, and a good dataset can reduce the instability of the models [1]. The quality of hydrological data significantly impacts hydrological data mining models, with stable and anomaly-free hydrological time series typically yielding more valuable patterns [2]. When mining hydrological data, addressing data quality issues such as missing values and anomalies within a controllable range is crucial to ensure the accuracy and credibility of experiments and applications. Factors such as inaccuracies in the installation, upgrading, and sensor interference of

monitoring systems at hydrological stations, as well as abnormal jumps or failures of instruments caused by extreme weather conditions, may lead to anomalies, errors, and other phenomena in hydrological data. Such errors and anomalies in data records can sometimes mislead disaster management assessments, leading to serious consequences. However, there are still challenges in the current hydrological data quality control:

1. A strong reliance on manual intervention and high manual cleaning costs. In China, most hydrological data quality control methods are still in the theoretical research and modeling stage. The control of data quality often requires manual intervention [3], lacking intelligent data quality control algorithms and models.
2. Low credibility of short-term missing data imputation. For hydrological data with short-term missing values due to machine failures, commonly used interpolation methods such as the average method, weighted method, or simple spatial interpolation method [4,5] are employed. The effectiveness of linear interpolation methods [6,7] varies based on data collection density, rendering the credibility of imputed data difficult to assess.
3. Lack of reliable replacement values for anomalous data. Currently, many methods [8–13] detect anomalous data through basic checks and standard settings but fail to provide reliable replacement values for substituting anomalous data.

To address the aforementioned issues, this study investigates how to establish an intelligent hydrological data control method based on the continuity, periodicity, seasonality, and real-time characteristics of hydrological data. The aim is to reduce reliance on manual intervention, optimize hydrological data quality, and consequently enhance the stability and precision of data mining models.

2. Related Works

Data quality [14] is a measure of the accuracy, completeness, and consistency of data [15], forming the foundation for data applications. The advent of the big data era has brought challenges to data management and information extraction while ensuring data quality is a prerequisite for effective data analysis [16]. Data quality control involves taking specific measures to ensure that data collected, stored, and transmitted meet certain requirements. Implementing quality control measures on hydrological data can effectively detect missing and suspicious data, maintaining data quality within a controllable range. This lays the groundwork for hydrological predictions, flood forecasts, and other applications. Therefore, controlling the quality of hydrological data is essential to ensure the reliability of hydrological data.

Numerous scholars in the field of hydrology have proposed various schemes for the control, management, and assessment of hydrological data quality. Sciuto et al. [8] conducted control on temperature data by setting a sliding window for data consistency. They checked whether observed data fell within confidence intervals of two different fixed probabilities, aiming to identify anomalies in temperature data. Steinacker et al. [9] analyzed types of data errors and proposed an approach based on the spatial location of monitoring stations to establish a topological structure for determining weights between stations and guiding artificial assistance in control directions. Sciuto [10] and Abbot [11] used climate indices, atmospheric temperatures, and historical data from surrounding stations to forecast rainfall data at monitoring stations and set detection intervals, identifying data outside the intervals as anomalies. Fu Fangjing and Luo Xi [12] independently conducted a thorough examination and detection of hydrological data entering the database, focusing on overall rationality, completeness, and consistency, in order to identify anomalous data. Yu Yufeng [13] employed Benford's Law to detect the distribution pattern and overall rationality of hydrological data, proposing a semi-automatic cleaning approach to improve hydrological data quality [17], offering innovative insights for hydrological data quality control. Tang Dandan et al. [18], addressing the issue of detecting anomalies in continuous hydrological data, proposed two combination models. They used the confidence intervals provided by the models and statistical methods to determine data anomalies. Zhao et al. [19]

presented two control methods for different types of hydrological data, setting suspicious levels and recommended values for data reference by hydrological personnel. Li et al. [20] use the TOPographic Kinematic Approximation and Integration (TOPKAPI) distributed model for deterministic forecasts with low-quality input data and then use the Hydrologic Uncertainty Processor (HUP) to provide the probabilistic forecast results for operational practices. Yu et al. [21] detected abnormal patterns in hydrological time series based on weighted probability suffix trees, effectively identifying anomalies by combining trend characteristic symbols and clustering approximate patterns. Lattawit et al. [22] developed a median-based statistical outlier detection approach using a sliding window technique. The results show that the Spline interpolation method yielded a superior performance on non-cyclical data while the LSTM outperformed other interpolation methods on a distinct tidal data pattern. Hyojoong K and Heeyoung K proposed a method for contextual anomaly detection, particularly in the case where the response and contextual variables are both high-dimensional and complex [23], and a contextual anomaly detection method for multivariate time series data [24]. The experiments showed that the proposed methods all perform well.

However, there are still numerous challenges in the current hydrological data quality control. For instance, the current hydrological data quality control methods in China are mostly in the theoretical research and modeling stage. There is strong dependence on manual cleaning, and there is a lack of intelligent data quality control algorithms and models for real-time intelligent control of hydrological data.

3. Proposed Method

Hydrological time series data, such as water level and discharge, exhibit characteristics such as continuity, periodicity, seasonality, and real-time properties. These features make hydrological data subject to temporal variations. Moreover, the changes in hydrological data over a period of time exhibit spatial proximity on the time scale, displaying a certain similarity in the variations of samples with close temporal distances and data with the same spatial positions in different periods.

For historical hydrological data that have not undergone reorganization, we use deep learning methods [25,26] to establish a predictive control model. The model, based on the temporal continuity, periodicity, seasonality, and spatial positions on the time scale of hydrological data, establishes a Periodic Temporal LSTM Predictive Control Model (PT-LSTM). The model provides a prediction line, constructs confidence intervals to detect anomalous data, and offers recommended values. Additionally, considering the real-time nature of hydrological data, an improved version of the PT-LSTM model is proposed, termed the Online Optimized Combination Predictive Control Model (OCPT-LSTM). This model incorporates real-time data for real-time model correction, providing confidence intervals to detect data quality and recommending values for replacement.

On the foundation of basic quality controls (such as missing data and format checks, and extreme value checks), this study constructs a PT-LSTM Model and an OCPT-LSTM Model based on the characteristics of hydrological data, for predictive control. The models provide confidence intervals for detected data, and any data outside these intervals is considered suspicious and requires manual judgment on whether to replace it with the recommended values provided by the models. The process is illustrated in Figure 1.

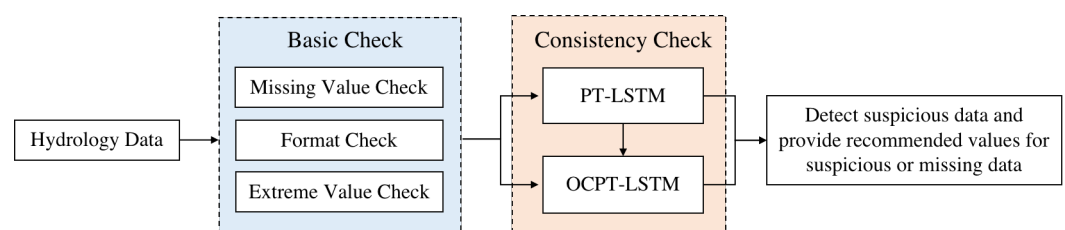


Figure 1. Quality control method of hydrological data.

3.1. Periodic Temporal Predictive Control Model

Hydrological data exhibit continuity and periodicity within a certain timeframe. Addressing their continuity, predictions for the data at time $n + 1$ can be made using historical data from the previous n time steps. Regarding their periodicity, where data transformations at the same time on adjacent days fall within a certain range, predicting future hydrological conditions at the same time using historical data at the corresponding time is feasible.

3.1.1. Dataset Construction

The continuity of hydrological data ensures that data changes within the same period will fall within a certain range. Periodicity guarantees that data follow similar patterns across different cycles. Seasonality ensures that the magnitude of hydrological data changes is related within the same season. Constructing datasets based on the temporal characteristics and spatial positions on the time scale of hydrological data allows for the exploration of more patterns within the data and enables better analysis.

In addressing the continuity of hydrological data, predictions for the data at time $t + 1$ are made based on the real-time data at historical times up to t . The historical hydrological data are divided into M groups continuously, each containing N^T data points. The input vector is represented as $[x_1, x_2, x_3, \dots, x_{MN^T}]$. These are arranged into an $M \times N^T$ matrix X , where the matrix has a periodicity of N^T , and it is structured as follows:

$$X = \begin{bmatrix} x_1 & x_2 & \dots & x_{N^T} \\ x_{N^T+1} & x_{N^T+2} & \dots & x_{2N^T} \\ \vdots & \vdots & \ddots & \vdots \\ x_{(M-1)N^T+1} & x_{(M-1)N^T+2} & \dots & x_{MN^T} \end{bmatrix} \quad (1)$$

The above matrix is constructed from m subsequences with a periodicity of N^T , where x_{N^T+2} is one element. Apart from the adjacent x_{N^T+1} and x_{N^T+3} in both directions, which are in spatial proximity to x_{N^T+2} , there are also data points from the same time in the previous and subsequent cycles, such as x_2 and x_{2N^T+2} for the same time within the same day. Regarding the periodicity of hydrological data, the data at the same time (reference time) for the previous m days exhibit a similar trend to the data at the same time for the $m + 1$ day. A day consists of 24 time points, and here N^T can be taken as 24 for a daily cycle.

In response to the seasonality of hydrological data, it is observed that different seasons exhibit distinct hydrological conditions. Specifically, summers are often characterized by flood seasons with heavy rainfall, leading to significant fluctuations in flow. Autumns experience continuous drizzle, causing a delayed end to the flood season and maintaining a high flow rate. Springs and winters are mostly dry seasons with less rainfall, resulting in relatively stable flow changes. Figure 2 depicts the daily flow rates for the WaiZhou station in 2007 and the AoXiaping station in 2008. The WaiZhou Station is located on the Ganjiang River in Poyang Lake, Jiangxi Province, at 115.83 east longitude and 28.63 north latitude. The AoXiaping Station is located on the Heyuan River in Poyang Lake, Jiangxi Province, at 114.43 east longitude and 26.18 north latitude. The AoXiaping Station is located in the southwest direction of the WaiZhou Station, with a relatively long latitude interval, which is of reference significance for comparison.

Despite the considerable difference in flow rates between the two stations, there are similarities in their trend patterns. For the WaiZhou station, significant flow variations occur from April to September, followed by a relatively steady period after September. Similarly, the AoXiaping station exhibits pronounced flow changes from May to November, particularly during the period from June to August, when flow fluctuations are significant. In summary, summers and autumns are often characterized by high flow periods, while springs and winters are generally dry seasons.

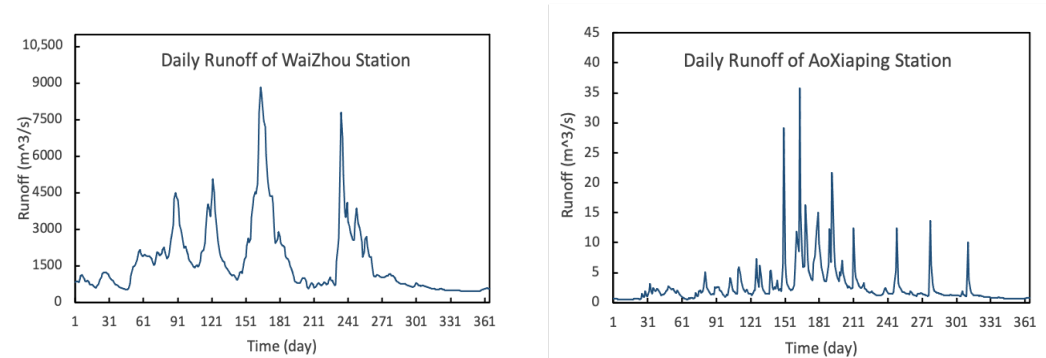


Figure 2. Daily discharge of WaiZhou station and AoXiaping station.

Given the significant differences in hydrological conditions across seasons, the dataset can be divided into two categories based on seasons: spring and winter (dry season) as one category, and summer and autumn (wet season) as another category. This allows for separate training based on distinct seasonal patterns.

3.1.2. PT-LSTM Model Construction

The LSTM neural network can retain long-term information and effectively capture hydrological patterns, providing certain advantages in modeling hydrological sequences. In this paper, considering the specific temporal characteristics of hydrological data, we propose the PT-LSTM model. The model based on the continuity, periodicity, and seasonality of time series, and their spatial proximity in the time scale, divides the time series into multiple sub-sequences according to periods. These subsequences are then arranged in matrices to facilitate spatial exploration. Next, the model linearly maps these sub-sequences by multiplying them with a weight matrix to obtain data with continuous and periodic characteristics. Subsequently, the data are fed into a recursive layer for exploration to obtain the prediction results. The network structure of the PT-LSTM model is illustrated in Figure 3. The PT-LSTM model not only considers the temporal characteristics of time series data and the spatial position in the time scale, reducing training complexity, but also minimizes the reduction in the neural network time step, thereby preserving the original data information to the maximum extent.

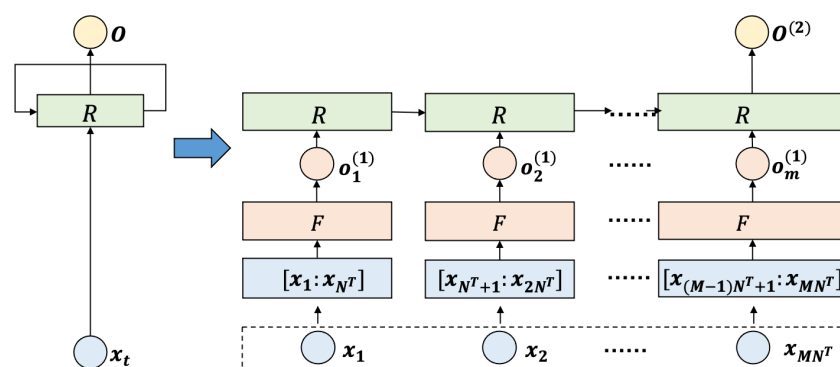


Figure 3. Network structure of PT-LSTM.

Where x_t, x_1, x_2, x_{MN^T} represent input data; $[x_1 : x_{N^T}], [x_{N^T+1} : x_{2N^T}], [x_{(M-1)N^T+1} : x_{MN^T}]$ are sub-sequences of the processed periodic matrix; F is the fully connected layer in PT-LSTM for linear mapping of the periodic matrix; $o_1^{(1)}, o_2^{(1)}, o_m^{(1)}$ are the outputs after linear processing; R is the recursive layer in PT-LSTM; and $O, O^{(2)}$ is the output after the recursive layer.

The following provides a detailed explanation of the PT-LSTM model:

(1) Data Preprocessing and Parameter Setting

Choose the data. First, divide the data into two categories based on the wet season and dry season. Process each category of data separately. Divide the historical hydrological data into M groups, each with N^T data points. The input vector is $[x_1, x_2, x_3, \dots, x_{MN^T}]$. Arrange it into an $M \times N^T$ matrix with a period of N^T according to Equation (1). Normalize the data to be within the range of $[-1, 1]$. Shuffle the order randomly and select 60% of the data as the training set, 20% as the validation set, and 20% as the test set.

Initialize the network weights. Specify the activation functions as the sigmoid function and tanh function. Use the mean squared error function as the loss function, calculated as follows:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - Y_i)^2 \quad (2)$$

where y_i is the predicted value at time step i , Y_i is the true value at time step i , and n is the number of samples input into the model.

(2) Linear Mapping Processing

Perform linear mapping on the above M sets of subsequences, and input the dataset into the fully connected layer according to the pre-divided groups. The number of input nodes in this layer is N^T , and the number of output nodes is 1. The mathematical expression is as follows:

$$O^{(1)} = XW^{(1)} = \begin{bmatrix} x_1 & x_2 & \dots & x_{N^T} \\ x_{N^T+1} & x_{N^T+2} & \dots & x_{2N^T} \\ \vdots & \vdots & \ddots & \vdots \\ x_{(M-1)N^T+1} & x_{(M-1)N^T+2} & \dots & x_{MN^T} \end{bmatrix} \begin{bmatrix} w_{1,1}^{(1)} \\ \vdots \\ \vdots \\ w_{N^T,1}^{(1)} \end{bmatrix} \quad (3)$$

$$O^{(1)} = \begin{bmatrix} w_{1,1}^{(1)}x_1 & w_{2,1}^{(1)}x_2 & \dots & w_{N^T,1}^{(1)}x_{N^T} \\ w_{1,1}^{(1)}x_{N^T+1} & w_{2,1}^{(1)}x_{N^T+2} & \dots & w_{N^T,1}^{(1)}x_{2N^T} \\ \vdots & \vdots & \ddots & \vdots \\ w_{1,1}^{(1)}x_{(M-1)N^T+1} & w_{2,1}^{(1)}x_{(M-1)N^T+2} & \dots & w_{N^T,1}^{(1)}x_{MN^T} \end{bmatrix} = \begin{bmatrix} o_{1,1}^{(1)} \\ o_{2,1}^{(1)} \\ \vdots \\ o_{M,1}^{(1)} \end{bmatrix} \quad (4)$$

where $W^{(1)}$ is the weight matrix of this layer and X is the input matrix of $M \times N^T$. $O^{(1)}$ is used as the input to the recursive layer. In the sequence input, each element represents one time step, and the total number of steps in the network is M . For each time step, the output of the previous layer is used as the input for the current layer. $o_t^{(1)} = o_{t,1}^{(1)}$, $t = 1, 2, 3, \dots, M$. The output of the second hidden layer at time step t is given by:

$$o_t^{(2)} = \sigma(o_{t-1}^{(2)}, o_t^{(1)}) \quad (5)$$

The output of the second hidden layer is denoted as $O^{(2)} = [o_1^{(2)}, o_2^{(2)}, \dots, o_M^{(2)}]$, where $\sigma(\cdot)$ represents the activation function. The output at time step t for the n -th hidden layer is given by $o_t^{(n)} = \sigma(o_{t-1}^{(n)}, o_t^{(n-1)})$, and the output of the n -th hidden layer is $O^{(n)} = [o_1^{(n)}, o_2^{(n)}, \dots, o_M^{(n)}]$. Compared to traditional neural networks, this construction reduces the time step from MN^T to M , reducing the time step and lowering training complexity. It effectively addresses the vanishing gradient problem of LSTM while modeling long sequences as short sequences. This overcomes LSTM's limitations on the time step and retains a significant amount of original sequence information.

(3) Model Training

Take the output $O^{(1)}$ of the fully connected layer as input and perform forward calculation for each unit in the hidden layer using the following formula:

$$\text{Forget Gate: } F_t = \sigma(W_f \times [h_{t-1}, o_t^{(1)}] + b_f)$$

$$\text{Input Gate: } I_t = \sigma(W_i \times [h_{t-1}, o_t^{(1)}] + b_i)$$

$$\text{Output Gate: } O_t = \sigma(W_o \times [h_{t-1}, o_t^{(1)}] + b_o)$$

$$\text{Cell State: } c_t^{\sim} = \tanh(W_c \times [h_{t-1}, o_t^{(1)}] + b_c), c_t = F_t \odot c_{t-1} + i_t \odot c_t^{\sim}$$

$$\text{Hidden State: } h_t = O_t \odot \tanh(c_t)$$

Calculate the prediction error of the model, then backpropagate the prediction error through each neuron, update the network weights, and iterate through the training process. When the mean squared error no longer decreases or meets certain criteria, the iteration concludes, and the network training is completed.

After obtaining the prediction reference line, confidence intervals can be constructed. The confidence interval is determined based on the estimated value to identify the range within which the true value may exist. With the predicted value y and mean squared error MSE , the confidence interval is formed as follows: $Z = [y - \delta \cdot MSE, y + \delta \cdot MSE]$. Here, δ is a weighted value with $\delta \in [2, 5]$, adjusted according to seasonal variations [27]. If $y - \delta \cdot MSE < 0$, then the confidence interval becomes $Z = [0, y + \delta \cdot MSE]$.

After obtaining the confidence interval, data control is performed using the confidence interval. Data outside the confidence interval are considered suspicious, and the predicted values from the model serve as suggested values for hydrological staff to decide whether to replace them.

3.2. Online Optimization Combination Control Model

The hydrological station needs to upload a large number of data every day, and these data have real-time characteristics. When detecting hydrological data, ordinary prediction control models may introduce errors gradually due to the lack of timely updates in the dataset, leading to biases in data prediction control and weak model timeliness. Building an online, real-time optimization model can provide timely and effective control of the data.

3.2.1. OCPT-LSTM Model Concept

This section combines the PT-LSTM model with the adaptive differential evolution algorithm to establish the OCPT-LSTM model. Drawing on the idea of preserving excellent individuals and their fitness values in a temporary array within the population, this approach involves selecting the top m excellent parameters as the initial parameters for the model. The goal is to establish a weighted combination prediction and control model, thereby enhancing the robustness of prediction and control.

The Online Optimization Combination Control Model employs the OCPT-LSTM model as the prediction model and the PT-LSTM model as the base model. After a certain period, a temporary base model group is constructed, using the $m + 1$ to $m + k$ excellent parameters from the temporary array to establish k temporary models. These models are trained using historical real-time data from the same period as the dataset. By comparing the temporary base models with the base models in the combination model, if the best model in the temporary models performs better than the worst model in the combination control model in terms of prediction, timely replacement occurs. Finally, the combination prediction and control model is reweighted based on prediction errors. The OCPT-LSTM model dynamically replaces the worst-performing excellent model with an outstanding temporary model. This process ensures the stability of the combination model with minimal changes, continually adjusting the timeliness of the control model. The model framework is illustrated in Figure 4.

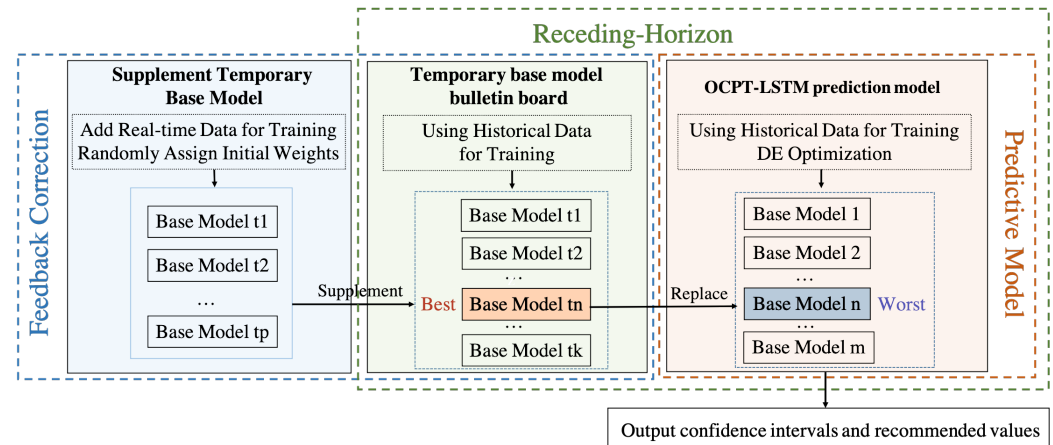


Figure 4. Framework of online optimal combined predictive control method.

3.2.2. OCPT-LSTM Model Construction

The Differential Evolution (DE) algorithm is an optimization algorithm based on human theories. It simulates the process of cooperation and competition among individuals in a population, guiding the group towards optimal solutions through the selection of cooperative and competitive interactions. Compared to other optimization algorithms, the DE algorithm demonstrates superior performance in local search for complex optimization problems.

The convergence performance of the DE algorithm is largely influenced by the selection value of mutation factors. The Adaptive Differential Evolution (ADE) algorithm adds an adaptive scaling factor on the basis of the DE algorithm. We use adaptive mutation factors to enhance the convergence performance of the model. In the early iterations, the mutation factor is large, helping to maintain the diversity of the population. As the number of iterations increases, the mutation factor decreases, preserving excellent population information and avoiding disruption to the optimal solution. The Adaptive Mutation Factor is as follows:

$$F = F_{max} - \frac{t(F_{max} - F_{min})}{T} \quad (6)$$

where F is the mutation factor, t is the iteration count, T is the maximum iteration count, F_{max} is the maximum value of the mutation factor, and F_{min} is the minimum value of the mutation factor.

Due to the need for real-time updates in the online predictive control model, the periodic time perspective may not effectively control data generated by sudden extreme weather events. Therefore, this paper adopts a continuous real-time time perspective for predictive control, predicting future conditions based on the preceding n time steps, such as the hydrological condition at the $n + 1$ time step. In this section, a single predictive control model is constructed based on the idea of an adaptive differential optimization algorithm, and an online optimized combination control model is developed, as illustrated in Figure 5.

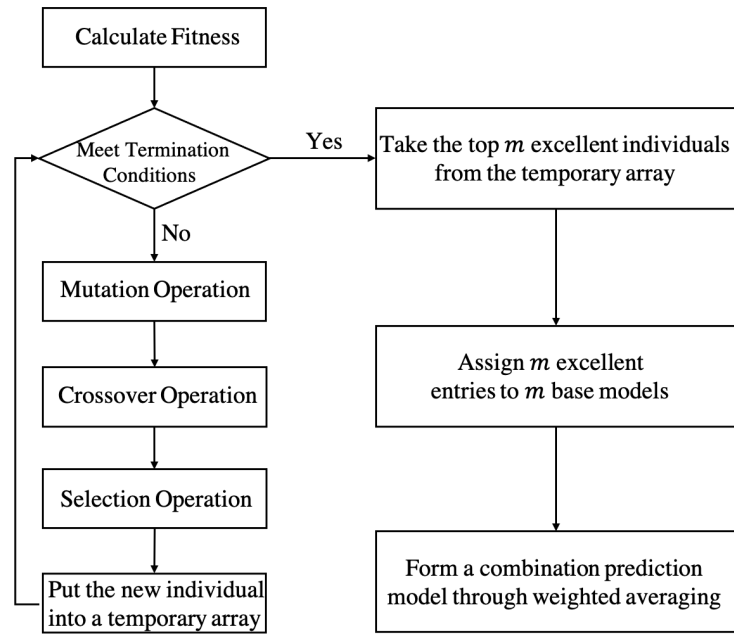


Figure 5. Online optimal combined prediction model process.

The specific steps of the online optimized combination control model are as follows:

- (1) Training set construction. Select the dataset, apply the periodic matrix processing as described in Section 3.1.1, and normalize the data. Shuffle the order randomly and choose 60% of the data as the training set, 20% as the validation set, and 20% as the test set.
- (2) Encoding. Uniformly encode the parameters of PT-LSTM into individuals.
- (3) Initialization operation.

$$X_j^i(0) = x_j^{il} + rand(0, 1)(x_j^{iu} - x_j^{il}) \quad (7)$$

where $x_j^i(0)$ is the i th individual in the j th dimension, x_j^{il} is the lower bound of the j th dimension, x_j^{iu} is the upper bound of the j th dimension, and $rand(0, 1)$ represents a random number on $[0, 1]$.

- (4) Population Generation. Randomly generate the initial population in step 3 with a size of np and initialize mutation factor F , individual gene dimension D , maximum number of iterations T , and crossover probability CR .

- (5) Fitness Calculation. Calculate the fitness of each individual using mean squared error as the fitness criterion.

- (6) Mutation Operation. Randomly select two individuals from the population and perform adaptive mutation on the individuals, generating mutated individuals.

$$V_i(t) = x_j^{r1}(t) + F(x_j^{r2}(t) - x_j^{r3}(t)) \quad (8)$$

where $r1$, $r2$, and $r3$ are randomly generated integers, and $r \in 1, 2, \dots, NP$, $x_j^{r1}(t)$, $x_j^{r2}(t)$, $x_j^{r3}(t)$ are the j th dimensional components of the three individuals, respectively. F is the variation factor corresponding to Formula (6).

- (7) Crossover Operation. Perform a crossover operation on the mutated individuals to obtain new individuals. When $rand(0, 1) \geq CR$, $u_j(t) = v_j(t)$, otherwise $u_j(t) = x_j(t)$. Generally speaking, the larger the CR , the faster the convergence speed, which is beneficial for local search. However, excessive crossover probability can lead to premature convergence and the poor robustness of the algorithm. The smaller the CR , the better the population diversity, which is beneficial for global search. After crossing, the new individual $U_i(t) = (u_1(t), u_2(t), \dots, u_2(t))$.

(8) Selection Operation. Employ a greedy algorithm to select individuals from both the original population and the newly generated individuals. Individuals with lower fitness values are chosen to replace the corresponding individuals in the original population. The fitness values of the selected excellent individuals are saved in a temporary array.

(9) Selecting Excellent Individuals. Iteratively execute the mutation, crossover, and selection operations on the population until the iteration limit is reached or the position of the best individual is determined. Sort the temporary array in ascending order of fitness values, and use the top m individuals to construct the model.

(10) Establishing Base Models. Decode the m excellent individuals generated in (9) and build m base models. Assign differential initial weights to these base models and train them to construct m PT-LSTM models.

(11) Combination Model. Weightedly combine the PT-LSTM base models to establish the OCPT-LSTM model.

Once the model is established, every certain period, the m excellent individuals in the temporary array, along with k randomly generated individuals, form the initial population with a size of $n(m + k)$. The steps (5) to (10) are then executed. The individual with the minimum fitness among them is selected for decoding, and the optimal initial threshold is assigned for training. If the error of this new network is less than that of the network in the ensemble model, replace the base model with the maximum error in the ensemble model and re-establish the model; otherwise, keep the original ensemble model unchanged.

After obtaining the predicted values, a confidence interval is established for the detection of data. First, use the predicted data at different time points to draw a time series prediction reference line. For each time point on the prediction reference line, add the positive and negative error values of the 95% confidence interval of the model's prediction error as the upper and lower bounds. Finally, use wavelet transform to smooth the control boundaries, taking the part between the two smooth curves as the confidence interval.

This paper builds two models, PT-LSTM and OCPT-LSTM, for predictive control on the foundation of basic quality control. The models construct confidence intervals based on the prediction results to detect anomalous data. The detection interval is dynamically adjusted according to the data variations. Data outside the interval are considered suspicious and require manual judgment on whether to replace it with the suggested values provided by the model. Both models can effectively provide confidence intervals and advisory values. The choice between the models depends on the specific requirements of the control situation. When only detecting suspicious data or pursuing efficiency, the PT-LSTM model can be chosen. When accuracy is crucial and replacement values need to be provided, the OCPT-LSTM model is a suitable choice. Both models can also construct confidence intervals separately, and a comprehensive confidence interval can be created for data control. Additionally, the two models can cross-validate each other. Data corrected by the PT-LSTM model, which are based on cyclical temporal prediction control, can undergo a secondary quality confirmation through the OCPT-LSTM ensemble model in the online prediction control approach.

4. Experiment and Analysis

4.1. Evaluation Metrics

This section employs four evaluation metrics as performance indicators: Coefficient of Determination (R^2), Root Mean Squared Error (RMSE), Absolute Error (AE), and Maximum Error (ME).

Coefficient of Determination (R^2) represents the degree of agreement between predicted and observed values, with a range of $[0, 1]$. A higher R^2 value indicates higher prediction accuracy. The formula for calculation is:

$$R^2 = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \quad (9)$$

where Y_i is the observed value, $\hat{Y}_i$ is the predicted value, and $\bar{Y}$ is the mean of the observed values.

Root Mean Square Error (RMSE) calculates the mean of the squared differences between predicted and actual data errors, and a smaller value indicates better performance. The formula for calculation is:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2} \quad (10)$$

Absolute Error (AE) is the absolute difference between predicted and actual monitored values, while Maximum Error (ME) is the maximum absolute difference between predicted and actual values.

4.2. Experimental Dataset

Jiangxi Province is surrounded by mountains on three sides, with higher terrain in the central and southern parts, mostly consisting of hilly plateaus. The northern part is surrounded by mountains with lower terrain. The water flows from high to low areas, and the unique topography of Jiangxi Province makes it prone to water accumulation in low-lying areas. Excessive water accumulation, coupled with untimely reservoir scheduling, can easily lead to flooding disasters. Poyang Lake, located in the central-southern part of Jiangxi Province, is the largest freshwater lake in China and a frequent and severely affected area by flooding in the middle and lower reaches of the Yangtze River.

For the experiment, the hydrological stations selected are WaiZhou (Station code: 62302250, longitude 115.83, and latitude 28.63) and ShiShang (Station code: 62312050, longitude 116.2, and latitude 26.53) in the Gan River basin of Poyang Lake, Jiangxi Province. WaiZhou Station is located in WaiZhou Village, Taohua Township, Nanchang City, while ShiShang Station is located in ShiShang Town, Ningdu County, under Ganzhou City. Both stations belong to the Ganjiang River, experiencing large fluctuations in flow, high peak flows, and are prone to anomalous data. The experiment considered these two hydrological stations with relatively large flow rates as the experimental points and involved selecting a portion of the hourly flow rate data from the WaiZhou hydrological station for the period 2004 to 2008 after data reorganization, totaling 23,136 records. Sixty percent of the reorganized data was used for training, 20% for validation, and 20% for testing and adjusting confidence intervals. Similarly, the Stone Shang station's hourly data were selected for partial periods from 2004 to 2008 for experimentation. Among these, 60% of the reorganized data was used for training and 20% for validation, and 20% of the non-reorganized data was used for testing, totaling 17,520 records.

Before conducting the experiments, preprocessing was carried out on the experimental dataset. Hourly flow data that has not been controlled may contain missing values. For such data, linear interpolation was used for missing value imputation. Linear interpolation involves calculating the value of an unknown variable by connecting two known values with a straight line. The formula for linear interpolation is as follows:

$$y = y_b + \frac{y_a - y_b}{x_a - x_b} (x - x_b) \quad (11)$$

where (x_a, y_a) and (x_b, y_b) represent the coordinates of the known points, and y denotes the value of the unknown variable at a position x within the interval $[x_a, x_b]$ along the straight line.

Next, the data will undergo a normalization process. Data normalization is a procedure to scale the data to a specific range, aiming to shorten the convergence time during training. Max–Min normalization and mean normalization are two commonly used normalization methods. However, Max–Min normalization scales the data to the range $[0, 1]$, while the size of the data after mean normalization cannot be fixed within a specific interval. To improve

the training speed of the network, this study adopts the Max–Min normalization method to normalize the data, as shown in the following formula:

$$X_i^o = \frac{X_i - X_{min}}{X_{max} - X_{min}} \quad (12)$$

In the time series data X , where X_{max} is the maximum value and X_{min} is the minimum value.

4.3. Model and Parameter Settings Comparison

In the experiment, the comparative models include Support Vector Machine Regression (SVR), Long Short-Term Memory Network (LSTM), PT-LSTM, and OCPT-LSTM.

The Support Vector Machine (SVM) [28,29] analyzes the model structure to find the minimum risk and optimal structure, constructing the optimal classification hyperplane. This hyperplane maximizes the distance between the hyperplane and different types of sample sets while ensuring classification accuracy. SVR is a variant of SVM designed for regression tasks and prediction calculations. SVR characterizes the non-linear relationship between the vectors to be predicted and support vectors in the training set to predict the vectors to be predicted in the test set. LSTM has been described in detail in Section 2. PT-LSTM and OCPT-LSTM are the proposed models in this chapter, which have been detailed earlier.

Experimental Environment: Operating System Windows 10, Processor Intel i7-9700, 2.4 GHz, Memory 32 GB, and GPU GTX 1050Ti.

Training parameters based on experimental tuning, the optimal hyperparameter configurations are as follows: the penalty coefficient of SVM is 0.3, the kernel function is the radial basis kernel function, and the parameter g is 0.05. The number of neuron layers of LSTM is 2, the number of hidden nodes is 50, the value of the learning factor lr is 0.004, the batchsize size is 64, and the number of iterations t is 50. The PT-LSTM in the period is 24, the number of neuron layers is 2, the number of hidden nodes is 150, the learning rate lr is 0.001, the batchsize is 64, and the number of iterations t is 50. In OCPT-LSTM, the range of the number of nodes in the hidden layer of the base model is 150, the learning factor lr is 0.005, the batchsize is 64, the model population size NP is set to 20, the dimensionality D is set to 30, the maximum variance factor F_{max} is 0.6, the minimum variance factor F_{min} is 0.2, the crossover probability CR is 0.5, and the maximum number of iterations T is 30.

4.4. Experimental Comparison and Result Analysis

This experiment conducts quality control on the data from the WaiZhou Station and ShiShang Station separately. The training data consist of cleaned hydrological data, and for the WaiZhou Station, the test data remain the cleaned data. The purpose of the WaiZhou Station experiment is to assess model accuracy and compare the prediction results of various models. Additionally, this experiment examines the usability of the prediction control model's confidence interval to observe whether the measured real data fall within the prediction control model's confidence interval and whether the interval effectively contains the measured real data. For the ShiShang Station, the training data comprise cleaned hydrological data, while the test data consist of raw hydrological data. This setup aims to assess whether the prediction control model can effectively detect abnormal data.

Firstly, the experiment predicts the flow at the WaiZhou Station, and the results are shown in Figure 6. It can be observed that the SVM model has significant fluctuations and intense shaking. The LSTM model, compared to the SVM model, shows considerable improvement, but there is still an issue with the peak values not reaching high points, and the prediction line is not sufficiently fitted. PT-LSTM improves upon the performance of LSTM, with most of the fitted parts of the peak values achieving a reasonably good effect. Constructing a confidence interval using such prediction results can completely envelop the true values. The OCPT-LSTM model, built on the basis of PT-LSTM, undergoes combined parameter optimization, resulting in more accurate predictions, especially fitting

well during peak rise and fall. The specific evaluation metrics for the prediction results are shown in Table 1.

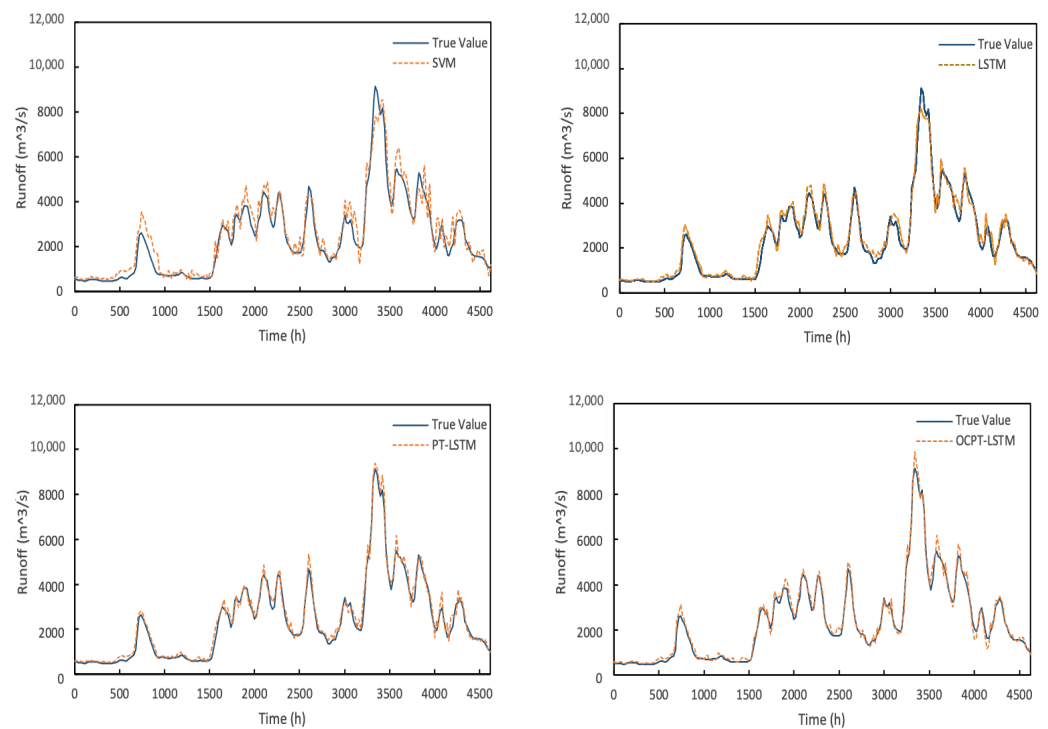


Figure 6. Comparison of flow prediction results of WaiZhou station in different models.

Table 1 presents a comparative evaluation of prediction results for different models. It can be observed that the SVM model performs the worst, with a low coefficient of determination, high root mean square error, and significantly larger maximum error for peak values compared to other models. The LSTM model, as an improved recurrent neural network model, shows much higher accuracy than the SVM model, with reduced root mean square error and maximum error. Both the PT-LSTM and OCPT-LSTM models are iteratively improved based on the basic LSTM model, resulting in better performance than the LSTM model. Among them, the OCPT-LSTM model exhibits the best performance with the fewest errors and more accurate predictions. As the OCPT-LSTM model combines and optimizes the PT-LSTM model, although it increases complexity, it also simultaneously improves accuracy. Both models, compared to other models, effectively enhance prediction performance.

Table 1. Comparison of WaiZhou station prediction results with different models.

Comparison Model	Evaluation Index	Coefficient of Certainty (R^2)	Root Mean Square Error (RMSE)/ m^3	Maximum Error (ME)/ m^3
SVM		0.940	213.12	1370
LSTM		0.974	106.85	1000
PT-LSTM		0.981	98.23	771
OCPT-LSTM		0.989	87.11	675

The construction results of the confidence intervals for the predictive control models are presented below. Figure 7 illustrates the confidence interval for PT-LSTM, while Figure 8 displays the confidence interval for OCPT-LSTM.

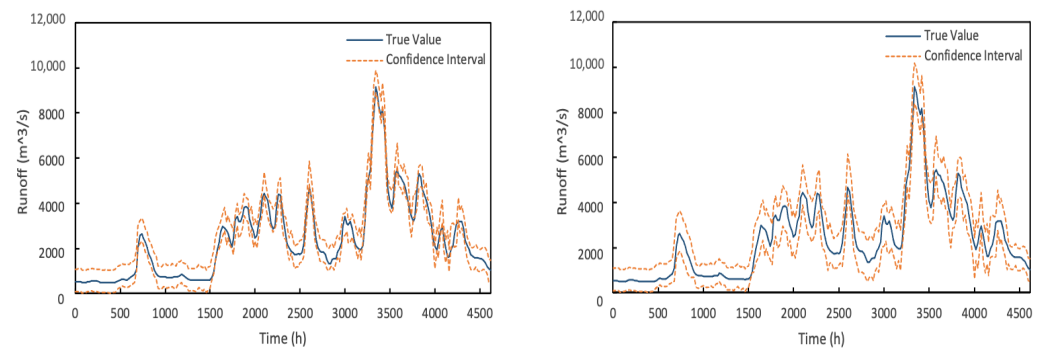


Figure 7. Confidence interval of PT-LSTM model before and after adjustment.

The first graph represents the conventional confidence interval for PT-LSTM, where the confidence interval coefficients are consistent. It can be observed that in situations where the flow rate experiences rapid increases and decreases, the measured values are prone to fall outside the confidence interval, leading to potential deviations in confidence interval control. Therefore, for parts of the flow with significant variations, the confidence interval is dynamically adjusted, as shown in the second graph. The adjusted confidence interval in the graph can easily envelop the measured values. Data falling between these adjusted confidence intervals are considered suspicious and require inspection for potential replacement.

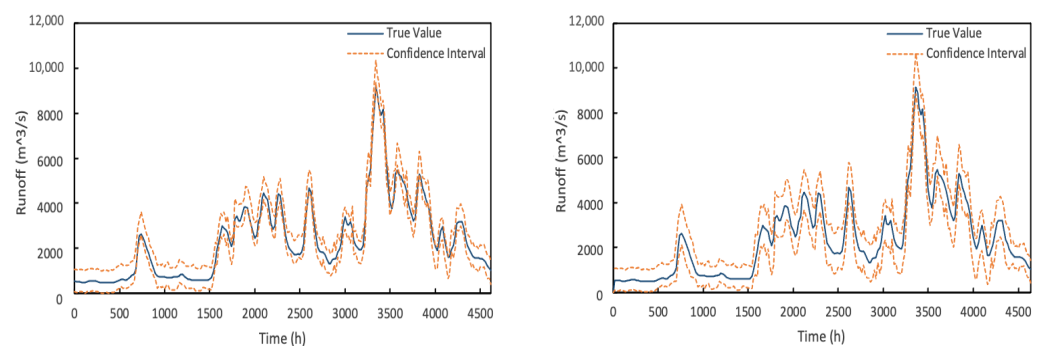


Figure 8. Confidence interval of OCPT-LSTM model before and after adjustment.

Similar to Figure 7, the two graphs in Figure 8 depict one with a conventional confidence interval and the other with an adjusted confidence interval. The first graph represents the conventional confidence interval for OCPT-LSTM. It can be observed that, similarly to the situation in the PT-LSTM case, in instances of large variations in flow rates, measured values tend to fall outside the confidence interval, leading to potential deviations in confidence interval control. Therefore, the confidence interval is adjusted for this portion, as shown in the second graph. In this graph, all measured values fall within the adjusted confidence interval. Data between these adjusted confidence intervals are considered suspicious and require inspection for potential replacement. After the establishment of the OCPT-LSTM model, it needs to be updated periodically, and the worst-performing base model should be replaced to ensure the timeliness of the model.

From the Table 2, it can be observed that the training time of the PT-LSTM model is approximately halved compared to LSTM, while the OCPT-LSTM model is the most complex and requires the longest training time. PT-LSTM simplifies the input based on LSTM, improving model efficiency and reducing training time. The OCPT-LSTM model involves the combination and continuous optimization replacement of PT-LSTM, resulting in longer training times and higher complexity. Therefore, when only detecting suspicious data or pursuing efficiency, the PT-LSTM model can be chosen. When accuracy and

the provision of replacement values are crucial, the OCPT-LSTM model is more suitable. Both models can independently construct confidence intervals and, in the end, build a comprehensive confidence interval for data control. Additionally, the two models can validate each other, with data corrected by the Periodic Temporal LSTM (PT-LSTM) model being confirmed through the secondary quality assurance provided by the OCPT-LSTM combined model in online predictive control methods.

Table 2. Comparison of PT-LSTM and OCPT-LSTM training time.

Model	LSTM	PT-LSTM	OCPT-LSTM
Training time (s)	543	180	1298

Next, the experiment conducts quality control on the data from the ShiShang station. Confidence intervals are constructed separately for the two models, and then the two confidence intervals are combined as follows: $Z = [(y_1 - \delta_1 \cdot mse_1) + (y_2 - \delta_2 \cdot mse_2), (y_1 + \delta_1 \cdot mse_1) + (y_2 + \delta_2 \cdot mse_2)]/2$. This allows for the simultaneous evaluation of the usability of the two confidence intervals. After obtaining the confidence intervals, adjustments are made to the intervals in areas with significant changes in flow. Finally, using the confidence intervals, data are controlled, and values outside the intervals are identified as suspicious. The predicted values obtained from the models are suggested values; hydrological personnel decide whether or not to replace them. The results of the confidence interval construction are shown in Figure 9. It is evident that some data points fall outside the confidence intervals. These data points are marked with red asterisks.

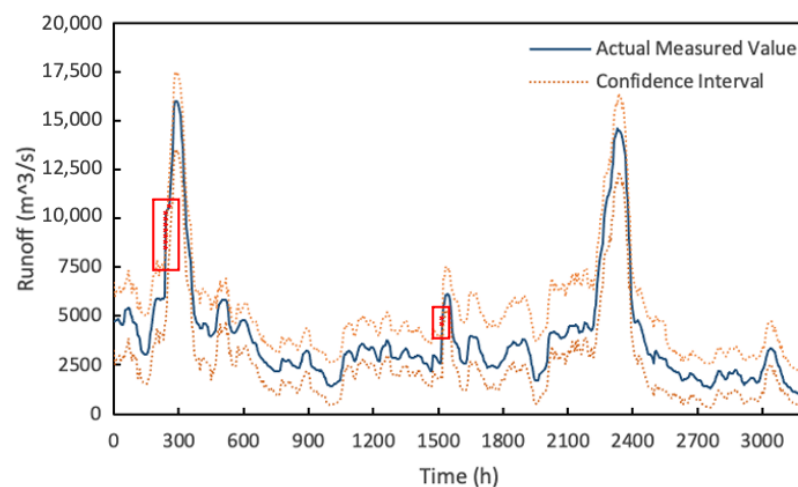


Figure 9. Abnormal data detected by combined confidence interval.

The data points in the red box in the figure represents the detected abnormal data. For the data points marked as anomalies, a search was conducted in the corresponding datasets. The identified time periods are from 2:00 a.m. to 10:00 p.m. on 8 March 2008, and from 6:00 p.m. to 10:00 p.m. on 12 May 2008. A comparison between the unprocessed data and processed data for these time periods is presented in Table 3. It also includes the interpolation effect of the Linear Interpolation method for comparison.

From the above table, it can be observed that the data detected by the confidence interval indeed exhibits some deviation from the cleaned data. Moreover, the predicted values are generally closer to the values of the clean data compared to the unprocessed data, resulting in overall smaller differences. Table 4 compares several sets of Absolute Error, which can more intuitively show that the suggested values provided by OCPT-LSTM are closer to clean data.

Table 3. Comparison of detected abnormal runoff data.

Date	Processed Data/m ³	Unprocessed Data/m ³	Estimated Value of OCPT-LSTM/m ³	Estimated Value of Linear Interpolation/m ³
8/3/2008 2:00	7050	10,200	6802	6236
8/3/2008 3:00	7103	10,209	6879	6442
8/3/2008 4:00	7159	10,227	6970	6648
8/3/2008 5:00	7213	10,248	7026	6735
8/3/2008 6:00	7288	10,263	7110	6964
8/3/2008 7:00	7322	10,280	7186	7147
8/3/2008 8:00	7386	10,296	7263	7249
8/3/2008 9:00	7442	10,312	7339	7556
8/3/2008 10:00	7498	10,328	7415	7693
8/3/2008 11:00	7554	10,345	7492	7875
8/3/2008 12:00	7611	10,361	7568	8058
8/3/2008 13:00	7649	10,378	7644	8240
8/3/2008 14:00	7790	10,400	7705	8422
8/3/2008 15:00	7970	10,411	7812	8604
8/3/2008 16:00	8150	10,419	7934	8786
8/3/2008 17:00	8354	10,429	8197	8969
8/3/2008 18:00	8542	10,439	8356	9151
8/3/2008 19:00	8734	10,448	8507	9333
8/3/2008 20:00	8926	10,450	8676	9515
8/3/2008 21:00	9118	10,461	8844	9697
8/3/2008 22:00	9310	10,469	9013	9880
12/5/2008 18:00	2583	3600	2733	2894
12/5/2008 19:00	2578	3850	2697	3188
12/5/2008 20:00	2570	4380	2592	3482
12/5/2008 21:00	2650	4670	2980	3776
12/5/2008 22:00	2770	4960	3200	4070

Table 4. Comparison of Absolute Error in detected abnormal runoff data.

Date	Processed Data/m ³	Absolute Error of Unprocessed Data/m ³	Absolute Error of OCPT-LSTM/m ³	Absolute Error of Linear Interpolation/m ³
8/3/2008 2:00	7050	3150	248	814
8/3/2008 3:00	7103	3106	224	661
8/3/2008 4:00	7159	3068	189	511
8/3/2008 5:00	7213	3035	187	478
8/3/2008 6:00	7288	2975	178	324
8/3/2008 7:00	7322	2958	136	175
8/3/2008 8:00	7386	2910	123	137
8/3/2008 9:00	7442	2870	103	114
8/3/2008 10:00	7498	2830	83	195
8/3/2008 11:00	7554	2791	63	321
8/3/2008 12:00	7611	2750	43	447
8/3/2008 13:00	7649	2729	5	591
8/3/2008 14:00	7790	2610	85	632
8/3/2008 15:00	7970	2441	158	634
8/3/2008 16:00	8150	2269	216	636
8/3/2008 17:00	8354	2075	157	615
8/3/2008 18:00	8542	1897	186	609
8/3/2008 19:00	8734	1714	227	599
8/3/2008 20:00	8926	1524	250	589
8/3/2008 21:00	9118	1343	274	579
8/3/2008 22:00	9310	1159	297	570

Table 4. Cont.

Date	Processed Data/m ³	Absolute Error of Unprocessed Data/m ³	Absolute Error of OCPT-LSTM/m ³	Absolute Error of Linear Interpolation/m ³
12/5/2008 18:00	2583	1017	150	311
12/5/2008 19:00	2578	1272	119	610
12/5/2008 20:00	2570	1810	22	912
12/5/2008 21:00	2650	2020	330	1126
12/5/2008 22:00	2770	2190	430	1300

5. Conclusions

This paper presents a method for online hydrological data quality control based on the characteristics of hydrological data and an adaptive differential evolution algorithm. Combining the principles of predictive control, a single predictive control model, PT-LSTM, is constructed by considering temporal features and spatial proximity at different time scales. Building upon PT-LSTM, an online optimization combination control model, OCPT-LSTM, is proposed to perform real-time rolling optimization and feedback correction. The introduced data quality control method reduces reliance on manual intervention, provides imputed values for missing data, effectively detects suspicious data, and offers suggested values. This approach contributes to enhancing the reliability of hydrological data analysis. In this study, it was observed that, besides anomaly detection, imputing missing data is another noteworthy issue. Currently, commonly used methods for missing data imputation include linear interpolation and mean imputation. However, these methods have accuracy limitations, prompting the consideration of further research into missing value imputation in future work.

Author Contributions: Conceptualization, Q.Z.; data curation, Y.Z.; formal analysis, X.Z. and S.C.; funding acquisition, Q.Z. and S.C.; investigation, Q.Z. and R.L.; methodology, Q.Z.; project administration, Y.Z. and R.L.; resources, Y.Z.; software, Q.Z.; supervision, Y.Z., X.Z. and S.C.; validation, Q.Z.; visualization, Q.Z. and X.Z.; writing—original draft, Q.Z.; and writing—review and editing, X.Z. and S.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been partially supported by the Talent Startup project of Nanjing Institute of Technology (No. YKJ202117), the Jiangsu Provincial Department of Education's University Philosophy and Social Science Research Project (No. 2023SJYB0434), and the Talent Startup project of Nanjing Institute of Technology (No. YKJ202316).

Data Availability Statement: The datasets presented in this article are not available due to the confidentiality of the data provided by the relevant hydrological department.

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

References

1. Soni, S.; Singh, A. Improving Data Quality using Big Data Framework: A Proposed Approach. *IOP Conf. Ser. Mater. Sci. Eng.* **2021**, *12*, 012092. [\[CrossRef\]](#)
2. Shanmugam, D.B.; Dhilipan, J.; Vignesh, A.; Prabhu, T. Challenges in Data Quality and Complexity of Managing Data Quality Assessment in Big Data. *Int. J. Recent Technol. Eng. (IJRTE)* **2020**, *9*, 589–593. [\[CrossRef\]](#)
3. Gou, J.; Miao, C.; Samaniego, L.; Xiao, M.; Wu, J.; Guo, X. CNRD v1.0: A High-Quality Natural Runoff Dataset for Hydrological and Climate Studies in China. *Bull. Am. Meteorol. Soc.* **2021**, *102*, 929–947. [\[CrossRef\]](#)
4. Veeraswamy, G.; Nagaraju, A.; Balaji, E.; Sreedhar, Y.; Narasimhlu, K.; Harish, P. Data sets on spatial analysis of hydro geochemistry of Gudur area, SPSR Nellore district by using inverse distance weighted method in Arc GIS 10.1. *Data Brief* **2019**, *22*, 1003–1011.
5. Yang, H. Random distributional response model based on spline method. *J. Stat. Plan. Inference* **2020**, *207*, 27–44. [\[CrossRef\]](#)
6. Mousa, M.; Kkhrajan, M.; Merzah, H.Y. Construct Polynomial of Degree n by Using Repeated Linear Interpolation. *IOP Conf. Ser. Mater. Sci. Eng.* **2020**, *928*, 042009.
7. Huang, G. Missing data filling method based on linear interpolation and lightgbm. *J. Phys. Conf. Ser.* **2021**, *1754*, 012187. [\[CrossRef\]](#)

8. Sciuto, G.; Bonaccorso, B.; Cancelliere, A.; Rossi, G. Probabilistic quality control of daily temperature data. *Int. J. Climatol.* **2013**, *33*, 1211–1227. [[CrossRef](#)]
9. Steinacker, R.; Mayer, D.; Steiner, A. Data Quality Control Based on Self-Consistency. *Mon. Weather Rev.* **2011**, *139*, 3974–3991. [[CrossRef](#)]
10. Sciuto, G.; Bonaccorso, B.; Cancelliere, A.; Rossi, G. Quality control of daily rainfall data with neural networks. *J. Hydrol.* **2008**, *364*, 13–22. [[CrossRef](#)]
11. Abbot, J.; Marohasy, J. Application of artificial neural networks to rainfall forecasting in Queensland, Australia. *Adv. Atmos. Sci.* **2012**, *29*, 717–730. [[CrossRef](#)]
12. Fu, F.; Luo, X. Study on Quality Control Method of Hydrological Data. *Water Resour. Informatiz.* **2012**, *5*, 12–15+19.
13. Yu, Y.; Wang, D. An Application Research of Benford's Law in Hydrological Data Quality Mining. *Microelectron. Comput.* **2011**, *8*, 180–183+186.
14. Ding, X.; Wang, H.; Zhang, X.; Li, J.; Gao, H. Association relationships study of multi-dimensional data quality. *J. Softw.* **2016**, *27*, 1626–1644.
15. Juliusz, L.; Kulikowski. Data Quality Assessment: Problems and Methods. *Int. J. Organ. Collect. Intell. (IJOICI)* **2014**, *4*, 24–36.
16. Reynolds, M.; Bourke, A.; Dreyer, N. Considerations when evaluating real-world data quality in the context of fitness for purpose. *Pharmacoepidemiol. Drug Saf.* **2020**, *29*, 1316–1318. [[CrossRef](#)]
17. Yu, Y.; Zhang, J.; Zhu, Y.; Wang, D. Data Quality Control and Management for Hydrological Database. *J. China Hydrol.* **2013**, *3*, 65–68.
18. Tang, D.; Cheng, X.; Wang, D.; Zhu, Y. Research and Application of Continuous Hydrologic Data Quality Control. *Inf. Technol.* **2017**, *4*, 8–12+16.
19. Zhao, Q.; Zhu, Y.; Wan, D.; Yu, Y.; Cheng, X. Research on the Data-Driven quality control method of hydrological time series data. *Water* **2018**, *10*, 1712. [[CrossRef](#)]
20. Binqun, L.; Zhongmin, L.; Qingrui, C.; Wei, Z.; Huan, W.; Jun, W.; Yiming, H. On the Operational Flood Forecasting Practices Using Low-Quality Data Input of a Distributed Hydrological Model. *Sustainability* **2020**, *12*, 8268. [[CrossRef](#)]
21. Yu, Y.; Wan, D.; Zhao, Q.; Liu, H. Detecting Pattern Anomalies in Hydrological Time Series with Weighted Probabilistic Suffix Trees. *Water* **2020**, *12*, 1464. [[CrossRef](#)]
22. Lattawit, K.; Chantana, C.; Montri, M.; Wongchaisuwat, P.; Wimala, S.; Sarinnapakorn, K.; Boonya-aroonnet, S. Anomaly Detection Using a Sliding Window Technique and Data Imputation with Machine Learning for Hydrological Time Series. *Water* **2021**, *13*, 1862. [[CrossRef](#)]
23. Hyojoong, K.; Heeyoung, K. Contextual anomaly detection for high-dimensional data using Dirichlet process variational autoencoder. *IIEE Trans.* **2023**, *55*, 433–444.
24. Hyojoong, K.; Heeyoung, K. Contextual anomaly detection for multivariate time series data. *Qual. Eng.* **2023**, *35*, 686–695.
25. Li, D.; Cui, S.; Li, Y.; Xu, J.; Xiao, F.; Xu, S. Pad: Towards principled adversarial malware detection against evasion attacks. *IEEE Trans. Dependable Secur. Comput.* **2023**, *21*, 920–936. [[CrossRef](#)]
26. Cui, S.; Li, T.; Chen, S.C.; Shyu, M.L.; Li, Q.; Zhang, H. DISL: deep isomorphic substructure learning for network representations. *Knowl. Based Syst.* **2020**, *189*, 105086. [[CrossRef](#)]
27. Liu, X.; Ju, X.; Fan, S. A Research on the Applicability of Spatial Regression Test in Meteorological Datasets. *J. Appl. Meteor. Sci.* **2006**, *17*, 37–43.
28. Olivier, C.; Patrick, H.; Vladimir, V. Support vector machines for histogram-based image classification. *IEEE Trans. Neural Netw.* **1999**, *10*, 1055–1064.
29. Brown, M.; Steve, R.G.; Hugh, G. Support vector machines for optimal classification and spectral unmixing. *Ecol. Model.* **1999**, *20*, 167–179. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.