Abstract: Water level prediction in large dammed rivers is an important task for flood control, hydropower generation, and ecological protection. The variations of water levels in large rivers are traditionally simulated based on hydrological models. Recently, most studies have begun applying deep learning (DL) models as an alternative method for forecasting the dynamics of water levels. However, it is still challenging to directly apply DL to the simultaneous prediction of water levels across multiple sites. This study attempts to develop a hybrid framework by combining the Physical-based Hydrological model (PHM) and Long Short-Term Memory (LSTM). This study hypothesizes that our hybrid model can enhance the predictive accuracy of water levels in large rivers, because it considers the temporal-spatial information of mainstream-tributaries relationships. The effectiveness of the proposed model (PHM-BP-LSTM) is evaluated using the daily water levels from 2012 to 2018 in the Three Gorges Reservoir (TGR), China. Firstly, we use a hydrological model to produce a large amount of water level data to solve the limited training data set. Then, we use the Back Propagation (BP) neural network to capture the mainstream-tributaries relationship. The future changes in water levels in the different mainstream stations are simultaneously predicted by the LSTM model. We reveal that our hybrid model yields satisfactory accuracy for daily water level simulations at fourteen mainstream stations of the TGR. We further demonstrate the proposed model outperforms the traditional machine learning methods in different prediction scenarios (one-day-ahead, three-day-ahead, seven-day-ahead), with RMSE values ranging from 0.793 m to 1.918 m, MAE values ranging from 0.489 m to 1.321 m, and the average relative errors at each mainstream station are controlled below 4%. Overall, our PHM-BP-LSTM, combining physical process and deep learning, can be viewed as a potentially useful approach for water level prediction in the TGR, and possibly for the rapid forecast of changes in water levels in other large rivers.

Keywords: water level prediction; Three Gorges Reservoir; deep learning; hybrid model

1. Introduction

Large-scale hydraulic engineering projects on rivers have brought considerable changes to the water environment and have affected the utilization and protection of water resources [1]. Hydraulic engineering projects can regulate water levels and improve the utilization efficiency and security of water resources. Still, they also change the hydrological characteristics, water quality, and aquatic ecosystems of rivers [2,3]. To balance out the
benefits and costs of artificial dams, it is necessary to reasonably regulate the water levels; that is, to plan the storage and release of water according to different objectives: such as power generation, flood control, water supply, irrigation, ecological protection, etc. [4], to achieve optimal effects. For example, in the Adda River basin in Italy, Moisello et al., investigated the effects of different man-made basin changes on water resources and highlighted how the water resources management of a basin must reconcile different needs [5]. To realize effective water level regulation, it is necessary to accurately monitor and predict the artificial dams and their surrounding environment, and take corresponding measures. Among them, water level prediction is a critical technology that can provide a scientific basis and decision support for water level regulation. At present, the commonly used water level prediction methods can be divided into two categories: mechanism models and data-driven models.

Traditionally, water level prediction relies on mechanism models that simulate the physical processes of water movement in a river system. Mechanism models are based on hydrological principles and simulate river hydrological processes by solving equations and boundary conditions [6]. These models have physical solid meaning and universality, but also have some disadvantages. For instance, it is hard to determine accurate boundary conditions due to various factors affecting river hydrological processes (complex boundary conditions). The input data (such as rainfall, flow rate, etc.) is uncertain and random, and may have missing data (variable inputs). The river needs to be divided into smaller units and consider their interaction, which leads to large amounts of calculation (large number of calculation units). The calculation speed is slow because of complex equations, boundary conditions, and many iterative calculations (long calculation time) [7,8].

In recent years, the data-driven model has emerged as an alternative approach for hydrological calculations. These models use historical data to establish empirical relationships between input and output variables, without requiring detailed knowledge of the underlying physical mechanism. Data-driven models can overcome some limitations of hydrological models and achieve high prediction accuracy and efficiency. Various machine learning methods have been widely used in water level prediction tasks, including single models such as the Autoregressive Integrated Moving Average model (ARIMA) [9], Genetic Programming (GP) [10], Support Vector Machines (SVM) [11], and tree-based models [12], as well as hybrid models such as the hybrid Extreme Learning Machine combined with hybrid Particle Swarm Optimization and Grey Wolf Optimization (ELM-PSOGWO) [13], the hybrid support vector regression with the simulated annealing algorithm, and the mayfly optimization algorithm (SVR-SAMOA) [14]. Among them, Artificial Neural Networks (ANNs) are always suitable for water level prediction because they can learn from data and capture complex non-linear patterns [15,16]. However, conventional ANNs have some drawbacks, such as overfitting, local minima, and a black-box nature [17].

Deep learning (DL) techniques have been proposed to address these issues as an advanced extension of ANNs. DL techniques can construct multiple hidden layers with different activation functions and learning algorithms, enhancing the models’ representation and generalization abilities. Recently, some DL techniques have been successfully applied to water level forecasting problems, including Convolutional Neural Networks (CNN) [18,19], Long Short-Term Memory (LSTM) [20], Long Short-Term Memory-weighted mean of vectors optimizer (LSTM-INFO) [21], and CNN-LSTM [22,23], etc. However, the DL model is often limited due to the scarcity of monitoring data. Hybrid models that combine the hydrological process with DL have garnered widespread attention in the field of water resource management. These models capitalize on the unique strengths of traditional physical models and deep learning methods by combining them, thus overcoming the limitations of using a single approach and enhancing predictive capabilities. For instance, Yang et al., (2020) integrated ANN computer vision methods with hydrological models, resulting in the increased precision of river runoff simulations [24]; Li et al., (2023) effectively improved flood forecasting accuracy by employing a method that combined LSTM and hydrological models [25]. However, current research efforts mainly focus on
water level prediction at a few specific sites or particular regions [26]. They only use the historical data of the target site to train and test the model, without considering the spatial and temporal correlation among different areas. Consequently, it is difficult to directly apply them to the simultaneous prediction of water levels across multiple sites in large rivers, typically for the Yangtze River (China).

Water level predictions in large rivers need to consider the spatial relationships between the mainstream and its tributaries. The tributaries’ water level changes affect the mainstream’s water level changes, and vice versa. Using this relationship, multiple stations’ water level data can be used as input variables to improve the prediction accuracy of target stations. For instance, some researchers try to predict the water level of multiple sites based on the physical-based hydrological model (PHM) [27,28]. With respect to the data-driven models, Li et al., (2016) used Random Forests (RF) to predict the water level of Poyang Lake in China, and they found that the prediction accuracy was improved when considering the tributary’s relationship to the lake [29]. Similarly, Pan et al., (2020) used GRU to understand the changing trend in water level and the CNN to understand the spatial correlation among water level data observed from adjacent water stations to predict the multi-station water level [30]. These studies can prove that considering the mainstream-tributaries relationship can improve the predictive performance of the purely data-driven model. However, these studies are still limited by the availability and quality of historical data, which may affect the reliability and generalization of the models.

In this study, we develop a hybrid hydrological model combining a physical-based model and deep learning models for daily water level prediction. The effectiveness of the proposed model is validated through comparative work based on monitoring data from the Three Gorges Reservoir (TGR), China. The attributions of our study lie in: (1) using the PHM to provide sufficient samples for training the subsequent DL model; (2) using a back propagation neural network (BP) to capture the hidden and inherent mainstream-tributaries relationship; (3) using an LSTM model to simultaneously predict water level changes across multiple stations.

2. Materials and Methods
2.1. Study Sites

With population growth and economic development, the water resources in the Yangtze River (China) are under increasing pressure, making it extremely important to predict water levels in the basin. The Three Gorges Dam (TGD), shown in Figure 1, is one of the largest projects in China, and provides enormous socioeconomic benefits, including power generation, flood control, and shipping [31]. Significant impacts on the local hydrological environment have been observed with the construction and operation of the Three Gorges Dam on the Yangtze River. According to relevant survey data, the dam has reduced downstream flood peaks, increased water volume during dry seasons, and improved water quality. However, the dam has also had an impact on the ecological environment. The dam reservoir has widened the water surface and increased humidity, lowering the temperature in the Three Gorges reservoir area, thus affecting the local climate [32].

Additionally, the dam has affected aquatic plants and animals. Since the construction of the dam, many tributaries and rivers have been blocked, resulting in changes in the living environment of some species. Some unique aquatic plants and animals have even disappeared. To better protect the local ecological environment and water resources, accurate prediction of the water level in the Three Gorges Basin is essential to take timely measures and ensure local ecological and economic development.
\[ \frac{dQ}{dt} + a \frac{\partial Q}{\partial x} - q = 0 \]

\[ \frac{\partial A}{\partial t} + \frac{\partial Q}{\partial x} = 0 \]

\[ \frac{\partial A}{\partial t} + \alpha_0 \frac{\partial Q}{\partial x} - q = 0 \]

\[ \frac{\partial A}{\partial t} + \alpha_0 \frac{\partial Q}{\partial x} + \left( gA - B \left( \frac{Q}{A} \right)^2 \right) \frac{\partial^2 z}{\partial x^2} + \left( \frac{Q}{A} \right)^2 \frac{\partial A}{\partial x} = 0 \]
2.3. Deep Learning Model

This study established a PHM-BP-LSTM model to forecast the water level in the mainstream of the Yangtze River Three Gorges reservoir basin. The overall development process of the model is presented in Figure 2. Firstly, we established the PHM mechanism model to simulate the water level data of the mainstream stations, which could provide a large amount of simulated water level data for the subsequent deep learning model, thus solving the problem of insufficient data. Subsequently, the BP neural network model was constructed based on historical data from known tributary stations and simulated data from the mechanism model. This model predicted the historical water level data of the mainstream stations, utilizing simulated data from the mechanism model for model training and validation. Next, the LSTM model was established for time series forecasting of the water level at the mainstream stations using the prediction results from the BP network. Simulated data from the mechanism model were utilized for model testing. Finally, the established models were individually evaluated. Using a hybrid model reduced the data requirements for training a deep learning model, as it can use simulated data from the PHM as input. This can address issues with limited or missing historical data, which is often a challenge for purely data-driven models. To facilitate the training of all models, we performed Z-Score normalization on the input data. We used PyTorch’s deep learning framework and Python’s sklearn library. All models were trained using the Adam optimization algorithm [36].

Figure 2. The building process of the proposed model (PHM-BP-LSTM). The arrows points towards the direction of data generation from PHM and subsequent deep learning analysis.

2.3.1. Back Propagation Neural Network

The Back Propagation (BP) neural network is a widely used neural network model that is trained using the error backpropagation algorithm [37]. It has simple learning algorithms and powerful learning capabilities, which enable it to learn and store a great deal of nonlinear mapping relations of the input-output model without disclosing the mathematical equation that describes these relations. Therefore, we used it to model the mainstream-tributary relationships. The BP neural network model is a multilayer feedforward network that consists of two main processes: the forward propagation of information and the backward propagation of error [38]. The network comprises three primary layers: the input layer, the hidden layer, and the output layer. Information from external sources is transmitted through the input layer to the network’s hidden layer for processing, and the final result is obtained from the output layer. During training, if the error between the output result of the output layer and the pre-set input value of the BP neural network is
large, the network enters the backpropagation stage and updates its weights until the error between the output and the desired result meets certain conditions.

2.3.2. Long Short-Term Memory

The backpropagation through time algorithm used in traditional Recurrent Neural Network (RNN) models suffers from the problem of gradient dispersion, especially when dealing with long-term data. This issue leads to slow weight updates, resulting in an inability to effectively capture long-term memory in RNNs [36]. The Long Short-Term Memory (LSTM) network model was proposed to address this challenge. It can remember values over arbitrary intervals, making it well-suited to predict time series given time lags of unknown duration. It is relatively insensitive to gap length and can maintain a stable error gradient across long sequences without suffering from gradient dispersion or explosion. The LSTM model is a particular form of RNN that introduces memory blocks to replace hidden neurons for connecting hidden layers [39]. Each memory block comprises a memory cell (C), an input gate (i), a forget gate (f), and an output gate (o). The LSTM model overcomes the gradient exploding and vanishing issues of conventional RNN by controlling the flow of information between memory cells and gates. The LSTM can learn remote dependencies and effectively capture long-term memory in sequential data by updating or removing previously accumulated information. Its calculation formula is as follows.

\[
i_t = \sigma \left(W_i[h_{t-1}, x_t] + b_i\right)
\]

\[
f_t = \sigma \left(W_f[h_{t-1}, x_t] + b_f\right)
\]

\[
o_t = \sigma \left(W_o[h_{t-1}, x_t] + b_o\right)
\]

\[
\tilde{C}_t = \tanh(W_c[h_{t-1}, x_t] + b_c)
\]

\[
\tilde{C}_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t
\]

\[
h_t = o_t \cdot \tanh(C_t)
\]

where \(x_t\) represents the input at moment \(t\), \(h_{t-1}\) represents the hidden state at moment \(t - 1\), that is, the output state at the previous time, \(i_t\) represents the output of the input gate at moment \(t\) to control the influence of the input on the internal memory unit, \(f_t\) represents the output of the forgetting gate at moment \(t\), which controls which information in the memory unit at the previous time needs to be forgotten, \(o_t\) represents the output of the output gate at moment \(t\), and determines which information needs to be output to the state at the next time, \(C_t\) represents the state of the internal memory unit at moment \(t\), storing the long-term memory of the network, \(\tilde{C}_t\) represents the candidate state at moment \(t\) and the update information of the internal state at the current time. \(h_t\) represents the output at moment \(t\), \(W_i, W_f, W_o, W_c\) represent the weight matrix, \(b_i, b_f, b_o, b_c\) are the corresponding biases, \(\sigma(\cdot)\), \(\tanh(\cdot)\) represents the activation function.

2.4. Comparative Modes

2.4.1. Support Vector Regression

The Support Vector Machine (SVM) regression algorithm is called Support Vector Regression (SVR), which converts nonlinear features into linear features and increases dimensionality using kernel functions [40]. SVR is a supervised learning algorithm that follows the same principle as SVM: finding the best-fitting curve or hyperplane. Generally, SVR follows the structural risk minimization (SRM) theory instead of the empirical risk minimization (ERM) employed by most traditional ANNs. SRM aims to decrease the upper
limit of the generalization error, while ERM seeks to reduce the training error. Therefore, the SVR model achieves an optimum network structure and avoids overfitting [41].

2.4.2. Classification and Regression Tree

The Classification and Regression Tree (CART) algorithm partitions a set of samples into two child nodes by identifying one input variable and one break-point [42]. The algorithm begins at the root node, which is the entire set of available training samples. It performs recursive binary partitioning for each node until no further split is possible or a certain termination criterion is satisfied. The best split is identified at each node by an exhaustive search, testing all potential splits on each input variable and break-point. The split corresponding to the minimum deviations is selected by predicting two child nodes of samples with their mean output variables. Typically, an overly large tree is constructed, and pruning is employed to sequentially remove the splits that insufficiently contribute to training accuracy. After constructing a tree, an inquiry sample is assigned to one of the terminal leaves (non-splitting leaf nodes) and is then predicted with the mean output value of the samples belonging to the leaf node [43]. The CART algorithm’s simple structure and good interpretability have made it widely used in practice.

2.5. Model Evaluation Index

To measure the degree of fit between predicted and observed values and to evaluate the performance of the model, the root mean square error (RMSE), mean absolute error (MAE), and goodness of fit ($R^2$) of the model were calculated.

\[
RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}
\]

\[
MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|
\]

\[
R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}
\]

where, $n$ is the number of data, and $y$, $\hat{y}$, $\bar{y}$ are the observed data, predicted data, and mean observed data, respectively.

Model accuracy is measured by the MAE and RMSE, which range between $-\infty$ and $+\infty$ and between 0 and $+\infty$, respectively, and have an ideal value of 0 [44]. The goodness of fit refers to the degree to which the regression line fits the observed values [45]. The statistic that measures goodness of fit is the coefficient of determination $R^2$, which ranges between 0 and 1.

3. Results and Discussion

In this section, we first present the simulation results of the PHM mechanism model. Next, we present and discuss the water level prediction results of the BP neural network and LSTM models. Finally, we compare the performance with other machine learning techniques (SVR and CART models) for different prediction scenarios (T + 1 step, T + 3 step, T + 7 step). We also analyze the advantages and limitations of these techniques in water level prediction.

3.1. Building the Connections between Mainstream and Its Tributaries in TGR

We used the BP neural network to model the mainstream-tributaries relationships in the TGR area. We compared its performance with the mechanism model (PHM) that simulates the water level based on physical equations. Taking the historical hydrological data of the known tributary stations as input and the water level of the mainstream station as the output, we randomly divided the data into a 50% training set and a 50%
validation set to train and validate the BP model. Then, we used RMSE, MAE, and $R^2$ as evaluation criteria.

Figure 3 shows the convergence of the mean square error (MSE) for both the training and validation sets of the mainstream water level prediction model. It can be observed that after 100 epochs of training, the MSE values of the training and validation sets tend to stabilize and converge to nearly 0. This indicates that the BP model can effectively learn the inherent relationship between mainstream and tributary water levels and avoid overfitting or underfitting problems.

![Figure 3](image-url)

**Figure 3.** (a) Loss plot for BP model for predicting water level (T) of mainstream stations. (b–d) Loss plot for LSTM model for predicting multi-step (T + 1, T + 3, T + 7) water level (WL) of mainstream stations.

Table 1 presents the model’s prediction accuracy for each station along the mainstream. As can be seen, the water level prediction model exhibited relatively high accuracy at each mainstream station. The prediction accuracy remained consistently high in both the training and validation sets, with both RMSE and MAE relatively low (on the original data scale), and with an $R^2$ value above 0.9. Within the validation set, the smallest $R^2$ was 0.939 and the largest $R^2$ was 0.999. These results indicate that the BP model can reliably predict the water level of mainstream stations by modeling the mainstream-tributaries relationships.
Table 1. Performance of BP model in forecasting historical water levels of mainstream stations (the location index of mainstream stations is shown as the green dots in Figure 1).

<table>
<thead>
<tr>
<th>Location Index</th>
<th>Training Set</th>
<th>Validation Set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMSE</td>
<td>MAE</td>
</tr>
<tr>
<td>3</td>
<td>0.883</td>
<td>0.414</td>
</tr>
<tr>
<td>5</td>
<td>0.818</td>
<td>0.526</td>
</tr>
<tr>
<td>7</td>
<td>0.355</td>
<td>0.239</td>
</tr>
<tr>
<td>9</td>
<td>0.361</td>
<td>0.271</td>
</tr>
<tr>
<td>11</td>
<td>0.648</td>
<td>0.504</td>
</tr>
<tr>
<td>13</td>
<td>0.751</td>
<td>0.459</td>
</tr>
<tr>
<td>15</td>
<td>1.025</td>
<td>0.686</td>
</tr>
<tr>
<td>17</td>
<td>0.446</td>
<td>0.296</td>
</tr>
<tr>
<td>19</td>
<td>1.453</td>
<td>0.939</td>
</tr>
<tr>
<td>21</td>
<td>0.773</td>
<td>0.48</td>
</tr>
<tr>
<td>23</td>
<td>0.412</td>
<td>0.272</td>
</tr>
<tr>
<td>25</td>
<td>0.232</td>
<td>0.142</td>
</tr>
<tr>
<td>27</td>
<td>0.639</td>
<td>0.392</td>
</tr>
<tr>
<td>29</td>
<td>0.241</td>
<td>0.153</td>
</tr>
</tbody>
</table>

To illustrate the performance of the BP model in predicting the water level changes of the mainstream stations in the TGR area by modeling the mainstream-tributary relationships, we selected the dam-front (No.29 station) water level prediction as an example. As shown in Figure 4, the red line represents the observed values, while the light blue line represents the model-predicted values. The data covers the period from 2012 to 2019, with the first half used for training the model and the last used for validation. The BP model can capture the temporal variation and peak values of the water level very well, and is very close to the observed data. The top-left graph displays the relative error of each sample, which shows that the relative error for both the training and validation sets remained below 1%, indicating that the predictions are highly reliable.

![Figure 4](image-url) Figure 4. The fitting effect of the prediction (predicted data) dam-front (No.29 station) water level and the simulated water level (observed data) of the mechanism model. The inner plot represents the relative error after taking the absolute value.

Traditionally, mechanism models simulate the hydrological changes of the mainstream by using the temporal information of multiple tributaries [46,47]. At the same time, machine learning methods are rarely applied to establish the relationship between the mainstream and tributaries. Our study demonstrates that deep learning models are an effective way to build mainstream-tributaries relationships. Our results are consistent with some previous studies. For example, Lallahem et al., (2005) used the BP neural network to simulate water level changes, finding that the BP neural network had high accuracy and stability [48]. Furthermore, our study shows that the BP model can capture the nonlinear relationship between mainstream and tributary water levels and has high prediction accuracy. However, the BP model has limitations in dealing with time series data. It does not have a memory.
mechanism, which means it cannot capture the long-term dependencies and temporal patterns in the data. Numerous studies reveal that the temporal changes of the mainstream time series require introducing time series forecasting models [49,50].

3.2. Water Level Forecasting Based on the Proposed Model at Different Time Tasks

We applied the LSTM model to predict the time series water level at the mainstream stations. Taking the historical mainstream water level data predicted by the BP neural network as input, and the water level at T + 1, T + 3, or T + 7 steps in each mainstream station as output, we randomly split the data into a 50% training set and a 50% validation set for LSTM training and validation, respectively. Then, we used the data simulated by the mechanism model for each mainstream station to test the LSTM model. Figure 3 shows the convergence of the MSE for the training and validation sets of the LSTM-based mainstream water level time series prediction model (T + 1, T + 3, and T + 7). As shown in Figure 3, after training, the MSE values of both the training and validation sets tended to stabilize, indicating that the LSTM model can effectively predict the time series of water levels and avoid over-fitting or under-fitting problems.

Figure 5 shows the box plots of relative error for each mainstream station at T + 1, T + 3, and T + 7 steps using the LSTM model, which reflect the distribution and degree of dispersion of relative errors between the test set data and the predicted data at each mainstream station. As can be seen, the relative error of all mainstream stations in the time series prediction at T + 1, T + 3, and T + 7 steps is lower than 4%, and the median relative error is lower than 1%. It can also be seen that all mainstream stations have similar relative error ranges at T + 1, T + 3, and T + 7 step predictions, indicating that the LSTM model has relatively average prediction capability for each station in water level prediction tasks at mainstream stations for all steps. Figure 5 also shows that as the prediction steps increase, the relative error ranges also increase but remain at a low level. Therefore, the LSTM model can be considered reliable in time series prediction tasks.

![Figure 5. Box plots for the absolute value of the relative error between the water level predicted by the LSTM multi-step model and the simulated value of the mechanism model](image)

The analysis of the relative error box plots indicates that LSTM has a relatively average prediction capability at each mainstream station. To illustrate the performance of the LSTM model in predicting multi-step (T + 1, T + 3, T + 7) water levels of the mainstream stations in the TGR area, we selected the dam-front (No.29 station) water level prediction as an example. As shown in Figure 6, the red line represents observed values and the light blue line represents predicted values by the LSTM model. The plot in the upper left corner shows the relative error of each sample. The LSTM model can capture temporal variation and peak values of water level very well and is very close to the observed data. Relative errors are smaller for predictions at T + 1, T + 3, and T + 7 steps, indicating reliable predictions and better fitting effects.
The analysis of the relative error box plots indicates that LSTM has a relatively average prediction capability at each mainstream station. To illustrate the performance of the LSTM model in predicting multi-step (T + 1, T + 3, T + 7) water levels of the mainstream stations in the TGR area, we selected the dam-front (No. 29 station) water level prediction as an example. As shown in Figure 6, the red line represents observed values and the light blue line represents predicted values by the LSTM model. The plot in the upper left corner shows the relative error of each sample. The LSTM model can capture temporal variation and peak values of water level very well and is very close to the observed data. Relative errors are smaller for predictions at T + 1, T + 3, and T + 7 steps, indicating reliable predictions and better fitting effects.

Figure 6. The fitting effect of the multi-step (T + 1, T + 3, T + 7) time series prediction (predicted data) water level of the dam-front (No. 29 station) and the simulated water level (observed data) of the mechanism model. The inner plot represents the relative error after taking the absolute value.

Our study confirms that LSTM is an effective way to characterize water level changes, consistent with the findings of Liu et al., (2021), who developed a real-time rolling forecast approach for the short-term water levels of urban inland and external rivers using LSTM, addressing the high uncertainty of river water level prediction in Fuzhou city, China [51]. The results verified the feasibility of LSTM in water level forecasting. In addition, the results show that the PHM-BP-LSTM model can well capture the time-varying and peak water levels, and the model has high accuracy in simulating long-term diurnal water levels. Similarly, the predictive performance of LSTM is affected by the increase in time steps, manifested as a significant increase in model error with the increase in prediction task time. Previous studies have shown that model hyperparameters affect the predictive effect of the model [52]. In this study, we used the Bayesian optimization technique to optimize the hyperparameters [53]. The results of LSTM hyperparameter optimization are shown in Table 2.
Table 2. The hyperparameters of the LSTM model for different prediction steps.

<table>
<thead>
<tr>
<th>Step</th>
<th>Time Lag</th>
<th>Hidden Size</th>
<th>Learning Rate</th>
<th>Num Layers</th>
<th>Batch Size</th>
<th>Activation Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>T + 1</td>
<td>4</td>
<td>32</td>
<td>0.001</td>
<td>1</td>
<td>32</td>
<td>relu</td>
</tr>
<tr>
<td>T + 3</td>
<td>18</td>
<td>32</td>
<td>0.001</td>
<td>1</td>
<td>32</td>
<td>relu</td>
</tr>
<tr>
<td>T + 7</td>
<td>30</td>
<td>32</td>
<td>0.0005</td>
<td>1</td>
<td>32</td>
<td>relu</td>
</tr>
</tbody>
</table>

Notably, our study develops a multi-site collaborative prediction strategy, which models the spatial relationship between mainstream and tributaries using the BP neural network, and feeds the results into the LSTM to capture the temporal dynamics of the water level time series. In this way, the spatio-temporal correlation of water level changes among different stations is established, and finally, the future water level data of multiple stations are simultaneously output, achieving the goal of multi-site collaborative prediction. The results also confirm the effectiveness of this method for large-scale river water level prediction.

3.3. Model Comparisons with Conventional Machine Learning Approaches

Table 3 presents the accuracy statistics for the LSTM, SVR, and CART models. Regarding RMSE and MAE, the LSTM model showed greater accuracy than the SVR or CART models for all time steps. In contrast, the lowest RMSE- and MAE-based errors were achieved by the LSTM model followed by the SVR and CART models. Therefore, the LSTM model outperformed the SVR and CART models. In forecasting the one-day-ahead water level for TGR, the LSTM model (RMSE = 1.054 m, MAE = 0.489 m) showed the best performance among the developed standalone models, while the CART model (RMSE = 0.89 m, MAE = 0.619 m) showed the poorest performance. In forecasting the three-days-ahead and one-week-ahead water level for TGR, the LSTM model (RMSE = 1.282 m, MAE = 0.833 m and RMSE = 1.981 m, MAE = 1.321 m) demonstrated the best performance among the developed models, while the SVR model (RMSE = 1.39 m, MAE = 0.981 m and RMSE = 2.136 m, MAE = 1.53 m) showed the poorest performance.

Table 3. Performance of the models in multi-step time series forecasting water levels of mainstream stations. A represents the increase in RMSE and MAE of SVR and CART compared to LSTM, and— represents the performance decline of SVR and CART compared with LSTM.

<table>
<thead>
<tr>
<th>Step</th>
<th>Model</th>
<th>Training Set</th>
<th>Validation Set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>RMSE</td>
<td>MAE</td>
</tr>
<tr>
<td>T + 1</td>
<td>LSTM</td>
<td>0.717</td>
<td>0.412</td>
</tr>
<tr>
<td></td>
<td>SVR</td>
<td>0.783 (−Δ9.2%)</td>
<td>0.524 (−Δ27.2%)</td>
</tr>
<tr>
<td></td>
<td>CART</td>
<td>0.764 (−Δ6.6%)</td>
<td>0.503 (−Δ22.1%)</td>
</tr>
<tr>
<td>T + 3</td>
<td>LSTM</td>
<td>1.23</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>SVR</td>
<td>1.377 (−Δ11.9%)</td>
<td>0.923 (−Δ30%)</td>
</tr>
<tr>
<td></td>
<td>CART</td>
<td>1.237 (−Δ0.6%)</td>
<td>0.781 (−Δ10%)</td>
</tr>
<tr>
<td>T + 7</td>
<td>LSTM</td>
<td>1.909</td>
<td>1.168</td>
</tr>
<tr>
<td></td>
<td>SVR</td>
<td>2.132 (−Δ11.7%)</td>
<td>1.443 (−Δ23.5%)</td>
</tr>
<tr>
<td></td>
<td>CART</td>
<td>2.008 (−Δ5.2%)</td>
<td>1.333 (−Δ14.1%)</td>
</tr>
</tbody>
</table>

Our research results show that LSTM has certain advantages over traditional machine learning models in multi-site and multi-step simultaneous prediction tasks. Similarly, the prediction performance of LSTM, SVR, and CART will also be affected by the increase in time steps, which shows that the model error significantly increases with the increase in the prediction task time, but compared with SVR and CART, the prediction performance of LSTM is always the best. This shows that the PHM-BP-LSTM model proposed in this paper has superior performance in predicting the water level of mainstream stations in the TGR area.
However, in machine learning, different research perspectives and methods have limitations, and our study is no exception. First, our study only used a single feature (i.e., water level) as input data without considering other factors that may affect water level changes, such as rainfall, evaporation, temperature, etc. This may cause the model to ignore some important information or have some biases [54,55]. Second, due to the length of the paper, our study focuses on the difference between traditional machine learning and LSTM methods in water level prediction and does not compare and analyze different deep learning models, which have also been reported in some recent studies [56–58]. Furthermore, the key challenge in this paper is how to interpret the processes behind the predictions made by hybrid models. While hybrid models may demonstrate higher accuracy and efficiency in predicting water level fluctuations, they might also exhibit lower interpretability in revealing underlying causal relationships. To address these issues, in future studies, we plan to improve and extend our work from the following aspects:

1. Introduce multi-feature data and construct water level prediction models based on multi-input multi-output or multi-task learning techniques;
2. Compare different deep learning techniques for water level prediction tasks, and design more suitable deep learning structures for hydrological data characteristics and patterns.
3. Investigate the interpretability issues of hybrid models in predicting water level fluctuations in order to reveal underlying causal relationships more effectively.

4. Conclusions

We proposed a hybrid hydrological model (PHM-BP-LSTM) in forecasting the daily water level of the TGR. Firstly, we used the physical-based hydrological model to simulate the water level of 14 stations in the mainstream. Then, the BP neural network model was constructed based on historical data from known tributary stations and simulated data from the mechanism model. Finally, the LSTM model could effectively predict water levels by using historical water level data predicted by BP neural network as input without requiring boundary conditions and operation rules. The results show that our PHM-BP-LSTM model achieved high prediction accuracy and stability in different prediction scenarios (one-day-ahead, three-days-ahead, seven-days-ahead) at 14 mainstream stations, with RMSE values ranging from 0.793 m to 1.918 m, MAE values ranging from 0.489 m to 1.321 m, and the average relative errors at each mainstream station were controlled below 4% in all three forecasting scenarios. The PHM-BP-LSTM model outperformed other machine learning models (SVR and CART) regarding the RMSE and MAE values in all-time series prediction scenarios at all mainstream stations. The PHM-BP-LSTM model could effectively capture the nonlinear and complex relationship between the mainstream and tributary water levels, as well as the temporal dynamics of water level changes. The developed multi-site collaborative forecasting strategy could simultaneously forecast multiple sites along the mainstream of the TGR area. This strategy can effectively utilize the spatio-temporal information of water level data at different locations to improve the prediction performance of large-scale river systems.

Author Contributions: Methodology, M.X., S.Z. and K.S.; validation, M.X., L.W. and B.Y.; investigation, M.X., X.W. and Z.G.; data curation, M.X., S.Z. and K.S.; writing—original draft preparation, M.X., L.W., X.W., B.Y., Z.G. and K.S.; writing—review and editing, M.X., K.S. and M.S.; visualization, M.X.; supervision, K.S. and M.S.; funding acquisition, K.S. and M.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (No. 52379081; 62072429), Chongqing Science and Technology Commission (CSTB2022TIAD-KPX0199; CSTC2020jcyj-msxmX0792). Yunnan Science and Technology Commission (202202AH10006-4), Chongqing Education Commission (HZ2021008). West Light Foundation of The Chinese Academy of Sciences (E1296001). Chongqing Ph.D. Zhitongche Project (No. sl20210000783).
Data Availability Statement: The data presented in this study are available on request from the corresponding authors.

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

References
22. Yang, X.Y.; Zhang, Z.R. A CNN-LSTM Model Based on a Meta-Learning Algorithm to Predict Groundwater Level in the Middle and Lower Reaches of the Heihe River, China. Water 2022, 14, 2377. [CrossRef]
24. Yang, S.; Yang, D.; Chen, J.; Santisirisomboon, J.; Lu, W.; Zhao, B. A physical process and machine learning combined hydrological model for daily streamflow simulations of large watersheds with limited observation data. J. Hydrol. 2020, 590, 125206. [CrossRef]


42. Timofeev, R. Classification and Regression Trees (CART) Theory and Applications; Humboldt University: Berlin, Germany, 2004; p. 54.

43. Morgan, J. *Classification and Regression Tree Analysis*; Boston University: Boston, MA, USA, 2014; p. 298.


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.