water

MDPI

Article

Improving Daily Streamflow Forecasting Using Deep Belief Net-Work Based on Flow Regime Recognition

Jianming Shen [1,2], Lei Zou [1,*], Yi Dong [1,2], Shuai Xiao [1,2], Yanjun Zhao [1,2] and Chengjian Liu [1,2]

[1] Key Laboratory of Cycle and Related Land Surface Processes, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China; shenjm.19b@igsnrr.ac.cn (J.S.); dongy.19b@igsnrr.ac.cn (Y.D.); xiaos.17b@igsnrr.ac.cn (S.X.); zhaoyj.20b@igsnrr.ac.cn (Y.Z.); chengjianliu@stumail.nwu.edu.cn (C.L.)

[2] University of Chinese Academy of Sciences, Beijing 100049, China

* Correspondence: zoulei@igsnrr.ac.cn

Abstract: Streamflow forecasting is of great significance for water resources planning and management. In recent years, numerous data-driven models have been widely used for streamflow forecasting. However, the traditional single data-driven model ignores the utilization of different streamflow regimes. This study proposed an integrated framework for daily streamflow forecasting based on the regime recognition of flow sequences. The framework integrates self-organizing maps (SOM) for identifying streamflow sub-sequences, the random forests (RF) algorithm to select input variables for different streamflow sub-sequences, and a deep belief network (DBN) for establishing complex relationships between the selected input variables and streamflows for different sub-sequences. Specifically, the integrated framework was applied to forecast daily streamflow at the Xiantao hydrological station in the Hanjiang River Basin, China. The results show that the developed integrated framework has higher streamflow prediction accuracy than the single data-driven model (i.e., the DBN model in this study), with Nash efficiency coefficient (NSE) of 0.91/0.81 and coefficient of determination (R^2) of 0.93/0.89 for the integrated framework/DBN model during the validation period, respectively. Additionally, the prediction accuracy of the peak flood was also improved. The relative error of the peak flood derived from the integrated framework was reduced by 4.6%, compared with the single DBN model. Overall, the constructed integration framework, considering the complex characteristic of different flow regimes, could improve the accuracy for daily streamflow forecasting.

Keywords: daily streamflow forecasting; regime recognition; SOM; RF; DBN

check for updates

Citation: Shen, J.; Zou, L.; Dong, Y.; Xiao, S.; Zhao, Y.; Liu, C. Improving Daily Streamflow Forecasting Using Deep Belief Net-Work Based on Flow Regime Recognition. *Water* **2022**, *14*, 2241. https://doi.org/10.3390/w14142241

Received: 13 May 2022
Accepted: 13 July 2022
Published: 16 July 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 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/).

1. Introduction

Streamflow forecasting plays an important role in water resources planning and management in both the short and long term [1,2]. Accordingly, developing a precise and reliable model for streamflow forecasting is of high significance [3]. To date, a large number of data-driven models have been developed for streamflow forecasting [4–6]. However, improving the accuracy and reliability of these data-driven models still remains difficult, especially for streamflow with dramatic changes. The reason is that streamflow, which is influenced by various factors such as precipitation, soil moisture and evaporation, is characterized by its nonlinearity and non-stationary status, and the input-output relationship also changes in different periods [6–9].

Traditional streamflow prediction models are mostly physically based models including multiple physical processes establishing the physical relationship between rainfall and runoff [10–13]. Examples of physically based hydrological models are the SWAT [10], VIC [11], MIKE-SHE [12], and Xinanjiang models [13]. These models describe the physical mechanism in explaining rainfall-runoff processes. However, these models have limited

forecasting capabilities in capturing the non-stationary and non-linear characteristics of hydrologic datasets owing to the high variability of spatial and temporal features involved in streamflow forecasting [1]. In recent years, with the acceleration of information processing speed, data-driven models have been widely used for streamflow forecasting, making full use of hydro-meteorological datasets. Compared to the hydrological model, data-driven models do not need to describe the physical mechanisms of multiple hydrological processes and have the potential to achieve high accuracy for streamflow forecasting. As a data-driven model, artificial neural networks (ANN) can extract the characteristics of streamflow processes by learning and training known information from historical data. ANNs can be directly used to construct the relationships between inputs and outputs. According to Sulaiman et al. (2017), ANN models are reliable tools for streamflow forecasting [14]. Compared with most traditional models, ANNs provide acceptable generalization capabilities and speed [15]. Thus far, ANN models such as the Backpropagation Artificial Neural Network (BPNN), Extreme learning machine (ELM), Multi-Layer Perceptron (MLP), and Deep Belief Networks (DBN) have been successfully used for streamflow prediction [16–20]. However, efforts at physically interpreting the streamflow process and selecting input variables for a single data-driven model has limited [21–23].

The process of streamflow generation involves several stages (e.g., rising limbs, falling limbs, and base flows of a hydrograph). It exhibits distinct characteristics during different stages [24,25]. For example, the characteristics at the beginning of a rainfall event are very different from those at the falling limb of the same hydrograph, and even further from those of the low flows. To consider such distinct characteristics of the streamflow generation process, data analysis techniques (e.g., self-organizing map (SOM), Fuzzy C-means (FCM)) for flow regime recognition have been used to group data into clusters, and separate neural network models have been developed for each cluster [26–31]. Studies implementing data analysis techniques have focused on performance improvement, while efforts to physically interpret streamflow processes have been limited [32]. In addition, the relationship between the input variables and streamflow during different stages has been determined according to the overall characteristics of the entire basin, largely ignoring changes in the relationships for different stages. As the parameterization of hydrological models, the input data belonging to a particular stage is extremely important to obtain the characteristics of streamflow. Undoubtedly, redundant input variables are likely to aggravate the underlying complexity of models [22]. Therefore, it is important to set specific input variables for a particular stage in order to improve the accuracy and reliability of the model for streamflow forecasting.

In this study, an integrated framework, which incorporates the physical interpretation of streamflow processes into data-driven streamflow forecasting models and considers changes in relationships between selected input variables and streamflow of different sub-processes, is proposed. In this framework, streamflow sub-sequences are identified using the self-organizing map (SOM), and input variables corresponding to different streamflow regimes are selected using a random forests (RF) algorithm; complex relationships between the selected input variables and streamflow for different streamflow sub-sequences are constructed based on a DBN model. The developed integrated framework was applied to forecast daily streamflow at Xiantao hydrological station in the Hanjiang River Basin. In addition, the performance of the framework was compared with a single neural network model (hereafter referred to as DBN).

This paper is organized as follows. The study area and data are described in Section 2. The integrated framework, which includes SOM, RF algorithm, and DBN model, is proposed in Section 3. Section 4 presents and discusses the main results for streamflow forecasting at the Xiantao hydrological station. Finally, the conclusions are presented in Section 5.

2. Study Area and Data

The Hanjiang River, the largest tributary of the middle reaches of the Yangtze River, is the water source for large inter-basin water transfer projects such as the mid-route of the South-to-North Water Transfer Project and the Hanjiang-to-Weihe River Diversion Project [33]. It is located at 106°15′~114°20′ E and 30°10′~34°20′ N, with the basin covering about 159,000 km^2 (Figure 1). The Hanjiang River Basin (HJRB) is divided into three regions by the Huangjiagang and Huangzhuang hydrological stations: the upper sub-basin, the middle sub-basin, and the lower sub-basin. The topography in HJRB is high in the west and low in the east. The HJRB is located in a subtropical monsoon region and the rainfall is unevenly distributed throughout the year, with rainfall from May to October accounting for about 75% of the yearly rainfall.

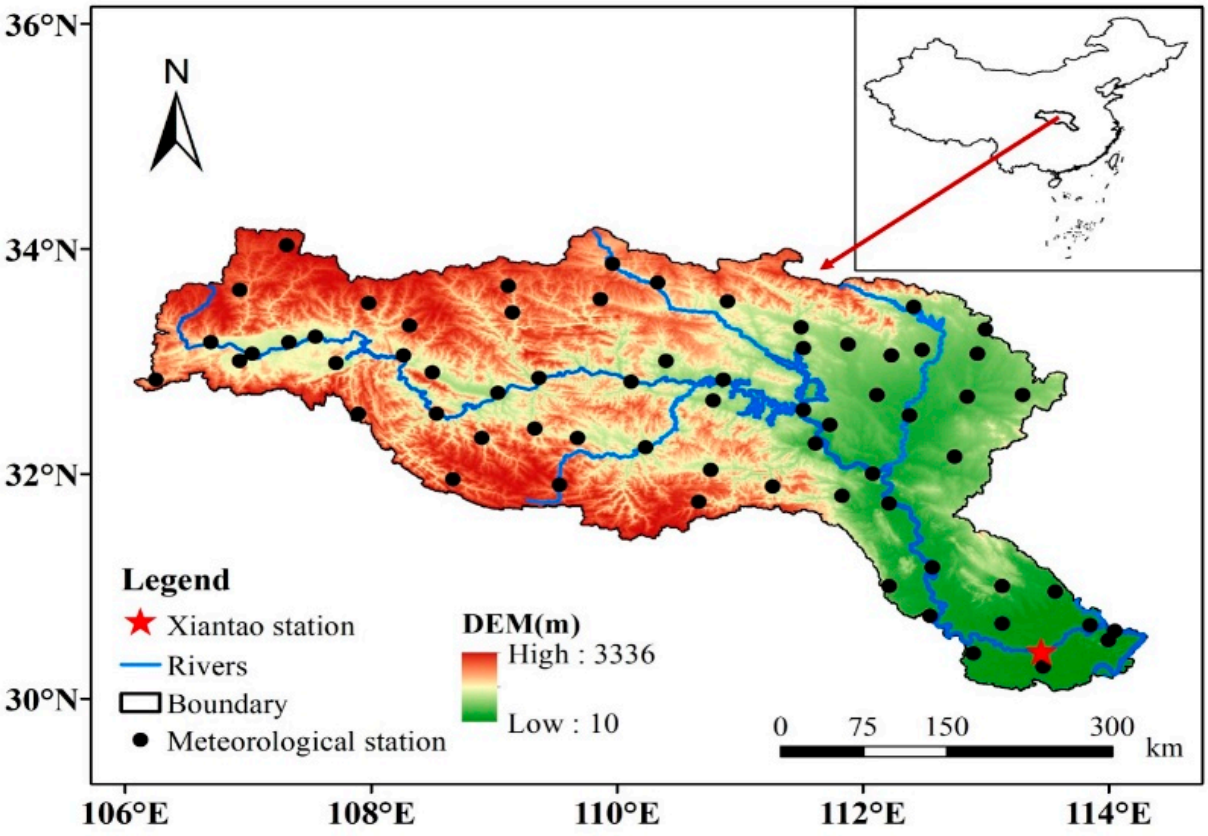

Figure 1. Location of the study area, showing the hydrological and meteorological stations.

Daily precipitation amounts during 1980–2014 from 66 national meteorological stations (Figure 1) in the HJRB were obtained from the National Meteorological Information Center (NMIC) of the China Meteorological Administration (CMA). The daily streamflow data of the flood seasons (late June to early October) [34] at the Xiantao hydrological station in the Hanjiang River Basin from 1980 to 2014 were collected from the Changjiang Water Resources Commission of the Ministry of Water Resources and the hydrological statistical yearbook. The Thiessen Polygons interpolation method was used to obtain the average rainfall over the study area. We also collected evaporation and surface soil moisture datasets from 1980 to 2014 estimated by the Global Land Evaporation Amsterdam Model (GLEAM) with a spatial resolution of 0.25° and a temporal resolution of one day. Table 1 shows the statistical characteristics of the datasets (Rainfall, Streamflow, Soil moisture, Evaporation) of the flood seasons from 1980 to 2014 over the Hanjiang River Basin.

Table 1. Statistical characteristics of datasets (Rainfall, Streamflow, Soil moisture, Evaporation) of the flood seasons from 1980 to 2014 over the Hanjiang Basin.

Datasets	MEAN	CV	SKEW	KURT
Streamflow	1260	0.85	2.79	9.94
Rainfall	4.2	1.51	2.41	7.22
Soil moisture	0.3	0.06	0.02	0.09
Evaporation	2.6	0.30	−0.09	−0.67

Note: CV: Coefficient of Variation; SKEW: Coefficient of skewness; KURT: Coefficient of kurtosis. Unit of Streamflow: m^3/s; Unit of Rainfall: mm/d; Unit of Soil moisture: m^3/m^3; Unit of Evaporation: mm/d.

3. Methodology

3.1. Streamflow Process Clusters Based on Hydro-Meteorological Conditions

The streamflow generation process can be divided into several periods corresponding to respective streamflow sub-processes, such as base flows, rising limbs, and falling limbs. The different streamflow sub-processes are dominated by the hydrological and meteorological conditions corresponding to different periods [30]. A part of a hydrograph recorded before, during, and after a rainfall event is shown in Figure 2. Before the rainfall event, continuous evaporation depleted soil moisture, and the streamflow sub-process presented a low flow (baseflow). At the beginning of the rainfall event, rainfall replenished the soil water deficit, and the streamflow rose slowly. As high intensity rainfall continued to occur and soil moisture became saturated, the streamflow increased rapidly to reach the peak, and the streamflow sub-process of the period presented a rising limb state. After cessation of the rainfall event, the streamflow declined sharply, and the streamflow sub-process presented a falling limb state. In the lower part of the recession limb, the streamflow decreased slowly to the base flow. Considering this pattern, the hydro-meteorological data were grouped into several clusters to represent different streamflow sub-processes.

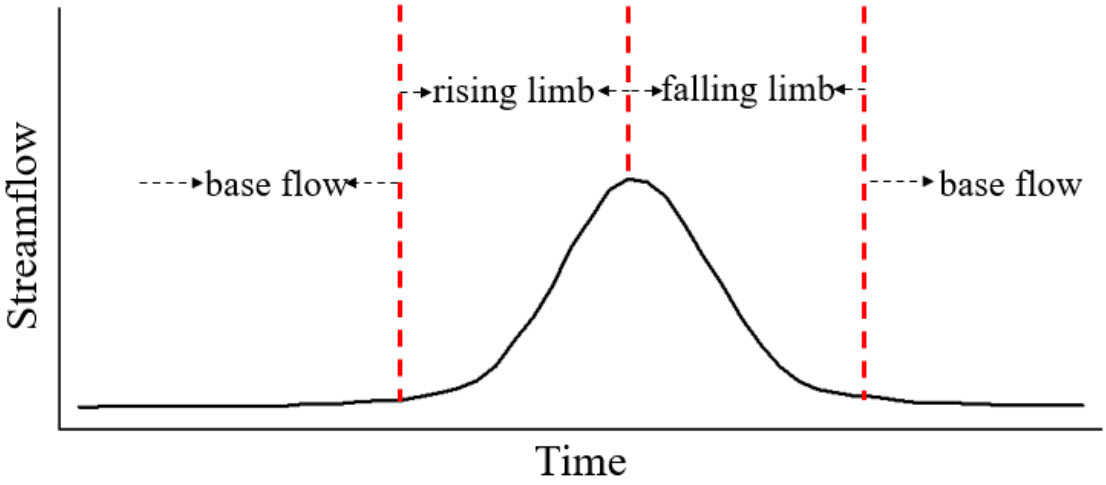

Figure 2. Part of a hydrograph recorded before, during, and after a rainfall event.

3.2. Integrated Neural Network Framework (SOM-RF-DBN)

The integrated framework (SOM-RF-DBN) for daily streamflow forecasting was developed through the following three steps: (i) data cluster analysis; (ii) input variable selection; and (iii) model development (Figure 3). In the first step, an SOM was used to identify streamflow sub-processes, and the datasets were partitioned into specific clusters corresponding to different streamflow sub-processes. Each cluster represented a segment of the hydrograph (e.g., the rising limbs, falling limbs, and base flows). In the second step, an RF algorithm was used for selecting input variables corresponding to different streamflow sub-processes. Finally, separate DBN models were developed for each cluster. Details of the SOM, RF, and DBN are provided in the following Sections 3.2.1–3.2.3.

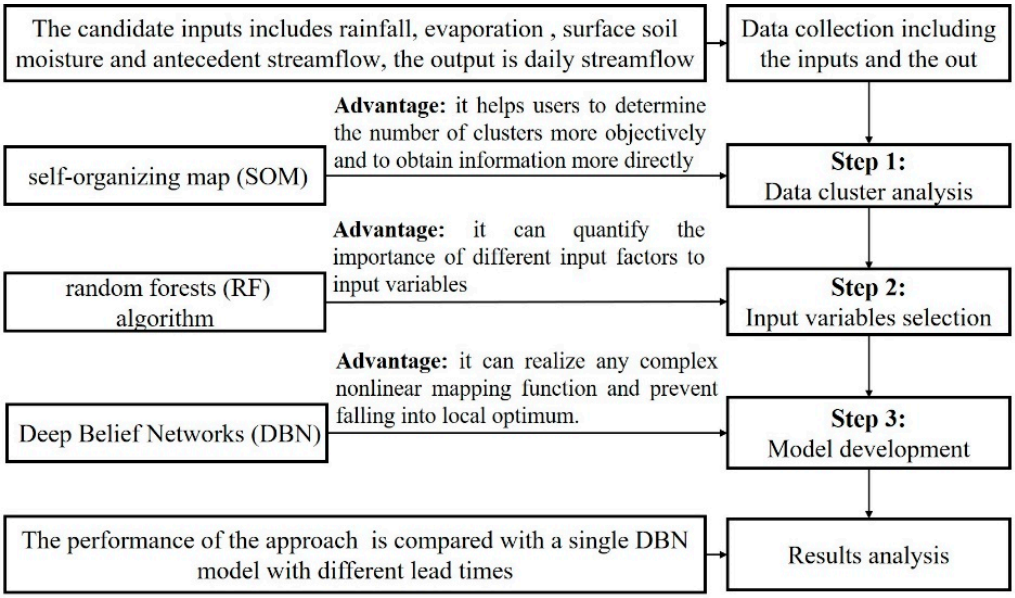

Figure 3. Illustration of the integrated framework for daily streamflow forecasting.

3.2.1. Self-Organizing Map (SOM)

The SOM is an unsupervised learning neural network that used a competitive learning strategy to map the input data onto a two-dimensional topological map [35]. The SOM comprises one input layer and one output layer (the topological map): The input layer contains a node for each of the n variables characterising the unit to classify and the output layer is a two-dimensional array whose nodes are connected, by weighted connections, to the input layer. Each input vector "activates" only one output node, representing its class, using the three processes: the competitive process, the cooperative process, and the adaptive process.

The competition process seeks an optimal match between the input vector $x(t)$ and weight vector $w_i(t)$. This process can be expressed as

$$i^*(t) = \text{argmin}\| x(t) - w_i(t)\|, i = 1, 2, \cdots m \tag{1}$$

where $\| * \|$ is the Euclidean distance commonly used as the similarity measure, m is the total number of neurons, $i^*(t)$ which is called the winning neuron, is an index used to identify the neuron that best matches the input data $x(t)$, and t is the discrete time step corresponding to the iteration of the algorithm.

In the cooperative process, the influence of the winning neuron is delivered to its neighboring neurons. The amplitude of the influence should decrease monotonically with the lateral distance. A time-varying topological form is defined by the Gaussian weighted neighborhood function (2):

$$\mathrm{h}_{i,i^*(t)}(t) = \exp\left(-\frac{\|r_{i(t)} - r_{i^*(t)}\|^2}{2\sigma^2(t)}\right) \tag{2}$$

where $r_i(t)$ and $r_{i^*}(t)$ respectively determine the position of i and i^* in the output array of the SOM network and the effective width of neighborhood function $\sigma^2 > 0$. To ensure the convergence of the weight vector to a stable state, $\eta(t)$ and $\sigma(t)$ should gradually decay with time. The following exponential attenuation can be adopted:

$$\eta(t) = \eta_0\left(\frac{\eta_T}{\eta_0}\right)^{t/T}, \sigma(t) = \sigma_0\left(\frac{\sigma_T}{\sigma_0}\right)^{t/T} \tag{3}$$

η_0 and η_T represent the initial and final values of $\eta(t)$, respectively. σ_0 and σ_T represent the initial and final values of $\sigma(t)$, respectively.

The adaptive process sequentially adjusts the weights of the neurons in the neighborhood of the winning neuron, taking the winning neuron as the center. The adjusted weight at iteration time $t+1$ is defined as:

$$w_i(t+1) = w_i(t) + \eta(t)\mathrm{h}_{i,i^*(t)}(t)(x(t) - w_i(t)) \quad (4)$$

where $\eta(t)$ is the learning rate.

The weight was adjusted until the global sorting became stable. The results maintain the topology of the input vector, and similar patterns were mapped to adjacent areas of the network. These topological relationships can be stored.

After the SOM training is done, feeding the trained SOM with all input data can lead to the feature map (the topological relationship). The way to obtain the feature map is to label all winning neurons in the output array with the identities of corresponding input data. Figure 4a shows that the feature map was obtained by feeding the trained SOM with 20 pieces of data. Each hexagon grid in the feature map represents one neuron. The numbers in grids are the identities of input data. With the feature map, the relative topological relationships between input data can be identified. The location of a winning neuron in the feature map shows the topological location of corresponding input data in the input space. Data that are close in the input space tend to map onto same or adjacent winning neuron in a certain place of the feature map. According to this property, the feature map can reveal the grouping of input data. Thus, a proper number of clusters can be determined objectively. Figure 4b shows the density map that was obtained by counting the number of members in each grid of the feature map. The numbers in grids in the density map represent the amount of input data that was mapped onto the winning neuron. The blue lines that surround the non-blank grids represent two specific clusters.

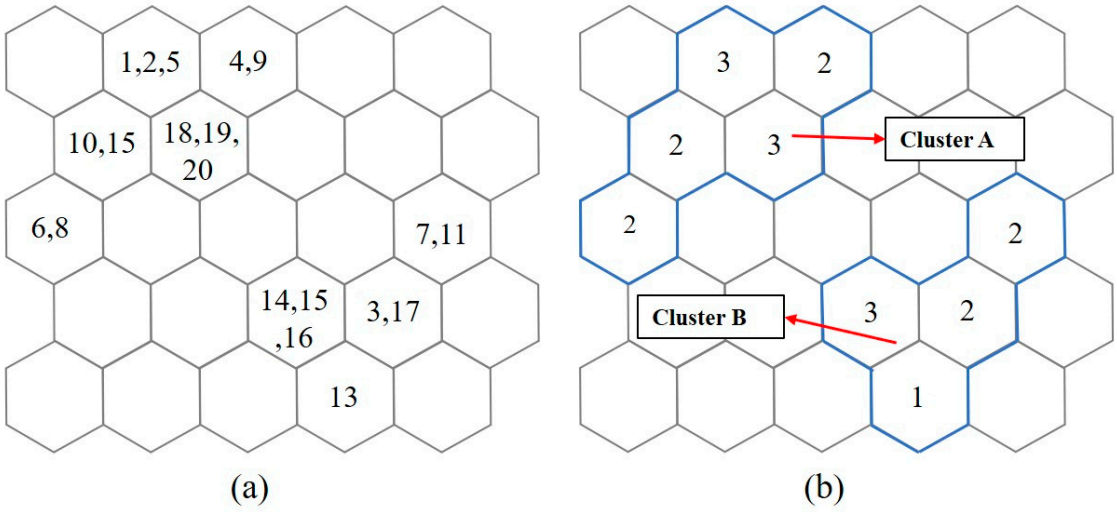

Figure 4. Illustrations of (**a**) the feature map, (**b**) the density map.

3.2.2. Random Forests (RF)

The RF algorithm was applied to select the optimal input variables corresponding to different streamflow sub-processes [36]. The RF algorithm uses a combination of independent decision trees to model data and measure variable importance. It has good prediction ability and relatively simple coding with high performance for noisy data. Moreover, it has unique advantages in the selection and identification of high-dimensional feature factors. Compared with traditional methods for selecting input factors, such as the correlation coefficient method and the intelligent search algorithm, the RF algorithm does not require the number of optimal input factors to be preset, and the amount of calculations is relatively

small. It can consider the complex nonlinear correlation between factors and quantify the importance of different input factors to predict variables.

The RF model uses the importance score of input factors to measure the impact of each input factor on streamflow. The two commonly used indicators for measuring the importance of factors are the Gini index and the out of bag (OOB) error rate [37]. In this study, the importance of variables was measured on the basis of the OOB error rate. Firstly, OOB data were used as the test set to test the performance of all regression trees, and the corresponding mean square deviation was obtained, which is recorded as $\{MSE_i, i = 1, 2, \cdots, b\}$. Secondly, noise disturbance was added to each characteristic variable x_j of the OOB dataset to generate new OOB data. Subsequently, all regression trees were retested with the new dataset to obtain the mean square deviation matrix after random disturbance:

$$\begin{bmatrix} MSE_{11} & MSE_{12} & \cdots & MSE_{1b} \\ MSE_{21} & MSE_{22} & \cdots & MSE_{2b} \\ \vdots & \ddots & & \vdots \\ MSE_{m1} & MSE_{m2} & \cdots & MSEmb \end{bmatrix} \tag{5}$$

The variable importance measure (VIM) of the k-th input characteristic variable is defined as:

$$VIM_{\mathrm{k}} = \frac{\frac{1}{b}\sum_{i=1}^{b}\left(MSE_j - MSE_{kj}\right)}{SE} \tag{6}$$

SE is the standard error of b regression trees.

3.2.3. Deep Belief Network (DBN)

The widely used DBN, first proposed by Geoffrey Hinton in 2006, is an effective modeling method for mapping non-linear relationships [38]. Compared with other ANN models, DBN can prevent results from falling into a local optimum and can thereby accelerate the training process [39]. A typical DBN structure can be decomposed into multiple restricted Boltzmann machines (RBM). Each RBM contains a hidden layer $h = (h_1, h_2, \cdots h_n)$ and a visual layer $v = (v_1, v_2, \cdots v_n)$, where the hidden layer of the former RBM structure is the input layer of the next RBM structure, and the gradient descent algorithm and back-propagation algorithm are used to optimize the model. The energy function of the joint distribution of the visible layer and the hidden layer cells can be defined as:

$$E(v, h) = -\sum_{j=1}^{m}\sum_{i=1}^{n} v_i w_{ij} h_j - \sum_{i=1}^{n} b_i v_i - \sum_{j=1}^{n} c_j h_j \tag{7}$$

where w_{ij} is the connection weight matrix, is visual layer unit bias, $b = (b_1, b_2, Lb_n)$ and $c = (c_1, c_2, \cdots c_n)$ is the hidden unit bias.

The joint probability between the visible layer and the hidden layer can be defined as:

$$\rho(v, h) = \exp\left(\frac{-E(v, h)}{\sum_v \sum_h \exp(-E(v, h))}\right) \tag{8}$$

Given v_i and h_j defines as the states of visible layer i and hidden layer j, and it can be seen that the cells of hidden layer are conditionally independent of the cells of visible layer. Therefore, the probability of each cell in the visual layer can be calculated as:

$$\rho(v_i = 1|h) = \frac{1}{1 + \exp\left(-\sum_{j=1}^{m} w_{ij} h_j - b_i\right)} \tag{9}$$

The probability of each cell in the hidden layer can therefore be calculated as:

$$\rho(h_i = 1|v) = \frac{1}{1 + \exp\left(-\sum_{i=1}^{n} v_i w_{ij} - c_j\right)} \tag{10}$$

Finally, the parameters $\theta(w_{ij}, b_i, \cdots c_j)$ are optimized by the logarithm likelihood function, which maximizes the visible layer $\rho(v)$.

As a deep network structure, the DBN training process includes two steps: unsupervised pre-training and supervised back-propagation network fine-tuning. In the unsupervised pre-training stage, each layer of the RBM network is trained separately. In order to extract important feature information, the reconstruction error function is taken as the objective function to map the unit feature vectors to different feature spaces, and the initial weights of the pre-training network are then obtained. In the fine-tuning stage, the entire network is optimized via reverse propagation.

3.3. Experiment Setup

The integrated framework was developed for daily streamflow forecasting and its applicability to the Xiantao hydrological station in the Hanjiang River Basin was tested. The daily streamflow forecasting model may be generalized as:

$$Q_t = f(Q_{t-1}, \ldots, Q_{t-m}; R_t, R_{t-1}, \ldots, R_{t-m}; E_t, E_{t-1}, \ldots, E_{t-m}; S_t, S_{t-1}, \ldots, S_{t-m}) \quad (11)$$

where Q_t represents current streamflow, $Q_{t-m}(m = 1, 2, 3, \cdots)$ denotes antecedent streamflow, R_t denotes current rainfall, $R_{t-m}(m = 1, 2, \cdots)$ denotes antecedent rainfall in the study area, E_t denotes current evaporation, $E_{t-m}(m = 1, 2, \cdots)$ denotes antecedent evaporation, S_t denotes current surface soil moisture and $S_{t-m}(m = 1, 2, \cdots)$ denotes antecedent surface soil moisture. The set of candidate inputs usually includes variables which might be weakly relevant to the problem. Weakly relevant input variables only serve to add complexity into the model. In this study, according to the tests we carried out, the value of m was set as 10. Accordingly, 43 sets of hydro-meteorological data were used as initial inputs: daily rainfall (R_t), 10 antecedent daily rainfall $(R_{t-1}, R_{t-2}, \cdots, R_{t-10})$, 10 antecedent streamflow $(Q_{t-1}, Q_{t-2}, \cdots, Q_{t-10})$, current daily evaporation (E_t), 10 antecedent evaporation $(E_{t-1}, E_{t-2}, \cdots, E_{t-10})$, current daily evaporation (S_t), and 10 antecedent surface soil moisture $(S_{t-1}, S_{t-2}, \cdots, S_{t-10})$. Current daily streamflow (Q_t) was set as the sole output variable. The Xiantao station included 3920 samples. In order to eliminate the influence of different-scale variables on the training of the model, the hydro-meteorological data were first normalized to fall within a specified range (from 0 to 1). Through this preprocessing step, the trained streamflow forecasting model was deemed to be more efficient.

The dimension of SOM is a key parameter for determining the size of the output space, which affects the number of clustering results. In general, the number of the clustering results increases with the increasing dimension of SOM. We trained the SOM model with four dimensions (5×5, 7×7, 10×10, and 15×15). The epoch and learning rate of SOM were 200 and 0.01, respectively. After the cluster results were determined, input variables were selected for each cluster (streamflow sub-process) using the RF algorithm. Finally, a DBN model was trained separately using the selected input variables and the streamflow of different streamflow sub-processes. The number of hidden layers, number of neurons, and epoch were determined through trial and error. The MAE loss function and the efficient Adam version of the stochastic gradient descent were used to train the DBN model.

3.4. Performance Evaluation Criteria

Model performance was evaluated in terms of the root mean squared error (*RMSE*), Nash–Sutcliffe efficiency coefficient (*NSE*), coefficient of determination (R^2), mean absolute error (*MAE*), and error of peak discharge (EQ_p). These indicators can be defined as follows:

$$NSE = 1 - \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (y_i - \overline{y}_i)^2} \quad (12)$$

$$R^2 = \frac{\left(\sum_{i-1}^{N} (y_i - \overline{y}_i)(\hat{y}_i - \overline{\hat{y}}_i)\right)^2}{\sum_{i-1}^{N} (y_i - \overline{y}_i)^2 \sum_{i-1}^{N} (\hat{y}_i - \overline{\hat{y}}_i)^2} \quad (13)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{N}} \quad (14)$$

$$MAE = \frac{\sum_{i=1}^{N} | \hat{y}_i - y_i |}{N} \quad (15)$$

$$EQ_p = \frac{| Q_p - \hat{Q}_p |}{| Q_p |} \times 100\% \quad (16)$$

where $\hat{y}_i$ and y_i represent the streamflow of the predicted and observed hydrographs at time i, $\overline{y}_i$ denotes the average observed streamflow at time i, $\overline{\hat{y}}_i$ is mean of the predicted streamflow, n is the number of data points, $\hat{Q}_p$ and Q_p are the predicted and observed peak streamflow, respectively. NSE ranges from 1 (a perfect fit) to $-\infty$. Values less than zero indicate that the observation mean would be a better predictor than the model. The RMSE may range from 0 (a perfect fit) to $+\infty$ (no fit) based on the relative range of the data. R^2 has a range of 0–1, with higher values indicating a higher degree of co-linearity.

4. Results and Discussion

4.1. Data Clustering

The SOM model was firstly used to group the hydro-meteorological data into a certain number of clusters. The number of clustering results can be objectively obtained. Figure 5 shows the density maps obtained from SOM with four dimensions (5 × 5, 7 × 7, 10 × 10, and 15 × 15). According to the density maps of Figure 5a–d, the hydro-meteorological data were clustered into two sub-groups, three sub-groups, seven sub-groups, and 10 sub-groups, respectively. After the SOM training is done, the members of each cluster can be also obtained. A DBN model was trained for each cluster. Table 2 shows the values of the performance parameters for the four dimensions of SOM during the calibration and validation periods. For the 7 × 7 dimension of SOM, all input data were clustered into three sub-groups, and the performance parameters showed the best values. With the increasing number of dimensions, the number of clustering results increased and the performance of the model declined rapidly. With the total number of samples remaining unchanged and the number of dimensions of SOM increasing, the number of clustering results increased, and the numbers of each cluster decreased. Consequently, the model performance deteriorated and the separate DBN model could not learn the sufficient sample features of each cluster. Therefore, it is very important for streamflow forecasting to select an appropriate dimension for training the SOM according to the number of samples.

Considering the abovementioned results, the SOM with the 7 × 7 dimension was adopted. All input data were grouped into three clusters, as shown in Figure 5b. Cluster B showed the maximum number of members which is 1542 (the sum of the numbers in all grids surrounded by the blue line of cluster B). Next, the clustering was explored for streamflow during a typical flood season (Figure 6). During this typical flood season, the maximum daily streamflow closed to 12,000 (m^3/s). It is interesting to note that the hydro-meteorological data during the stage of base flows (no rainfall or sprinkling rainfall) were grouped into cluster B. The hydro-meteorological data during the rising limb stage of the streamflow were grouped into Cluster A, which is influenced by the duration and intensity of rainfall. The hydro-meteorological data during the stage of the falling limb were grouped into Cluster C, which is dominated by the sprinkling rainfall and storage characteristics of the basin. The clustering results suggest that the hydro-meteorological data were grouped according to the different hydro-meteorological conditions corresponding to respective streamflow sub-processes.

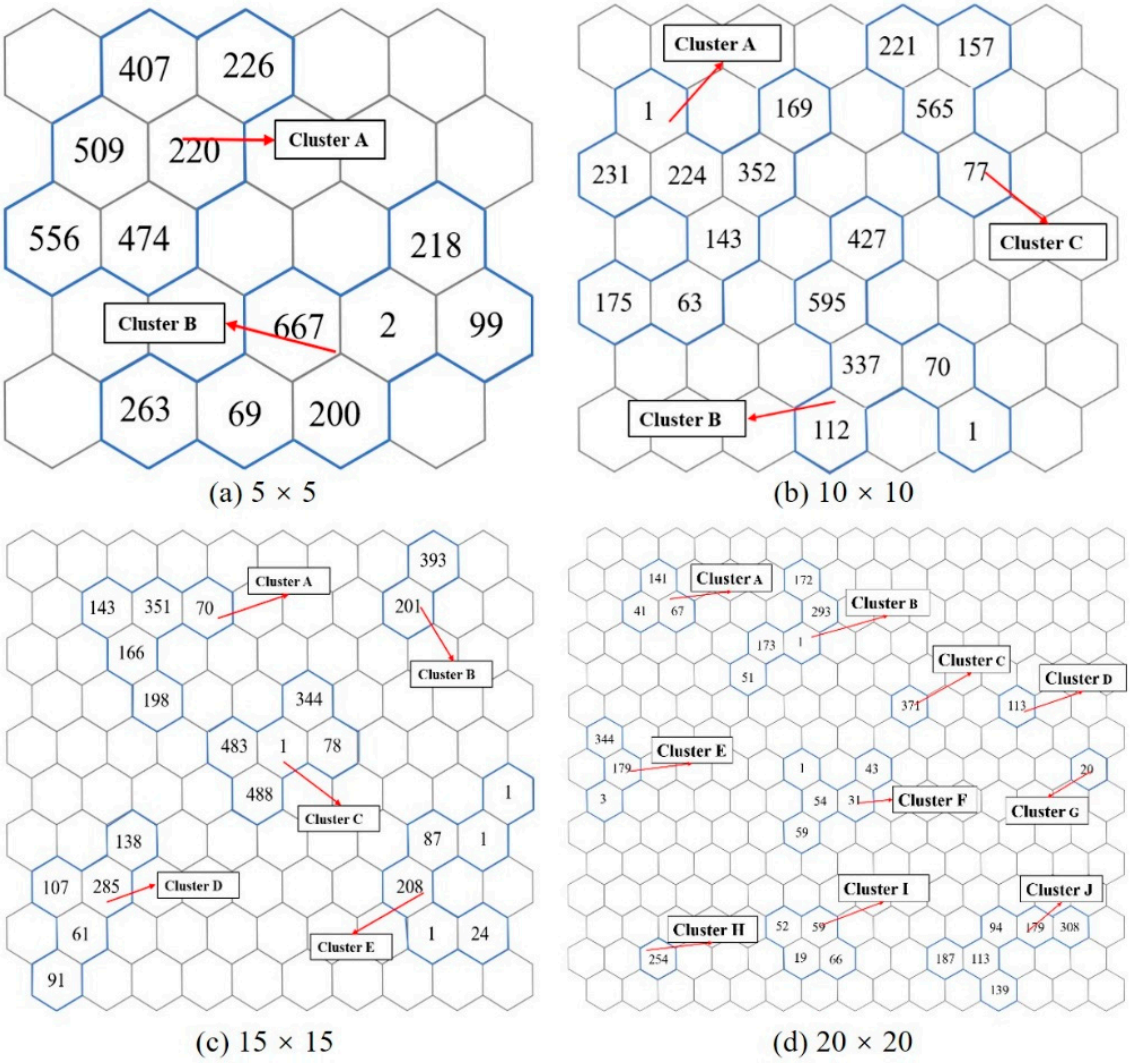

Figure 5. Density maps from SOM with four different dimensions.

Table 2. Performance of the DBN model for four dimensions of SOM during the calibration and validation periods.

Period	Dimension of SOM	*NSE*	R^2	*RMSE*	*MAE*
Calibration	5 × 5	0.90	0.90	262.79	140.45
	7 × 7	0.94	0.95	256.12	111.71
	10 × 10	0.72	0.87	471.45	221.42
	15 × 1 5	0.63	0.82	612.35	361.87
Validation	5 × 5	0.89	0.90	263.35	141.98
	7 × 7	0.91	0.93	261.66	129.17
	10 × 10	0.73	0.86	442.13	241.67
	15 × 15	0.65	0.80	601.23	354.13

4.2. Input Variable Selection

The RF algorithm-based input variable selection method was used to determine the final input variables for streamflow during different stages. The importance scores of candidate input variables were calculated using the RF algorithm. Candidate input variables were then sorted according to the importance scores. The input variables at the top of the order were selected, and the total scores of the selected input variables were greater than 0.95. Table 3 shows the order of the selected input variables for different

clusters. According to the results, Q_{t-1} has a closer relationship with Q_t than other inputs. The selected input variables varied for different clusters to a certain extent. Cluster A, which represents data during the rising limb stage, has a strong relationship with the previous streamflow, previous rainfall, and previous surface soil moisture, but has a weak relationship with evaporation. Cluster B, which represents data during the base flow stage, shows a strong relationship with previous streamflow, rainfall, surface soil moisture, and evaporation. Similar to Cluster A, cluster C, which represents data during the falling limb stage, also shows a poor relationship with evaporation.

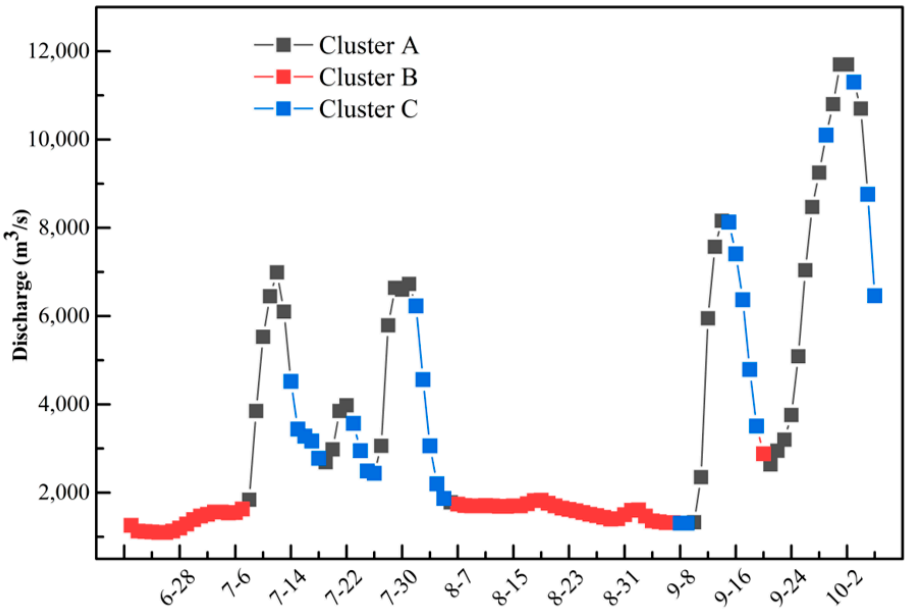

Figure 6. SOM clustering results of streamflow during a typical flood season.

Table 3. Input variables for different clusters selected using the RF algorithm.

Out Variable	Group	Input Variables and Importance Scores
Q (t)	Cluster A	*Q* (t − 1), 0.67; *R* (t − 3), 0.10; *Q* (t − 2), 0.07; *R* (t − 2), 0.04; *Q* (t − 3), 0.04; S (t − 3), 0.04;
	Cluster B	*Q* (t − 1), 0.65; *R* (t − 3), 0.08; *Q* (t − 2), 0.08; *S* (t − 3), 0.05; *R* (t − 2), 0.04; *E* (t − 3), 0.04; *R* (t − 4), 0.03;
	Cluster C	*Q* (t − 1), 0.69; *Q* (t − 2), 0.12; *S* (t − 3), 0.07; *R* (t − 3), 0.05; *R* (t − 2), 0.02;

Note: Q: streamflow; R: rainfall; S: surface soil moisture; E: evaporation.

4.3. Performance Comparison between the Integrated Framework and Single DBN Model

Figure 7 shows the statistical characteristics of the validation dataset (2005–2014). In Figure 7a, the upper boundary does not exceed 3500 m^3/s. Due to the storage characteristics and continuous rainfall in the catchment during the flood season, the validation dataset contained many outliers. In the validation dataset, the upper boundary (3500 m^3/s) was exceeded on 113 days (10.1%), and daily streamflow exceeded 10,000 m^3/s on five days. As shown in Figure 7a, values forecasted by the single DBN model were smaller than those observed when the daily streamflow exceeded 10,000 m^3/s. In contrast, the SOM-RF-DBN framework provided forecasts more consistent with the observation data. The cumulative distribution of the modeled and observed data is shown in Figure 7b. When the cumulative distribution value is less than 0.85, the three curves of cumulative distribution nearly coincide. However, when the value is greater than 0.85, the curve of DBN deviates from the other two curves. The single DBN model and SOM-RF-DBN framework exhibited similar performances in small-volume streamflow forecasting (Figure 7). However, the

single DBN model showed poorer performance in large-volume streamflow forecasting. The statistical characteristics of the validation dataset therefore not only showed that the daily streamflow characteristics of the Xiantao hydrological station are highly complex and difficult to forecast but also preliminarily showed that the constructed integrated framework is better than the single DBN model for daily streamflow forecasting.

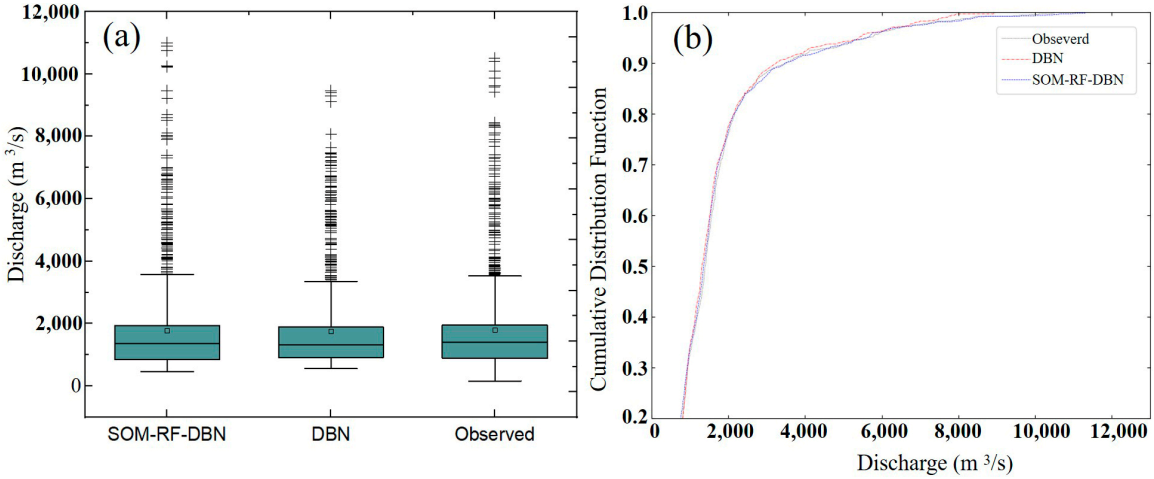

Figure 7. Box-plots (**a**) and cumulative distribution function (**b**) of observed and forecasted streamflow with the validation dataset for the single DBN model and SOM-RF-DBN framework.

The forecasting performance of the two models was also quantitatively analyzed. Table 4 compares the performances of the single DBN and SOM-RF-DBN models for streamflow forecasting. The DBN and SOM-RF-DBN models were quantitatively analyzed according to five performance parameters. The NSE value of the SOM-RF-DBN model was greater than 0.91, while that of the single DBN model was less than 0.85 for the calibration and validation periods. Notably, the EQ_P value of the SOM-RF-DBN model was only half that of the single DBN model. These results indicate that the SOM-RF-DBN model can accurately forecast peak flow. The accurate forecasting of flood peak streamflow is critical for forecasting hydrological processes.

Table 4. Comparison of the performances of the single DBN and the integrated framework (SOM-RF-DBN) for daily streamflow forecasting during the calibration and validation periods.

Datasets	Models	*NSE*	R^2	*RMSE*	*MAE*	EQ_p
			Calibration			
1980–2004	DBN	0.85	0.89	446.20	194.83	9.95%
	SOM-RF-DBN	0.94	0.95	256.29	111.71	4.84%
			Validation			
2005–2014	DBN	0.81	0.89	404.77	197.53	10.34%
	SOM-RF-DBN	0.91	0.93	261.66	129.17	5.74%

Figure 8 shows scatter plots of the observed and forecasted streamflow values. The R^2 values of the DBN and SOM-RF-DBN models were 0.89 and 0.93, respectively. The higher R^2 of the SOM-RF-DBN framework indicates its capability in accurately constructing non-linear relationships between the selected input variables and observed streamflow. As shown in Figure 8a, the forecasting results of the single DBN model exhibits more scattering with many abnormal values. In contrast, the results of the SOM-RF-DBN model are less scattered (Figure 8b). Overall, the integrated framework is more suitable for streamflow forecasting than the single DBN model, and it also shows a better correlation with the observation data.

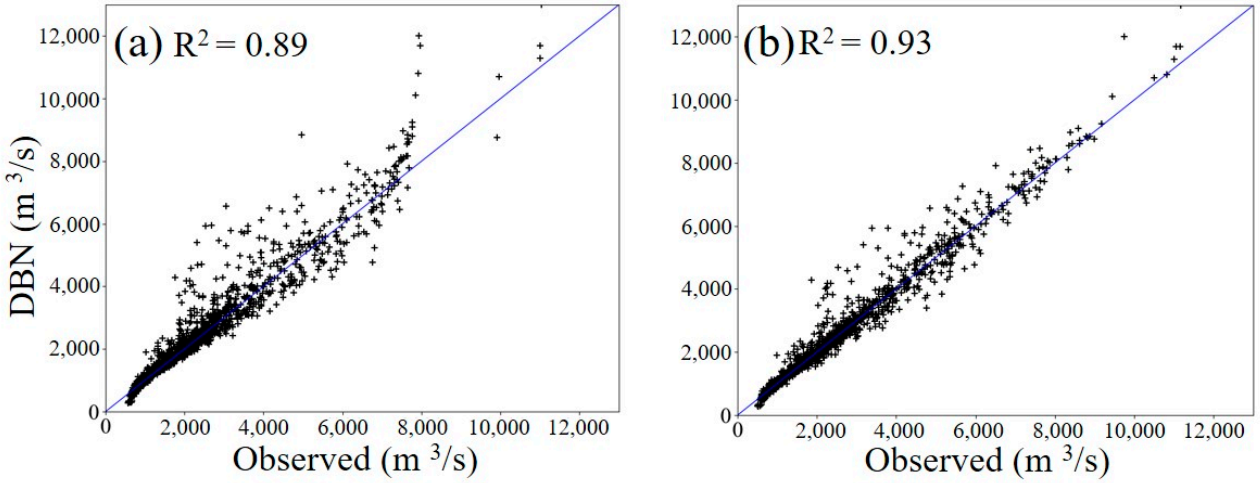

Figure 8. Scatter plots of the observed and forecasted streamflow with the validation dataset: (**a**) DBN model; (**b**) SOM-RF-DBN framework.

Figure 9 shows hydrographs of the observed and forecasted streamflow using the single DBN model and SOM-RF-DBN framework for the validation period. During the flood seasons throughout the ten years, the peak streamflow in 2005 and 2011 exceeded 10,000 m^3/s. The SOM-RF-SOM framework could forecast peak streamflow better than the single DBN model. In 2009, 2011, and 2014, the EQ_p values of SOM-RF-SOM framework were approximately half those of the single DBN model, approximately one-third that of the single DBN model in 2006. Even for the remaining years, the EQ_p values of the SOM-RF-SOM framework remained lower than those of the single DBN model. This proves that the single DBN model has weaker ability to forecast flood peaks than the SOM-RF-DBN framework. The forecasted values of the single DBN model fluctuated abnormally compared to the observed values, irrespective of whether the streamflow volume was large or small. In contrast, the SOM-RF-DBN framework did not exhibit such fluctuations. The fluctuations of forecasted values by the single DBN model were especially more pronounced in 2009, 2012, 2013, and 2014, while the flood hydrograph of the SOM-RF-DBN framework almost coincided with the hydrograph of the observation data. The comparison of daily streamflow forecasting between the single DBN model and SOM-RF-DBN framework proved that the SOM based analysis of hydrological data can improve the performance of DBN models in simulating and forecasting daily streamflow.

Table 5 presents the results of streamflow forecasting at different lead times (one to two days) for the single DBN and SOM-RF-DBN models. In general, the SOM-RF-DBN framework provided better forecasting results than the single DBN model at different lead times. The values of performance parameters were better for the SOM-RF-DBN framework than for the single DBN model in both the calibration and validation periods. The EQ_p values of the SOM-RF-DBN framework at one and two days lead time were also less than those of the DBN model to varying degrees. These results further confirm that the SOM-RF-DBN model has a good performance in flood peak flow forecasting. With the increase of the forecasting period, the performance of the two models decreased, but the values of NSE, R^2, RMSE, MAE, and EQ_p did not show very clear trends. Based on the above results, it can be concluded that the SOM-RF-DBN framework can accurately forecast highly complex streamflow. Accordingly, the integrated SOM-RF-DBN modeling framework can be considered suitable for hydrological research.

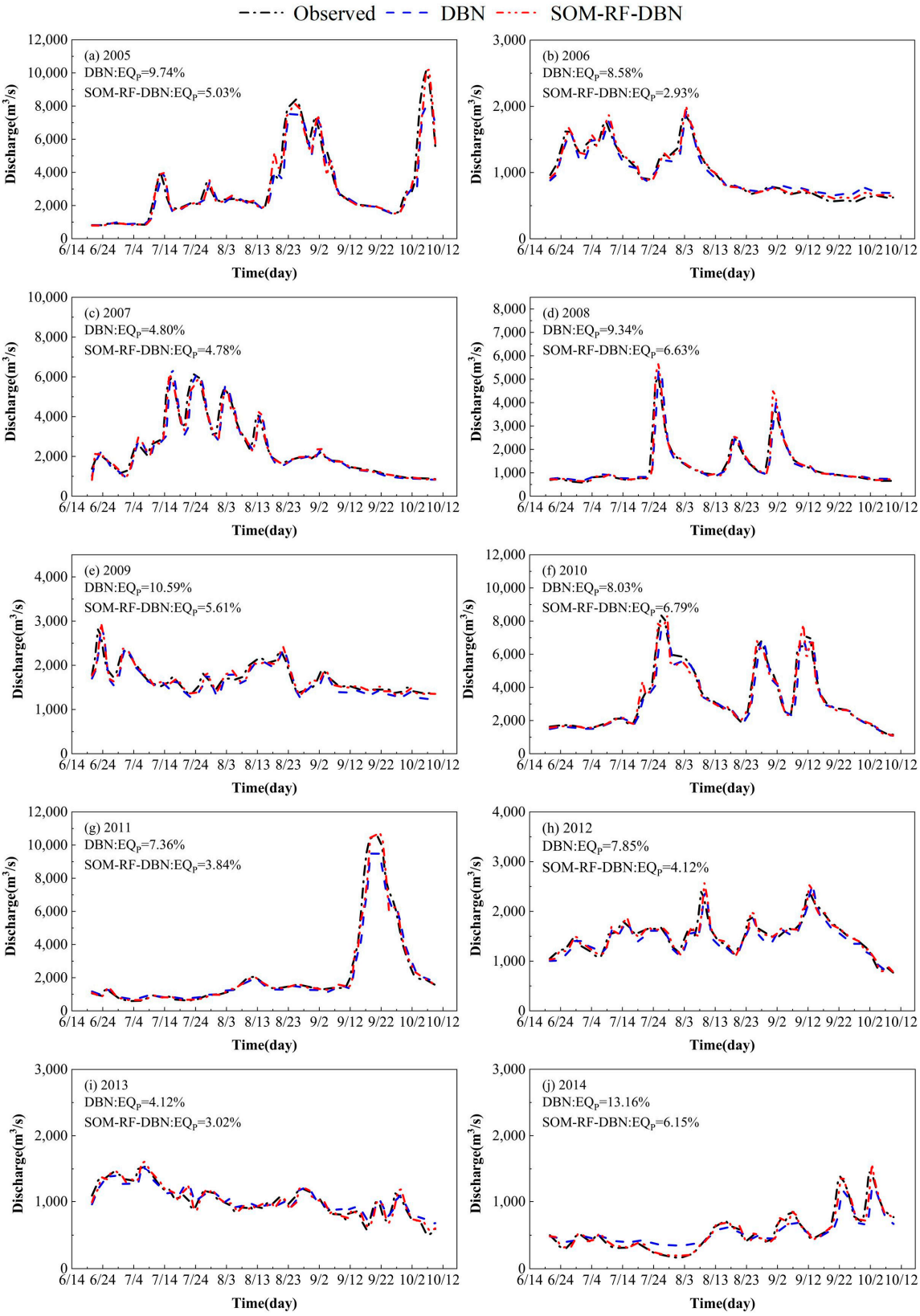

Figure 9. Hydrographs of observed and forecasted streamflow by the DBN and SOM-RF-DBN models for the validation period (2005–2014).

Table 5. Performances of the single DBN model and SOM-RF-DBN framework in streamflow forecasting at different lead times (1–2 d).

Time	Period	Models	*NSE*	R^2	*RMSE*	*MAE*	EQ_p
$t+1$	Calibration	DBN	0.80	0.83	474.83	235.13	12.81%
		SOM-RF-DBN	0.86	0.88	332.54	154.67	8.42%
	Validation	DBN	0.77	0.81	464.30	228.22	12.05%
		SOM-RF-DBN	0.87	0.89	324.45	139.27	7.89%
$t+2$	Calibration	DBN	0.68	0.70	661.08	323.73	18.30%
		SOM-RF-DBN	0.72	0.74	578.45	291.78	15.64%
	Validation	DBN	0.64	0.64	658.06	317.03	18.36%
		SOM-RF-DBN	0.70	0.71	610.76	301.21	16.81%

Kratzert et al. (2018) used the LSTM model to simulate rainfall-runoff processes and obtained better results than those based on physical mechanisms [40]. Hu et al. (2018) used an artificial neural network as well as an LSTM network model to simulate the rainfall runoff process for flood events, and the LSTM network model provided better results than the ANN model [41]. We therefore, used the same input data with our model to train the LSTM model for forecasting streamflow. The NSE values of the LSTM model were 0.91 and 0.90 for the calibration and validation periods, respectively, which are better than those of the single DBN model. Compared with the LSTM model, the NSE values of the SOM-RF-DBN model is also higher. Although the LSTM model has a memory unit for simulating complex streamflow, it still could not provide deep insight into the characteristics of streamflow processes.

5. Conclusions

This study proposed an integrated framework (SOM-RF-DBN) for daily streamflow forecasting by integrating the physical understanding of different flow regimes into the modelling process. The integrated framework integrates SOM for identifying streamflow sub-processes, a RF algorithm for selecting input variables, and DBN for constructing the complex relationships between the selected input variables and streamflow for different sub-processes. This framework was successfully applied to the forecasting of continuous daily streamflow at the Xiantao hydrological station in the Hanjiang River Basin. The main findings are as follows:

(1) The integrated framework can improve the performance of the single neural network model for daily streamflow forecasting. The NSE values of the integrated framework exceeded 0.9 for both the calibration and validation periods. The relative error of the peak flood derived from the integrated framework were reduced by 4.6%, compared with the single DBN model.

(2) The integrated framework, which incorporates the physical interpretation of streamflow processes and considers changes of relationships between the selected input variables and the streamflow of different sub-processes, can provide a better understanding of the complex characteristics of different flow regimes during streamflow generation processes, thus improving the forecasting accuracy and reliability of streamflow forecasting based on data-driven models.

In the future, we will improve the performance of the integrated framework for long-term prediction by applying more input variables. In addition, we will add rigorous structural validity to this framework because it is often ignored in neural network models.

Author Contributions: Formal analysis and writing—original draft, J.S.; methodology, conceptualization, writing—review and editing, L.Z.; investigation, Y.D. and Y.Z.; data analysis, Y.D. and S.X.; methodology, C.L.; conceptualization, J.S. and C.L. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDA23040304) and National Natural Science Foundation of China (41890823).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used or analyzed during the current study are available on reasonable request.

Acknowledgments: We thank the anonymous reviewers for their constructive feedback.

Conflicts of Interest: The authors declare that they have no conflicts of interest.

References

1. Cheng, M.; Fang, F.; Kinouchi, T.; Navon, I.M.; Pain, C.C. Long lead-time daily and monthly streamflow forecasting using machine learning methods. *J. Hydrol.* **2020**, *590*, 125376. [CrossRef]
2. Kilinc, H.C.; Yurtsever, A. Short-Term Streamflow Forecasting Using Hybrid Deep Learning Model Based on Grey Wolf Algorithm for Hydrological Time Series. *Sustainability* **2022**, *14*, 3352. [CrossRef]
3. Yazid, T.; Doudja, S.G.; Ali, N.A.; Ozgur, K.; Ahmed, E.S. Improving artificial intelligence models accuracy for monthly streamflow forecasting using grey Wolf optimization (GWO) algorithm. *J. Hydrol.* **2020**, *582*, 124435.
4. Ibrahim, K.; Huang, Y.F.; Ahmed, A.N.; Koo, C.H.; El-Shafie, A. A review of the hybrid artificial intelligence and optimization modelling of hydrological streamflow forecasting. *Alex. Eng. J.* **2022**, *61*, 279–303. [CrossRef]
5. Mo, R.; Xu, B.; Zhong, P.A.; Zhu, F.; Huang, X.; Liu, W.F.; Xu, S.Y.; Wang, G.Q.; Zhang, J.Y. Dynamic long-term streamflow probabilistic forecasting model for a multisite system considering real-time forecast updating through spatio-temporal dependent error correction. *J. Hydrol.* **2021**, *601*, 126666. [CrossRef]
6. Meng, E.H.; Huang, S.Z.; Huang, Q.; Fang, W.; Wu, L.Z.; Wang, L. A robust method for non-stationary streamflow prediction based on improved EMD-SVM model. *J. Hydrol.* **2019**, *568*, 462–478. [CrossRef]
7. Huang, S.Z.; Huang, Q.; Leng, G.Y.; Liu, S.Y. A nonparametric multivariate standardized drought index for characterizing socioeconomic drought: A case study in the Heihe River Basin. *J. Hydrol.* **2016**, *542*, 875–883. [CrossRef]
8. Boucher, M.A.; Quilty, J.; Adamowski, J. Data Assimilation for Streamflow Forecasting Using Extreme Learning Machines and Multilayer Perceptrons. *Water Resour. Res.* **2020**, *56*, e2019WR026226. [CrossRef]
9. Chu, H.B.; Wei, J.H.; Wu, W.Y.; Jiang, Y.; Chu, Q.; Meng, X.J. A classification-based deep belief networks model framework for daily streamflow forecasting. *J. Hydrol.* **2021**, *595*, 125967. [CrossRef]
10. Johan, V.T.; Bieger, K.; Arnold, J.G. A hydropedological approach to simulate streamflow and soil water contents with SWAT+. *Hydrol. Processes* **2021**, *35*, e14242.
11. Maza, M.; Srivastava, A.; Bisht, D.S.; Raghuwanshi, N.S.; Bandyopadhyay, A.; Chatterjee, C.; Bhadra, A. Simulating hydrological response of a monsoon dominated reservoir catchment and command with heterogeneous cropping pattern using VIC model. *J. Earth Syst. Sci.* **2020**, *129*, 200. [CrossRef]
12. Aredo, M.R.; Hatiye, S.D.; Pingale, S.M. Impact of land use/land cover change on stream flow in the Shaya catchment of Ethiopia using the MIKE SHE model. *Arab. J. Geosci.* **2021**, *14*, 114–128. [CrossRef]
13. Wang, J.; Bao, W.M.; Gao, Q.Y.; Si, W.; Sun, Y.Q. Coupling the Xinanjiang model and wavelet-based random forests method for improved daily streamflow simulation. *J. Hydroinform.* **2021**, *23*, 589–604. [CrossRef]
14. Sulaiman, J.; Wahab, S.H. Heavy Rainfall Forecasting Model Using Artificial Neural Network for Flood Prone Area. In *IT Convergence and Security*; Springer: Singapore, 2017; Volume 2018, pp. 68–76.
15. Minocha, V.K. Discussion of "Comparative Analysis of Event-based Rainfall-runoff Modeling Techniques—Deterministic, Statistical, and Artificial Neural Networks" by Ashu Jain and S. K. V. Prasad Indurthy. *J. Hydrol. Eng.* **2004**, *9*, 550–551. [CrossRef]
16. Riad, S.; Mania, J.; Bouchaou, L.; Najjar, Y. Rainfall-runoff model using an artificial neural network approach. *Math. Comput. Model.* **2004**, *40*, 839–846. [CrossRef]
17. Lima, D.B.D.; Lima, M.D.C.E.; Salgado, R.M. An Empirical Analysis of MLP Neural Networks Applied to Streamflow Forecasting. *IEEE Lat. Am. Trans.* **2011**, *9*, 295–301. [CrossRef]
18. Lima, A.R.; Cannon, A.J.; Hsieh, W.W. Forecasting daily streamflow using online sequential extreme learning machines. *J. Hydrol.* **2016**, *537*, 431–443. [CrossRef]
19. Yaseen, Z.M.; Jaafar, O.; Deo, R.C.; Kisi, O.; Adamowski, J.; Quilty, J.; El-Shafie, A. Stream-flow forecasting using extreme learning machines: A case study in a semi-arid region in Iraq. *J. Hydrol.* **2016**, *542*, 603–614. [CrossRef]
20. Chu, H.B.; Wei, J.H.; Qiu, J. Monthly Streamflow Forecasting Using EEMD-Lasso-DBN Method Based on Multi-Scale Predictors Selection. *Water* **2018**, *10*, 1486. [CrossRef]
21. Ghaith, M.; Siam, A.; Li, Z.; El-Dakhakhni, W. Hybrid Hydrological Data-Driven Approach for Daily Streamflow Forecasting. *J. Hydrol. Eng.* **2020**, *25*, 04019063–04019071. [CrossRef]

22. Li, X.Y.; Maier, H.R.; Zecchin, A.C. Improved PMI-based input variable selection approach for artificial neural network and other data driven environmental and water resource models. *Environ. Model. Softw.* **2015**, *65*, 15–29. [CrossRef]
23. Prasad, R.; Deo, R.C.; Li, Y.; Maraseni, T. Input selection and performance optimization of ANN-based streamflow forecasts in the drought-prone Murray Darling Basin region using IIS and MODWT algorithm. *Atmos. Res.* **2017**, *197*, 42–63. [CrossRef]
24. Li, M.; Wang, Q.J.; Robertson, D.E.; Bennett, J.C. Improved error modelling for streamflow forecasting at hourly time steps by splitting hydrographs into rising and falling limbs. *J. Hydrol.* **2017**, *555*, 586–599. [CrossRef]
25. Li, F.F.; Cao, H.; Hao, C.F.; Qiu, J. Daily Streamflow Forecasting Based on Flow Pattern Recognition. *Water Resour. Manag.* **2021**, *35*, 4601–4620. [CrossRef]
26. Hsu, K.l.; Gupta, H.V.; Gao, X.G.; Sorooshian, S.; Imam, B. Self-organizing linear output map (SOLO): An artificial neural network suitable for hydrologic modeling and analysis. *Water Resour. Res.* **2002**, *38*, 1302–1319. [CrossRef]
27. Lin, G.F.; Wang, C.M. Performing cluster analysis and discrimination analysis of hydrological factors in one step. *Adv. Water Resour.* **2006**, *29*, 1573–1585. [CrossRef]
28. Jain, A.; Srinivasulu, S. Integrated approach to model decomposed flow hydrograph using artificial neural network and conceptual techniques. *J. Hydrol.* **2006**, *317*, 291–306. [CrossRef]
29. Lin, G.F.; Wu, M.C. A hybrid neural network model for typhoon-rainfall forecasting. *J. Hydrol.* **2009**, *375*, 450–458. [CrossRef]
30. Toth, E. Classification of hydro-meteorological conditions and multiple artificial neural networks for streamflow forecasting. *Hydrol. Earth Syst. Sci.* **2009**, *13*, 1555–1566. [CrossRef]
31. Zaher, Y.M.; Isa, E.; Hossein, B.; Ravinesh, C.D.; Ali, D.M.; Wan, H.M.W.M.; Lamine, D.; Ahmed, E.; Singh, V.P. Novel approach for streamflow forecasting using a hybrid ANFIS-FFA model. *J. Hydrol.* **2017**, *554*, 263–276.
32. Jhong, Y.D.; Chen, C.S.; Lin, H.P.; Chen, S.T. Physical Hybrid Neural Network Model to Forecast Typhoon Floods. *Water* **2018**, *10*, 632. [CrossRef]
33. Chen, H.; Guo, J.; Xiong, W.; Guo, S.L.; Xu, C.Y. Downscaling GCMs using the Smooth Support Vector Machine method to predict daily precipitation in the Hanjiang Basin. *Adv. Atmos. Sci.* **2010**, *27*, 274–284. [CrossRef]
34. Wang, D.; Wu, D.; Xie, X.; Li, X. Study on Spatio-Temporal Variation of Runoff in Flood Season in Hanjiang River Basin. *Pearl River* **2020**, *41*, 30–39.
35. Kohonen, T. Self-Organized Formation of Topologically Correct Feature Maps. *Biol. Cybern.* **1982**, *43*, 59–69. [CrossRef]
36. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
37. Genuer, R.; Poggi, J.-M.; Tuleau-Malot, C. Variable selection using random forests. *Pattern Recognit. Lett.* **2010**, *31*, 2225–2236. [CrossRef]
38. Hinton, G.E.; Salakhutdinov, R.R. Reducing the dimensionality of data with neural networks. *Science* **2006**, *313*, 504–507. [CrossRef]
39. Huang, W.H.; Song, G.J.; Hong, H.K.; Xie, K.Q. Deep Architecture for Traffic Flow Prediction: Deep Belief Networks with Multitask Learning. *IEEE Trans. Intell. Transp. Syst.* **2014**, *15*, 2191–2201. [CrossRef]
40. Kratzert, F.; Klotz, D.; Brenner, C.; Schulz, K.; Herrnegger, M. Rainfall-runoff modelling using Long Short-Term Memory (LSTM) networks. *Hydrol. Earth Syst. Sci.* **2018**, *22*, 6005–6022. [CrossRef]
41. Hu, C.H.; Wu, Q.; Li, H.; Jian, S.Q.; Li, N.; Lou, Z.Z. Deep Learning with a Long Short-Term Memory Networks Approach for Rainfall-Runoff Simulation. *Water* **2018**, *10*, 1543. [CrossRef]