


Regional characteristics' impact on the performances of the gated recurrent unit on streamflow forecasting

Qianyang Wang , Yuexin Zheng, Qimeng Yue, Yuan Liu and Jingshan Yu*

College of Water Sciences, Beijing Normal University, Beijing, China

*Corresponding author. E-mail: jingshan@bnu.edu.cn

 QW, 0000-0001-6186-4695

ABSTRACT

The gated recurrent unit (GRU) has obtained attention as a potential model for streamflow forecasting in recent years. Common patterns and specialties when employing it in different regions, as well as a comparison between different models still need investigation. Therefore, we examined the performances of GRU for one, two, and three-day-ahead streamflow forecasting in seven basins in various geographic regions in China from the aspect of robustness, overall accuracy, and accuracy of streamflow peaks' forecasting. The robustness and accuracy of it are closely related to correlations between the input and forecasting target series. Also, it outperforms the benchmark machine learning models in more cases, especially for one-day-ahead forecasting (NSE of 0.88–0.96 except for the unsatisfactory result in the Luanhe River basin). The deterioration of its accuracy along the increasing lead time depends on the dominant time lags between the rainfall and streamflow peaks. Recommendations were proposed for further applications.

Key words: artificial neural networks, assorted regions, gated recurrent unit, performances, recurrent neural networks, streamflow forecasting

HIGHLIGHTS

- An evaluation of GRU versus benchmark models for streamflow forecasting in diverse regions.
- The problem that how do data and basins' characteristics affect the model's performance was discussed.
- The summarized patterns are valuable for a quick applicability evaluation and data selection process in further applications.

INTRODUCTION

Accurate streamflow simulation and forecasting have profound significance for water resources management, which includes some aspects such as irrigation and domestic water allocation, ecological water utilization, flood alarming, and water transportation. However, as a result of the high nonlinearity of the hydrological process (Parisouj *et al.* 2020), it is challenging to obtain an accurate streamflow simulation result. In the past decades, considerable numbers of studies have focused on attaining a better modelling accuracy utilizing different techniques, which can be mainly categorized into physics-based models and data-driven models.

Well developed physics-based models such as SWAT (Busico *et al.* 2020), BTOPMC (Wang *et al.* 2007), MIKE SHE (Paparrizos & Maris 2017), ATHYS (Laganier *et al.* 2014), and HEC-HMS (Khatri *et al.* 2018) have been shown to perform well in runoff simulation and prediction. Nevertheless, these models usually acquire not only the observed rainfall and runoff data as the input, but also need additional data that contain the watershed's physical information. The quality of those additional data, for example, the resolution of the digital elevation model and land use will affect the simulation result (Sadeghi *et al.* 2021). Conversely, data-driven models, which only consider the statistical relationships between the input and output, do not necessarily need additional data when applied in streamflow simulation and forecasting (Parisouj *et al.* 2020). Although the physical meaning of their parameters is not yet fully understood, their excellent performance has been supported by several studies (Zuo *et al.* 2020; Zhao *et al.* 2021). Some studies also reported that the data-driven models outperformed physics-based or conceptual hydrological models (Zhou *et al.* 2019; Fan *et al.* 2020; Ji *et al.* 2021) when simulating runoff in their

This is an Open Access article distributed under the terms of the Creative Commons Attribution Licence (CC BY 4.0), which permits copying, adaptation and redistribution, provided the original work is properly cited (<http://creativecommons.org/licenses/by/4.0/>).

study areas, although the physics-based models also have the advantage, which is the good applicability in ungauged basins with limited observed flow data (Li *et al.* 2009).

Artificial neural networks (ANNs) are data-driven models that have received much attention in recent years since that their ability to fit complicated nonlinear relationships. For ANNs, different architectures result in several sub-divisions with various advancements and limitations to fulfill the need of the specific assignments (Kao *et al.* 2020). As a kind of specific sub-division, the recurrent neural networks (RNNs) calculate step by step and have a structure called 'state' to store previous information. Therefore, this kind of architecture is appropriate for time series in which the data depend on previous time steps. But the basic RNN cell also has its limitations. When dealing with long-term dependencies, gradient vanishing and explosion often occur, making the performance unsatisfactory. To solve this problem, researchers have proposed a variant of RNNs termed the 'long short-term memory (LSTM)' neural network (Hochreiter & Schmidhuber 1997). Subsequently, the gated recurrent unit (GRU) (Cho *et al.* 2014) was proposed as a variant of LSTM, which requires fewer calculations and can provide a higher training efficiency, as well as an equivalent accuracy (Chen *et al.* 2020; Shahid *et al.* 2020). These architectures' performances are superior to others when processing sequences as the result of considering long-term dependencies (Wang *et al.* 2020).

As the streamflow at the prediction target partially depends on the rainfall and upstreaming streamflow at previous time steps, the LSTM and GRU are appropriate choices for its forecasting. According to recent studies, the LSTM and GRU models can perform well when forecasting or simulating the flood or streamflow. For example, the LSTM model has been implemented in the Da River Basin, Vietnam for one-day-ahead, two-day-ahead, and three-day-ahead flood forecasting, the Nash–Sutcliffe efficiency (NSE) reaches at least 0.87 (Le *et al.* 2019); a study utilized the LSTM model in Anhe catchment in Henan province, China, for 75 flood events' forecasting, the qualification rate (QR) value of the flooding process is above 84%; the GRU and LSTM models were used for runoff prediction in the Shaxi River basin in China for 2–12 hours of prediction lead time, both the models performed well (NSE values range from 0.65 to 0.99), and required fewer time-step optimizations compared with other RNN models (Gao *et al.* 2020).

The abovementioned studies mainly focused on discussing the patterns or the applicabilities in a specific research area, however, some patterns would vary when carrying out the streamflow forecasting in basins with different properties. When applying the LSTM or GRU on streamflow forecasting, some studies found that the incorporation of rainfall data would hinder the forecasting accuracy (Le *et al.* 2019). Another study in a basin with different rainfall–runoff mechanisms found different patterns that suggested that the inclusion of the rainfall data can improve the robustness and accuracy for long-lead-time runoff forecasting as the result of the lag effect in the rainfall–runoff mechanism (Wang *et al.* 2020). In addition, the applicability and performance of a model in different kinds of regions would vary. These very recent studies have usually employed the model and focused on the patterns in a specific research area, however, the common patterns in the implementations of the model in different regions can still be explored and summarized. Also, the applicability of the GRU compared with traditional machine learning models in various types of regions still needs investigation.

To seek out the potential common patterns of the impact of assorted basins' characteristics (for instance, the area, topography, time lags between the rainfall and streamflow peaks, and rainfall–runoff correlations) on the performance of the GRU streamflow forecasting model, as well as to evaluate the applicability of the GRU in different regions compared with machine learning models, this study employed the random forest (RF), support vector regressor (SVR), and GRU for one-day-ahead ($T+1$), two-day-ahead ($T+2$), and three-day-ahead ($T+3$) streamflow forecasting in seven basins in different geographical regions in China. For the GRU model, the performance was evaluated from the aspect of training convergence, robustness, and optimized accuracy. Most importantly, the analysis was carried out quantitatively combining with the hydrological attributes of each basin and provided an insight into the performance of the model from the aspect of hydrological processes. When making the comparison between different models, the optimized models were selected. The knowledge about the specific impact of the basins' characteristics, as well as the common patterns in this model's applications in different regions, will be valuable for the practical utilization of the data-driven models on streamflow forecasting and a better understanding of the mechanism of the GRU streamflow forecasting model.

METHODOLOGY

GRU and stacked-GRU

A basic GRU cell contains two gate structures, the update gate and the reset gate. The structure of a basic GRU cell is shown in Figure 1. This structure carries out a recursive algorithm that continuously puts the hidden state into calculations. The

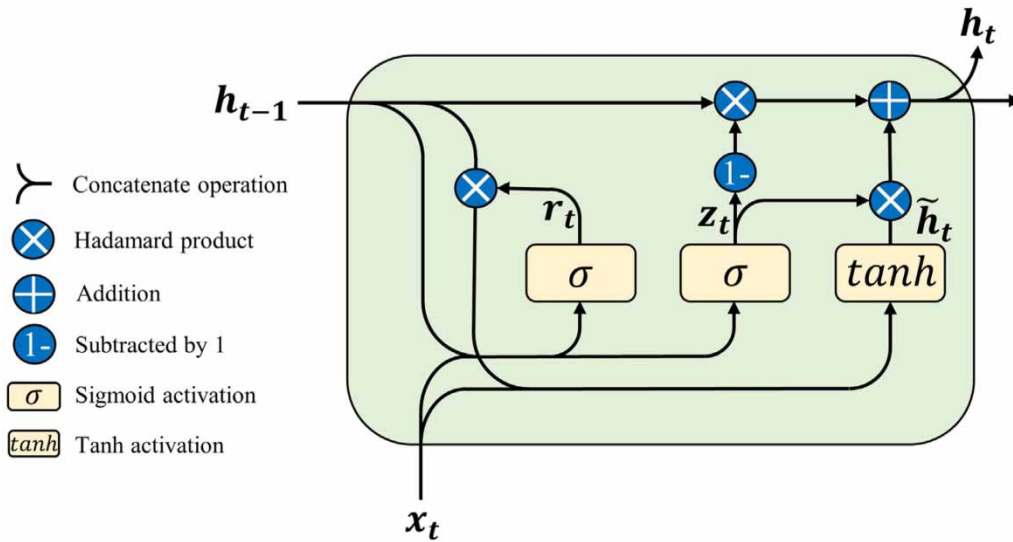


Figure 1 | The structure of a basic GRU cell.

weighting matrix results in an effect that the unimportant information at previous timesteps will be forgotten through the calculations, while the useful information will be retained. The calculation processes can be described by Equations (1)–(4):

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]) \quad (1)$$

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \quad (2)$$

$$\tilde{h}_t = \tanh(W_h \cdot [r_t \odot h_{t-1}, x_t]) \quad (3)$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \quad (4)$$

where r_t and z_t denote the output of the reset gate and the update gate, respectively. W_r , W_z , and W_h are the weighting matrix in their corresponding operations. σ represents the sigmoid activation function. \odot represents the Hadamard product (element-wise product). $[a, b]$ denotes the concatenate operation for two vectors. h_t represents the current hidden state. In these operations, the shapes of W_r , W_z , and W_h are equal to the summation of the input sequence (x_t) ('timesteps') and the length of h_t , times another hyperparameter named 'the number of neurons'. These weighting matrices act as coefficients in the linear mapping operation and are modified continuously in the training process according to the error. The nonlinear activation functions, which are \tanh and σ , enable the model to fit the complex nonlinear relationships. Typically, the more 'neurons' a GRU cell has, the better nonlinear fitting ability it will obtain, however, excessive neurons will let the overfitting problem occurs. The hidden state h_t of the cell is a row vector with the same number of columns as the input matrix. In our streamflow forecasting cases, the output of the GRU cell, which was a single value representing the predicted streamflow at the target station, was generated from the hidden state at the last time step through a mapping operation.

The whole calculation steps of a GRU cell can be presented as a 'chain' structure, and several chains can be stacked to form a stacked structure (Figure 2) to improve the nonlinear fitting ability. In the structure, the output layer uses the state of the previous layer as its input. The number of layers can also be optimized.

Benchmark models

Two classical machine learning models, which are RF (Tyrallis *et al.* 2019) and SVR (Smola & Scholkopf 2004) with a radial basis function kernel (Ji *et al.* 2021), were selected as benchmark models. RF is an ensemble learning model consisting of numerous decision tree estimators. SVR is a kernel-based algorithm that is capable of fitting either the nonlinear relationships or linear relationships.

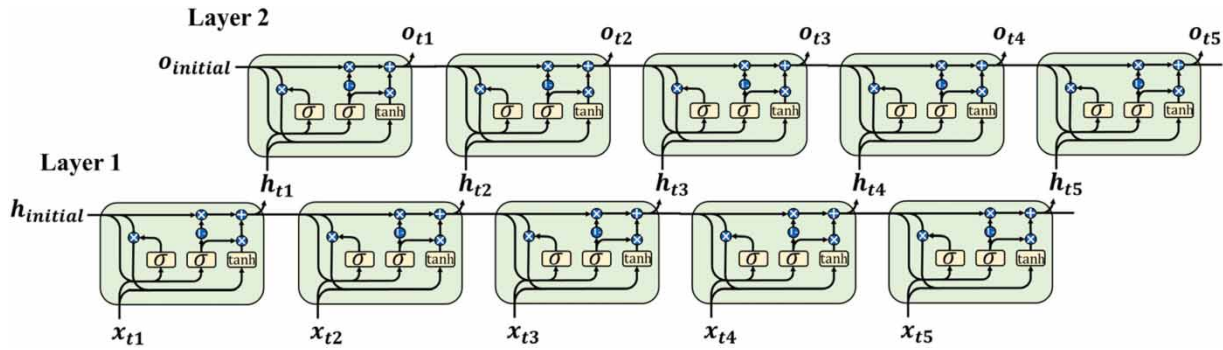


Figure 2 | The structure of stacked-GRU with two layers.

Data preprocessing

In streamflow forecasting tasks using data-driven models, the input data set can include the observed streamflow data, rainfall data, evaporation data, and other variables that are closely related to the future streamflow.

For the GRU model, the observed data should be in a time-series format. In this study, the raw data set was prepared as column vectors; each column represents the data at a specific hydrological/meteorological station (or reservoir), while each number in the column vector is a record at a particular time step. Then, the column vectors were concatenated as a matrix. The sliding window method (Figure 3) was used to generate sample data X in a matrix and sample data Y (the observed data) in a vector. In this method, the number of the records in a sample is a hyperparameter named ‘timesteps’ that were optimized during the training and validation processes. Missing values in the time series are accepted, but the samples with missing values would better be discarded in case of interfering with the training of weighting matrices, although interpolation methods can be considered in some cases (Zhao *et al.* 2021).

The input of two benchmark models was the streamflow and rainfall data in a single time step since the inclusion of data in multiple time steps caused a high dimension input matrix, which resulted in bad forecasting performances.

Before the training process, the rainfall and streamflow data were normalized with Equation (5), which is called Min–Max normalization. The normalized data has a value range of [0,1]:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (5)$$

where, x' denotes the normalized data, $\min(x)$ and $\max(x)$ are the minimum and maximum values of each column in the input data set.

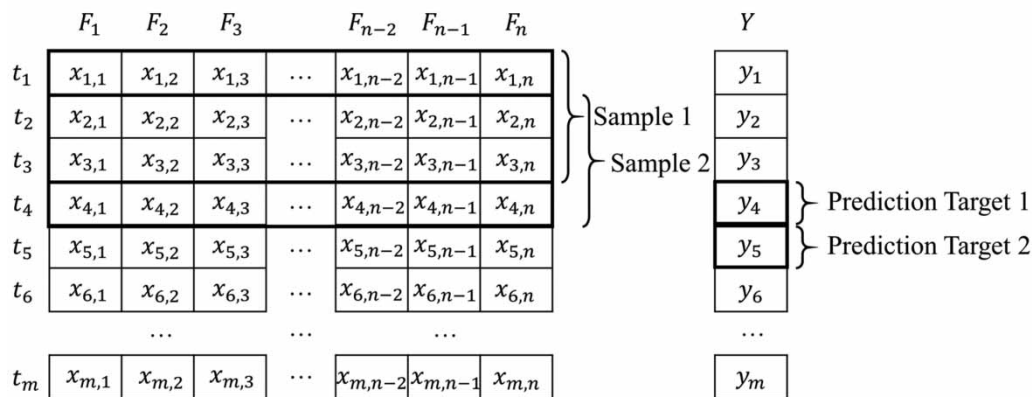


Figure 3 | An example of the T+1 sliding window sampling method. Using the sample matrix as the input to predict the daily streamflow one day after the last time step. For the T+2 or T+3 forecasting, the prediction targets are shifted to the data at the corresponding timestep.

Model evaluation

Nash–Sutcliffe efficiency (NSE), root mean square error (RMSE), and mean absolute error (MAE) has been used in the overall accuracy evaluation in each simulation, as shown in Equations (6)–(8), respectively. The mean absolute percentage error (MAPE) as shown in Equation (9) has been used for flood peak evaluations:

$$NSE = 1 - \frac{\sum_{i=1}^n (O_i - P_i)^2}{\sum_{i=1}^n (O_i - \bar{O})^2} \quad (6)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (O_i - P_i)^2} \quad (7)$$

$$MAE = \frac{1}{n} \times \sum_{i=1}^n |O_i - P_i| \quad (8)$$

$$MAPE = \frac{100\%}{n} \times \sum_{i=1}^n \left| \frac{O_i - P_i}{O_i} \right| \quad (9)$$

where, O_i is the observed streamflow, P_i is the predicted streamflow, \bar{O} is the mean value of the observed streamflow, n is the total number of observations. The NSE values range from $-\infty$ to 1. A higher NSE value indicates a better fitting result. Both the RMSE and MAE range from 0 to $+\infty$ (with a unit of m^3/s in this study), the lower are the RMSE and MAE, the better the model performs. The RMSE is more sensitive to abnormal values than the MAE. The MAPE has a value ranging from 0% to 100%.

CASE STUDIES

Study areas

Seven basins, including the Wei River basin (WRB) (104° E–110° E, 33° N–37° N), the Liao River basin (LRB) (117° E–125° E, 38° N–44° N), the Wujiang River basin (WJRB) (104° E–109° E, 26° N–30° N), the Luan River basin (LURB) (115° E–120° E, 39° N–43° N), the Xiangjiang River basin (XJRB) (110° E–115° E, 25° N–29° N), the Dongjiang River basin (DJRB) (114° E–116° E, 22° N–26° N), and the Huai River basin (HRB) (112° E–121° E, 31° N–36° N) were selected from different geographic regions with different climate types. The properties of the selected research areas, which include the basin area, geographic region, climate type, mean annual temperature, the period of the rainy season, annual rainfall, annual evaporation, and the length of the main river channel, are shown in Table 1. The topography, the locations of the collected hydrological/

Table 1 | Properties of the selected research areas

Basin	Area (km ²)	Geographic region	Climate type	Mean temperature (°C)	Rainy season	Annual rainfall (mm)	Annual evaporation (mm)	Length of main channel (km)
WRB	1,36,013.7	NW	Continental monsoon	7.8–13.5	Jun–Oct	350–700	660–1,600	818
LRB	2,20,938.4	NE	Continental monsoon	7.0–13.0	Jun–Sep	352–954	850–1,200	1,345
WJRB	87,918.7	SW	Subtropical monsoon	13.0–17.5	May–Sep	900–1,400	500–800	1,018
LURB	44,750.0	N	Continental monsoon	7.6	Jul–Sep	500–600	950–1,150	877
XJRB	94,240.9	C	Subtropical monsoon	16.0–18.0	Apr–Jun	1,400–1,700	1,200–1,700	948
DJRB	27,402.9	S	Subtropical monsoon	21.0	Apr–Sep	1,500–2,400	1,400–1,600	562
HRB	1,59,275.1	E	Climatic transition	13.2–15.7	Jul–Sep	600–1,600	750–1,300	1,000

Note: NW, Northwest China; NE, Northeast China; SW, Southwest China; N, North China; C, Central China; S, South China; E, East China.

meteorological stations, and the main reservoirs' outlets in each basin are shown in Figure 4. The last hydrological station near the outlet in each basin has been set as the prediction target for streamflow forecasting.

Except for the differences and similarities between the selected basins that can be found in Table 1 and Figure 4, other elements are of note. From the aspect of topography, the Wujiang River basin is a karst mountainous basin (75.6% of which consists of carbonate rock) with few plain areas, while the LRB and the HRB have large plain areas. Other basins fall in between. For the reservoir regulations, the LURB differs from others as streamflow at its downstream region is dominated by the Taolinkou (which controls approximately 11.3% of the watershed area) reservoir and the Daheiting reservoir (which controls approximately 78.4% of the watershed area).

Data

In this study, the collected data include the daily streamflow data from 2007 to 2014 at the hydrological stations and some main reservoirs (extracted from the 'Hydrological Data Yearbook of the People's Republic of China'), as well as the daily rainfall data at the meteorological stations (extracted from China Meteorological Data Service Center, <http://data.cma.cn>). Snowfall is not considered in this study, while the impact of snow melting would be recorded in the data at the stations located upstream of the prediction targets. Table 2 shows a summary of the rainfall and streamflow data set. Table 3 presents a summary of the Pearson correlation coefficient (R) between the input series and the forecasting target series with different lead times in each basin. Obvious regional differences of the data, as well as the temporal variation of the relationships between the input and forecasting targets, can be found according to the given tables.

Data Set partitioning, scenarios, and model settings

After the sampling process, the whole data set (with 2,909 samples) was divided into a training set, a validation set, and a testing set according to an approximate ratio of 6:2:2 (Song *et al.* 2020). The training set was used to modify the weighting matrix to fit the relationships between input and output. The validation set was used to evaluate the model's robustness to various hyperparameters combinations and to select appropriate hyperparameter combinations with relatively accurate

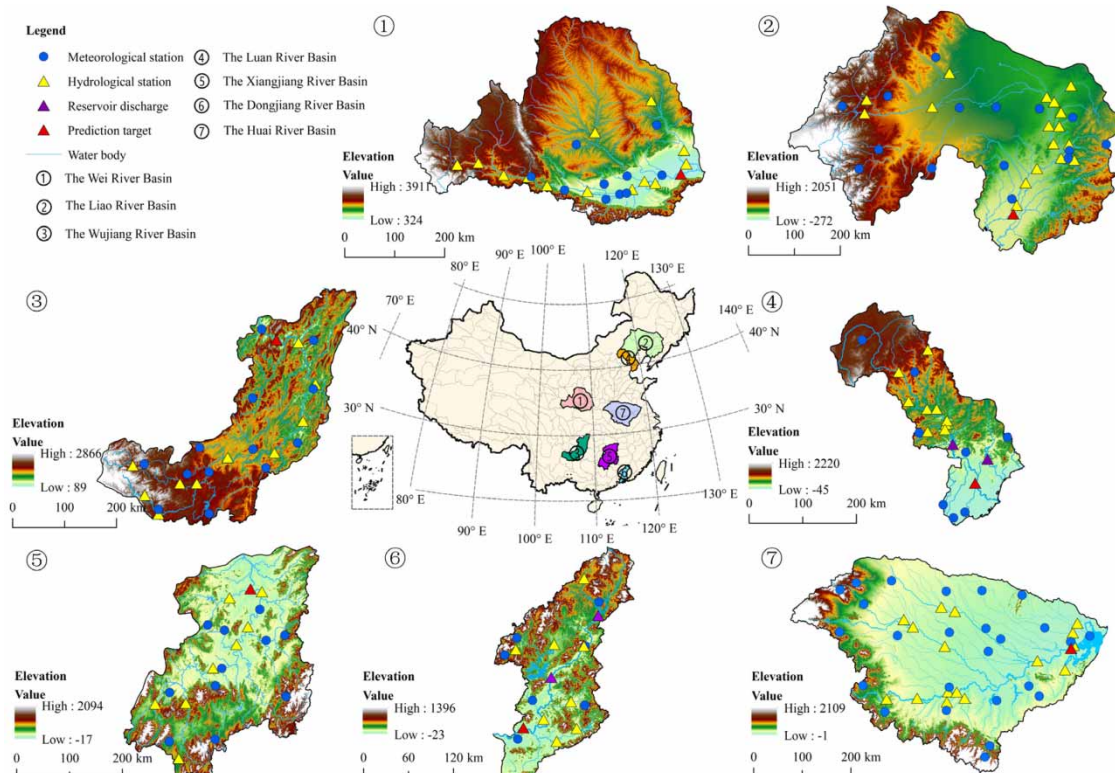


Figure 4 | A summary of the study areas, including the topography, hydrological stations' locations, meteorological stations' locations, reservoirs' discharges, and the selected prediction target of each basin.

Table 2 | Summary of the rainfall and streamflow data set in each basin

Basin	Num. of hydrological stations (streamflow data)	Num. of meteorological stations (rainfall data)	Rainfall (mm/d)		Streamflow (m ³ /s)	
			Mean	Max	Mean	Max
WRB	14	10	1.60	62.34	171.73	4,410
LRB	18	15	1.29	55.18	98.61	1,840
WJRB	11	13	2.91	63.35	1,374.89	15,400
LURB	12	8	1.47	73.95	33.23	3,410
XJRB	9	11	3.92	82.15	1,868.08	18,400
DJRB	11	7	4.65	103.51	698.15	7,360
HRB	17	25	2.34	70.51	692.13	7,740

Table 3 | A summary of the Pearson correlation coefficients between the input series at different hydrological/meteorological stations and the forecasting target series in each basin

Basin	Lead time	Rainfall		Streamflow	
		Mean	Max	Mean	Max
WRB	T+1	0.22	0.26	0.63	0.96
	T+2	0.41	0.48	0.60	0.87
	T+3	0.42	0.50	0.52	0.73
LRB	T+1	0.09	0.15	0.57	0.99
	T+2	0.13	0.20	0.58	0.96
	T+3	0.17	0.26	0.58	0.94
WJRB	T+1	0.32	0.39	0.59	0.91
	T+2	0.31	0.39	0.54	0.80
	T+3	0.28	0.36	0.49	0.74
LURB	T+1	0.15	0.31	0.39	0.89
	T+2	0.18	0.30	0.34	0.73
	T+3	0.18	0.28	0.30	0.57
XJRB	T+1	0.32	0.38	0.80	0.95
	T+2	0.39	0.45	0.75	0.84
	T+3	0.37	0.44	0.67	0.74
DJRB	T+1	0.46	0.53	0.75	0.92
	T+2	0.45	0.52	0.66	0.79
	T+3	0.37	0.43	0.57	0.70
HRB	T+1	0.15	0.20	0.59	0.98
	T+2	0.17	0.22	0.57	0.96
	T+3	0.19	0.23	0.59	0.93

results. The testing set was for evaluating the capability for extrapolating (Ripley 2014) of the model with selected hyperparameter combinations, also for evaluating the overall performance combined with the modelling results in the validation set. Two input scenarios S1 (with rainfall data) and S2 (without rainfall data) were set to compare the impact of the rainfall data in different basins for different lead-time (T+1, T+2, and T+3) forecasting. For the GRU model, Adam optimizer was selected in weighting matrices' training, 72 hyperparameters' combinations were set for grid search-based hyperparameter tuning in the validation set. A summary of the GRU model settings can be found in Table 4. For the benchmark models, the hyperparameters were optimized by the sparrow search algorithm (Xue & Shen 2020) with a population size of 10 and an iteration of 500 in the training stages. The searching spaces of the hyperparameters of these two models are given in Table 5. To make a comparison between the GRU model and benchmark models, the input selection scenario which had the best performance for each lead time was selected.

Table 4 | A summary of the GRU model settings and hyperparameters searching spaces in the forecasting cases in each basin

Num. of layers	Epoch	Batch size	Learning rate	Timesteps	Num. of neurons		
					Layer 1	Layer 2	Layer 3
1	500	10	0.005	3–10	1	–	–
2					16, 32, 64, 128	1	–
3					16, 32, 64, 128	16	1

Table 5 | Hyperparameters searching spaces of RF and SVR

Model	Hyperparameters	Lower bound	Upper bound
RF	Max. depth	1	100
	Num. estimators	1	3,000
	Min. samples split	2	10
	Min. samples leaf	1	10
	Max. features	1	Num. of stations
SVR	Gamma	0	100
	C	0	100

RESULTS AND DISCUSSION

Input scenarios' impacts on convergence and robustness of the GRU model

The convergence of the model with different input scenarios and lead times is assessed through the training loss. Figure 5 shows the bands of training losses of (the top 50%) models according to their performances in the training stage. In all the basins, the converging process slowed down with the increased lead time. In addition, the models which included the rainfall data tended to obtain more convergent training loss curves compared with the models without the rainfall data. The advantages of the rainfall data's inclusion are not evident and are even reversed in T+1 forecasting cases, except in the WJRB. The results indicate that the inclusion of rainfall data will enhance the training convergence of the model for a relatively long-lead-time forecasting task.

The robustness of the model is evaluated in the validation stage. A robust model tends to obtain a higher NSE value, lower RMSE value, and a lower MAE value compared to other models in the validation stage; meanwhile, the value ranges of its metrics tend to be less sensitive to the changing hyperparameters. A robust model will need less effort on optimization and will have less uncertainty. Figure 6(a)–6(c) shows the distribution of NSE, RMSE, and MAE in the validation set 6. The outliers in the Figures are represented by cross markers and are the metrics with a value lower than the lower quartile minus 1.5 interquartile range (IQR) or higher than the upper quartile plus 1.5 IQR (Schwertman & de Silva 2007). In those figures, the descending NSE value ranges, ascending RMSE value ranges, as well as ascending MAE value ranges are observed in all the basins when the lead time of the forecasting increases. The mean and median values of these metrics also follow the same pattern. This phenomenon has also been reported in relevant studies that employ the LSTM or GRU for streamflow forecasting (Le *et al.* 2019; Kao *et al.* 2020; Wang Q *et al.* 2020). It is caused by the increasing amount of useful information contained in the data at the previous timesteps being excluded with the increased lead time. A shorter lead time makes the forecasting model more robust. The impact of the rainfall data on the robustness of the model varies in different basins and the forecasting with different lead times. According to Figure 6, the rainfall data resulted in the higher median and mean NSE values, lower median and mean RMSE values, as well as lower median and mean MAE values in the WRB, the WJRB, the XJRB, and the DJRB. Similar to the pattern found in training losses, this phenomenon became more significant when the lead time of the forecasting increased, although it was not evident in the T+1 forecasting in the WRB. This can be partially explained by the time lag effect of the hydrologic responses (Ross *et al.* 2019; Iwasaki *et al.* 2020). The time lag of the rainfall-runoff events makes the rainfall data meaningful to the target streamflow data. In these basins, the models which include the rainfall data are more robust than the ones without the rainfall data. By contrast, the inclusion of the rainfall data

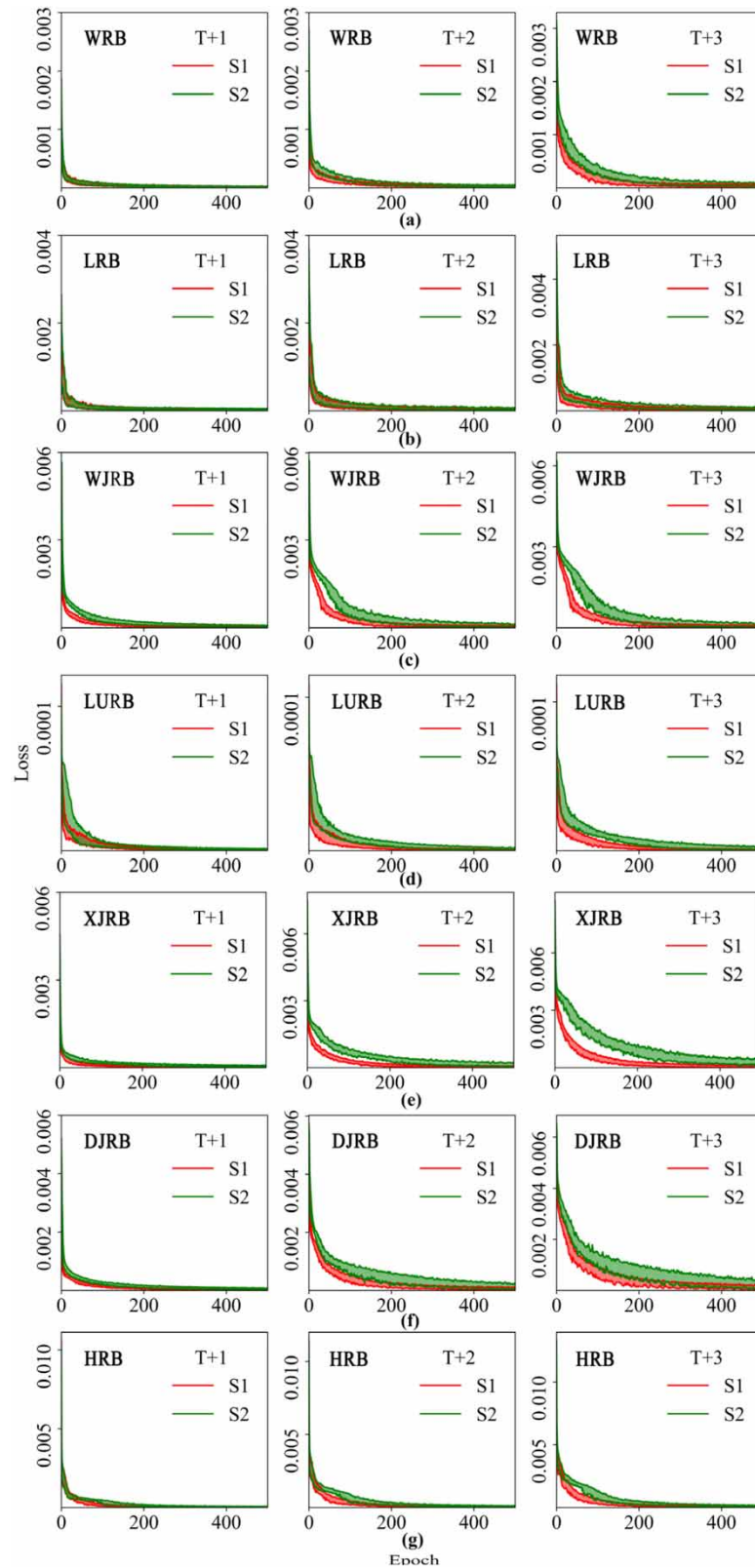


Figure 5 | The band of training losses of the GRU model (top 50% according to the performances) under different input scenarios in each basin.

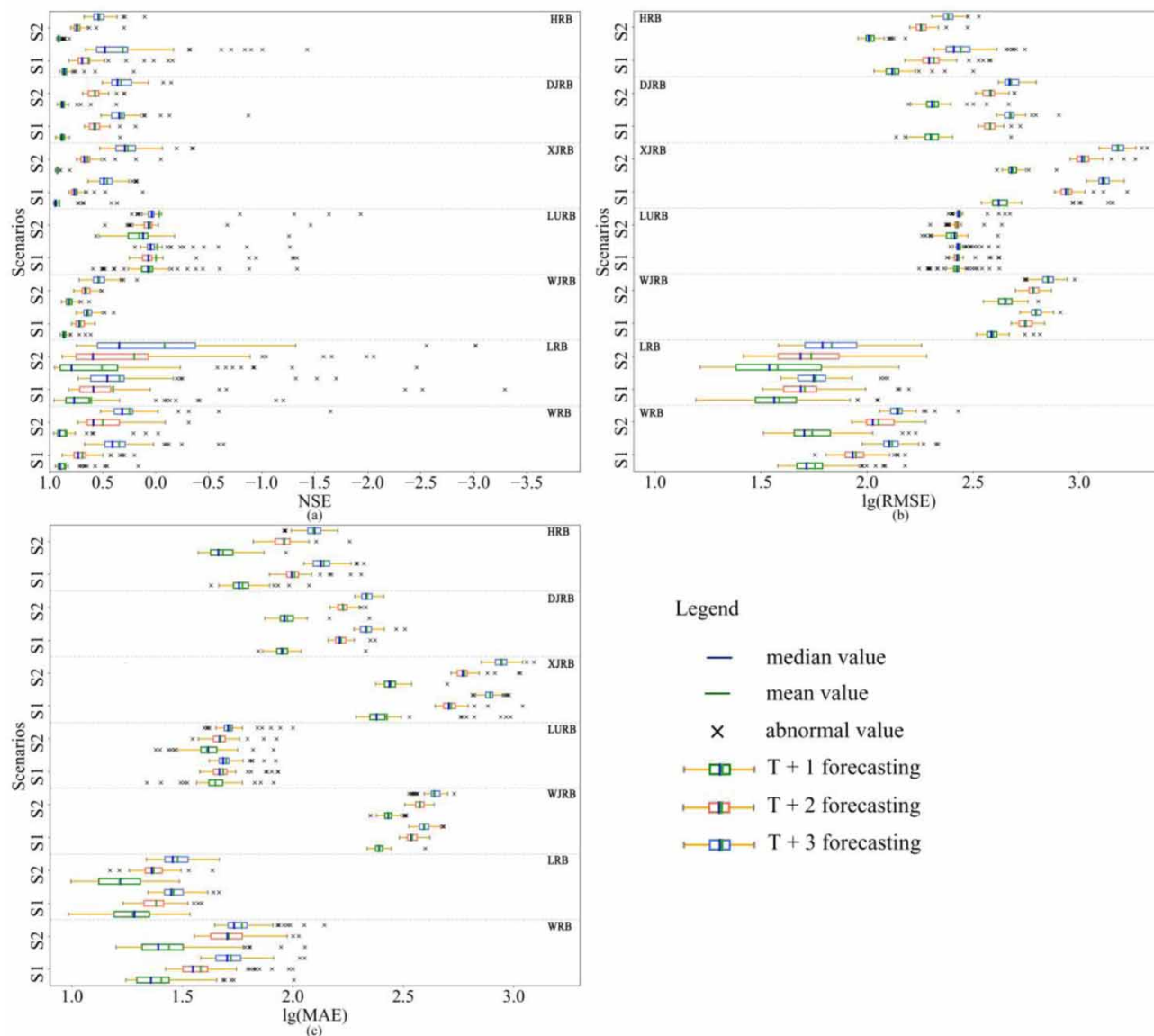


Figure 6 | The distribution of NSE, RMSE, and MAE in the validation set. The values of RMSE and MAE are presented as logarithms of 10 (lg).

significantly interferes with the robustness of the forecasting model in the HRB, the distributions of errors (including RMSE and MAE) had higher mean and median values, while the NSE had lower mean and median values when the rainfall data were incorporated into the training process (S1). This pattern can be found in all the lead time cases in this region. Also, the model with rainfall data yielded more outliers in its evaluation metrics. For the LRB, the impact of the rainfall data had no significant pattern, although the distribution of the evaluation metrics is more aggregated when including the rainfall data in forecasting, the median values of the metrics showed advantages when excluding the rainfall data in T+1 and T+2 forecasting. The LURB is a special case, the distributions of the NSE showed that the fittings are not satisfactory in each forecasting case. No significant impact of the inclusion of the rainfall data on the robustness has been observed.

The patterns of robustness can be related to the correlations between the input and forecasting target dataset. According to Table 3, the Pearson correlation coefficients between the input and forecasting target streamflow series revealed a descending trend as the lead time increases. This trend is positively related to the above mentioned trend of NSE ranges, and negatively related to the trends of RMSE and MAE ranges simultaneously with the increased lead time. The various impacts of the rainfall data in different basins can also be explained from the aspect of correlations. According to Table 3, the correlation

coefficients between the rainfall data and the forecasting target streamflow data are significantly higher (in T+2 and T+3 forecasting cases) in the WRB (a mean value of 0.41 and 0.42, respectively), the WJRB (a mean value of 0.32, 0.31, 0.28 for the corresponding forecasting cases), the XJRB (a mean value of 0.32, 0.39, and 0.37 for the corresponding forecasting cases), and the DJRB (a mean value of 0.46, 0.45, and 0.37 for the corresponding forecasting cases) than in other basins. Nevertheless, for the T+1 forecasting in the WRB, the correlation coefficient is lower (a mean value of 0.22). While in the HRB, the LRB, and the LURB, in which the inclusion of the rainfall data hinders the robustness of the model or does not have a significant pattern, the correlation coefficients of the rainfall data are relatively low, the mean value ranges from 0.09 to 0.19 in all cases. These phenomena suggest that the impact of the rainfall data on the models' robustness is closely related to the correlation between the collected rainfall series and the forecasting target streamflow series. The potential reasons for the relatively low correlation coefficients between the rainfall and outlet streamflow in these three regions are that both the HRB and LRB are large plain basins, in which the local rainfall does not dominate the streamflow at the outlet, and the LURB is dominated by reservoirs that result in the interfered rainfall-runoff relationship.

Therefore, in large plain basins and reservoir-dominated basins with low rainfall-runoff correlations, the rainfall data are not suggested to be directly included in the model's training stage, although it can improve the convergence in the training stage. For the long-lead-time forecastings in other types of basins, the rainfall data are suggested to be considered to not only improve the convergence but also improve the probability to obtain a robust model. Moreover, the correlation evaluation can be executed in the vector selection stage to get a more robust model, which tends to obtain better fitting results and is less sensitive to the interference of the various hyperparameter combinations.

Overall accuracy evaluation

Table 6 shows the overall evaluation metrics of the best RF, SVR, and GRU model (the hydrographs of the simulation results for different lead times and scenarios in the training set, validation, and testing set in each basin can be found in Supplementary data Appendix A) that consider the accuracy in both the validation set and testing set to make sure that the model is accurate and has good generalization ability. Except for the LURB, the GRU model performed well in all the other basins

Table 6 | Overall evaluation metrics in the validation set and testing set for different models.

Basin	Lead time	NSE			RMSE (m ³ /s)			MAE (m ³ /s)		
		RF	SVR	GRU	RF	SVR	GRU	RF	SVR	GRU
WRB	T+1	0.81	0.89	0.94	89.65	67.18	48.22	45.45	28.47	27.03
	T+2	0.75	0.80	0.85	102.36	90.47	78.34	45.41	36.18	33.97
	T+3	0.68	0.68	0.63	115.20	116.62	120.61	53.00	43.63	47.16
LRB	T+1	0.94	0.98	0.96	35.03	20.92	19.91	17.38	11.47	12.43
	T+2	0.93	0.94	0.91	39.48	34.15	35.50	20.81	16.98	16.84
	T+3	0.89	0.89	0.84	48.49	48.19	41.67	23.91	22.40	23.30
WJRB	T+1	0.81	0.65	0.88	576.01	776.20	437.92	306.55	324.43	246.46
	T+2	0.70	0.56	0.71	715.97	870.65	701.12	376.74	387.09	361.73
	T+3	0.60	0.48	0.64	821.93	940.69	784.48	432.95	423.70	381.74
LURB	T+1	0.07	0.02	0.63	188.35	194.05	107.91	27.95	28.79	14.07
	T+2	0.08	0.01	0.32	187.53	195.12	135.96	27.38	31.88	27.17
	T+3	0.04	0.01	0.27	191.71	195.45	131.52	30.44	32.76	26.72
XJRB	T+1	0.95	0.90	0.95	394.36	561.20	357.37	217.71	296.02	201.33
	T+2	0.85	0.73	0.81	680.73	902.56	729.40	368.51	464.38	405.76
	T+3	0.68	0.57	0.63	983.47	1,138.93	1,037.44	546.98	591.71	601.99
DJRB	T+1	0.88	0.81	0.93	237.72	298.50	179.39	107.12	101.76	87.61
	T+2	0.72	0.65	0.69	361.82	407.87	376.71	165.97	153.40	178.98
	T+3	0.57	0.50	0.46	452.14	485.82	504.32	212.93	188.72	246.84
HRB	T+1	0.83	0.93	0.94	198.53	127.70	105.08	103.57	69.74	50.16
	T+2	0.73	0.81	0.83	247.51	207.76	181.78	131.80	105.99	88.13
	T+3	0.64	0.69	0.73	289.21	268.67	234.04	158.07	135.44	123.24

The value is bold when GRU outperforms others.

and has the NSE values range from 0.88 to 0.96 for T+1 forecasting, 0.69 to 0.91 for T+2 forecasting, and 0.46 to 0.84 for T+3 forecasting. The best fitting result occurs in the LRB (T+1 forecasting has an NSE of 0.96, an RMSE of $19.91 \text{ m}^3/\text{s}$, and an MAE of $12.43 \text{ m}^3/\text{s}$), while the worst fitting occurs in the LURB, in which the best forecasting result only has an NSE of 0.63 (T+1). The model can achieve a good result even for long-lead-time forecasting in the LRB (T+3 forecasting has an NSE of 0.84, an RMSE of $41.67 \text{ m}^3/\text{s}$, and an MAE of $23.30 \text{ m}^3/\text{s}$) and the HRB (T+3 forecasting has an NSE of 0.73, an RMSE of $234.04 \text{ m}^3/\text{s}$, and an MAE of $123.24 \text{ m}^3/\text{s}$). But in the basins other than these two, the accuracy of the forecasting results decreased significantly faster when the lead time was increased, especially in the LURB (the best NSE value of T+2 and T+3 forecasting were only 0.32 and 0.27, respectively) and the DJRB (the best NSE value of T+2 and T+3 forecasting were 0.69 and 0.46, respectively). One of the reasons for the unsatisfactory forecasting results in the LURB is that the streamflow at the prediction target is mainly controlled by two adjacent upstream reservoirs (the locations have been marked in Figure 4). The manually controlled runoff processes are significantly different from the natural process. In addition, the reservoir regulations in flood seasons are different from those in dry seasons, thus, those differences result in interferences and uncertainties. Although other basins also have a considerable amount of reservoirs, their size and the upstreaming locations of those reservoirs make the dominance of these reservoirs unlike the two in the LURB. Another reason is the extreme flood events in the validation set and the testing set. According to gathered data, the peak value of the daily streamflow that occurred on 3 August 2012 reached $3,410 \text{ m}^3/\text{s}$, and another event that occurred on 10 July reached $617 \text{ m}^3/\text{s}$. However, the maximum peak value in the training set is only $266 \text{ m}^3/\text{s}$ which occurred on 1 August 2011. The rainfall-runoff processes in those two flood events in the validation set and testing set are significantly different from the processes in the training set (see the hydrographs in the appendix). This finding indicates that the applicability of the GRU forecasting model may be confined in some basins, in which the streamflow is dominated by manual control, especially when the control point is close to the prediction target. Additionally, the lack of similar flood events for training can also introduce uncertainty and may cause serious underestimations. The relatively slow decreasing trend of the forecasting accuracy along the lead time in the LRB and HRB, as well as the fast trend in the LURB and DJRB, is partly attributed to the length of the main channel. According to Table 2, the length of the main channel of the Liao River and Huai river is over 1,000 km, while the length of the Dongjiang River is only 562 km. Although the Luan River also has a total length of 877 km, the streamflow at the prediction target is dominated by two reservoirs close to the downstream regions, thus, the actual length of the channel should be considered from the control points to the prediction target. The streamflow tends to spend more time to the outlet in long river channels, therefore, in the LRB and the HRB, the streamflow data at upstream stations several days ago can be meaningful for long-lead-time forecasting. However, in the LURB and the DJRB, the upstream flows to the outlet in a relatively short period and leads to difficulty in long-lead-time forecasting.

Compared with RF and SVR, the GRU model significantly outperformed others in all the forecasting cases in the WJRB, the HRB, and the LURB. Interestingly, the RF and SVR were totally not capable of forecasting the streamflow in the LURB, while the GRU model could capture the patterns in the historical information to some extent (with an NSE of 0.63, an RMSE of $107.91 \text{ m}^3/\text{s}$, and an MAE of $14.07 \text{ m}^3/\text{s}$). The GRU model will potentially perform better if the necessary reservoir operation factors are considered, or more flood events are included in the training stage. In the WRB, the GRU model outperformed others in T+1 and T+2 forecasting cases. In the DJRB and the XJRB, the GRU model performed the best in T+1 forecasting. Overall, it can be concluded that, although the selected machine learning models can also perform well, the GRU model is outstanding in more cases (13 in 21) as the result of its ability to learn the context information in a time series.

Compared with other studies, the GRU model obtained similar performances to LSTM and tended to have better accuracy than other data-driven models. In addition, both of them are superior to physics-based models from the perspective of evaluation metrics. For example, a study compared the SWAT and LSTM in the XJRB, the result showed that the performance of LSTM is significantly better (Fan *et al.* 2020). This is mainly because that the physics-based models are generalizations of the real world with a lot of approximations, in contrast, the data-driven models directly learn the relationships between the observed data, although with errors. However, the physics-based models also have the advantage, which is that they do not need plenty of observed streamflow for calibration. Therefore, they have been widely used in ungauged basins (Piman & Babel 2013).

Accuracy of streamflow peaks' forecasting

Streamflow peaks, especially the flood peaks are the portion with most of the concerns in streamflow forecasting. To evaluate the accuracy of GRU when forecasting the flood peaks, the maximum streamflow of all the flood events in each year in the

validation and testing stage (2012–2014) has been selected as representative flood peaks. In addition, all the rainfall and their corresponding flow peaks in the validation and testing period were recognized, the time lags between them were calculated according to the definition in the published literature (Lombard & Holtzschlag 2018) to explore the performances of the GRU model on different types of rainfall-runoff events. Table 7 shows the dates, the observed streamflow values, and the MAPE of three selected flood events of RF, SVR, and GRU for all the forecasting cases in each basin. The pattern of the MAPE is similar to the patterns of RMSE and MAE that have been described in section ‘Overall Accuracy Evaluation’. Of note, the MAPE in the Wujiang River basins for T+1 and T+2 forecasting are significantly higher than in other basins (except for the LURB that obtains bad fitting results). Even the best case in T+1 forecasting in this basin has the MAPE over 30%. The result suggests that during the flood events, the correlations between the streamflow data at the upstreaming stations and the one at the prediction target station are not good. The probable reason is that the Wujiang River basin is located in a karst mountainous region (Hou & Gao 2019), which is underlain by carbonate rocks (Yang *et al.* 2020). It is also an area with heterogeneous geological structures including countless preferential underground flow paths, sinks, springs, and ponors (Sezen *et al.* 2019). The temporal variability of recharge and hydraulic connectivity (Zhou *et al.* 2019) in this kind of region results in it being difficult to streamflow forecast. In some rainfall-runoff events, the relationships between the input and forecasting target can differ from the statistical relationships used by the model and can cause errors. Gao (2012) employed the backward propagation neural network for streamflow forecasting in two karst mountainous sub-basins in the upstream region of the Wujiang River basin; the determination coefficients (R^2) of 0.538 and 0.420 were obtained, and the model tended to underestimate flood events. Darras *et al.* (2015) have also reported inaccurate flash flood forecasting results by neural networks in the Lez River basin, which is a karst basin in southern France. They suggest that the karst discharge has a different dynamic than the surface discharge, and can have a different time lag effect due to the saturation of the hydrosystem prior to the event. Hence, although the overall accuracy of the streamflow forecasting in the Wujiang River basin

Table 7 | The dates, observed values, and mean absolute percentage error of the selected flood peaks forecasted by different models in each basin

Basin	Date (month/day/year)	Observed streamflow (m ³ /s)	MAPE (%)								
			T+1			T+2			T+3		
			RF	SVR	GRU	RF	SVR	GRU	RF	SVR	GRU
WRB	9/3/2012	2,020	19.95	34.97	8.30	24.41	55.28	51.40	61.66	72.37	74.20
	7/24/2013	2,200									
	9/17/2014	1,520									
LRB	8/7/2012	498	16.67	14.12	5.80	27.14	31.23	10.30	28.93	41.79	28.80
	8/22/2013	1,340									
	6/22/2014	428									
WJRB	6/4/2012	5,710	42.12	60.13	30.60	60.02	66.79	54.50	65.18	70.20	65.70
	9/13/2013	4,310									
	7/17/2014	15,400									
LURB	8/3/2012	3,410	66.23	64.84	43.00	62.49	68.01	53.60	71.63	74.38	64.20
	7/10/2013	617									
	5/13/2014	57.4									
XJRB	6/13/2012	11,600	19.68	35.43	6.20	30.45	52.16	30.50	54.62	61.33	50.80
	8/25/2013	9,590									
	5/26/2014	11,900									
DJRB	4/29/2012	2,530	25.54	39.49	17.50	39.61	57.69	39.50	63.31	72.18	64.90
	8/18/2013	7,360									
	5/21/2014	4,570									
HRB	9/12/2012	2,700	14.35	6.55	4.40	25.50	26.36	14.00	34.00	38.94	33.60
	9/28/2013	1,960									
	9/30/2014	3,240									

The bold values denote the lowest MAPE in their corresponding cases.

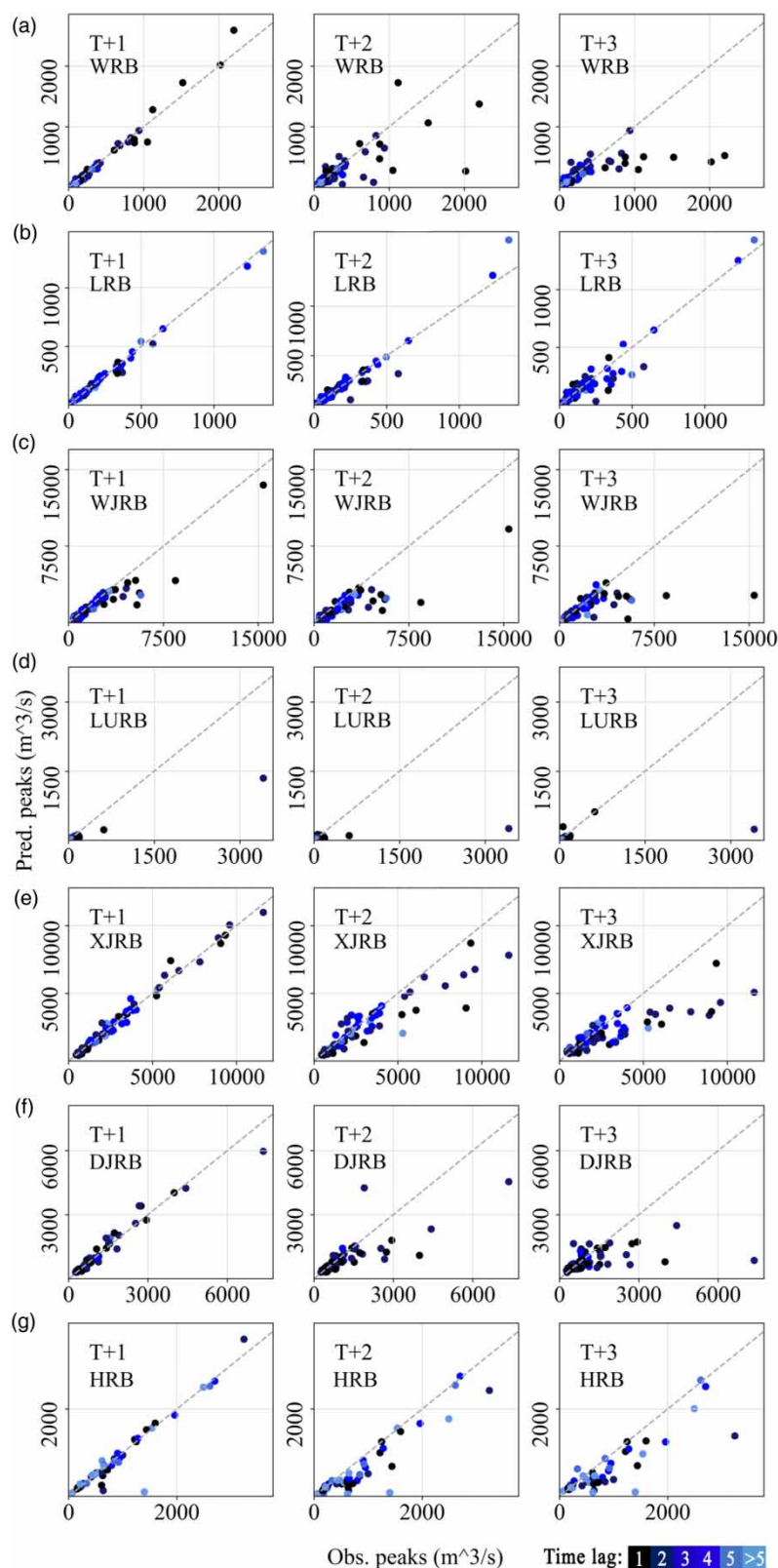


Figure 7 | The streamflow peaks with different time lags and their fitting results of the GRU model in each basin.

is acceptable (according to Table 5), the modelling process should still be improved by considering the specific dynamics in the karst basins.

Compared with the benchmark models, the GRU model outperformed in most of the forecasting cases, especially in T+1 forecastings. In long-lead-time forecastings in some basins (WRB, T+2 and T+3; WJRB, T+3; XJRB, T+2; DJRB, T+3), the performances of the RF model were better than the GRU. That is caused by the mechanisms' difference between these two models. The RF model consists of decision trees with several judgments operations and a voting mechanism, the core principle of it is 'the similar input produces the similar output'. The GRU model is mainly based on the context information in the time series. For an extreme flood event, the response of the runoff process occurs rapidly, thus, some important context would be lacking in the input series. However, the principle of the RF would make it have a relatively poor generalization ability, especially when the similar samples are not sufficient in the training stage. That is also an explanation of the significant out-performance of the GRU in the LURB.

Figure 7 shows the recognized streamflow peaks versus the GRU models' prediction results, and the time lags of the streamflow peaks to the corresponding rainfall events represented by markers with different colors. The GRU model obtained accurate forecasting results for streamflow peaks in all the basins except for the LURB. When the lead time increased, the underestimations on peak flows with short time lags (1–2 days) and large volumes became more evident. For the basins, which have more streamflow peaks with short time lags (for example, the WRB and the DJRB), the performances on their estimations deteriorated rapidly. In contrast, in the large plain basins, such as the LRB and HRB, that are dominated by long-time-lag rainfall-runoff events, the model's accuracy can remain satisfactory even in T+3 forecastings. This pattern also well supports the discussed reason for the overall accuracies and deterioration in the WRB, the XJRB, and DJRB, as well as the good performances in the HRB and the LRB for long-lead-time forecasting cases. There are two potential ways to enhance the model's performance for long-lead-time forecasting. The first one is collecting more flood events and establishing the event-based models, also, an error rectification model based on ensemble learning would be effective.

CONCLUSIONS

In this study, the GRU model was employed to forecast the streamflow for different lead times in seven basins to investigate the model's data selection effects and performances affected by the basins' characteristics, also the performances of the GRU model were compared with two benchmark machine learning models including RF and SVR. The results have been evaluated from three aspects, including the convergence and robustness of the model, the overall accuracy of the model, and the accuracy of streamflow peaks forecasting. In summary, the main findings include:

- (1) The trend of correlation coefficients of the upstream streamflow series along the lead time controls the trend of evaluation metrics' value ranges. The correlations between the rainfall series and forecasting target series, which are affected by the basin's characteristics, are positively related to the rainfall data's impact on the robustness of the model. Although the rainfall data can improve the convergence of the training loss curve, it is not recommended to be included in the correlation between it and the target streamflow series is not good.
- (2) The GRU streamflow forecasting model performs well in most of the basins, but for those basins in which the prediction target is close to and dominated by reservoirs, the performance will be significantly affected. Compared with RF and SVR, the GRU model tends to outperform others, especially in T+1 forecasting cases.
- (3) The deterioration of streamflow peaks' prediction accuracy of the GRU model with the increased lead time mainly depends on the regional patterns of the time lag of rainfall-runoff events, in large plain basins with more long-time lag rainfall-runoff events, the model tends to perform better in long-lead-time forecastings. For flood peaks forecasting, the GRU model outperforms other models in most of the forecasting cases, although the RF may perform better in some long-lead-time forecasting cases.

These findings are meaningful for application and study in other regions. In the model selection stage, correlation analysis can be carried out to preliminarily assess the model's applicability for different lead-time forecasting; also, the rainfall data's impact can be judged in the vector selection process. For the streamflow peaks' and flood peaks' forecasting, the patterns observed are a valid reference. There remain some topics for further investigation. For example, the vector selection process based on the correlation coefficient can be detailed for the data at each station. Furthermore, the potential possibility of event-based modelling, ensemble forecasting, and error rectification models to improve the accuracy can still be considered.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China, under Grant 51779007; and the National Key Research and Development Program of China under Grant 2016YFC0401308.

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

REFERENCES

- Busico, G., Colombani, N., Fronzi, D., Pellegrini, M., Tazioli, A. & Mastrocicco, M. 2020 Evaluating SWAT model performance, considering different soils data input, to quantify actual and future runoff susceptibility in a highly urbanized basin. *Journal of Environmental Management* **266**, 110625. <https://doi.org/10.1016/j.jenvman.2020.110625>.
- Chen, X., Huang, J., Han, Z., Gao, H., Liu, M., Li, Z., Liu, X., Li, Q., Qi, H. & Huang, Y. 2020 The importance of short lag-time in the runoff forecasting model based on long short-term memory. *Journal of Hydrology* **589**, 125359. <https://doi.org/10.1016/j.jhydrol.2020.125359>.
- Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. & Bengio, Y. 2014 Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: *EMNLP 2014–2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*. pp. 1724–1734.
- Darras, T., Raynaud, F., Borrell Estupina, V., Kong-A-Siou, L., Van-Exter, S., Vayssade, B., Johannet, A. & Pistre, S. 2015 Neural network modeling and geochemical water analyses to understand and forecast karst and non-karst part of flash floods (case study on the Lez river, Southern France). *IAHS-AISH Proceedings and Reports* **369**, 43–48.
- Fan, H., Jiang, M., Xu, L., Zhu, H., Cheng, J. & Jiang, J. 2020 Comparison of long short term memory networks and the hydrological model in runoff simulation. *Water (Switzerland)* **12** (1), 1–15.
- Gao, Y. 2012 Application of BP neural network to runoff forecasting in Karst mountainous area in Guizhou. *Ground Water (China)* **34** (2), 63–65.
- Gao, S., Huang, Y., Zhang, S., Han, J., Wang, G., Zhang, M. & Lin, Q. 2020 Short-term runoff prediction with GRU and LSTM networks without requiring time step optimization during sample generation. *Journal of Hydrology* **589**, 125188. <https://doi.org/10.1016/j.jhydrol.2020.125188>.
- Hochreiter, S. & Schmidhuber, J. 1997 Long short-term memory. *Neural Computation* **9** (8), 1735–1780.
- Hou, W. & Gao, J. 2019 Simulating runoff generation and its spatial correlation with environmental factors in Sancha River Basin: the southern source of the Wujiang River. *Journal of Geographical Sciences* **29** (3), 432–448.
- Iwasaki, K., Katsuyama, M. & Tani, M. 2020 Factors affecting dominant peak-flow runoff-generation mechanisms among five neighbouring granitic headwater catchments. *Hydrological Processes* **34** (5), 1154–1166.
- Ji, H., Chen, Y., Fang, G., Li, Z., Duan, W. & Zhang, Q. 2021 Adaptability of machine learning methods and hydrological models to discharge simulations in data-sparse glaciated watersheds. *Journal of Arid Land* **13** (6), 549–567.
- Kao, I. F., Zhou, Y., Chang, L. C. & Chang, F. J. 2020 Exploring a long short-term memory based encoder-decoder framework for multi-step-ahead flood forecasting. *Journal of Hydrology* **583**, 124631. <https://doi.org/10.1016/j.jhydrol.2020.124631>.
- Khatri, H. B., Jain, M. K. & Jain, S. K. 2018 Modelling of streamflow in snow dominated Budhigandaki catchment in Nepal. *Journal of Earth System Science* **127** (7), 1–14. <https://doi.org/10.1007/s12040-018-1005-5>.
- Laganier, O., Ayrat, P. A., Salze, D. & Sauvagnargues, S. 2014 A coupling of hydrologic and hydraulic models appropriate for the fast floods of the Gardon River basin (France). *Natural Hazards and Earth System Sciences* **14** (11), 2899–2920.
- Le, X. H., Ho, H. V., Lee, G. & Jung, S. 2019 Application of Long Short-Term Memory (LSTM) neural network for flood forecasting. *Water (Switzerland)* **11**, 7.
- Li, H., Zhang, Y., Chiew, F. H. S. & Xu, S. 2009 Predicting runoff in ungauged catchments by using Xinanjiang model with MODIS leaf area index. *Journal of Hydrology* **370** (1–4), 155–162. <http://dx.doi.org/10.1016/j.jhydrol.2009.03.003>.
- Lombard, P. J. & Holtschlag, D. J. 2018 Estimating lag to peak between rainfall and peak streamflow with a mixed-effects model. *Journal of the American Water Resources Association* **54** (4), 949–961.
- Paparrizos, S. & Maris, F. 2017 Hydrological simulation of Sperchios River basin in Central Greece using the MIKE SHE model and geographic information systems. *Applied Water Science* **7** (2), 591–599. <http://dx.doi.org/10.1007/s13201-015-0271-5>.
- Parisouj, P., Mohebzadeh, H. & Lee, T. 2020 Employing machine learning algorithms for streamflow prediction: a case study of four river basins with different climatic zones in the United States. *Water Resources Management* **34** (13), 4113–4131.
- Piman, T. & Babel, M. S. 2013 Prediction of rainfall-runoff in an ungauged basin: case study in the mountainous region of Northern Thailand. *Journal of Hydrologic Engineering* **18** (2), 285–296.
- Ripley, B. D. 2014 *Pattern Recognition and Neural Networks*. Cambridge University Press, Oxford, UK.
- Ross, C. A., Ali, G., Spence, C., Oswald, C. & Casson, N. 2019 Comparison of event-specific rainfall-runoff responses and their controls in contrasting geographic areas. *Hydrological Processes* **33** (14), 1961–1979.

- Sadeghi, S. H., Moradi Dashtpagerdi, M., Moradi Rekabdarkoolai, H. & Schoorl, J. M. 2021 Sensitivity analysis of relationships between hydrograph components and landscapes metrics extracted from digital elevation models with different spatial resolutions. *Ecological Indicators* **121**, 107025. <https://doi.org/10.1016/j.ecolind.2020.107025>.
- Schwertman, N. C. & de Silva, R. 2007 Identifying outliers with sequential fences. *Computational Statistics and Data Analysis* **51** (8), 3800–3810.
- Sezen, C., Bezak, N., Bai, Y. & Šraj, M. 2019 Hydrological modelling of karst catchment using lumped conceptual and data mining models. *Journal of Hydrology* **576**, 98–110.
- Shahid, F., Zameer, A. & Muneeb, M. 2020 Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM. *Chaos, Solitons and Fractals* **140**, 110212. <https://doi.org/10.1016/j.chaos.2020.110212>.
- Smola, A. J. & Scholkopf, B. 2004 A tutorial on support vector regression. *Statistics and Computing* **14**, 199–222. Available from: http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=1CAD92EF8CCE726A305D8A41F873EEFC?doi=10.1.1.114.4288&rep=rep1&type=pdf%0Ahttp://download.springer.com/static/pdf/493/art%3A10.1023%2FB%3ASTCO.0000035301.49549.88.pdf?auth66=1408162706_8a28764ed0fae9.
- Song, T., Ding, W., Wu, J., Liu, H., Zhou, H. & Chu, J. 2020 Flash flood forecasting based on long short-term memory networks. *Water (Switzerland)* **12** (1), 109. <https://doi.org/10.3390/w12010109>.
- Tyralis, H., Papacharalampous, G. & Langousis, A. 2019 A brief review of random forests for water scientists and practitioners and their recent history in water resources. *Water (Switzerland)* **11** (5), 910. <https://doi.org/10.3390/w11050910>.
- Wang, G., Zhou, M., Takeuchi, K. & Ishidaira, H. 2007 Improved version of BTOPMC model and its application in event-based hydrologic simulations. *Journal of Geographical Sciences* **17** (1), 73–84.
- Wang, C., Du, W., Zhu, Z. & Yue, Z. 2020 The real-time big data processing method based on LSTM or GRU for the smart job shop production process. *Journal of Algorithms and Computational Technology* **14**, 1–9.
- Wang, Q., Liu, Y., Yue, Q., Zheng, Y., Yao, X. & Yu, J. 2020 Impact of input filtering and architecture selection strategies on GRU runoff forecasting: a case study in the Wei river basin, Shaanxi, China. *Water* **12** (12), 3532. <https://doi.org/10.3390/w12123532>.
- Xue, J. & Shen, B. 2020 A novel swarm intelligence optimization approach: sparrow search algorithm. *Systems Science and Control Engineering* **8** (1), 22–34. <https://doi.org/10.1080/21642583.2019.1708830>.
- Yang, X., Li, Y., Wang, B., Xiao, J., Yang, M. & Liu, C. Q. 2020 Effect of hydraulic load on thermal stratification in karst cascade hydropower reservoirs, Southwest China. *Journal of Hydrology: Regional Studies* **32**, 100748. <https://doi.org/10.1016/j.ejrh.2020.100748>.
- Zhao, X., Lv, H., Wei, Y., Lv, S. & Zhu, X. 2021 Streamflow forecasting via two types of predictive structure-based gated recurrent unit models. *Water (Switzerland)* **13** (1), 1–17.
- Zhou, Q., Chen, L., Singh, V. P., Zhou, J., Chen, X. & Xiong, L. 2019 Rainfall-runoff simulation in karst dominated areas based on a coupled conceptual hydrological model. *Journal of Hydrology* **573**, 524–533. <https://doi.org/10.1016/j.jhydrol.2019.03.099>.
- Zuo, G., Luo, J., Wang, N., Lian, Y. & He, X. 2020 Two-stage variational mode decomposition and support vector regression for streamflow forecasting. *Hydrology and Earth System Sciences* **24** (11), 5491–5518.

First received 7 June 2021; accepted in revised form 23 January 2022. Available online 3 February 2022