Correction model for rainfall forecasts using the LSTM with multiple meteorological factors

Chang-Jiang Zhang | Jing Zeng | Hui-Yuan Wang | Lei-Ming Ma | Hai Chu

Abstract
The goal of this study was to improve the accuracy of model forecasting, such that forecasters could use model products to make more efficient daily weather predictions. Historical data of the 12 hr following a given time for various meteorological factors from the control forecasts of the European Centre for Medium-Range Weather Forecasting (ECMWF) between 20° N latitude and 110°–130° E longitude were used to verify the performance of the proposed method. Eight major meteorological factors were selected via correlation analysis between control forecast meteorological factors and real-time rainfall. The samples were divided into four types using the K-means clustered method. Each type was respectively modelled by long short-term memory (LSTM) in order to correct rainfall forecasts for eastern China. The eight major meteorological factors were used as the model input, and the differences between real-time rainfall data and model-forecast rainfall were used as the model output. The corrected results revealed that the root mean square error decreased by 0.65, and the threat scores of light rainfall and rainstorms were improved.

KEYWORDS
ECMWF, K-means, LSTM, model prediction, threat score

1 INTRODUCTION
In recent years, numerical forecasting has become increasingly important in modern weather prediction. In particular, the dependence on numerical short-term forecasting products has increasingly become more pronounced. Model forecasting products, as an important tool for weather forecasting and climate prediction, provide an essential decision-making basis for a forecaster’s daily tasks (Shu et al., 2013). Therefore, comparing the prediction effects of each forecasting centre, as well as determining how to make full use of the ensemble forecast data of each centre in order to improve forecasting accuracy, have become very important research topics (Lin et al., 2009).

1.1 Application research of model-forecast rainfall
At present, precipitation forecasting primarily relies on numerical modelling, and precipitation is the model
diagnostic quantity that involves multiple iterations for cumulus convection and microphysical processes, as opposed to direct integration of the model. The model itself has uncertainties, however, including initial value and model errors. Ma and Bao (2017) pointed out that these uncertainties are related to the grid process, in which there is a large parameterization error, and the uncertainty is very large. Ensemble prediction provides strong support for describing forecasting uncertainty. Many studies have shown that ensemble forecasting has more potential economic value in weather decisions than single deterministic forecasts. In developed countries, the construction of ensemble forecasting operating systems is relatively mature. Given the uncertainty of numerical forecasts, ensemble forecasting is considered to be one of the important tools for daily operational forecasting. The World Meteorological Organization (WMO) has incorporated ensemble prediction into one of the four developmental directions of numerical forecasting in the 21st Century (Wang and Chen, 2007). In 2005, the WMO proposed the THORPEX Interactive Grand Global Ensemble (TIGGE) forecasting program (Zhi and Chen, 2010), involving 10 countries: China, the United Kingdom, France, the European Union, Australia, the United States, Canada, Japan, South Korea and Brazil. The program collects the real-time data of operational forecasts from all 10 centres. For example, the National Center for Environmental Prediction (NCEP) and the Canadian Meteorological Centre (CMC) have developed the North American Ensemble Forecasting System (NAEFS) (Ma et al., 2011). Several related studies on ensemble forecasting services have also been carried out in China. For example, the National Meteorological Center has developed an ensemble forecasting toolbox to interpret ensemble forecast products and is developing a quad-weather warning and refined meteorological element forecasting system based on ensemble forecasting (Xiong et al., 2015).

The European Centre for Medium-Range Weather Forecasting (ECMWF) is one of the most sophisticated global numerical forecasting models in the world. It has been using constantly replenished unconventional satellite observations and radar since the 1990s, and its continuously improving assimilation program, quality control methods and bias correction terms give its reanalysis results a stepwise ascendancy. At present, the horizontal resolution of the ECMWF’s new-generation atmospheric high-resolution numerical prediction model has achieved a $0.125^\circ \times 0.125^\circ$ latitude and longitude grid. The China Meteorological Administration provides the model product for operational use through CMAcast (Wang et al., 2018). Given the high resolution of the ECMWF numerical models and China’s higher rainfall prediction accuracy rate, the high-resolution ECMWF rainfall forecast product from June–September 2015 to 2017 was used as the experimental data in an attempt to improve model rainfall forecasting.

1.2 Research progress for the correction of model-forecast rainfall

The ECMWF ensemble forecasting product is a superimposed tool that uses the optimal score set statistics for different precipitation levels. Because it has obvious correction effects for the deterministic forecasting of heavy precipitation, and its precipitation area and magnitude forecasting are better than those of deterministic forecasting, it has become an important reference product for short- and medium-term weather forecasting (Chen et al., 2015). However, whether it is the ensemble prediction fusion product issued by the China Meteorological Administration or the regional fusion product whose design was based on the Haihe River Basin testing results, as the forecasting time increases, the dispersion between the members of the ensemble forecast increases, leading to large prediction errors at both precipitation extremes, particularly for heavy precipitation (Xu et al., 2018).

Similar to deterministic forecasts, a variety of interpretation techniques have been developed based on ensemble forecasting products. For example, Wang (2015) used Bayesian probability decision theory, historical observation data and ensemble prediction data to correct the rainstorm set probability forecasting in the Sichuan Basin of China, which improved the accuracy of rainstorm forecasting in this region. Liu et al. (2013) used the cumulative probability distribution of ensemble precipitation forecasts to establish an extreme precipitation forecasting model based on ensemble prediction and model historical forecasting, which effectively improved the forecasting ability of heavy precipitation events. Li et al. (2014), Chen (2016) and Tang (2019) investigated the frequency-matching precipitation forecasting method. In these studies, it was assumed that the precipitation forecasting frequency was consistent with the observed precipitation frequency, and the corrected value of the precipitation forecasting was obtained by frequency matching. Zhou et al. (2015) matched the cumulative probability distribution of ensemble precipitation forecasts with that of observed precipitation. The average of the ensemble forecasts for 100 mm precipitation events was corrected, thereby effectively correcting the systematic error of rainstorm forecasts.

In recent years, deep learning has demonstrated its extraordinary ability and great potential in many different fields. Given its extremely wide range of application possibilities, the development of deep learning has also promoted the historical process of artificial intelligence
Deep learning has made breakthroughs in the fields of computer vision, speech recognition and natural language processing. New technological innovations not only bring challenges but also present opportunities in the development of meteorological prediction technology. Deep learning provides a new approach for meteorological problems that were difficult to solve based on shallow neural networks. For example, Guan (2017) applied convolutional neural networks to short-term rainfall prediction. Zambrano et al. (2018) used a multi-layer feedforward neural network to predict the degree of drought in Chile in 2018. However, in the correction of rainfall forecasting using modelling, deep learning has rarely been applied.

### 1.3 Problem analysis

As a result of many factors, including the initial field of the model, boundary conditions, physical processes, terrain, vegetation and the design of the model itself, there is inevitably some error in model predictions (Cui et al., 2009). Common precipitation ensemble prediction correction methods include the frequency matching method, model output statistics (MOS) (Wilks, 2007), Bayesian model averaging (BMA) (Sloughter et al., 2010), the back propagation (BP) neural network method (Li et al., 2018), and others. Each method, however, has its own disadvantages. For example, BMA sometimes yields poor prediction results due to its hypothetical prior model; the analogue observation similarity method consumes fewer computing resources, but produces more noise in the correction of light rain; and, as a local search optimization method, it is easy for the BP neural network to drop into a local extremum.

Given the various shortcomings of the common correction methods, as well as the powerful abilities of deep learning, the deep-learning method was used to correct model rainfall forecasting in order to improve the accuracy of precipitation prediction. In addition, considering that the weather parameters appearing in common neural networks are considered to be independent of each other, timing relationships are not generally considered in the construction of meteorological forecasting models. Yang (2017) pointed out that weather forecasting models with time-series features added to the depth feedforward network are better than models without time-series features. Therefore, this type of model is the one most likely to eliminate forecast bias by correcting aggregate precipitation forecasting. Hochreiter and Schmidhuber (1997) used long short-term memory (LSTM) to solve the long-term dependence problem of the ordinary cyclic neural network in order to construct the corrected model of rainfall forecasting.

Based on the above analysis, combined with the characteristics of the data, the model based on the LSTM was adopted. In combination with K-means clustering, multiple meteorological factors were used to correct model forecast rainfall.

### 2 EXPERIMENTAL DATA

In the present study, the real-time rainfall data from the automated stations in East China and the near-coastal areas of China from the Central Meteorological Observatory of Shanghai were used as reference values. There are some differences in the collection of weather site's rainfall data every 12 hr, sometimes only hundreds of a site's data, sometimes thousands of a site's data. For example, the live site distribution at 2000 on July 23, 2017, and at 0800 on July 27, 2017, are shown in Figure 1. The data consisted of 6 hr of precipitation observation data at 0800 and 2000 hours (Beijing local time), covering a total of 12 months from June–September 2015 to 2017. The forecast data covered the same time span as the observed rainfall data from the ECMWF website, and the spatial span of the forecast data was $20^\circ - 40^\circ$ N latitude and $110^\circ - 130^\circ$ E longitude. Considering that there were too many data samples at a resolution of 0.5°, the data samples at a resolution of 1° were taken to be sufficient. Moreover, since data with a grid resolution of 1° could satisfy the correction requirement, these data were selected as the present study. The forecast data included a total of 24 ground and high-altitude meteorological factors for the subsequent 12 hr from 0000 and 1200 Coordinated Universal Time (UTC). The ground data included nine meteorological factors and the high-altitude data included five meteorological factors at three altitudes. The ground data included the total precipitation (tp) predicted by the ECMWF model, comprising 287,973 samples. The specific meteorological factors for the high-altitude and ground data are listed in Table 1.

#### 2.1 Data processing

Because the primary focus was on correcting the model forecast rainfall for the following 12 hr, the experimental data need to add two 6 hr cumulative precipitation values in order to obtain the 12 hr cumulative precipitation. The nearest-neighbour interpolation method was then employed to interpolate the data onto the grid between $20^\circ$ and $40^\circ$ N latitude and $110^\circ - 130^\circ$ E longitude. Since the observation times of the precipitation data were 0800 and 2000 hours (Beijing local time), and the ECMWF control forecast reported the next 12 hr of
rainfall at 0000 and 1200 hours, the observation data needed to be correlated with the model forecast data. The 0800 hours precipitation observation data corresponded to the control forecast data reported from the previous day at 1200 hours, and the 2000 hours precipitation observation data corresponded to the control forecast data reported at 0000 hours on the same day. The correspondence relationship is shown in Table 2.

### 2.2 Data analysis

There were various relationships between the 24 meteorological factors in the control forecast data and the actual rainfall observations. The relationships between the meteorological factors and the actual rainfall were analysed in order to make use of the data and avoid the adverse effects of useless meteorological factors on the rainfall correction. In the present study, the correlation coefficient was used to analyse the correlation between meteorological factors and actual rainfall.

The correlation coefficient is a statistical index used to reflect the degree of close correlation between variables. It is calculated by the product difference method, which is based on the difference between two variables and their respective averages and reflects the degree of correlation between two variables by multiplying their two deviations. The expression for the correlation coefficient $R$ is as follows:

$$ R = \frac{\sum_{i=1}^{n} (I_a(i) - \bar{I}_a)(I_g(i) - \bar{I}_g)}{\sqrt{\sum_{i=1}^{n} (I_a(i) - \bar{I}_a)^2} \sqrt{\sum_{i=1}^{n} (I_g(i) - \bar{I}_g)^2}} $$

The closer the absolute $R$ is to 1, the stronger the correlation between the two variables. The closer the absolute value of $R$ is to 0, the weaker the correlation between the two variables. The correlations between the 24 meteorological factors and the actual rainfall are shown in Figure 2.
The experimental results revealed that the correlation between model-forecast rainfall and actual rainfall was the largest, close to 0.6. In addition, the wind speed (v700) in the north–south direction and the temperature (t700) at the 700 hPa height, as well as the temperature at the surface (t) and the total column water vapor content (tcwat), were also highly correlated with the actual rainfall.

In view of the correlation co-efficients between the meteorological factors and the actual rainfall, the present study the meteorological factors with correlation co-efficients > 0.3 were used as the model input. Therefore, the input consisted of eight dimensions, including the model-forecast rainfall (tp) with the highest correlation co-efficient; the northerly wind speed (v700), temperature (t700) and potential height (hg700) at the 700 hPa height; and the temperature (t850) and potential height (hg850) at the 850 hPa height. The surface meteorological factors included temperature (t) and total column water vapor content (tcwat).

3 | ESTABLISHMENT OF RAINFALL CORRECTION MODEL BASED ON K-MEANS AND LSTM

In the present study, an LSTM, which was designed to solve the long-term dependence problem of ordinary cyclic neural networks, was used to construct the corrected model for rainfall forecasting.

First, the K-means clustering method was used to divide the samples into four categories. Next, the LSTM was used respectively to build models for the different data types. Eight types of meteorological factors (including the model-forecast rainfall) were used as inputs, and the difference between the actual rainfall and the model-forecast rainfall was taken as the output. The constructed model was then used to correct the model-forecast rainfall. Finally, the corrected rainfall following the classification was integrated in order to obtain the final corrected rainfall. The overall flowchart of the proposed corrected model is shown in Figure 3.

3.1 | K-means clustering

The K-means clustering algorithm was proposed by MacQueen (1967) to classify and cluster large data sets efficiently. The idea of the algorithm is that given n d-dimensional data, when K is the number of subsets of data to be generated, the K-means algorithm divides the given data set into K groups. Each group is a class, and each class Ck has a centre Oi. Using Euclidean distance as the criterion for judging the similarity between data, the square of the sum of the distance between data points in a class Ck with clustering centre Oi is calculated. The calculation formula is as follows:

\[ J(C_k) = \sum_{m_i \in C_k} ||m_i - o_k||^2. \]  

The K-means clustering algorithm is an iterative process. The ultimate goal of clustering is to minimize the sum of the squares of all data elements in the class to their cluster centre distance. The flowchart of the K-means clustering algorithm is shown in Figure 4.
Clustering is an unsupervised algorithm. In the present study, the eight meteorological factors that had been selected by correlation analysis were used as input for the $K$-means clustering algorithm. Four samples were randomly selected as initial clustering centres, and other samples were added to the nearest classes. The means of each sample class were calculated as new clustering centres, and the clustering centres were reclustered based on distance until the clustering centres did not change. After many experiments, it was determined that the effect of subsequent correction was best when the input samples were clustered into four categories.

\[
J(C) = \sum_{k=1}^{K} J(C_k) = \sum_{i=1}^{n} d_{kl} \| m_i - o_k \|^2
\]  
\[
d_{kl} = \{1, m_i \in c_l \}, 0, m_i \notin c_l\}
\]  

### 3.2 LSTM model

Since the process of rainfall is complex, it is influenced by many factors. It is also highly correlated with time series. Given that the LSTM primarily deals with time-related complex relationships. In the present study, four classes of input samples from $K$-means clustering were subsequently modelled by the LSTM. After modelling, the relationship between the meteorological factors predicted by the model and the actual rainfall could be further analysed. The proposed model could be used to correct the model-forecast rainfall better.

The LSTM is a special cyclic neural network proposed by Hochreiter and Schmidhuber (1997). Similar to the traditional cyclic neural network model, it is also based on time-series data. Through the cyclic connection between neurons, the intrinsic relationships between time-series data are mined, and the time-series data are modelled (Gao, 2019). However, the LSTM cycle differs
from that of the traditional neural network model in that it features a special neuronal structure known as a “memory cell”. Through this hidden layer structure, the LSTM network is capable of storing information for an arbitrary length of time in order to obtain a more precise time-series model.

The memory cell structure of the LSTM network is shown in Figure 5. The solid line is the structure to which the network element must be connected; the dotted line is the structure that can be connected.

The memory unit module is composed of three “door” structures (input gate, forget gate and output gate) and one loop-connecting unit. The core idea is to control the switches of each “gate” through a nonlinear function in order to protect and control the state of the memory unit, further controlling the increase and decrease of information.

Therefore, the key to the LSTM network is the long-term storage of data through the state of the memory unit. In general, the three “gates” output a value ranging from 0 to 1 via the sigmoid function, thereby determining how much new information can be input to the memory unit. At the same time, the forget gate reads the values of \( x_t \) and \( h_{t-1} \), and \( b_f \), \( b_g \), \( b_o \) and \( b_c \) are the offset vectors of the input gate, forget gate, output gate and input transition, respectively.

The inputs of \( x_t \) and \( h_{t-1} \) are converted in order to generate a new information quantity via the tanh function; the input gate reads \( x_t \) and \( h_{t-1} \) and outputs a value ranging from 0 to 1 via the sigmoid function (0 means no information is allowed to pass; 1 means all of the information is allowed to pass); multiplying this value by the information quantity determines how much new data information can be input into the memory unit. At the same time, the forget gate reads the values of \( x_t \) and \( h_{t-1} \), outputs the values of 0–1 via the sigmoid function, multiplies them with the instantaneous state \( C_{t-1} \) on the memory unit in order to determine the retention and abandonment of the original

\[
\begin{align*}
  i_t &= \sigma(W_{ix}x_t + W_{hi}h_{t-1} + b_i), \quad (5) \\
  f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f), \quad (6) \\
  o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o), \quad (7) \\
  \tilde{C}_t &= \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_C), \quad (8) \\
  C_t &= f_t \odot C_{t-1} + i_t \odot \tilde{C}_t, \quad (9) \\
  h_t &= o_t \odot \tanh(C_t), \quad (10)
\end{align*}
\]

where \( \sigma \) is the sigmoid function; \( \tanh \) is the hyperbolic tangent function; \( i_o, f_o, O_t \) and \( C_t \) are the input gate, forget gate, output gate and conversion unit of the input, respectively; \( W_{xi}, W_{xf}, W_{xo}, W_{xc} \) and \( W_{hi}, W_{hf}, W_{ho} \) and \( W_{hc} \) are, respectively, the input weights for the input gate, forget gate, output gate and input transition corresponding to \( x_t \) and \( h_{t-1} \); and \( b_f, b_g, b_o \) and \( b_c \) are the offset vectors of the input gate, forget gate, output gate and input transformation, respectively.
information in the memory unit, and then adds the new information to the retained original information to update the status of the current memory unit. Finally, the output gate reads the values of \( x_t \) and \( h_{t-1} \), calculates them using the sigmoid function and multiplies them with the state \( C_t \) of the memory unit processed by the tanh function, so as to determine the output information of the state of the memory unit (Zhou et al., 2015).

The LSTM network is a special kind of cyclic neural network. Compared with traditional circulating neural networks, the LSTM network increases one unit state in order to preserve long-term information and solve the problem of the traditional circulation network gradient either disappearing or blowing up during training. Therefore, the structure of the LSTM network is the same as that of the traditional cyclic neural network, and is composed of an input layer, a hidden layer and an output layer. The hidden layer structure adds the connection from the hidden layer at the previous moment to the hidden layer at the next moment. The unfolded structure of the single-layer LSTM network is shown in Figure 5 (Cao et al., 2018). During the calculation process of the LSTM network in the LTSM unit, there is an input sequence, and an output sequence is obtained through the hidden layer. The fully connected neural network is set to calculate the output of each hidden layer and obtain the output of each node. The node output is connected in order to yield the final output \( y \).

In the present study, 287,973 samples from 12 months of data were used to verify the performance of the proposed model. The sample data were arranged in chronological order. During the training process, 235,053 samples from 10 months of data were used as input for the proposed model. The current rainfall samples \( x_t \) were processed by the LSTM hidden layer in order to obtain the output \( h_t \). The current hidden layer output \( h_t \) and the next 12 hr rainfall sample \( x_{t+1} \) were then input into the LSTM hidden layer, and the difference of the next node was predicted. Finally, the output sequence of each node’s hidden layer was connected by the full connection layer in order to obtain the final output. The final corrected rainfall results were obtained by adding the difference between the predicted rainfall and control-forecast rainfall with the model-forecast rainfall.

4 | EXPERIMENTAL RESULTS

In the present study, 10 months of data were used as training samples and two months were used as testing samples. The method was implemented on the Python platform. The LSTM network, in combination with K-means clustering, was then used to correct the model-forecast rainfall. Finally, the results of the proposed corrected rainfall model were compared with those of the frequency matching method, linear regression method, support vector machine (SVM), and deep belief network (DBN).

4.1 | Evaluation parameters

After the model forecast had been corrected, the root mean square error (RMSE) and threat scores (TSs) were used to reflect the overall error of the estimated results. The RMSE formula is as follows:

\[
RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (I_a(i) - I_g(i))^2}
\]

which includes the observed rainfall \( I_a \) and the estimated rainfall \( I_g \). The RMSE describes the overall accuracy of the precipitation estimation and can reflect the estimated error to a certain extent.

According to the “Important Weather Forecast Quality Assessment Method” formulated by the China Meteorological Administration, the TS is an important evaluation index for guiding forecasting products. Its formula is as follows:

\[
TS = Na/(Na + Nb + Nc)
\]

where \( Na, Nb \) and \( Nc \) are the number of lattice points for correct prediction, underestimation and overestimation, respectively. In the present study, the precipitation grade was classified according to the 12 hr cumulative precipitation. The specific criteria are listed in Table 3. For different precipitation levels, when the model predicted rainfall with the same magnitude as the actual rainfall, the rainfall forecast at the current grid points was classified as correct; when the magnitude of the forecast rainfall exceeded that of the actual rainfall, the rainfall forecast at the current grid points was classified as over-estimation; when the magnitude of the forecast rainfall was less than the actual rainfall, the rainfall forecast at the current grid points was classified as underestimation. The specific standards are shown in Table 4.

The TS ranges from 0 to 1 and reflects the accuracy of the effective precipitation forecast. At present, the evaluation of internal guidance forecast products is scored according to this evaluation method, which is the main basis for operational assessment.

4.2 | Debugging process

Since the data used in the present study consisted of eight meteorological factors (including model-forecast rainfall),
The relationships between these meteorological factors and actual rainfall, so K-means was used for clustering. Four samples were randomly selected as clustering centres, and the samples were clustered into four categories according to Euclidean distance between samples. The 10 month training data and two month test data used. 

### Table 3

| Precipitation level | Light rain | Moderate rain | Heavy rain | Rainstorm |
|---------------------|------------|---------------|------------|-----------|
| 12 hr rainfall (mm) | 0.1–4.9 mm | 5.0–14.9 mm   | 15.0–29.9 mm | ≥ 30 mm   |

### Table 4

Correct prediction, underestimation and overestimation of specific standards

| Real-time rainfall \ forecast rainfall | Light rain | Moderate rain | Heavy rain | Rainstorm |
|---------------------------------------|------------|---------------|------------|-----------|
| Light rain                            | Na         | Nc            | Nc         | Nc        |
| Rain                                  | Nb         | Na            | Nc         | Nc        |
| Heavy rain                            | Nb         | Nb            | Na         | Nc        |
| Rainstorm                             | Nb         | Nb            | Nb         | Na        |

### Table 5

Comparisons of results clustered into different types

| Result \ types of samples | 3  | 4  | 5  | 6  |
|----------------------------|----|----|----|----|
| Root mean square error (RMSE) | 7.82 | 7.45 | 7.50 | 8.10 |
| Average threat score (TS)    | 0.110 | 0.140 | 0.137 | 0.092 |

### Table 6

Clustering results after K-means clustering

| Samples \ rainfall type after clustering | Type 1 | Type 2 | Type 3 | Type 4 |
|-----------------------------------------|--------|--------|--------|--------|
| Training samples                        | 58,019 | 84,410 | 50,556 | 42,068 |
| Test sample                             | 6,344  | 17,703 | 23,913 | 4,960  |

### Figure 6

Long short-term memory (LSTM) network expansion diagram
for modelling were clustered. After clustering into three, four, five and six categories, the results of clustering into different types were compared, as shown in Table 5, and it was determined that clustering into four categories was the optimal model approach. Because the four classes after K-means clustering differed from the light rain,
moderate rain, heavy rain and rainstorm categories used for TS evaluation, the classes after clustering were designated Type 1, Type 2, Type 3 and Type 4. The specific clustering results are listed in Table 6.

The TS and RMSE of the results were improved by correcting each rainfall type after clustering. The RMSE before clustering was 7.85, while after clustering it was 7.45. The comparison of TSs before and after clustering is shown in Figure 6. The TSs for light rain, heavy rain and rainstorm all improved. Overall, clustering exhibited a good effect on the correction of model-forecast rainfall.

Four types of clustering were modelled using a two-layer LSTM network. After continuous debugging, the four types of LSTM models had different internal nodes; the numbers of specific nodes are listed in Table 7. In addition, the number of iterations required to converge each clustering type varied. The convergence process of each type is shown in Figures 7a–d.

It can be seen from Figure 8 that the iteration processes of Types 1–3 converged relatively slowly, while Type 4 required only approximately 50 iterations to converge, thus exhibiting a faster iteration process.

### 4.3 RMSE comparison

The model forecasting rainfall data for August–September 2017 were corrected by the proposed model. After clustering, the RMSE of each type improved a small amount, Type 1 decreased noticeably and the overall RMSE decreased by 0.7. The specific data are listed in Table 8.

Before using the LSTM to correct rainfall, several other methods were used for modelling, including classical frequency matching, linear regression, SVM learning and the DBN. A comparison of the various methods revealed that the LSTM had a better overall performance correcting model-forecast rainfall.

Although the RMSE is the smallest after the correction with the linear regression model, in practice TS, which is a statistical method, is usually used as the evaluation standard of the rainfall forecasting in actual operational weather forecasting. As can be seen from Figure 8, except for the TS of light rain, the other TS after linear regression correction are lower. The RMSE difference between the proposed model in the present paper and that linear regression correction is only 0.02, and the TS of the proposed model is better than that of linear regression, so the overall performance of the proposed LSTM model is the best. The method used the present study had an RMSE that was slightly higher than that of linear regression. Compared with the SVM and frequency matching methods, the RMSE of the proposed model was much lower. Comparison of the RMSE for each model is shown in Table 9.

### 4.4 Comparison of TSs

In the present study, the TSs consisted of four scores: light rain, moderate rain, heavy rain and rainstorm. After correction by various methods, the TSs of each rainfall type improved to varying extents. A comparison of the specific TSs is shown in Figure 9.

The most classic frequency matching method yielded the TS that was closest to that of the model forecast. The heavy rain TS increased a small amount, while the other scores did not improve. The linear regression and SVM methods were similar: the light rain TS improved, but the heavy rain and rainstorm TSs were very low. After the model rainfall was corrected by the DBN method, the moderate rain score increased slightly, while the TSs of the other types of rainfall did not increase. After correction by the LSTM, the light rain TS effectively improved, and the rainstorm TS also improved. Overall, the LSTM model demonstrated the best correction performance in rainfall forecasting.

### 4.5 Case comparison

The data in present study consisted of rainfall information from 20° to 40° N latitude and 110°–130° E
longitude. From the test data, the rainfall that fell between 0800 and 2000 on August 7, 2017, was selected for comparative analysis. The regional real-time precipitation map was obtained using the nearest-neighbour interpolation method based on the observed data of the national surface meteorological station. The regional distributions of the original forecast and the observed precipitation of the ECMWF model are shown in Figures 10a,b. The rainfall distribution after LSTM correction is shown in Figure 10c.

It can be seen from the regional distribution map that more 30 mm heavy rain events occurred in the border area between Henan, Hubei and Anhui provinces, with the rainfall at some stations in southern Henan > 60 mm. The central part of Hubei is generally prone to heavy rain and storms. The rainfall centre of the original forecast of the EC model was west of this area. It only predicted moderate rain amounts of < 20 mm in the southern part of Henan. The torrential rain that occurred in the area was not predicted, and the overall rain band
range was narrow. After correction with the proposed LSTM model, the range of heavy rain from southern Henan to northeastern Hubei increased significantly, and the rainstorm area in southwestern Hubei increased as well. Overall, the corrected precipitation was closer to the real-time precipitation distribution. It is noteworthy that in the area near Poyang Lake, northwest of the Yangtze River, the proposed LSTM model corrected the local light rain rainfall predicted by the original model to moderate rain, which was closer to the actual occurrence. In addition, a comparison of the precipitation range of Hunan and Zhejiang before and after correction revealed that the original forecast of the EC model had a larger forecast range for small-scale precipitation and a large number of vacant forecasts were produced, while the phenomena of vacant and omitted forecasts were effectively reduced by the LSTM correction.

The regional rainfall distributions after frequency matching, stepwise linear regression, SVM and DBN correction are shown in Figures 11a–d, respectively. By comparing the regional rainfall distribution maps after correction by different methods, it was found that the results of the frequency matching method based on mathematical statistics and the model-forecast rainfall were essentially similar, thus their TSs were the closest. The rainfall distributions yielded by the SVM and linear regression were similar. Both methods basically reduced the amount of heavy rainfall in the model forecast, making the RMSE smaller, although the heavy rain and rainstorm forecasts basically disappeared, and light rain covered the entire area, leading to more vacant forecasts. After correction by the DBN method, several areas of heavy rain forecast by the model had expanded. For instance, in the central part of Hubei province, the rainfall reached 79 mm, and the range of rainfall was disorderly, characteristics that were not consistent with the actual rainfall. Overall, the proposed LSTM model was found to reduce RMSE effectively and improve the TSs of light and heavy rain.

5 | CONCLUSIONS

The present study used the K-means clustering algorithm to classify model forecast data, and then used long short-term memory (LSTM) to perform subsequent modelling for different types of rainfall data. Using feature extraction and continuous debugging training of the LSTM model, the relationships between the meteorological factors related to the model forecast and the actual rainfall were learned, allowing the effective correction of model-forecast rainfall.

For the proposed correction model, the threat scores (TSs) of moderate and heavy rain were both inferior to those of the model forecast. In future, it is hoped that the data from multiple meteorological factors predicted by the model will be converted into pseudo-colour maps, corresponding to actual rainfall values, and the input
data will be transformed into images. In addition to analysing temporal relationships, spatial relationships will also be analysed in subsequent research. The relationship between rainfall at a particular site and the model-forecast rainfall within a certain surrounding area will be learned, such that the model-forecast rainfall can be corrected in order to improve the TSs of moderate and heavy rain.

ACKNOWLEDGEMENTS

A portion of this research was supported by the National Natural Science Foundation of China (grant numbers 41575046 and 41975069), the Project of Commonwealth Technique and Application Research of Zhejiang Province, China (grant number 2016C33010), the Project of the Shanghai Meteorological Center of China (grant number SCMO-ZF-2017011), and Project of China National Meteorological Center (grant number KYH06Y19169).

ORCID

Chang-Jiang Zhang 👤 https://orcid.org/0000-0002-2170-3878

REFERENCES

Cao, G.Q., Zhang, X.M. and Chen, Y.F. (2018) A study on landslide warning of multivariable mine dumping site based on PCA-LSTM. Computer Systems & Applications, 27, 252–258. https://doi.org/10.15888/j.cnki.csa.006646.

Chen, B.Y., Dai, C. and Guo, Y.Q. (2015) Precipitation forecast test and analysis of ECMWF set statistical products in flood season in 2013. Heavy Rain Disaster, 34, 64–73. https://doi.org/10.3969/j.issn.1000-9045.2015.01.009.

Cui, J., Zhou, X.S., Zhang, A.Z. and Yan, Q. (2009) Application of weather test in autumn and winter precipitation forecasting of numerical model in Northeast China. Journal of Meteorology and Environment., 25, 17–21. https://doi.org/10.3969/j.issn.1673-503X.2009.04.004.

Fan, Y.Q. (2015) A review of domestic research on deep learning. Distance Education in China (Comprehensive Edition), 01, 27–33. https://doi.org/10.3969/j.issn.1009-458X.2015.06.007.

Gao, Y. (2019) Location prediction algorithm of moving object based on LSTM. Computer Science and Exploration, 13, 23–34. https://doi.org/10.3778/j.issn.1673-9418.1801029.

Guo, P.Z. (2017) Short-term rainfall prediction model based on integrated learning and deep learning. China Statistical Education Society, 2017, 22.

Hochreiter, S. and Schmidhuber, J. (1997) Long short-term memory. Neural Computation., 9, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

Li, J., Du, W. and Chen, C.J. (2014) Introduction and analysis of frequency (or area) matching method for precipitation deviation correction. Meteorology, 40, 580–588. https://doi.org/10.7519/j.issn.1000-0526.2014.05.008.

Xi, J. (2004) Study on the correction of the precipitation data based on topographic factor in the source region of the Yellow River. Journal of Basic Science and Engineering, 26, 1147–1163. https://doi.org/10.16058/j.issn.1005-0930.2018.06.001.

Lin, C.Z., Zhi, X.F., Han, Y. and Wang, J.Y. (2009) Multi-mode super-aggregate forecasting of ground temperature based on TIGGE data. Quarterly Journal of Applied Meteorology., 20, 706–712. https://doi.org/10.11898/1001-7313.20090608.

Li, Q., Wei, J.H., An, J., Zhang, B. and Ren, Y. (2018) Study on the fusion correction of TRMM 3B43 precipitation data based on topographic factor in the source region of the Yellow River. Journal of Basic Science and Engineering, 26, 1147–1163. https://doi.org/10.16058/j.issn.1005-0930.2018.06.001.
fusion products in Haihe River basin. *Journal of Meteorology and Environment*, 34, 11–17. https://doi.org/10.3969/j.issn.1673-503X.2018.04.002.

Yang, H. (2017) *Research on Weather Prediction Based on Deep Learning*. Harbin Institute of Technology. Harbin, Heilongjiang Province, China.

Zambrano, F., Anton, V., Andy, N. and Michele, M. (2018) Prediction of drought-induced reduction of agricultural productivity in Chile from MODIS, rainfall estimates, and climate oscillation indices. *Remote Sensing of Environment*, 219, 15–30. https://doi.org/10.1016/j.rse.2018.10.006.

Zhi, X.F. and Chen, W. (2010) New achievements of international atmospheric research in THORPEX program. *Journal of Atmospheric Science*, 33, 504–511. https://doi.org/10.3969/j.issn.1674-7097.2010.04.015.

Zhou, D., Chen, J., Chen, C.P. and Wang, J.Y. (2015) Application research of rainstorm ensemble prediction-observation probability match correction method in Sichuan Basin. *Rainstorm Disaster*, 34, 97–104. https://doi.org/10.3969/j.issn.1004-9045.2015.02.001.

---

**How to cite this article:** Zhang C-J, Zeng J, Wang H-Y, Ma L-M, Chu H. Correction model for rainfall forecasts using the LSTM with multiple meteorological factors. *Meteorol Appl*. 2020;27: e1852. https://doi.org/10.1002/met.1852