A revised forecast method of ECMWF precipitation based on CNN feature extraction

Huasheng Zhao¹*, Xiaoyan Huang¹, Ying Huang¹ and Jian Jin²

¹Guangxi Institute of Meteorological Sciences, Nanning, 530022, China
²School of Computer Science and Technology, East China Normal University, Shanghai, 200062, China

*Corresponding author’s e-mail: 57870954@qq.com; jjin@cs.ecnu.edu.cn

Abstract. The paper proposes a method for correcting the precipitation forecast of the European Centre for Medium-Range Weather Forecasts (ECMWF) model. The method first uses the ECMWF physical quantity field to draw the flow field diagram in the form of weather, and uses it as the input of the convolutional neural network (CNN) and performs feature extraction on it. Secondly, select several characteristic factors that are highly correlated with the forecast object from the many characteristic factors. Finally, the selected characteristic factors and the precipitation forecast factors of the ECMWF model are used as the input factors of the random forest regression model for modeling and forecasting. Through the correction forecast test of the next 24h precipitation at 10 forecast test stations, the results show that the ECMWF precipitation correction forecast method proposed in this paper is 15% less than the forecast method using ECMWF interpolation to the station on the MAE and RMSE indicators. And 17%. At the same time, on the TS scores of rainstorms above 24h, the revised forecast method proposed in this paper has a false alarm rate significantly lower than the forecast method of ECMWF interpolation to the station.

1. Introduction
Precipitation is an extremely important link in the hydrological cycle of the earth. Compared with other atmospheric variables, precipitation has a particularly high temporal and spatial variability. These irregular features are related to the individual formation and growth of precipitation clouds, which in turn There is a complicated coupling relationship with the surrounding atmospheric fluid dynamics[1]. For this reason, people continue to improve the numerical weather prediction model through various methods to improve the precipitation forecasting ability[2-3]. However, compared with the prediction of atmospheric factors such as air pressure, temperature, humidity and wind, the prediction ability of numerical models for precipitation is still far behind that of other atmospheric factors. For example, studies by Stephens[4] and Tapiador[5] have shown that the precipitation prediction of numerical models usually fails to reveal many key aspects such as the location, time, and intensity of precipitation. Therefore, it is important to carry out correction studies on the precipitation forecast of numerical models. Realistic meaning. Practice has also proved that the revised numerical model precipitation products can usually provide more accurate precipitation forecasts[6]. So far, a large number of researches on interpretation and application of numerical forecast products for objective correction of precipitation forecasts have been carried out at home and abroad, and they have been applied in actual forecasting operations[7]. For example, Pan Baoxiang[8] trained the model by optimizing the hierarchical set of the spatial convolution kernel to learn the dynamic features related to precipitation from the surrounding
dynamic field. In this way, the daily precipitation forecast of the numerical model is corrected, and the results show that if there are enough data, the corrected forecast method of precipitation will be better than the corrected forecast effect of reanalyzing precipitation products and using linear regression.

From the above researches at home and abroad, it can be seen that the current precipitation correction forecast research of numerical models mainly uses statistical downscaling (SD) methods to make corrections from multiple angles and methods, and has achieved certain results. Generally, the statistical downscaling (SD) method is used to correct the model precipitation, and the effect of the correction depends largely on how to obtain the characteristic factors closely related to the prediction object. In recent years, the CNN network is the most prominent method in terms of feature extraction. Compared with traditional neural networks, CNN has greatly enhanced people's ability to process structured high-dimensional data. It uses the internal structure of the data to reduce the structure of the model and achieves effective information extraction. This article attempts to propose a ECMWF precipitation correction forecast method based on the combination of CNN and Random Forest. In order to explore a new method for correcting precipitation in numerical models.

2. Method and principle

2.1. Convolutional Neural Network
Convolutional neural network is composed of convolutional layer, down-sampling layer, fully connected layer, etc. It has the characteristics of local connection, weight sharing, and down-sampling in time or space in structure. In the convolutional layer, the pixels are weighted and summed through the convolution kernel to obtain the feature map of the original image to achieve feature extraction. In order to extract different features more fully, multiple different convolution kernels can be used in the same network. In the convolutional layer, there are features such as local connection and weight sharing. Each neuron in the same feature map shares the weight when calculating with the convolution kernel, which greatly reduces the number of parameters and the risk of overfitting. The form of the convolutional layer is as follows:

$$x^l_j = f \left( \sum_{i \in M_j} x^{l-1}_{ij} k_{ij} + b_j^l \right)$$

Among them: $M_j$ is the feature map of the middle layer; $l, k, b$ respectively represent the number of layers, convolution kernel and bias value, and $f$ represents the activation function, often taking functions such as sigmoid, tanh and Relu.

After the feature is extracted by the convolutional layer, in order to further reduce the amount of calculation, the obtained feature map is usually down-sampled. Downsampling will not change the number of feature map layers, but will reduce the size of the feature map. This can not only retain the main features, but also achieve the effect of dimensionality reduction, while also preventing overfitting. Down-sampling methods include mean sampling, maximum sampling, overlapping sampling, mean-square sampling, normalized sampling, random sampling, deformation-constrained sampling, etc. The down-sampling method used in this article is maximum sampling. Its form is as follows:

$$x^l_j = f (\beta^l_j \text{down}(x^{l-1}_{j} + b_j^{l}))$$

Among them, $\text{down}()$ is the downsampling function. The fully connected layer is usually the last layer of the convolutional neural network. The output neurons of this layer and each output neuron are connected together. When CNN is used in the regression problem, the output features of this layer are used as the input of the regression layer. Figure 1 shows the structure of the CNN model in this article.
2.2. Random forest algorithm

This paper uses the random forest algorithm[7] to construct a nonlinear mapping relationship between the feature factors obtained from the convolutional neural network in section 2.1 and the forecasting station. Random forest algorithms can be divided into classification and regression models. Random forest regression (RFR) is a nonlinear statistical method proposed by Leo Breiman in 2001[13]. RFR uses Bootstrap to extract multiple training sample subsets from the original sample, and model each sample subset separately. Further, the prediction is made by combining multiple decision trees, and the average value is taken as the final prediction result[14]. Its essence is similar to the idea of ensemble forecasting in meteorology, that is, integrating the prediction results obtained by modeling multiple decision trees. The algorithm has the advantages of fast calculation speed, good generalization
performance and fewer parameters. At the same time, it is not prone to "overfitting". For this reason, this article attempts to use this method to model and forecast precipitation.

The RFR algorithm uses bootstrap sampling technology to grow a random vector (i.e. regression decision tree) to form a combined model of \( h(X, \theta_k), k = 1, \ldots, p \). The predictor variable is different from its classification model and is a numerical variable. The predicted value is obtained by averaging the predicted results of k trees.

3. Experimental and its data

The data used in this paper are the precipitation field released by ECMWF daily at 08:00 and 20:00, the altitude field of 500-850Pha, and the 48h forecast field data of the U/V wind field. The time period for selecting the test sample is from January 2011 to October 2018. After removing the missing samples, the total sample size was 5675.

This paper selects the cumulative precipitation of 10 sites in the next 24 hours as the test object. Table 1 shows the latitude and longitude of the 10 test stations.

| Station number | St 1 | St 2 | St 3 | St 4 | St 5 | St 6 | St 7 | St 8 | St 9 | St 10 |
|----------------|------|------|------|------|------|------|------|------|------|-------|
| lon (E)        | 108.21| 110.30| 106.61| 110.17| 109.40| 109.14| 108.35| 108.04| 111.30| 107.35|
| lat (N)        | 22.65 | 25.32 | 23.90 | 22.64 | 24.34 | 21.46 | 21.78 | 24.69 | 23.48 | 22.41 |

For each test station’s T-day time report, when the ECMWF third-order polynomial interpolation results in the precipitation of the forecast station>10mm, first use the CNN model to perform characteristic factor analysis on the T-1 day 48h forecast field of many physical elements in the ECMWF model. Extract and select a number of characteristic factors and key rainfall forecast grid factors as the input factors of the random forest, and carry out the revised forecast of the final numerical model. The specific construction steps:

1. Selection of training samples. Select all samples with EC greater than 10mm in the historical samples of the forecast site for modeling experiment (sample size is about 1000), and use the last 200 samples of this part of samples as independent samples, and other samples as training samples.

2. CNN model and its input. Use ECMWF’s 850Pha height field, 850Pha wind field and 500Pha wind field to draw the weather situation field map, and use this as the input of the CNN model.

3. Determine the CNN model structure. The specific network structure is shown in Figure 1.

Input to the random forest regression model. After the CNN model is trained, in order to avoid excessive human subjective influence, this paper selects the five feature factors with the highest correlation with the forecasting site for all test sites, together with the precipitation factor predicted by the EC model, as the random forest regression model Enter the feature factor.

4. Test results and analysis.

In this paper, when conducting prediction experiments on 10 different stations, considering that the weight initialization of the CNN model has a certain impact on the prediction results of the model during training, the CNN-RFR method uses the average value of its 5 experiments As the final forecast value of the model, the specific statistics are shown in Figure 2.
Figure 2 Statistics of 200 independent samples by several forecasting methods when the EC forecast value is less than or equal to 10mm

From the statistical results in Figure 2:
(1) In the statistical results of 10 test sites, the average absolute error (MAE) of the CNN-RFR method is smaller than that of the EC interpolation method and reduced by 10 to 33%. In addition, the MAE error of the CNN method is smaller than the MAE error of the EC interpolation method by 4~28%.
(2) For TS statistics above heavy rain: the TS scores of the CNN-RFR method are significantly higher than the other two forecasting methods. In addition, in the independent sample test of 10 stations, the CNN method has 9 stations with TS scores higher than or equal to the EC interpolation method, which shows that the revised model also has good correction forecasting ability.

5. Summary and discussion.
In the experiment of using EC precipitation interpolation greater than 10mm to model and correct the forecast, the CNN-RFR numerical model precipitation correction forecast method proposed in this paper has a lower MAE than the EC interpolation method, which shows that the precipitation correction forecast of this method has a positive forecast accuracy. Skills. At the same time, the TS scores of the CNN-RFR method are significantly higher than the EC interpolation method in terms of precipitation forecasts and sunny rain forecasts of the magnitude above heavy rain. At the same time, although the performance of the CNN model given in this article is inferior to that of the CNN-RFR method, its prediction accuracy and the TS score of precipitation forecasts above heavy rainfall are also significantly better than EC in most cases. Interpolation method.

Acknowledgments
This work was supported by the Guangxi Natural Science Foundation (Grants 2018GXNSFAA294128, 2018GXNSFAA281229, 2017GXNSFDA198030) and the National Natural Science Foundation of China (Grant 41765002).

References
[1] Houze Jr, Robert A. Cloud dynamics[M], volume 104. New York: Academic press. 2014.
[2] Jing Sun, Guangguang Cheng, Xiaoling Zhang. An improved Bias Removal Method for Precipitation Prediction and Its Application [J]. Journal of Applied Meteorological Science,2015,26(02):173-184.
[3] Feng Xue, Xing Wang, Shuang Wu, et al. Radar extrapolation and numerical model dynamic fusion prediction method of precipitation probability [J]. Meteorological Science and Technology, 2017, 45(06):1036-1042.

[4] Stephens G L, Lecuyer T, Forbes R G, et al. Dreary state of precipitation in global models [J]. Journal of Geophysical Research, 2010: 1-13.

[5] Tapiador F J, Roca R, Genio A D, et al. Is Precipitation a Good Metric for Model Performance? [J]. Bulletin of the American Meteorological Society, 2019, 100(2): 223-233.

[6] Zaiwen Wang, Zuofang Zheng, Min Chen, et al. Forecast of Meteorological Elements Based on Nonlinear Regression Method of Support Vector Machine [J]. Journal of Applied Meteorological Science, 2012, 23(05): 562-570.

[7] Huasheng Zhao, Xiaoyan Huang, Ying Huang. Application of ECMWF Ensemble Forcast Products to Rainstorm Forecast in Guangxi [J]. Journal of Applied Meteorological Science. 2018, 29(03): 344-352.

[8] Pan B, Hsu K, AghaKouchak A, et al. Improving precipitation estimation using convolutional neural network [J]. Water Resources Research. 2019, 55(3): 2301-2321.

[9] Liu Y, Racah E, Prabhat, et al. Application of Deep Convolutional Neural Networks for Detecting Extreme Weather in Climate Datasets [J]. ArXiv preprint, 2016, arXiv:1605.01156.

[10] Rui Chen, Xiang Wang, Weimin Zhang, et al. A hybrid CNN-LSTM model for typhoon formation forecasting. 2019, 23(3): 375-396.

[11] Pradhan R, Aygun R S, Maskey M, et al. Tropical Cyclone Intensity Estimation Using a Deep Convolutional Neural Network [J]. IEEE Transactions on Image Processing. 2018, 27(2): 692-702.

[12] Krizhevsky A, Sutskever I, Hinton G. ImageNet classification with deep convolutional neural networks [J]. Communications of the ACM, 2017, 60(6): 84-90.

[13] Breiman L. Random Forests. Machine Learning [J]. 2001, 45(1): 5-32.

[14] Liang Zhou, Qiong Huang, AO Li, et al. GWA study of Alzheimer's disease based on random forest and enrichment aalysis [J]. Scientia Sinica(Vitae). 2012, 42(8): 639-647.