STUDY ON CORRECTION OF DAILY PRECIPITATION DATA OF THE QINGHAI-TIBETAN PLATEAU WITH MACHINE LEARNING MODELS

Chen Ning[a], Yudan Wang[b], Zhuotong Nan[c], Hao Chen[a], Canran Liu[a]

[a]Baoji University of Science and Art, Baoji, 721013, China
[b]Cold and Arid Regions Environmental and Engineering Research Institute, Lanzhou, 730000, China
[c]School of Geography Science, Nanjing Normal University, Nanjing, 210023, China

* Corresponding author. E-mail: nanzt@njnu.edu.cn

ABSTRACT

The daily precipitation datasets of the Qinghai-Tibetan plateau (QTP) are mainly assimilated from remote sensing products and *in-situ* observations. The accuracy of those datasets needs further improvement with environmental and meteorological factors. This paper selected the related environmental and meteorological factors as input; k-Nearest Neighbor (KNN), Multivariate Adaptive Regression Splines (MARS), Support Vector Machine (SVM), Multinomial Log-linear Models (MLM) and Artificial Neural Networks (ANN) as correction models; 112 upscaled daily precipitation observations from the standard meteorological stations as ground truth to correct the commonly used ITPCAS and CMORPH daily precipitation of the QTP. Results show that the KNN model has the highest correction accuracy. The distribution of the corrected ITPCAS precipitation is nearer to the spatial pattern of the precipitation over the QTP than the corrected CMORPH precipitation. The correction accuracy is influenced by the precipitation distribution pattern of the original dataset.

Index Terms— machine learning model, precipitation data, the Qinghai-Tibetan Plateau, data correction

1. INTRODUCTION

Current daily precipitation datasets over the Qinghai Tibetan Plateau (QTP) are mainly assimilated from remote sensing products and in-situ observations. Previous studies have shown significant errors of the datasets in hydrological and climatic model simulation [1], which became challenges in modelling the future change of the QTP under climate change. Previous studies showed that environmental factors such as DEM and NDVI were used in the scale changes of precipitation [2] and meteorological factors were used in the simulation of occurrence, development and variation of precipitation [3]. Those multi-dimension data are highly correlated with the daily precipitation of the QTP. However, the mechanisms of correlation are not clear. Machine learning models can correct the daily precipitation with less priori knowledge and multi-dimension variables when proper algorithm is selected [3]. In section 2, 5 commonly used machine learning models including k-Nearest Neighbor (KNN) [4], Multivariate Adaptive Regression Splines (MARS), Support Vector Machine (SVM), Multinomial Log-linear Models (MLM) and Artificial Neural Networks (ANN) were evaluated in correction of the daily precipitation of the QTP by 5-fold cross validation. In section 3, the commonly used ITPCAS and CMORPH precipitation dataset of the QTP were corrected by the model with best performance. Errors of the corrected precipitation datasets were discussed. Finally, a summary of the study was made in section 4.

2. METHOD

2.1. Study area

The QTP (75.73°~104.33°E; 26.01°~39.69°N) is the highest and largest plateau in the world with the elevation above 4000m. The climate of the QTP is typical plateau continental with average annual temperature below 5 °C in most regions. The annual precipitations of most regions of the QTP are below 400 mm except in the southeastern QTP, where the annual precipitation is used to be more than 1000 mm. The precipitation of the QTP distributes unevenly with

extreme arid region in the northwest and pluvial region in the southeast.

2.2. Data

The Climate Prediction Center Morphing Method (CMORPH) precipitation is a global satellite precipitation product fused by the passive microwave data of the Tropical Rainfall Measuring Mission (TRMM) and the Infra-red data. It was downloaded from the NOAA website (ftp://ftp.cpc.ncep.noaa.gov/precip/CMORPH_V1.0/), in 3 hour time step and 8km×8km resolution. Air temperature, humidity, wind speed and precipitation were extracted from the ITPCAS China Meteorological Forcing Dataset. The ITPCAS precipitation is fused by the TRMM data and the precipitation data of the China Meteorological Administration standard meteorological stations [5]. Elevation, slope and aspect were extracted from the "China 1km DEM dataset" and NDVIs were extracted from the MODIS NDVI data with 250m×250m resolution and 16 days time steps. All the above data were downloaded from the Scientific Data Center of the Cold and Arid Regions and resampled into 8km×8km. The precipitation observations of the 112 standard meteorological stations of the QTP were downloaded from the China Meteorological Data sharing Service System (http://cdc.nmic.cn/) and used in cross validation.

2.3. Machine learning models

Given sample data and algorithms, machine learning models can quickly determine the dependency between input and output variables. The simulation accuracy of model varied with different algorithm. MARS divide environmental factors at break points to create piecewise function and optimize the contribution of each function to improve the RMSE of cross validation. KNN [4] can assign different weights to neighboring samples with different distances and shows superior performance in processing samples of uneven spatial distributions. SVM has concise structure and small computation cost. It is suitable for nonlinear simulation and small sample data. MLM analyzes the mathematical expectation of dependent variables to test the dependency of dependent and independent variables. It is suitable for nonlinear samples. ANN can process multi-variables and large sample set. But misconvergence happens occasionally. In this study, 8 envrionmental and meterological factors, including elevation, slope, aspect and daily NDVI, air temperature, humidity, wind speed and ITPCAS daily precipitation in 8km×8km resolution, were used as model input to simulate the corrected daily precipitation by the 5 models. Then the precipitation observations of the 112 standard meterological stations of the QTP were upscaled to 8km×8km as ground truth. The 5 models were evaluated in the 5-fold cross validation. The running parameters of the 5 models can be found in Table 1.

Table 1 Parameters of MARS, SVM, MLM, ANN and KNN models

Model Name	Package	Core function	Other parameters
MARS	earth	Piecewise function	default
SVM	e1071	Radial basis function	type=eps-regression
MLM	nnet	--	contrasts= contr.treatment
ANN	nnet	--	size=2; maxit=200
KNN	kknn	Gaussian Function	kmax=15; distance=1

2.4. Spatial pattern of the corrected precipitation datasets

The spatial pattern of the corrected ITPCAS and CMORPH precipitation were evaluated by comparing the original and corrected annual precipitation in 8 typical precipitation regions in the QTP.

3. RESULTS AND DISCUSSIONS

3.1. The RMSE of the 5-fold cross validation

The RMSE of the 5-fold cross validation of ITPCAS daily precipitation corrected with the 5 models was shown in Table 2 and Figure 1.

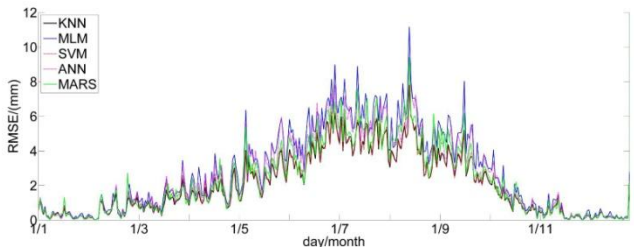

Figure 1 The RMSE of the 5-fold cross validation on daily precipitation of 112 stations corrected with 5 machine learning models

Table 2 The RMSE of the **5-fold cross validation on daily precipitation of 112 stations corrected with 5 machine learning models**

Model name	RMSE (mm)		
	Max	Min	Average
MLM	11.17615	0.002889	2.804681
ANN	8.864609	0.003175	2.52553
MARS	9.140058	0.003079	2.205364
SVM	7.797978	0.003134	2.049306
KNN	7.61608	0.00297	2.00799

Table 2 shows that ANN, MARS and MLM have larger RMSE than the KNN and SVM. The SVM and KNN have smaller RMSE in maxmium, minmuim and average RMSE. As KNN is suitable for large sample set with uneven spatial distribution, it is selected to evaluate the correction of the ITPCAS and CMORPH precipitation. Figure 1 show that the maxmium RMSE of the 5 models all happens in 18th August and the minimium happens in 5th Febrary, which means the occurance of the RMSE extremal is not related with algorithsm. The correction algorithm can only affect the value of the RMSE.

3.2. Temoral distribution of the RMSE

Taken the original ITPCAS and CMORPH precipitation as model input respectively, the daily precipitation of the two datasets are corrected with the KNN model. Then the RMSE of the original and the corrected precipitation were calculated against the ground truth of the 112 meteorological stations. Results can be seen in figure 2.

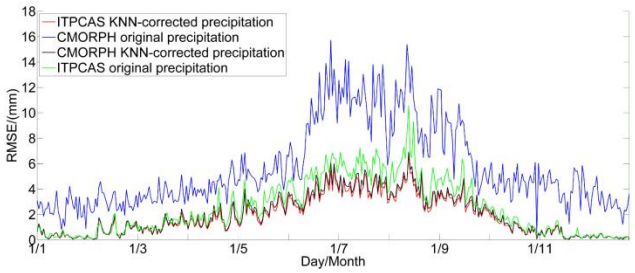

Figure 2 The RMSE of the ITPCAS and CMORPH original and corrected daily precipitation with the KNN model

Figure 2 show that the RMSE of the ITPCAS and the CMORPH precipitation all decrease significantly after correction, especially in the peak time of precipitation from June to September. The original CMORPH precipitation has the largest RMSE throughout the year. After correction, the RMSE of the corrected CMORPH precipitation are near to that of the corrected ITPCAS precipitation, which means that the KNN model can significantly correct the two precipitation datasets. The accuracy of the corrected ITPCAS precipitation is higher than the corrected CMOPH precipitation.

3.3. Spatial distribution of the corrected and original ITPCAS annual precipitation

Previous study [6] showed that there were 8 typical regions representing the precipitation spatial pattern of the QTP: (1) rain shadow region of the cold and arid core of the QTP; (2) rain shadow region of Karakoram Mountain; (3) rain shadow region of the northern slope of Himalayan; (4) relative pluvial region of Qiangtang Plateau; (5) arid region of Qaidam basin; (6) relative pluvial region of the southern slope of the Qilian Mountain; (7) relative rain shadow region of the central region of the Hengduan Mountain; (8) pluvial region of the YarlungTsangpo Great Canyon.

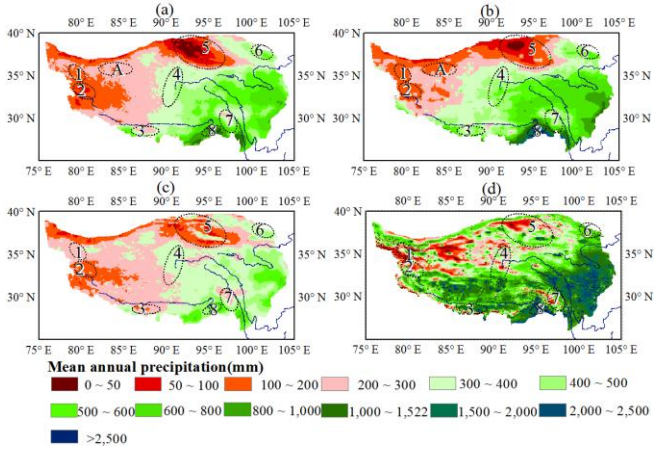

Figure 3 The spatial distributions of ITPCAS original (a) and KNN-corrected (b) precipitation and CMORPH original (c) and KNN-corrected (d) precipitation

.

Figure 3 show the distribution of the original and corrected CMORPH and ITPCAS annual precipitation in the QTP. Figure 3a and 3b tell that the original and the corrected ITPCAS precipitation all agree with the precipitation spatial pattern of the QTP in some extent. The corrected ITPCAS precipitation is nearer to the ground truth than the original ones in most of the typical regions except in region (7), where the rain shadow is overestimated by the corrected ITPCAS precipitation. Figure 3c and 3d show that the orignal CMORPH precipitation has large errors in precipitation distribution, while the accuracy of the corrected CMORPH is significantly improved. It agrees well with the ground truth in the eastern and southern QTP, but still presents large errors in the arid region of the western and northern QTP. For example, the corrected CMORPH presents 300-400 mm precipitation in some areas of region (5), which actually has annual precipitation below 100 mm. Generally speaking, the corrected CMORPH annual precipitation is less accurate than the corrected ITPCAS annual precipitation in spatial distribution. It is related with the inacurate spatial pattern of the original CMORPH precipitation, which presents too much precipitation in the arid region of the western and northern QTP. On the other hand, the lack of observation data in those regions made it difficult for the KNN model to correct the errors of the CMORPH precipitation.

4. SUMMARY

This study carried out correction studies on the ITPCAS and CMORPH daily precipitation of the QTP with 5 popular machine learning models, including KNN, MARS, MLM, SVM and ANN. KNN was proved to be more suitable in the correction of daily precipitation of the QTP with relatively smaller RMSE and fairly good spatial patterns of annual precicpitation of the QTP. Among the two corrected precipitation datasets, the ITPCAS was proved to more accurate than the CMORPH. The correction accuracy of the CMORPH dataset was greatly influenced by the inaccurate distribution of the original CMORPH precipitation in the arid regions, where meteorological stations are scarce. The KNN model has significantly decreased the RMSE and improved the spatial distribution of the original ITPCAS precipitation. Machine learning models are proved to be effective in the correction of daily precipitation datasets of the QTP.

.

5. REFERENCES

[1] Y. Shen, A.Y. Xiong, Y. Wang, et al., "Performance of high-resolution satellite precipitation products over China," *J. Geophys. Res.,* vol. 115, D2, DOI:10.1029/2009jd012097, 2010.

[2] S.F. Jia, W.B. Zhu, A.F. Lu, et al., "A statistical spatial downscaling algorithm of TRMM precipitation based on NDVI and DEM in the Qaidam Basin of China," *Remote. Sens. Environ.,* vol. 115, no. 12, pp. 3069-3079, 2011.

[3] X. Beuchat, B. Schaefli, M. Soutter, et al., "Toward a robust method for subdaily rainfall downscaling from daily data," *Water. Resour. Res.,* vol. 47, no. 9, DOI:10.1029/2010wr010342, 2011.

[4] Hart. "The Condensed Nearest Neighbor Rule," *IEEE T Inform Theory*., vol. 14, no. 3, pp. 515-516, 1968.

[5] Y. Y. Chen, K. Yang and J. He et al., "Improving land surface temperature modeling for dry land of China," *J. Geophys. Res.,* vol.116, D20104, doi:10.1029/2011JD015921, 2011.

[6] W.W. Qi, B. P. Zhang, Y. Pang, et al., "TRMM-data-based spatial and seasonal patterns of precipitation in the Qinghai-Tibet Plateau," *Sci. Geogr. Sin*., vol. 33, no. 8, pp. 999-1005, 2013. (in Chinese)