CORRECTION OF DAILY PRECIPITATION DATA OF ITPCAS DATASET OVER THE QINGHAI-TIBETAN PLATEAU WITH KNN MODEL

Yudan Wang[a,b], Zhuotong Nan[c], Hao Chen[d], Xiaobo Wu[a]

[a]Cold and Arid Regions Environmental and Engineering Research Institute, Lanzhou, 730000, China
[b]University of Chinese Academy of Sciences, Beijing, 100049, China
[c]School of Geography Science, Nanjing Normal University, Nanjing, 210023, China
[d]Baoji University of Science and Art, Baoji, 721013, China
* Corresponding author. E-mail: nanzt@njnu.edu.cn

ABSTRACT

As the meteorological stations in the Qinghai-Tibetan plateau (QTP) are scarcely and unevenly distributed, daily precipitation datasets generated from observation data and remote sensing inversion models are not accurate. The data accuracy can be improved by environmental and meteorological factors. This study selected k-Nearest Neighbor (KNN), a machine learning model, to correct the commonly used ITPCAS precipitation data over the QTP by establishing the relationship between daily precipitation and environmental (elevation, slope, aspect, vegetation) as well as meteorological factors (air temperature, humidity, wind speed). Error analysis shows that the KNN-corrected ITPCAS precipitation is more accurate than the original one. The spatial distribution of the corrected ITPCAS precipitation agrees well with the precipitation distribution pattern of the QTP. The error distribution of the corrected ITPCAS precipitation shows significant seasonal and regional characteristics.

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

1. INTRODUCTION

Precipitation datasets play an important role in simulating and forecasting the climatic and hydrological changes over the Qinghai-Tibetan Plateau (QTP) under climate change. As the remote sensing products of precipitation showed significant errors and *in situ* observations were scarcely and unevenly distributed in the QTP, precipitation datasets generated from those multi-sources data had unneglectable errors [1]. The controlling factors of the precipitation of the QTP varied greatly at different spatial and temporal scales. Previous studies showed that environmental and meteorological factors had significant correlationship with daily precipitation of the QTP [2, 3] and machine learning models can simulate the spatial and temporal distribution of daily precipitation with those multi-dimensional factors. In this study, k-nearest neighbor (KNN) model, a machine learning model that shows superior performance in processing samples of uneven spatial distributions, was used to correct the daily precipitation of the commonly used ITPCAS precipitation with the input of environmental (elevation, slope, aspect, vegetation) and meteorological (air temperature, wind speed, humid, surface pressure) factors. In section 2, the data sources and the running parameters of the KNN model were elaborated. Cross validation with ground truth upscale from the observations of 112 meteorological stations was employed in the KNN model. In section 3, ITPCAS precipitation of the QTP was corrected according to the established KNN model and errors of the corrected ITPCAS precipitation were analyzed. Spatial distribution of the corrected and original ITPCAS precipitation was discussed with knowledge of 8 typical precipitation regions of the QTP. Finally, a summary of this 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 plateau in the world with an area over 2.6 million square kilometers and elevation above 4000m. The mean air temperature is -6.4°C in winter and 10.5°C in summer. The annual precipitation of most regions of the QTP is below 400 mm and precipitation mainly occurs in summer. The spatial heterogeneity of precipitation is strong with annual precipitation around 800 mm to 1000 mm in the southeast regions and 50 mm to 100 mm in the northwest regions.

2.1. Data

Air temperature, humidity, wind speed and precipitation were extracted from the ITPCAS China Meteorological Forcing Dataset. The ITPCAS precipitation of the QTP is widely used and is fused from Tropical Rainfall Measuring Mission (TRMM) data and precipitation data of the standard meteorological stations [4]. 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 used in the cross validation and downloaded from China Meteorological Data sharing Service System (http://cdc.nmic.cn/).

2.3. The KNN model

As a popular machine learning model, KNN assigns different weights to neighboring samples with different distances and shows superior performance in processing samples with uneven spatial distributions. In this study, 8 most correlated environmental and meteorological factors to daily precipitation of the QTP, 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. The KNN model was developed in R. The R package is kknn. The core function is Gaussian function. The maximum neighboring point kmax is 15 and the distance to target point is 1.

2.4. Correction and error analysis of ITPCAS precipitation

Precipitation data of 112 meteorological stations were upscaled into 8km×8km grids by MicroMet [5], a meteorological model that provides precipitation scale changes according to elevation, vegetation and precipitation lapse rate. The upscaled data were used as ground truth in the model validation. The KNN model was validated by the R-based 5-fold cross validation. The errors of the KNN model were analyzed in the following way:

(1) Temporal distribution of the errors: mean values of the corrected daily precipitation from 2009 to 2012 were calculated by the mean daily precipitation of the 112 meteorological stations simulated by the KNN model. Then the RMSE, bias, relative bias and correlation coefficient of the corrected and original precipitation to the observed precipitation were calculated respectively. The seasonal variation of the simulation errors were analyzed accordingly.

(2) Spatial distribution of the errors: mean values of the corrected annual precipitation from 2009 to 2012 were calculated by the accumulated daily precipitation of each meteorological stations simulated by the KNN model. Then the RMSE, bias, relative bias and correlation coefficient of the corrected and original precipitation to the observed precipitation were calculated respectively. The spatial variation of the simulation errors were analyzed accordingly.
The spatial distribution of the corrected ITPCAS 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. Temporal distribution of the errors

Temporal distribution of the errors was shown in Table 1 and Figure 1.

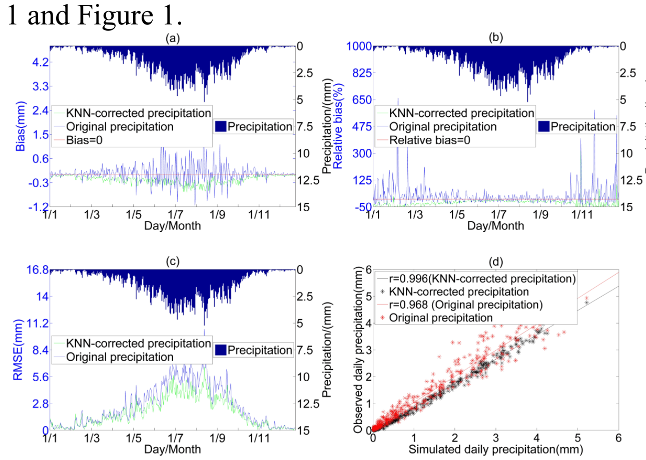

Figure 1 The daily (a) bias, (b) relative bias, (c) RMSE and (d) correlation coefficient of ITPCAS original and KNN-corrected daily precipitation

Table 1 The analysis of daily bias and relative bias of original and KNN-corrected ITPCAS precipitation

error	overestimate (day)	underestimate (day)	overestimate (max)	underestimate (max)
Bias (original)*	215	150	1.63 mm	1.14 Mm
Bias (corrected)**	7	358	0.09 mm	0.65 Mm
Relative bias(original)	215	150	659%	50.1%
Relative bias(original)	7	358	316%	81.8%

*original means the original ITPCAS precipitation
**corrected means the corrected ITPCAS precipitation

Table 1 show that the corrected ITPCAS precipitations mainly underestimate the ground truth. Overestimation only occurs in 7 days of the year, while the original ITPCAS precipitations mainly show overestimation with more than 255 days of the year. Fig 1a, 1b show that from June to September, when daily precipitations are abundant, biases of corrected and original ITPCAS precipitation are high while the relative bias are low. From November to February, when the daily precipitations are little, biases of the corrected and original ITPCAS precipitation are low, while the relative biases are high. Table 1 and Figure 1 show that the biases, relative biases, RMSE and correlation coefficients of the ITPCAS precipitation significantly decreased after correction, especially from June to September when precipitation is abundant. The corrected ITPCAS precipitation still shows significant seasonal trend with large bias, large RMSE and small relative bias in summer; small bias, small RMSE and large relative bias in winter. Finger 5d shows that the correlation coefficient of the corrected ITPCAS precipitation with the ground truth is higher than that of the original one. It can be concluded that the corrected ITPCAS daily precipitation data are more accurate than the original data.

3.2. Spatial distribution of the errors

Table 2 The analysis of per-station bias and relative bias of original and KNN-corrected ITPCAS annual precipitation

error	overestimate (number)	underestimate(number)	overestimate (max)	underestimate(max)
Bias (original)*	90	22	245.5 mm	384.8 mm
Bias (corrected)**	31	81	101.4 mm	348.9 mm
Relative bias(original)	90	22	420%	79%
Relative bias(original)	31	81	244%	37%

*original means the original ITPCAS precipitation
**corrected means the corrected ITPCAS precipitation

Table 2 shows that the corrected ITPCAS precipitations mainly underestimate the ground truth with 72.3% or 81 underestimated stations, while the original ITPCAS precipitations mainly show overestimation with 80.4% or 90 underestimated stations. The maximum underestimation and overestimation of biases and relative biases of the corrected ITPCAS precipitation significantly decrease compared to those of the original ITPCAS precipitation. The correlation coefficient of the corrected ITPCAS precipitation and the original ITPCAS precipitation with ground truth are all around 0.995. It can be concluded that the corrected ITPCAS annual precipitation data of each station is more accurate than the original ones.

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 characteristics 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.

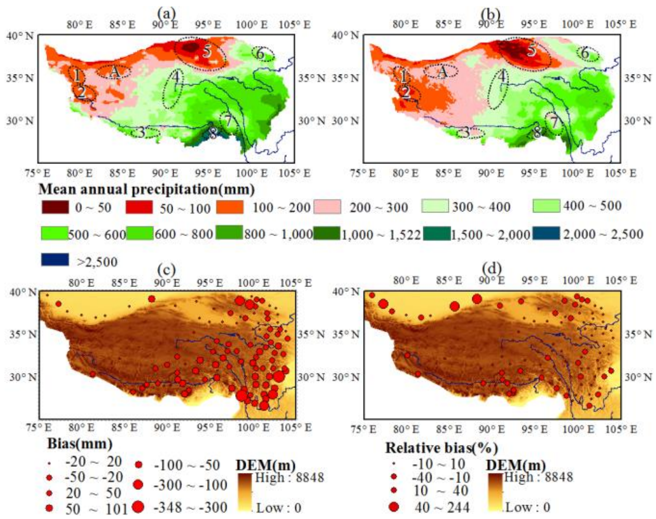

Figure 2 The spatial distributions of the original (a) and (b) the KNN-corrected ITPCAS precipitation and (c) the bias and (d) relative bias of the corrected ITPCAS precipitation

The spatial distributions of the corrected and original ITPCAS annual precipitation of the 8 typical regions can be found in Figure 2. It shows that the corrected and original ITPCAS precipitation agrees well with the precipitation distribution characteristics of the QTP. The annual precipitations of the eastern, central and southern QTP decrease significantly after correction, which makes the corrected precipitation in region (3), (4), (6) and (8) nearer to the ground truth and current knowledge. Only in region (7), the original ITPCAS precipitation is nearer to the ground truth. In the northern and western QTP, the corrected precipitation increases the area of arid region (2) and (5), which is nearer to the ground truth. In region (1), the corrected and original ITPCAS precipitation is both inaccurate. Generally, the corrected ITPCAS annual precipitations are nearer to the ground truth than the original ones in spatial distribution.

Figure 2c and 2d show that the distribution of errors of the corrected ITPCAS annual precipitation presents regional characteristics. The bias is large in the Hengduan Mountain and Qilian Mountain, while the relative bias is large in the arid region of the northern border of the QTP.

4. SUMMARY

This study carried out corrections of the ITPCAS daily precipitation of the QTP with the KNN model and environmental and meteorological factors. Error analysis showed that the correction significantly improved the overestimation of the ITPCAS precipitation and increased the accuracy in temporal and spatial scale. The error distribution of the corrected ITPCAS precipitation showed remarkable seasonal and regional characteristics. The summer precipitation had large bias, small RMSE and relative bias, while the winter precipitation has large relative bias, small bias and RMSE. The bias was large in Hengduan and Qilian Mountains. The relative bias was large in the arid region of the northern border of the QTP. The spatial distribution of the corrected ITPCAS annual precipitation is nearer to ground truth than the original ones. The KNN model is proved to be effective in the correction of the daily ITPCAS precipitation.

.

5. REFERENCES

[1] B.Y. Kan, F.G. Su, K. Tong, et al., "Analysis of the Applicability of Four Precipitation Datasets in the Upper Reaches of the Yarkant River,the Karakorum," *J. Glaciol. Geocryol.,* vol. 35, no.3, pp. 710-722, 2013.

[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] 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.

[5] G. E. Liston, K. Elder, "A meteorological distribution system for high-resolution terrestrial modeling (MicroMet)," *J. Hydrometeorol.,* vol. 7, no. 2, pp. 217-234. 2006.

[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)