Remote Sensing of Environment xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

A spatial data mining algorithm for downscaling TMPA 3B43 V7 data over the Qinghai–Tibet Plateau with the effects of systematic anomalies removed

Ziqiang Ma[a], Zhou Shi[a,*], Yin Zhou[a], Junfeng Xu[b], Wu Yu[c], Yuanyuan Yang[a]

[a] *Institute of Agricultural Remote Sensing and Information Technology Application, College of Environmental and Resource Sciences, Zhejiang University, Hangzhou 310058, China*
[b] *Institue of Remote Sensing and Earth Science, Hangzhou Normal University, Hangzhou 310029, China*
[c] *College of Resources and Environment, Tibet Agriculture and Animal Husbandry College, Linzhi, Tibet 860114, China*

ARTICLE INFO

Keywords:
Downscaling
TMP A3B43 V7
Precipitation
Data mining
Qinghai–Tibet Plateau

ABSTRACT

Precipitation plays an important role in the water cycle and in matter and energy exchanges. Acquiring accurate information on precipitation over the Qinghai–Tibet Plateau, which has a limited rain gauge network, has been a great challenge. Downscaling the TRMM Multisatellite Precipitation Analysis (TMPA) 3B43 Version 7 dataset (0.25° resolution) provided an optimal approach to estimating precipitation at 1 km resolution over this highland plateau. Our downscaling assumptions were that non-stationary relationships between precipitation and land surface characteristics occur and have varying two-dimensional scale effects, and that the relationships vary in different sub-regions having differing combinations of land surface characteristics, including vegetation index, topographical factors, and land surface temperatures. We used Cubist (a spatial data mining algorithm) to implement our assumption. Cubist separated the Qinghai–Tibet Plateau into sub-regions according to geographical similarities, and selected the most effective variables over each sub-region to build models. We found that: (1) the downscaled results using this algorithm were more accurate and precise than other commonly used algorithms (e.g., geographically weighted regression) and the original TMPA data at 0.25° resolution; (2) DEM showed limited correlation with precipitation over the Qinghai–Tibet Plateau; and (3) the effects of systematic anomalies in the original TMPA data were removed in the downscaled results based on Cubist. We conclude that Cubist is a promising algorithm able to take hundreds of variables into consideration, and in this study was used to retrieve precipitation estimates at approximately 1 km resolution.

1. Introduction

Precipitation plays an important role in the water cycle and in matter and energy exchanges. Accurate information on precipitation amounts and patterns is necessary to enable exploration of the impacts of precipitation on soil moisture, vegetation cover and growth, and water supply to glaciers, and to enable correlations between precipitation and climate change (Guo et al., 2004; Michaelides et al., 2009; Teng et al., 2014; Liu et al., 2016). The Qinghai–Tibet Plateau is referred to as the Earth's "third pole" because it has the largest glaciers outside the south and north pole areas, and as the "water tower of Asia" because it is the headwater area for many major Asian rivers, including the Indus, Ganges, Brahmaputra, Yangtze, and Yellow rivers (Lin and Zhao, 1996; Yao et al., 2012); > 1.4 billion people depend on water from these rivers (Immerzeel et al., 2010). However, accurate information on precipitation over the Qinghai–Tibet Plateau is difficult to obtain; because of its harsh environment, remoteness, and poor transportation networks, and the sparse distribution of rain gauge stations (Xie and Xiong, 2011). Therefore, new approaches are needed to provide accurate high spatial resolution (approximately 1 km) estimates of precipitation over the Qinghai–Tibet Plateau.

Precipitation information at high spatial resolution is typically generated using spatial interpolation methods based on ground measurements. However, interpolation errors occur for areas where rain gauges are sparse, especially those regions having complex topography (Kyriakidis et al., 2001; McVicar et al., 2007; Nastos et al., 2016). In these situations remote sensing has commonly been used to produce gridded precipitation estimates (Joyce et al., 2004; Michaelides et al., 2009). Over the last three decades, projects including the Global Precipitation Climatology Project (GPCP) (Huffman et al., 1997, 2001, 2009; Adler et al., 2003), the Tropical Rainfall Measuring Mission (TRMM) (Kummerow et al., 1998, 2000; Huffman et al., 2007), and the Global Satellite Mapping of Precipitation (GSMaP) project (Kubota et al., 2007) have gathered satellite-based precipitation estimates.

* Corresponding author.
E-mail address: shizhou@zju.edu.cn (Z. Shi).

http://dx.doi.org/10.1016/j.rse.2017.08.023
Received 3 December 2016; Received in revised form 4 August 2017; Accepted 14 August 2017

Please cite this article as: Ma, Z., Remote Sensing of Environment (2017), http://dx.doi.org/10.1016/j.rse.2017.08.023

Although these satellite-based estimates provide more information on precipitation for areas having limited rain gauges, their spatial resolutions (e.g., 0.25–0.5°) are too coarse for application in localized regions and watersheds, or for parameterization of hydrological and meteorological models at the local scale (Tao and Barros, 2010). Therefore, the use of downscaling models and differing combinations of environmental variables as auxiliary data have been explored to derive precipitation estimates at finer spatial scales (Kustas et al., 2003; Agam et al., 2007; Merlin et al., 2010). To downscale the data from TRMM Multisatellite Precipitation Analysis (TMPA) 3B43 Version 7, Immerzeel et al. (2009) demonstrated that there was an exponential relationship between precipitation and the normalized difference vegetation index (NDVI). Because of the effects of topography in the Qaidam Basin, Jia et al. (2011) introduced a multiple linear regression model to downscale TMPA data for the area, using both the NDVI and the digital elevation model (DEM). To improve accuracy and precision, Duan and Bastiaanssen (2013) proposed two downscaling-calibration methods (geographical difference analysis: GDA; geographical ratio analysis: GRA) for use with rain gauge observations following downscaling of TPMA data using the NDVI. Previous downscaling algorithms were based on the assumption that the relationships between precipitation and environmental variables (the NDVI and/or DEM) were constant over a study area. However, Xu et al. (2015) and Chen et al. (2015) considered that these relationships varied spatially, and so introduced a moving window regression method (geographically weighted regression: GWR), in which each regression point has a regression relationship with neighboring data points. These studies mainly explored the NDVI and/or DEM as the environmental covariates in downscaling the coarse TMPA data.

However, other studies have demonstrated that the spatial patterns of precipitation are influenced by other land surface characteristics (Shi et al., 2015). In simulating precipitation at a global scale, Schultz and Halpert (1994) found that coupling land surface temperatures with the NDVI improved the precision and accuracy compared with use of the NDVI alone. Chen et al. (2015) used land surface temperatures to improve spatial downscaling using GWR, while Fang et al. (2013) suggested that topographical factors influence the patterns of spatial distribution of precipitation.

Cubist is a spatial data mining algorithm that applies a divide-and-conquer strategy. We used this algorithm to downscale TMPA 3B43 V7 data over the Qinghai–Tibet Plateau at an annual scale, for the period 2000 to 2013, based on the assumptions that the non-stationary relationships between precipitation and land surface characteristics occur with varying two-dimensional scale effects, and that the relationships vary among sub-regions having differing land surface variables. The main objectives of this study were: (1) to provide a spatial data mining algorithm to downscale coarse satellite-based estimates; (2) to compare the downscaled results based on Cubist with those based on GWR; and (3) to assess whether the effects of anomalies in the original TMPA data could be removed using Cubist.

2. Study area and materials

2.1. Study area

This study focused on the Qinghai–Tibet Plateau, in the southwest region of China. It occupies an area of approximately 2.57 million km^2 (Zhang et al., 2002) extending from 26°00′12″ N to 39°46′50″ N, and 73°18′52″ E to 104°46′59″ E (Fig. 1). The Qinghai–Tibet Plateau is the highest plateau in the world, having a mean elevation > 4000 m, and is commonly referred to as “the roof of the world”. Because of its unique topography and location, the Qinghai–Tibet Plateau has complex climatic conditions and various vegetation types (An et al., 2001). Two main atmospheric circulation patterns affect the Qinghai–Tibet Plateau: the Indian monsoon in summer, and the westerlies in winter (Yao et al., 2012). The central and southern parts of the plateau are dominated by the Indian monsoon, and receive more precipitation than the western and northern parts of the plateau, which receive relatively little precipitation under the influence of the westerlies. Although the southeastern Asian monsoon prevails in the eastern part of the Qinghai–Tibet Plateau, it brings limited precipitation into this area because the Hengduan Mountains block most of the moist air (Shen et al., 2011). Because of these atmospheric patterns, there is a clear demarcation between the dry and wet seasons; approximately 90% of precipitation occurs in the wet season, mainly from March to August (Shi et al., 2015). During the period 1982–2000, the average air temperature in the warmest month (July) ranged from 7 to 15 °C, and in the coldest month (January) ranged from − 1 to − 7 °C (Zhong et al., 2011).

Grasslands, forest, and shrub-lands cover the Qinghai–Tibet Plateau (Shen et al., 2011). Grasslands, including meadows and steppes, occupy approximately 70% of the plateau, and are mostly distributed in the central and eastern regions (Piao et al., 2011). The growing season in the Qinghai–Tibet Plateau coincides with the annual precipitation patterns (Che et al., 2014). Almazroui (2011) suggested that at the annual scale in the TRMM era, among the satellite-based precipitation estimates the TMPA data shows the most similar rainfall patterns to those observed from rain gauges (Chen et al., 2013). Thus, in this study we downscaled the TMPA data at the annual scale.

2.2. Datasets and processing

2.2.1. Rain gauge data

The observational data used in this study were provided by the Chinese Meteorological Data Sharing Service System (http://cdc.nmic.cn/home.do). The rain gauge network on the Qinghai–Tibet Plateau is sparse, because of the harsh environmental conditions and complicated topography. We obtained data from 106 meteorological stations that are distributed unevenly across the plateau (Fig. 1).

2.2.2. TRMM satellite precipitation dataset

The TRMM satellite was jointly sponsored in 1997 by the National Aeronautics and Space Administration (NASA) and the Japanese Space Agency (JAXA) to monitor and investigate tropical and subtropical rain systems (Kummerow et al., 1998). The TRMM satellite contains a precipitation radar, a TRMM microwave imager, and a visible and infrared scanner. The TMPA 3V43 V7 data were first published in May 2012 (http://mirador.gsfc.nasa.gov) at a spatial resolution of 0.25° for latitudes from 50°N to 50°S worldwide (Huffman et al., 2007). To determine the dry and wet years, we averaged all TMPA values and the rainfall observations over the Qinghai–Tibet Plateau for each year from 1998 to 2013 (Fig. 2). The TMPA data and the ground observations were similar, and the wettest and driest years were 1998 and 2006, respectively. Since 18 February 2000 the moderate resolution imaging spectroradiometer (MODIS) aboard the Terra satellite (MODIS Terra) has provided high temporal and spatial resolution earth observations (Li et al., 2013), including land surface temperatures, so we downscaled the TMPA data from 2000 to 2013.

2.2.3. NDVI datasets

The NDVI is an indicator of vegetative productivity, and is correlated with precipitation over spatial and temporal scales. The NDVI from MOD13A1 in the latest MODIS Collection 6 is generated from the MODIS Terra at a resolution of 16 days and 500 m, and was obtained from the Land Processes Distributed Active Archive Center (LP DAAC) (https://ladsweb.nascom.nasa.gov/data/search.html). There were twenty 16-day periods in 2000 following initiation of MODIS Terra, and in the following years there have been twenty-three 16-day periods each year. The MODIS data are processed from atmospherically-corrected bi-directional surface reflectance in the red and near infrared wavebands, but the reflectance can be masked by water, clouds, dense atmospheric aerosols, and cloud shadows. For each year of the study we averaged all the NDVI values over the Qinghai–Tibet Plateau, to