Postprint: Variation Characteristics of Summer Extreme High Temperature in Central Asia Based on Random Forest Interpolation

Authors: Meng Xinning, Jiao Ruili, Liu Nian, Xia Jiangjiang, YAN Zhongwei, Yu Shuang, Lou Xiao, Li Haochen, Wang Lizhi, Chen Liang, Zheng Ziyan, Zhao Na, Meng Xinning, Xia Jiangjiang

Date: 2020-07-02T00:00:00+00:00

Abstract

Using daily maximum temperature data from 65 meteorological stations across Central Asia, combined with ERA-Interim reanalysis data and latitude, longitude, and elevation data, a random forest interpolation model was constructed and its reliability validated. Based on this model, missing values at the meteorological stations were filled to obtain a complete station daily maximum temperature dataset TStation_f, and interpolation was performed to generate a daily maximum temperature gridded dataset TRFIM_G for Central Asia from 1979 to 2016 with a spatial resolution of $0.75°$\times$0.75°$. $The spatiotemporal variation characteristics of summer extreme high temperature indices in Centr$ $(10a)^{-1}$, with significant warming primarily distributed in western Kazakhstan, most of Turkmenistan, southeastern Uzbekistan, and other regions. The increasing trend of summer extreme high temperature indices derived from TRFIM_G is significantly greater than that derived from TStation_f, suggesting that estimates of summer extreme high temperature trends in this region based on station observation data are notably underestimated. The dataset obtained in this study can to some extent compensate for the limitation of using station observation data to unilaterally characterize extreme high temperature changes in Central Asia, and can help guide more accurately the implementation of appropriate mitigation and adaptation measures in response to extreme weather and climate events.

Full Text

Preamble

Change of Summer Extreme High Temperature in Central Asia Based on Random Forest Interpolation

MENG Xin-ning[1], JIAO Rui-li[1], LIU Nian[2],[3], XIA Jiang-jiang[2],[3]*, YAN Zhong-wei[2],[3], YU Shuang[2],[3], LOU Xiao[2], LI Hao-chen[4],[5], WANG Li-zhi[2], CHEN Liang[2], ZHENG Zi-yan[2], ZHAO Na[6]

[1]Beijing Information Science and Technology University, Beijing 100101, China
[2]Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing 100029, China
[3]University of Chinese Academy of Sciences, Beijing 100049, China
[4]School of Science, Beijing University of Posts and Telecommunications, Beijing 100876, China
[5]Peking University, Beijing 100871, China
[6]Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China

Abstract

Using daily maximum temperature data from 65 meteorological stations in Central Asia, combined with ERA-Interim reanalysis data and geographic information (latitude, longitude, and altitude), we constructed a random forest interpolation model and validated its reliability. Based on this model, we filled missing values at meteorological stations to obtain a complete daily maximum temperature dataset (TStation_f), and subsequently generated a gridded daily maximum temperature dataset for Central Asia (TRFIM_G) at 0.75°$\times$0.75°$spatial resolution for the period 1979–2016. Using TRFIM_G, we further analyzed the spatio temporal average extreme hight temperature indices increased at rates of 0.22–0.30℃ · (10a)^{-1}$, with significant warming primarily distributed in western Kazakhstan, most of Turkmenistan, and southeastern Uzbekistan. The warming trends derived from TRFIM_G are substantially greater than those based on TStation_f, indicating that estimates of summer extreme high temperature trends using station observations alone are significantly underestimated. The dataset obtained in this study can help overcome the limitations of using sparse station data to characterize extreme high temperature changes in Central Asia, thereby providing more reliable guidance for mitigation and adaptation measures against extreme weather and climate events.

Keywords: random forest interpolation; machine learning; summer extreme high temperature; Central Asia

Introduction

The increasing frequency of extreme high temperature events can negatively impact human health, ecosystems, and socioeconomic systems. In recent years, extreme high temperature events have shown an increasing trend globally. Central Asia (Kazakhstan, Tajikistan, Kyrgyzstan, Uzbekistan, and Turkmenistan; see [Figure 1: see original paper]) serves as a core region of the Belt and Road Initiative. This region experiences dramatic temperature variations, large elevation differences, and frequent high-temperature weather in summer. Characterized by a spatially heterogeneous oasis-desert pattern, Central Asia is highly sensitive to global climate change and prone to rapid hydrological and geomorphological changes. Studying summer extreme high temperature changes in Central Asia not only deepens our understanding of regional climate change patterns but also provides a foundation for developing targeted mitigation and adaptation strategies to ensure the sustainable development of the Belt and Road Initiative.

Previous studies have demonstrated a steady warming trend in Central Asia over recent decades, with a faster warming rate than the Northern Hemisphere average. The frequency, intensity, and duration of historical extreme high temperature events such as heatwaves have also increased. However, these studies relied on either meteorological station data or existing gridded datasets. Central Asia suffers from sparse and unevenly distributed meteorological stations, making it difficult for station-based analyses to represent regional climate changes. While gridded temperature datasets can substitute for station data in analyzing regional extreme weather and climate events, previous gridded temperature datasets for Central Asia were either non-daily, did not specifically represent daily maximum temperature, or had insufficient temporal coverage, rendering them unsuitable for analyzing summer extreme high temperature events. Consequently, under the backdrop of global warming over the past century, no high spatiotemporal resolution data have been available to accurately characterize historical changes in summer extreme high temperatures in Central Asia.

To better understand the characteristics of summer extreme high temperature changes in Central Asia, it is essential to interpolate the existing sparse station maximum temperature data onto a grid to obtain a high-resolution gridded daily maximum temperature dataset. Traditional interpolation techniques in meteorology and climatology, such as nearest neighbor, spline, regression, and kriging, are primarily based on statistical methods that require subjective prior knowledge and typically involve specific variables. These methods may produce flawed interpolated datasets due to incomplete understanding of physical processes. In contrast, machine learning techniques operate through adaptive mechanisms, enabling them to learn from data without relying on assumptions, capture unknown or difficult-to-describe functional relationships, and easily handle diverse data sources. Machine learning is thus becoming a common interpolation technique. Previous studies have shown that the random forest algorithm in machine learning performs excellently in spatial interpolation of environmental variables and monthly temperatures.

This study uses ERA-Interim reanalysis data, elevation, latitude and longitude data as input features, and daily maximum temperatures from 65 Central Asian stations as output labels to construct an interpolation model using the random forest algorithm. The model is then applied to fill missing station data and perform spatial interpolation of gridded data, enabling analysis of summer extreme high temperature characteristics in the region.

1.1 Station Observation Data

Daily maximum temperature station data used in this study were obtained from NCDC (ftp://ftp.ncdc.noaa.gov/pub/data/gsod) and GHCN-D (https://www1.ncdc.noaa.gov/pub/data/ghcn/daily/). Quality control procedures were applied as follows: if the percentage of missing daily maximum temperature data during June–August (summer) for a given station in a particular year exceeded 14%, that station-year was deemed unusable. Additionally, if a station had more than 10% unusable years during 1979–2016, the entire station record was discarded. This process yielded daily maximum temperature data for summer (June–August) from 1979 to 2016 for 65 stations (see [Figure 1: see original paper]). The multi-year missing data patterns for each station are shown in [Figure 2: see original paper].

1.2 Reanalysis Data

Reanalysis data are historical gridded meteorological datasets produced by numerical weather prediction models and data assimilation techniques driven by multiple observation sources. Due to their spatial continuity, reanalysis data can overcome the limitations of sparse station observations and serve as an important data source for regional climate change research. This study employs the ERA-Interim global atmospheric reanalysis dataset from the European Centre for Medium-Range Weather Forecasts (ECMWF). ERA-Interim has been shown to match station observations well across different regions globally and is suitable for climate research in Central Asia due to its high spatial precision and quality. The dataset has a spatial resolution of 0.75°$×$0.75° and includes 48 meteorological variables (https://apps.ecmwf.int/datasets/data/interim-full-daily/levtype=sfc/). For this study, we extracted all meteorological variables corresponding to the time of day with highest temperature (12:00 UTC) for the period 1979–2016.

Table 1 Dataset information used in this study

Dataset	Source	Spatial Resolution
Station observations	NCDC/GHCN-D	Point data
Reanalysis	ERA-Interim (ECMWF)	$0.75°\$\times\$0.75°\|\|Interpolatedstationdata\|RFIMmodeloutput\|Pointdat$

2.1 Interpolation Algorithm and Evaluation Methods

This study employs random forest (RF) to "predict" daily maximum temperatures in Central Asia, which constitutes both station data imputation and gridded data interpolation. RF is an ensemble learning algorithm based on decision trees. For regression problems, the final prediction is the average of multiple decision tree predictions. The algorithm offers advantages including fast computation, high robustness, and low susceptibility to overfitting. We use grid search to tune RF parameters and root mean square error (RMSE) to evaluate model performance. The optimal model is then used to fill missing station data and perform spatial interpolation of gridded data.

2.2 Research Scheme

The training, validation, and testing datasets were constructed from seven elements: daily maximum temperatures at 65 stations, longitude, latitude, elevation (Table 1), year, month, and day, plus all 48 meteorological variables from the nearest ERA-Interim grid point to each station. This yielded a total of 54 explanatory variables (features) with station daily maximum temperature as the response variable (label).

2.2.1 RFIM Model Establishment and Evaluation Model construction utilized station daily maximum temperatures and their corresponding features (see [Figure 3: see original paper]). Station observations and features were divided into three parts: training set (80%), validation set (20%), and test set (2016 data). The training dataset was used to develop the initial RFIM (Random Forest Interpolation Model), which typically required hyperparameter tuning (e.g., maximum features, number of trees) using validation set error as reference. When RMSE on the validation set ceased to decrease significantly, the model was considered optimally fitted. The test dataset (2016 data) was then used to assess generalization capability by comparing predicted daily maximum temperatures against actual observations and calculating RMSE. The final optimized RFIM model was saved for subsequent missing data imputation and spatial interpolation.

For clarity, we define: TStation as observed station daily maximum temperature, TERA as ERA-Interim grid temperature at the nearest grid point, TRFIM_S as RFIM-predicted station daily maximum temperature, TRFIM_G as RFIM-predicted gridded daily maximum temperature, and TStation_f as the completed station daily maximum temperature dataset.

2.2.2 Missing Data Imputation and Gridded Data Interpolation For missing station data, explanatory variables from the nearest grid point to the missing station were input into the RFIM model to predict the missing daily maximum temperature values, thereby completing each station' s record. By inputting explanatory variables for all Central Asian grid points into the RFIM model, we generated a gridded daily maximum temperature dataset

(TRFIM_G) at 0.75°$×$0.75° resolution for the region. This dataset was then used to analyze summer extreme high temperature characteristics.

Results

3.1 RFIM Model Performance Evaluation

As shown in [Figure 4: see original paper], the RMSE between RFIM-predicted station temperatures (TRFIM_S) and observations (TStation) averaged 1.87°C across all 65 stations, significantly lower than the RMSE between ERA-Interim grid temperatures (TERA) and station observations (3.81°C). Moreover, TRFIM_S versus TStation RMSE showed little variation across stations, indicating high reliability and stability of RFIM predictions.

To visually verify RFIM accuracy, we compared regional averages of TRFIM_S and TERA against observations for 2016 ([Figure 5: see original paper]). TERA exhibited systematic underestimation of station temperatures, consistent with previous findings. In contrast, TRFIM_S closely matched observed TStation values, demonstrating that RFIM predictions better approximate actual observations and are suitable for climate change analysis in Central Asia.

3.2 Characteristics of Summer Extreme High Temperature Changes in Central Asia

Using the TRFIM_G dataset (Section 2.2.2), we calculated summer extreme high temperature intensity indices by averaging the top n (n = 1, 5, 10, 15) daily maximum temperatures each summer (June–August), denoted as TXn. Linear trends were computed for these four indices and tested for significance using moving t-tests ($\alpha = 0.05$). Corresponding TXn values based on TStation_f were also calculated for comparison.

As shown in [Figure 6: see original paper], region-averaged summer extreme high temperature intensity based on TRFIM_G exhibits increasing trends, with TX1, TX5, TX10, and TX15 increasing at rates of 0.22, 0.27, 0.30, and 0.30℃ · $(10a)^{-1}$, respectively. The trend magnitudes increase with n, indicating that more "average" extreme indices show greater warming rates than more "extreme" indices. All four trends are statistically significant. In contrast, trends based on TStation_f are 0.02, 0.12, 0.16, and 0.19℃ · $(10a)^{-1}$ for TX1, TX5, TX10, and TX15, respectively, with only TX15 reaching statistical significance. This demonstrates that using station data alone significantly underestimates the increasing trend of summer extreme high temperature intensity in Central Asia.

Spatial patterns of linear trends for the four indices ([Figure 7: see original paper]) show consistent warming across most of Central Asia based on TRFIM_G. Significant warming is concentrated in western Kazakhstan, most of Turkmenistan, and southeastern Uzbekistan, while Tajikistan shows non-significant warming trends. Most areas of Kyrgyzstan exhibit cooling

trends for all four indices, though these are non-significant for TX1, TX5, and TX10, with only some areas showing significant cooling for TX15.

For TX15 specifically, region-averaged trends in the significantly warming areas are $0.6℃ \cdot (10a)^{-1}$ in western Kazakhstan, $0.30℃ \cdot (10a)^{-1}$ across most of Turkmenistan, and $0.32℃ \cdot (10a)^{-1}$ in southeastern Uzbekistan. Previous studies of summer mean temperature changes in Central Asia have also found faster warming in the west than in the east, consistent with our results.

Based on TStation_f ([Figure 7: see original paper]), 8, 9, 17, and 22 stations show significant trends for TX1, TX5, TX10, and TX15, respectively. The spatial patterns are generally consistent with TRFIM_G results, confirming the reliability of the gridded interpolation. However, the limited station data cannot capture fine-scale spatial variability in trend patterns.

Conclusion

Previous research on temperature interpolation and climate change impacts has predominantly used station observations directly, which cannot adequately characterize climate change in data-sparse regions. While some studies have employed gridded meteorological data, these were either derived from traditional interpolation methods or used reanalysis data directly. Such statistically based approaches introduce subjectivity, and reanalysis data can introduce spurious climate signals due to numerical schemes and assimilation methods, affecting analysis results.

This study constructed a random forest interpolation model using daily maximum temperature data from 65 Central Asian meteorological stations, ERA-Interim reanalysis data, and geographic information. The resulting gridded daily maximum temperature dataset (TRFIM_G) for Central Asia at 0.75°$×$0.75° resolution for 1979–2016 reveals that summer extreme high temperature intensity has increased at rates significantly greater than those based on TStation_f. This indicates that limited station observations substantially underestimate summer extreme high temperature trends, leading to inadequate understanding of historical climate change and underestimation of risks from associated extreme events (heatwaves, etc.), potentially rendering mitigation and adaptation measures less effective. The proposed random forest-based approach ensures objectivity in interpolation data while avoiding analysis errors introduced by reanalysis data, enabling high-resolution climate change studies even in data-sparse regions.

References

[1] Yu Shuang, Xia Jiangjiang, Yan Zhongwei, et al. Loss of work productivity in a warming world: Differences between developed and developing countries[J]. Journal of Cleaner Production, 2019, 208(6): 1219-1225.

[2] Xia Jiangjiang, Tu Kai, Yan Zhongwei, et al. The super-heat wave in eastern

China during July-August 2013: a perspective of climate change[J]. International Journal of Climatology, 2016, 36(3): 1291-1298.

[3] Jiao Wenhui, Zhang Bo, Huang Tao, et al. Spatiotemporal change of extreme temperature in the Hedong Region in recent 30 years[J]. Arid Zone Research, 2019, 36(6): 1466-1477.

[4] Gong Zitong, Chen Hongzhao, Yang Fan, et al. Pedogeochemistry and environment of aridisol regions in Central Asia[J]. Arid Zone Research, 2017, 34(1): 1-9.

[5] Xu Ting, Shao Hua, Zhang Chi. Temporal pattern analysis of air temperature change in Central Asia during 1980-2011[J]. Arid Land Geography, 2015, 38(1): 25-35.

[6] Shen Weifeng, Miao Qilong, Wei Tiexin, et al. Analysis of temperature variation in recent 130 years in Central Asia[J]. Journal of Arid Meteorology, 2013, 31(1): 32-36.

[7] Lioubimtseva E, Cole R. Uncertainties of climate change in arid environments of Central Asia[J]. Reviews in Fisheries Science, 2006, 14(1-2): 29-49.

[8] IPCC. Climate Change 2013: The Physical Science Basis[M]. Cambridge, UK: Cambridge University Press, 2013: 159-254.

[9] Perkins S E, Alexander L V, Nairn J R. Increasing frequency, intensity and duration of observed global heatwaves and warm spells[J]. Geophysical Research Letters, 2012, 39(20): L20714.

[10] Yu S, Yan Z W, Freychet N, et al. Trends in summer heatwaves in central Asia from 1917 to 2016: Association with large-scale atmospheric circulation patterns[J]. International Journal of Climatology, 2020, 9(1): 115-127.

[11] Alexander L V, Zhang X B, Peterson T C, et al. Global observed changes in daily climate extremes of temperature and precipitation[J]. Journal of Geophysical Research: Atmospheres, 2006, 111(D5): 1042-1063.

[12] Piper S C, Stewart E F. A gridded global data set of daily temperature and precipitation for terrestrial biospheric modeling[J]. Global Biogeochemical Cycles, 1996, 10(4): 757-782.

[13] New M, Lister D, Hulme M, et al. A high resolution data set of surface climate over global land areas[J]. Climate Research, 2002, 21(1): 1-25.

[14] Hijmans R J, Cameron S E, Parra J L, et al. Very high resolution interpolated climate surfaces for global land areas[J]. International Journal of Climatology, 2005, 25(15): 1965-1978.

[15] Kilibarda M, Hengl T, Heuvelink G B M, et al. Spatio-temporal interpolation of daily temperatures for global land areas at 1km resolution[J]. Journal of Geophysical Research: Atmospheres, 2014, 119(5): 2294-2313.

[16] Li J, Heap A D, Potter A, et al. Application of machine learning methods to spatial interpolation of environmental variables[J]. Environmental Modelling & Software, 2011, 26(12): 1647-1659.

[17] Appelhans T, Mwangomo E, Hardy Douglas R, et al. Evaluating machine learning approaches for the interpolation of monthly air temperature at Mt. Kilimanjaro, Tanzania[J]. Spatial Statistics, 2015, 14(5): 91-113.

[18] Fan Binbin, Luo Geping, Zhang Chi, et al. Evaluation of summer precipitation of CFSR, ERA-Interim and MERRA reanalyses in Xinjiang[J]. Geographical Research, 2013, 32(9): 1602-1612.

[19] Dee D P, Uppala S M, Simmons A J, et al. The ERA-Interim reanalysis: configuration and performance of the data assimilation system[J]. Quarterly Journal of the Royal Meteorological Society, 2011, 137(656): 553-597.

[20] Ma Huijuan, Gao Xiaohong, Gu Xiaotian. Random forest classification of landsat 8 imagery for the complex terrain area based on the combination of spectral, topographic and texture information[J]. Journal of Geo-information Science, 2019, 21(3): 359-371.

[21] Wang Yisen, Xia Shutao. A survey of random forests algorithms[J]. Information and Communications Technologies, 2018, 12(1): 49-55.

[22] Wen Xiaole, Zhong Ao, Hu Xiujuan. The classification of urban greening tree species based on feature selection of random forest[J]. Journal of Geo-information Science, 2018, 20(12): 1777-1786.

[23] Cui Dongwen, Jin Bo. Comprehensive evaluation of water ecological civilization based on random forests regression algorithm[J]. Advances in Science and Technology of Water Resources, 2014, 34(5): 56-60+79.

[24] Chen Tao, Zhi Hai, Bian Duo. Investigation on the discrepancy between observed surface temperature and ERA-Interim over the Qinghai-Tibet Plateau and its attribution[J]. Mountain Research, 2019, 37(1): 1-8.

[25] Dong Guanghui, Zhao Liuru, Huang Haibo, et al. Application of hypothesis test in supplier change[J]. Chinese Pharmaceutical Affairs, 2017, 31(10): 1142-1146.

[26] Zhang Ying, Xu Jianhua, Chen Zhongsheng, et al. Spatial and temporal variation of temperature in Central Asia[J]. Journal of Arid Land Resources and Environment, 2016, 30(7): 133-137.

[27] Shen Haojun, You Qinglong, Wang Pengling, et al. Analysis on heat waves variation features in China during 1961–2014[J]. Journal of the Meteorological Sciences, 2018, 38(1): 28-36.

[28] Nie Yu, Han Zhenyu, Han Rongqing, et al. Interannual variation of heat wave frequency persistence over China and the associated atmospheric circulation anomaly[J]. Meteorological Monthly, 2018, 44(2): 294-303.

[29] Jin Hongmei, Yan Pengcheng, Bai Qingshun, et al. Spatial and temporal distribution of extreme high temperature events in Central Asia over the last 70 years[J]. Journal of Arid Meteorology, 2019, 37(4): 550-556.

[30] Hairiguli Namaiti, Yusufujiang Rusuli, Madiniyati Dilixiati, et al. Adaptability analysis of ERA-Interim and GHCN-CAM reanalyzed temperature data in Tianshan Mountains Area, China[J]. Mountain Research, 2019, 37(4): 613-621.

Note: Figure translations are in progress. See original paper for figures.