**Journal**

Climatic Change

**Title**

Changes of diurnal temperature range over East Asia from 1901 to 2018 and its relationship with precipitation

**Supplementary Material**

1. Figure S1
2. CMA-LSAT v1.1 dataset
3. Classification method of urban and rural stations in East Asia (Figure S1)

**Authors**

Xiubao Sun[1 2 3], Chunzai Wang[1 2 3] *, Guoyu Ren[4 5]

[1] *State Key Laboratory of Tropical Oceanography, South China Sea Institute of Oceanology, Chinese Academy of Sciences, Guangzhou, China 510301*

[2] *Southern Marine Science and Engineering Guangdong Laboratory (Guangzgou), Guangzhou, China 511458*

[3] *Innovation Academy of South China Sea Ecology and Environmental Engineering, Chinese Academy of Sciences, Guangzhou, China 510000*

[4] *Department of Atmospheric Science, School of Environmental Studies, China University of Geosciences, Wuhan, China 430074*

[5] *Laboratory for Climate Studies, National Climate Center, China Meteorological Administration, Beijing, China 100081*

*Corresponding author: Dr. Chunzai Wang (E-mail: cwang@scsio.ac.cn)

## 1. Figure S1

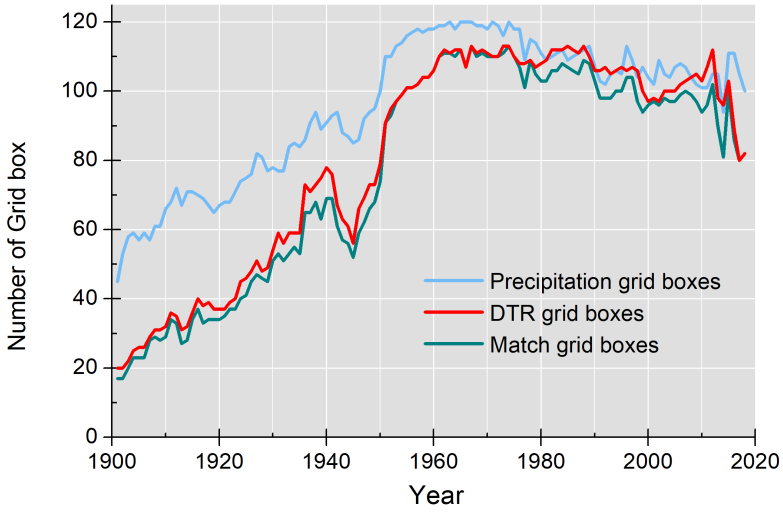


**Figure S1.** Long-term changes in the number of precipitation grid boxes (blue curve), DTR grid boxes (red curve), and match grid boxes (green curve).

## 2. China Meteorological Administration - global land surface air temperature dataset (CMA-LSAT v1.1)

The original data sources of CMA-LSAT data include 3 global datasets (Global Historical Climatology Network-V3 dataset (GHCN-V3), Climatic Research Unit Temperature 4.0 dataset (CRUTEM4.0), and Berkeley Earth Surface Temperature dataset (BEST)); 2 regional datasets (European Climate Assessment & Dataset (ECA&D), and Historical Instrumental Climatological Surface time series of the greater Alpine region database (HISTALP)); and 8 national datasets (China, USA, Russia, Canada, Australia, Korea, Japan, Vietnam) (Sun et al. 2017; Xu et al. 2014). Compared with the other existing global land surface air temperature datasets, CMA-LSAT dataset shows similar ability in describing global temperature changes, despite there are some differences when describing regional temperature changes (Xu et al. 2017). Compared with other global datasets, the obvious improvement of CMA-LSAT data spatial coverage is mainly in South America, Africa, East Asia especially in mainland China and its neighboring countries and regions (Sun et al. 2017).

The CMA-LSAT dataset has been processed by quality-control and homogenization. The method of quality-control was based on that used for GHCN dataset (Menne et al. 2010; Lawrimore et al. 2011), which mainly includes climatological outlier test, spatial inconsistency test, and internal inconsistency test (Sun et al. 2017; Xu et al. 2018). Data from non-homogenized stations in the CMA-LSAT data were also detected for temporal inhomogeneities by applying the RHtest-V3 system developed by Environment Canada (Wang et al. 2007; Wang and Fang 2011). The RHtest-V3 model consists of two parts: the penalized maximal t-test (PMT), and the penalized maximal F test (PMF, Wang and Fang 2011). Firstly, the station with breakpoint is detected as the target station, and then the reference stations of the breakpoint station are established according to the method invented by Peterson et al (1998). By comparing the target series with reference time series of neighboring stations, the background climate signal could be removed from the target station series. Finally, the PMF algorithm in RHtest-V3 system is used to correct the breakpoints in the temperature series of the target station.

The dataset contains 11,036 stations, including maximum, minimum and mean temperature data. There are 2,051 stations in the East Asian region (80°–150°E; 0°–60°N).

## 3. Classification method of urban and rural stations

The station classification model used in this paper is a machine learning method developed by Zhang (2020) and Zhang et al (2020). The United States National Climate Data Center (NCDC) has built the United States climate reference network (USCRN) since 1997, and it can be considered that the observation environment around the USCRN will not be affected by urbanization in the next 50 years (Diamond et al. 2013). The USCRN network is recognized to be able to accurately characterize regional climate change in the United States (Diamond et al. 2013; Ren and Chu 2019). Therefore, we use USCRN data to train machine learning algorithm model. Our goal is to select a global rural station network similar to the USCRN from all global stations. Based on this global rural station network, we can also select the rural station network in East Asia.

The “isolated forest” algorithm in machine learning is used to select rural stations, which is often used to identify outliers or abnormal data from all data (Liu et al. 2012). The first step of

the calculation method is to use USCRN fit model, and then to excavate the land use data around global stations through the model, and finally determine the dense parts (rural stations) and outliers (urban stations) in global stations network (Zhang 2020).

The isolated forest algorithm does not use any method based on distance or density measurement. Instead, it divides the data randomly by recursion, and stops after calculating that all the data are isolated. In this random segmentation strategy, outliers often have short cutting paths (isolated by fewer cuts), while the points with higher density are cut multiple times isolated. It can be divided into three stages:

(1) The first step is to construct the classification model of urban and rural stations. The data excavated by the station classification model is the percentage of land use around the station in 2018 (data available from www.esa-cci.org). The land use data around the station can reflect the intensity of the station affected by urbanization, and the classification of station types based on the land use data has been applied in some studies (Hu et al. 2009). The reason for using only the land use data around the stations in 2018 is to ensure that the selected rural stations can represent the stations that have not been affected by urbanization so far. If the impact of station migration is ignored, it can be considered that the selected rural stations are similar to the USCRN. The disadvantage of the classification method is that the selected rural stations may include stations that have moved from urban areas to rural areas. However, the data used in our study have been processed by homogenization process, which can be considered that the impact of urbanization in relocated stations is removed through the homogenization correction process.

Determining the station buffer range is another key point in the model construction. Taking China as an example, the observation specifications of China Meteorological Administration (2003) have strict restrictions on the building height and surrounding observation environment within 2 km of the national basic station. However, these observational specifications only consider the direct impact of the station surrounding environment on the detection, and do not fully consider the impact of urbanization effect. A large number of studies have found that the influence area of urban heat island effect is not only limited to the local stations, but also can reach the outskirts of city, and it increases with the size of city (Zhou and Shu 1994; Souch and Grimmond 2006). Therefore, it is difficult to determine the

urbanization impact range for a large number of stations, and there has never been a consensus on the impact range of urbanization. The station buffer range we used in this study is 12 km around the station in the current definition method, mainly referring to the station buffer range used in the previous study of Tysa et al. (2019).

Since the station buffer area is determined as 12km, the land use data of 1-12 km in the station buffer zone can be defined as 12 dimensions in the model. Among these 12 dimensions, a cutting point (or a hyperplane) can be generated randomly in a random dimension, and the data space can be cut into two subspaces by cutting points. The data larger than the cutting point and the data smaller than the cutting point are divided into two parts:

$$min(x_{i,j} = q, x_{ij} \in X^{'}) < p < max(x_{i,j} = q, x_{ij} \in X^{'})$$

where $p$ represents a stochastic cutting point, $q$ represents 12 dimensions, $X^{'}$ represents all data in the classification model, $x_{i,j}$ represents a station data in the model. All the data points can be divided into two categories: one is the dense part points which can be isolated after many times of cutting (rural stations), and the other is the outliers which are isolated after few cuts (urban stations).

USCRN stations are used as training data to fit the model. Then, the algorithm iterates until each station on the global land belongs to an independent space. In other words, according to the calculation results of the model, the global land stations can be divided into dense part (rural network similar to the USCRN) and outliers (urban stations).

(2) Verify the reliability of the model and training data, and determine the pollution parameters in the model. The pollution parameters in the model represent the degree of station affected by urbanization, which is a key parameter in the model. 70% of the station data are randomly selected from USCRN station network to train the model, and the remaining 30% data are used as the verification dataset to verify the accuracy of the model results. The pollution parameters were set to 0-0.5 (interval 0.05). The results show that the abnormal ratio of the validation dataset and the training dataset is very close. This also shows that USCRN data training model is reliable.

(3) All the USCRN data are used to fit the model. Then the 1-12 km land use data of global stations are put into the model for calculation. Finally, the global stations are divided into two

categories, dense stations (rural stations) and outlier stations (urban stations). When the pollution parameter is set at 0.3, the selected rural stations are similar to those of USCRN stations. With the increase of pollution parameter, the number of rural stations gradually decreases. Based on the above method, East Asian stations can be divided into 1436 urban stations and 346 rural stations (Figure S2a-b). The urbanization land in the buffer range of 1-12 km around the East Asian rural stations is less than 5%.

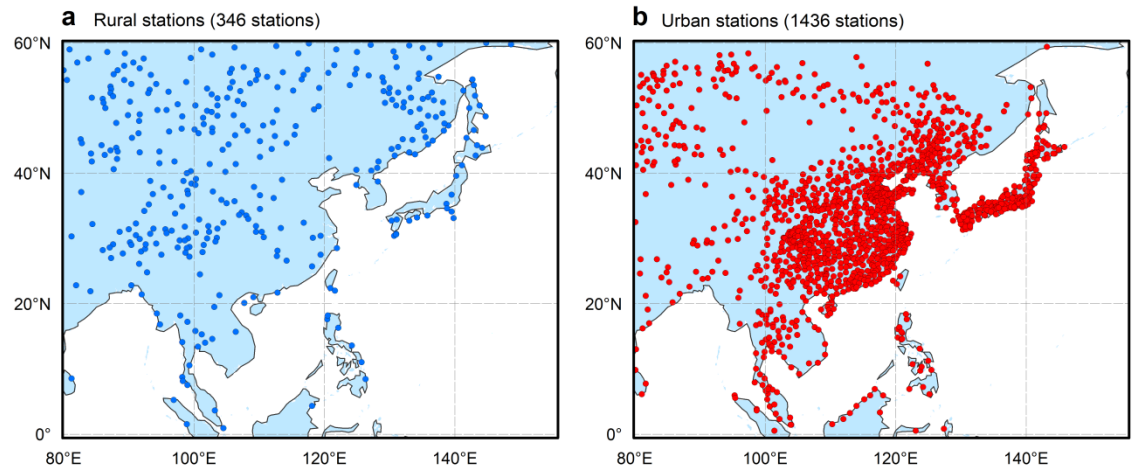


**Figure S2.** Spatial locations of urban (a) and rural (b) stations. The classification method of urban and rural stations is the isolated forest algorithm in machine learning.

## References

China Meteorological Administration (2003) Criterion of surface meteorological observation. China Meteorological Press, Beijing. (in Chinese)

Diamond HJ, Karl TR, Palecki MA, et al (2013) U.S. Climate Reference Network after one decade of operations: status and assessment. Bulletin of the American Meteorological Society 94(4): 485–498

Hu Y, Jia G, Guo H, (2009) Linking primary production, climate and land use along an urban–wildland transect: a satellite view. Environmental Research Letters 4(4):044009

Makowski K, Jaeger EB, Chiacchio M, et al (2009) On the relationship between diurnal temperature range and surface solar radiation in Europe. Journal of Geophysical Research 114: D00D07. doi:10.1029/2008JD011104

Lawrimore JH, Menne MJ, Gleason BE, et al (2011) An overview of the global historical

climatology network monthly mean temperature data set, version 3. Journal of Geophysical Research: Atmospheres 116:5454-5466

Liu FT, Ting KM, Zhou ZH (2012) Isolation-Based Anomaly Detection. ACM Transactions on Knowledge Discovery from Data, 2012, 6(1): 1–39.

Menne MJ, Williams CN, Vose RS, (2010) The U.S. Historical Climatology Network monthly temperature data, version 2. Bulletin of the American Meteorological Society 90:993-1007

Peterson TC, Easterling DR, Karl TR, et al (1998) Homogeneity adjustments of in situ atmospheric climate data: a review. International Journal of Climatology 18:1493-1517

Ren GY, Chu ZY (2019) A brief introduction to the U.S. Climate Reference Network. Advances in Meteorological Science & Technology 9(04): 56–61 (in Chinese)

Souch C, Grimmond S (2006) Applied climatology: urban climate. Progress in Physical Geography 30(2): 270-279

Sun XB, Ren GY, Xu WH, et al (2017) Global land-surface air temperature change based on the new CMA GLSAT data set. Science Bulletin 62(4):236-238

Sun XB, Ren GY, You QL, et al (2019) Global diurnal temperature range (DTR) changes since 1901. Climate Dynamics 52(5):3343-3356

Tysa SK, Ren GY, Qin Y, et al (2019) Urbanization effect in regional temperature series based on a remote sensing classification scheme of stations. Journal of Geophysical Research: Atmospheres 124(20): 10646–10661

Wang XL, Feng Y (2011) RHtestsV3 user manual. Climate Research Division Atmospheric Science and Technology Directorate Science and Technology Branch, Environment Canada Toronto, Ontario, Canada

Wang XL, Wen QH, Wu Y, (2007) Penalized maximal t test for detecting undocumented mean change in climate data series. Journal of Applied Meteorology & Climatology 46:916-931

Xu WH, Li QX, Jones PD, et al (2017) A new integrated and homogenized global monthly land surface air temperature dataset for the period since 1900. Climate Dynamics (15):1-24. doi: 10.1007/s00382-017-3755-1

Xu WH, Li QX, Yang S, et al (2014) Overview of global monthly surface temperature data in

the past century and preliminary integration. Advances in Climate Change Research 5:111-117

Zhang PF (2020) Observed trend changes in extreme temperature over the global land, 1951-2018, China University of Geosciences (Wuhan). (in Chinese)

Zhang PF, Ren GY, Qin Y, et al. (2021) Urbanization effects on estimates of global trends in mean and extreme air temperature. Journal of Climate, 2021, 34(5):1923–1945.

Zhou SZ, Shu J (1994) Urban Climatology. China Meteorological Press, Beijing. (in Chinese)