Hindawi
Advances in Meteorology
Volume 2022, Article ID 9895576, 15 pages
https://doi.org/10.1155/2022/9895576

Research Article

An Integrated Framework for Mapping Nationwide Daily Temperature in China

Shaobo Zhong,[1,2] **Xinlan Ye**,[3] **Mingxing Wang**,[3] **Xin Mei**,[3] **Dunjiang Song**,[4] **and Wenhui Wang**[3]

[1] *Urban Construction School, Beijing City University, Beijing 10083, China*
[2] *Institute of Urban Systems Engineering, Beijing Academy of Science and Technology, Beijing 100035, China*
[3] *Faculty of Resources and Environment Science, Hubei University, Wuhan, China*
[4] *Institute of Science and Development, Chinese Academy of Sciences, Beijing 100190, China*

Correspondence should be addressed to Xin Mei; 55296119@qq.com

Received 11 August 2021; Revised 31 January 2022; Accepted 4 April 2022; Published 14 May 2022

Academic Editor: Stefano Federico

Air temperature (T_a) is an essential parameter for science research and engineering practice. While the traditional site-based approach is only able to obtain observations in limited and discrete locations, satellite remote sensing is promising to retrieve some environmental variables with spatially continuous coverage. Nowadays, land surface temperature (T_s) measurements can be obtained from some satellite sensors (e.g., MODIS), further enabling us to estimate T_a in view of the relationship between T_a and T_s. In this article, we proposed a two-phase integrated framework to estimate daily mean T_a nationwide. In the first phase, multivariate linear regression models were fitted between site-based observations of daily mean air temperature ($T_{a\text{-mean}}$) and MODIS land surface temperature products (including Terra day: $T_{\text{MOD-day}}$, Terra night: $T_{\text{MOD-night}}$, Aqua day: $T_{\text{MYD-day}}$, and Aqua night: $T_{\text{MYD-night}}$) conditional on some covariates of environmental factors. The fitted models were then used to predict $T_{a\text{-mean}}$ from those covariates at unobserved locations. The predicted $T_{a\text{-mean}}$ were looked on as stochastic variables, and their distributions were also obtained. In the second phase, Bayesian maximum entropy (BME) methods were used to produce spatially continuous maps of $T_{a\text{-mean}}$ taking the meteorological station observations as hard data and the predicted $T_{a\text{-mean}}$ in the first phase as soft data. It is shown that the proposed approach is promising to improve the interpolation accuracy significantly, comprehensively considering the prior knowledge and the context of space variability and correlation, which will enable it to compile spatially continuous air temperature products with higher accuracy.

1. Introduction

Air temperature (T_a) is one of the most important variables in scientific research and engineering practice. Short-term effects of air temperature include disease spreading, crop growth, snowmelting, and inundation, while the long-term effects of it on regional and global development such as global warming, drought, extreme weather, and food safety are concerned. In meteorology, air temperature is one of the basic meteorological factors commonly observed at 2 m above the ground in weather observation sites. Therefore, their spatial distribution depends on the sites built for weather data collection [1]. T_a is obtained as point data that cannot directly depict the range of climate variability within a region. Weather observation sites are dense in cities and sparse in sparsely populated and underdeveloped regions. In China, most of weather observation sites are distributed in the east. To densify air temperature observations, the remotely sensed land surface temperature (LST, T_s), retrieved from thermal images, has been used in estimating T_a. However, the major limitations in estimating T_a using remotely sensed T_s are the uncertainties of T_s estimation, nonlinear relationship between T_s and T_a, and trade-off between the temporal and the spatial resolution. In addition, the thermal remote sensing approach is not applicable under cloudy conditions.

To produce high-resolution T_a data, many methods involving in station observations and remotely sensed T_s are proposed. Three types of methods—interpolation, regression analysis, and simulation—were reviewed and have been proven to be useful in mapping high-resolution T_a [2]. Among them, satellite remote sensing T_s-based regression estimation and spatial interpolation of observed T_a are most popular approaches. Wang et al. [3] compared spatial interpolation and regression analysis models for estimates of monthly near-surface T_a from station observations in China [4]. Their results indicated that the higher standard deviation and the lower mean of near-surface T_a from sample data would be associated with a better performance of predicting monthly near-surface T_a using spatial interpolation models. Considering filling in data gaps in the time series of T_a, Shtiliyanova et al. applies a Kriging-based interpolation in the temporal dimension to predict missing T_a data [5] against temperature datasets from five sites in Europe and one site situated in the Indian Ocean (Réunion Island, France overseas). To provide long-term grid historical temperature datasets based on sparse historical stations, a temperature spatial interpolation based on the Biased Sentinel Hospitals Areal Disease Estimation (P-BSHADE) method was proposed, which successfully interpolate 1-km grids of monthly T_s in the historical period of 1900–1950 in China [6]. Considering the characteristics of spatial autocorrelation and nonhomogeneity of the temperature distribution to obtain unbiased and minimum error variance estimates, the proposed method shows better accuracy compared to inverse distance weighting (IDW), spline. Cho et al. proposed a so-called stacking ensemble model consisting of multilinear regression (MLR), support vector regression (SVR), and random forest (RF) optimized by the SVR to interpolate the daily maximum T_a during summertime in Seoul, the capital of South Korea [7]. Some other case studies can also be seen, which have selected either satellite remote sensing-based T_a estimation or spatial interpolation of site T_a to conduct spatially continuous T_a prediction. However, thanks to either of them has limitations and shortcomings, it is necessary to integrate advantages of potential methods to enhance the knowledge of T_a patterns, and to map spatially continuously covered and high-resolution T_a.

It is well known that remote sensing takes some advantages such as continuous space tessellation, wide observation scope, and low cost over other observation methods. In recent decades, several well-known sensors such as AVHRR and MODIS were launched into orbit, which enables us to obtain moderate resolution images with many spectrums (e.g., up to 36 for MODIS) ranging from infrared to microwave. Simultaneously, researchers begin to develop a variety of algorithms for retrieving T_s from these remotely sensed data. In view of the fact that, T_s and T_a have a strong association with each other, remote sensing has been combined with meteorological sites to improve the spatial coverage and accuracy of T_a. Kloog et al. presented work on predicting T_a from T_s in Massachusetts by predicting 24 h T_a means on a 1-km grid across the Northeast and Mid-Atlantic states demonstrating how T_s can be used reliably to predict daily T_a at high resolution in large geographical areas even in nonretrieval days [8]. Alonso et al. proposed to estimate T_a from 28 explanatory variables (covariates), using multiple linear regressions, which integrates variables from remote sensing and the variables traditionally used like the ones from the Land Use Land Cover [9]. Some other researchers proposed statistical models to estimate T_a using MODIS land surface temperature data [10–13]. Furthermore, with a daytime T_a variation model, it is possible to estimate daily T_a at any time. For example, Chen et al. [14] first estimate daytime T_a at the times of Terra and Aqua satellites overpass, and subsequently, the maximum T_a is inferred from the daytime T_a. A general approach to obtaining T_a from images involves two steps: (1) retrieve the land surface temperature at the satellite overpass time and then (2) calculate the T_a from the land surface temperature according to the physical or statistical function relationship between them.

Spatial interpolation is a class of methods that interpolate the values of unobserved locations from the values of observed locations. In a geographical context, spatial auto-association is universally present, which differs spatial problems from nonspatial ones [15]. Therefore, the focus of most of the spatial interpolation methods is devoted to dealing with this nuisance. These methods estimate the most likely values through spatial trend analysis and spatial correlation analysis of discrete observed data. Geostatistics is developed from Kriging interpolation techniques along this thread. The recent advances in geostatistics have been striving to obtain better estimates by blending as much information as possible (including observations and prior knowledge). Christakos [16] extended Kriging techniques into a new methodological framework called BME (Bayesian maximum entropy) [16]. BME takes some types of data and different types of knowledge into spatial interpolation. According to BME methodology, these data and knowledge are divided into general knowledge (GK) and site-specific knowledge (SK). Two types of data are involved in BME interpolation procedure: hard data and soft data. Hard data are generally measured with instruments and considered having definite values. Soft data have indefinite values relative to hard data. Soft data are generally depicted by value interval, probability distribution, etc. The application of BME methods in some fields can be referenced in [3, 17, 18]. It has been also shown that BME is promising in blending multiple sensor data and multiresolution satellite products [19, 20].

In this study, T_s from MODIS sensors and T_a from weather stations were integrated to densify T_a, considering that T_s is spatially continuous and is able to provide additional information for spatial estimation of T_a. The purpose of this study was to estimate T_a with a high spatial resolution and accuracy simultaneously, coupling high spatial resolution remote sensing data with high accuracy (but spatially sparse) station data. We proposed a two-phase approach to estimating T_a at unobserved locations and demonstrated the estimation and evaluation of daily mean T_a ($T_{a\text{-mean}}$) on four days of 2004 (the vernal equinox (VE), the summer solstice (SS), the autumnal equinox (AE), and the winter solstice (WS)).

2. Study Area and Materials

The study area is Mainland China. There is a total of 839 sites distributed over the area. Observations from these sites are taken as hard data. A total of 334 locations are supplemented to obtain soft data through multivariate linear regression estimate techniques. These supplemented locations were generated at spatial random conditional on meteorological site locations (constrained in areas with sparse sites). Figure 1 shows the study area and the distribution of meteorological observation sites and the supplemented locations within it. According to the elevation, the whole study area can be divided into three terrain ranges. The first terrain range is in the western China and has a mean altitude higher than 4000 m. High mountains are main features in this range. The second range is in the central and northern China and has a mean altitude from 1000 m to 2000 m. It consists mainly of plateaus, large basins, and some high mountains. And the third range is in the eastern China and has mean altitude from 200 m to 1000 m. It is mainly composed of plains, hills, and low mountains. We also divided the whole study area into the west part and the east part in terms of the Hu Huanyong line in the name of Chinese scholar Hu Huanyong, which is going to be used to examine the effects of areas with different geographical characteristics on the T_a estimation.

The T_a observations from meteorological sites and the T_s products derived from MODIS in the 80th day (VE), the 172th day (SS), the 266th day (AE), and the 356th day (WS) of 2014 were used. These are four representative days in a whole year. Daily T_a variables are collected at those sites from China Meteorological Service Center. $T_{a\text{-mean}}$ was calculated based on a published standard by CMA (China Meteorological Administration) from the original weather site observations, which is the arithmetic mean of the T_a observations at four time points: 02 : 00, 08 : 00, 14 : 00, and 20 : 00 a day.

Two MODIS products—MOD11C1 and MYD11C1—were used, which are global T_s data retrieved from MODIS sensors aboard Terra and Aqua satellites. They are estimated at time when satellites pass according to an algorithm developed by NASA. The nominal equatorial passing time of Terra is around 10:30 am and 10:30 pm of local solar time, while Aqua passes in the opposite direction at about 1:30 am and 1:30 pm. Therefore, Terra MOD11C1 and Aqua MYD11C1 include both day and night T_s (Terra day: $T_{\text{MOD-day}}$, Terra night: $T_{\text{MOD-night}}$, Aqua day: $T_{\text{MYD-day}}$, and Aqua night: $T_{\text{MYD-night}}$). These two products cover the main continents on the Earth surface, and the spatial resolution is 0.05 degrees. Therefore, the real ground extent is 5 km × 5 km or so near the equator. We downloaded the T_s data from a remotely sensed data distribution website created by the United States National Aeronautics and Space Administration (NASA) (https://disc.gsfc.nasa.gov/). The global products were trimmed to the north latitude range (39.4°, 41.1°) and east longitude range (115.3°, 117.6°), which is the bounding box of the study area. According to the quality control information of the products, we excluded the missing data and the data whose nominal errors are greater than 3K. As a result, we got some data with acceptable quality (nominal errors ≤3K). The analysis will be carried out based on these remaining data.

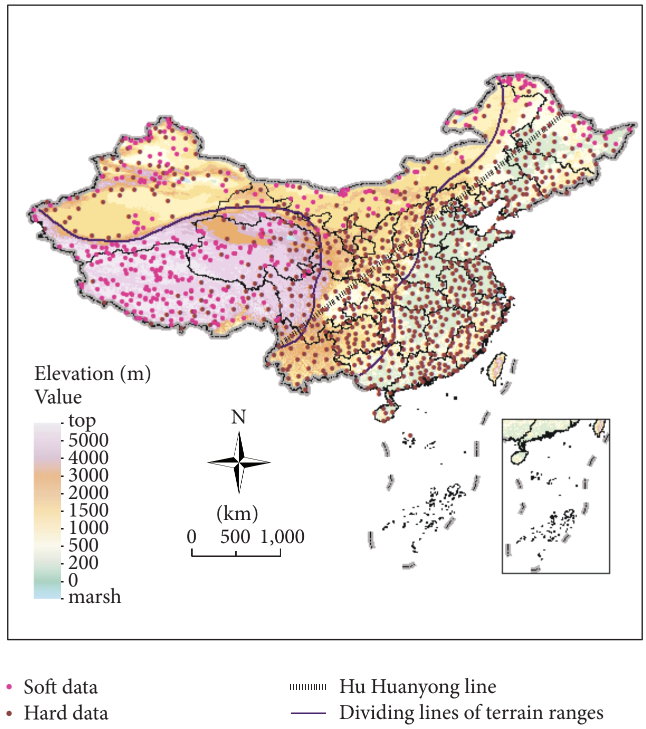


Figure 1: Map of the study area: provincial administration regions of Mainland China and the geographical locations of all 839 meteorological observation sites (brown dots) within it. These sites provide T_a observations called hard data. The study area was divided into two parts: the west part and the east part with a dividing line (called Hu Huanyong line in the name of Chinese scholar Hu Huanyong). According to the elevation, three terrain ranges are demarcated (brown lines). The pink dots are supplemented locations to calculate soft data (totally 334 locations).

We also acquired a DEM data of China from the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) Global Digital Elevation Model (GDEM), which was developed jointly by the Ministry of Economy, Trade and Industry (METI) of Japan and NASA under the agreement of contribution to GEOSS (Global Earth Observation System of Systems), and a public release was started on June 29, 2009 [21]. Version 1 of the DEM data is produced in 2009 and reproduced and improved in 2011 (version 2). We downloaded version 2 of the DEM data. This dataset is provided in raster format, and its spatial resolution is 30 m. The absolute vertical accuracy is within 0.20 meters on average.

Two vegetation index data provided by MODIS were used. One is the classical NDVI (normalized difference vegetation index) product and the other is EVI (enhanced vegetation index) product. The latter is more sensitive to areas with high vegetation coverage than the former. The MOD13C1 product was selected to extract these two indices. Since the MOD13C1 product is a 16-day composite one. Corresponding to the four days when $T_{a\text{-mean}}$ was estimated in this study, the MOD13C1 data on the 65th day, the 161th day, the 257th day, and the 353th day of 2014 were selected and downloaded.

A LUCC (Land Use/Cover Change) dataset from the Chinese Academy of Sciences Resources and Environment Science Data Center (https://www.resdc.cn/) was selected to extract the land use/cover. The dataset covers 1990 to 2015, and the time interval of production is 5 years. Its spatial resolution is 1 km × 1 km. We selected the LUCC data in 2015 that is closest to 2014. The format of the data is raster format. In this dataset, the land use/cover types of level 1 include forest, grassland, wetland, farmland, and artificial surface.

3. Methods

The proposed methodology includes two phases. The first phase regressed the T_a observations from weather sites on T_s from MODIS conditional on some environmental factors such as topological, meteorological, vegetative, and LUCC as covariates. In this phase, multivariate linear regression models are fitted. Once proper models are fitted, we can predict the T_a at a specific location given covariates. According to the statistical theory of multivariate linear regression analysis, the predicted T_a has an uncertainty, and we can also estimate its confidence interval. Then, in the second phase, we make an ensemble estimation of T_a by integrating weather site observations and the prediction of T_a from the regression analysis in the first phase.

3.1. Multivariate Linear Estimates. The multivariate linear estimation is formulated as

$$Y = \mathbf{x}\widehat{\boldsymbol{\beta}} + e, \tag{1}$$

where Y is the dependent variable, $\mathbf{x}$ is the vector of the prediction variables, $\widehat{\boldsymbol{\beta}}$ is the parameter vector of the multivariate linear regression model, and e is the residual error. Thus, given a group of samples X_0 (land surface temperatures and environmental variables), we can estimate the corresponding T_a as

$$\widehat{\mathbf{Y}}_0 = \mathbf{X}_0\widehat{\boldsymbol{\beta}}. \tag{2}$$

For a linear regression, the estimate of variance is

$$\widehat{\sigma}_{e_0}^2 = \widehat{\sigma}^2\left(1 + \mathbf{X}_0\left(\mathbf{x}'\mathbf{x}\right)^{-1}\mathbf{X}_0'\right), \tag{3}$$

where $\widehat{\sigma}$ is the estimated standard error, which is given as

$$\widehat{\sigma}^2 = \frac{\sum e_i^2}{n-k-1}, \tag{4}$$

where n is the number of samples, and k is the number of covariates.

We further construct a statistic:

$$t = \frac{\widehat{Y}_0 - Y_0}{\widehat{\sigma}_{e_0}} \sim t\,(n-k-1). \tag{5}$$

In probability theory, this statistic is *t-distributed* with a degree of freedom n–k–1. Therefore, we can calculate the confidence interval of Y_0:

$$\widehat{Y}_0 - t_{\alpha/2} \times \sigma_{e_0} < Y_0 < \widehat{Y}_0 + t_{\alpha/2} \times \sigma_{e_0}. \tag{6}$$

When $n > 35$, the t distribution can be approximated with the normal distribution $N\,(\widehat{Y}_0, \widehat{\sigma}_{e_0}^2)$.

In this study, three groups of variables are taken as covariates: T_s from Terra and Aqua, including $T_{\text{MOD\_Day}}$, $T_{\text{MOD\_Night}}$, $T_{\text{MYD\_Day}}$, and $T_{\text{MYD\_Night}}$; vegetation data including EVI and NDVI; and topological data including latitude, altitude, and longitude. All variables used in the regression analysis are listed in Table 1. The dependent variable $T_{a\text{-mean}}$ is regressed on the covariates. The first-level classification of LUCC was represented by setting up a nominal covariate LUCC. To carry out the regression analysis, we expanded it into 7 dummy variables according to 6 types of the first-level classification.

Taking weather sites as reference locations, we extracted the values of covariates from corresponding data sources such as MOD11C1, MYD11C1, MOD13C1, LUCC, and DEM using GIS Extraction tools (Extract Values to Points in ArcMap). For a certain weather site, if there exist missing values or outliers in either of the dependent variable and covariates, the sample of this site was excluded.

3.2. BME Interpolation. We selected the BME method as the interpolation estimator of the T_a, which is firstly established by Christakos [16, 22]. BME uses many types of data for spatial estimation. These data are classified into two types: GK and SK, whose specific definitions and meaning in meteorological data can be identified in some research examples [3, 17, 23]. In this study, we intend to estimate the values of a meteorological variable $X(s)$ (a spatial random field) at unmeasured locations s_k given acquired data: $\chi = [x_1, x_2, \ldots, x_m]'$, where χ represents a set of meteorological data x_i at spatial locations $s_i\,(i = 1, \ldots, m;\, s_i \neq s_k)$. These datasets can be divided into two main groups: hard data and soft data. The former is observations of T_a from meteorological stations, and the latter is T_a estimation obtained from the multivariate linear regression analysis. According to the regression modeling, the soft data are expressed in normal distribution (an approximation of t distribution) with the corresponding estimated T_a and their variances as distribution parameters. We let the vector $\chi_{\text{hard}} = [x_1, x_2, \ldots, x_{m_h}]'$ denote the hard data and $\chi_{\text{soft}} = [x_{m_h+1}, x_{m_h+2}, \ldots, x_m]'$ is the soft data. Furthermore, let $\chi_{\text{map}} = [x_1, x_2, \ldots, x_m, x_k]'$ denote the random vector including the hard data, soft data, and unknown value x_k. There are three stages to finish BME estimation.

3.2.1. Prior Stage. The aim of this stage was to determine the prior probability density function (pdf) $f_G\,(\chi_{\text{map}})$ based on GK, called prior pdf. Prior knowledge of GK is taken as statistics constraint conditions derived from χ_{map}, which is expressed mathematically as

$$\overline{g}_\alpha = \int d\chi_{\text{map}} f_G\left(\chi_{\text{map}}\right) g_\alpha\left(\chi_{\text{map}}\right), \tag{7}$$

where $\alpha = 0, 1, \ldots, N_c$, N_c is the number of constraints, $g_\alpha\,(\chi_{\text{map}})$ are known functions, and the case $\alpha = 0$, $g_0 = 1$ is a

Table 1: Dependent variable and covariates considered in the multivariate linear regression analysis.

Variable group	Variable category (n)	Units
Dependent variable	Daily T_a (1): daily mean T_a ($T_{a\text{-mean}}$)	Degree Celsius
Covariates	T_s (4): $T_{\text{MOD\_day}}$ (o_d), $T_{\text{MOD\_night}}$ (o_n), $T_{\text{MYD\_day}}$ (y_d), $T_{\text{MYD\_night}}$ (y_n)	Degree Celsius
	Vegetation index (2): NDVI (ndvi), EVI (evi)	n/a
	Land cover type (1): LUCC (lucc)	Nominal
	Terrain (3): longitude (lon), latitude (lat), altitude (ele)	Decimal degree, decimal degree, meter

Table 2: Best models with $T_{a\text{-mean}}$ regressed on different sets of covariates and their specifications for the selected four days.

Day	Model	R^2	S.E.	Intercept	Lat	Lon	Ele	ndvi	evi	y_d	y_n	o_d	o_n
VE	1	0.958	2.755	7.276							0.812		
	2	0.967	2.449	4.009							0.699	0.201	
	3	0.970	2.341	3.868							0.426	0.184	0.294
	4	0.971	2.304	2.878					6.614		0.402	0.186	0.275
	5	0.972	2.282	6.287		−0.030			7.124		0.408	0.170	0.281
SS	1	0.891	2.622	12.035							0.689		
	2	0.904	2.468	10.964							0.438		0.288
	3	0.911	2.387	14.143			−0.001				0.370		0.223
	4	0.922	2.242	21.342	−0.134		−0.002				0.262		0.211
	5	0.935	2.060	21.386	−0.201		−0.002			0.149	0.214		0.131
	6	0.937	2.030	26.713	−0.195	−0.041	−0.002			0.124	0.218		0.126
AE	1	0.939	2.625	6.798									0.856
	2	0.946	2.460	7.338							0.364		0.491
	3	0.952	2.328	4.274						0.156	0.338		0.433
	4	0.956	2.227	2.131					6.808	0.197	0.316		0.380
	5	0.959	2.166	5.496	−0.089				5.919	0.216	0.286		0.363
	6	0.960	2.143	7.176	−0.119		−0.001		6.323	0.232	0.246		0.340
	7	0.960	2.132	10.446	−0.116	−0.027	−0.001		6.641	0.217	0.242		0.347
	8	0.961	2.125	10.223	−0.118	−0.027	−0.001		6.853	0.171	0.231	0.065	0.344
WS	1	0.958	2.972	4.400									0.891
	2	0.971	2.495	1.271								0.342	0.617
	3	0.973	2.390	1.702							0.338	0.342	0.287
	4	0.975	2.331	7.975	−0.184						0.308	0.263	0.274
	5	0.975	2.312	9.985	−0.248		0.000				0.287	0.283	0.228
	6	0.976	2.288	16.134	−0.294	−0.040	−0.001				0.266	0.270	0.224

normalization constraint. In this article, other constraints include means and covariances of χ_{map} and probability of soft data. The corresponding forms of $g_\alpha(\chi_{\text{map}})$ can be referred to [22].

An entropy function of $f_G(\chi_{\text{map}})$ is defined as

$$\text{Inf}\left(\chi_{\text{map}}\right) = -\int d\chi_{\text{map}} f_G\left(\chi_{\text{map}}\right) \log f_G\left(\chi_{\text{map}}\right). \quad (8)$$

Thus, prior pdf may be derived by means of a procedure that maximizes the entropy function and takes into consideration the constraints of equation (7), which represent prior knowledge.

3.2.2. Preposterior Stage. The objective of this stage is to collect and organize additional auxiliary information in appropriate forms to produce SK. These are then used in the BME model. Hard data were incorporated in the prior stage indirectly and used directly in the preposterior stage. Soft data were generated according to the responding Gaussian distributions.

3.2.3. Posterior Stage. The aim of this stage is to update the prior pdf based on the Bayesian conditional probability theorem and SK, thereby attaining the posterior pdf. When the distribution of hard and soft data is certain, the posterior pdf $f_K(x_k)$ of spatial variable x_k at location s_k is

$$f_K(x_k) = f_G\left(x_k|\chi_{\text{hard}}, \chi_{\text{soft}}\right) = \frac{f_G\left(x_k, \chi_{\text{hard}}, \chi_{\text{soft}}\right)}{f_G\left(\chi_{\text{hard}}, \chi_{\text{soft}}\right)}. \quad (9)$$

In practice, only hard and soft data within the scope of maximum distance $d_{\max}$ to the estimation point are used to calculate x_k. If the soft data are in the form of a pdf, then

$$f_K(x_k) = \frac{\int_\alpha^\beta f_G\left(x_k, \chi_{\text{hard}}, \chi_{\text{soft}}\right) f_S\left(\chi_{\text{soft}}\right) \mathrm{d}\chi_{\text{soft}}}{\int_\alpha^\beta f_G\left(\chi_{\text{hard}}, \chi_{\text{soft}}\right) f_S\left(\chi_{\text{soft}}\right) \mathrm{d}\chi_{\text{soft}}}. \quad (10)$$

where $f_S(\chi_{\text{soft}})$ is the pdf of soft data.

3.3. Prediction Accuracy Evaluation. We used the leave-one-out cross-validation to evaluate the goodness of the results. Each time, one out of m_h observations were left as the validation point to calculate the error between the estimated value and the observed value, and the other m_h−1 points were used to build the model and predict the value at the

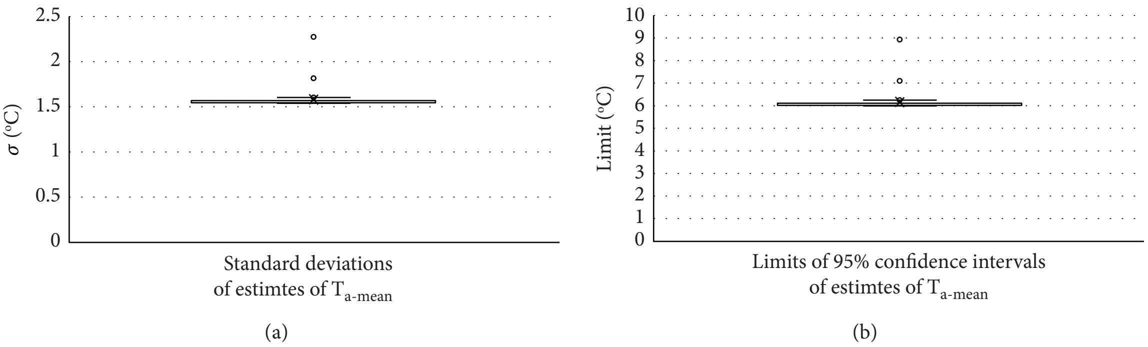


Figure 2: Boxplots of (a) the standard deviations (σ) and (b) the limits of the 95% confidence intervals of the estimated values of $T_{a\text{-mean}}$ with the applicable regression models in the 334 supplemented locations.

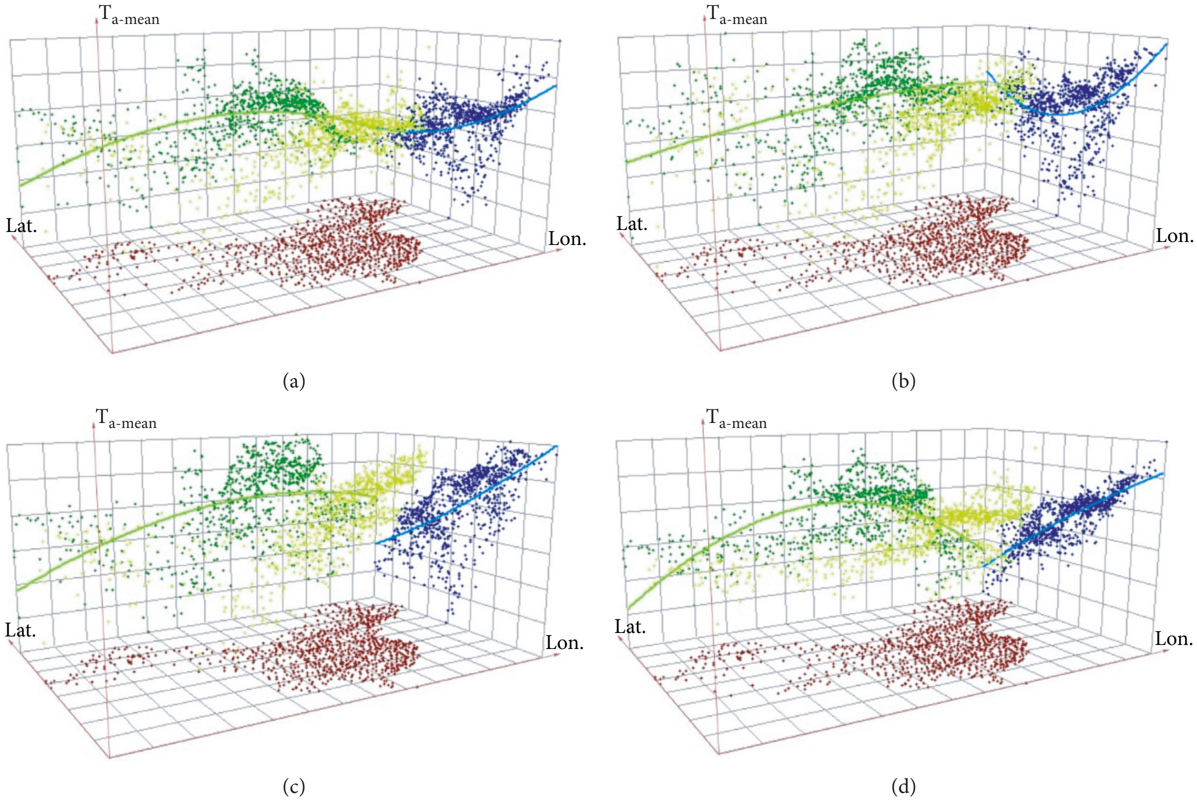


Figure 3: Three-dimensional scatterplots of $T_{a\text{-mean}}$. Longitude is the X axis, latitude is the Y axis, and $T_{a\text{-mean}}$ is the Z axis. The observations of $T_{a\text{-mean}}$ of (a) VE, (b) SS, (c) AE, and (d) WS were projected onto the XZ (green dots) and YZ (blue dots) planes to analyze the trends of them. According to the 3D scatterplots and projections, a polynomial with order of 2 was used to fit the trend surfaces of the observations.

leave-one-out location. Based on the prediction errors, two indices are used to evaluate the interpolation accuracy, and one is mean absolute error (MAE), defined as

$$\text{MAE} = \frac{\sum_{i=1}^{m_h} \text{abs}\left(x_i - \hat{x}_i\right)}{m_h}. \qquad (11)$$

And the other is root mean square error (RMSE), which is defined as

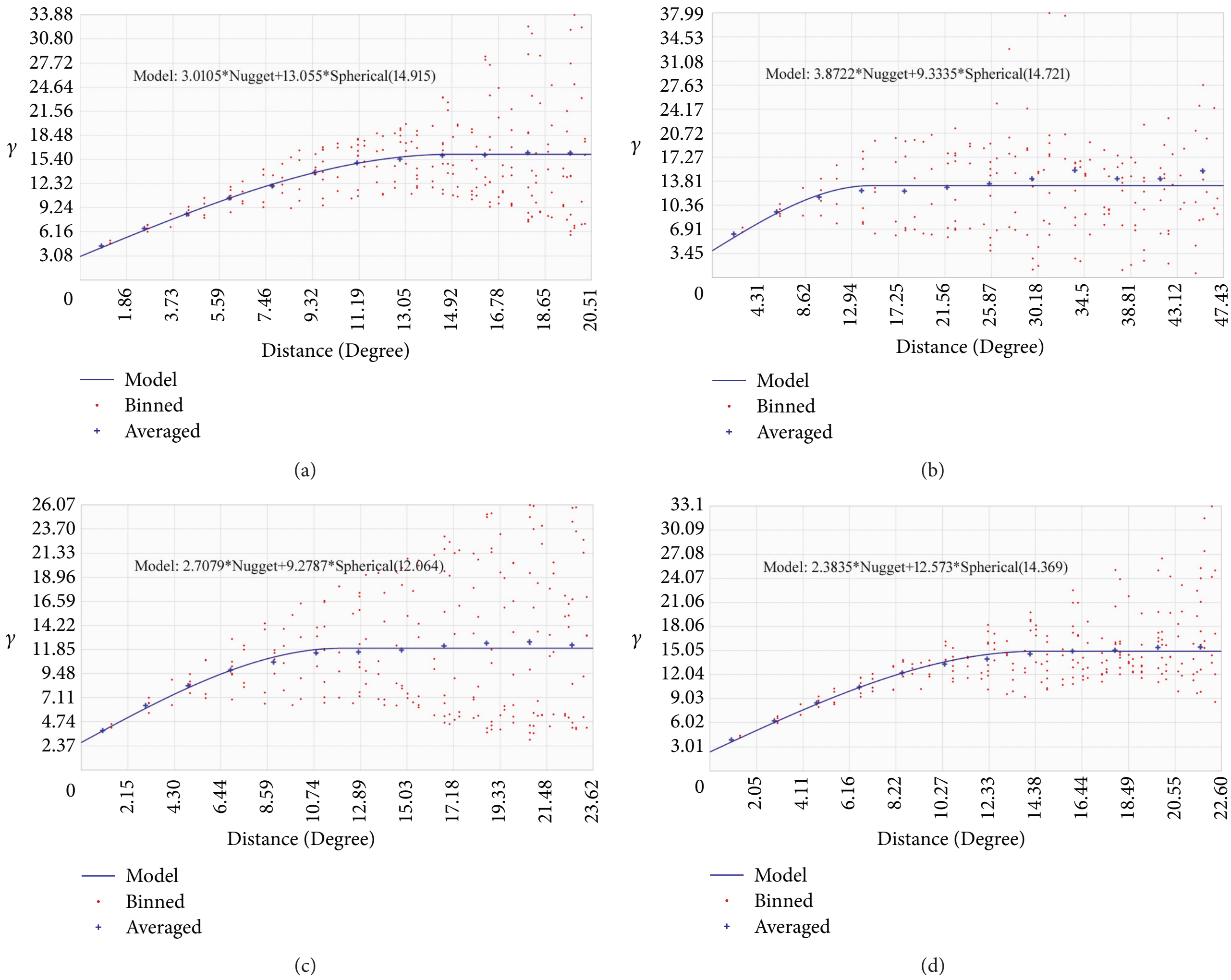
$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{m_h} \left(x_i - \hat{x}_i\right)^2}{m_h}}, \qquad (12)$$

where $\hat{x}_i$ represents the estimated T_a value by the spatial interpolation at the spatial location s_i.

MAE is mainly used to evaluate the upper and lower limits of errors, while RMSE is better at evaluating the sensitivity of spatial interpolation results and the maximal minimum effect of some points.

4. Results

4.1. Regression Analysis and Prediction. We made regression analysis and prediction for each of the four selected days. With a stepwise regression procedure, $T_{a\text{-mean}}$ was



(a)

(b)

(c)

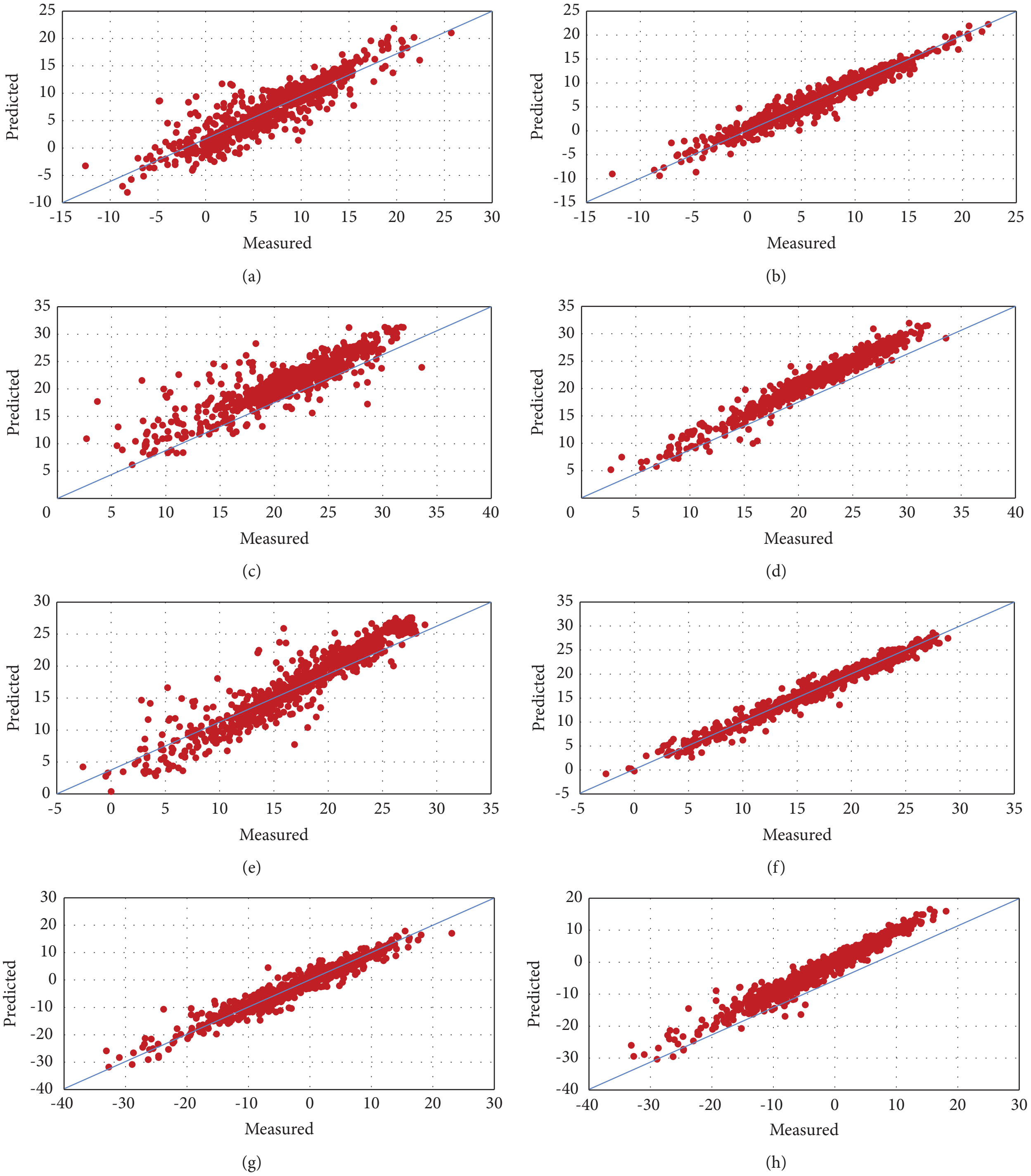
(d)

Figure 4: Observation binning and experimental semivariograms, and the fitted semivariogram models of residuals after detrending for (a) VE, (b) SS, (c) AE, and (d) WS. γ (Distance) is the semivariogram function, where Distance is the distance of pairs of locations and Nugget is called the nugget.

regressed on those selected covariates. At each step, a temporary best model (with maximal adjusted R-square) was chosen given the certain covariates. When adding any new covariate always leads to a decrease in the adjusted R-square, the final best model was obtained. Table 2 shows the best fitted models. Among all these models, Model 1 of VE has the smallest R-square value of 0.891, and Model 6 of WS has the biggest R-square value of 0.976. Furthermore, the F test shows the p-values of these models are all ≤ 0.001, indicating these models are statistically significant. Also, the t test for regression coefficients shows all these models have coefficients with significance levels <0.01. Thus, these models were used to predict the corresponding T_a at the supplemented locations. Table 2 shows the best models with $T_{a\text{-mean}}$ regressed on different sets of covariates and their specifications for the selected four days.

According to the fitted models, the prediction values at the supplemented locations (pink dots in Figure 1) and their corresponding confidence intervals were calculated taking the extracted values of covariates. Perhaps not all the values of the covariates are able to be extracted due to potential missing data and outliers. In this case, we turned to the model whose covariates are available and R^2 is largest (we call it the applicable model) to calculate the estimated T_a and its corresponding confidence intervals. Figure 2 shows the boxplots of the standard deviations (σ) of the estimated values of $T_{a\text{-mean}}$. From Figure 2, we can evaluate the regression accuracy; for example, the regression estimates of $T_{a\text{-mean}}$ indicate that most of estimates have an estimation error with σ approximating 1.6 degrees Celsius. The boxplots on the right show the limits of the confidence intervals of the estimated values of $T_{a\text{-mean}}$, showing most of estimates have limits of about 6 degrees Celsius (95% confidence interval). These estimated values and their corresponding confidence intervals were taken as soft data in the coming BME interpolation.

According to the estimate u and the standard deviation σ at a certain supplemented location, a normal



(a)

(b)

(c)

(d)

(e)

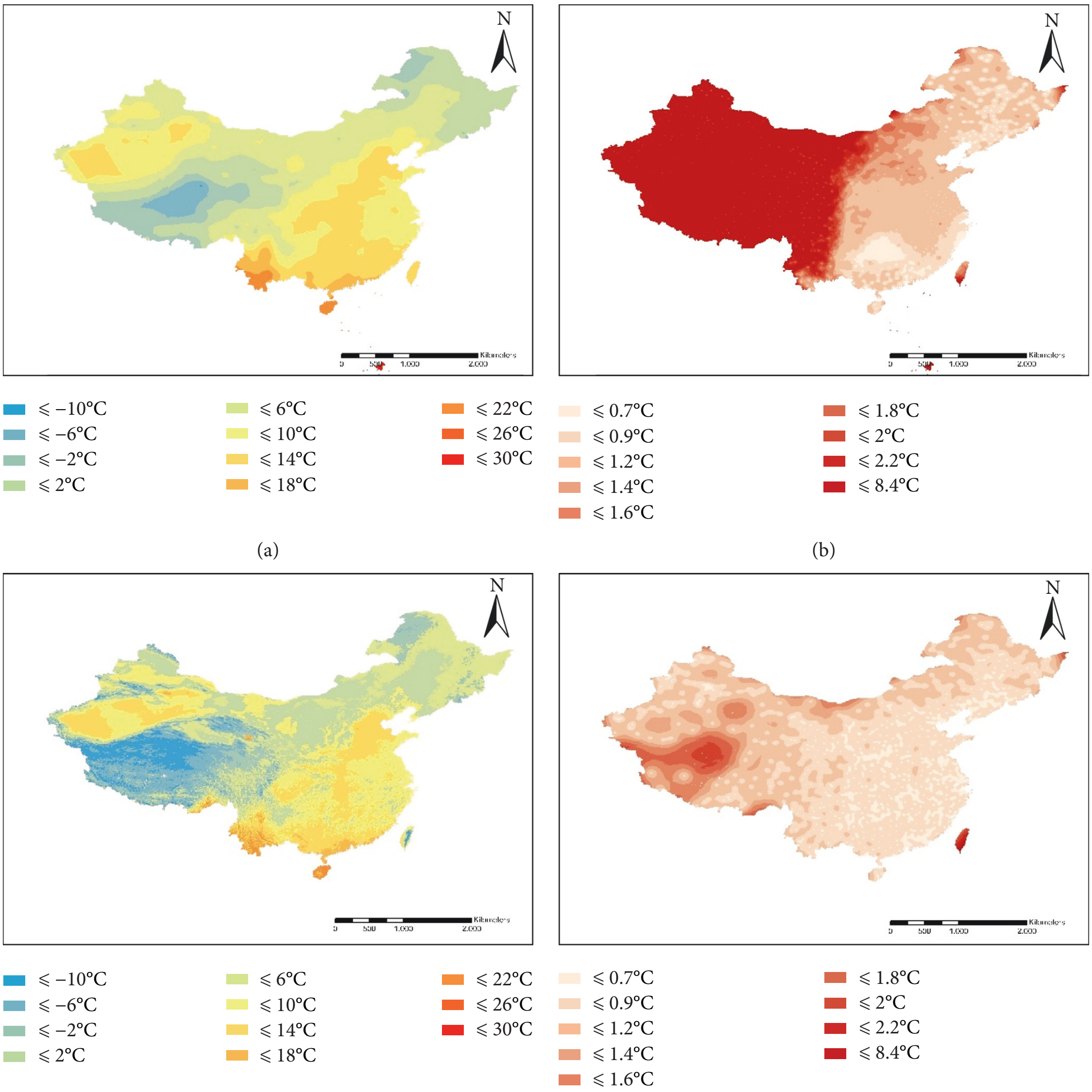
(f)

(g)

(h)

Figure 5: Scatter plots of measured and predicted values for the four days and the two methods, respectively. Two graphs in a row correspond to a certain day, and the graph on the left is plotted according to the results of the BME-hard method, while the graph on the right is plotted according to the results of the BME-both method.

distribution (Gaussian distribution) $N\ (u,\ \sigma^2)$ is used to approximate the distribution of the variable. A graphical BME tool, the SEKS-GUI software library, was used to carry out the interpolation operation, which accepts soft data in the form of Gaussian, uniform, or triangular distributions [24, 25]. An Excel file was created to feature the soft data of the supplemented locations. In the case of Gaussian distribution, for a certain supplemented



(a)

(b)

(c)

(d)

FIGURE 6: VE interpolation maps of $T_{a\text{-mean}}$ with (a) BME-hard and (c) BME-both. (b, d) The corresponding prediction standard error maps.

location s (x, y), its soft data are portrayed in consecutive cells in the same row: $x \quad y \quad u \quad \sigma^2$.

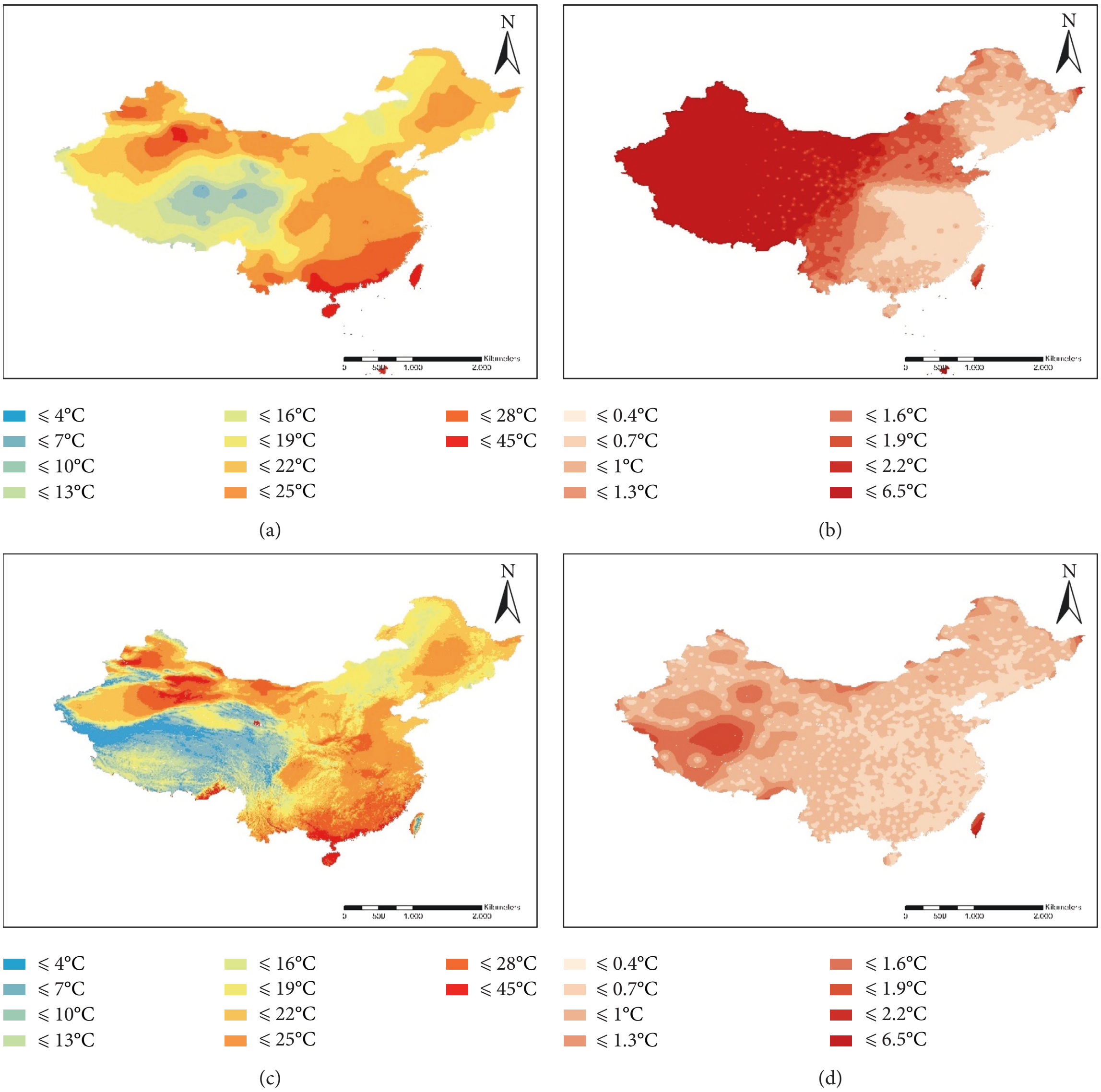
4.2. Spatial Interpolation and Evaluation. Taking longitude as X axis, latitude as Y axis, and $T_{a\text{-mean}}$ as Z axis, in a XYZ coordinate system, the 3D scatterplots of $T_{a\text{-mean}}$ and their projections on XZ and YZ planes are plotted as shown in Figure 3. From the spatial distribution of the T_a observations and the characteristics of the projections on XZ and YZ planes, quadratic polynomial was used in the coming interpolation operation of BME.

After removing the global trends from the observations of T_a, we analyzed the semivariograms of the residuals. With a technique called binning, the experimental semivariograms of the residuals were plotted. Through observing the experimental semivariograms, we selected the spherical model to fit the semivariograms. Isotropy was applied for the fitting of the experimental semivariograms. Figure 4 shows the fitted results. The fitted models are also given in each subgraph.

After identifying the semivariograms, two methods—BME with only hard data (BME-hard) and BME with both hard and soft data (BME-both)—were applied to predict $T_{a\text{-mean}}$.

The scatter plots of measured and predicted values are shown in Figure 5. Following the scatter plots, the BME-both achieved higher accuracy than BME-hard. Generally, scatter plots of SS and AE are more clustered and better fitted with the line $y = x$, indicating their predictions are more accurate and stabler than SEs and WSs.

Spatially continuous distribution maps of $T_{a\text{-mean}}$ in the study area were produced by setting the output grid size to 0.1degree*0.1degree, as well as of the corresponding



(a) (b) (c) (d)

FIGURE 7: SS interpolation maps of $T_{a\text{-mean}}$ with (a) BME-hard and (c) BME-both. (b, d) The corresponding prediction standard error maps.

prediction standard errors as shown in Figures 6–9 for VE, SS, AE, and WS, respectively. Each figure shows two pairs of maps. Each pair of maps consists of an estimation map and a prediction standard error map, which are produced by BME-hard (maps in the upper part) and BME-both (maps in the lower part). From these maps, we see the BME-both-produced maps are more accurate than those produced by BME-hard overall. Comparing the prediction standard error maps, we found BME-both-produced maps have significantly smaller and stabler prediction standard error than BME-hard-produced maps, which means the estimation from BME-both has less uncertainty. Furthermore, interpolation errors are closely associated with the density of the site distribution. The first terrain range has the largest errors, where meteorological sites are very sparse because of high altitude and sparse population. Furthermore, according to the results of the four selected days, interpolation errors are probably dependent of days or seasons. The prediction standard error maps show that interpolation results in SS and AE have higher accuracy than those in SE and WS. WS has the highest interpolation errors among the four selected days.

To explore the effects of different geographic characteristics on the interpolation accuracy, the prediction accuracy in six areas—the whole study area, the west and east of Hu Huanyong line (Hu W. and Hu E.), the high terrain range (Terrain 1), the medium terrain range (Terrain 2), and the low terrain range (Terrain 3)—were evaluated with BME-both and BME-hard. Table 3 shows the calculated RMSE and MAE for each area and method. Our two-phase method, BME-both, achieved significant improvement on nationwide T_a prediction. Quantitative calculation further

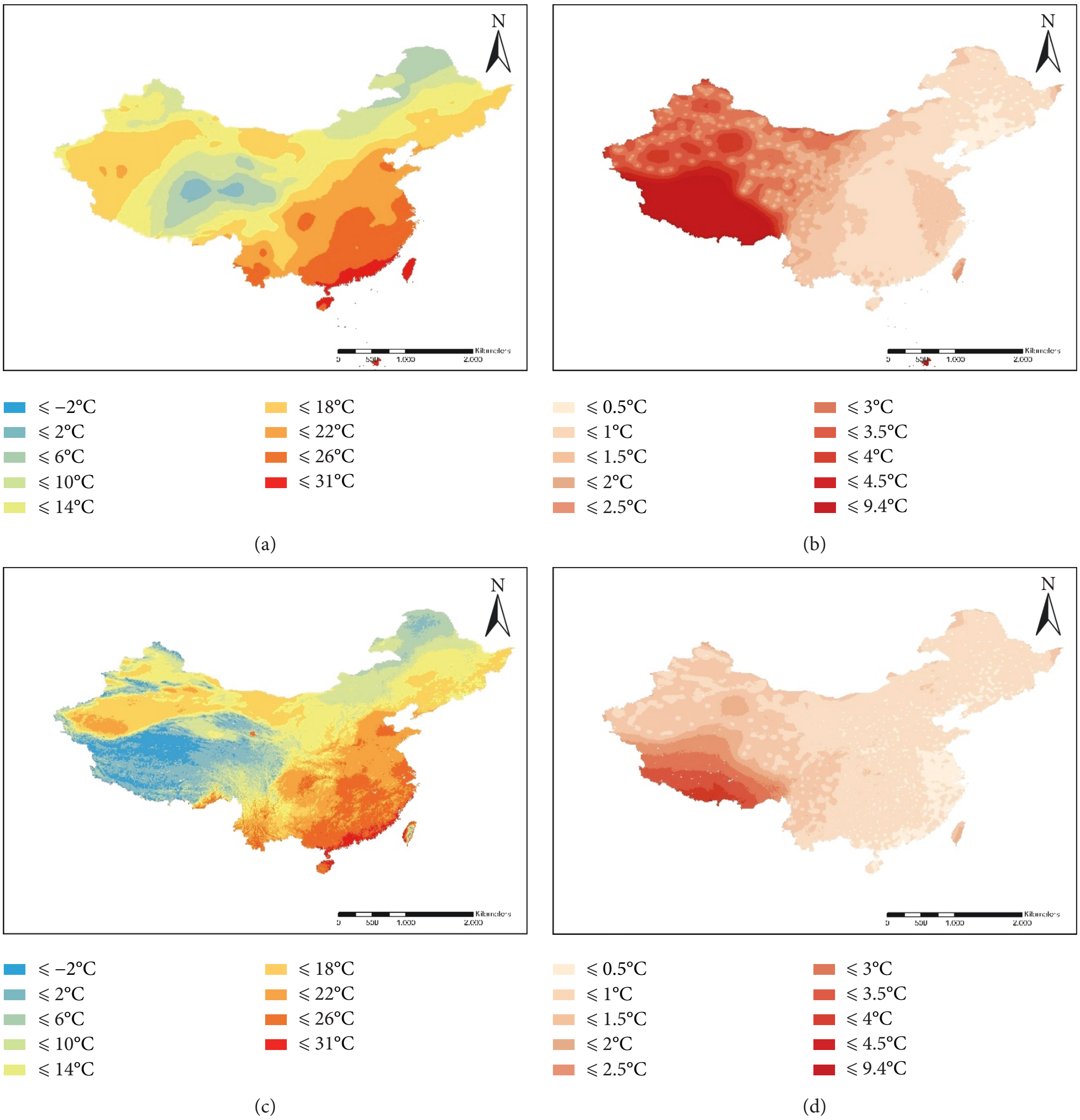
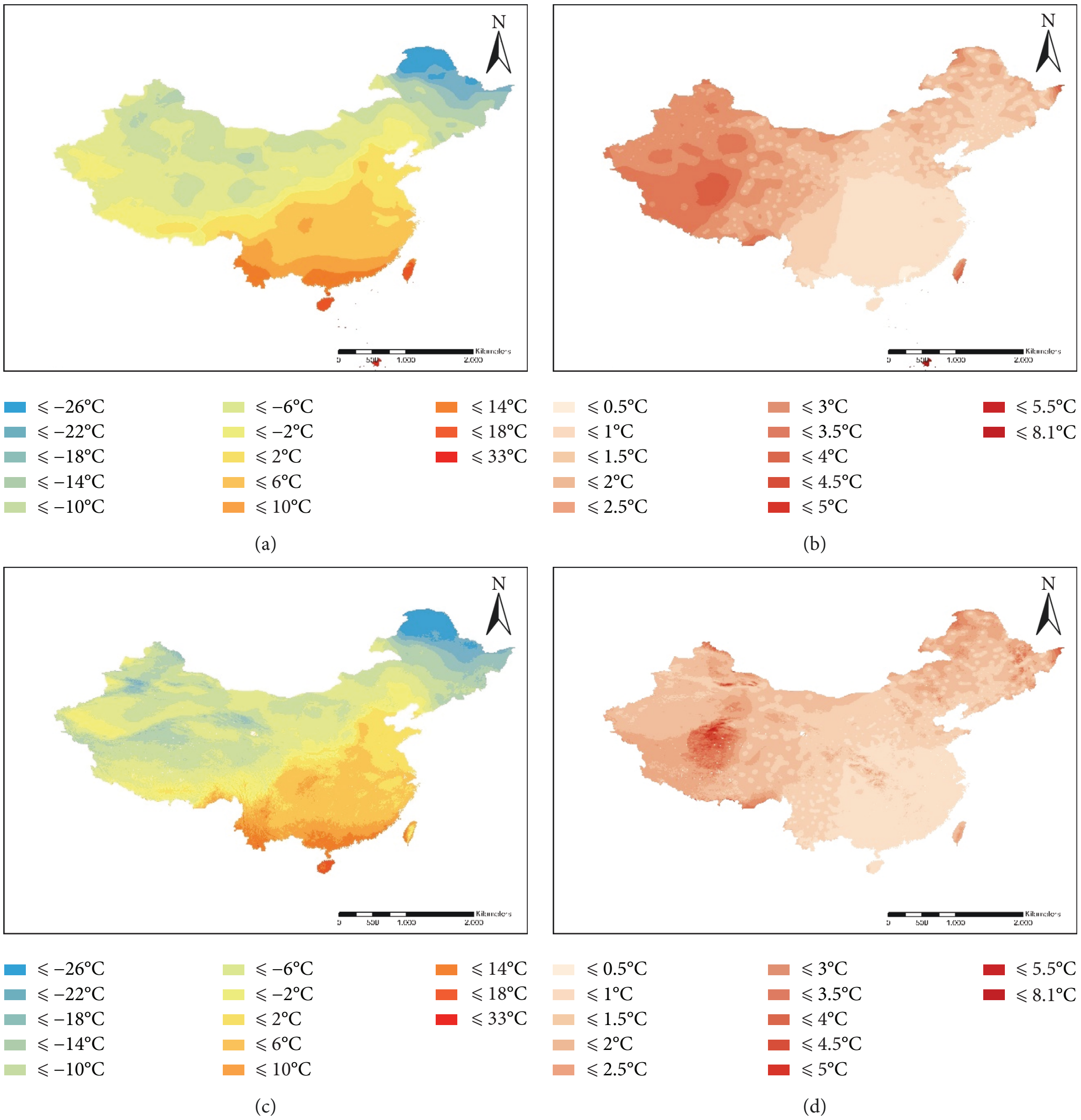


Figure 8: AE interpolation maps of $T_{a\text{-mean}}$ with (a) BME-hard and (c) BME-both. (b, d) The corresponding prediction standard error maps.

confirmed our judgement that interpolation accuracy is associated with some geographic factors such as site density and altitude. In areas with low altitude and dense population, the interpolation accuracy is better than that in other areas.

5. Discussion

Soft data are one of the two kinds of input data in BME interpolation, which differ from other interpolation methods. The general framework of BME explains some approaches to the acquisition and representation of soft data. Transforming prior knowledge into probability distribution is the most used technique in practice. It is well known that regression analysis can predict a response variable from some covariates in terms of a function relationship. Furthermore, as a random variable, the predicted variable (dependent variable) obeys a certain probability distribution. Taking advantage of this characteristic, in this article, we employed a multivariate linear regression method to retrieve the distribution of T_a at some unobserved locations and used Gaussian distribution to approximate the t distributions of the predicted variable of the regression model. Thus, additional information is incorporated into the interpolation process as prior knowledge of a probabilistic distribution. The results indicate that significant improvement is made on the interpolation accuracy. In some studies other than T_a interpolation, researchers practiced this idea and showed enhanced effects, which confirmed that it is an effective approach to retrieve soft data and implement BME interpolation operation. For example, Cao et al. used a binary logistic regression to model the occurrence of the highly pathogenic avian influenza and interpolate the risk, indicating an improved



(a)

(b)

(c)

(d)

Figure 9: WS interpolation maps of $T_{a\text{-mean}}$ with (a) BME-hard and (c) BME-both. (b, d) The corresponding prediction standard error maps.

result [26]. Some scholars also regressed the precipitation on historical records or other factors to generate soft data for better interpolation [3, 27, 28].

We compared the results with several studies contributed to daily T_a estimation. Janatian and Zeng investigated daily T_a retrieval from remotely sensed T_s considering the T_a–T_s relationship. The former proposed a statistical framework for estimating T_a using MODIS T_s data and made a case study in the eastern part of Iran [10], which achieved a RMSE of 3.0°C in daily T_a estimation. The latter evaluated the ability of MODIS T_s data to estimate daily temperature over the Corn Belt in the United States [12]. They found that the RMSE of different land covers ranges from 2°C to 5°C. As seen, the errors of these results are much larger than ours. This can be explained that our methods take far more explanatory variables and integrate hard data and soft data. There are also some studies considering some potential explanatory variables. For example, Alonso et al. used multiple regression models to estimate the daily T_a from some remotely sensed explanatory variables and the variables traditionally used like the ones from the Land Use Land Cover in Rhône-Alpes county, located in southeastern France. They achieved a RMSE of 1.20. This result is consistent with our results. However, our study area has a larger extent and more complicated environment, which may cause that our results have a little lower accuracy than theirs. In their study, remote sensing variables were incorporated and made significant contribution to the T_a prediction model. In another study, a stacking ensemble model was proposed to interpolate daily maximum T_a during summertime over Seoul [7]. They achieved a RMSE of 0.7°C, which is lower than our results. In our study, however, BME-both has an improvement of 40%, which is much better than theirs (less than 20%).

In the first phase of the proposed two-phase process, we used regression modeling to predict $T_{a\text{-mean}}$ in unobserved locations and took them as soft data. From the multivariate linear estimates, the best models show good fitting. However, as explained in the regression analysis operation, the best models are sometimes unusable due to failure to extract data of some variables from the corresponding data sources, and as a result, inferior models have to be selected for calculating estimates in unobserved locations, which degrades the prediction accuracy. The observation errors and unavailability of covariates are of great concerns in the regression phase of our two-phase model. For example, T_s retrieved from satellite remote sensing includes errors depending on the underlying surface. And even T_s is unavailable when the surface is sheltered by clouds. Furthermore, since limited cognition for the problem domain and data unavailability, it is difficult to build a perfect model. For example, we considered 10 putative factors as covariates in our multivariate linear regression analysis. However, there are certainly other unknown (or untaken) factors affecting the dependent variable and the relationship between T_a and covariates is likely nonlinear. Therefore, regression analysis generally is taken as a primary approach in some practices where rough results are acceptable. Nowadays, some alternative advanced approaches such as neural networks are prevalent [29, 30]. These methods can depict the complex nonlinear relationships between outputs and inputs and can improve the prediction accuracy. Nonetheless, enormous input parameters and large data are needed to achieve good results in neural network models.

From the comparison of the interpolation accuracy in six areas, it is shown that BME-both is consistently better than BME-hard in any certain area. This can be expected from the characteristics of spatial distribution of meteorological observation stations: the meteorological observation stations in the east are spatially dense, more hard data (with high quality) are incorporated into the interpolation, and the soft data contribute less to the interpolation results. In contrast, more soft data (considered high uncertainty) are used in the interpolation and lead to bigger errors of the interpolation results.

6. Conclusions

T_a is one of the most important parameters for science and practice. While traditional meteorological station observations are limited and spatially distributed unevenly, environmental factors and remote sensing images are promising to predict T_a with spatially continuous coverage. For example, T_s retrieval has undergone long-term studies shortly after the first Landsat launched. And then in terms of the statistical relationship between T_s and T_a, T_a can be predicted at the pixel locations. With advances in the acquisition of remotely sensed data and retrieval models, some full-fledged algorithms for retrieving T_s from remotely sensed data have been developed. As a newer remotely sensed data source, MODIS has been put into use in many fields since 2000. Especially benefiting from well-developed products with sound algorithms, the width and depth of MODIS data application are promoted largely. The distribution website of MODIS data products created by NASA is capable of providing some land products covering the globe (e.g., T_s). And the global users can freely download these product data according to their requirements.

To densify T_a from discrete meteorological observations, two popular approaches, interpolation and regression, are widely used. Interpolation uses the observed data, while regression fits the observed data and covariates and then predicts the T_a in unobserved locations with covariate inputs. There are advantages and disadvantages for them. Incorporating the relationship between T_a and environmental factors into the interpolation procedure is supposed to supplement additional information into the observations, and thus, higher accuracy is expected to achieve. The BME methods can well blend observations and prior knowledge, which use two kinds of data called hard data and soft data.

Utilizing the relationship between T_a and environmental variables, in this article, a two-phase approach was proposed to increase the accuracy of T_a estimates. Taking $T_{a\text{-mean}}$ prediction as example, first, multivariate linear regression models were fitted between $T_{a\text{-mean}}$ and T_s ($T_{\text{MOD-day}}$, $T_{\text{MOD-night}}$, $T_{\text{MYD-day}}$, $T_{\text{MYD-night}}$) conditional on some environmental factors including vegetative and topographical ones. The fitted models were then used to predict $T_{a\text{-mean}}$ from those factors. The predicted $T_{a\text{-mean}}$ were looked on as stochastic variables, and their distributions were also estimated. In the second phase, BME methods were used to interpolate the meteorological observations of $T_{a\text{-mean}}$ taking the meteorological station observations as hard data and the predicted $T_{a\text{-mean}}$ in the first phase as soft data. It is approved that the proposed methods supplement new information to the observations and thus reduce uncertainty of the estimated results. Particularly, for some stations distributed spatially sparse, our method significantly improves the accuracy. The proposed approach is supposed to map T_a with spatially continuous coverage and higher accuracy simultaneously through blending multisource information.

In the coming work, we are going to investigate other advanced methods for producing soft data, which is expected to further improve the accuracy of T_a estimates. Deep learning regression methods are also promising approaches to achieving T_a estimates with high precision in view of their performance in the prediction of environmental parameters, which are of great interest in our next work. According to the improved methods, we plan to produce high spatial resolution T_a products and distribute them for free use.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors acknowledge the support from the National Key R&D Program of China (Grant no. 2019YFF0301301), National Natural Science Foundation of China (Grant no. 72174031), and Youth Scholar Program of Beijing Academy of Science and Technology (Contract no. YS202004). The authors also appreciate the support for this paper from the Beijing Key Laboratory of Operation Safety of Gas, Heating and Underground Pipelines.

References

[1] C. Vancutsem, P. Ceccato, T. Dinku, and S. J. Connor, "Evaluation of MODIS land surface temperature data to estimate air temperature in different ecosystems over Africa," *Remote Sensing of Environment*, vol. 114, no. 2, pp. 449–465, 2010.

[2] H. Taheri-Shahraiyni and S. Sodoudi, "High-resolution air temperature mapping in urban areas: a review on different modelling techniques," *Thermal Science*, vol. 21, no. 6, pp. 2267–2286, 2017.

[3] C.-L. Wang, S.-B. Zhong, G.-N. Yao, and Q.-Y. Huang, "BME spatiotemporal estimation of annual precipitation and detection of drought hazard clusters using space–time scan statistics in the Yun-Gui-Guang region, Mainland China," *Journal of Applied Meteorology and Climatology*, vol. 56, no. 8, pp. 2301–2316, 2017.

[4] M. Wang, G. He, Z. Zhang et al., "Comparison of spatial interpolation and regression analysis models for an estimation of monthly near surface air temperature in China," *Remote Sensing*, vol. 9, no. 12, p. 1278, 2017.

[5] A. Shtiliyanova, G. Bellocchi, D. Borras, U. Eza, R. Martin, and P. Carrère, "Kriging-based approach to predict missing air temperature data," *Computers and Electronics in Agriculture*, vol. 142, pp. 440–449, 2017.

[6] C. Xu, J. Wang, and Q. Li, "A new method for temperature spatial interpolation based on sparse historical stations," *Journal of Climate*, vol. 31, no. 5, pp. 1757–1770, 2018.

[7] D. Cho, C. Yoo, J. Im, Y. Lee, and J. Lee, "Improvement of spatial interpolation accuracy of daily maximum air temperature using stacking ensemble technique," in *Proceedings of the American Geophysical Union, Fall Meeting*, San Francisco, California, 2019.

[8] I. Kloog, F. Nordio, B. A. Coull, and J. Schwartz, "Predicting spatiotemporal mean air temperature using MODIS satellite surface temperature measurements across the northeastern USA," *Remote Sensing of Environment*, vol. 150, pp. 132–139, 2014.

[9] L. Alonso and F. Renard, "Integrating satellite-derived data as spatial predictors in multiple regression models to enhance the knowledge of air temperature patterns," *Urban Science*, vol. 3, no. 4, p. 101, 2019.

[10] N. Janatian, M. Sadeghi, S. H. Sanaeinejad et al., "A statistical framework for estimating air temperature using MODIS land surface temperature data," *International Journal of Climatology*, vol. 37, no. 3, pp. 1181–1194, 2017.

[11] Y. Yang, W. Cai, and J. Yang, "Evaluation of MODIS land surface temperature data to estimate near-surface air temperature in northeast China," *Remote Sensing*, vol. 9, no. 5, p. 410, 2017.

[12] L. Zeng, B. Wardlow, T. Tadesse et al., "Estimation of daily air temperature based on MODIS land surface temperature products over the Corn Belt in the US," *Remote Sensing*, vol. 7, no. 1, pp. 951–970, 2015.

[13] W. Zhu, A. Lű, and S. Jia, "Estimation of daily maximum and minimum air temperature using MODIS land surface temperature products," *Remote Sensing of Environment*, vol. 130, pp. 62–73, 2013.

[14] Y. Chen, H. Sun, and J. Li, "Estimating daily maximum air temperature with MODIS data and a daytime temperature variation model in Beijing urban area," *Remote Sensing Letters*, vol. 7, no. 9, pp. 865–874, 2016.

[15] W. R. Tobler, "A computer movie simulating urban growth in the Detroit region," *Economic Geography*, vol. 46, no. sup1, pp. 234–240, 1970.

[16] G. Christakos, "A bayesian/maximum-entropy view to the spatial estimation problem," *Mathematical Geology*, vol. 22, no. 7, pp. 763–777, 1990.

[17] C. Wang, S. Zhong, F. Zhang, and Q. Huang, "Precipitation interpolation by multivariate bayesian maximum entropy based on meteorological data in Yun-Gui-Guang region, Mainland China," *Earth and Environmental Science*, vol. 46, 2016.

[18] F. S. Zhang, S. B. Zhong, Z. T. Yang, C. Sun, C. L. Wang, and Q. Y. Huang, "Spatial estimation of losses attributable to meteorological disasters in a specific area (105.0°E–115.0°E, 25°N–35°N) using bayesian maximum entropy and partial least squares regression," *Advances in Meteorology*, vol. 2016, Article ID 1547526, 16 pages, 2016.

[19] X. Kou, L. Jiang, Y. Bo, S. Yan, and L. Chai, "Estimation of land surface temperature through blending MODIS and AMSR-E data with the bayesian maximum entropy method," *Remote Sensing*, vol. 8, no. 2, p. 105, 2016.

[20] A. Li, Y. Bo, Y. Zhu, P. Guo, J. Bi, and Y. He, "Blending multi-resolution satellite sea surface temperature (SST) products using bayesian maximum entropy method," *Remote Sensing of Environment*, vol. 135, pp. 52–63, 2013.

[21] M. Hato, H. Tsu, T. Tachikawa, M. Abrams, and B. Bailey, *The ASTER Global Digital Elevation Model (GDEM)-for Societal Benefit*, AGU Fall Meeting, San Francisco, California, 2009.

[22] G. Christakos and X. Li, "Bayesian maximum entropy analysis and mapping: a farewell to kriging estimators?" *Mathematical Geology*, vol. 30, no. 4, pp. 435–462, 1998.

[23] Z. Yu, S. Zhong, C. Wang, Y. Yang, G. Yao, and Q. Huang, "Mapping comparison and meteorological correlation analysis of the air quality index in mid-eastern China," *ISPRS International Journal of Geo-Information*, vol. 6, no. 2, p. 52, 2017.

[24] A. Kolovos, H. Yu, and G. Christakos, *SEKS-GUI V. 0.6 User Manual*, Department of Geography, San Diego State University, San Diego, CA, 2006.

[25] H.-L. Yu, A. Kolovos, G. Christakos, J.-C. Chen, S. Warmerdam, and B. Dev, "Interactive spatiotemporal modelling of health systems: the SEKS–GUI framework," *Stochastic Environmental Research and Risk Assessment*, vol. 21, no. 5, p. 647, 2007.

[26] C. Cao, M. Xu, C. Chang et al., "Risk analysis for the highly pathogenic avian influenza in Mainland China using meta-modeling," *Chinese Science Bulletin*, vol. 55, no. 36, pp. 4168–4178, 2010.

[27] F. Zhang, Z. Yang, S. Zhong, and Q. Huang, "Exploring mean annual precipitation values (2003–2012) in a specific area

(36N–43N, 113E–120E) using meteorological, elevational, and the nearest distance to coastline variables," *Advances in Meteorology*, vol. 2016, Article ID 2107908, 13 pages, 2016.

[28] S. Zhong, C. Wang, Y. Yang, and Q. Huang, "Risk assessment of drought in Yun-Gui-Guang of China jointly using the standardized precipitation index and vulnerability curves," *Geomatics, Natural Hazards and Risk*, vol. 9, no. 1, pp. 892–918, 2018.

[29] X. Li, L. Peng, Y. Hu, J. Shao, and T. Chi, "Deep learning architecture for air quality predictions," *Environmental Science and Pollution Research*, vol. 23, no. 22, pp. 22408–22417, 2016.

[30] Z. Xiong, J. Zheng, D. Song, S. Zhong, and Q. Huang, "Passenger flow prediction of urban rail transit based on deep learning methods," *Smart Cities*, vol. 2, no. 3, pp. 371–387, 2019.