Theoretical and Applied Climatology
https://doi.org/10.1007/s00704-021-03537-9

ORIGINAL PAPER

An imputation method for the climatic data with strong seasonality and spatial correlation

Yun Qin[1] · Guoyu Ren[1,2] · Panfeng Zhang[1] · Lixiu Wu[3,4] · Kangmin Wen[1]

Received: 27 April 2020 / Accepted: 14 January 2021

Abstract
Missing data were frequently found in the instrumental climatic records, which hindered the statistical analyses on climate change. A novel imputation method, called Imputation Based on Decomposition of Time Series (IBDTS), was developed in this article for the climatic data with strong seasonality and spatial correlation. It was to decompose the time series into three components first, and then to predict the missing values in each component. The trend component was predicted by regression analysis, the seasonal component was predicted by spectral analysis, and the remainder component was predicted by spatial interpolation. The IBDTS imputation method showed relatively small errors in performance, and kept the real attributes of climatic series, including the amplitude and phase with the cycle period of 12 months, and the linear trend. The sensibility to station distance for the IBDTS method was relatively small. In addition, the IBDTS method had the ability to deal with the data with none of or only a few of complete series, and it was possible to be applied not only in the field of climatology but also in other fields as long as the data had the intrinsic properties of strong seasonality and spatial correlation.

Keywords Imputation · Missing data · Climatic data · Seasonality · Spatial correlation

1 Introduction

Climatic data can be missing for many reasons (Shen and Somerville 2019). For instrumental time series, the missing data may be caused by the loss of yearbooks due to wars or fire accidents etc. in the early period, and the occasional interruptions of automatic stations, instrument malfunctions, and network reorganizations etc., in the most recent period (Simolo et al. 2010). Missing data may lead to inaccurate estimation in climate research (Stooksbury et al. 1999; Schneider 2001; Massetti 2014; Domonkos and Coll 2019). Moreover, as most statistical methods assume that the dataset is complete, it is necessary to eliminate missing data before addressing the substantive questions (Hopke et al. 2001; Mudelsee 2014).

There were two ways to eliminate missing data: one was to remove the stations with missing data; the other was to replace missing data with reasonable substituted values (Kabacoff 2015). As removing the stations with missing records would lose large amounts of information, the second way was usually chosen to deal with the missing data. In statistics, the process of replacing missing data was called imputation (Little and Rubin 2002; van Buuren 2012).

In terms of the number of variates with missing data, the imputation methods could be classified into univariate imputation and multivariate imputation (Little and Rubin 2002; van Buuren 2012). As the climatic data was characterized by huge quantities, complex relationships between climatic elements, and mixed missingness mechanisms (Little and Rubin 2002; Wallace and Hobbs 2006; Zhang 2018), multivariate imputation was difficult to accurately deal with the variety of missingness with uncertain statistical distribution, and it might cost a lot of computational resources and time, especially for the high temporal resolution datasets, e.g., hourly records, which reduced the imputation efficiency. Moreover, as many observation analyses on climate change were based on single

✉ Guoyu Ren
guoyoo@cma.gov.cn

[1] Department of Atmospheric Science, School of Environmental Studies, China University of Geosciences, Wuhan, China

[2] Laboratory for Climate Studies, National Climate Center, China Meteorological Administration, Beijing, China

[3] Department of Applied Statistics, School of Science, Guangxi University of Science and Technology, Liuzhou, China

[4] Department of Statistics, School of Mathematics and Statistics, Northeast Normal University, Changchun, China

Published online: 30 January 2021

element (Bindoff et al. 2013), in this article we only focused on the univariate imputation to single climatic element rather than the multivariate imputation to all.

There were many traditional univariate imputation methods, including mean imputation, regression imputation, stochastic regression imputation, hot-deck imputation, cold-deck imputation, etc. (Dempster 1977; Ford 1983; Little and Rubin 2002). Though those traditional imputation methods were simple and convenient, limitations and drawbacks were shown in the practical applications (van Buuren 2012; Kang et al. 2012). For instance, they lacked the utilization of the temporal information (Luo et al. 2018). In recent years, model-based methods and machine learning were introduced for imputation of time series, such as Kalman filtering, K-nearest neighbor (KNN), recurrent neural network (RNN), and auto-associative neural network (AANN) (Grewal and Andrews 2008; García-Laencina et al. 2010; Mudelsee 2014). In the fields of atmospheric and climate sciences, empirical orthogonal function (EOF) and expectation maximization (EM) algorithm were also frequently used to fill missing climatic data (von Storch and Zwiers 1999; Navarra and Simoncini 2010; Wilks 2019). However, most of the advanced algorithms and software packages were designed and developed at the service of multivariate data, which could not be applied to the univariate data directly (Moritz and Bartz-Beielstein 2017). Moreover, climatic data were characterized by not only the properties of time series but also the spatial correlation related to the geographical location, but few imputation methods took both of them into consideration at the same time.

Therefore, it was necessary to develop an imputation method to deal with the climatic data due to its distinctive spatio-temporal attributes. On the one hand, as many climatological processes were linked to externally enforced deterministic cycles, e.g., the annual cycle, the temporal variation of climate was usually characterized by strong seasonality (von Storch and Zwiers 1999; Mudelsee 2014; Deng and Fu 2019). If one value was missing, it was hopeful to be filled based on the periodicity of climatic time series. In order to find a reasonable value to replace the missing value, spectral analysis was a good way to extract the periodic components of time series (von Storch and Zwiers 1999; Smith 1999; Alessio 2016; Shumway and Stoffer 2017). On the other hand, there was strong spatial correlation for climatic data. Generally, the closer the meteorological stations were located, the closer the values of climatic elements were observed. If there was a missing value at any one station, it usually could be filled by the records from its neighbor stations with some methods. Those kinds of methods which were to fill the missing data based on spatial relationship were usually called interpolation, including inverse distance weighting (IDW), Kriging, two-dimensional splines, etc. (Dobesch et al. 2007; Fischer and Getis 2010; Li and Heap 2014; Kisaka et al. 2016).

In this article, the data of surface air temperature, a representative climatic element with strong seasonality and spatial correlation, was taken as an example to make imputation. Based on the decomposition of time series, a novel imputation method was developed with the consideration of both temporal and spatial information of the climatic data. This paper was organized as follows: the data and the imputation method were described in Section 2; the imputation results and the comparisons with three previous methods were shown in Section 3; discussion was made in Section 4 on the advantages of the imputation method developed in this article and the disadvantages which were hopeful to be improved in future; and conclusions were drawn in Section 5.

2 Materials and methods

2.1 Data

The dataset of monthly surface air temperature series was provided by National Meteorological Information Center, China Meteorological Administration. The number of meteorological stations in the dataset was 2419, with 143 national reference climate stations, 682 national basic meteorological stations, and 1594 national ordinary stations, respectively. The observed time ranged from January 1951 to December 2015 (780 months in total). Quality control and homogeneity adjustment had been made by Cao et al. (2016). The spatial distribution of stations was featured as dense in the east part of China and sparse in the west part (Fig. 1a).

The missing rate of the dataset was 13.74%. In terms of the length of monthly series, though it was short at some stations, the stations with records $\geq$660 months accounted for 81.48% (Fig. 1a). In terms of the number of stations, it ranged from 142 to 2419 throughout the 780 months (Fig. 1b). It was noted that the stations were few during the first decade, and from then on, the number of stations increased to $\geq$2028. Therefore, the monthly series from January 1961 to December 2015 (660 months in total) was used in this research. The missing rate during this period was 4.22%. There were 1600 stations with complete records, and for the other 819 stations, the number of missing values in the monthly series ranged from 1 to 598. The annual values were considered as missing as long as there were any missing monthly values during the year. Accordingly, the number of missing values in the annual series ranged from 1 to 50, at those stations with incomplete records.

2.2 Methods

The meteorological stations were divided into two groups: G_1 with complete records, and G_2 with at least one missing value at the station. It was necessary to make imputation for the missing data in G_2. A novel imputation method developed in

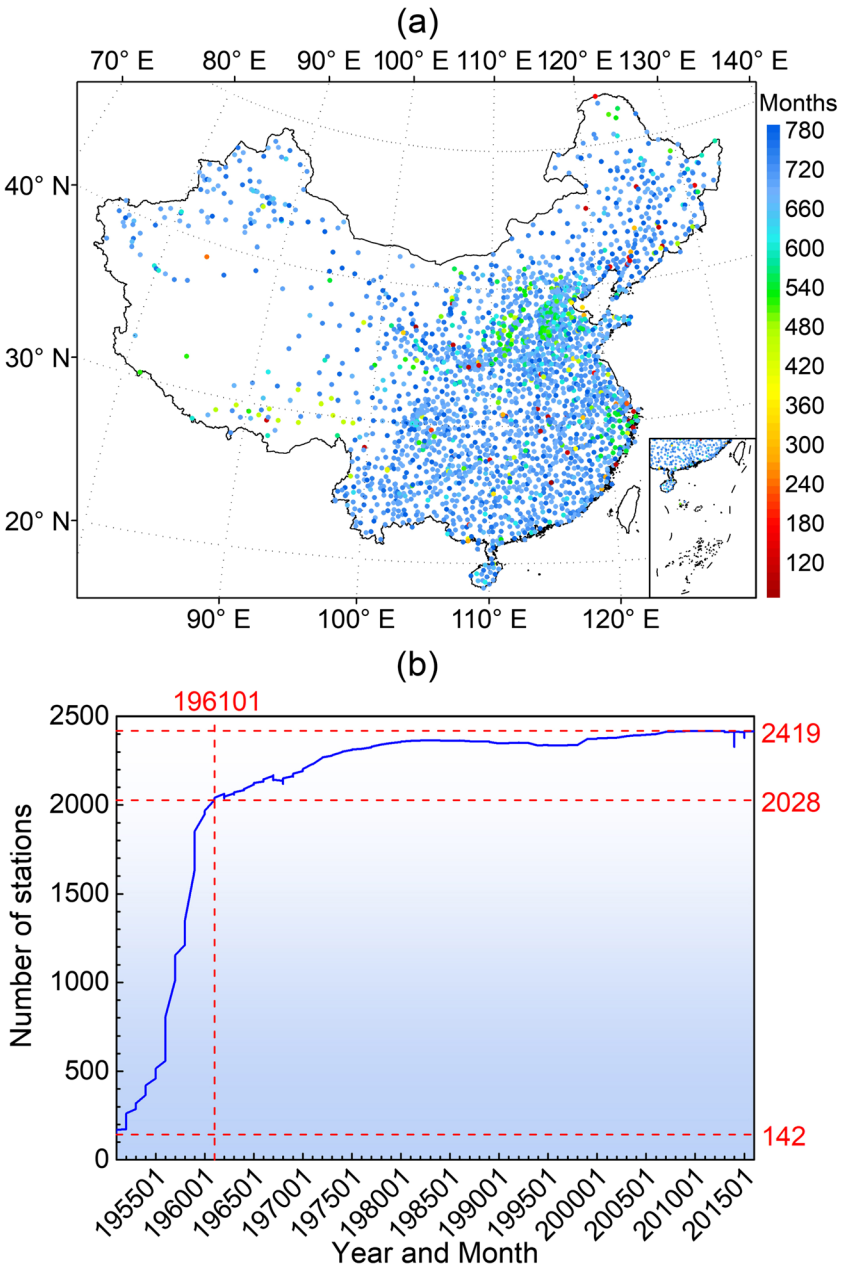

Fig. 1 The spatial distribution of meteorological stations (**a**) and the number of stations varied with time (**b**). The color at each station in patch (**a**) represented the length of monthly series with records: the bluer the color, the longer the monthly series; the redder the color, the shorter the monthly series. The maximum length was 780, i.e., complete monthly series; the minimum length was 62

this article, called Imputation Based on Decomposition of Time Series (IBDTS), was used to predict the values of monthly series in G_2. The IBDTS method used the data information from not only G_1 but also G_2 itself (Fig. 2). Finally, the missing values in G_2 would be replaced by the predicted values.

2.2.1 Decomposition of time series

The monthly temperature series, denoted by $TEM_{[\mathrm{mon}],t}$ at time point t, could be decomposed into three components: (1) the trend component, denoted by T_t, which reflected the long-term progression of temperature change (secular variation); (2) the seasonal component, denoted by S_t, which reflected the seasonality (seasonal variation); (3) the remainder component, denoted by R_t, which was related to the stochastic weather variability, measurement error, etc. (irregular variation) (Kendall 1976; von Storch and Zwiers 1999; Hyndman and Athanasopoulos 2018). Therefore, the monthly temperature series of a station in G_1 at time point t could be expressed as

$$TEM_{[\mathrm{mon}],t} = T_t + S_t + R_t \quad (1)$$

Because of the relatively small climatic change on time scale during the research period in this article, linear trend was assumed. Hence, the T_t could be expressed as

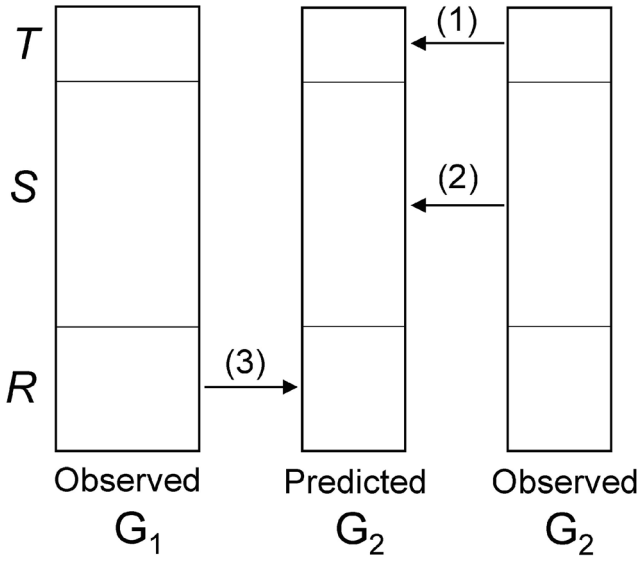

Fig. 2 Data stream between the observed and the predicted values in the two groups. *T*, *S*, and *R* represented the trend component, the seasonal component, and the remainder component, respectively (see Section 2.2.1). The direction of arrows represented the direction of data stream, and the numbers above the arrows reflected the steps of data processing

$$T_t = b_0 + b_1 \cdot t \tag{2}$$

where b_0 and b_1 were the intercept and the linear trend of linear regression fitted by the least-squares, respectively; t was a sequence number of chronological month ($t = 1, 2, 3, \cdots, 660$).

The discrete Fourier transform (DFT) was used to transform the time domain signal, the sequence $X_t (=TEM_{[\mathrm{mon}],t} - T_t)$, into the frequency domain signal, the sequence Y_f, and the process was expressed as

$$Y_f = \sum_{t=1}^{L} X_t \cdot e^{-i\frac{2\pi}{L}(t-1)f} \tag{3}$$

where i was the imaginary unit; L was the length of t; f was a normalized frequency ($f = 0, 1, 2, \cdots, L-1$) (Proakis and Manolakis 1996; von Storch and Zwiers 1999; Alessio 2016). In consideration of the Euler's formula (Moskowitz 2002), the expression-(3) was expressed as

$$Y_f = \sum_{t=1}^{L} X_t \cdot \left[\cos\left(\frac{2\pi}{L}(t-1)f\right) - i \cdot \sin\left(\frac{2\pi}{L}(t-1)f\right)\right] \tag{4}$$

The normalized frequency f was a sinusoid's frequency with f cycles per L samples. Y_f was a complex number that encoded both amplitude and phase of a complex sinusoidal signal. The amplitude and the phase at frequency f were respectively

$$A_f = \left|Y_f\right|/N \tag{5}$$

$$c_f = -i \cdot \ln\left(Y_f/\left|Y_f\right|\right) \tag{6}$$

where $|Y_f|$ was the modulus of complex number Y_f, and N was the number of stations in G_1. When f=0, $A_{f=0}$ was the direct current component of Y (i.e., the composite of Y_f at all frequency f), which was equal to the arithmetic mean value of the sequence X_t. In addition, the inverse discrete Fourier transform (IDFT) was derived as

$$X_t = \frac{1}{L}\sum_{f=0}^{L-1} Y_f \cdot e^{i\frac{2\pi}{L}(t-1)f} \tag{7}$$

As f referred to the number of cycles in L months, there were L/f months in each cycle, which was defined as cycle period. Due to the symmetrical characteristic of Y_f, the actual amplitude was equal to the double value of A_f when f varied from 1 to $L/2$. When the cycle period was equal to 12 months (i.e., $f = 55$), based on the expression-(7), the component of $Y_{f=55}$ could be transformed into the time domain signal, which was the half of the seasonal component (i.e., $S_t/2$). Then, we could get the S_t. After the T_t and S_t were extracted from the $TEM_{[\mathrm{mon}],t}$, the remaining component was R_t.

2.2.2 Prediction for the trend component

The annual temperature series of the stations in G_2, denoted by $TEM_{[\mathrm{ann}],yr}$ at the yr^{th} year without any missing monthly values, the value of which was the arithmetic mean of the 12 monthly values in the year. Although there were some missing values in the annual temperature series, linear regression could also be applied to G_2. The expression of trend component of the stations in G_2 was similar to that in G_1, by using the annual temperature instead of the monthly temperature at the time point of optimal month. The optimal month for a station was

$$\widehat{m} = \arg\min_{m}\left\{\sum_{yr}\left(TEM_{[\mathrm{ann}],yr} - TEM_{[\mathrm{mon}],yr}(m)\right)^2\right\}(m = 1, 2, \cdots, 12) \tag{8}$$

where $TEM_{[\mathrm{mon}],yr}(m)$ was the monthly temperature at the m^{th} month within the yr^{th} year. Note that the total number of yr varied with stations, and the largest yr was less than $L/12$.

2.2.3 Prediction for the seasonal component

The DFT was also used to calculate the amplitude and phase where cycle period was 12 months in G_2. If there were any missing monthly values in the year, the whole monthly values in this year were regarded as missing. By omitting the months with missing values, a new monthly temperature series was generated which included the completed monthly temperature values in the years that the value of annual temperature existed. Note that corresponding trend component had been subtracted before the amplitude and phase were calculated. With the amplitude and phase at the frequency where the cycle

period was 12 months, the seasonal components of stations in G_2 could be calculated by the expression (7).

2.2.4 Prediction for the remainder component

Based on the remainder components of stations at time point t in G_1, the values of remainder components of stations at time point t in G_2 were predicted through the IDW interpolation (Philip and Watson 1982; Watson and Philip 1985). The values of 12 nearest neighbor stations were used to calculate the value of the interpolated station, and the power of weight function of IDW was set as 1.

2.3 Assessment

2.3.1 Criteria for performance evaluation

The root mean squared error (RMSE), mean absolute error (MAE), and mean bias error (MBE) were used to evaluate the performance of imputation method, which were expressed as follows (Willmott and Matsuura 2005; Du et al. 2020)

$$\mathrm{RMSE} = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(\widehat{TEM}_{[\mathrm{mon}],k} - TEM_{[\mathrm{mon}],k}\right)^2} \tag{9}$$

$$\mathrm{MAE} = \frac{1}{K}\sum_{k=1}^{K}\left|\widehat{TEM}_{[\mathrm{mon}],k} - TEM_{[\mathrm{mon}],k}\right| \tag{10}$$

$$\mathrm{MBE} = \frac{1}{K}\sum_{k=1}^{K}\left(\widehat{TEM}_{[\mathrm{mon}],k} - TEM_{[\mathrm{mon}],k}\right) \tag{11}$$

where $TEM_{[\mathrm{mon}],k}$ and $\widehat{TEM}_{[\mathrm{mon}],k}$ were the observed and the predicted monthly temperature at time point k, respectively; K was the number of observed values in the monthly temperature series ($K < L = 660$).

In addition, three special indicators for climatic data were used to measure whether or not the imputation changed the attributions of true data. They were the amplitude and the phase with the cycle period of 12 months, and the linear trend, respectively. As the missing rate during the period of 1961–2015 was small, we assume that the observed values kept the attributions of true data.

2.3.2 Comparisons with previous imputation methods

Based on the above-mentioned six indicators, three kinds of previous imputation methods were compared with the IBDTS method, which were hot-deck imputation, RNN imputation, and IDW interpolation. The first one utilized none of the spatiotemporal information of the climatic data, the second one utilized only the temporal information, and the third one utilized only the spatial information to fill the missing values.

The hot-deck imputation was to predict the missing data with the values from a similar complete data series, which was one of the most common method in practice (Little and Rubin 2002; García-Laencina et al. 2010). In this article, for each station in G_2, the most similar station in G_1 was selected to make the prediction. The selection criteria were the correlation coefficient of anomaly series between the predicted station in G_2 and all the stations in G_1; the higher the correlation coefficient was, the more similar they were. Note that the anomaly here was referring to the difference between the time series and its mean (i.e., multi-year mean). In G_2, the values in the years with missing monthly values were ignored when calculating the mean of monthly temperature series. The anomaly series of the station selected from G_1 was then taken as that of the predicted station in G_2. With its mean added, the monthly temperature series could be predicted.

The RNN imputation was to predict the missing data by the feedforward networks with backpropagation training, which had a dynamical memory to keep temporal information (Lukoševičius and Jaeger 2009; García-Laencina et al. 2010; Pasini 2015). In this article, for each station in G_2, the three most similar stations in G_1 were selected to make the prediction, which was also based on the correlation coefficient of anomaly series described above. The RNN with 3 input layers, 2 hidden layers with 5 neurons each, and 1 output layer were created. According to the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm (Broyden 1970; Fletcher 1970; Goldfarb 1970; Fletcher 1987; Nocedal and Wright 2006; Haghighi 2014), the monthly temperature series of the three stations in G_1 (input) and that of the predicted station in G_2 (output) at the time points (i.e., months) with records in the predicted station were used as training dataset. The RNN toolbox for MATLAB was downloaded from GitHub (Atabay 2016).

IDW interpolation was a spatial interpolation method, which was to predict the missing data by calculating a weighted average with the inverse distance between the predicted station and its surrounding stations (Myers 1994; Dobesch et al. 2007; Xu et al. 2013). In this article, for each time point, the missing data of stations in G_2 were only predicted by the complete series in G_2. The parameters of IDW were the same with those in the IBDTS imputation.

In addition, an experiment was designed to test the sensibility to station distance for the imputation methods. Firstly, n_1 (={5, 10, 15, 20, 30, 40, 50, 75, 100, 200, 300, 500, 700, 1000, 1300, 1600}) stations from G_1 and n_2 (={5, 10, 15, 20, 30, 40, 50, 75, 100, 200, 300, 500, 700}) stations from G_2 were picked out at random. Secondly, 50 repeated trials were made for each combination of n_1 and n_2, and the average station distances in each group for each trial were calculated. The average station distance was the mean of distance between any two connected stations under the division of triangulated irregular network (TIN) (Delaunay 1934). Thirdly, the mean and standard deviation of RMSE were calculated for each combination of n_1 and n_2.

3 Results

3.1 Amplitude and phase

The spectrogram showed that the largest amplitude occurred in the cycle period of 12 months, followed by the cycle period of 6 months (Fig. 3), which was in accordance with the previous studies (Deng et al. 2018). The amplitude with the cycle period of 12 months at all 2419 stations ranged from 2.8 to 24.4 °C, which showed clear gradual increase from low latitude to high latitude (Fig. 4a). The higher the amplitude was, the larger the temperature fluctuated within a year.

The phase with the cycle period of 12 months also showed a gradual spatial variation across China's mainland, with smaller values in the southeast part of China (Fig. 4b). Note that theoretically, the phase ranged from $-\pi$ to π. However, as the phases at all the 2419 stations were near the two values of $-\pi$ and π, a numeral of 2π was added to the phases which were less than 0. The larger the phase was, the earlier the temperature extremes in the cosine wave occurred within a year; vice versa. Note that the phase of π represented the hottest and coldest condition, i.e., the crest and trough in the cosine wave, occurred in July and in January, respectively, and meanwhile, June and August tied for the 2nd hottest, May and September tied for the 4th hottest, April and October tied for the 6th hottest, etc.

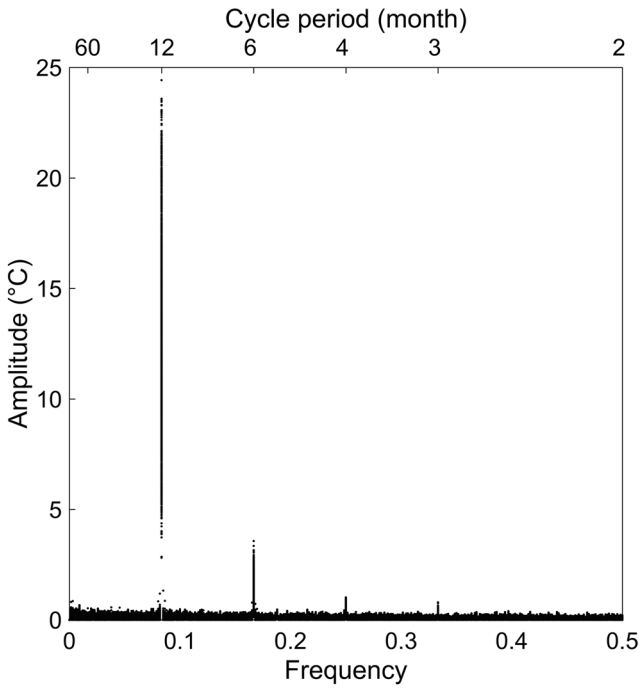

Fig. 3 Spectrogram of the monthly temperature series of the 2419 stations. The ordinate represented the actual amplitude calculated with DFT; the lower abscissa represented the frequency, which was equal to f/L, with f ranging from 0 to $L/2$ (see Section 2.2); the upper abscissa represented the cycle period, which was equal to the reciprocals of the values on the lower abscissa

3.2 Imputation results

As was shown in Fig. 5, the RMSE of IBDTS method was relatively small at most of the stations (97.19%). The RMSEs larger than 1.0 were mainly found in the west part of China where the station distribution was sparse. The negative correlation coefficient between RMSE and station density also showed that the lower the station density was, the larger the RMSE was, which indicated that the station distance had an effect on the accuracy of imputation. These results may be related to the bad performance of the prediction for the remainder component in the imputation process. In this article, the remainder component was predicted by the IDW interpolation, which was a simple spatial interpolation method with limited accuracy.

3.3 Comparisons with previous imputation methods

3.3.1 RMSE, MAE, and MBE

According to the results of RMSE and MAE, the IBDTS method performed better than the hot-deck and the IDW method, but a little worse than the RNN method (Table 1). Both the mean and the standard deviation of RMSE and MAE of the IDW method were the largest, which indicated that there would be much inaccurate for the imputation of climatic data with only the consideration of spatial information. However, by the means of decomposing the time series into three components, the missing values in each component were filled separately, with only the remainder component filled by the IDW interpolation, and the results could be greatly improved (see the RMSE and MAE of IBDTS method). As the hot-deck method utilized the information of similar complete series, the errors of performance were relatively small, even if it did not utilize any spatiotemporal information. Compared with the hot-deck method, the RNN method performed better. These results showed that the accuracy of imputation could be improved with temporal information included. In addition, in contrast with other three methods, the MBE of IBDTS method showed that it had negative average bias. Though the absolute value of the mean of MBE of IBDTS method was a little large than the IDW method, the standard deviation of MBE was much smaller than its. Overall, in the aspect of three common statistical indicators, the performance of the IBDTS method was better than the hot-deck and the IDW method, but worse than the RNN method.

3.3.2 Amplitude, phase, and linear trend

As was shown in Fig. 6, the amplitude, phase, and linear trend could be changed more or less by the imputation process. In terms of the amplitude and phase, the RNN method performed much better than the hot-deck and the IDW method. These

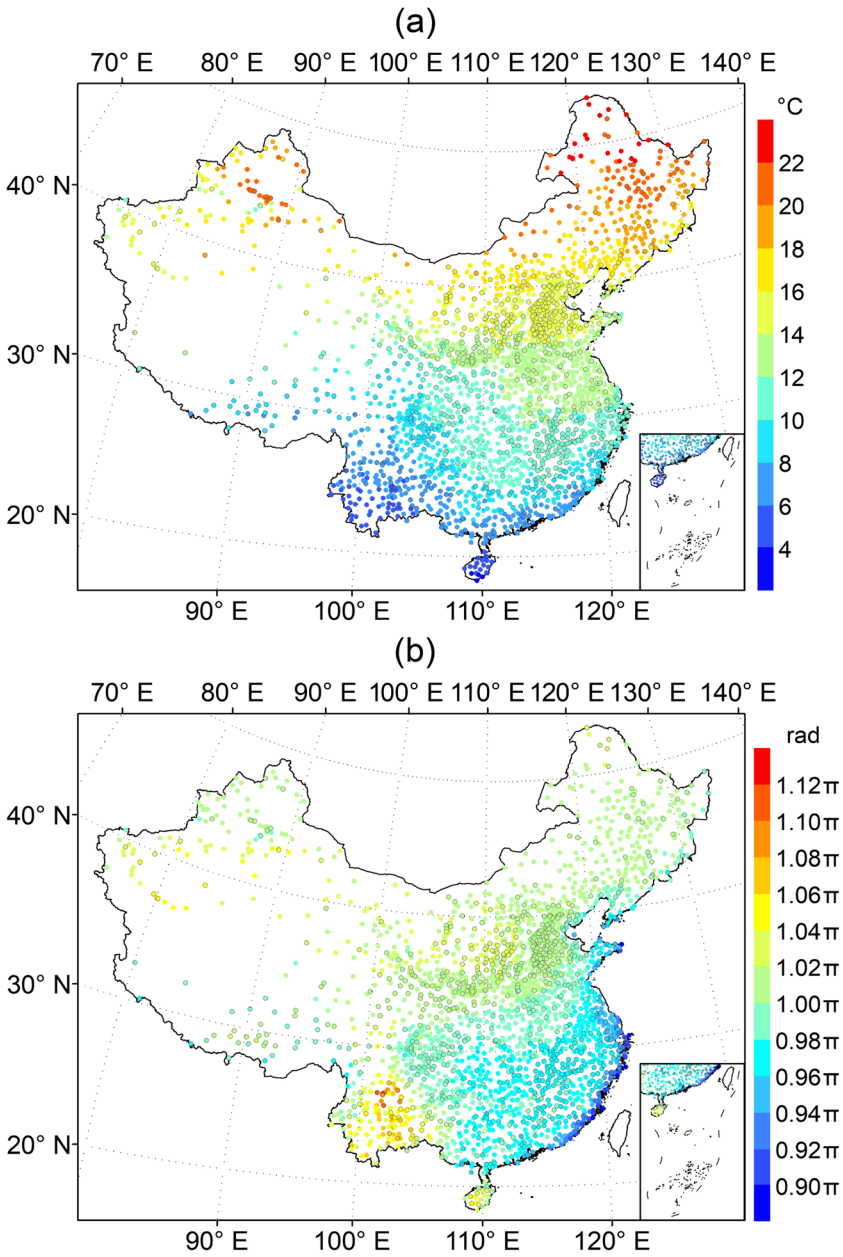

Fig. 4 Spatial distribution of the amplitude (**a**) and phase (**b**) with the cycle period of 12 months. There were 2419 stations in total: the dots with outlines represented the stations in G_2; otherwise, in G_1

results should be related to that both the hot-deck and the IDW method did not utilize the temporal information of incomplete series itself. As to the hot-deck method, the amplitude and phase at the predicted station in G_2 were just replaced by that at the station most related to it in G_1, leading to a large error for amplitude and phase in the area with sparse stations. In addition, all of the three previous imputation methods made the phase smaller, with the medians lower than the zero-value line, respectively (Fig. 6b). These results were very possibly caused by the fact that there were a lot of stations with missing data in the west part of China. The information which were used to fill the missing data at those stations mostly came from the stations in G_1 located in the east part of China, where the phase was smaller (see Fig. 4b).

As to the linear trend, the difference on performance between the three imputation methods was small, with the RNN method slightly better than the other two (Fig. 6c). The stations at which the absolute differences of linear trend between the predicted series and original series ranged within 0.2 °C/decade accounted for 94.63%, 95.48%, and 93.04% based on the hot-deck, RNN, and IDW methods, respectively, and the large differences mostly occurred when missing rate was larger than a half. Whether the linear trends of original series were positive or negative, those of predicted series based on the three methods were mainly ranging from 0.0 to 0.5 °C/decade (not shown). These results may be highly related to the fact that the linear trends of 97.81% stations in G_1 were in that range.

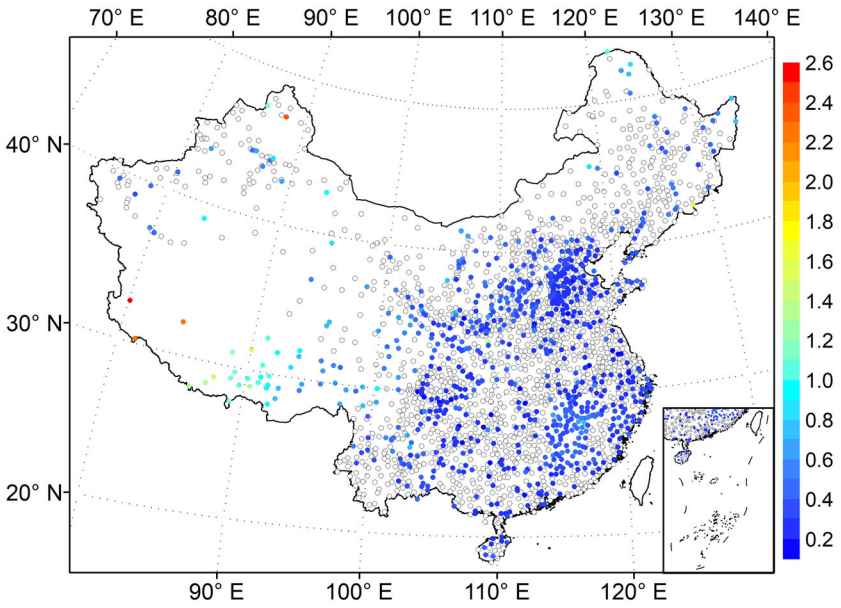

Fig. 5 Spatial distribution of the RMSE of IBDTS method. The colored dots represented the stations in G_2, and the stations in G_1 were marked with hollow

Compared with the three previous imputation methods, the IBDTS method developed in this article kept the real attributes of climatic data in original series, including the amplitude, phase, and linear trend.

3.3.3 Sensibility to station distance

The experimental results showed that the station distance in G_2 had a little effect on the RMSE, but the effect of station distance in G_1 was significant (not shown). For all of the four imputation methods, the mean of RMSE showed systematic increase with the increase of station distance in G_1, and the increase rates of mean and standard deviation of RMSE of IBDTS method were lower than that of hot-deck and IDW methods, but a little higher than that of RNN method (Fig. 7). When the average station distance in G_1 was about 100 km, the performance of the four imputation methods were relatively close; however, when the distance became large, the advantage of IBDTS and RNN methods was apparent. Overall, compared with the hot-deck method, the IBDTS and the RNN methods were more reasonable for the imputation in the areas with sparse stations.

Table 1 Comparison results of RMSE, MAE, and MBE between four imputation methods. The mean and the standard deviation (the values in the brackets) were calculated

Methods	Hot-deck	RNN	IDW	IBDTS
RMSE	0.472 (0.264)	0.277 (0.148)	1.162 (1.364)	0.396 (0.232)
MAE	0.376 (0.214)	0.213 (0.110)	1.065 (1.336)	0.309 (0.179)
MBE	0.068 (0.167)	0.000 (0.001)	0.021 (1.667)	−0.023 (0.043)

4 Discussion

Due to the imputation mechanism of the IBDTS method, any outliers in the series might lead to wrong prediction to the remainder component, and the inhomogeneous series had an impact on the trend estimation (Wang and Swail 2001; Vincent et al. 2012); therefore, it was necessary to make quality control and homogeneity adjustment before imputation. As the imputation method had taken not only the neighboring stations but also the predicted station itself into account and the trend and the seasonal components which accounted for dominant proportion of the whole components were predicted based on the predicted station itself, any one missing value in the sequence of remainder component could be predicted based on the its neighboring stations with records at that time point; therefore, the IBDTS method could deal with the data in which none of or only a few of climatic series were complete.

In terms of RMSE, MAE, and MBE, the IBDTS method behaved better than the hot-deck and IDW methods but worse than the RNN method. However, all of the three previous methods had changed the attributes of the original series, including amplitude, phase, and linear trend, which may have an impact on the detection and attribution of climate change. Moreover, both the hot-deck and the RNN methods depended on the stations with complete series, which may made them unavailable to the imputation of the climatic series observed in the early period characterized by large amounts of missing data with few complete series (Brönnimann et al. 2018).

Nevertheless, there were some flaws of IBDTS method, which could be improved in the future. Firstly, as the linear trend during the recorded years may not represent that throughout the whole time series, the decomposed linear trend from the original series may be not correct when there were

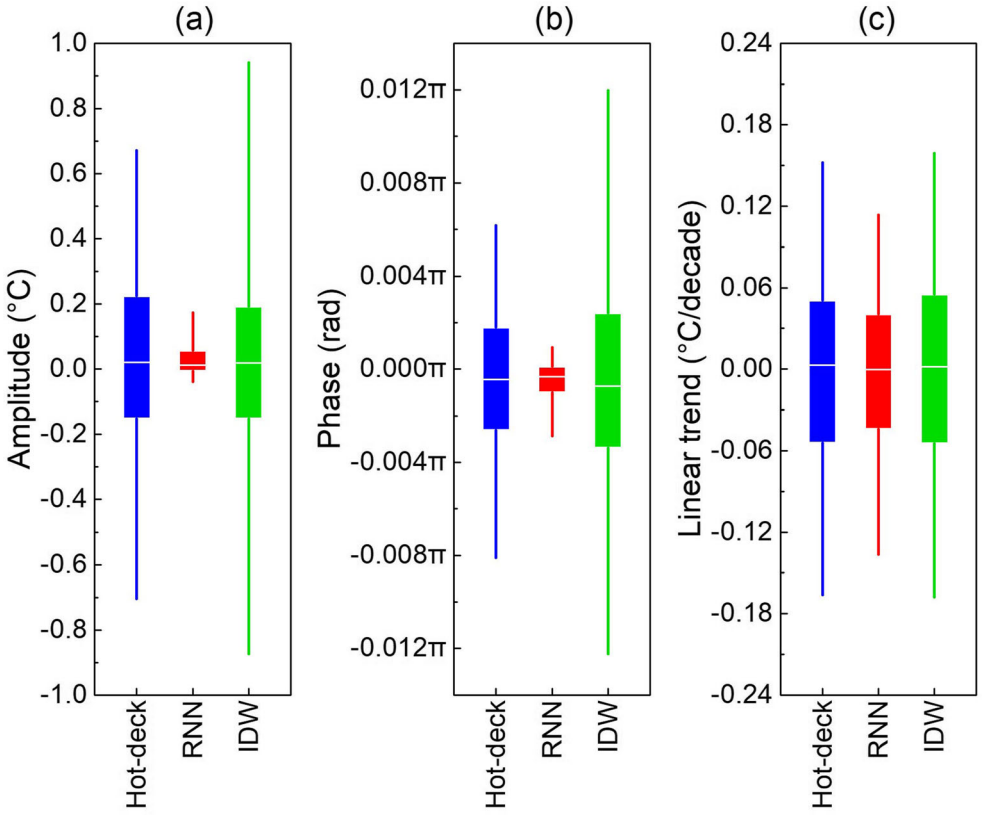

Fig. 6 Box plot of the amplitude (**a**) and phase (**b**) with the cycle period of 12 months, and the linear trend (**c**). The values represented the differences between the predicted series and the original series. The low and the high edge of the boxes represented the positions of the lower quartile (25%) and the upper quartile (75%), respectively. The endpoints of the lines extending vertically from the boxes represented the positions of the percentage of 5% and 95%, respectively. The white lines across the boxes represented the medians. Note that the box plots for IBDTS imputation were not shown due to zero values

many missing values, therefore, the extraction of linear trend should be improved. Secondly, it should be compared that the difference of accuracy to estimate the remainder component between different spatial interpolation methods to check if other methods behaved better. Thirdly, the adjustment to the bandwidth of weight function might improve the accuracy for the prediction of remainder component. In this article, 12 nearest neighbor stations were selected to interpolate, but it was necessary to check whether it behaved better when more or less neighbor stations were included, and whether the fixed bandwidth (i.e., the neighbor stations were the stations within the radius of fixed distance) behaved better than the adaptive bandwidth (i.e., the distance was changeable with the number of stations, and the bandwidth was set according to the number) (Fischer and Getis 2010). Fourthly, it remained to answer if the weight function was reasonable. In this article, the weight function was only a function of distance, and it remained to check if other factors included could improve the accuracy, such as direction and altitude. Finally, as this research was focused on only the monthly series, it was necessary to make it clear whether the imputation for the series with higher temporal resolution (e.g., daily series, hourly series) could still behave well by modifying the parameters of method.

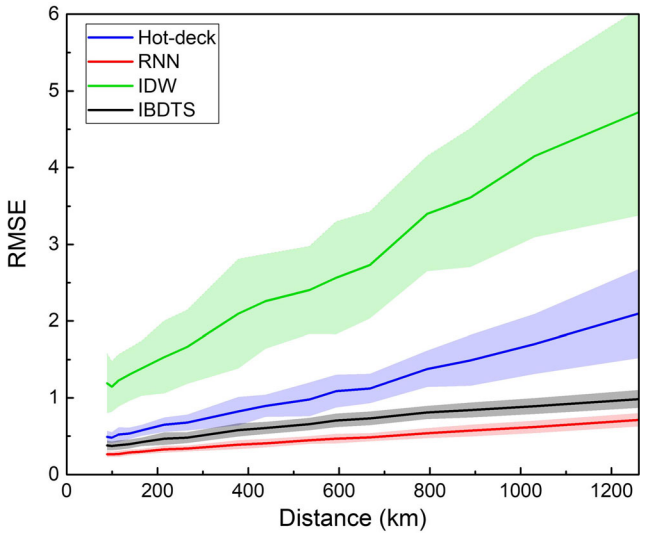

Fig. 7 Relationship between RMSE and the average station distance in G_1. Only the situation of the average station distance of 790 km (at about medium distance when n_2 varied from 5 to 700, described in Section 2.3) in G_2 was shown in the figure. The solid lines represented the mean of RMSE varied with the increase of distance, and the standard deviation of RMSE was shown with the shadow regions on the two sides of the solid lines

As the IBDTS method was designed for the data with strong seasonality and spatial correlation, it was potential to be applied not only in the field of climatology but also in other fields as long as the time series had the attributions of strong seasonality and spatial correlation.

5 Conclusions

The IBDTS imputation method developed in this article had taken both of the temporal and spatial information into consideration, which had a relatively small RMSE, MAE, and MBE in performance, and kept the real attributes of the original series, including the amplitude and phase with the cycle period of 12 months, and the linear trend.

The station distance had an effect on the accuracy of IBDTS method, with larger RMSE in the areas where the stations distribution was sparse. The sensibility to station distance for the IBDTS method was a little stronger than that for the RNN method but weaker than for the hot-deck and IDW methods. Compared with the hot-deck and RNN methods, the IBDTS method had the ability to deal with the data with none of or only a few of complete series, e.g., the climatic data in the early period of instrumental records.

The IBDTS imputation method could be improved in the future on the aspects of the prediction for each component and the application for the series with higher temporal resolution. It was potential to be applied not only in the field of climatology but also in other fields as long as the time series had the attributions of strong seasonality and spatial correlation.

Acknowledgements This work was supported by the Chinese Ministry of Science and Technology (MOST) National Key R&D Program (No.2018YFA0605603) and the Science Foundation Program of Guangxi University of Science and Technology (No.1711311). The authors thank Yunxin Huang, Tianlin Zhai, and Huqiang Qin for their kind assistance during manuscript writing, and the reviewers for providing constructive comments, which greatly improved this manuscript.

References

Alessio SM (2016) Digital signal processing and spectral analysis for scientists: concepts and applications. Springer, New York

Atabay D (2016) Pyrenn: first release (version v0.1). Zenodo. https://doi.org/10.5281/zenodo.45022

Bindoff NL, Stott PA, AchutaRao KM, Allen MR, Gillett N, Gutzler D, Hansingo K, Hegerl G, Hu Y, Jain S. Mokhov II, Overland J, Perlwitz J, Sebbari R, Zhang X (2013) Detection and attribution of climate change: from global to regional. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Stocker TF, Qin D, Plattner G-K, Tignor M, Allen SK, Boschung J, Nauels A, Xia Y, Bex V, Midgley PM (eds)]. Cambridge University Press, Cambridge, and New York

Brönnimann S, Brugnara Y, Allan RJ, Brunet M, Compo GP, Crouthamel RI, Jones PD, Jourdain S, Luterbacher J, Siegmund P, Valente MA, Wilkinson CW (2018) A roadmap to climate data rescue services. Geosci Data J 5:28–39. https://doi.org/10.1002/gdj3.56

Broyden CG (1970) The convergence of a class of double-rank minimization algorithms. IMA J Appl Math 6:76–90. https://doi.org/10.1093/imamat/6.1.76

Cao L, Zhu Y, Tang G, Yuan F, Yan Z (2016) Climatic warming in China according to a homogenized data set from 2419 stations. Int J Climatol 36:4384–4392. https://doi.org/10.1002/joc.4639

Delaunay B (1934) Sur la sphère vide. A la mémoire de Georges Voronoï. Bulletin de l'Académie des Sciences de l'URSS 6:793–800

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc 39:1–38

Deng Q, Fu Z (2019) Comparison of methods for extracting annual cycle with changing amplitude in climate series. Clim Dyn 52:5059–5070. https://doi.org/10.1007/s00382-018-4432-8

Deng Q, Nian D, Fu Z (2018) The impact of inter-annual variability of annual cycle on long-term persistence of surface air temperature in long historical records. Clim Dyn 50:1091–1100. https://doi.org/10.1007/s00382-017-3662-5

Dobesch H, Dumolard P, Dyras I (eds) (2007) Spatial interpolation for climate data: the use of GIS in climatology and meterology. ISTE, London

Domonkos P, Coll J (2019) Impact of missing data on the efficiency of homogenisation: experiments with ACMANTv3. Theor Appl Climatol 136:287–299. https://doi.org/10.1007/s00704-018-2488-3

Du Z, Wang Z, Wu S, Zhang F, Liu R (2020) Geographically neural network weighted regression for the accurate estimation of spatial non-stationarity. Int J Geogr Inf Sci 34:1353–1377. https://doi.org/10.1080/13658816.2019.1707834

Fischer MM, Getis A (eds) (2010) Handbook of applied spatial analysis: software tools, methods and applications. Springer, Heidelberg

Fletcher R (1970) A new approach to variable metric algorithms. Comput J 13:317–322. https://doi.org/10.1093/comjnl/13.3.317

Fletcher R (1987) Practical methods of optimization, 2nd edn. Wiley, New York

Ford BL (1983) An overview of hot-deck procedures. Incomplete Data in Sample Surveys. 2:185–207

García-Laencina PJ, Sancho-Gómez J-L, Figueiras-Vidal AR (2010) Pattern classification with missing data: a review. Neural Comput & Applic 19:263–282. https://doi.org/10.1007/s00521-009-0295-6

Goldfarb D (1970) A family of variable-metric methods derived by variational means. Math Comput 24:23–26. https://doi.org/10.1090/S0025-5718-1970-0258249-6

Grewal MS, Andrews AP (2008) Kalman filtering: theory and practice using MATLAB, 3rd edn. Wiley, Hoboken

Haghighi AD (2014) Numerical optimization: understanding L-BFGS. URL: http://aria42.com/blog/2014/12/understanding-lbfgs. Accessed 2 Dec 2014

Hopke PK, Liu C, Rubin DB (2001) Multiple imputation for multivariate data with missing and below-threshold measurements: time-series concentrations of pollutants in the Arctic. Biometrics 57:22–33

Hyndman RJ, Athanasopoulos G (2018) Forecasting: principles and practice, 2nd edn. OTexts, Melbourne

Kabacoff RI (2015) R in action: data analysis and graphics with R, 2nd edn. Manning, Shelter Island

Kang HM, Yusof F, Mohamad I (2012) Imputation of missing data with different missingness mechanism. Jurnal Teknologi 57:57-67. https://doi.org/10.11113/jt.v57.1523

Kendall MG (1976) Time-series, 2nd edn. Griffin, London

Kisaka MO, Mucheru-Muna M, Ngetich FK, Mugwe J, Mugendi D, Mairura F, Shisanya C, Makokha GL (2016) Potential of deterministic and geostatistical rainfall interpolation under high rainfall variability and dry spells: case of Kenya's central highlands. Theor Appl Climatol 124:349–364. https://doi.org/10.1007/s00704-015-1413-2

Li J, Heap AD (2014) Spatial interpolation methods applied in the environmental sciences: a review. Environ Model Softw 53:173–189. https://doi.org/10.1016/j.envsoft.2013.12.008

Little RJA, Rubin DB (2002) Statistical analysis with missing data, 2nd edn. Wiley, Hoboken

Lukoševičius M, Jaeger H (2009) Reservoir computing approaches to recurrent neural network training. Comp Sci Rev 3:127–149. https://doi.org/10.1016/j.cosrev.2009.03.005

Luo Y, Cai X, Zhang Y, Xu J, Yuan X (2018) Multivariate time series imputation with generative adversarial networks. In: 32nd Conference on Neural Information Processing Systems. Montréal, Canada

Massetti L (2014) Analysis and estimation of the effects of missing values on the calculation of monthly temperature indices. Theor Appl Climatol 117:511–519. https://doi.org/10.1007/s00704-013-1024-8

Moritz S, Bartz-Beielstein T (2017) imputeTS: time series missing value imputation in R. R J 9:207–218. https://doi.org/10.32614/RJ-2017-009

Moskowitz MA (2002) A course in complex analysis in one variable. World Scientific, River Edge

Mudelsee M (2014) Climate time series analysis: classical statistical and bootstrap methods, 2nd edn. Springer, New York

Myers DE (1994) Spatial interpolation: an overview. Geoderma 62:17–28. https://doi.org/10.1016/0016-7061(94)90025-6

Navarra A, Simoncini V (2010) A guide to empirical orthogonal functions for climate data analysis. Springer, Dordrecht

Nocedal J, Wright SJ (2006) Numerical optimization, 2nd edn. Springer, New York

Pasini A (2015) Artificial neural networks for small dataset analysis. J Thoracic Dis 7:953–960. https://doi.org/10.3978/j.issn.2072-1439.2015.04.61

Philip GM, Watson DF (1982) A precise method for determining contoured surfaces. Appea J 22:205–212. https://doi.org/10.1071/AJ81016

Proakis JG, Manolakis DG (1996) Digital signal processing: principles, algorithms, and applications, 3rd edn. Prentice-Hall, Upper Saddle River

Schneider T (2001) Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values. J Clim 14:853–871. https://doi.org/10.1175/1520-0442(2001)014<0853:AOICDE>2.0.CO;2

Shen SSP, Somerville RCJ (2019) Climate mathematics: theory and applications, 1st edn. Cambridge University Press, Cambridge

Shumway RH, Stoffer DS (2017) Time series analysis and its applications: with R examples, 4th edn. Springer, New York

Simolo C, Brunetti M, Maugeri M, Nanni T (2010) Improving estimation of missing values in daily precipitation series by a probability density function-preserving approach. Int J Climatol 30:1564–1576. https://doi.org/10.1002/joc.1992

Smith SW (1999) The scientist and engineer's guide to digital signal processing, 2nd edn. California Technical Publishing, San Diego

Stooksbury DE, Idso CD, Hubbard KG (1999) The effects of data gaps on the calculated monthly mean maximum and minimum temperatures in the continental United States: a spatial and temporal study. J Clim 12:1524–1533. https://doi.org/10.1175/1520-0442(1999)0122.0.CO;2

van Buuren S (2012) Flexible imputation of missing data, 2nd edn. Chapman and Hall/CRC, Boca Raton

Vincent LA, Wang XL, Milewska EJ, Wan H, Yang F, Swail V (2012) A second generation of homogenized Canadian monthly surface air temperature for climate trend analysis. J Geophys Res 117:D18110. https://doi.org/10.1029/2012JD017859

von Storch H, Zwiers FW (1999) Statistical analysis in climate research. Cambridge University Press, Cambridge

Wallace JM, Hobbs PV (2006) Atmospheric science: an introductory survey, 2nd edn. Elsevier Academic Press, Amsterdam

Wang XL, Swail VR (2001) Changes of extreme Wave Heights in northern hemisphere oceans and related atmospheric circulation regimes. J Clim 14:2204–2221. https://doi.org/10.1175/1520-0442(2001)014<2204:COEWHI>2.0.CO;2

Watson DF, Philip GM (1985) A refinement of inverse distance weighted interpolation. Geoprocessing 2:315–327

Wilks DS (2019) Statistical methods in the atmospheric sciences, 4th edn. Elsevier, Cambridge

Willmott CJ, Matsuura K (2005) Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim Res 30:79–82. https://doi.org/10.3354/cr030079

Xu C, Wang J, Hu M, Li Q (2013) Interpolation of missing temperature data at meteorological stations using P-BSHADE. J Clim 26:7452–7463. https://doi.org/10.1175/JCLI-D-12-00633.1

Zhang Z (2018) Multivariate time series analysis in climate and environmental research. Springer International Publishing, Cham