Supplement of

Construction of homogenized daily surface air temperature for Tianjin city during 1887-2019

Peng Si, Qingxiang Li, and Philip D Jones

Correspondence to: Qingxiang Li (liqingx5@mail.sysu.edu.cnv)

In this supplementary material, we report how to extend the daily maximum temperature time series for Tianjin from Jan 1 1887 to Aug 31 1890. From the information of reference data sources, only the Berkeley Earth daily maximum/minimum temperatures are available. So the maximum temperature Berkeley Earth-daily data corresponding to the site level Tianjin station is selected as the extension data. Two methods of standardized series and linear regression are both applied to the extension of the daily maximum series during the period of Jan 1 1887-Aug 31 1890 at Tianjin station (Si et al., 2017) and finally the reasonable extension series is determined by comparing the results of an error analysis between the two methods.

S1 Standardized series method

This method assumes that for all the stations in the same climate region, the meteorological information at a certain time point is similar to its anomaly from the multi-year average values. The method can be expressed as follows:

$$Z_{Berkeley} = \frac{(x_i - \overline{x_i})}{s_i} \quad \text{(S1)}$$

$$y_{Tj} = Z_{Berkeley} s_{Tj} + \overline{y_{Tj}} \quad \text{(S2)}$$

Where, $Z_{Berkeley}$ is the Berkeley Earth-daily standardized series, x_i is the Berkeley Earth-daily data for some date, $\overline{x_i}$ and s_i are the mean and standard deviation of Berkeley Earth-daily data for some

date during 1961-1990. y_{Tj} is the fitted daily series at Tianjin station while $\overline{y_{Tj}}$ and s_{Tj} are the mean and standard deviation of observed Tianjin-daily data at some date covering 1961-1990.

S2 Linear regression method

Linear regression method is used to interpolate the candidate time series ($\widehat{y_i}$) by establishing a quantitative linear statistical relationship between the candidate station and its neighbour. It can be expressed as follows:

$$\widehat{y_i} = b_0 + bx_i \quad \text{(S3)}$$

In which, regression coefficients of b_0 and b are obtained by the following formulas,

$$\begin{cases} nb_0 + b\sum_{j=1}^{n} x_j = \sum_{j=1}^{n} y_j \\ b_0\sum_{j=1}^{n} x_j + b\sum_{j=1}^{n} x_j^2 = \sum_{j=1}^{n} x_j y_j \end{cases} \quad \text{(S4)}$$

Where $n = 30$, x_j is Berkeley Earth-daily data, and y_j is observed Tianjin-daily data at the same date. Similar to standardized series method (described in the earlier subsection), a quantitative linear statistical relationship is also established by multi-year average values during 1961-1990.

The significance test of the regression equation is carried out using the *F* statistic defined in Eq. (S5), to investigate whether the established linear statistical relationship defined in Eq.(S3) is significant. The *F* statistic follows an *F*-distribution with numerator degree of freedom of 1 and denominator degree of freedom $(n-2)$. If $F > F_{0.05}$, then the linear relation of regression equation is considered to be significant at 5% significance level.

$$F = \frac{\frac{U}{1}}{\frac{Q}{(n-2)}} \quad \text{(S5)}$$

$$U = \sum_{j=1}^{n}\left(\widehat{y_j} - \overline{y}\right)^2 \quad \text{(S6)}$$

$$Q = \sum_{j=1}^{n}\left(y_j - \widehat{y_j}\right)^2 \quad \text{(S7)}$$

Where $n = 30$. In Eq. (S6) - (S7), $\widehat{y_j}$, y_j and $\overline{y}$ are the fitted daily data at Tianjin station, observed Tianjin-daily data and the mean of observed Tianjin-daily data at some date during 1961-1990, respectively.

At the same time, the significance of the regression coefficient is also tested to further confirm the credibility of established linear statistical relationship. The *t* statistic defined in Eq. (S8) satisfies a *t*-distribution with $(n-2)$ degrees of freedom. If $|t| > t_{0.05}$, then the linear relation is considered to be significant at the 5% significance level.

$$t = \frac{\frac{b}{\sqrt{c}}}{\sqrt{\frac{Q}{n-2}}} \quad \text{(S8)}$$

$$c = \left[\sum_{j=1}^{n}\left(x_j - \overline{x}\right)^2\right]^{-1} \quad \text{(S9)}$$

Where $n = 30$, b is one of the regression coefficients in Eq. (S4) and Q is calculated using Eq. (S7). In Eq. (S9), x_j is Berkeley Earth-daily data and $\overline{x}$ represents the mean of Berkeley Earth-daily series at some date during 1961 - 1990.

S3 Error analysis

In order to make the sample for error analysis large enough, the fitted daily maximum temperature series formed by the above two methods are estimated by comparing errors between the fitted and the actual observations over the same time period. The error assessment indicators include the Standard Mean Square Error ($SMSE$), Standard Error (*SE*) and the proportion of the differences in the range of ±0.5 ℃ between the fitted and observed data (*P*). These indicators have been used in construction of the monthly surface air temperature series for Baoding city during 1913 - 2014 (Si et al., 2017).

Standard Mean Square Error ($SMSE$) is similar to Root Mean Square Error (*RMSE*), which is used to measure the deviation between fitted and true values. But in order to eliminate the differences from

units of different elements and climate change of the same element at different site location as well as easy comparison of different elements, the $SMSE$ is superior indicator compared to *RMSE* (Huang et al., 2004). $SMSE$ is shown as follows:

$$SMSE = \sqrt{\sum_{i=1}^{m}\left(\frac{\hat{y}_i - y_i}{s}\right)^2} \qquad \text{(S10)}$$

Where, m is the sample size, $\hat{y}_i$ is the fitted daily data at Tianjin station and y_i is the actual observed Tianjin-daily data at the same date，similarly hereinafter. s is the standard deviation of observed Tianjin-daily data at the same date during 1961 - 1990.

Standard Error (*SE*) is an indicator inferring statistical reliability. The smaller the values of *SE* implies the closeness of the sample statistics to the population parameter, thereby makes the sample as ideal representative of the population and naturally therefore extrapolation of the population parameter using sample statistics became more reliable. It can be expressed as follows:

$$SE = \sqrt{\frac{1}{(m-1)}\sum_{i=1}^{m}\left(\varphi_i' - \overline{\varphi'}\right)^2} \qquad \text{(S11)}$$

In Eq. (S11), $\varphi_i' = \varphi_i - \overline{\varphi_i}$, $\overline{\varphi'} = \frac{1}{m}\sum_{i=1}^{m}\varphi_i'$, $\varphi_i = \hat{y}_i - y_i$

$$P = \frac{m_p}{m} \times 100\% \qquad \text{(S12)}$$

In Eq. (S12), m_p is the number of differences between fitted data and the actual observations within ±0.5 ℃.

Figure S1 displays the errors between fitted and actual observations of daily maximum temperature series calculated using Eq. (S10) - (S12) in the period of Jan 1 1891 - Dec 31 2018 at Tianjin station.

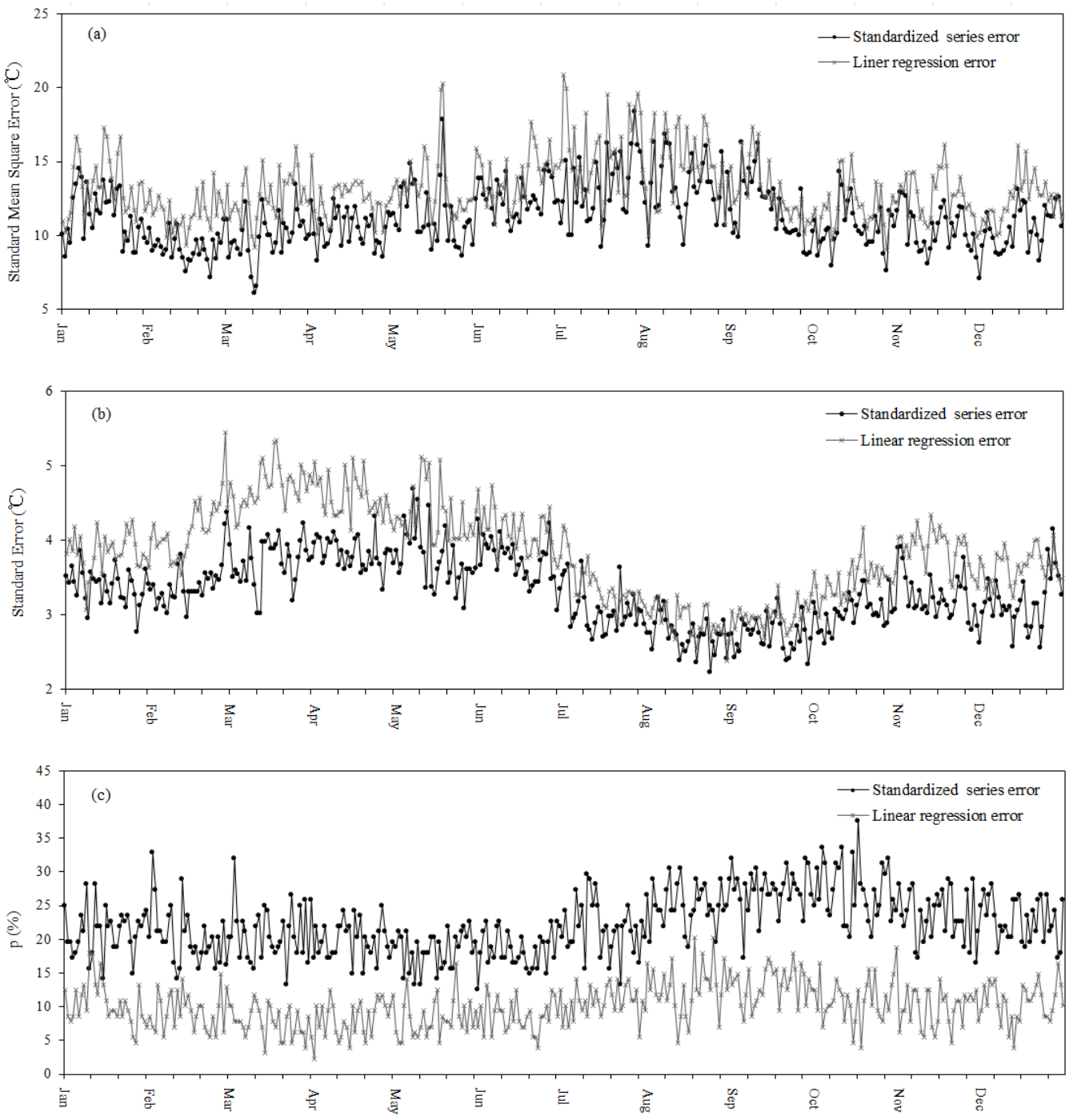

Figure S1. The indices of Standard Mean Square Error (*SMSE*) (a), Standard Error (*SE*) (b), and *P* (c) for the two fitted daily maximum temperatures at Tianjin station from Jan 1 1891 to Dec 31 2018.

Table S1. The statistics of *SMSE*, *SE*, and *P* for the two fitted data during Jan 1 1891-Dec 31 2018.

SMSE (℃)		*SE* (℃)		*P* (%)	
standardized	linear	standardized	linear	standardized	linear

	series	regression	series	regression	series	regression
Maximum	18.4	20.9	4.7	5.4	37.5	20.3
Minimum	6.1	9.0	2.2	2.4	12.5	2.3
Median	11.0	12.8	3.3	3.8	21.9	10.2

As shown in Fig. S1, it is obvious that errors are different in the case of the two fitted daily data at Tianjin station. Both of them are above 93% of *SMSE* (Fig. S1a) and *SE* (Fig. S1b) using the standardized series method which appears to be smaller than that from the linear regression method covering 1891-2018. Index *P* in both the cases is over 99% of difference ratio within ±0.5 ℃ from standardized series method which is higher than that from the linear regression method (Fig. S1c). As shown in Table S1, standardized series exhibits smaller error which agrees well with the statistics *SMSE*, *SE* and *P* in Fig. S1. The maximum, minimum and median values of *SMSE* and *SE* from the standardized series method are all smaller than those from the linear regression method. For index *P*, the median of difference ratio within ±0.5 ℃ between the standardized series and the actual observation is 21.9%, but it is just 10.2% for the linear regression fitted series. The minimum value of index *P* from the linear regression method is only 2.3%, while it is 12.5% from the standardized series method. Hence based on error analysis, the fitted daily maximum temperature series by the standardized series method is selected for the extension during the periods from Jan 1 1887 to Aug 31 1890 at Tianjin station. However, the daily minimum series still begins with the date on Sep 1 1890 due to scarcity of reference data sources.