

Mathematical Proof of the Synthetic Running Correlation Coefficient and Its Ability to Reflect Temporal Variations in Correlation

ZHAO Jinping^{1), 2), *}, CAO Yong¹⁾, SHI Yanyue³⁾, and WANG Xin¹⁾

1) College of Oceanic and Atmospheric Sciences, Ocean University of China, Qingdao 266100, China

2) Physical Oceanography Laboratory, Ministry of Education, Qingdao 266100, China

3) School of Mathematical Sciences, Ocean University of China, Qingdao 266100, China

(Received November 5, 2020; revised January 20, 2021; accepted January 29, 2021)

© Ocean University of China, Science Press and Springer-Verlag GmbH Germany 2021

Abstract The running correlation coefficient (RCC) is useful for capturing temporal variations in correlations between two time series. The local running correlation coefficient (LRCC) is a widely used algorithm that directly applies the Pearson correlation to a time window. A new algorithm called synthetic running correlation coefficient (SRCC) was proposed in 2018 and proven to be reasonable and usable; however, this algorithm lacks a theoretical demonstration. In this paper, SRCC is proven theoretically. RCC is only meaningful when its values at different times can be compared. First, the global means are proven to be the unique standard quantities for comparison. SRCC is the only RCC that satisfies the comparability criterion. The relationship between LRCC and SRCC is derived using statistical methods, and SRCC is obtained by adding a constraint condition to the LRCC algorithm. Dividing the temporal fluctuations into high- and low-frequency signals reveals that LRCC only reflects the correlation of high-frequency signals; by contrast, SRCC reflects the correlations of high- and low-frequency signals simultaneously. Therefore, SRCC is the appropriate method for calculating RCCs.

Key words running correlation coefficient; synthetic running correlation coefficient; time window; comparability; standard value

1 Introduction

Correlation describes the degree of consistency between two time series. The correlation coefficient (CC) is an important statistical quantity (Pearson, 1896) that reflects the overall correlation between two data series. Because knowledge of temporal variations in correlation may sometimes be useful, the running correlation coefficient (RCC) was proposed to reflect varying correlations (Kuznets, 1928).

In most cases, RCC simply applies the CC to a pair of data pieces of the complete dataset. The length of the data piece is called the time window, and the window is moved stepwise to obtain the RCC (*e.g.*, Kodera, 1993). RCC is a time series with values greater than -1.0 and less than $+1.0$. The RCC obtained by this method was called local running correlation coefficient (LRCC) by Zhao *et al.* (2018). LRCC is widely used to study varying correlations between two time series, such as the correlations of the Arctic Oscillation and sea level pressure (Zhao *et al.*, 2006), water transport in the Labrador Sea and the North Atlantic Oscillation (Varotsou *et al.*, 2015), atmospheric

circulation and air temperature (Hynčica and Huth, 2020), Australian rainfall and El Niño (Brown *et al.*, 2016), solar variability and paleoclimate records (Turner *et al.*, 2016), equatorial quasi-biennial oscillations and stratospheric temperatures (Kodera, 1993; Soukhearev, 1997), solar cycle (Salby *et al.*, 1997), and solar UV irradiance (Elias and Zossi de Artigas, 2003).

Zhao *et al.* (2018) observed that the LRCC algorithm uses the mean values determined by the data within the time window, which means the mean values also vary with time. LRCC only reflects the correlation of anomalies corresponding to the means and does not reflect the correlation between varying means. Therefore, the authors proposed a new algorithm for RCC called synthetic running correlation coefficient (SRCC). SRCC reflects correlations for anomalies and means using the global means calculated for the whole dataset. The relationship between LRCC and SRCC could be derived to illustrate the consistency and differences of these two algorithms. Some authors, such as Zhao *et al.* (2019) and Ji and Zhao (2019), have obtained remarkable results using SRCC.

The definition, calculation method, application examples, and physical significance of SRCC were addressed by Zhao *et al.* (2018) in effort to prove that the method is valid and credible. However, SRCC still lacks the support

* Corresponding author. Tel: 0086-532-66782096

E-mail: jpzhao@ouc.edu.cn

of mathematical theory. In the present study, the validity of SRCC is demonstrated theoretically. The analysis based on geometric and physical significances proves that SRCC is an appropriate method for measuring varying correlations. In Section 2, the background of the two RCCs is introduced. Comparability as the basic requirement for RCCs is then proposed in Section 3. A mathematical demonstration of SRCC is given in Section 4. Finally, the physical significance of SRCC is discussed in Section 5.

2 Background of Running Correlation Algorithms

All of the CCs discussed in this study are linear correlations; nonlinear correlations (*e.g.*, Geng *et al.*, 2018) are not discussed. A simple CC defined as the Pearson product-moment correlation coefficient (Pearson, 1896) was first introduced by Francis Galton (Galton, 1888) for linear correlation. The more common form of this CC was developed and applied by Karl Pearson (Pearson, 1938; Merrington *et al.*, 1983). For two time series of data lengths N with equal intervals:

$$\begin{cases} X = \{x_k : k = 1, 2, \dots, N\} \\ Y = \{y_k : k = 1, 2, \dots, N\} \end{cases} \quad (1)$$

The simple correlation coefficient R is written as follows:

$$R = \frac{\sum_{k=1}^N (x_k - \bar{X})(y_k - \bar{Y})}{\sqrt{\sum_{k=1}^N (x_k - \bar{X})^2} \sqrt{\sum_{k=1}^N (y_k - \bar{Y})^2}} \quad (2)$$

where

$$\bar{X} = \frac{1}{N} \sum_{k=1}^N x_k \quad \text{and} \quad \bar{Y} = \frac{1}{N} \sum_{k=1}^N y_k \quad (3)$$

are the means calculated based on all data; thus, these means are called ‘global means’. R in Eq. (2) calculated from all of the data is referred to as the ‘global CC’.

RCC is a useful tool for understanding temporal variations in the correlation between two time series. The CC of two time series centered at i is:

$$R_r(i) = \frac{\sum_{k=i-n}^{i+n} (x_k - \bar{X}_i)(y_k - \bar{Y}_i)}{\sqrt{\sum_{k=i-n}^{i+n} (x_k - \bar{X}_i)^2} \sqrt{\sum_{k=i-n}^{i+n} (y_k - \bar{Y}_i)^2}} \quad (4)$$

$i = 1+n, \dots, N-n,$

where $i \in [1+n, N-n]$ and the time window is $[i-n, i+n]$; that is:

$$\begin{cases} X_i = \{x_k : k = i-n, i-n+1, \dots, i+n-1, i+n\} \\ Y_i = \{y_k : k = i-n, i-n+1, \dots, i+n-1, i+n\} \end{cases} \quad (5)$$

An RCC is obtained by moving the window i . $R_r(i)$ is LRCC to distinguish it from other RCCs. The means of LRCC are obtained from the data within the time window:

$$\bar{X}_i = \frac{1}{2n+1} \sum_{k=i-n}^{i+n} x_k, \quad \bar{Y}_i = \frac{1}{2n+1} \sum_{k=i-n}^{i+n} y_k \quad (6)$$

Hereafter, these means are referred to as ‘local means’.

The algorithm for LRCC in Eq. (4) is the direct application of the definition of the global CC. This algorithm only changes the data length with the limit of the time window. The algorithm assumes that the definition used for the global CC could also be applied to the RCC, but no theoretical evidence proving that this direct application is reasonable has been obtained. Zhao *et al.* (2018) indicated that the means in Eq. (6) also vary with time. LRCC reflects only the correlation between two anomalies within the time window and does not capture the contributions of two varying means. Some important signals contained in the means are clearly missing, which raises further issues whether LRCC reflects the significance of statistics despite ignoring variations in the local means. The RCCs in different time windows should be comparable to each other; however, $R_r(i)$ is obtained only from the data within the time window and independent of the data of other time windows. Thus, LRCCs in different windows lack common information and, therefore, are not comparable.

Zhao *et al.* (2018) identified this problem and proposed a new algorithm to calculate RCC:

$$R_s(i) = \frac{\sum_{k=i-n}^{i+n} (x_k - \bar{X})(y_k - \bar{Y})}{\sqrt{\sum_{k=i-n}^{i+n} (x_k - \bar{X})^2} \sqrt{\sum_{k=i-n}^{i+n} (y_k - \bar{Y})^2}} \quad (7)$$

$i = 1+n, \dots, N-n,$

where \bar{X} and \bar{Y} are the global means as defined by Eq. (3). This RCC is referred to as the SRCC. $R_s(i) \in [-1, 1]$ can easily be verified using the Cauchy inequality:

$$\left| \sum_{k=i-n}^{i+n} (x_k - \bar{X})(y_k - \bar{Y}) \right| \leq \sum_{k=i-n}^{i+n} |(x_k - \bar{X})(y_k - \bar{Y})| \leq \sqrt{\sum_{k=i-n}^{i+n} (x_k - \bar{X})^2} \sqrt{\sum_{k=i-n}^{i+n} (y_k - \bar{Y})^2} \quad (8)$$

Although SRCC was first proposed by Zhao *et al.* (2018), this algorithm has actually existed for a long time in the following form:

$$R'_s(i) = \frac{\sum_{k=i-n}^{i+n} x_k y_k}{\sqrt{\sum_{k=i-n}^{i+n} (x_k)^2} \sqrt{\sum_{k=i-n}^{i+n} (y_k)^2}} \quad (9)$$

$i = 1+n, \dots, N-n,$

where the means of the two data series have been removed beforehand and the calculation in Eq. (9) does not include

the means. This procedure is equivalent to adopting the global means. Therefore, the algorithm in Eq. (9) is equivalent to SRCC.

Zhao *et al.* (2018) attempted to prove which RCC is better by proposing and adopting a criterion, that is, the temporal average of the RCC should be close to the global CC. In fact, the average RCC is not exactly equal to the global CC because the amounts of data used for both algorithms differ but could be very close to each other. In general, the temporal average of SRCC is close to the global CC; by contrast, in most cases, the temporal average of LRCC cannot fulfill this criterion. Thus, according to the temporal average criterion, SRCC is better than LRCC for measuring running correlations.

3 Comparability of the RCC Values of Different Windows

An RCC is meaningful only when its values at different times are comparable with each other. For example, if the temperatures of two cities are compared, a large RCC should reflect a consistent variation, and a low one should reflect lower consistency. In this situation, the RCCs are comparable. The standard for each city, which should be an unchanged constant, must be established in advance to meet the requirement of comparability. In the above example, the mean air temperature of each city is used as the comparison standard for cold or warm events, and these events should differ between southern and northern cities. If the standard changes over time, for example, if different temperatures for summer and winter are selected as standards, the RCCs in winter and summer would not be comparable.

Mathematically, a physical component can be decomposed into standard and comparative quantities. The standard quantities should be two unchanged constants not involved in the comparison, and the comparison is conducted on the two comparative quantities (Burdun and Markov, 1972). The means of the temperatures are qualified standard quantities for indicating whether the environment is warmer or cooler at a given time. The comparison is applied to the anomalies of the temperature obtained by subtracting the standard quantities. Here we derive the standard quantities mathematically.

According to Eq. (2), the global means (\bar{X}, \bar{Y}) are the qualified standards for comparison. Therefore, the global CC of the two time series meets the above comparability requirements. When Eq. (2) is directly applied to a time window, the resultant local means (X_i, Y_i) only use the data in the time window, as shown in Eq. (6). In general, the local means are different when $i \neq j$, so LRCCs in different time windows use different standard quantities and, thus, do not meet the comparison criteria. According to the definition of Eq. (7), SRCC meets the requirements of comparability because the global means (\bar{X}, \bar{Y}) are constant standards.

Even if the global means (\bar{X}, \bar{Y}) are replaced by any real constant numbers (\bar{X}_0, \bar{Y}_0) , Eq. (7) holds true, which means the constant chosen as the standard is somewhat

arbitrary, and the RCCs obtained using different standard values will differ. Thus, more physical constraints must be applied to obtain a unique RCC. Let us verify that (\bar{X}_0, \bar{Y}_0) equals (\bar{X}, \bar{Y}) in SRCC.

Let (\bar{X}_0, \bar{Y}_0) in Eq. (7) be any arbitrary values. Then, a new time series can be expressed as follows:

$$\begin{cases} F_{xi} = \frac{1}{2n+1} \sum_{k=i-n}^{i+n} (x_k - \bar{X}_0) \\ F_{yi} = \frac{1}{2n+1} \sum_{k=i-n}^{i+n} (y_k - \bar{Y}_0) \end{cases}, i = 1+n, \dots, N-n. \quad (10)$$

The physical definitions of F_{xi} and F_{yi} are the average deviations relative to the standard values (\bar{X}_0, \bar{Y}_0) in a time window, which vary with time and the values of (\bar{X}_0, \bar{Y}_0) .

That the means of F_{xi} and F_{yi} are equal to zero is proposed here as a new constraint:

$$\begin{cases} \frac{1}{N} \sum_{i=1}^N F_{xi} = \frac{1}{N(2n+1)} \sum_{i=1}^N \left[\sum_{k=i-n}^{i+n} (x_k - \bar{X}_0) \right] = 0 \\ \frac{1}{N} \sum_{i=1}^N F_{yi} = \frac{1}{N(2n+1)} \sum_{i=1}^N \left[\sum_{k=i-n}^{i+n} (y_k - \bar{Y}_0) \right] = 0 \end{cases}. \quad (11)$$

The rationality of this constraint is that the averaged deviation should be equal to zero regardless of the chosen data window n . That is, the positive and negative parts of the averaged deviation should be completely equal. Then, (\bar{X}_0, \bar{Y}_0) can be uniquely determined:

$$\begin{cases} \bar{X}_0 = \frac{1}{N(2n+1)} \sum_{i=1}^N \sum_{k=i-n}^{i+n} x_k \\ \bar{Y}_0 = \frac{1}{N(2n+1)} \sum_{i=1}^N \sum_{k=i-n}^{i+n} y_k \end{cases}. \quad (12)$$

The significance of Eq. (12) is that the standard values equal the averages of the local means in the whole data domain $[1, N]$. For clarity, the global means (\bar{X}, \bar{Y}) are incorporated into Eq. (12) so that the corresponding relations become:

$$\begin{cases} \bar{X}_0 = \bar{X} - \frac{1}{N(2n+1)} \sum_{i=1}^N \left[(2n+1)x_i - \sum_{k=i-n}^{i+n} x_k \right] \\ \bar{Y}_0 = \bar{Y} - \frac{1}{N(2n+1)} \sum_{i=1}^N \left[(2n+1)y_i - \sum_{k=i-n}^{i+n} y_k \right] \end{cases}. \quad (13)$$

Notice that the range of k is $[1, N]$ and that of i is $[1+n, N-n]$ in Eq. (10). The range of i must be extended to $[1, N]$. Therefore, the time series $X = \{x_k; k=1, 2, \dots, N\}$ and $Y = \{y_k; k=1, 2, \dots, N\}$ must be extended beyond both limits:

$$\begin{cases} \{\dots, x_{N-1}, x_N, x_1, \dots, x_{N-1}, x_N, x_1, \dots\} \\ \{\dots, y_{N-1}, y_N, y_1, \dots, y_{N-1}, y_N, y_1, \dots\} \end{cases}$$

Although any extension could be applied because the

extended data do not affect the SRCC calculation in Eq. (7), the extension must satisfy the condition that the mean of the extended data equals the mean of the original data series. This condition is necessary to adopt the same signal of the original data in the extended data. The ideal method is to perform a periodic extension, which is equivalent to a circular extension (e.g., Woods and Oneil, 1986). The general periodic extension is a sinusoidal extension (e.g., Huybrechs, 2010); it may also be a cosinoidal extension, such as the mirror extension of Zhao and Huang (2001).

Because the means are unchanged after data extension, the sums of the second terms on the right-hand side of Eq. (13) are always zero. Thus, the unique standard, which is different from the arbitrary standard, is:

$$\begin{cases} \bar{X}_0 = \bar{X} \\ \bar{Y}_0 = \bar{Y} \end{cases} \quad (14)$$

The global means (\bar{X}, \bar{Y}) are proven to be the unique standard quantities satisfying the condition of Eq. (11). Although the standards for comparison can be arbitrarily

selected, only one pair of standards, namely the global means, satisfies the condition that the averaged deviations are equal to zero. Therefore, the SRCC algorithm given in Eq. (7) is the only RCC expression that satisfies Eq. (11).

For further comparison, an example with two randomly generated white noise data series, $f_1(t)$ and $f_2(t)$, for the time range 0–500 is shown in Figs.1a and 1b. The global CC between the two white noise data series is zero, and the average values of LRCC and SRCC are 0.01 and 0.02, respectively. LRCC and SRCC for these two white noise data are quite similar, as shown in Figs.1c and 1d, respectively.

If a constant value is added between 150 and 350, the two time series are defined as:

$$\begin{cases} A_1(t) = f_1(t) + a_1 \\ A_2(t) = f_2(t) + a_2 \end{cases} \quad (15)$$

The constants a_1 and a_2 are set as:

$$a_1 = \begin{cases} 3 & 200 \leq t \leq 300 \\ 0 & \text{other time} \end{cases}, \quad a_2 = \begin{cases} 2 & 200 \leq t \leq 300 \\ 0 & \text{other time} \end{cases} \quad (16)$$

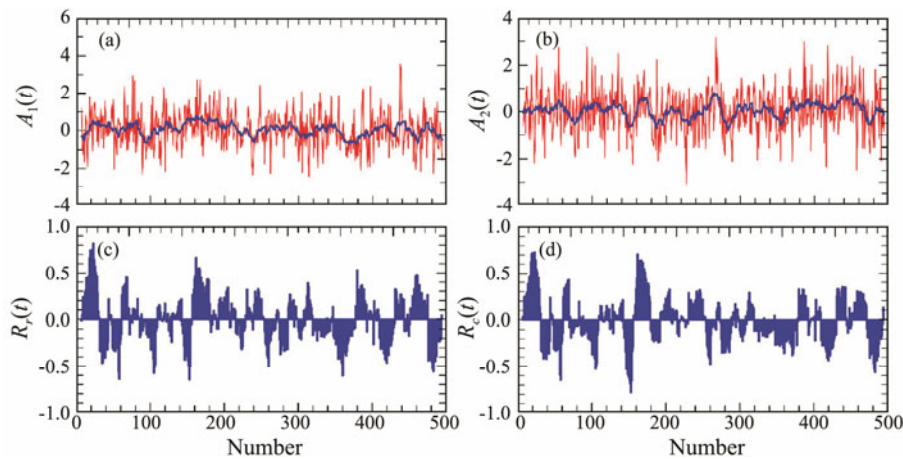


Fig.1 Two running correlation coefficients of a white noise data series. (a) and (b), Two series of white noise (red lines) and the local means (blue lines); (c), LRCC; (d), SRCC.

The new time series are shown in Figs.2a and 2b. The

global correlation coefficient is 0.548. RCCs in the time

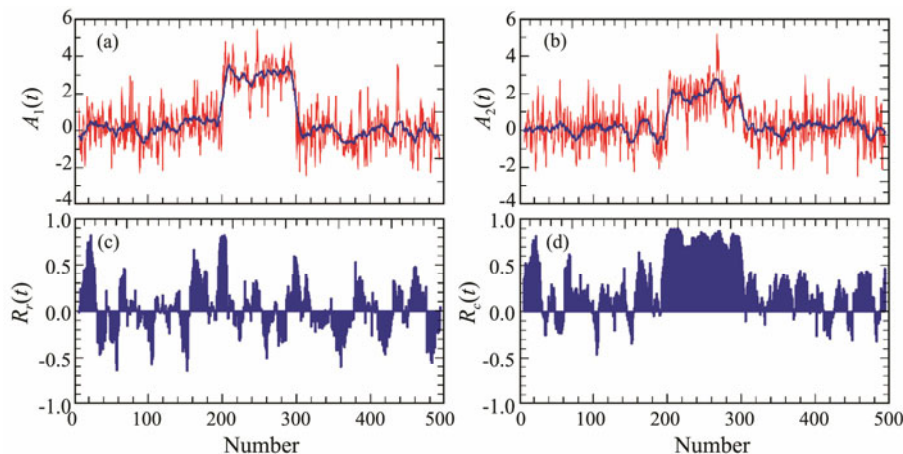


Fig.2 Two running correlation coefficients. (a) and (b), Two series defined by Eq. (15) (red lines) and the local means (blue lines); (c), LRCC; (d), SRCC.

interval with non-zero constants may be expected to show high correlations. SRCC remarkably increases in the presence of these constants and shows an average value of 0.518. By comparison, LRCC changes minimally, as shown in Fig.2c, with an average value of only 0.030. Therefore, whereas SRCC is a suitable metric that could reflect the expected running correlation, LRCC appears to lose important information.

Therefore, from the perspective of comparability, an invariant value must be chosen as the standard value. Because the local mean varies with time, LRCC is not a comparable CC. The global mean is a qualified and unique standard value, and SRCC is the unique RCC satisfying the comparability criterion.

4 Mathematical Difference Between LRCC and SRCC

Although SRCC has been proven to be a qualified RCC by Zhao *et al.* (2018) and a unique form of an RCC with comparability as demonstrated in Section 3, mathematically verifying that SRCC is a unique form of RCC based on the original geometric definition of statistical quantities remains necessary.

In the linear correlation framework, a linear correlation can be used to calculate the CC. Consider a pair of time series in Eq. (1) with data length N in a scatterplot in x - y space and draw a straight line through this cloud of points that approaches all of the points ‘as closely as possible’.

If $a+bx$ is used to estimate y and $c+dy$ is used to estimate x , then the deviations of the two lines from the data are:

$$\begin{cases} Q(a,b) = \sum_{k=1}^N [y_k - (a + bx_k)]^2 \\ Q(c,d) = \sum_{k=1}^N [x_k - (c + dy_k)]^2 \end{cases} \quad (17)$$

Calculating the minimum value of $Q(a, b)$ by the least-squares method yields:

$$\begin{cases} b = \frac{\sum_{k=1}^N (x_k - \bar{X})(y_k - \bar{Y})}{\sum_{k=1}^N (x_k - \bar{X})^2} \\ d = \frac{\sum_{k=1}^N (x_k - \bar{X})(y_k - \bar{Y})}{\sum_{k=1}^N (y_k - \bar{Y})^2} \end{cases} \quad (18)$$

The global correlation coefficient R can be expressed as:

$$R^2 = bd = \frac{\left[\sum_{k=1}^N (x_k - \bar{X})(y_k - \bar{Y}) \right]^2}{\left[\sum_{k=1}^N (x_k - \bar{X})^2 \right] \left[\sum_{k=1}^N (y_k - \bar{Y})^2 \right]}, \quad (19)$$

which is identical in form to Eq. (2). Fig.3 shows that the two empirical regression lines pass through the global means (\bar{X}, \bar{Y}) .

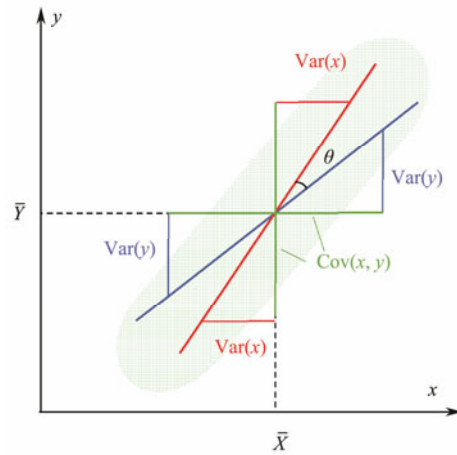


Fig.3 Geometric interpretation of the linear correlation coefficients (redrawn from Schmid, 1947). The shadow area represents the scatterplot of the data, and θ is the angle between the regression lines.

LRCC expresses the correlation of the data series in the time window $[i-n, i+n]$ as shown in Eq. (5). The linear regression of the data is calculated using a similar method. Here, $a'+b'x$ is used to estimate y , $c'+d'y$ is used to estimate x , and the deviations of the two lines from the data are:

$$\begin{cases} Q_i(a', b') = \sum_{k=i-n}^{i+n} [y_k - (a' + b'x_k)]^2 \\ Q_i(c', d') = \sum_{k=i-n}^{i+n} [x_k - (c' + d'y_k)]^2 \end{cases} \quad (20)$$

where quantities marked ‘ $'$ ’ are constants related to the length of the time window. The least-squares method can be used to calculate b' and d' as follows:

$$\begin{cases} b' = \frac{\sum_{k=i-n}^{i+n} (x_k - \bar{X}_i)(y_k - \bar{Y}_i)}{\sum_{k=i-n}^{i+n} (x_k - \bar{X}_i)^2} \\ d' = \frac{\sum_{k=i-n}^{i+n} (x_k - \bar{X}_i)(y_k - \bar{Y}_i)}{\sum_{k=i-n}^{i+n} (y_k - \bar{Y}_i)^2} \end{cases} \quad (21)$$

where (\bar{X}_i, \bar{Y}_i) is the local mean of the i time window. Thus,

$$R_r^2(i) = b'd' = \frac{\left[\sum_{k=i-n}^{i+n} (x_k - \bar{X}_i)(y_k - \bar{Y}_i) \right]^2}{\sum_{k=i-n}^{i+n} (x_k - \bar{X}_i)^2 \sum_{k=i-n}^{i+n} (y_k - \bar{Y}_i)^2}, \quad (22)$$

which is identical to the expression of LRCC in Eq. (4). Eq. (22) is obtained from the slope of the lines fitted by the data in the window. In general, $(X_{t_1}, Y_{t_1}) \neq (X_{t_2}, Y_{t_2})$ when $t_1 \neq t_2$. Geometrically, the cross points (equal to the local means) of the regression lines for different time windows appear at different positions, as shown in Fig.4a. Because a cross point corresponds to a standard and only RCCs in time windows at the same cross point are comparable, the values of LRCC at different i cannot be compared with each other.

Because the regression lines of the global correlation cross the global means (\bar{X}, \bar{Y}) , this correlation may also be a constraint for RCC in Eq. (20), that is, the empirical regression lines fitted to all time windows must cross the

point (\bar{X}, \bar{Y}) .

$$\begin{cases} \bar{Y} = a' + b'\bar{X} \\ \bar{X} = c' + d'\bar{Y} \end{cases} \quad (23)$$

Eq. (23) should be an additional condition for the deviation of the two regression lines in Eq. (20). Substituting $a' = \bar{Y} - b'\bar{X}$, $c' = \bar{X} - d'\bar{Y}$ into Eq. (20) yields:

$$\begin{cases} Q_i(b') = \sum_{k=i-n}^{i+n} [(y_k - \bar{Y}) - b'(x_k - \bar{X})]^2 \\ Q_i(d') = \sum_{k=i-n}^{i+n} [(x_k - \bar{X}) - d'(y_k - \bar{Y})]^2 \end{cases} \quad (24)$$

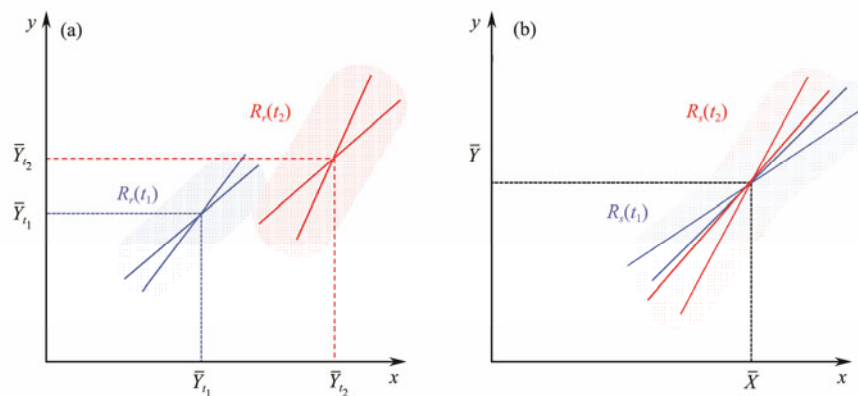


Fig.4 Geometric interpretation of two RCCs at different times t_1 and t_2 . (a), LRCC with different means $(\bar{X}_{t_1}, \bar{Y}_{t_1})$ and $(\bar{X}_{t_2}, \bar{Y}_{t_2})$; (b), SRCC with the same means (\bar{X}, \bar{Y}) . The shadow area represents the scatterplot of the data.

When the minimum $Q_i(b')$ and $Q_i(d')$ are calculated by the least-squares method:

$$\begin{cases} b' = \frac{\sum_{k=i-n}^{i+n} (x_k - \bar{X})(y_k - \bar{Y})}{\sum_{k=i-n}^{i+n} (x_k - \bar{X})^2} \\ d' = \frac{\sum_{k=i-n}^{i+n} (x_k - \bar{X})(y_k - \bar{Y})}{\sum_{k=i-n}^{i+n} (y_k - \bar{Y})^2} \end{cases} \quad (25)$$

Thus:

$$R_s^2(i) = b'd' = \frac{\left[\sum_{k=i-n}^{i+n} (x_k - \bar{X})(y_k - \bar{Y}) \right]^2}{\sum_{k=i-n}^{i+n} (x_k - \bar{X})^2 \sum_{k=i-n}^{i+n} (y_k - \bar{Y})^2} \quad (26)$$

which is the expression for SRCC.

Eq. (26) clearly shows that SRCC is simply obtained by adding a constraint condition, Eq. (23), when calculating LRCC. The geometric expression of SRCC is shown in Fig.4b. The regression lines of the data in all time windows cross the global means (\bar{X}, \bar{Y}) . The significance of

this result is that a unified frame of reference is established with (\bar{X}, \bar{Y}) , in which the variations in SRCC at different times can be effectively compared. This relationship also means that SRCC takes local and global information into account and, therefore, demonstrates the close connection of the local correlation with the global correlation.

5 Physical Significance of the SRCC

The comparability and geometric consistency of SRCC were demonstrated in Sections 3 and 4 to improve the understanding of the physical significance of SRCC. The following examples present the differences between LRCC and SRCC and explain the reasons behind these differences. Because the annual variation is generally the strongest signal in geoscience, all of the data series used in the following examples are averaged by a 12-point running mean to filter the annual signal.

1) Contribution of the means and the anomalies of SRCC

The relationship between SRCC $R_s(i)$ and LRCC $R_r(i)$ was simply expressed by Zhao *et al.* (2018) as follows:

$$R_s(i) = R_r(i) \cos \gamma_x \cos \gamma_y + \sin \gamma_x \sin \gamma_y, \quad (27)$$

where:

$$\left\{ \begin{aligned} \cos \gamma_x &= \frac{\sigma_{rx}(i)}{\sqrt{[\sigma_{rx}^2(i) + (\bar{X}_i - \bar{X})^2]}} \\ \sin \gamma_x &= \frac{\bar{X}_i - \bar{X}}{\sqrt{[\sigma_{rx}^2(i) + (\bar{X}_i - \bar{X})^2]}} \\ \cos \gamma_y &= \frac{\sigma_{ry}(i)}{\sqrt{[\sigma_{ry}^2(i) + (\bar{Y}_i - \bar{Y})^2]}} \\ \sin \gamma_y &= \frac{\bar{Y}_i - \bar{Y}}{\sqrt{[\sigma_{ry}^2(i) + (\bar{Y}_i - \bar{Y})^2]}} \end{aligned} \right. \quad (28)$$

where $\bar{X}_i - \bar{X}$ and $\bar{Y}_i - \bar{Y}$ are the mean differences between the local and global means and $\sigma_{rx}(i)$ and $\sigma_{ry}(i)$ are the local variances defined as follows:

$$\left\{ \begin{aligned} \sigma_{rx}^2(i) &= \frac{1}{2n+1} \sum_{k=i-n}^{i+n} (x_k - \bar{X}_i)^2 \\ \sigma_{ry}^2(i) &= \frac{1}{2n+1} \sum_{k=i-n}^{i+n} (y_k - \bar{Y}_i)^2 \end{aligned} \right. \quad (29)$$

Eq. (27) reveals that SRCC comprises $R_r(t)$ and 1 with certain weights. The weight of $R_r(t)$ is $\cos\gamma_x\cos\gamma_y$ (cosine-weight), and the weight of 1 is $\sin\gamma_x\sin\gamma_y$ (sine-weight). A larger variance benefits the cosine-weight, and a larger mean difference benefits the sine-weight. When the mean difference is zero in extreme cases, the two correlation coefficients are equal; by contrast, when the variance of the anomaly approaches zero, SRCC equals 1.

As an example, the averaged air temperature anomalies for the North Atlantic at 2 m and 500hPa and their means are shown in Figs.5a and 5b. LRCC presents a high-frequency variation (Fig.5c), whereas SRCC shows a positively dominant running correlation (Fig.5d). Figs.5e and 5f reveal that sine-right is dominant in most time windows and that cosine-right is only apparent in some years. When cosine-right is apparent, SRCC becomes weak or opposite; otherwise, it is strong and close to 1. When the variance is dominant, the anomalous variation is dominant, and the variation of the mean is neglected. When the mean difference is dominant, the variations in anomalies are not important. This example shows that SRCC reflects the combined effects of the variance and mean difference simultaneously.

2) Contribution of low- and high-frequency signals

Besides the contributions of the means and anomalies, the double-frequency signal is another factor that decisively impacts LRCC and SRCC. Although fluctuations may contain various frequencies, all of the frequencies considered in the present example can be roughly divided into two groups: one with a period shorter than the time window (high frequency) and the other with a period longer than the time window (low frequency). LRCC usually represents the correlation between high-frequency signals because the low-frequency signals included in the local means are removed from the calculations. By contrast, SRCC still considers the correlation of low-frequency signals. For example, the Arctic Oscillation Index (Fig.6a) and the latent heat flux in the Greenland Sea (Fig.6b) are compared

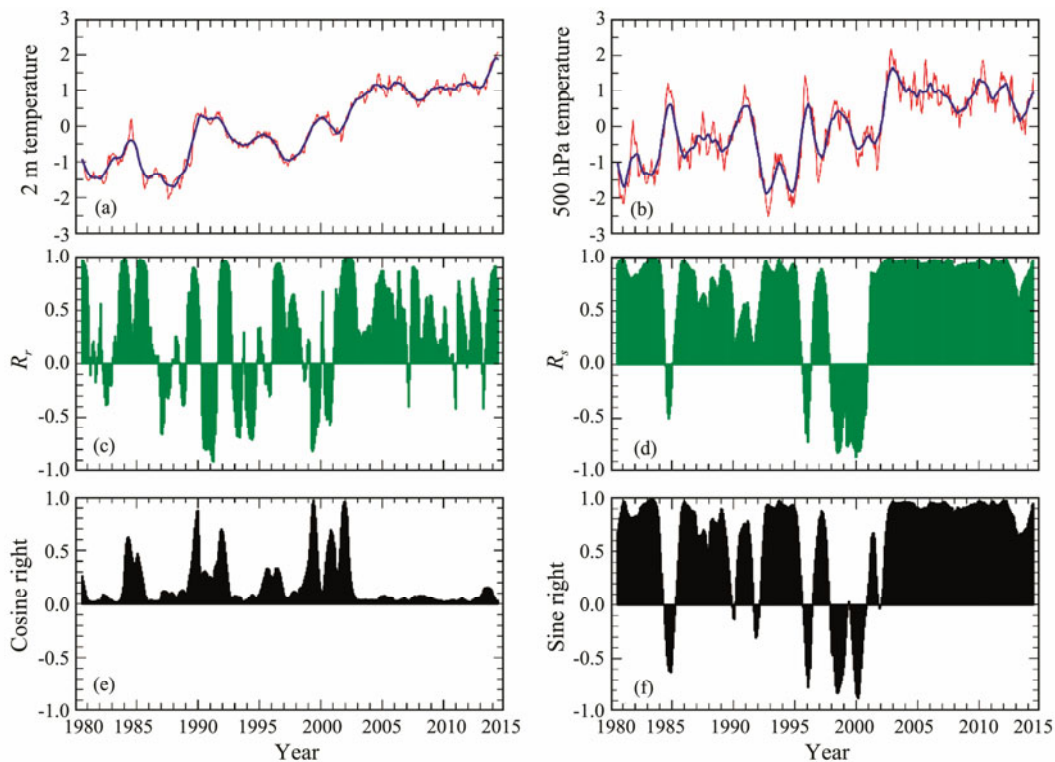


Fig.5 Two running correlation coefficients between 2 m and 500hPa air temperature anomalies averaged for the North Atlantic over the period 1980–2015. The 2 m and 500hPa air temperature data are obtained from NCEP/NCAR Reanalysis 1. (a), 2 m temperature anomalies (red line) and the local mean (blue line); (b), 500 hPa temperature anomalies (red line) and the local mean (blue line); (c), LRCC $R_r(t)$; (d), SRCC $R_s(t)$; (e), cosine-right $\cos\gamma_x\cos\gamma_y$; (f), sine-right $\sin\gamma_x\sin\gamma_y$.

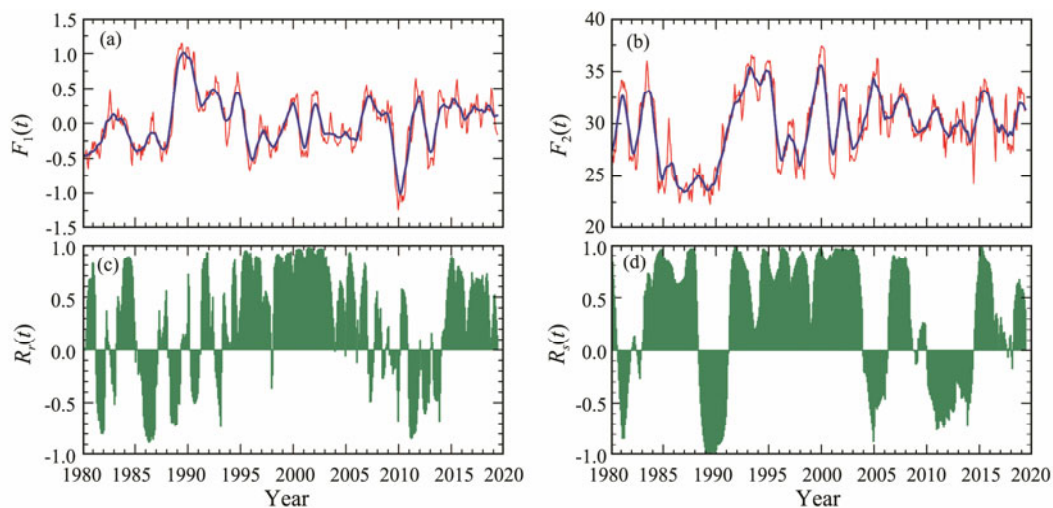


Fig.6 Running correlation dominated by high frequency signals. (a), Arctic Oscillation Index (red line) and the local mean (blue line); (b), Latent heat flux (unit: W m^{-2}) in the Greenland Sea (red line) and the local mean (blue line); (c), LRCC $R_r(t)$; (d), SRCC $R_s(t)$. The data of the Arctic oscillation index are obtained from the National Weather Service Climate Prediction Center of NOAA. The latent heat flux data are obtained from NCEP-DOE Reanalysis 2 from the National Centers of Environment Prediction.

The appearances of LRCC (Fig.6c) and SRCC (Fig.6d) are quite similar, which means high-frequency signals dominate the data set. This result indicates that the latent heat responds better to the high-frequency variations rather than the low-frequency variations of the Arctic Oscillation. According to Eq. (27), cosine-right is dominant in this example. However, the LRCC and SRCC results of 2 m and 500 hPa air temperatures are quite different, as shown in Fig.5. This finding indicates that the correlation of low-frequency signals is significant, with sine-right being dominant (Fig.5f).

Although SRCC includes the correlation of low-frequency signals, it does not filter out high-frequency signals. Therefore, LRCC includes only the correlation of high-frequency signals, while SRCC includes the correlations both high- and low-frequency signals.

The correlation of monthly sea level air pressures in Beijing and Guangzhou (Figs.7a and 7b) is described here as another example. The appearances of LRCC (Fig.7c) and SRCC (Fig.7d) are quite different. LRCC reflects high-frequency features with mostly positive correlations, whereas SRCC shows negative correlations, which means low-frequency signals are dominant in the data set. In this example, the global CC is -0.755 , consistent with the averaged SRCC. This result strongly exhibits the advantages of SRCC over LRCC. Indeed, in the present example, LRCC appears to have notable shortcomings.

SRCC presents the reverse variation in low-frequency, seesaw-like oscillations between Beijing and Guangzhou. This phenomenon is fairly similar to the North Atlantic Oscillation in that the seesaw-like oscillation appears in the surface air pressure difference between Iceland and Azores Island. LRCC cannot detect this low-frequency phenomenon.

The following example presents another correlation of

low-frequency phenomena revealed by SRCC. The latitudes of Beijing and New York are located 40°N . The central longitude of Beijing is 116°E , and that of New York is 74°W . Comparison of the surface temperatures of the two cities can help improve the understanding of their long-term variations. The monthly averaged temperature data for 1989–2017 are selected from the NCEP, and a 12-point running average is adopted to eliminate seasonal variations. The variations in temperatures are shown in Figs.8a and 8b. The temperature variations in the two cities are highly similar before 2009, and even their extremes occur in the same years. However, the temperature variations in the two cities are nearly contradictory from 2009 to 2016. The maximum value of a city's temperature often corresponds to the minimum value of the other city. SRCC (Fig.8d) reveals these characteristics well. Prior to 2009, positive correlations are dominant; after 2009, negative correlations are dominant. LRCC cannot accurately reflect this shift in correlation (Fig.8c).

In this example, positive correlation was the regular situation and showed the consistent low-frequency variation of global air temperature. The negative correlation obtained in 2009–2016 is an abnormality that can be explained by Arctic amplification (Francis and Vavrus, 2012). Arctic amplification has a sizable impact on the climate of the mid-latitudes, such as those seen in the severe winters experienced in New York (2009–2013) and Beijing (2014–2015), as shown in Figs.8a and 8b. Francis and Vavrus (2012) explained the occurrence of severe winters using Rossby wave theory. Specifically, the amplitude of the Rossby wave markedly increases as a result of Arctic warming, which allows the cold air in higher latitudes to flow out to the mid-latitude areas along the fronts. The negative correlation observed in 2009–2016 indicates that the cold air phenomenon occurs alternately in New York and Beijing.

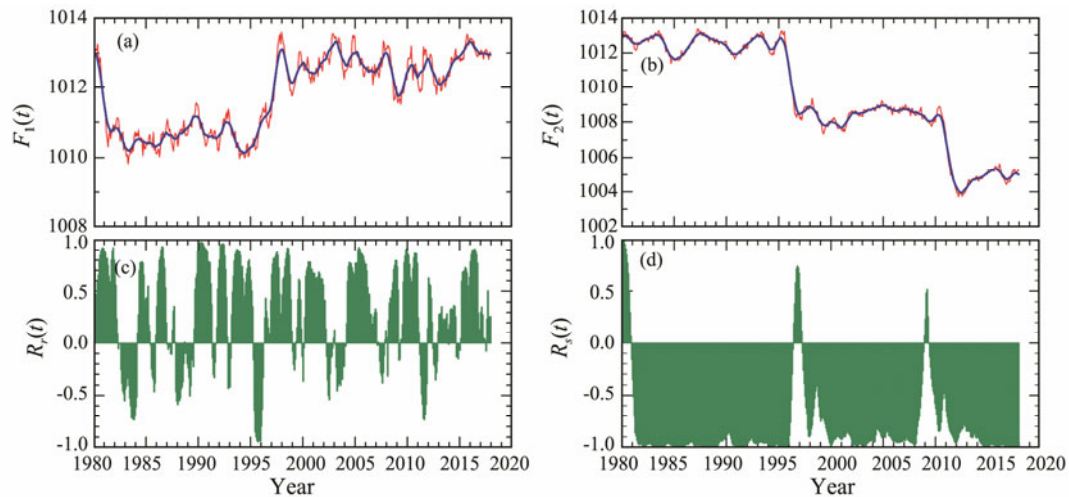


Fig.7 Running correlation dominated by low frequency. (a), Monthly air pressure (unit: hPa) in Beijing (red line) and the local mean (blue line); (b), Monthly air pressure (unit: hPa) in Guangzhou (red line) and the local mean (blue line); (c), LRCC $R_L(t)$; (d), SRCC $R_S(t)$. The monthly air pressure data are obtained from the China Meteorological Data Service Center.

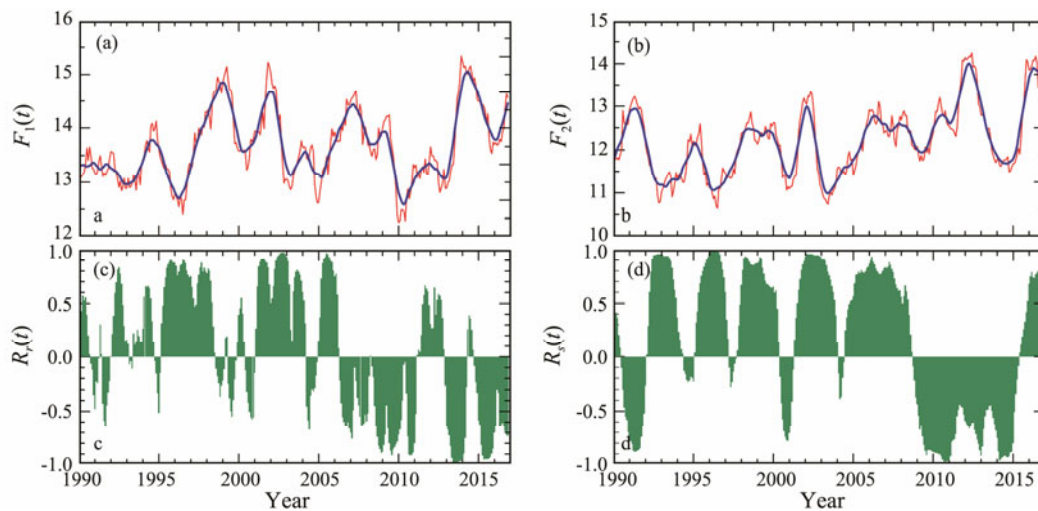


Fig.8 Running correlation of surface air temperatures in Beijing and New York. (a), Surface air temperature (unit: °C) with a 12-point average in Beijing (red line) and the local mean (blue line); (b), Surface air temperature (unit: °C) with a 12-point average in New York (red line) and the local mean (blue line); (c), LRCC $R_L(t)$; (d), SRCC $R_S(t)$. The surface air temperature data are obtained from NCEP/NCAR Reanalysis 1.

Therefore, LRCC only reflects the correlation between high-frequency signals, whereas SRCC reflects the correlation of both high- and low-frequency signals. In particular, if the correlation of phenomena with low-frequency variations or long-term trends is to be studied, SRCC is the inevitable choice.

6 Discussion and Conclusions

A running correlation coefficient (RCC) is calculated by moving the time window to study temporal variations in the correlations of two time series. The local running correlation coefficient (LRCC) is obtained by the direct application of the general definition of a correlation coefficient to the data within a time window. However, LRCC only reflects the correlation between two anomalies within the time window and fails to reflect the contributions

of two varying means. Thus, a new method called synthetic running correlation coefficient (SRCC) was proposed by Zhao *et al.* (2018), which is calculated using the means of all data (global means) instead of the varying local means. SRCC reflects the correlation between varying anomalies and varying means. However, as a recently proposed method, SRCC lacks the support of mathematical theory. In the present study, the validity of SRCC is demonstrated theoretically by considering the comparability of RCC values in different time windows.

RCC is only meaningful when its values at different times can be compared. Thus, a pair of constants must be determined prior to the actual calculation as the standard for comparison. The unique standard quantities are demonstrated to be the global means. This result indicates that SRCCs of different time windows are comparable, but LRCCs are not. Comparability may also be expressed in

the x - y geometric space of the two data series. The cross points of the fitted lines of LRCC at different times are found at different positions; by contrast, the cross points of SRCC are located at the same position. Therefore, the magnitudes of SRCC at different time windows are comparable, whereas those of LRCC are not.

In this study, the relationship between LRCC and SRCC was derived by statistical methods, and SRCC was obtained by adding a constraint condition to the LRCC algorithm. Specifically, the cross points of the regression lines must pass through the center of all data represented by the global means in the geometric space. This constraint condition provides the mathematical basis of the comparison standards. Thus, SRCC is proven to be the unique RCC satisfying the comparability criterion.

When the temporal fluctuations are divided into high-frequency (*i.e.*, periods shorter than the time window) and low-frequency (*i.e.*, periods longer than the time window) signals, some examples show that LRCC only reflects the correlation of high-frequency signals. By contrast, SRCC reflects the correlation of both high and low frequencies.

Our findings do not mean that previous results obtained using LRCC are questionable. Many studies have focused on the correlation between seasonal and sub-seasonal signals, which are high-frequency variations, and LRCC is, in fact, a good measure of the relevant correlation. Nevertheless, if the correlation of phenomena with low-frequency variations or long-term trends is to be studied, SRCC is the inevitable choice because it provides the complete information of the running correlation between various periods.

More importantly, SRCC embodies the physical fact that any piece of data is a part of the whole dataset. The global means are simply parameters that include the information of the whole data. SRCC establishes the connection between local and global variations via the global means. Therefore, SRCC is the correct approach to calculate RCC.

The dependence of SRCC on global means gives rise to a unique feature of SRCC. Alterations in a data domain result in changes in the global mean. Thus, SRCC varies for different data lengths even if the same data series are used.

Acknowledgements

This study was supported by the National Natural Science Foundation of China (Nos. 41976022, 41941012), and the Major Scientific and Technological Innovation Projects of Shandong Province (No. 2018SDKJ0104-1).

References

Brown, J. R., Hope, P., Gergis, J., and Henley, B. J., 2016. ENSO teleconnections with Australian rainfall in coupled model simulations of the last millennium. *Climate Dynamics*, **47** (1-2): 79-93, DOI: <http://dx.doi.org/10.1007/s00382-015-2824-6>.

Burdun, G. D., and Markov, B. N., 1972. *Osnovy Metrologii (Fundamentals of Metrology)*. Izd-vo Standartov, Moscow, 196-206.

Elias, A. G., and Zossi de Artigas, M., 2003. A search for an association between the equatorial stratospheric QBO and solar UV irradiance. *Geophysical Research Letters*, **30**: 1841.

Francis, J. A., and Vavrus, S. J., 2012. Evidence linking Arctic amplification to extreme weather in mid-latitudes. *Geophysical Research Letters*, **39**: L06801, DOI: [10.1029/2012GL051000](https://doi.org/10.1029/2012GL051000).

Galton, F., 1888. Correlations and their measurement, chiefly from anthropometric data. *Proceedings of the Royal Society of London*, **45**: 135-145.

Geng, X., Zhang, W., Jin, F. F., and Stuecker, M. F., 2018. A new method for interpreting nonstationary running correlations and its application to the ENSO-EAWM relationship. *Geophysical Research Letters*, **45**: 327-334.

Huybrechs, D., 2010. On the Fourier extension of non-periodic functions. *SIAM Journal on Numerical Analysis*, **47** (6): 4326-4355.

Hynčica, M., and Huth, R., 2020. Gridded versus station temperatures: Time evolution of relationships with atmospheric circulation. *Journal of Geophysical Research: Atmospheres*, **125**: e2020JD033254, <https://doi.org/10.1029/2020JD033254>.

Ji, X. P., and Zhao, J. P., 2019. Transition periods between sea ice concentration and sea surface air temperature in the Arctic revealed by an abnormal running correlation. *Journal of Ocean University of China*, **18** (3): 633-642, DOI: [10.1007/s11802-019-3909-3](https://doi.org/10.1007/s11802-019-3909-3).

Kodera, K., 1993. Quasi-decadal modulation of the influence of the equatorial quasi-biennial oscillation on the north polar stratospheric temperatures. *Journal of Geophysical Research: Atmospheres*, **98**: 7245-7250.

Kuznets, S., 1928. On moving correlation of time sequences. *Journal of the American Statistical Association*, **23** (162): 121-136.

Merrington, M., Blundell, B., Burrough, S., Golden, J., and Hogarth, J., 1983. A list of the papers and correspondence of Karl Pearson (1857 – 1936). Publications Office, University College London.

Pearson, E. S., 1938. *Karl Pearson: An Appreciation of Some Aspects of His Life and Work*. Cambridge University Press, Cambridge, 193-257.

Pearson, K., 1896. Mathematical contributions to the theory of evolution. – On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London*, **60** (3): 489-498.

Salby, M., Callaghan, P., and Shea, D., 1997. Interdependence of the tropical and extratropical QBO: Relationship to the solar cycle versus a biennial oscillation in the stratosphere. *Journal of Geophysical Research: Atmospheres*, **102** (D25): 29789-29798.

Schmid, J., 1947. The relationship between the coefficient of correlation and the angle included between regression lines. *The Journal of Educational Research*, **41** (4): 311-313.

Soukhearev, B., 1997. The sunspot cycle, the QBO, and the total ozone over northeastern Europe: A connection through the dynamics of stratospheric circulation. *Annales Geophysicae*, **15**: 1595-1603.

Turner, T. E., Swindles, G. T., Charman, D. J., Langdon, P. G., Morris, P. J., Booth, R. K., Parry, L. E., and Nichols, J. E., 2016. Solar cycles or random processes? Evaluating solar variability in Holocene climate records. *Scientific Reports*, **6**: 23961, <https://doi.org/10.1038/srep23961>.

Varotsou, E., Jochumsen, K., Serra, N., Kieke, D., and Schneider, L., 2015. Interannual transport variability of upper Labrador Sea water at Flemish Cap. *Journal of Geophysical Research: Atmospheres*, **120**: 1035-1048.

- Oceans*, **120**: 5074-5089.
- Woods, J. W., and Oneil, S. D., 1986. Subband coding of images. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **34**: 1278-1288.
- Zhao, J. P., and Huang, D. J., 2001. Mirror extending and circular spline function for empirical mode decomposition method. *Journal of Zhejiang University (Science)*, **2** (3): 247-252.
- Zhao, J. P., and Drinkwater, K., 2014. Multiyear variation of the main heat flux components in the four basins of Nordic Seas. *Periodical of Ocean University of China*, **44** (10): 9-19 (in Chinese with English abstract).
- Zhao, J. P., Cao, Y., and Shi, J. X., 2006. Core region of Arctic oscillation and the main atmospheric events impact on the Arctic. *Geophysical Research Letters*, **33**: L22708.
- Zhao J. P., Cao, Y., and Wang, X., 2018. The physical significance of the synthetic running correlation coefficient and its applications in oceanic and atmospheric studies. *Journal of Ocean University of China*, **17** (3): 451-460.
- Zhao, J. P., Drinkwater, K., and Wang, X., 2019. Positive and negative feedbacks related to the Arctic oscillation revealed by air-sea heat fluxes. *Tellus A: Dynamic Meteorology and Oceanography*, **71** (1): 1-21.

(Edited by Chen Wenwen)