J. Ocean Univ. China (Oceanic and Coastal Sea Research)
https://doi.org/10.1007/s11802-018-3798-x
ISSN 1672-5182, 2018 17 (3): 451-460
http://www.ouc.edu.cn/xbywb/
E-mail:xbywb@ouc.edu.cn

The Physical Significance of the Synthetic Running Correlation Coefficient and Its Applications in Oceanic and Atmospheric Studies

ZHAO Jinping[1), 2), *], CAO Yong[1), *], and WANG Xin[1)]

1) *College of Oceanic and Atmospheric Sciences*, *Ocean University of China*, *Qingdao* 266100, *China*
2) *Physical Oceanography Laboratory*, *Ministry of Education*, *Qingdao* 266100, *China*

(Received January 15, 2018; revised March 15, 2018; accepted March 27, 2018)

Abstract In order to study the temporal variations of correlations between two time series, a running correlation coefficient (RCC) could be used. An RCC is calculated for a given time window, and the window is then moved sequentially through time. The current calculation method for RCCs is based on the general definition of the Pearson product-moment correlation coefficient, calculated with the data within the time window, which we call the local running correlation coefficient (LRCC). The LRCC is calculated *via* the two anomalies corresponding to the two local means, meanwhile, the local means also vary. It is cleared up that the LRCC reflects only the correlation between the two anomalies within the time window but fails to exhibit the contributions of the two varying means. To address this problem, two unchanged means obtained from all available data are adopted to calculate an RCC, which is called the synthetic running correlation coefficient (SRCC). When the anomaly variations are dominant, the two RCCs are similar. However, when the variations of the means are dominant, the difference between the two RCCs becomes obvious. The SRCC reflects the correlations of both the anomaly variations and the variations of the means. Therefore, the SRCCs from different time points are intercomparable. A criterion for the superiority of the RCC algorithm is that the average value of the RCC should be close to the global correlation coefficient calculated using all data. The SRCC always meets this criterion, while the LRCC sometimes fails. Therefore, the SRCC is better than the LRCC for running correlations. We suggest using the SRCC to calculate the RCCs.

Key words running correlation coefficient; time window; anomaly; varying mean; synthetic running correlation coefficient

1 Introduction

A correlation coefficient is a quantity that describes the degree of correlation of two different physical quantities. The correlation coefficient of a linear correlation is called the simple correlation coefficient and is also known as the Pearson product-moment correlation coefficient (Pearson, 1896). The Pearson's correlation coefficient was first introduced by Francis Galton (Galton, 1888), but its well-known form for effect sizes was developed and applied by Karl Pearson (Pearson, 1938). For two time series, X_k and Y_k, with data lengths N, the Pearson's correlation coefficient R is commonly represented as follows:

$$R=\frac{\sum_{k=1}^{N}(X_k-\overline{X})(Y_k-\overline{Y})}{\sqrt{\sum_{k=1}^{N}(X_k-\overline{X})^2}\sqrt{\sum_{k=1}^{N}(Y_k-\overline{Y})^2}}, \tag{1}$$

where $\overline{X}$ and $\overline{Y}$ are means defined by

$$\overline{X}=\frac{1}{N}\sum_{k=1}^{N}X_k;\ \overline{Y}=\frac{1}{N}\sum_{k=1}^{N}Y_k\ . \tag{2}$$

The two variables in Eq. (1) can be centralized by subtracting the respective means as follows:

$$X_k^{'}=X_k-\overline{X};\ Y_k^{'}=Y_k-\overline{Y}\ . \tag{3}$$

Eq. (1) can be written as follows:

$$R=\frac{\sum_{k=1}^{N}X_k^{'}Y_k^{'}}{\sqrt{\sum_{k=1}^{N}X_k^{'2}}\sqrt{\sum_{k=1}^{N}Y_k^{'2}}}\ . \tag{4}$$

For the sake of distinguishing the correlation coefficients from the RCCs shown below, the correlation coefficients expressed by Eq. (1) and Eq. (4) are called global correlation coefficients, and the mean expressed by Eq. (2) is called the global mean.

The global correlation coefficient has a definite geometric meaning. In X-Y space, two fitting straight lines can be obtained by

* Corresponding author. Tel: 0086-532-66782096
E-mail: jpzhao@ouc.edu.cn; caoyong@ouc.edu.cn

$$X_k = aY_k + b;\ Y_k = cX_k + d\,. \tag{5}$$

The global correlation coefficient is equal to the cosine of the acute angle θ between the two straight lines. The smaller the angle θ is, the greater the global correlation coefficient becomes (Schmid, 1947).

Being discontent with the global correlation coefficient, some scientists tried to study the temporal variations of correlations. Thus, the running correlation coefficient (RCC) has become a useful tool for describing the varying correlations. The RCC is usually calculated *via* an algorithm similar to Eq. (1) by using a subset of the data centered at i within the time window $\pm n$,

$$R_r(i) = \frac{\sum_{k=i-n}^{i+n}(X_k - \overline{X}_i)(Y_k - \overline{Y}_i)}{\sqrt{\sum_{k=i-n}^{i+n}(X_k - \overline{X}_i)^2}\sqrt{\sum_{k=i-n}^{i+n}(Y_k - \overline{X}_i)^2}}$$

$$i = 1+n,\cdots,N-n\,, \tag{6}$$

where n is chosen to be large enough to depress high-frequency signals and small enough to express the period we are concerned with (Zhao *et al.*, 2006), and $\overline{X}_i$ and $\overline{Y}_i$ are called the local means because they are obtained from the data within the time window:

$$\overline{X}_i = \frac{1}{2n+1}\sum_{k=i-n}^{i+n} X_k;\ \overline{Y}_i = \frac{1}{2n+1}\sum_{k=i-n}^{i+n} Y_k\,. \tag{7}$$

By sequentially moving the window through time, the varying correlation coefficient with time, $R_r(t)$, is obtained and reflects the temporal variations of the correlation characteristics. The time series of the RCC is shorter than the length of the total data and has dead zones with n points at both ends of the data when the time window length is $\pm n$. When an abnormal correlation appears in the RCC time series, this value usually indicates a special event that scientists are likely to be interested in. As all the means and anomalies in $R_r(t)$ are obtained from the data within the time window, we call $R_r(t)$ the local running correlation coefficient (LRCC).

The LRCC expressed by Eq. (6) is widely used to study varying correlations between two time series, such as the correlation of the sea level pressure of the Nordic Seas to the Arctic Oscillation Index (Zhao *et al.*, 2006), the association of the equatorial stratospheric QBO and solar activity (Kodera, 1993; Salby *et al.*, 1997; Soukhearev, 1997; Elias and Zossi de Artigas, 2003), the connection of the oceanic transport with several systems (Varotsou *et al.*, 2015), the temperature reconstructions for the periods of the 16th to 18th centuries (Maurer *et al.*, 2009), the correlations between cloud cover and sea ice concentrations (Ji and Zhao, 2015), the inter annual variations of river discharge (Chen *et al.*, 2016), *etc*.

However, calculating the LRCC using Eq. (6) is still questionable: is this system a reasonable method for deriving the RCC? In fact, Eq. (6) lacks proof for calculating the RCC. This study first defines the defect of the LRCC and then proposes a new method for calculating the RCC.

2 Two Kinds of Running Correlation Coefficients

The LRCC expressed by Eq. (6) has been the general and sole form of an RCC up until now. An example of the LRCC of two time series according to Eq. (6) is shown in Fig.1: one time series is the spatially averaged solar radiation arriving at the Greenland Sea (Fig.1a), and the other is the Arctic Oscillation (AO) Index (Fig.1b). The LRCC $R_r(t)$ calculated for the two time series is shown in Fig.1c, and is dominated by positive correlations in most years, but there is irrelevant or negative correlation in some years.

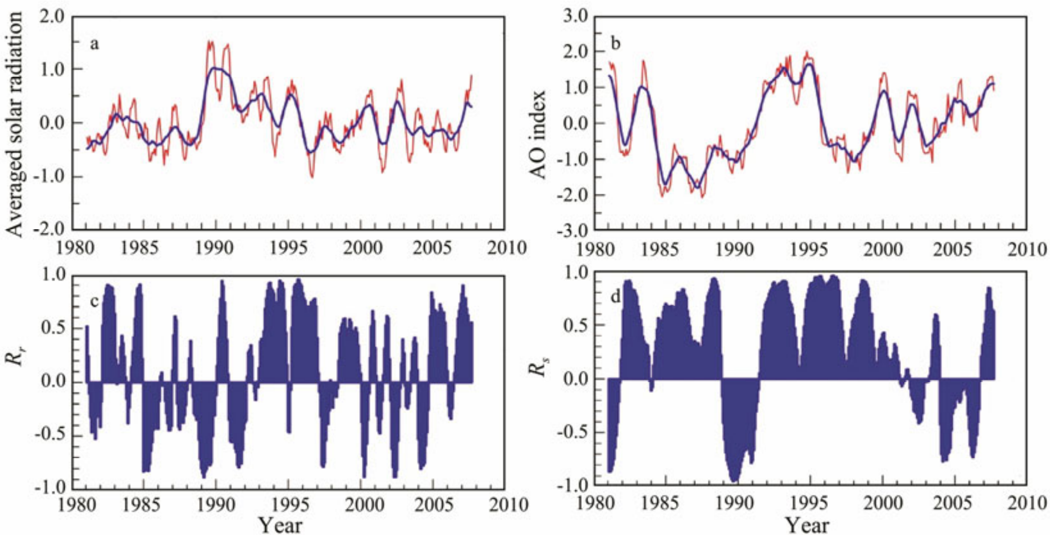

Fig.1 Comparison of the two running correlation coefficients. a), Averaged arriving solar radiation in the Greenland Sea (red line) and its local mean (blue line); b), AO index (red line) and its local mean (blue line); c), $R_r(t)$; d), $R_s(t)$.

However, there is a problem with $R_r(t)$. Not only the anomalies change with time but the means also change with time (red lines in Figs.1a and 1b). The value $R_r(t)$ reflects only the correlation between the anomalies of the two time series, but the correlations of the varying local means are not considered because these values have been subtracted when calculating $R_r(t)$ (see Eq. (6)). Thus, whether the LRCC is significant when only the anomaly variations are considered remains unclear.

Here, we give an example to illustrate the problem. Let the temperatures of two cities during a summer vary with an LRCC, $R_0(t)$. Assuming that the temperatures of the two cities increased 2℃ and 1℃ within the same time window, respectively, the LRCC is still $R_0(t)$. This is an inevitable result of Eq. (6) since the LRCC is determined by the anomalies, not by the varying means. Although the two anomalies did not change, the local means did change during this period. The varying means usually reflect important physical processes. In this example, the temperatures of the two cities in the summer rose, indicating the emergence of a warm summer. A higher positive correlation coefficient is expected in this situation, but the LRCC calculated by Eq. (6) changes little as it cannot reflect the actions of varying means.

The contribution of the varying local means is not negligible, so an unavoidable problem in calculating the RCC lies in how to present the impacts of varying means. This problem does not appear when calculating the global correlation coefficients with global means, as the mean has a singular value. However, when the RCC is calculated by using local means, the varying local means become explicit and are excluded from the LRCC. The LRCC cannot fully reflect the changing correlations if the varying means are not considered. One key question lies in finding an algorithm that reflects the correlations caused by both the varying anomalies and varying local means.

To present the physical significance of the running correlation, unchanged means are adopted to calculate the RCC when attempting to model the variations of local means. These unchanged means are chosen as the global means, as defined by Eq. (2), which are calculated using all available data. Another RCC $R_s(t)$ is defined as

$$R_s(i)=\frac{\sum_{k=i-n}^{i+n}(X_k-\bar{X})(Y_k-\bar{Y})}{\sqrt{\sum_{k=i-n}^{i+n}(X_k-\bar{X})^2}\sqrt{\sum_{k=i-n}^{i+n}(Y_k-\bar{Y})^2}}$$

$$i=1+n,\cdots,N-n. \quad (8)$$

Here, the $R_s(t)$ for the two sets of data shown in Figs.1a, and 1b is shown in Fig.1d and is obviously different from the LRCC shown in Fig.1c. The positive correlations are greatly improved by removing some high-frequency pulses, and only two large negative correlation periods occur. These improvements are due to the consideration of varying means, as the frequency of the varying mean is lower than that of the anomalies. Obviously, the local means are not used when calculating the RCC *via* Eq. (8). A portion of the local mean is retained in the new anomaly value after subtracting the global mean and is included in the calculation of $R_s(t)$. As $R_s(t)$ includes the contribution of varying anomalies and varying local means, here this result is called a synthetic running correlation coefficient (SRCC). This method was first used by Zhao and Su (2004) to calculate the RCC of two time series and found inconsistent time periods and their dynamic mechanisms, but the details of the rationality of the SRCC use have not been demonstrated.

In fact, the reason for using the method for global correlation coefficients to calculate the RCC and LRCC has not been mathematically demonstrated. The need to derive a rational form to obtain the RCC remains. The SRCC then becomes a candidate in addition to the LRCC. However, the data range for the global mean is the length of the data, but the data range of the SRCC is the length of the time window as expressed by Eq. (8). We hope to know if $R_s(t)$ is a rational correlation coefficient. How is this coefficient understood mathematically and physically? Here, we try to explain the SRCC and explore its physical significance.

To better reflect the differences of the two RCCs, we use two functions to simulate the running correlation:

$$A_1(t)=1.0\times\sin(2\pi t/5.0)+a_1(t);$$
$$A_2(t)=5.0\times\sin(2\pi t/12.0)+a_2(t). \quad (9)$$

Taking the time range of 0–700, the means $a_1(t)$ and $a_2(t)$ are set as

$$a_1(t)=\begin{cases}1 & t\in[300,400]\\ 0 & \text{other time}\end{cases};\ a_2(t)=\begin{cases}5 & t\in[300,400]\\ 0 & \text{other time}\end{cases}. \quad (10)$$

The red lines in Figs.2a and 2b are the two functions, and the blue lines are their local means. The LRCC and SRCC values with 13 time windows are shown in Fig.2c and Fig.2d, respectively.

The two RCCs are very small when the local means are equal to zero. When the non-zero local mean appears, the response of the LRCC to the changing mean is still small, except at the points with shifts of the local means (Fig.2c). However, the SRCC becomes large in the presence of non-zero local means, reflecting high correlations. Obviously, compared with $R_r(t)$, $R_s(t)$ can reflect the correlations caused by varying local means, which approaches our results concerning the possible relevance. In addition, the average of the LRCC is 0.03, and the average value of the SRCC is 0.17. Thus, the latter is closer to the global correlation coefficient of 0.29.

Since the local means are non-zero between 300 and 400, only one-seventh of the data range of the total sequences has global means that slightly deviate from zero. However, for the non-zero means between 100 and 600 (Figs.3a, b), the global means will deviate further from zero. In this situation, $R_r(t)$ is still high only at the transition points (Fig.3c), and $R_s(t)$ is high near both ends (Fig.3d). The global correlation coefficient is 0.39, the average value of $R_r(t)$ is 0.02, and the average value of $R_s(t)$ is 0.31.

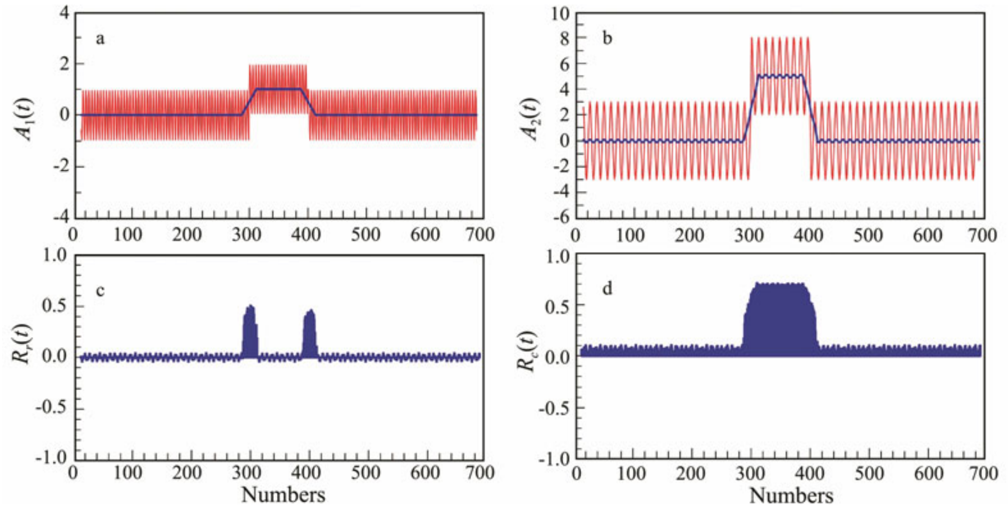

Fig.2 Two kinds of running correlation coefficients with varying means (Example 1). a) and b) show the functions defined by Eq. (9) (red line) and their local means (blue lines) with non-zero a_1 and a_2 during $300 \leq t \leq 400$; c), $R_r(t)$; d), $R_s(t)$.

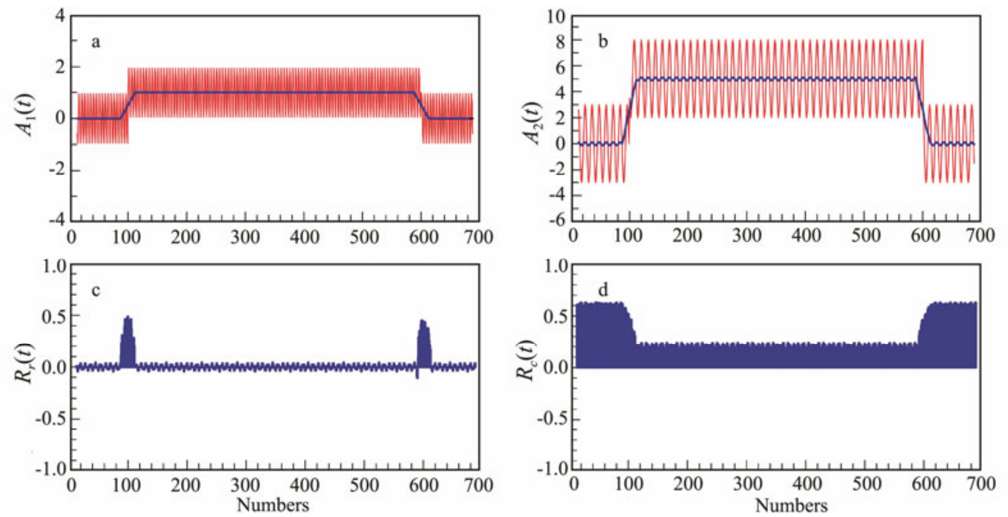

Fig.3 Two kinds of running correlation coefficients with varying means (Example 2). a) and b) show the functions defined by Eq. (9) (red line) and their local means (blue lines) with non-zero a_1 and a_2 during $100 \leq t \leq 600$; c), $R_r(t)$; d), $R_s(t)$.

The $R_s(t)$ of the various conditions can be simulated by adjusting the amplitude, period, a_1 and a_2 in Eq. (9). The results show that the contributions of the local means to $R_s(t)$ are related to their relative magnitudes. If the values of a_1 and a_2 decrease in Eq. (9), $R_s(t)$ also decreases. The local variance is also an important factor. If both amplitudes decrease, the variances will be reduced and $R_s(t)$ will increase. No matter how the parameters in Eq. (9) are adjusted, the results are similar, *i.e.*, the average of the LRCC is small because this value increases only at the transition zones, and the average of the SRCC is more influenced by changing means.

3 The Relationship Between Two RCCs

The examples in Section 2 show the differences between the LRCC and SRCC as well as the effects of the local means and local variances on $R_s(t)$. The examples indicate that $R_s(t)$ is better able to reflect the running correlations between the two time series. $R_s(t)$ not only reflects the correlation between the varying anomalies but also the correlation between the varying means. In this section, the relationships between the two RCCs will be established.

The numerator part of Eq. (8) is rewritten as

$$\sum_{k=i-n}^{i+n}(X_k-\bar{X})(Y_k-\bar{Y})=\sum_{k=i-n}^{i+n}(X_k-\bar{X}_i)(Y_k-\bar{Y}_i)+(2n+1)(\bar{X}_i-\bar{X})(\bar{Y}_i-\bar{Y}) \,. \quad (11)$$

Thus,

$$R_s(i)=\frac{\sum_{k=i-n}^{i+n}(X_k-\bar{X}_i)(Y_k-\bar{Y}_i)}{\sqrt{\sum_{k=i-n}^{i+n}(X_k-\bar{X})^2}\sqrt{\sum_{k=i-n}^{i+n}(Y_k-\bar{Y})^2}}+\frac{(2n+1)(\bar{X}_i-\bar{X})(\bar{Y}_i-\bar{Y})}{\sqrt{\sum_{k=i-n}^{i+n}(X_k-\bar{X})^2}\sqrt{\sum_{k=i-n}^{i+n}(Y_k-\bar{Y})^2}} \,. \quad (12)$$

This equation can be further written as follows:

$$R_s(i) = R_r(i)\frac{\sqrt{\sum_{k=i-n}^{i+n}(X_k-\overline{X}_i)^2}\sqrt{\sum_{k=i-n}^{i+n}(Y_k-\overline{Y}_i)^2}}{\sqrt{\sum_{k=i-n}^{i+n}(X_k-\overline{X})^2}\sqrt{\sum_{k=i-n}^{i+n}(Y_k-\overline{Y})^2}}$$

$$+\frac{(2n+1)(\overline{X}_i-\overline{X})(\overline{Y}_i-\overline{Y})}{\sqrt{\sum_{k=i-n}^{i+n}(X_k-\overline{X})^2}\sqrt{\sum_{k=i-n}^{i+n}(Y_k-\overline{Y})^2}}. \quad (13)$$

The local variances $\sigma_{rx}(t)$ and $\sigma_{ry}(t)$ are

$$\sigma_{rx}^2(i) = \frac{1}{2n+1}\sum_{k=i-n}^{i+n}(X_k-\overline{X}_i)^2 ;$$

$$\sigma_{ry}^2(i) = \frac{1}{2n+1}\sum_{k=i-n}^{i+n}(Y_k-\overline{Y}_i)^2 , \quad (14)$$

and the global variances $\sigma_{sx}(t)$ and $\sigma_{sy}(t)$ are

$$\sigma_{sx}^2(i) = \frac{1}{2n+1}\sum_{k=i-n}^{i+n}(X_k-\overline{X})^2 = \sigma_{rx}^2(i)+(\overline{X}_i-\overline{X})^2 ;$$

$$\sigma_{sy}^2(i) = \frac{1}{2n+1}\sum_{k=i-n}^{i+n}(Y_k-\overline{Y})^2 = \sigma_{ry}^2(i)+(\overline{Y}_i-\overline{Y})^2 . \quad (15)$$

Substituting Eqs. (14) and (15) into Eq. (13), we have

$$R_s(i) = R_r(i)\frac{\sigma_{rx}\sigma_{ry}}{\sqrt{[\sigma_{rx}^2(i)+(\overline{X}_i-\overline{X})^2]}\sqrt{[\sigma_{ry}^2(i)+(\overline{Y}_i-\overline{Y})^2]}}$$

$$+\frac{(\overline{X}_i-\overline{X})(\overline{Y}_i-\overline{Y})}{\sqrt{[\sigma_{rx}^2(i)+(\overline{X}_i-\overline{X})^2]}\sqrt{[\sigma_{ry}^2(i)+(\overline{Y}_i-\overline{Y})^2]}} . \quad (16)$$

The relationship between the two RCCs is established by Eq. (16), which is related to two factors: one is the local variance and the other is the difference between the local mean and global mean (hereafter referred to as the 'mean difference').

Eq. (16) can be further simplified by using the geometric relationship shown in Fig.4. Let the local variances σ_{rx} or σ_{ry} be the horizontal side of the triangle, and let the mean differences be the vertical sides of the triangle. Then, the term in Eq. (16) can be expressed as

$$\tan\gamma_x = \frac{(\overline{X}_i-\overline{X})}{\sigma_{rx}};\ \tan\gamma_y = \frac{(\overline{Y}_i-\overline{Y})}{\sigma_{ry}} . \quad (17)$$

$$\cos\gamma_x = \frac{\sigma_{rx}}{\sqrt{[\sigma_{rx}^2(i)+(\overline{X}_i-\overline{X})^2]}} ;$$

$$\sin\gamma_x = \frac{(\overline{X}_i-\overline{X})}{\sqrt{[\sigma_{rx}^2(i)+(\overline{X}_i-\overline{X})^2]}} ;$$

$$\cos\gamma_y = \frac{\sigma_{ry}}{\sqrt{[\sigma_{ry}^2(i)+(\overline{Y}_i-\overline{Y})^2]}} ;$$

$$\sin\gamma_y = \frac{(\overline{Y}_i-\overline{Y})}{\sqrt{[\sigma_{ry}^2(i)+(\overline{Y}_i-\overline{Y})^2]}} . \quad (18)$$

The relation between the two RCCs is simplified as

$$R_s(i) = R_r(i)\cos\gamma_x\cos\gamma_y + \sin\gamma_x\sin\gamma_y . \quad (19)$$

Eq. (19) shows that the SRCC $R_s(t)$ comprises $R_r(t)$ and 1 for 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). Both weights are not independent but are connected by trigonometric functions. The two angles γ_x and γ_y can be simply determined by the ratio of the mean difference to the local variance using Eq. (17). According to Eq. (18), the larger variance supports a dominant 'cosine weight', and the large mean difference benefits the 'sine weight'. In the extreme cases when the mean difference is zero, the two correlation coefficients are equal, while when the variance of the anomaly is zero, the SRCC is equal to 1.

The relationship between the LRCC and SRCC is expressed by Eq. (19), but in practice, Eq. (19) is used only for understanding $R_s(t)$, not for real calculations. The SRCC is calculated using Eq. (8). The physical significances of the two RCCs are discussed in the next section.

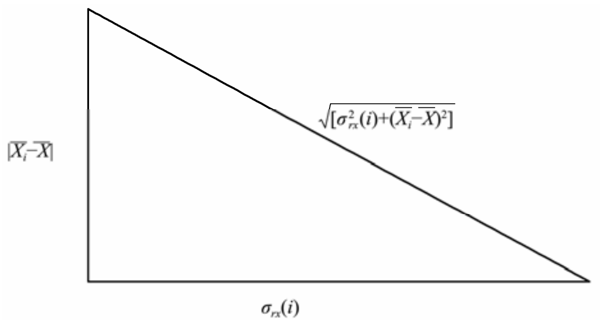

Fig.4 The geometric relationship between local variance and mean difference.

4 The Physical Significance of SRCC

As proven above, the LRCC given by Eq. (6) is a reasonable way to reflect the anomaly-related correlation properties. The contribution of the varying mean is not included in the LRCC, which will result in some missing important information. Thus, determining an improved form of the RCC is necessary. The SRCC $R_s(t)$ has proven to be a type of correlation coefficient including the contributions of the varying anomaly and varying local mean, which is what we want to show. The physical significance of the SRCC needs to be further discussed. Here, we use an analogy to enhance our understanding of the SRCC via the form given in Eq. (19).

The position of the sun in the sky in the equatorial coordinate system was described by (Cooper, 1969)

$$\cos\Theta = \cos\tau\cos\delta\cos\varphi + \sin\delta\sin\varphi , \quad (20)$$

where Θ is the solar zenith angle, τ is the sun time angle,

δ is the sun declination, and φ is the geographical latitude. These parameters are sketched for the equatorial coordinate system in Fig.5. Eq. (19) for SRCC is very similar to Eq. (20) which could be analogously compared. In Eq. (20), the solar time angle τ is the angle between the meridians of the sun and the observer. For an observer standing at the meridian of the local time, at the noon moment, τ is 0˚, and at 6 am/pm, it is ±90˚. Variations of the time angle are caused by the Earth's rotation, such that the time angle reflects only the meridian of where the sun is located. To affirm the position of the sun on the meridian, two other important factors need to be considered: the sun's decline and the geographical latitude of the observer, which are independent of the time angle. The sun's declination δ varies between ±23.27˚ with an annual period, which directly determines the latitudinal distribution of the length of a day. For example, the length of a day at the equator is 12 hours, while in polar summer the length of a day can be up to 24 hours. When the observer's geographical latitude is given and the sun's declination is known, the solar zenith can be completely affirmed.

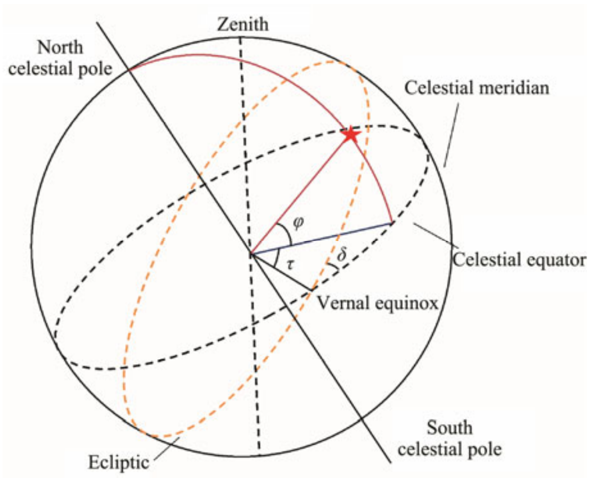

Fig.5 The sun's position in the equatorial coordinate system. τ is the sun time angle, δ is the sun declination, and φ is the geographical latitude.

When the sun's declination and the observer's latitude are both zero, *i.e.*, when the sun is above the equator and the observer is on the equator, the sun zenith angle is equal to the sun time angle τ by Eq. (20). If the sun declination or the geographical latitude, or both, is not equal to zero, the zenith angle will deviate from the time angle. Therefore, the time angle reflects only the meridian of the sun, while the zenith reflects the sun's position at the meridian. If only the time angle is considered, it is impossible to give the sun's position.

Similarly, $R_r(t)$ in Eq. (19) corresponds to the time angle, and $R_s(t)$ corresponds to the solar zenith angle. The two angles of γ_x and γ_y are independent, corresponding to the sun's declination and the observed latitude. The two angles are determined by their mean difference and the local variance of the two data sets. When and only when the two factors are considered, the running correlation coefficient, *i.e.*, $R_s(t)$, could express the complete correlation, including the contributions of both variations in the anomaly and the changing mean. However, $R_r(t)$ only reflects the contribution of the anomaly variation, meaning this value is incomplete for use as the running correlation.

Let us go back to the example of the variations of air temperature. When the air temperatures in the two cities increase simultaneously, $R_r(t)$ changes a little, but $R_s(t)$ is obviously larger during this period. Therefore, $R_s(t)$ could reflect the warm event, while $R_s(t)$ would not.

5 Verifying that $R_s(t)$ Exceeds the $R_r(t)$ Uses as An RCC

Although SRCC is reasonable and has a clear physical significance, it is still necessary to prove that $R_s(t)$ is a better RCC than $R_r(t)$. In fact, there is a good criterion for this purpose: the temporal average of the RCCs should approach the global correlation coefficient. Because the running correlation has a smoothing effect, its average will not be equal to the global correlation coefficient, but their difference should not be large. Some examples of two RCCs calculated from the actual data are shown below to prove that $R_s(t)$ is a better RCC.

5.1 The Correlation Between the Arctic Oscillation and North Atlantic Oscillation Indices

The AO and North Atlantic Oscillation (NAO) are considered to be different aspects of the same process. The correlation of the AO and NAO indices (Figs.6a, b) is considerably high. The difference between $R_r(t)$ and $R_s(t)$ is not remarkable (Figs.6c, d). The global correlation coefficient of the two indices is 0.60, while the averages of these two RCCs are 0.56 and 0.58 for the AO and NAO, respectively. Both are close to the global correlation coefficient. The $\cos\gamma_x\cos\gamma_y$ and $\sin\gamma_x\sin\gamma_y$ are 0.87 and 0.09, respectively, showing that the variations of the anomaly are more significant. Therefore, the two RCCs are close to each other.

5.2 The Correlation Between the Mean Temperature Anomalies at 2 m and 500 hPa

Fig.6 shows an example with a greater variance and smaller mean difference. Fig.7 shows an opposite example, with a smaller variance and greater mean difference, showing the obvious difference between $R_r(t)$ and $R_s(t)$.

The monthly air temperature anomalies averaged over the North Atlantic at 2 m and 500 hPa are shown in Figs.7a, b, and their $R_r(t)$ and $R_s(t)$ values are shown in Figs.7c, d. The global correlation coefficient between the mean temperature anomalies is 0.68. The average of $R_r(t)$ is only 0.28, which is much lower than the global correlation coefficient. The average of $R_s(t)$ reaches 0.64, which is very close to the global correlation coefficient. $R_r(t)$ is rougher than $R_s(t)$ (Figs.7c, d), which results in a low average LRCC. However, $R_s(t)$ eliminates many high-frequency signals, showing a very high correlation between the two time series (Fig.7d).

The variation ranges of the two local variances are very

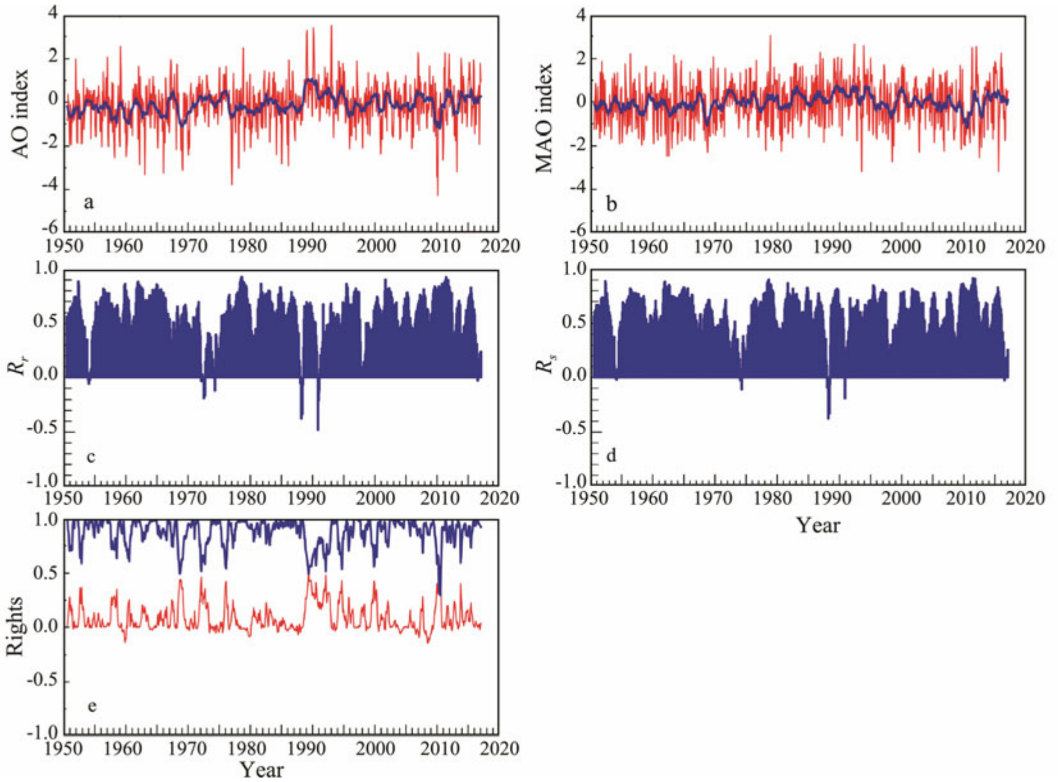

Fig.6 The two running correlation coefficients between the AO and NAO. a), AO index (red line) and its local mean (blue line); b), NAO index (red line) and its local mean (blue line); c), $R_r(t)$; d), $R_s(t)$; e), the time series of the weights of $\cos\gamma_x\cos\gamma_y$ (blue line) and $\sin\gamma_x\sin\gamma_y$ (red line).

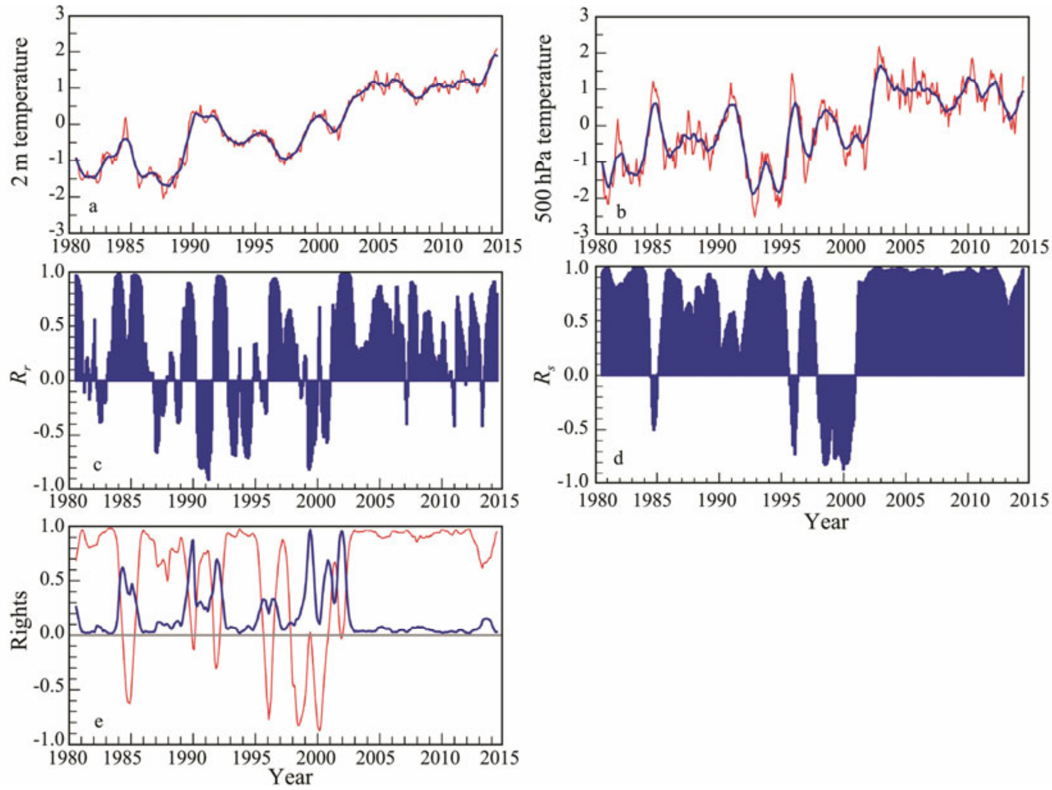

Fig.7 Two running correlation coefficients between the 2 m and 500 hPa air temperature anomalies averaged for the North Atlantic. a), 2 m temperature anomalies (red line) and their local mean (blue line); b), 500 hPa temperature anomalies (red line) and their local mean (blue line); c), $R_r(t)$; d), $R_s(t)$; e), the time series of the weights of $\cos\gamma_x\cos\gamma_y$ (blue line) and $\sin\gamma_x\sin\gamma_y$ (red line).

small, and the weight $\cos\gamma_x\cos\gamma_y$ is only 0.18 (blue line in Fig.7e), whereas the mean difference is much larger, with the weight of sin$\gamma_x$sinγ_y 0.59 (red line in Fig.7e). Obviously, a large mean difference leads to a great difference between $R_r(t)$ and $R_s(t)$.

Therefore, $R_s(t)$ satisfies the criterion of being consistent with the global correlation coefficient. Additionally, there are some high-frequency spiky peaks in $R_r(t)$ that should not exist, as the period of some peaks are shorter than the time window. Meanwhile, $R_s(t)$ diminishes the high-frequency peaks and includes the action of low-frequency signals as expected, increasing the attention paid to the physical background of the low-frequency variations. The results suggest that $R_s(t)$ is a better RCC.

5.3 The Correlation Between the Cloudiness and the Downward Shortwave Radiation in Nordic Seas

The correlation between the cloudiness and solar shortwave radiation anomalies in Nordic seas (Figs.8a, b) is very high, with a global correlation coefficient of −0.72. The RCCs also show a very good negative correlation, with slight differences in 2005–2006 (Figs.8c, d). The average values of $R_r(t)$ and $R_s(t)$ are −0.81 and −0.71, respectively. Although $R_r(t)$ is greater than $R_s(t)$, $R_s(t)$ remains closer to the global correlation coefficient and is a more believable RCC. The weights cos$\gamma_x$cosγ_y and sin$\gamma_x$sinγ_y are 0.54 and −0.25 (Fig.8e), respectively, showing that the contribution of the variance is greater than that of the mean difference.

Note that the large deviations of cloudiness and the arriving shortwave radiation in 2005–2006 are unreasonable. However, as these values are regional averages, these deviations might be caused by unique processes that we have not yet identified. This result suggests that the RCC can indicate the potential abnormal events in the data.

5.4 The Correlation Between the Cloudiness and Sea Ice Concentration in the Central Arctic

The correlation between the cloudiness and sea ice concentration in the central Arctic was discussed by Ji and Zhao (2015). The relationship between the cloudiness and sea ice concentration is interesting. If the cloud change is caused by the sea ice, *i.e.*, less sea ice results in more clouds, both should be negatively correlated. On the other hand, if the sea ice concentration is influenced by the cloudiness, increasing cloudiness results in increasing sea ice, and the two should be positively correlated.

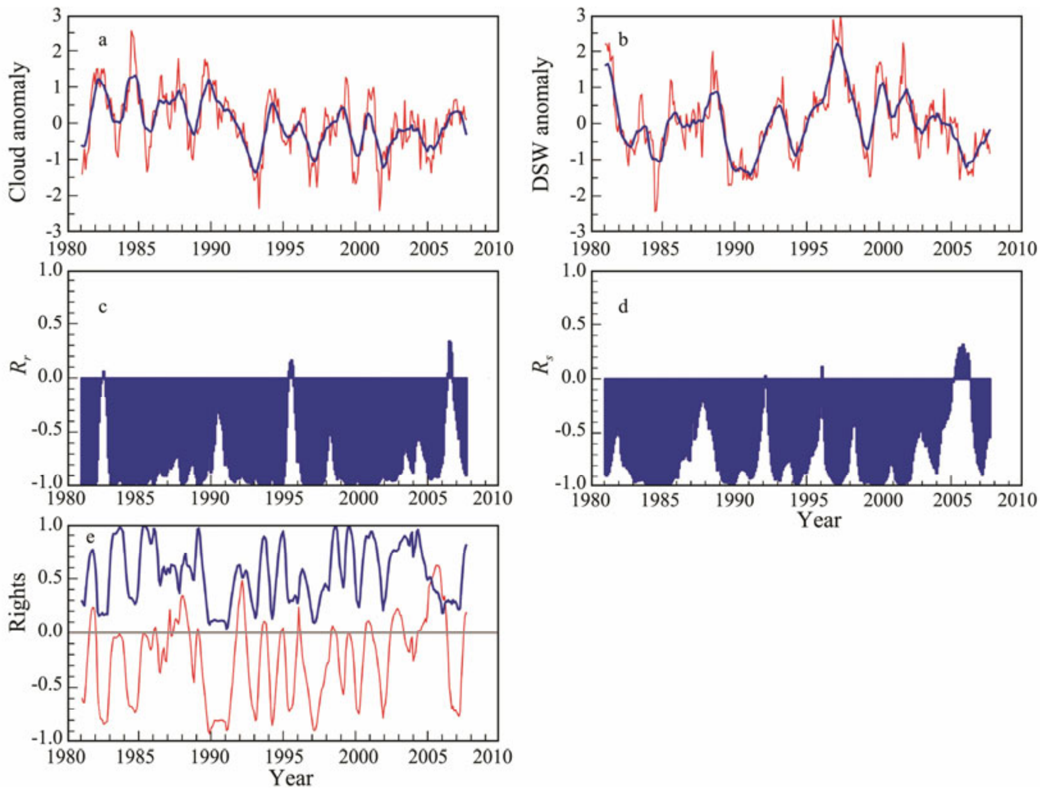

Fig.8 Two running correlation coefficients of the cloudiness and arriving solar short-wave radiation in Nordic seas. a), the cloudiness anomaly (red line) and its local mean (blue line); b), the arriving shortwave radiation anomaly (red line) and its local mean (blue line); c), $R_r(t)$; d), $R_s(t)$; e), the time series of the weights of cos$\gamma_x$cosγ_y (blue line) and sin$\gamma_x$sinγ_y (red line).

The daily low cloudiness and sea ice concentration values from 1992 to 2012 are presented in Figs.9a, b. Their global correlation coefficient is −0.21, showing that decreases of sea ice concentrations increase the generation of low clouds. The average values of $R_r(t)$ and $R_s(t)$ are −0.20 and −0.24, respectively, which are both close to the global correlation coefficient. The negative correlation is dominant, and the RCC can be more than −0.5 during negative correlation stages. However, since the past decade, the positive correlation stage has become dominant, which indicates a changed impact of the clouds on the sea ice. The positive correlation is related to the sea ice retreat in the Arctic, especially in 2010, when the sea ice concentration in the central Arctic reached its minimum (Zhao *et al*., 2018). This pattern implies that more solar energy enters the ocean to melt sea ice when the cloudiness decreases. The transition from a negative correlation to a positive correlation is interesting and is worth further study.

5.5 Advantage of the SRCC

The above examples for ocean and atmosphere exhibit the characteristics of the two RCCs. The SRCC $R_s(t)$ is the weighted average of $R_r(t)$ and 1. The weights are not independent, but comprise and are linked by the sine and cosine functions, such that the SRCC in any case is less than or equal to 1. The contributions of the variance of the anomaly and of the mean difference to the correlation are accounted for. The frequency of the mean variation is lower than that of the variation of the anomaly; thus, the former contributes the low-frequency part of the $R_s(t)$. When the anomaly variation contribution is dominant, $R_s(t)$ and $R_r(t)$ are quite similar and difficult to distinguish. However, when the contribution of the mean variation is dominant, $R_s(t)$ and $R_r(t)$ are markedly different, and the frequency of $R_s(t)$ is lower than that of $R_r(t)$.

Because the RCC is used to reveal temporal variations of the correlations of two time series, the most important issue is that the values of the RCC at different time points should be comparable. The LRCC has a serious flaw as it cannot reflect the correlations caused by varying means. Thus, the values of the LRCC at different time points are not comparable. However, the SRCC can reflect the correlations caused by both the anomaly variations and varying means, representing complete comparability of the correlations at different time points. The comparability of correlations reflects the similarities of the two physical processes at different times. The average of the SRCC values is always close to the total correlation coefficient, showing their mathematical and physical consistencies. Furthermore, SRCC is a simple algorithm and is easy to use. Therefore, we suggest the use of the SRCC instead of the LRCC when calculating the RCC.

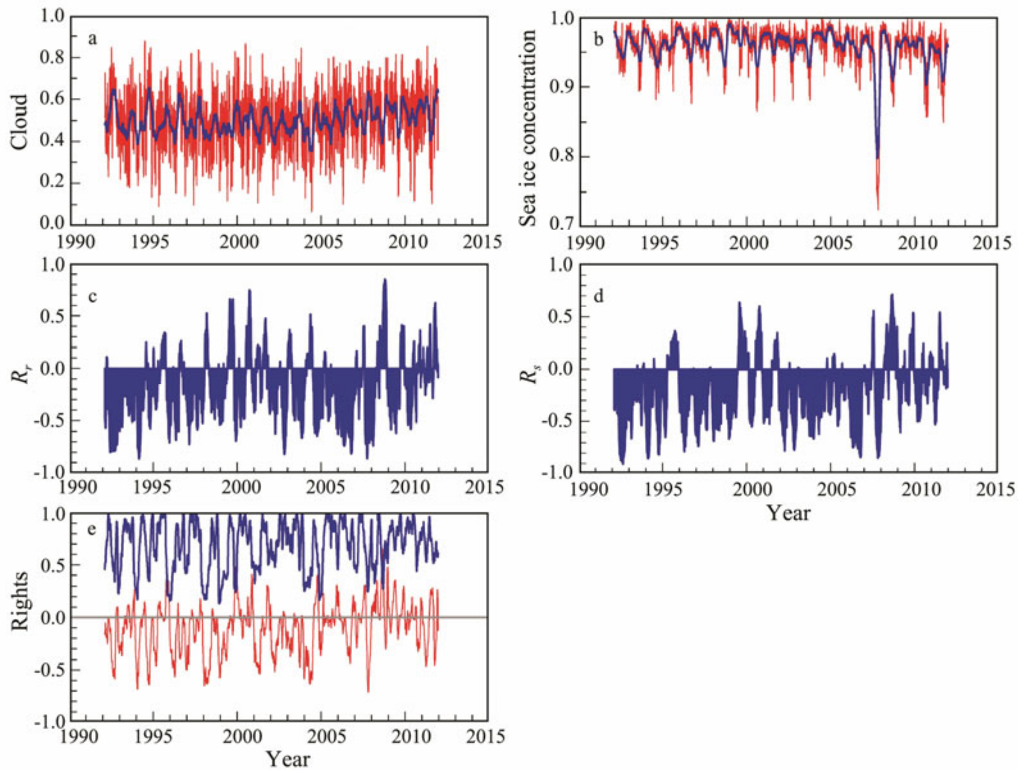

Fig.9 Two running correlation coefficients between the cloudiness and sea ice concentrations in the central Arctic. a), the cloudiness anomaly (red line) and its local mean (blue line); b), the sea ice concentration anomaly (red line) and its local mean (blue line); c), $R_r(t)$; d), $R_s(t)$; e), the time series of the weights of $\cos\gamma_x\cos\gamma_y$ (blue line) and $\sin\gamma_x\sin\gamma_y$ (red line).

6 Conclusions

To study the temporal variations of the correlations of two time series, a RCC must be calculated. An RCC is calculated for a given time window, and the window is then moved sequentially through time to obtain a varying correlation coefficient. The current calculation method for the RCC is based on the general definition of a correlation coefficient, *i.e.*, the RCC value is calculated using the data within the time window, which we refer to as the LRCC. It is noted in this paper that the LRCC reflects only the correlations between the two anomalies within the time window and fails to reflect the contributions of the two varying means. In fact, the lack of the contribution from the varying means is an unavoidable problem, and a solution cannot be provided by the algorithm of the LRCC.

To address this problem, the global mean for all data is used to calculate the RCC, which is called the SRCC. The SRCC is related not only to the anomaly variation described by the LRCC but also to the contribution of the varying means. An important advantage of the SRCC is that the contribution of the varying means to the correlation is adequately considered.

To prove that the SRCC is a better method for use in the running correlation, a criterion is proposed and adopted that the temporal average of the RCCs should be close to the global correlation coefficient. Some examples are discussed in this paper to further our understanding of the difference between the two RCCs. If the anomaly variations in the two processes are dominant, the two RCCs will be consistent. If the contribution of the varying means is dominant, the difference between the two RCCs will be large. In general, the average of the SRCC is close to the global correlation coefficient, while in most cases, the average of the LRCC cannot match the criterion. Additionally, the high-frequency noise in the LRCC does not appear in the SRCC, showing a reasonable low-frequency variation as an RCC. Importantly, as established by the use of the global mean, the SRCC values at different time points are intercomparable. Therefore, the SRCC is better than the LRCC for use in running correlations. We suggest the use of the SRCC instead of the LRCC to calculate the RCCs of two time series.

Although the SRCC has been proven to be a reasonable and a better running correlation coefficient, this method has not been proven to be an exclusive RCC. Hopefully, this result will lead to further research on running correlations.

Acknowledgements

This study is supported by the Key Program of the National Natural Science Foundation of China (No. 41330960) and the Global Change Research Program of China (No. 2015CB953900). We are thankful to Dr. Yanyue Shi for her helpful suggestions.

References

Chen, G. S., Wang, Y. M., Liu, D. F., and Liu, S. Y., 2016. Detection of basin runoff variation based on moving correlation coefficient method. *Journal of Natural Disasters*, **25** (1): 11-18.

Cooper, P. I., 1969. The absorption of radiation in solar stills. *Solar Energy*, **12** (3): 333-346.

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** (16): 337-356.

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

Ji, X. P., and Zhao, J. P., 2015. Analysis of correlation between sea ice concentration and cloudiness in the central Arctic. *Haiyang Xuebao*, **37** (11): 92-104 (in Chinese with English abstract)

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** (D4): 7245-7250.

Maurer, C., Koch, E., Hammerl, C., Hammerl, T., and Pokorny, E., 2009. BACCHUS temperature reconstruction for the period 16th to 18th centuries from Viennese and Klosterneuburg grape harvest dates. *Journal of Geophysical Research Atmospheres*, **114** (D22): D22106.

Pearson, E. S., 1938. Karl Pearson: An appreciation of some aspects of his life and work. *Biometrika*, **28** (3-4): 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*, **102** (D25): 29789-29798.

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

Soukharev, 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** (12): 1595-1603.

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 Oceans*, **120** (7): 5074-5089.

Zhao, J. P., and Jie, S. U., 2004. Causes for correlation and inconsistency between two kinds of low-frequent signals. *Periodical of Ocean University of China*, **34** (5): 697-703 (in Chinese with English abstract).

Zhao, J. P., Barber, D., Zhang, S. G., Yang, Q. H., Wang, X. Y., and Xie, H. J., 2018. Record low sea ice concentration in the central Arctic during summer of 2010. *Advances in Atmospheric Sciences*, **35** (1): 104-113, DOI: https://doi.org/10.1007/s00376-017-7066-6.

Zhao, J. P., Cao, Y., and Shi, J., 2006. Core region of Arctic oscillation and the main atmospheric events impact on the Arctic. *Geophysical Research Letters*, **33** (22): L22708.

(Edited by Chen Wenwen)