Reply to all Referee comments

Reply to reviewer#1:

C1: 9 (and everywhere): "principal" not "principle"
R1: All the word ***"principle"*** has been replaced by ***"principal"***.

C2: 68: Please rewrite, the current meaning is unclear
R2: Thanks! The sentence has been rewritten as ***"the SSAM approach developed by Schoellhamer (2001) computes the elements c(j) of the lagged correlation matrix by,"***

C3: 70: Please rewrite, the "pair of no missing data" is unclear
R3: Thanks! The sentence is revised as ***"where, both x_i and x_{i+j} must be observed rather than missed, N_j is the number of the products of x_i and x_{i+j} within the sample index $i \le N - j$".***

C4: 95: Please rewrite, the meaning is unclear
R4: Thanks! The sentence has been rewritten as ***"The solution of Eq. (10) is as follows,"***

C5: 111: Do you refer to white noise? If yes -- indicate this, if not -- explain the correlation structure of the time series.
R5: Thanks for your comment! The *R(t)* time series refers to the Gaussian white noise. And we have revised the sentence as ***"R(t) is a time series of Gaussian white noise with zero mean and unit standard deviation".***

Reply to reviewer#2:

C1: 13-16: Please split this long sentence into two. Also, please rephrase the statement to better explain what % improvement means. Rewrite "missing data reaches 60%" as "the number of the missing data reaches 60% of the trajectory length" or something similar.
R1: According to your suggestion, we have rewritten this sentence as ***"The result from the synthetic time series with missing data shows that the relative errors of the principal components reconstructed by ISSA are much smaller than those reconstructed by SSAM. Moreover, when the percentage of the missing data over the whole time series reaches 60%, the improvements of relative errors are up to 19.64, 41.34, 23.27 and 50.30% for the first four principal components, respectively."***

C2: 53-55: Please explain how the lagged matrix is constructed. Please define standardized time series.

R2: Thanks for your suggestion! First, we have added the lagged matrix ***C***, the correspondent revision is ***its Toeplitz lagged correlation matrix C is formulated by***

$$\boldsymbol{C}=\begin{bmatrix} c(0) & c(1) & \cdots & c(L-1) \\ c(1) & c(0) & \ddots & \vdots \\ \vdots & \vdots & \ddots & c(1) \\ c(L-1) & \cdots & \cdots & c(0) \end{bmatrix} \tag{1}$$

Each element $c(j)$ ***is computed by***

$$c(j)=\frac{1}{N-j}\sum_{i=1}^{N-j}x_i x_{i+j} \qquad j=0,1,2,\cdots,L-1 \tag{2}$$

The standardized time series here means the stationary time series, so we change the word ***"standardized"*** into ***"stationary"***.

C3: 65-66: Why "On the other hand"?
R3: Thanks! "***On the other hand***" is changed to "***Thus***".

C4: 75-76: What do you mean by "not proven"? Also, is this the only difference from Schoellhamer? Please discuss the novelty of your approach explicitly in the introduction and abstract.
R4: Thanks for your important suggestions. Schoellhamer [2001] used a scale factor L/L_i in calculating the principal component $a_{k,l}$, but he did not tell us the reason of using such a scale factor. In order to avoid confusion we have deleted the sentence "***However, this scale factor is not proved in Schoellhamer (2001)***" in the revised version.
Yes, the method of calculating the principal component is the main difference of our approach from Schoellhamer [2001]. We have pointed out in the abstract and introduction that our approach is derived based on the property that the original time series can be reproduced from its principal components, and Schoellhamer's approach is just a special case of our approach.

C5: 97-98: Please explain what you mean by "neglecting" elements.
R5: "neglecting elements" means the elements are set to zeros. The sentence is revised as "***If the non-diagonal elements of*** G_i ***are set to zero***".

C6: 203-204: Please rewrite as "as the number of missing points get larger"
R6: According to your suggestion the sentence "***As the missing data get larger***" has been revised as "***As the fraction of missing data increases,***".

C7: Thank you for addressing the comments in my initial review. The revised manuscript, however, still does not explain what are the differences of the proposed approach from that of Schoellhamer (2001). Please think of adding a separate paragraph that will clearly explain those differences. I'll proceed with publication in NPG Discussion section as soon as this change is implemented.

R7: We have added a separate paragraph at the end of section 2 to explain the differences as follows: "***The main difference of our ISSA approach from the SSAM approach of Schoellhamer (2001) is in calculating the PCs. We produce the PCs from observed data with Eq. (14) according to the power spectrum (eigenvalues) and eigenvectors of the PCs. While Schoellhamer (2001)***

calculates the PCs from observed data with Eq. (6) only according to the eigenvectors and uses the scale factor L/L_i to compensate the missing value. We have pointed out that this scale factor can be derived from Eq. (15), which is the simplified version of our ISSA approach, by supposing the missing data points with the same eigenvector elements. Therefore the performance of our ISSA approach will be better than SSAM of Schoellhamer (2001). The only disadvantage of our method is that it will cost more computational effort."

Reply to reviewer#3:

C1: ISSA improves SSAM by reformulating the calculation of PCs (equation 7) to incorporate RCs for missing values (equations 8 to 14). The improvement is small for mostly complete time series and increases as the quantity of missing data increases. I encourage the authors to post ISSA code for others to use.

R1: Thanks for your kindly suggestion. We will modify our code and post it soon.

C2: It appears that ISSA Eigenvectors v are calculated as they are in SSAM from the Toeplitz matrix formed from equation 5. This ISSA step should be added to the manuscript.

R2: We add the sentence in page 1951, line 9 **"*Then we compute the eigenvalues and eigenvectors from the lagged correlation matrix C*".**

C3: The eigenvectors are then used to create matrix G. It appears that matrix G must be created and equation 14 solved for each time step i. This is a large increase in computational effort compared to SSAM, which should be stated in the manuscript.

R3: We add the sentence in page 1953, line 23 **"*The only disadvantage of our method is that it will cost more computational effort.*"**

C4: In equation 11, the sums are for all times in the window with a missing value. The values of the eigenvector do not change with time, so the sum can be replaced with N_m, the number of missing values in the window (e.g. $\sum v_{1,j}v_{2,j} = N_m v_{1,j}v_{2,j}$). If N_m=0, equation 10 reduces to equation 3.

R4: The values of the eigenvector vary with the subscript j, so the $\sum v_{1,j}v_{2,j} \neq N_m v_{1,j}v_{2,j}$.

C5: p1953, line 8-13: Equation 15 is used to compare SSAM and ISSA which is good to include but the approach contains a contradiction that should be explained. To compare their results to SSAM, the authors set non-diagonal element in equation 11 to zero but also assume $v_{k,i} = L^{-1/2}$, in which case the diagonal elements would equal N_m/L where N_m is the number of missing data points in the window. The authors should explain this contradiction. For the case where $N_m/L \ll 1$, this contradiction would be minor. Is this contradiction inherently assumed in the formulation of SSAM, and if so, does it explain the relatively improving performance of ISSA as N_m/L (% missing data in table 1) increases? SSAM performance declines when $N_m/L > 0.5$ which is roughly when the diagonal elements of equation 11 become less than the non-diagonal elements—could this be the cause? Or does the ISSA assumption that missing values can be represented by an RC expression create this contradiction? Missing values are ignored when calculating the eigenvectors in both methods, but ISSA does not ignore missing values when calculating PCs.

R5: Thanks for your comment. The Schoellhamer (2001) did not tell us the reason to choose the scale factor L/L_i. And, we find when $v_{k,i} = L^{-1/2}$ and non-diagonal elements equal to zero are both satisfied, we can get the same formula as in Schoellhamer (2001). Thus, we assume it is he ignored this contradiction that makes his method poorer than ours.

C6: p1948 abstract: Add that the improvement is small for mostly complete time series and increases as the quantity of missing data increases. Because of this, I suggest changing 'much smaller' to 'smaller'.

R6: We have changed 'much smaller' to **"*smaller*"**.

C7: p1948, line 16: define SD

R7: SD means **"*standard deviation*"**

C8: p1948, line 17: A difference of 1.2 mg/L (~10%) is within typical measurement error.

R8: Although the percentage of missing data reaches 61%, but the distribution of observed data are very concentrated, thus the non-diagonal elements of matrix $\mathbf{G_i}$ is very small. Then the improvement is also very small.

C9: p1948, line 25-26: use 'wide' only once in the sentence.

R9: We have changed the sentence into **"*SSA has been widely used in geosciences to analyze a variety of time series*"**.

C10: p1949, line 9: Define GNSS

R10: GNSS represents **"*Global Navigation Satellite System*"**.

C11: p1951, line 14: Insert paragraph break where SSAM ends and ISSA starts.

R11: We have revised as above.

C12: p1953, line 8: Insert paragraph break where ISSA ends and comparison to SSAM begins.

R12: We have revised as above.

C13: p1954, line 2-3: This section is about synthetic time series, not the real time series, so delete this sentence.

R13: We have delete this sentence.

C14: p1954, line 20: Delete 'even'.

R14: We have delete the word **"*even*"**.

C15: Equation 18: define T (transpose?).

R15: T represents **"*transpose*"**.

C16: p1955, line 8: delete the word 'clear'.

R16: We have delete the word **"*clear*"**.

C17: p1956, line 19: the mean residual is not represented in table 2.
R17: We have added the mean residual in table 2.

C18: p1956, line 22: the difference of r2 of 0.9178 and 0.9046 seems to be minor- is this statistically significant? Autocorrelation would probably have to be considered.
R18: The reason is almost the same with C8.

C19: Delete 'As' in last row, replace with 'SF'
R19: We have replaced "As" with ***"SF"***.

C20: p1957, line 7-8: Change 'With the missing data gets more, the improvements of the relative errors becomes more evident.' to 'As the fraction of missing data increases, the improvement of the relative error becomes greater'.
R20: We have change the sentence into ***"As the fraction of missing data increases, the improvement of the relative error becomes greater".***

C21: p1957, line 12: The SSC improvements are minor and within measurement error.
R21: The reason is almost the same with C8.

Reply to reviewer#4:

C1: The abbreviation are not proper. For example improved SSA ISSA or similarly SSAM. These should be changed as it is not common.
R1: The SSAM is the abbreviationused by Schoellhamer (2001) to represent the approach of Singular Spectrum Analysis for the time series with Missing data. Our approach is an improved version of SSAM. Thus, we named our approach asISSA to represent improved singular spectrum analysis.

C2: The introduction is very poor. They need to inform what are the novelties of the proposed technique and why its work better than the previous approach. The definition and explanation in page 1953, just before section 3 should go to introduction as explained above. In fact, this motivates your work. Of course, it needs to be expended.
R2: Thanks for your suggestion. We have changed the last paragraph of introduction into ***"This paper is motivated by Schoellhamer (2001) and Shen et al. (2014) and will develop an improved SSA (ISSA) approach. In our ISSA, the lagged correlation matrix is computed with the same way as Schoellhamer (2001), the PCs are directly computed with both the eigenvalues and eigenvectors of the lagged correlation matrix. However, the PCs in Schoellhamer (2001) were calculated with the eigenvectors and a scale factor to compensate the missing value. Moreover, we do not need to fill the missing data recursively and iteratively as in Golyandina and Osipov (2007). The rest of this paper is organized as follows: the improvement of SSA for time series with missing data will be followed in Sect. 2, synthetic and real numerical examples are presented in Sects. 3 and 4 respectively, and then conclusions are given in last Sect. 5."***

C3: Page 1954: We use the 30 h window size (L=120). This is very important issue. Window length.

You did not mention about selection of window length and moreover, the sensitivity of your proposed method to L. The following source might are related to window length selection among many papers on this issue: 1. Multivariate Singular Spectrum Analysis: A General View and New Vector Forecasting Approach. 2. On the Separability Between Signal and Noise in Singular Spectrum Analysis. 3. Separability and Window Length in Singular Spectrum Analysis. 4. Hydroelectric Energy Forecast.

R3: Thank you for suggestion. This paper chooses the same window length as that in Schoellhamer (2001) in order to compare the results with Schoellhamer (2001). We agree with you that the window length is an important issue for singular spectrum analysis; therefore we add the following sentence in page 1954 line 17 "***Although the selection of window length is an important issue for SSA (Hassani 2012, 2013), this paper chooses the same window length (L=120) as that in Schoellhamer (2001) in order to compare the performance of the proposed method with that of Schoellhamer (2001). Using the synthetic time series we computed the lagged correlation matrix and the variances of each mode."***

C4: The performance of the new method should be evaluated with the simulation study. Here the authors use two series of the data sets. However, to have a comprehensive view, they need to consider several issues. Table 2: There is no mean and also mean absolute error. The results indicate that the new approach works better in terms of variance, but reporting mean is important to see the bias of the residual. Figure 2: is very informative. Accordingly, I would recommend having similar figure for simulated data.

R4: Thanks for your comments. We have added the mean absolute error in the table 2.Besides, Figure 2 is the results from simulated data.

Improved Singular Spectrum Analysis for Time Series with Missing Data

Y. Shen[1] F. Peng[1,2] B. Li[1]

1. College of Surveying and Geo-informatics, Tongji University, Shanghai, PR, China
2. Center for Spatial Information Science and Sustainable Development, Shanghai, PR, China

Abstract. Singular spectrum analysis (SSA) is a powerful technique for time series analysis. Based on the property that the original time series can be reproduced from its principal components, this contribution will develop an improved SSA (ISSA) for processing the incomplete time series and the modified SSA (SSAM) of Schoellhamer (2001) is its special case. The approach was evaluated with the synthetic and real incomplete time series data of suspended-sediment concentration from San Francisco Bay. The result from the synthetic time series with missing data shows that the relative errors of the principal components reconstructed by ISSA are smaller than those reconstructed by SSAM. Moreover, when the percentage of the missing data over the whole time series reaches 60%, the improvements of relative errors are up to 19.64, 41.34, 23.27 and 50.30% for the first four principal components, respectively. Besides, both the mean absolute error and mean root mean squared error of the reconstructed time series by ISSA are also smaller than those by SSAM. The respective improvements are 34.45 and 33.91% when the missing data accounts for 60%. The results from real incomplete time series also show that the standard deviation (SD) derived by ISSA is 12.27mg L^{-1}, smaller than 13.48 mg L^{-1} derived by SSAM.

Keywords: Time series analysis, Singular spectrum Analysis, Missing Data

1. Introduction

Singular spectrum analysis (SSA) introduced by Broomhead and King (1986) for studying dynamical systems is a powerful toolkit for extracting short, noisy and chaotic signals (Vautard et al., 1992). SSA first transfers a time series into trajectory matrix, and carries out the principal component analysis to pick out the dominant components of the trajectory matrix. Based on these dominant components, the time series is reconstructed. Therefore the reconstructed time series improves the signal to noise ratio and reveals the characteristics of the original time series. SSA has been widely used in geosciences to analyze a variety of time series, such as the stream flow and sea-surface temperature (Robertson and Mechoso, 1998; Kondrashov and Ghil, 2006), the seismic tomography (Oropeza and Sacchi, 2011) and the monthly gravity field (Zotova and Shum, 2010). Schoellhamer (2001) developed a modified SSA for time series with missing data (SSAM), which has been successfully applied to analyze the time series of suspended-sediment concentration (SSC) in San Francisco Bay (Schoellhamer, 2002). This SSAM approach doesn't need to fill missing data. Instead, it computes the each principal component (PC) with observed data and a scale factor related to the number of missing data. Shen et al. (2014) developed a new principal component analysis approach for extracting common mode errors from the time series with missing

data of a regional station network. The other kind of SSA approaches process the time series with missing data by filling the data gaps recursively or iteratively, such as the "Catterpillar"-SSA method (Golyandina and Osipov, 2007), the imputation method (Rodrigues and Carvalho, 2013) or the iterative method (Kondrashov and Ghil, 2006). This paper is motivated by Schoellhamer (2001) and Shen et al. (2014) and will develop an improved SSA (ISSA) approach. In our ISSA, the lagged correlation matrix is computed with the same way as Schoellhamer (2001), the PCs are directly computed with both the eigenvalues and eigenvectors of the lagged correlation matrix. However, the PCs in Schoellhamer (2001) were calculated with the eigenvectors and a scale factor to compensate the missing value. Moreover, we do not need to fill the missing data recursively and iteratively as in Golyandina and Osipov (2007). The rest of this paper is organized as follows: the improvement of SSA for time series with missing data will be followed in Sect. 2, synthetic and real numerical examples are presented in Sects. 3 and 4 respectively, and then conclusions are given in last Sect. 5.

2. Improved Singular Spectrum Analysis for Time Series with Missing Data

For a stationary time series x_i ($1 \le i \le N$), we can construct an $L \times (N-L+1)$ trajectory matrix with a window size L, its Toeplitz lagged correlation matrix $\boldsymbol{C}$ is formulated by

$$\boldsymbol{C} = \begin{bmatrix} c(0) & c(1) & \cdots & c(L-1) \\ c(1) & c(0) & \ddots & \vdots \\ \vdots & \vdots & \ddots & c(1) \\ c(L-1) & \cdots & \cdots & c(0) \end{bmatrix} \tag{1}$$

Each element $c(j)$ is computed by

$$c(j) = \frac{1}{N-j} \sum_{i=1}^{N-j} x_i x_{i+j} \qquad j = 0, 1, 2, \cdots, L-1 \tag{2}$$

For matrix $\boldsymbol{C}$, we can compute its eigenvalues λ_k and the corresponding eigenvectors $\boldsymbol{v}_k$ in descending order of λ_k ($1 \le k \le L$). Then the ith element of kth principal components (PCs) $\boldsymbol{a}_k$ is computed by

$$a_{k,i} = \sum_{j=1}^{L} x_{i+j-1} v_{j,k} \qquad 1 \le i \le N-L+1 \tag{3}$$

where $v_{j,k}$ is the jth element of $\boldsymbol{v}_k$. We compute the kth reconstructed components (RCs) of the time series with the kth PCs as (Vautard et al., 1992)

$$x_i^k = \begin{cases} \frac{1}{i}\sum_{j=1}^{i} a_{k,i-j+1} v_{j,k} & 1 \le i \le L-1 \\ \frac{1}{L}\sum_{j=1}^{L} a_{k,i-j+1} v_{j,k} & L \le i \le N-L+1 \\ \frac{1}{N-i+1}\sum_{j=i-N+L}^{L} a_{k,i-j+1} v_{j,k} & N-L+2 \le i \le N \end{cases} \quad (4)$$

Since λ_k, the variance of the kth RC, is sorted in descending order, the first several RCs contain most of the signals of the time series, while the remaining RCs contain mainly the noises of time series. Thus the original time series will be reconstructed with first several RCs.

The SSAM approach developed by Schoellhamer (2001) computes the elements $c(j)$ of the lagged correlation matrix by,

$$c(j) = \frac{1}{N_j} \sum_{i \le N-j} x_i x_{i+j} \qquad j = 0,\ 1,\ 2,\ \cdots,\ L-1 \quad (5)$$

where, both x_i and x_{i+j} must be observed rather than missed, N_j is the number of the products of x_i and x_{i+j} within the sample index $i \le N-j$. Then we compute the eigenvalues and eigenvectors from the lagged correlation matrix $\boldsymbol{C}$. The PCs are also calculated with observed data,

$$a_{k,i} = \frac{L}{L_i} \sum_{1 \le j \le L} x_{i+j-1} v_{j,k} \qquad 1 \le i \le N-L+1 \quad (6)$$

where L_i is the number of observed data within the sample index from i to i+L-1. The reconstruction procedure of time series from PCs is the same as SSA. The scale factor L/L_i is used to compensate the missing value.

In order to derive the expression of computing PCs for the time series with missing data, the Eq. (3) is reformulated as,

$$a_{k,i} = \sum_{i+j-1 \in S_i} x_{i+j-1} v_{j,k} + \sum_{i+j-1 \in \bar{S}_i} x_{i+j-1} v_{j,k} \quad (7)$$

where, $1 \le i \le N-L+1$, S_i and $\bar{S}_i$ are the index sets of sampling data and missing data respectively within the integer interval $[i,\ i+L-1]$, i.e. $S_i \cap \bar{S}_i = 0$ and $S_i \cup \bar{S}_i = [i,\ i+L-1]$. If PCs are available, we can reproduce the missing values. Therefore, the missing values in Eq. (7) can be substituted with PCs as,

$$x_{i+j-1} = \sum_{m=1}^{L} a_{m,i} v_{j,m} \quad (8)$$

Substituting Eq. (8) into the second term of the right hand of Eq. (7) yields,

$$\left(1-\sum_{i+j-1\in\bar{S}_i} v_{j,k}^2\right)a_{k,i}-\sum_{i+j-1\in\bar{S}_i}\sum_{m=1,m\neq k}^{L} v_{j,m}v_{j,k}a_{m,i}=\sum_{i+j-1\in S_i} x_{i+j-1}v_{j,k} \tag{9}$$

Collecting all equations of Eq. (9) for k =1,2,⋯, L, we have,

$$\boldsymbol{G}_i\boldsymbol{\xi}_i=\boldsymbol{y}_i \tag{10}$$

where,

$$\boldsymbol{G}_i=\begin{bmatrix} 1-\sum_{i+j-1\in\bar{S}_i} v_{j,1}^2 & -\sum_{i+j-1\in\bar{S}_i} v_{j,1}v_{j,2} & \cdots & -\sum_{i+j-1\in\bar{S}_i} v_{j,1}v_{j,L} \\ -\sum_{i+j-1\in\bar{S}_i} v_{j,2}v_{j,1} & 1-\sum_{i+j-1\in\bar{S}_i} v_{j,2}^2 & \cdots & -\sum_{i+j-1\in\bar{S}_i} v_{j,2}v_{j,L} \\ \vdots & \vdots & \ddots & \vdots \\ -\sum_{i+j-1\in\bar{S}_i} v_{j,L}v_{j,1} & -\sum_{i+j-1\in\bar{S}_i} v_{j,L}v_{j,2} & \cdots & 1-\sum_{i+j-1\in\bar{S}_i} v_{j,L}^2 \end{bmatrix}, \tag{11}$$

$$\boldsymbol{\xi}_i=\begin{bmatrix} a_{1,i} \\ a_{2,i} \\ \vdots \\ a_{L,i} \end{bmatrix},\ \boldsymbol{y}_i=\begin{bmatrix} \sum_{i+j-1\in S_i} x_{i+j-1}v_{j,1} \\ \sum_{i+j-1\in S_i} x_{i+j-1}v_{j,2} \\ \vdots \\ \sum_{i+j-1\in S_i} x_{i+j-1}v_{j,L} \end{bmatrix} \tag{12}$$

Since $\boldsymbol{G}_i$ is a symmetric and rank-deficient matrix with the number of rank-deficiency equaling to the number of missing data within the interval [x_i, x_{i+L-1}], the PCs $a_{k,i}$ (k=1, ,2,⋯,L) are solved with Eq. (10) based on the following criterion (Shen et al. 2014),

$$\min:\boldsymbol{\xi}_i^T\boldsymbol{\Lambda}^{-1}\boldsymbol{\xi}_i \tag{13}$$

where, $\boldsymbol{\Lambda}$ is diagonal matrix of eigenvalues λ_k, which is the covariance matrix of PCs. The solution of Eq. (10) is as follows,

$$\boldsymbol{\xi}_i=\boldsymbol{\Lambda}\boldsymbol{G}_i^T\left(\boldsymbol{G}_i^T\boldsymbol{\Lambda}\boldsymbol{G}_i\right)^{-}\boldsymbol{y}_i \tag{14}$$

The symbol '-' denotes the pseudo-inverse of a matrix.

If the non-diagonal elements of $\boldsymbol{G}_i$ are all set to zero, the Eq. (14) can be further simplified as,

$$a_{k,i}=\frac{1}{1-\sum_{i+j-1\in\bar{S}_i} v_{j,k}^2}\sum_{1\le j\le L} x_{i+j-1}v_{j,k}\qquad 1\le k\le L,\ 1\le i\le N-L+1 \tag{15}$$

Supposing $v_{1,k}=v_{2,k}=\cdots=v_{L,k}=1/\sqrt{L}$ at the missing data points, the solution of Eq. (15) will be reduced to Eq. (6). Therefore, the SSAM approach is a special case of our ISSA

approach. By the way, the first several PCs contain most variance; the element x_{i+j-1} can be approximately reproduced with the first several PCs in Eq. (8).

The main difference of our ISSA approach from the SSAM approach of Schoellhamer (2001) is in calculating the PCs. We produce the PCs from observed data with Eq. (14) according to the power spectrum (eigenvalues) and eigenvectors of the PCs. While Schoellhamer (2001) calculates the PCs from observed data with Eq. (6) only according to the eigenvectors and uses the scale factor L/L_i to compensate the missing value. We have pointed out that this scale factor can be derived from Eq. (15), which is the simplified version of our ISSA approach, by supposing the missing data points with the same eigenvector elements. Therefore the performance of our ISSA approach will be better than SSAM of Schoellhamer (2001). The only disadvantage of our method is that it will cost more computational effort.

3. Performance of ISSA with synthetic time series

The same synthetic time series as Schoellhamer (2001) are used to analyze the performance of ISSA compared to SSAM. The synthetic SSC time series is expressed as,

$$c(t) = 0.2R(t)c_s(t) + c_s(t) \tag{16}$$

where, $R(t)$ is a time series of Gaussian white noise with zero mean and unit standard deviation; $c_s(t)$ is the periodic signal expressed as,

$$\begin{aligned} c_s(t) = {} & 100 - 25\cos\omega_s t + 25(1-\cos 2\omega_s t)\sin\omega_{sn} t \\ & + 25\left(1 + 0.25(1-\cos 2\omega_s t)\sin\omega_{sn} t\right)\sin\omega_a t \end{aligned} \tag{17}$$

The periodic signal oscillates about the mean value 100mg/L including the signals with seasonal frequency $\omega_s = 2\pi/365\ day^{-1}$, spring/neap angular frequency $\omega_{sn} = 2\pi/14\ \text{day}^{-1}$ and advection angular frequency $\omega_a = 2\pi/(12.5/24)\ \text{day}^{-1}$. The one year of synthetic SSC time series $c(t)$, starting at October 1 with 15-minute time step, is presented on the bottom of Fig. 1, the corresponding periodic signal $c_s(t)$ is shown on the top of Fig. 1.

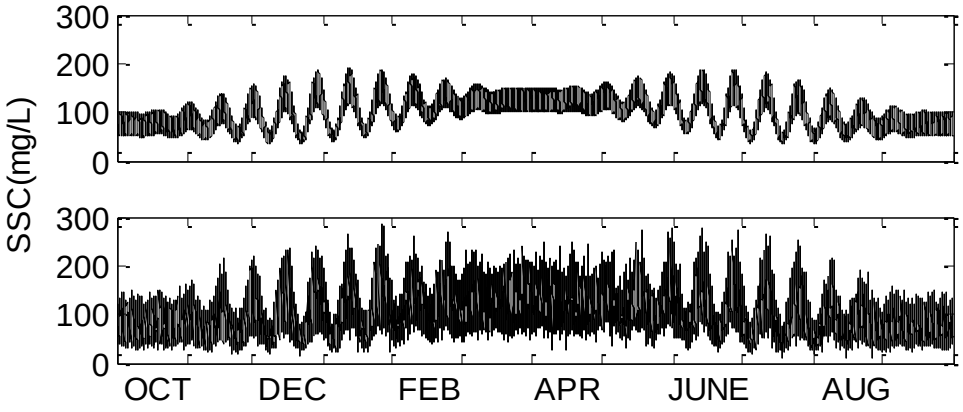

Figure 1. periodic signal $c_s(t)$（top）and Synthetic time series（bottom）

Although the selection of window length is an important issue for SSA (Hassani 2012, 2013), this paper chooses the same window length (L=120) as that in Schoellhamer (2001) in order to compare the performance of the proposed method with that of Schoellhamer (2001). Using the synthetic time series we computed the lagged correlation matrix and the variances of each mode. The first 4 modes contain the periodic components, which account for 72.3% of the total variance; particularly, the first mode contains 50.2% of the total variance. In order to evaluate the accuracies of reconstructed PCs from the time series with different percentages of missing data, following the way of Shen et al. (2014), we compute the relative errors of the first four modes derived by ISSA and SSAM with the following expression,

$$p = \frac{1}{N}\sum_{i=1}^{N}\sqrt{\frac{(\boldsymbol{a}_i - \boldsymbol{a}_0)^T(\boldsymbol{a}_i - \boldsymbol{a}_0)}{\boldsymbol{a}_0^T\boldsymbol{a}_0}} \times 100\% \tag{18}$$

where, the symbol 'T' denotes the transpose of a matrix; p denotes relative error; N is the number of repeated experiments; $\boldsymbol{a}_i$ is the reconstructed PCs of ith experiment from data missing time series, $\boldsymbol{a}_0$ denotes the PCs reconstructed from the time series without missing data. We design the experiment of missing data by randomly deleting the data from the synthetic time series. The percentage of deleted data is from 10% to 60% with an increase of 10% each time. Then, we reconstruct the first four PCs from the data deleted synthetic time series using both SSAM and ISSA, and repeat the experiments for 50 times. The relative errors of the first four PCs are presented in Fig. 2, from which we clearly see that the accuracies of reconstructed PCs by our ISSA are obviously higher than those by SSAM, especially for the second and fourth PCs. In the case of 60% missing data, the accuracy improvements are up to 19.64, 41.34, 23.27 and 50.30% for the first four PCs, respectively.

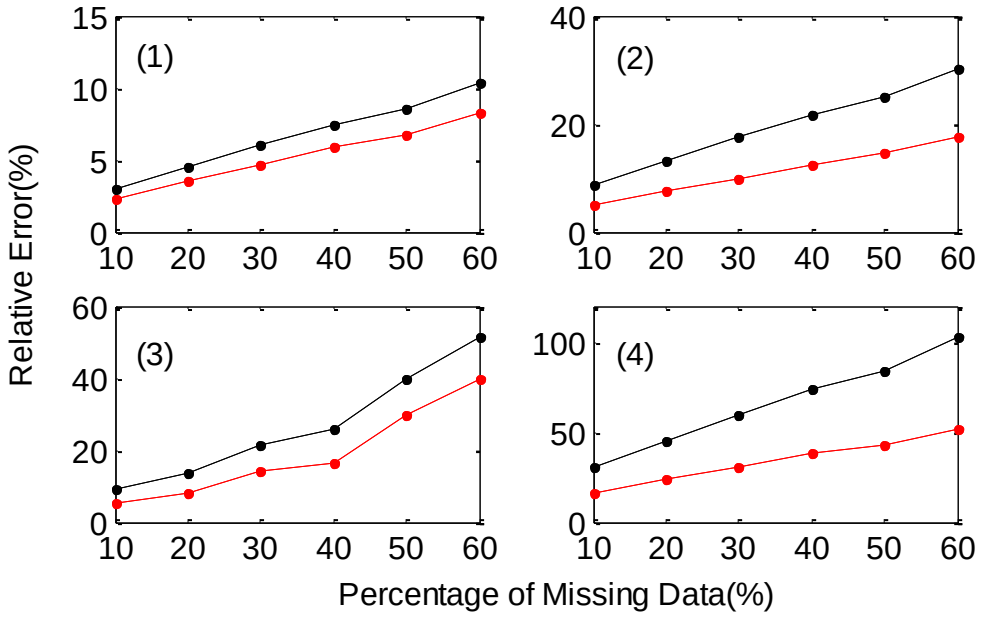

Figure 2. Relative errors of first four PCs (ISSA: red line; SSAM: black line)

We reconstruct the time series $\hat{c}(t)$ using the first four PC modes and then evaluate the quality of reconstructed series by examining the error $\Delta\hat{c}(t)=\hat{c}(t)-c_s(t)$. For the cases whose missing data are between 10% to 50% over the whole time series, the reconstructed component of the time series is calculated only when the percentage of missing data in the window size is less than 50%; while for the cases whose overall missing data already reach 60%, it is allowed 60% missing data in the window size. In Fig. 3, we demonstrate the root mean squared errors (RMSE) of each experiment of different percentages of missing data. The RMSE is computed with $\Delta\hat{c}(t)$ as

$$\mathrm{RMSE}=\sqrt{\sum_{j=1}^{M}\Delta\hat{c}^2\left(t_j\right)\Big/M} \tag{19}$$

where M is the number of data points involved in the experiment.

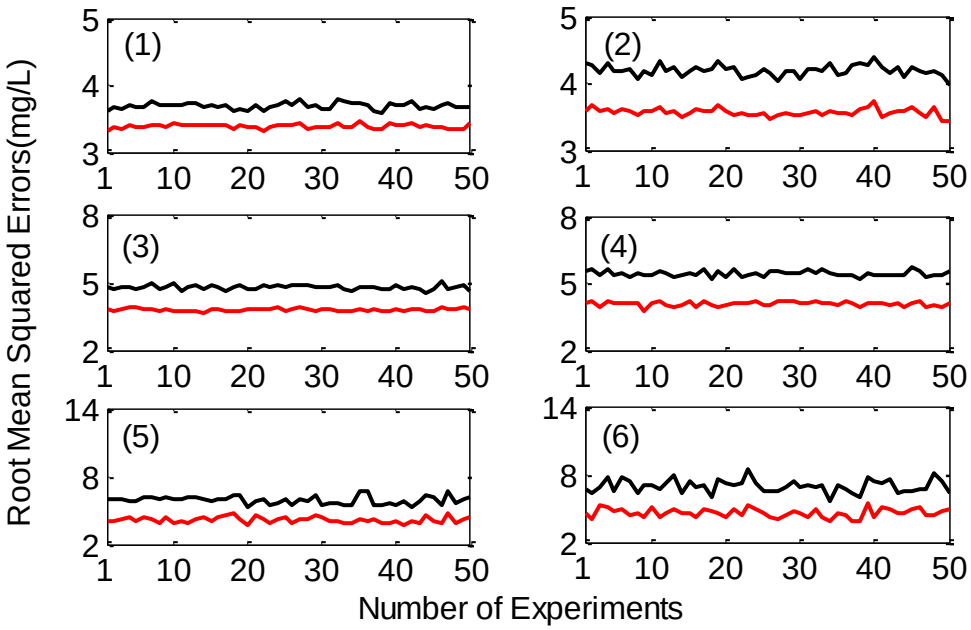

Figure 3. RMSE of 50 experiments, (1)~(6) represent percentage of missing data ranging from 10% to 60% with 10% increments.

As we can see from the Fig. 3, the RMSEs of ISSA are much smaller than those of SSAM for all same experiment scenarios. In Table 1, we present the mean absolute reconstruction error (MARE) and mean root mean squared errors (MRMSE) of 50 experiments with different percentages of missing data.

Table 1: Mean absolute reconstruction error and mean root mean squared error of simulated time series with different percentage of missing data (mg L^{-1})

Percentage of Missing Data (%)	MARE			MRMSE		
	SSAM	ISSA	IMP (%)	SSAM	ISSA	IMP (%)
0	2.48	2.48	0	2.06	2.06	0%
10	2.87	2.60	9.41	3.68	3.38	2.21
20	3.26	2.73	16.26	4.19	3.56	15.04
30	3.71	2.90	21.83	4.76	3.78	20.59
40	4.22	3.11	26.30	5.42	4.07	24.91
50	4.57	3.17	30.63	5.89	4.14	29.71
60	5.37	3.52	34.45	6.96	4.60	33.91
SF Bay Example	3.38	3.08	8.87	2.70	2.29	15.19

Obviously, if there is no missing data, the ISSA coincides with SSAM. If the percentage of missing data increases, both MARE and MRMSE will become larger. In Table 1, all the MARE and MRMSE of ISSA are smaller than those of SSAM. When the percentage of missing data reaches 50%, the MARE and MRMSE are 3.17mg L^{-1} and 4.14 mg L^{-1} for ISSA, and 4.57 mg L^{-1} and 5.89 mg L^{-1} for SSAM, respectively. The improved percentage (IMP) of ISSA with respect to SSAM is also listed in Table 1. As the missing data increases, the IMPs of both MARE and MRMSE

increase as well. Moreover, when the synthetic time series with the missing data is same as the real SSC time series of Fig. 4, the IMPs of MARE and MRMSE are 8.87% and 15.19%, respectively.

4. Performance of ISSA with real time series

The mid-depth SSC time series at San Mateo Bridge is presented in Fig. 4, which contains about 61% missing data. This time series was reported by Buchanan and Schoellhamer (1999) and Buchanan and Ruhl (2000), and analyzed by Schoellhamer (2001) using SSAM. We analyze this time series using our ISSA with the window size of 30h (L=120) comparing with SSAM. The first 10 modes represent dominant periodic components as shown in Schoellhamer (2001) which contain 89.1% of the total variance. Therefore, we reconstruct the time series with first 10 modes when the missing data in a window size is less than 50%.

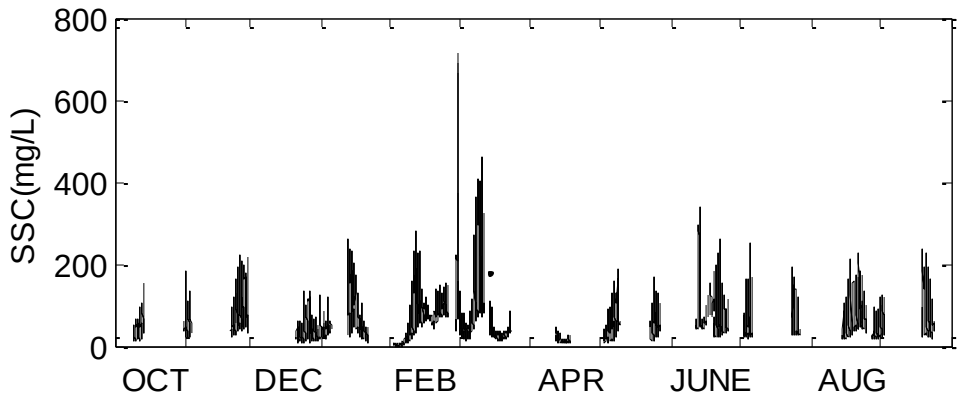

Figure 4. Mid-depth SSC time series at San Mateo Bridge during water year 1997

The residual time series, e.g. the differences of observed minus reconstructed data, are presented in Fig. 5. The maximum, minimum and mean absolute residuals as well as the SD are presented in Table 2. It is clear that both maximum and minimum residuals are significantly reduced by using ISSA approach. The SD of our ISSA is reduced by 8.6%. The squared correlation coefficients between the observations and the reconstructed data from ISSA and SSAM are 0.9178 and 0.9046, respectively, which reflect that the reconstructed time series with our ISSA can indeed, to very large extent, specify the real time series.

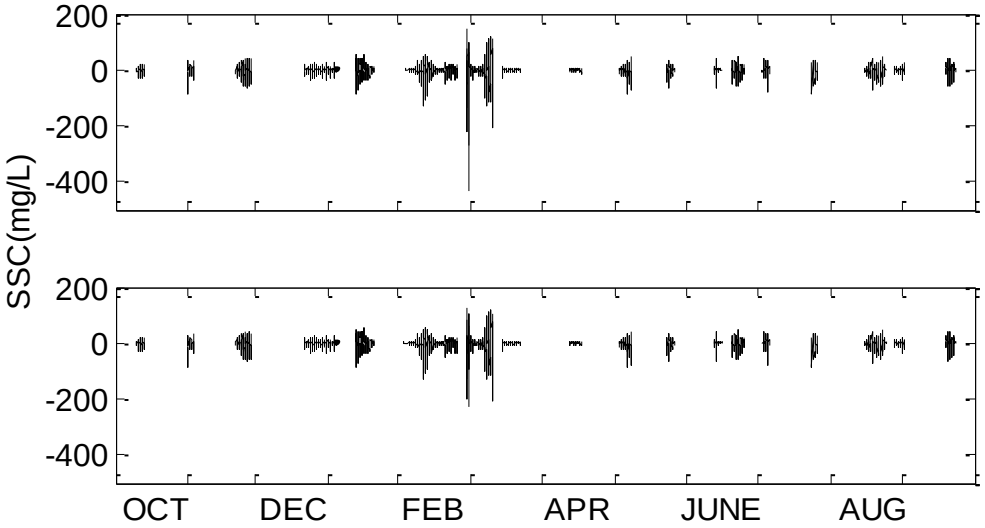

Figure 5. Residual series after removing reconstructed signals from first 10 modes (top: SSAM; bottom: ISSA)

Table 2: Maximum and minimum and mean absolute residuals of SSAM and ISSA

Residuals(mg L^{-1})	SSAM	ISSA
Maximum	145.05	126.61
Minimum	-432.20	-227.70
Mean absolute residuals	8.19	8.00
SD	13.48	12.27

5. Conclusions

We have developed the ISSA approach in this paper for processing the incomplete time series by using the principle that a time series can be reproduced by using its principal components. We proved that the SSAM developed by Schoellhamer (2001) is a special case of our ISSA. The performances of ISSA and SSAM were demonstrated with a synthetic time series, and the results show that the relative errors of the first four principal components by ISSA are significantly smaller than those by SSAM. As the fraction of missing data increases, the improvement of the relative error becomes greater. When the percentage of missing data reaches 60%, the improvements of the first four principal components are up to 19.64, 41.34, 23.27 and 50.30%, respectively. Moreover, when the missing data accounts for 60%, the MARE and MRMSE derived by ISSA are 3.52 mg L^{-1} and 4.60 mg L^{-1}, and by SSAM are 5.37 mg L^{-1} and 6.96 mg L^{-1}. The corresponding improvements of ISSA with respect to SSAM are 34.45 and 33.91%. When the missing data of synthetic time series is the same as the real SSC time series, the improvements of MARE and MRMSE are 8.87 and 15.19%, respectively. The SD derived from the real SSC time series at San Mateo Bridge by ISSA and SSAM

are 12.27 mg L^{-1} and 13.48 mg L^{-1}, and the squared correlation coefficients between the observations and the reconstructed data from ISSA and SSAM are 0.9178 and 0.9046, respectively. Therefore, ISSA can indeed, to a great extent, retrieve the informative signals from the original incomplete time series.

Author contribution

Y. Shen proposed the improved singular spectrum analysis and F. Peng carried out the FORTRAN program and performed the simulations. Y. Shen, F. Peng and B. Li prepared the manuscript.

Acknowledgements

This work was sponsored by Natural Science Foundation of China (Projects: 41274035, 41474017) and partly supported by State Key Laboratory of Geodesy and Earth's Dynamics (SKLGED2013-3-2-Z).

References

Broomhead, D.S., G.P. King, Extracting qualitative dynamics from experimental data. Physica D, 20, 217-236, 1986.

Buchanan, P.A., and C.A Ruhl, Summary of suspended-solids concentration data, San Francisco Bay, California, water year 1998, Open File Report 99-189, 41 pp., U.S. Geological Survey, 2000.

Buchanan, P.A., and D. H. Schoellhamer, Summary of suspended solids concentration data, San Francisco Bay, California, water year 1997, Open File Report 00-88 URL http://ca.water. usgs.gov/rep/ofr99189/, 52 pp., U.S. Geological Survey, 1999.

Golyandina, N., E. Osipov, The "Catterpillar"-SSA method for analysis of time series with missing data, J. Stat. Plan. Inf., 137, 2642-2653, 2007.

Hassani H., Mahmoudvand R., Zokaei M., et al. On the Separability between signal and noise in singular spectrum analysis, Fluct. Noise Lett. 11(2), 1-11, 2012.

Hassani H., Mahmoudvand R. Multivariate singular spectrum analysis: a general view and new vector forecasting approach, Int. J. Energy Stat., 1(1), 55-83, 2013.

Kondrashov, D. M. Ghil, Spatio-temporal filling of missing points in geophysical data sets, Nonlin. Processes Geophys., 13, 151-159, 2006.

Oropeza, V., M. Sacchi, Simultaneous seismic data denoising and reconstruction via multichannel singular spectrum analysis, Geophysics, 76(3), 25-32, 2011.

Robertson, A.W. and C. R. Mechoso, Interannual and decadal cycles in river flows of southeastern South America, Journal of Climate, 11(10), 2570-2581, 1998.

Rodrigues, P.C., M. de Carvalho, Spectral modeling of time series with missing data, 2013

Schoellhamer, D.H., Factors affecting suspended-solids concentrations in South San

532 Francisco Bay, California, J. Geophys. Res., 101(C5), 12087-12095, 1996.
533 Schoellhamer, D.H., Singular spectrum analysis for time series with missing data,
534 Geophys. Res. Lett. 28(16), 3187-3190, 2001.
535 Schoellhamer, D.H., Variability of suspended-sediment concentration at tidal to annual
536 time scales in San Francisco Bay, USA, Continental Shelf Research, 22, 1857-
537 1866, 2002
538 Shen, Y., W. Li, G. Xu, B. Li. Spatiotemporal filtering of regional GNSS network's
539 position time series with missing data using principal component analysis,
540 Journal of Geodesy, DOI 10.1007/s00190-013-0663-y, Vol.88: 1-12, 2014
541 Vautard, R., P. Yiou, and M. Ghil, Singular-spectrum analysis: A toolkit for short, noisy,
542 chaotic signals, Physica D, 58, 95-126, 1992.
543 Vautard, R. and M. Ghil, Singular spectrum analysis in nonlinear dynamics with
544 applications to paleoclimatic time series, Physica D, 35, 395-424, 1989.
545 Wang, X.L., J. Corte-Real, and X. Zhang, Intraseasonal oscillations and associated
546 spatial-temporal structures of precipitation over China, J. Geophys. Res.,
547 101(D14), 19035-19042, 1996.
548 Yiou, P., K. Fuhrer, L.D. Meeker, J. Jouzel, S. Johnsen, and P.A. Masked, Paleoclimatic
549 variability inferred from the spectral analysis of Greenland and Antarctic ice-
550 core data, J. Geophys. Res., 102(C12), 26441-26454, 1997.
551 Zotova, L.V., C.K. Shum, Multichannel singular spectrum analysis of the gravity field
552 from grace satellites, AIP Conf. Proc., 1206, 473-479, 2010