An Optimized Atmospheric Missed Data Recovery Algorithm Based on Singular Spectrum Iteration

Liu Wei, Jin Weidong
School of Information Science and Technology
Southwest Jiaotong University
Chengdu, P.R.China
ceocio@hotmail.com

Wang Huizan, Zhang Ren
Institute of Meteorology
PLA University of Science and Technology
Nanjing, P.R.China
zren63@126.com

***Abstract*—A new idea of interval quartering algorithm was proposed to improve the insufficiency of the conventional singular spectrum analysis iterative interpolation on parameter selection (including the number K of principal component and the embedding dimension M), and the applied test and comparative analysis recovery was carried out to the missing data. The experimental results showed that the improved method is very effective to the interpolation of missing data, and the computing speed of the improved algorithm is more rapid than that of conventional algorithm.**

***Keywords*—Singular Spectrum Analysis; Interpolation of missed data: Interval Quartering Algorithm**

I. INTRODUCTION

Missed data recovery is a useful technique for many sciences field. Some missed data recovery methods, such as statistic regression, Kriging interpolation, Kalman filtering, fractal interpolation and phase space reconstruction, are applied to the missed data recovery. Multivariate statistical analysis technologies[1], such as Empirical Orthogonal Function analysis (EOF), Principal Component-Canonical Correlation Analysis (PC-CCA), Singular Spectrum Analysis (SSA) and Multi-channel Singular Spectrum Analysis(MSSA), can be used to reveal the spatial correlative structure or temporal evolution of the scalar or vector fields, so been widely used in the time series analysis. How to integrate multivariate statistical analysis into the objectively accurate interpolation and filling of missing data has good prospects.

Aiming at the shortcomings and problems of general SSA and MSSA, a parameter optimization method called Interval Quartering Algorithm was proposed in this paper. It improved the conventional SSA/MSSA iterative interpolation methods. Interpolation experiments and comparative analysis were done by using it.

II. SSA/MSSA ITERATIVE INTERPOLATION

SSA/MSSA technique can be used for extracting out some simple modes containing important information from actual time series, and filtering out some random noises. The main idea of SSA iterative interpolation process is as follows: an inner-loop iteration is started by computing the leading empirical orthogonal function (EOF) of the centered, zero-padded data. Then the algorithm is performed again on the new time series in which the principal component corresponding to that EOF alone was used to obtain nonzero values in place of the missing point and correct the mean of the new time series. When this inner iteration has converged (the convergence of the iterative program has been proven mathematically [2]), an outer-loop iteration is performed by adding a second EOF for reconstruction and repeat the inner iteration. The embedding dimension M and the numbers of selected principal component K are optimized by the cross-validation method. Beckers et al. [3] discuss, in their Appendix A, how the bias introduced into the EOFs by missing data disappears as the iteration progresses.

The process of SSA iterative interpolation algorithm mainly contains the following two steps:

A. Optimization parameters of M and K by cross-validation

1) Give initial value $M=1, K=1$, and give the maximum embedding dimension value $M_{\max}$.

2) The data points of the original time series are classified as three categories: the training data X_{train} , the cross-validation data X_{cross_valid} and the missing data X_{fill} (i.e. when the points belong to X_{fill} , it means that we do not have data or they are unreliable). Thereinto, X_{train} , X_{cross_valid} are known data, but X_{cross_valid} are selected from the known data randomly and are seen as unknown data in the interpolation process. The effect of interpolation is evaluated by compare the interpolated values with the known values in X_{cross_valid} . X_{fill} are unknown data.

3) Let n=0, remove mean of time series X_{train} , record its average value X_{ave} , set X_{cross_valid} , X_{fill} to 0,and so we get the time series $X_n(t)$.

4) Perform SSA algorithm with embedding dimension M on the time series $X_n(t)$,select the former K principal components to reconstruct time series X_{recon} , replace the data X_{cross_valid} and X_{fill} with the data X_{recon} at its

corresponding position, and so we get the new time series $X_{n+1}(t)$.

5) If $\max\left|X_{n+1}(t)-X_n(t)\right|\le\varepsilon$, go to 6); otherwise, Let $n=n+1$ and return 4).

6) Let $X_{n+1}(t)=X_{n+1}(t)+X_{ave}$, and then compute the root mean square error $error(M,K)$ between the interpolated value of $X_{n+1}(t)$ at X_{cross_valid} and the known observed value at the same positions.

7) ① If $K<M$, then $K=K+1$, jump to ③; ② If $K=M$, then $M=M+1, K=1$;③ If $M=M_{\max}$, go to 8); Otherwise, return to 2) and start a new interpolation process.

8) Find out the optimal parameters M_{opt} and K_{opt} that makes the root mean square error is minimum. Program is over.

B. Interpolate missed data by SSA

1) The data points of the original time series are classified as two categories: the training data X_{train} (i.e. all known data including X_{cross_valid} thereinbefore) and the missing data X_{fill}.

2) Let n=0, remove mean of time series X_{train}, record its average value X_{ave}, set X_{fill} to 0, and so we get the time series $X_n(t)$.

3) Perform SSA algorithm with embedding dimension M_{opt} on the time series $X_n(t)$, select the former K_{opt} principal components to reconstruct time series X_{recon}, replace the data X_{fill} with the data X_{recon} at its corresponding position, and so we get the time series $X_{n+1}(t)$.

4) If $\max\left|X_{n+1}(t)-X_n(t)\right|\le\varepsilon$, $X_{n+1}(t)$ are the interpolation result of the time series, program is over; otherwise, Let $n=n+1$ and return to 3).

The process of MSSA iterative interpolation is similar to that of SSA.The differences between MSSA and SSA iterative interpolation are as follows: at the first step 4) and the second step 3), MSSA performs Multi-channel Singular Spectrum Analysis; at the second step 7), the condition that K should satisfy is not $K\le M$, but $K\le\min(M\times L,N')$.The specific algorithm of MSSA algorithm is omitted.

From said above, the deficiency of the conventional SSA iterative interpolation are as follows: because the optimization parameters M and K are exhausted much time, and for the total interpolation program, the goal of the first step is only providing the optimal parameters for the second step. Therefore, the deficiency of determining the optimal parameters should be improved.

III. AN IMPROVED ALGORITHM OF SSA/MSSA -INTERVAL QUARTERING ALGORITHM

To overcome the shortcomings of the general SSA/MSSA, an improved algorithm of SSA iterative interpolation-Interval Quartering Algorithm was presented.

A. The idea of Interval Quartering Algorithm

Firstly, the value range of independent variable,$[a_1,e_1]$,is divided into four small intervals with three divided points b_1,c_1,d_1,and then judge which one interval or two adjacent small intervals contain minimum of function (assume it is $[b_1,c_1]$);secondly, narrow the range of the interval to$[b_1,c_1]$, the interval$[b_1,c_1]$is subdivided into four smaller intervals and then judge which smaller interval containing minimum of function, and so on; repeat this process until no longer interval can be subdivided. At last, the values of each point in the interval that cannot be subdivided are computed, and then the location of the minimum point is found out.

How to find the interval where the minimum of the function is? Let five endpoints of four sub-intervals are a_i,b_i,c_i,d_i,e_i, and the corresponding function values are fa,fb,fc,fd,fe, then:

1) if $fb\le\min(fa,fc)$ *:* the minimum value must at interval$[a_i,c_i]$.

2) if $fc\le\min(fb,fd)$ *:* the minimum value must at interval$[b_i,d_i]$.

3) if $fd\le\min(fc,fe)$ *:* the minimum value must at interval$[c_i,e_i]$.

If all these three conditions are not satisfied, then $fb>\min(fa,fc)$, $fc>\min(fb,fd)$, $fd>\min(fc,fe)$ must be satisfied at the same time. Under these conditions, if $fa<fb$, according to the character of function $y=f(n)$ (in the character of "first decreases, and then increases" or "monotony"), there must be $fa<fb<fc<fd<fe$, the minimum value must at interval $[a_i,b_i]$; if $fa>fb$, according to the three inequalities $fb>\min(fa,fc)$, $fc>\min(fb,fd)$ and $fd>\min(fc,fe)$, then we can deduce $fa>fb>fc>fd>fe$, the minimum value must at interval $[d_i,e_i]$. Hence, two additional judgment conditions are as follows:

4) if $fb<fc$ *:* the minimum value must at interval$[a_i,b_i]$.

5) if $fc>fd$, the minimum value must at interval $[d_i,e_i]$.

The above five judgment conditions have included all sort situations.

Interval Quartering Algorithm has two major advantages:

1) Compared to the conventional method of finding the optimal parameters, Interval Quartering Algorithm is very timesaving. The time complexity of conventional algorithm is O(n), but that of Interval Quartering Algorithm is only O(log2n), the calculation speed was improved greatly.

2) It hardly fall into the local minimum. Because the algorithm is narrowing interval gradually, and the error function has only small fluctuations. For the large interval, small fluctuations do not affect the right selection of parameters. When the interval is reduced to a certain extent, computing all value of small interval and finding the minimum value can avoid the effect of small fluctuations.

B. The progress of Interval Quartering Algorithm

1) Let $i=1$,assign values to a_1 and e_1 (find minimum at the interval $[a_1, e_1]$); Let $fa=f(a_1)$, $fe=f(e_1)$, $flagC=0$ (if the middle point C needs recomputation,mark $flagC$ with 0).

2) Let $\tau=\left[\frac{e_i-a_i}{4}\right]$; if $\tau\geq 1$, then: ① $b_i=a_i+\tau$, $d_i=e_i-\tau$, $fb=f(b_i)$, $fd=f(d_i)$; ② if $flagc=0$,then $c_i=a_i+2\tau$, $fc=f(c_i)$; if $\tau<1$, then compute the value at points $a_i+1, a_i+2, ..., e_i-1$, and find the minimum from $a_i, a_i+1, ..., e_i-1, e_i$. Program is over.

3) if $fb\leq \min(fa, fc)$, then $a_{i+1}=a_i$, $c_{i+1}=b_i$, $e_{i+1}=c_i$, $fc=fb$, $fe=fc$, $flagC=1$, go to 4); if $fc\leq \min(fb, fd)$,then $a_{i+1}=b_i$, $c_{i+1}=c_i$, $e_{i+1}=d_i$, $fa=fb$, $fe=fd$, $flagC=1$, go to 4); if $fd\leq \min(fc, fe)$, then $a_{i+1}=c_i$, $c_{i+1}=d_i$, $e_{i+1}=e_i$, $fc=fc$, $fe=fd$, $flagC=1$,go to 4); if $fb<fc$,then $a_{i+1}=a_i$, $e_{i+1}=b_i$, $fe=fb$, $flagC=0$,go to 4); if $fc>fd$, then $a_{i+1}=d_i, e_{i+1}=e_i, fa=fd$, $flagC=0$,go to 4).

4) $i=i+1$, return to 2).

Interval Quartering Algorithm for MSSA is similar to SSA, so omitted.

IV. EXPERIMENT ON IMPROVED SSA/MSSA ALGORITHM

A. Area Coverage of Data

The NCEP/NCAR daily OLR (outgoing long wave radiation for describing cloud and convection activity) data was used in this paper, the data grid is 2.5 ° × 2.5 °,and rangement:90-140 ° E, 10 ° S-30 ° N, the temporal range of the time series is form May 1st ,2004 to April 30th,2006 (730 days in total), containing 260610 grid data.

B. OLR data interpolation experiment

1) Iteration Process Explanation: SSA iterative interpolation is for univariate time series, while MSSA iterative interpolation can handle multivariate time series problems. Therefore, we stretch the space grid points to one dimension, and the space grid points are seen as different variables. The Interval Quartering Algorithm is used to select parameters of MSSA iterative interpolation. We randomly select 40% of data from 260,610 grid points as missing data, and the remaining 60% as the known data. 10% of the known data (6% of the total data) are selected as cross-validation data, and the remaining 90% (54% of the total data) as training data.

Assign the parameter M=1, 2… in turn, and then apply the improved parameter selection method (Interval Quartering Algorithm) to search the corresponding optimal parameters K. According to the root mean square error of cross-validation data, we can identify the corresponding optimal parameter K for different embedding dimensions: M = 1, K = 32; M = 2, K = 52; M = 3, K = 68... , and get the corresponding root mean square error for different M values (see Table I).

TABLE I. CROSS-VALIDATION DATA ERROR COMPARISON OF OPTIMAL PARAMETER K IN DIFFERENT EMBEDDING DIMENSIONS M

Optimal parameters	correlation coefficient	root mean square error of missing data (W/S^2)
M=1,K=32	0.83780	22.864
M=2,K=52	0.85120	21.746
M=3,K=68	0.85076	21.880

M=1, EOF optimal iterative interpolation; M=2, MSSA optimal iterative interpolation.

The data of tab.1 showed that the MSSA iterative interpolation has a wider choice of parameters compared with EOF, and MSSA iterative interpolation method has an advantage over EOF iterative interpolation.

2) Interpolation Experiment Effect Analysis.

a) MSSA iterative interpolation effect and its comparison with EOF: In order to check up the interpolated effect of OLR spatial field, two days was selected for making the comparison between the actual fields and the recovery fields interpolation (Fig.1 and Fig.2).

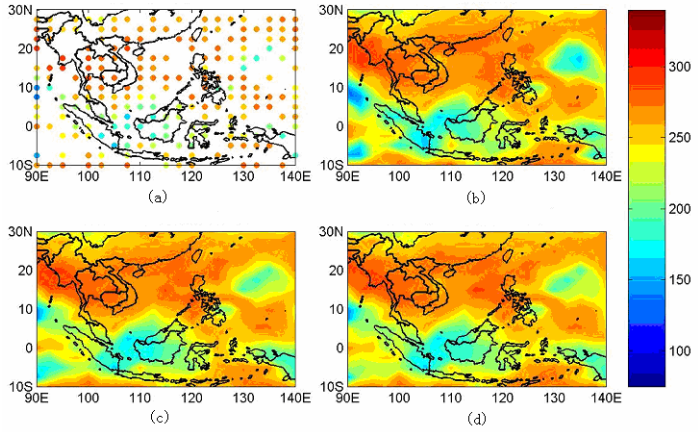

Figure 1. The comparison map of different OLR fields in December 12th, 2004. (a)OLR field missing 40% data; (b) actual OLR field; (c) reconstructed field by EOF; (d) reconstructed field by SSA

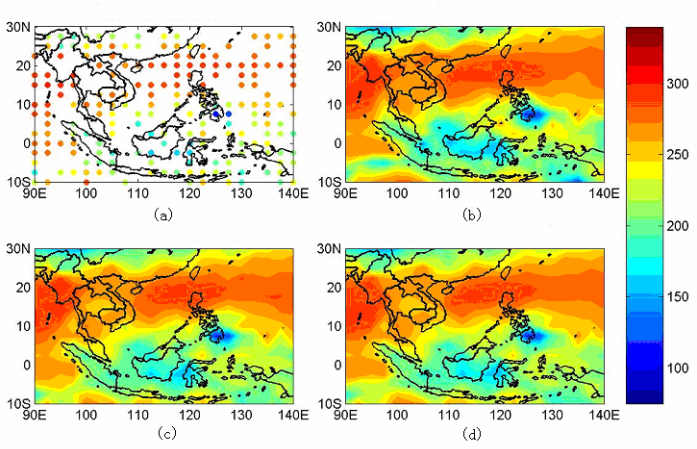

Figure 2. The comparison map of different OLR fields in April 5th, 2006. (a) OLR field missing 40% data ; (b) actual OLR field; (c) reconstructed field by EOF; (d) reconstructed field by SSA

It can be seen from Fig.1-2, EOF iterative interpolation is also a better interpolation technique.But the MSSA iterative interpolation is rather better than EOF,especially in detail describing. The differences between EOF and SSA showed clearly in Table II.

In order to check up the interpolated effect of OLR single-site time series, we take the time series at point [90 ° E, 30 ° N] as an example, and the correlation coefficient between EOF interpolated values and the corresponding poing actual time series is 0.7750, while that between MSSA interpolated values and the actual data time series 0.8251.

To further compare the effect of EOF iterative interpolation with that of MSSA iterative interpolation quantificationally, the correlation coefficient and the root mean square error between all interpolated values and the actual values for each method is computed (See Table 2). As can be seen from the table, the root mean square error of M = 2 (MSSA iterative interpolation method) is smaller than M = 1 (EOF iterative interpolation method), and the correlation coefficient of MSSA is higher (due to the large quantity of missing points containing 104,244 points, the result has good statistical significance), so the MSSA iterative interpolation is better than EOF.

TABLE II. THE INTERPOLATED EFFECT COMPARISON OF OPTIMAL PARAMETER K IN DIFFERENT EMBEDDING DIMENSIONS M

M and the corresponding optimal K	correlation coefficient	root mean square error of missing data (W/S^2)
M=1,K=32	0.84578	22.268
M=2,K=52	0.85530	21.483
M=3,K=68	0.85375	21.583

M=1, EOF optimal iterative interpolation ; M=2, MSSA optimal iterative interpolation.

b) Comparison between improved MSSA and general MSSA iterative interpolation: The superiority of the improved MSSA iterative interpolation compared with the conventional MSSA iterative interpolation is embodied mainly in computing time (see Table III) and calculation accuracy. It can be seen from Table III clearly that the computing speed of the improved MSSA iterative interpolation raises tens of times than that of the conventional MSSA iterative interpolation, and so the improved MSSA iterative interpolation has more obvious advantages, especially for large volume of data. As the conventional MSSA iterative interpolation needs more time, in order to reduce computing time, it often takes large time step of K and can not search the global optimal parameter K, which makes interpolated accuracy lower. By using the Interval Quartering Algorithm, the improved MSSA iterative interpolation can search the global optimal parameter K, the method makes the interpolated data is of high precision and accuracy. Therefore, the Interval Quartering Algorithm is a very effective algorithm for SSA/MSSA iterative interpolation, and it helps to improve the conventional SSA/MSSA iterative interpolation and develop the advantage of interpolation further.

TABLE III. THE COMPUTING TIME COMPARISON BETWEEN CONVENTIONAL AND IMPROVE MSSA ITERATIVE INTERPOLATION

M	conventional MSSA iteration that needs computing number of K	improved MSSA iteration that needs computing number of K	improved times of computing speed
1	357	19	17.8
2	714	22	31.5
3	728	22	32.1

V. SUMMARY

The general SSA/MSSA iterative interpolation can improve the interpolated accuracy compared with the EOF iterative interpolation, but the parameters selection contains some arbitrariness and blindness, which directly affect the quality of interpolated missing data and computational efficiency. Against the insufficiency of conventional SSA/MSSA iterative interpolation, this paper proposes an improved method of selecting the optimal parameters-Interval Quartering Algorithm. This method can effectively improve the efficiency and accuracy to SSA/MSSA iterative interpolation. Its main advantages are: (1) It is capable of finding the global optimal parameter to the error curve which has small local oscillation effectively; (2) The computing efficiency can be markedly improved by using SSA/MSSA iterative interpolation, the time complexity of general algorithm is O(n),but that of Interval Quartering Algorithm is only O(log2n), and the computing speed was greatly improved.

REFERENCES

[1] Wu Hongbao, Wu Lei. Methods for diagnosing and Forecasting Climate variability [M]. China Meteorological Press, Beijing, 2005(in Chinese).

[2] Jiang Zhihong, Ding Yuguo,Tu Qipu. Interpolation experiment of meteorological fields based on PC-CCA [J]. Journal of Nanjing institute of Meteorology, 1999, 22 (2):141-148(in Chinese).

[3] Beckers J M, Rixen M. EOF calculations and data filling from incomplete oceanographic datasets [J]. *Journal of Atmospheric and Oceanic Technology*, 2003, 20(12):1839-1856.