New reconstruction and forecasting algorithm for TEC data*

Wang Jun(王 俊)[a], Sheng Zheng(盛 峥)[b)c)†], Jiang Yu(江 宇)[b], and Shi Han-Qing(石汉青)[b]

[a] National Key Laboratory on Electromagnetic Environmental Effects and Electro-optical Engineering, PLA University of Science and Technology, Nanjing 210007, China

[b] College of Meteorology and Oceangraphy, PLA University of Science and Technology, Nanjing 211101, China

[c] Jiangsu Key Laboratory of Meteorological Observation and Information Processing, Nanjing University of Information Science and Technology, Nanjing 210044, China

(Received 6 December 2013; revised manuscript received 2 April 2014; published online 23 July 2014)

To reconstruct the missing data of the total electron content (TEC) observations, a new method is proposed, which is based on the empirical orthogonal functions (EOF) decomposition and the value of eigenvalue itself. It is a self-adaptive EOF decomposition without any prior information needed, and the error of reconstructed data can be estimated. The interval quartering algorithm and cross-validation algorithm are used to compute the optimal number of EOFs for reconstruction. The interval quartering algorithm can reduce the computation time. The application of the data interpolating empirical orthogonal functions (DINEOF) method to the real data have demonstrated that the method can reconstruct the TEC map with high accuracy, which can be employed on the real-time system in the future work.

Keywords: reconstruction, total electron content (TEC) data, empirical orthogonal function (EOF) decomposition, interval quartering algorithm

PACS: 94.20.Cf, 02.60.Ed **DOI:** 10.1088/1674-1056/23/9/099401

1. Introduction

In the whole solar-terrestrial space environment, the ionosphere couples tightly with the magnetosphere, lower and middle atmosphere in various forms. Any intense solar activity such as solar flares and coronal mass ejections can disturb the ionosphere. Meanwhile, ionosphere is a major region for human space activities, where a variety of processes like momentum, mass and energy transfer and severe disturbances, may have a significant impact on military activities and daily lives.[1,2]

The TEC can be retrieved from both Global navigation satellite system (GNSS) and ionosonde observations,[3,4] with special assumptions primarily about the plasma distribution in the topside ionosphere. The GNSS TEC is the slant TEC instead of vertical, and the thin-sheet approximation can introduce errors, which requires additional discussions. The ionosonde TEC has been existing for near half century while the GNSS has been introduced during recent years.[5,6] There are two main factors contributing to the missing data of TEC, one is that many ground based equipment do not update their digitization due to limited funds and age-old, the other is that some observations have been lost widely for quite a long time. The overall analysis and forecasting of space weather need a complete TEC database. Therefore, the reconstruction of the missing and invalid TEC data has important scientific value and practical significance.[7–9]

The ionosphere data inversion and reconstruction is a hot research topic in the world. The improved Kriging method with the Chinese ionospheric background as reference has been proposed, which is suitable for the construction of Chinese regional ionospheric TEC and critical frequency f_0F_2. This method reconstructs the regional ionospheric parameters by using the Kriging method with the introduction of the ionosphere distance and the regionalized variable.[1] The kalman filter of the Gauss Markov is used to reconstruct and assimilate TEC data.[3] However, the most drawbacks of the optimal interpolation and the Kriging interpolation are tremendous calculation, which need repeated priori information.[10,11] In order to solve the above problem, the data interpolating empirical orthogonal functions (DINEOF) method is proposed to reconstruct the missing data.[12–19] DINEOF method is self-adaptive and do not need any prior information, which has been used to successfully reconstruct the sea surface temperature (SST) in the cloud coverage via the satellite remote sensing. The physical basis of the DINEOF lies in that the EOF decomposition can extract several main modes and degrees of freedom that reflect the real physical variation from a large number of data, and then the retained optimal modes can be used to reconstruct the missing data efficiently. In this paper, based on the foregoing work, we introduce an interval quartering algorithm to reduce the computation time, and it shows that the precision and speed have been improved as compared to the Kriging interpolation method. After the reconstruction of TEC data is done, the TEC data can be forecasted with

*Project supported by the National Natural Science Foundation of China (Grant No. 41105013, 41375028, and 61271106), the National Natural Science Foundation of Jiangsu Province, China (Grant No. BK2011122), and the Key Laboratory of Meteorological Observation and Information Processing Scientific Research Fund of Jiangsu Province, China (Grant No. KDXS1205).

†Corresponding author. E-mail: 19994035@sina.com

© 2014 Chinese Physical Society and IOP Publishing Ltd

three steps. Firstly, the TEC data is decomposed with the EOF method. Secondly, non-linear time series forecasting of time field coefficient with genetic algorithm combined with the idea of chaotic prediction. Thirdly, with optimal reconstructed order, the space filed of EOF decomposition and the time filed of prediction is reconstructed. Finally, we can reduce the data noise and make forecast. The detailed process is beyond this paper, which will be presented in an accompanying paper.

The article is organized as follows. The methodology employed to reconstruct data is described in Section 2. The comparison with Kriging and DINEOF method to the simulation results are shown in Section 3.1, the real experiment data results are shown in Section 3.2, and the more details about comparison with Kriging and DINEOF method are shown in Section 3.3. A conclusion of the work is given in Section 4.

2. Principle of the DINEOF

Unlike the optimal interpolation method, which obtains the interpolating data by calculating the minimum error covariance, the DINEOF is an adaptive EOF decomposition method, which speculates the missing data through the empirical orthogonal function.[20]

2.1. Interpolation and adaptation of the EOFs

The $M \times M$ ($M > N$) matrix $\boldsymbol{X}_0$ is the original observation data, where M is spatial dimension and N is the temporal dimension. Each column of $\boldsymbol{X}_0$ corresponds a dataset of size M, and there are N observation data. The EOF decomposition of X_0 is written as $\boldsymbol{X}_0 = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathrm{T}}$, where $\boldsymbol{U}$ is $M \times r$, representing spatial decomposition and $\boldsymbol{V}$ is $N \times r$, representing temporal decomposition, $\boldsymbol{D}$ is a $r \times r$ ($r \leq \min(M,N)$) diagonal matrix filled with eigenvalues. Having value of the matrix $\boldsymbol{U}$, $\boldsymbol{D}$, and $\boldsymbol{V}$, we can resolve $\boldsymbol{X}_0$ via reconstruction formula $\boldsymbol{X}_0 = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathrm{T}} = \sum_{i=1}^{k} \rho_i U_i V_i^{\mathrm{T}}$, where ρ_i is the n-order eigenvalue and k ($k \leq r$) is order.

The method EOF data reconstruction presented above assumes that the matrix $\boldsymbol{X}_0$ is perfectly and completely known. Nevertheless, the amount of data has not been observed or invalid. Thus, additional processes are necessary for the missing data sets. Large-scale EOFs are not influenced by local changes in the values of a few points, and so we could try to acquire the missing data through the iterative EOF. The method can be described as follows. Suppose that $\boldsymbol{I}$ is the ensemble of discrete points where data are missing or invalid, when $[(i,j) \in I]$, we do not have data or they are invalid, and the number of missing data points is n_0. $\boldsymbol{X}_0$ is the observation data matrix with missing or invalid data. The anomaly matrix of the matrix $\boldsymbol{X}_0$ is $\boldsymbol{X}$, and it can be written as $\boldsymbol{X} = \boldsymbol{X}_0' - \overline{\boldsymbol{X}_0}$, where $\boldsymbol{X}_0'$ is the matrix that is calculated by setting the missing data in the original matrix $\boldsymbol{X}_0$ to be zero, and $\overline{\boldsymbol{X}_0}$ is the mean value of $\boldsymbol{X}_0$.

First of all, we use the EOF decomposition of this matrix to obtain the first estimation of the spatial and temporal EOFs $\boldsymbol{U}$ and $\boldsymbol{V}$ as well as their amplitudes D. Then the missing data $[(i,j) \in I]$ can be filled by the truncated EOF reconstruction, and it can be expressed as

$$(\boldsymbol{X}_{\mathrm{a}})_{ij} = (\boldsymbol{U}_N \boldsymbol{D}_N \boldsymbol{V}_N^{\mathrm{T}})_{ij} = (\boldsymbol{X}^{\boldsymbol{e}})_{ij} = \sum_{k=1}^{N} \rho_k (U_k)_i (V_k^{\mathrm{T}})_j, \quad (i,j) \in I, \tag{1}$$

where $\boldsymbol{X}_{\mathrm{a}}$ is the outcome of reconstruction of missing data. For any of these matrices, $\boldsymbol{X}^{\mathrm{e}}$ is the reconstruction of the corresponding matrix by the first N EOFs, $\delta\boldsymbol{X}$ is the modified matrix in each iteration process. The above reconstruction formula only keeps the first N order optimal EOFs. We thus obtain the truncated reconstruction of the matrix as the missing data points. Now, we obtain an analyzed field $\boldsymbol{X}_{\mathrm{a}}$. Compared with $\boldsymbol{X}$, the modified part is $\delta\boldsymbol{X}$, where the matrix $\delta\boldsymbol{X}$ is zero everywhere, except at the missing points. The relation between $\boldsymbol{X}_{\mathrm{a}}$ and $\boldsymbol{X}$ can be written as

$$\boldsymbol{X}_{\mathrm{a}} = \boldsymbol{X} + \delta X. \tag{2}$$

When the missing data points are filled by their new values, the same procedure can be applied again, because the EOFs as well as their amplitudes will be modified by the changes at the missing data points. The following set of repetition is performed until convergence, and it can be expressed as

$$\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathrm{T}} = \boldsymbol{X}_{\mathrm{a}}, \tag{3}$$

$$(\boldsymbol{X}_{\mathrm{a}})_{ij} = (\boldsymbol{U}_{\mathrm{N}} \boldsymbol{D}_{\mathrm{N}} \boldsymbol{V}_{\mathrm{N}}^{\mathrm{T}})_{ij} = (\boldsymbol{X}^{\mathrm{e}})_{ij} = \sum_{k=1}^{N} \rho_k (U_k)_i (V_k^{\mathrm{T}})_j, \quad (i,j) \in I. \tag{4}$$

It is important to analyze how the changes from one iteration to the next are structured. When the changes are very small, we can consider that the iteration is convergent.[24] The criteria of convergence is that the changes from one iteration to the next are less than 10^{-8}. More details about convergence criteria can be found in Ref. [21].

2.2. Determination of the optimal number of EOF

In the above section, we have the order N for the EOF data reconstruction. However, the optimal N is a key point. On one hand, with a rather small value of N, the reconstructed data could not reflect the inner physical distribution. On the other hand, a fairly large value of N will make errors even bigger and the calculation slower. Here we determine the optimal EOF order by adopting the cross validation algorithms based on the Monte Carlo as well as the objective analysis.[22,23] Referring to the empirical conclusions by Alvarez *et al.*, for an $M \times N$ observation matrix, we produce $\min(0.01 \times M \times N + 40, 0.03 \times M \times N)$ random data and replace some of them with valid data. Since these random

data are known, we can ascertain the error between the reconstructed and known data with different orders of N. The optimal reconstructed order N corresponds to the minimum error.

In the traditional algorithm, the optimal order N is selected one cycle in turn by comparing the error value of $N = i$ and $N = i+1$. The method to determine the optimal order N is largely arbitrary and blind, and the computation also costs much. In order to solve the problem, the interval quartering algorithm[21] is introduced to determine the optimal order N that is based on the overall change tendency and local existence fluctuation characteristic of the error curve. The error function is set for $e = f(N)$, where the function variable N is integer. The overall changing trend of the function is that it is first reduced, then increases, and exists between small fluctuations, and the purpose is to calculate the minimum value of the function in the integer interval $[a,e]$.

The basic idea of interval quartering algorithm is as follows: the interval $[a,e]$ is averagely divided into four small rooms by interval differentiate points b, c, and d, and determines the minimum value belong to which village room or which two adjacent village room. Then the new small room is subdivided into four more small rooms, and determine the minimum value belong to which village room or which two adjacent village room, etc. This process is repeated until the interval region cannot be segmented. At this time, the value of each point in the small room is computed, and find out the minimum point location.

The whole reconstruction procedure of the DINEOF is shown in Fig. 1.

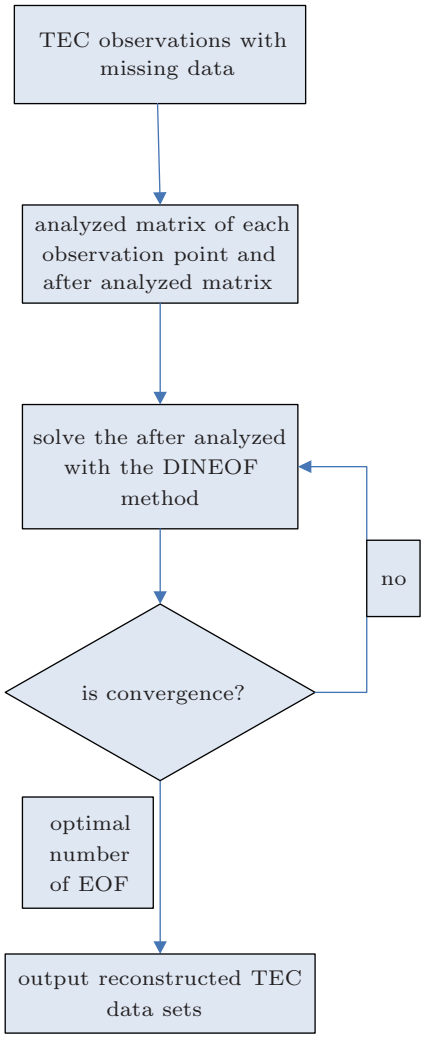


Fig. 1. (color online) Flow of reconstructed TEC data that includes missing points.

3. Analysis of experimental results

The global TEC data of the whole year of 2007 is used, which is provided by the GNSS observation network of Chinese Earthquake Administration. The TEC data are derived from the GNSS and ground ionosonde measurements, and are updated by two hours. The area of TEC data is from 180 °W (the same in the following texts) to 180 °E in longitude and from 87.5 °S to 87.5 °N in latitude. The resolution is 5° in longitudinal direction and 2.5° in the longitudinal direction.

In order to evaluate the precision of DINEOF, the root mean square error (RMS error) of reconstruction data is employed for the evaluation of the quality of the reconstruction. There are two evaluations: one is the reconstruction based on the simulation data, with the ratio of missing data being 30%, 50%, and 70%, respectively; and the other is the reconstruction based on the real experimental data.

3.1. Reconstruction precision analysis for simulation experiment

TEC distribute in three dimension as TEC(Lat,Lon,UT), the size of data is $71 \times 73 \times 4380$, the TEC image is updated every two hours. Since the change in TEC is closely related to the local time, the TEC change cycle is about one day for short-period variation, which provides a good opportunity to make advantage of temporal correlation of TEC data. Using the DINEOF method can not only reconstruct missing data both spatially and temporally, but also improves overall reconstructed precision, so it may be better than ordinary spatial interpolation methods (such as Kriging method, etc.). Three reconstructed results of the DINEOF method with data missing ratio being 30%, 50%, and 70% respectively are shown in Fig. 2, and for the Kriging method are presented in Fig. 3. Navy blue color on the left represents data missing. On the right is the reconstructed TEC image. The TEC value is shown as color with unit of TECU. Three pairs of TEC images are at 6:00UT on 1 June 2007. It can be seen from Fig. 2 that the reconstructed TEC exhibits global changes of ionosphere, with the real physics remained well. The comparison between Figs. 2 and 3 suggests that it is hard for the Kriging method to achieve the ideal result when the data missing ratio is high. Nevertheless, the DINEOF algorithm achieves better spatial and temporal characteristics as compared to the empirical or-

thogonal functions. The missing TEC data are effectively reconstructed by retaining the best modes, and the temporal and spatial characteristic distributions of reconstruction data are maintained well.

To illustrate the different reconstruction precisions with different number of EOFs employed, the optimal number of EOFs for different data missing percentage can be seen from Fig. 4. For instances, when 50 percent of data is missing and the optimal number of EOFs retained is 19, and the reconstruction precision can be as high as 0.83 TECU. It can be seen from Fig. 4, when the data missing percentage is high, the number of EOFs is larger. For instance, for 10 percent of data missing, the optimal EOF number is set to 13, for 50 percent it is 19, and for 80 percent it is 21.

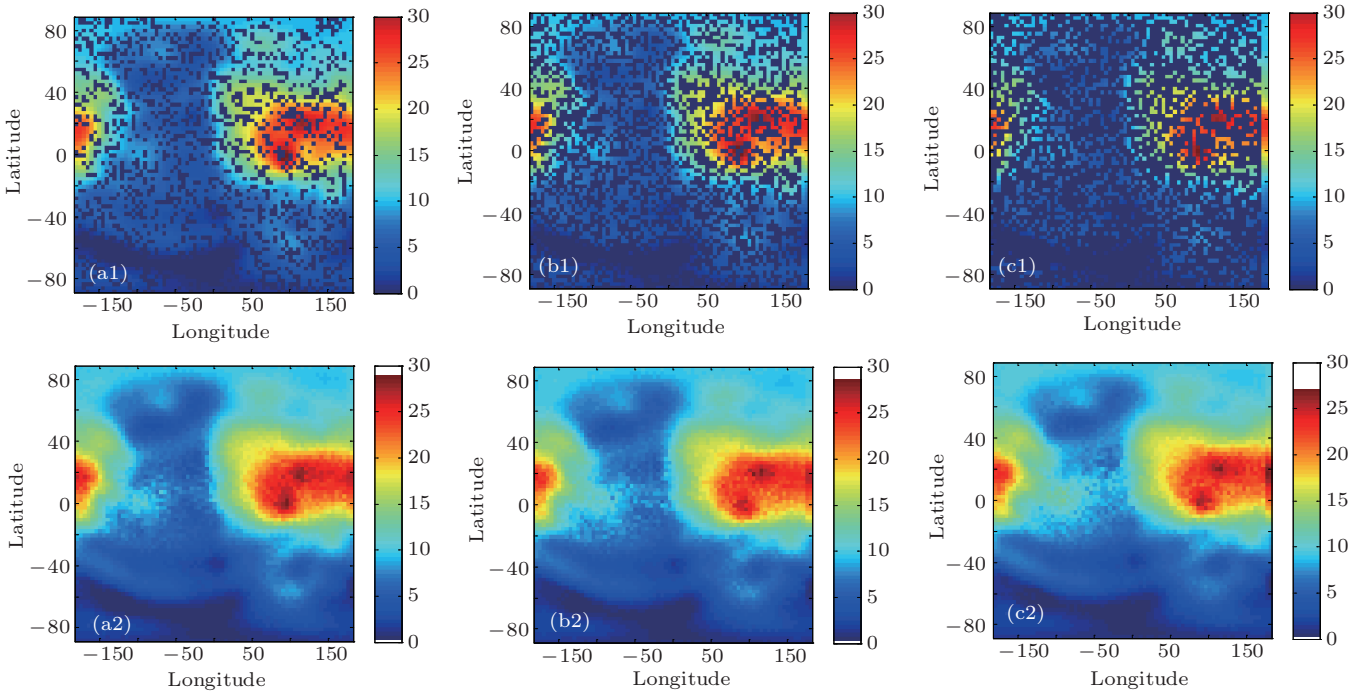


Fig. 2. (color online) (a) TEC images with (a) 30%, (b) 50%, and (c) 70% missing data reconstructed from the DINEOF method.

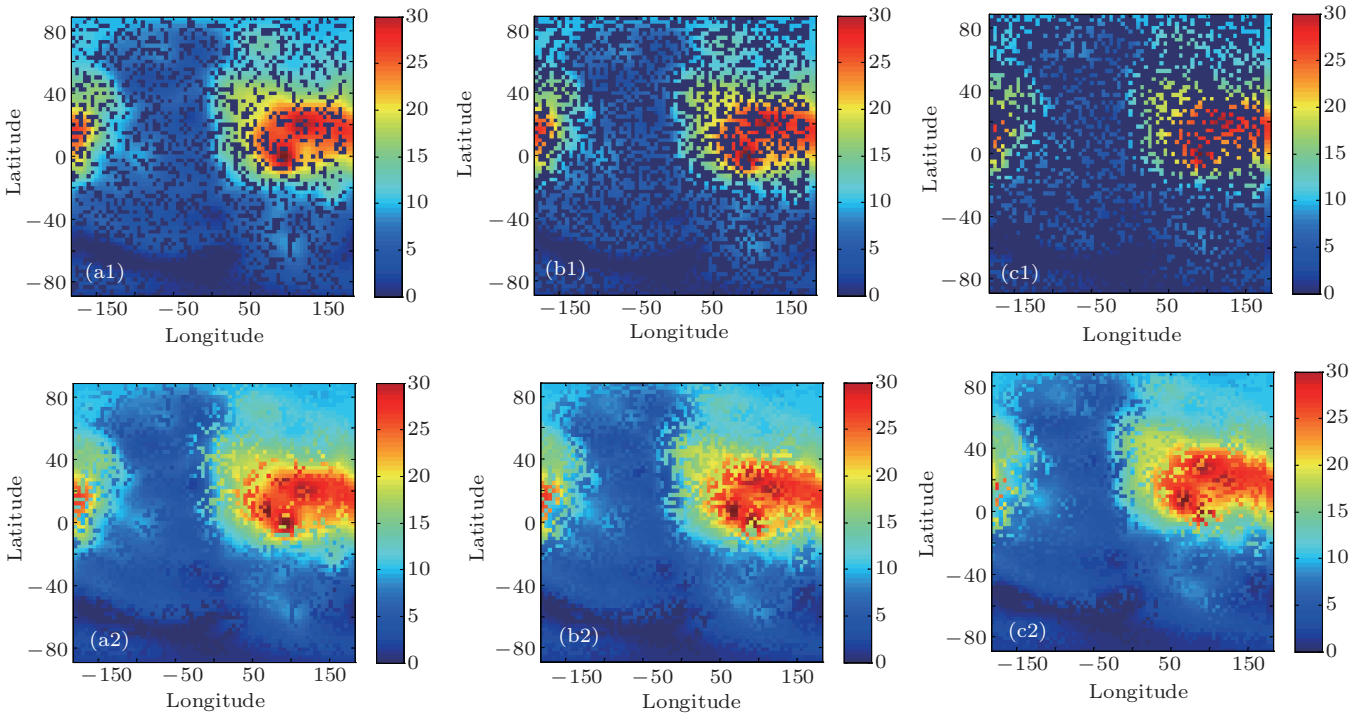


Fig. 3. (color online) TEC images with (a) 30%, (b) 50%, and (c) 70% missing data reconstructed from the Kriging method.

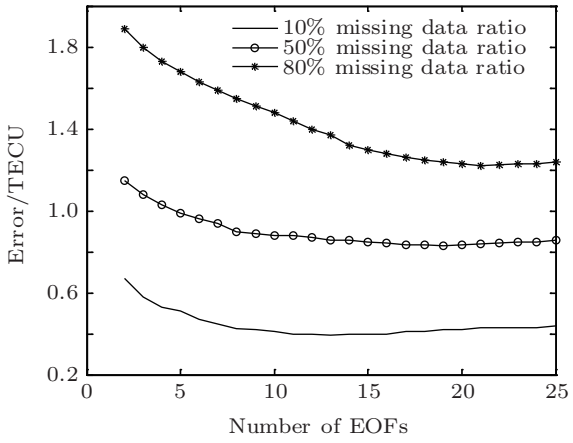


Fig. 4. Real error in the function of the number of EOFs retained.

3.2. Reconstruction precision analysis based on the real experimental data

There are no effective observations from 18:00UT to 24:00UT 20 on August 2007 in northwest China from the GNSS observation network of Chinese Earthquake Administration. The regions with missing TEC data are from 65 °E to 115 °E in longitude and 32.5 °N to 52.5 °N in latitude, the spatial range in longitude direction is 50° and in the latitude direction is 20°. The TEC data of four observation time periods (18:00UT, 20:00UT, 22:00UT, 24:00UT) are missing. The instance of missing data is representative in space and time distribution, and this paper as an example is to illustrate the DINEOF reconstruction effect for real experimental data. The observed TEC data from Chinese Academy of Sciences are used to evaluate the reconstruction results. The originality and reconstruction results can be seen in Fig. 5. Navy blue color on the left represents data missing, and on the right is the reconstructed TEC image. It can be seen from Fig. 5 that the area of data missing is larger and time of missing data is long, but the good reconstruction results can be obtained from the DINEOF method.

In order to further illustrate the difference between the Kriging method and the DINEOF method for the real experiment data, the reconstruction TEC image of the Kriging method is listed in Fig. 6. Comparing Fig. 5 with Fig. 6, it can be seen that the precision of DINEOF method is better than that of Kriging method, and the distribution and variation of TEC data physical characteristics is maintained well for the DINEOF method. Further statistics can show that the reconstruction error of the Kriging method is 1.06TECU for real experiment data, and the reconstruction error of the DINEOF method is 0.65TECU. Therefore, the DINEOF reconstruction accuracy is significantly better than the Kriging method for real experiment data reconstruction.

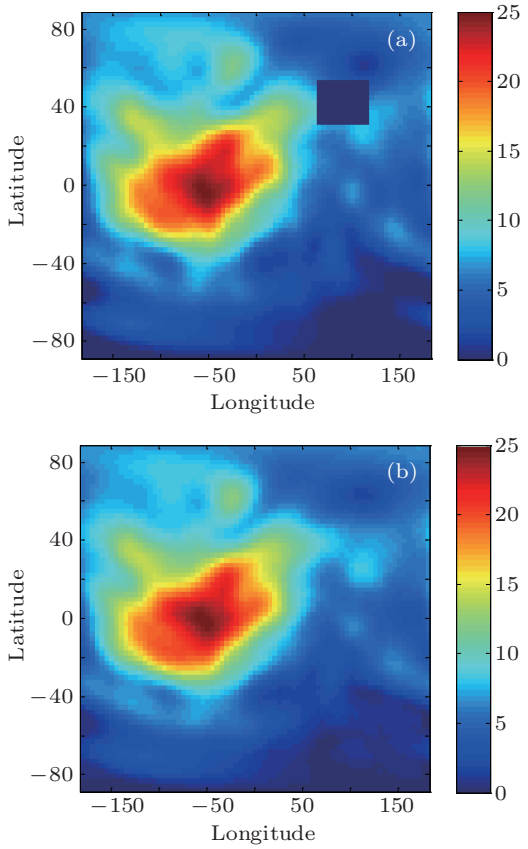


Fig. 5. (color online) (a) Real TEC images with missing data and (b) those reconstructed from the DINEOF method.

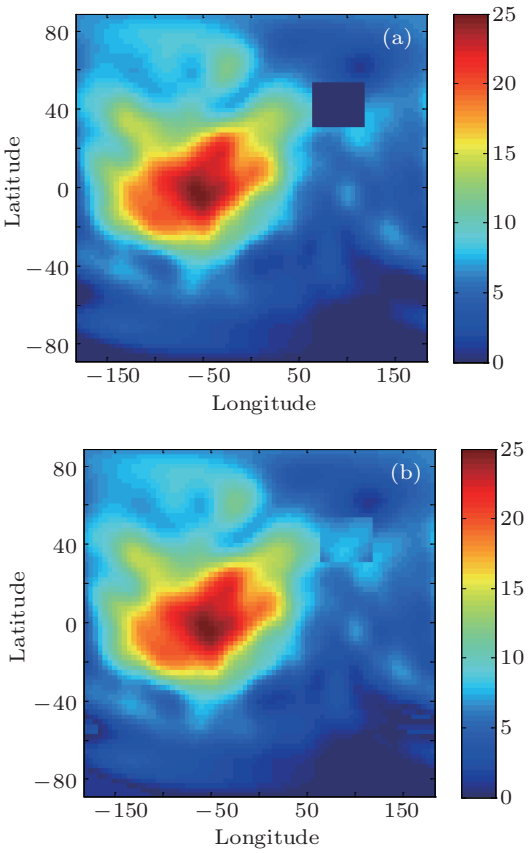


Fig. 6. (color online) (a) TEC images with missing data and (b) reconstruction of Kriging method for real experimental data.

3.3. More details about comparison with Kriging and DINEOF method

In order to show the advantages and disadvantages of DINEOF method and Kriging method, the reconstruction precision and computational time for the two methods were compared. As the interval quartering algorithm can only improve DINEOF computational efficiency, with no effects on the precision of DINEOF algorithm, the interval quartering algorithm is considered only for the improvement of the computation efficiency in comparison. Our computation equipment is a ThinkPad T420 equipped with four CPUs (I5-2520-m, 2.50 GHz) and 4 GB of EMS memory bank.

The reconstruction errors of the DINEOF and Kriging method for Simulation experiment are listed in Table 1 with the unit of TECU. The reconstruction precision of two methods for different data missing ratio is shown in Fig. 7 at the same time. It can be seen from Fig. 7, the reconstruction precision of Kriging method is better than that of the DINEOF method when the data missing ratio is about less than 30%, and the reconstruction precision of the DINEOF method is better than that of Kriging method when the data missing ratio is about greater than 30%. Thus, the DINEOF method is more suitable for the reconstruction of large data missing.

The computation times for the DINEOF and Kriging methods for simulation experiment are listed in Table 2, and the unit is s. In order to illustrate the improvement of computing efficiency for interval quartering algorithm, we consider two situations for DINEOF-IQA (DINEOF that with interval quartering algorithm) method and DINEOF method. It can be seen from Table 2 that the interval quartering algorithm can improve the computational time a lot.

Table 1. Reconstruction error of the DINEOF and Kriging method for simulation experiment.

Algorithm	TEC data missing ratio							
	10%	20%	30%	40%	50%	60%	70%	80%
Kriging	0.22	0.42	0.62	0.82	1.01	1.22	1.42	1.62
DINEOF	0.39	0.48	0.58	0.75	0.83	0.96	1.09	1.22

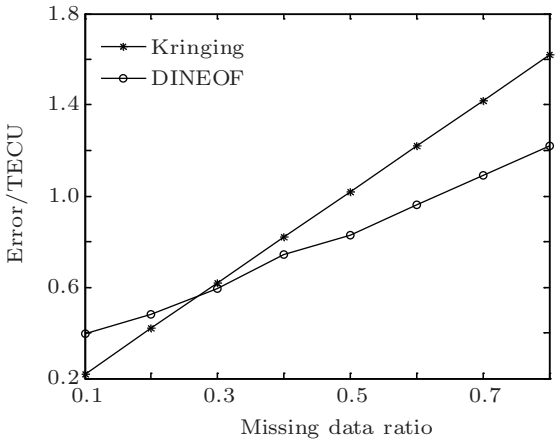


Fig. 7. Reconstruction precision for different methods.

Table 2. Computation time of DINEOF and Kriging method for simulation experiment.

Algorithm	TEC data missing ratio							
	10%	20%	30%	40%	50%	60%	70%	80%
Kriging	180.8	219.7	261.5	303.9	336.3	373.8	415.8	447.7
DINEOF	120.5	123.6	127.2	129.9	133.2	135.8	138.1	141.4
DINEOF-IQA	42.6	43.7	44.5	45.9	53.7	59.2	62.1	66.6

4. Conclusion

The DINEOF algorithm can be successfully applied to large database matrix reconstruction with good accuracy. It does not require any priori statistical error information and is proved to be stable. In the reconstruction of global TEC data of the whole year 2007, with the huge size of matrix of 22701540, we introduce the interval quartering algorithm to greatly improve the calculation speed and computation efficiency. In practice, the DINEOF algorithm has successfully reconstructed the global TEC data and the precision is 0.65TECU for real experimental data and 0.83TECU for simulation experimental data when 50 percent of data are missing.

The comparison between Kriging and DINEOF method suggests that the Kriging method can achieve better reconstruction results when the data missing ratio is less than about 30%, and the DINEOF can achieve the ideal result when the data missing ratio is higher. The DINEOF method can make full use of spatial and temporal relationship of TEC data distribution, and it can achieve better TEC data's physical characteristics. In the following research work, forecasting and additional optimization work such as improvement of convergence criteria in iterated reconstruction is necessary for dealing with much more larger data matrix. We can reconstruct other ionospheric remote sensing data like critical frequency f_0F_2. To reconstruct the critical frequency f_0F_2 is not easy, because its distribution not only tends to be clump but also its spatial correlation is weak. So, traditional interpolations such as the optimal interpolation or Kriging interpolation fail to reconstruct the missing data. In contrast, the characteristics of DINEOF make it suitable for the reconstruction of such data.[25]

References

[1] Liu R Y, Liu G H and Wu J 2008 *Chin. J. Geophys.* **51** 300 (in Chinese)
[2] Liu R Y, Xu Z and Wu J 2005 *J. Atmos. Sol. Terr. Phys.* **67** 1129

[3] Yue X A, Wan W X and Liu L B 2010 *Chin. J. Geophys.* **53** 787 (in Chinese)
[4] Zeng M, Wang W L and Bai X D 2013 *Chin. Phys. B* **22** 098105
[5] Guo P, Kuo Y H, Sokolovskiy S V and Lenschow D H 2011 *J. Atmos. Sci.* **68** 1703
[6] Guo P, Xu X and Zhang G X 2011 *J. Atmos. Sol-Terr. Phys.* **73** 839
[7] Sheng Z 2012 *Acta Phys. Sin.* **61** 219401 (in Chinese)
[8] Ma L M and Wu Z M 2010 *Chin. Phys. B* **19** 010201
[9] Sheng Z and Fang H X 2012 *Chin. Phys. B* **21** 029301
[10] Emery W and Thomson R 1998 *Data Analysis Methods in Physical Oceanography* (Pergamon: Elsevier Science Pub. Co.)
[11] Everson R, Cornillon P and Sirovich L 2002 *J. Phys. Ocean.* **27** 468
[12] Sheng Z and Fang H X 2013 *Chin. Phys. B* **22** 029301
[13] Sheng Z 2013 *Chin. Phys. B* **22** 029302
[14] Alvarez A, Barth A and Rixen M 2005 *Ocean Model.* **17** 325
[15] Beckers J and Rixen M 2003 *J. Atmos. Ocean Technol.* **20** 1839
[16] Chen L P and Lü X X 2009 *Chin. Phys. B* **18** 189
[17] Wan W X, Ding F and Ren Z P 2012 *Sci. China Tech. Sci.* **55** 1161
[18] Zhang A J, Ridley Z X and Hao Y 2012 *J. Geophys. Res.* **117** 03328
[19] Mao T, Wan W, Yue L S, Zhao B and Guo J 2008 *Radio Sci.* **43** 3629
[20] Bennett A 2002 *Inverse Modeling of the Ocean and the Atmosphere* (Cambridge: Cambridge University Press)
[21] Toumazou V and Cretaux J 2001 *Mon. Weath Rev.* **125** 1243
[22] Alvarez A 2003 *J. Atmos. Ocean Technol.* **20** 717
[23] Beckers J, Barth A and Alvarez A 2006 *Ocean Sci.* **16** 183
[24] Wang H Z, Zhang R, Liu W, Wang G H and Jin B G 2008 *Appl. Math. Mech.* **29** 1351
[25] Alvarez A, Barth A and Beckers J 2007 *JGR* **112** 3008