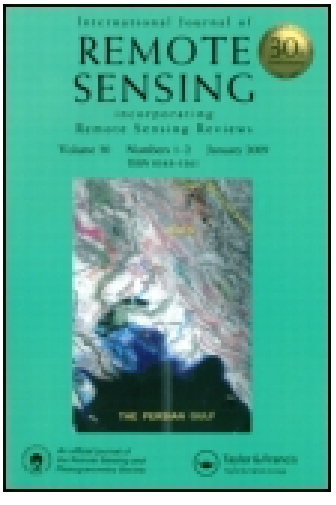

International Journal of Remote Sensing
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tres20

Reconstruction of satellite chlorophyll-a data using a modified DINEOF method: a case study in the Bohai and Yellow seas, China

Yueqi Wang[ab] & Dongyan Liu[a]
[a] Key Laboratory of Coastal Zone Environmental Processes and Ecological Remediation, Yantai Institute of Coastal Zone Research, Chinese Academy of Sciences, 264003 Yantai, Shandong, PR China
[b] University of the Chinese Academy of Sciences, 100049 Beijing, PR China
Published online: 04 Dec 2013.

To cite this article: Yueqi Wang & Dongyan Liu (2014) Reconstruction of satellite chlorophyll-a data using a modified DINEOF method: a case study in the Bohai and Yellow seas, China, International Journal of Remote Sensing, 35:1, 204-217, DOI: 10.1080/01431161.2013.866290

To link to this article: http://dx.doi.org/10.1080/01431161.2013.866290

International Journal of Remote Sensing, 2014
Vol. 35, No. 1, 204–217, http://dx.doi.org/10.1080/01431161.2013.866290

Reconstruction of satellite chlorophyll-*a* data using a modified DINEOF method: a case study in the Bohai and Yellow seas, China

Yueqi Wang[a,b] and Dongyan Liu[a]*

[a]*Key Laboratory of Coastal Zone Environmental Processes and Ecological Remediation, Yantai Institute of Coastal Zone Research, Chinese Academy of Sciences, 264003 Yantai, Shandong, PR China;* [b]*University of the Chinese Academy of Sciences, 100049 Beijing, PR China*

(*Received 8 October 2012; accepted 4 June 2013*)

A data-interpolating empirical orthogonal function (DINEOF) method was applied to 8 day composited satellite-derived chlorophyll-*a* (chl-*a*) images to produce a long-term, cloud-free chl-*a* data set over the Bohai Sea and Yellow Sea from 1997 to 2010. In this study, two additional procedures, a depth subdivision scheme and a new process of outlier detection and removal, improved the overall performance of this interpolating technique. The whole chl-*a* data set was divided into three subsets according to 20 and 50 m isobaths and the DINEOF reconstruction was performed on each subset. This subdivision scheme can significantly improve the accuracy of reconstruction, but is achieved with loss of computational efficiency due to the increased number of iterations required for reconstruction of the three subsets. A simple and new outlier detection method based on standardized residuals theory was developed to eliminate the spurious values (outliers) from the chl-*a* data set. The accuracy of the DINEOF reconstruction was significantly improved by the application of the outlier detection and removal process.

1. Introduction

Extensive space and time coverage of satellite images has made them indispensable to the study of oceanographic dynamics of marine ecosystems (Borzelli et al. 1999; Vantrepotte and Mélin 2011; Swardika, Tanaka, and Ishida 2012). Most satellite remote-sensing products are retrieved using spectral data from the visible and infrared bands, e.g. sea-surface temperature (SST) and sea-surface chlorophyll-*a* (chl-*a*) concentration. These satellite data sets commonly present with large-scale missing data due to cloud coverage or malfunctions in the satellite sensors. The extent of missing data can be higher than 95% or indeed completely missing over the study area. Incomplete satellite data sets significantly restrict their use in studying physical and biological ocean processes at both global ocean and regional scales. For example, a completed data set is necessary for much statistical analysis (e.g. cluster analysis and empirical orthogonal function (EOF) analysis) and use in forcing physical models.

Several methods have been used in the reconstruction of missing data from marine satellite images. Spline interpolation (Everson et al. 1997) and optimal interpolation (OI) (He et al. 2003; Hoer and She 2007) have been used to recover missing data and reconstruct SST and sea level anomaly (SLA) data sets (Fieguth et al. 1998). Beckers and Rixen (2003) and Alvera-Azcarate et al. (2005) described a technique for filling missing data based on the EOF algorithm, called the data-interpolating EOF (DINEOF).

*Corresponding author. Email: dyliu@yic.ac.cn

Compared to other interpolating methods, the DINEOF method is a self-consistent, parameter-free technique for reconstruction of gappy data, is computationally efficient, and presents the advantage of not requiring *a priori* information (Beckers and Rixen 2003; Alvera-Azcarate et al. 2005).

The cloud-gap filling techniques for ocean colour data sets are less developed than those for SST, perhaps because satellite chl-*a* data have become easily available only comparatively recently and the accuracy of chl-*a* retrieval is poor relative to SST and some other parameters. In recent years, several studies have attempted to reconstruct incomplete chl-*a* data using various methods, for example kriging (Saulquin, Gohin, and Garrello 2011) and other OI methods (Pukhtyar, Stanichny, and Timchenko 2009). Recently, the DINEOF method has been efficiently applied to reconstruct the missing data in chl-*a* data sets. For example, Alvera-Azcarate et al. (2007) used a multivariate approach based on the DINEOF method to reconstruct missing data in chl-*a*, SST, and sea-surface wind data. Sorjamaa et al. (2010) presented an improved version of the DINEOF method with an EOF pruning method and this modification can improve the accuracy of reconstruction while being computationally efficient. These limited previous studies indicated that the application of the DINEOF method in ocean colour data set reconstruction needs more testing and modification in order to yield more accurate results.

This study significantly improves the application of the DINEOF method. We present an application of the DINEOF method to reconstruct missing data in a long-term satellite chl-*a* data set over the Bohai and Yellow Seas (Figure 1). The study area is a shallow continental shelf with typical case II waters where inaccuracies in interpreting chl-*a*

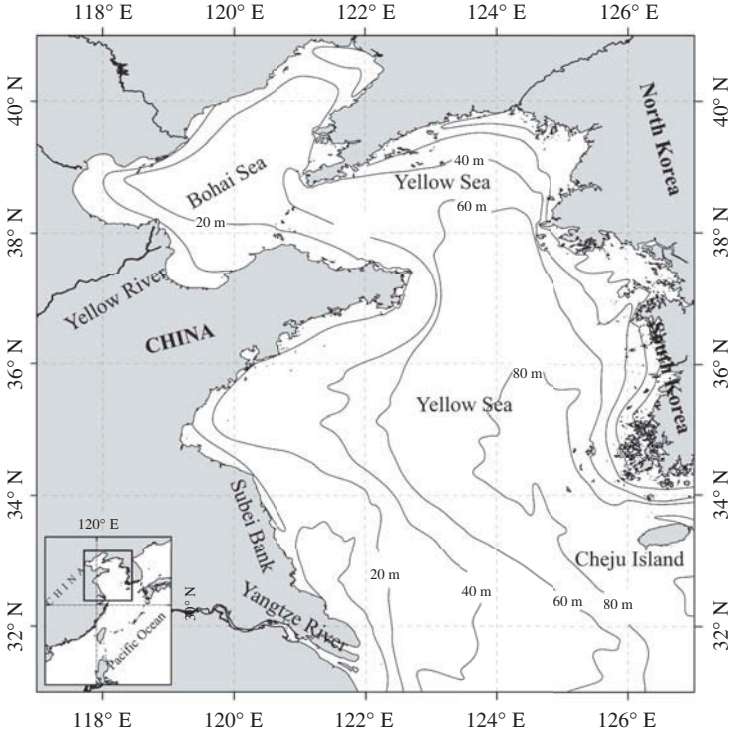

Figure 1. Bathymetric and geographic map of the study area.

products may be significant due to dissolved and suspended matter and atmospheric correction problems (Ruddick, Ovidio, and Rijkeboer 2000; Sun, Guo, and Wang 2010). These problems induce unacceptable errors (regarded as outliers) in the chl-*a* data set, making this area a perfect test site to evaluate the stability of the DINEOF method. In applying this method, we present here two additional procedures to improve the accuracy of the DINEOF method. First, we propose a modification of the ordinary DINEOF method, which involves a subdivision scheme using 20 and 50 m isobaths and second, a new outlier detection and removal method based on standardized residuals theory was performed and evaluated.

2. Data and methods

2.1. Satellite chlorophyll data set

The satellite-derived sea-surface chlorophyll-*a* (chl-*a*) concentration data set for this study comprises 8 day composite Level-3 global standard mapped images (SMIs) derived from the Sea-viewing Wide Field-of-view Sensor (SeaWiFS). These satellite images were obtained from the Ocean Biology Processing Group (OBPG) of the Goddard Space Flight Center (GSFC) (http://oceancolor.gsfc.nasa.gov/) in compressed hierarchical data format (HDF). The basic algorithm described by O'Reilly et al. (1998) and O'Reilly et al. (2000) was used to calculate the sea-surface chl-*a* concentrations and the reprocessing version is R2010. The time span of the data set is from September 1997 to December 2010 and the geographic area covers 117–127° N and 31–41° E (Figure 1). The initial size of this data set is 120 × 120 pixels and 609 images, with a resolution of 9 km × 9 km. Some of these images present extreme examples of missing data (more than 95%) and some of the pixels with less than 5% good data in temporal dimension were eliminated. Finally, a derived, hereafter termed the 'original' data set with 562 images and 7441 spatial pixels (covering about 94% of the sea area), was kept for the analysis. Since the satellite chl-*a* values spanned three orders of magnitude and chl-*a* retrievals are often log-normally distributed (Campbell 1995), raw chl-*a* data were $\log_{10}$-transformed prior to reconstruction and analysis in order to homogenize the variance and yield a nearly normal data distribution.

2.2. Algorithm of the DINEOF method

The DINEOF method is a self-consistent method for the reconstruction of missing data in oceanographic data sets. The DINEOF method was applied as follows in this study.

(1) The original data set was stored in the initial matrix with $m \times n$ dimensions, where m is the number of pixels and n the number of images; this matrix contains both existing and missing data. First, 3% of the existing data in the matrix were randomly selected and initially set aside (deemed as missing in the initial matrix) for progressive cross-validation. The mean temporal spatial value was then subtracted from the matrix and the missing data (including cross-validation data) set to zero.
(2) For reconstruction of the missing data, The EOF decomposition was computed by the singular value decomposition (SVD) method, and the spatial EOFs (**U**), singular values matrix (**S**) and temporal EOFs (**V**) were obtained. The missing data can be reconstructed by the truncated EOFs:

$$\mathbf{X}_{i,j} = \sum_{p=1}^{k} \mathbf{S}_p (\mathbf{U}_p)_i (\mathbf{V}_P^T), \tag{1}$$

where $\mathbf{X}_{i,j}$ are the missing data; i,j are the spatial and temporal indices of the missing data; $\mathbf{U}_p$ and $\mathbf{V}_p$ are the pth column of the spatial and temporal EOFs, respectively; $\mathbf{S}_p$ is the pth singular value; and k is the number of EOFs mode used for reconstruction.

(3) The first SVD decomposition was performed on original data set, and a new matrix was reconstructed by the existing data with original values and the missing data calculated by Equation (1) with $k = 1$, then the next decomposition performed on the new reconstructed matrix and the missing data were recalculated. This process was iterated until the predefined convergence criterion was reached, when the root mean square error (RMSE) at the cross-validation points was stabilized (the relative difference between previous and current iterations is smaller than the threshold value of 1.0×10^{-5}).

(4) The number of reconstructed EOFs (k) is increased with $k = 2, 3 \ldots k_{\max,}$ and procedure (3) is repeated. The optimal number of EOFs ($k_{\max}$) is retained when the minimum RMSE is obtained.

(5) Once the optimal number of EOFs is determined, the entire process is restarted including the 3% cross-validation data we set aside before. Finally, a cloud-free data set was constructed with values for the missing data are computed by truncated EOFs, and values for the existing data are kept for original values.

This is a general description of how the ordinary DINEOF method works. For a more detailed description, see Beckers and Rixen (2003) and Alvera-Azcarate et al. (2005).

2.3. Outlier detection

After the DINEOF procedure, the whole data set (include existing data points and missing data points) can be reconstructed using the truncated EOFs according to Equation (1). These truncated EOFs contain useful information on the variability, which can be used to detect outliers. The outliers can be detected within the DINEOF reconstruction as being those pixels for which the difference (the residuals) between the reconstructed and original values of the existing points is larger than statistically expected. To diagnose these statistical outliers based on the residuals (r_i) from all existing data points, we introduced the standardized residual (r_i/s^0), proposed by Rousseeuw and Leroy (1987). The preliminary scale estimator s^0 was calculated based on the objective function, multiplied by a finite sample correction factor dependent on m and p:

$$s^0 = 1.4826\left(1 + \frac{5}{m-p}\right)\sqrt{\text{median}(r_i^2)}, \tag{2}$$

where m is the number of existing data points and p is the number of estimators. With this scale estimator, the standardized residuals r_i/s^0 were computed and used to determine a weight w_i for the ith existing data point as follows:

$$w_i = \begin{cases} 1, & \text{if} \left|\frac{r_i}{s^0}\right| \le 2.5 \\ 0, & \text{otherwise} \end{cases}. \tag{3}$$

The procedure was repeated for the n points remaining after the outliers had been eliminated, but the second scale estimator (s^*) was defined as follows (Walczak and Massart 1995):

$$s^* = \sqrt{\left(\sum_{i=1}^{n} \frac{r_i^z}{n-p}\right)}. \tag{4}$$

Each of the n remaining points was evaluated again according to Equation (3) to perform the robust diagnosis of outliers.

2.4. Validation of reconstructed accuracy

In order to validate the accuracy of each reconstruction, several parameters were calculated, including the Pearson correlation coefficient (r), signal to noise ratio (SNR), RMSE, and mean absolute difference (MAD). After each reconstruction, the accuracy of the reconstruction was evaluated using these four statistical parameters calculated from the original values and the corresponding reconstructed values for the existing data points.

The SNR is defined as the ratio of standard deviation of the reconstructed values and the standard deviation of the errors (difference between original values and reconstructed values) for the existing data points.

The RMSE and MAD are defined as follows:

$$\mathrm{RMSE} = \sqrt{\frac{\sum [S-I]^2}{n}},\ \mathrm{MAD} = \frac{\sum |S-I|}{n}, \tag{5}$$

where S indicates the original chl-a value, I indicates the reconstructed chl-a value, and n is the number of samples.

2.5. Normality test

Skewness and kurtosis were used as a measurement of normality of chl-a data. The coefficients of skewness (sk) and kurtosis (ku) are defined as follows (Mardia 1970; Groeneveld and Meeden 1984):

$$\mathrm{sk} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^3}{(N-1)s^3},\ \mathrm{ku} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^4}{(N-1)s^4}, \tag{6}$$

where x_i is the ith observation, $\bar{x}$ is the mean, s is the standard deviation, and N is the number of observations.

Skewness is a measure of the asymmetry of a distribution with zero for a standard normal distribution; kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution, with the kurtosis for a standard normal distribution being three.

3. Results

3.1. chl-a data set description

After elimination of pixels and images with extremely high missing data coverage (>95%), we established a new chl-*a* data set with 7441 spatial pixels and 562 temporal images (hereafter called the 'original data set') – the average data coverage of this original data set is 50.2%. Figure 2 shows the mean percentage data coverage of the original chl-*a* data set over the Bohai and Yellow Seas. Data coverage is 30–90% over most of the domain, with the highest values in the northern area. Some of the coastal areas (depth < 20 m) display an extremely high missing data coverage, probably due to failure of atmospheric calibration and higher chl-*a* retrieved algorithm error in coastal waters (IOCCG 2000; O'Reilly et al. 2000).

To test the improvement of the DINEOF method achieved by subdivision, the whole chl-*a* data set was divided into three subsets corresponding to water depths of 0–20, 20–50, and >50 m, respectively. We selected this subdivision scheme because the chl-*a* variability of this study area exhibits significantly different spatial and temporal patterns with water depth (Shi and Wang 2012; Yamaguchi et al. 2012). Table 1 shows the characteristics of the whole chl-*a* data set and three chl-*a* subsets (the chl-*a* value is $\log_{10}$-transformed). It clearly shows that after $\log_{10}$-transformation, all four data sets are near to normally distributed; average chl-*a* values (average) and kurtosis are higher in shallow waters than in deep, whereas the standard deviation is lower in shallow waters than in deep. Even though we are not explicitly dividing the standard deviation by the average values, the standard deviation has already shown apparent high values in areas of low average chl-*a* values. These characteristics indicate that shallow waters present high

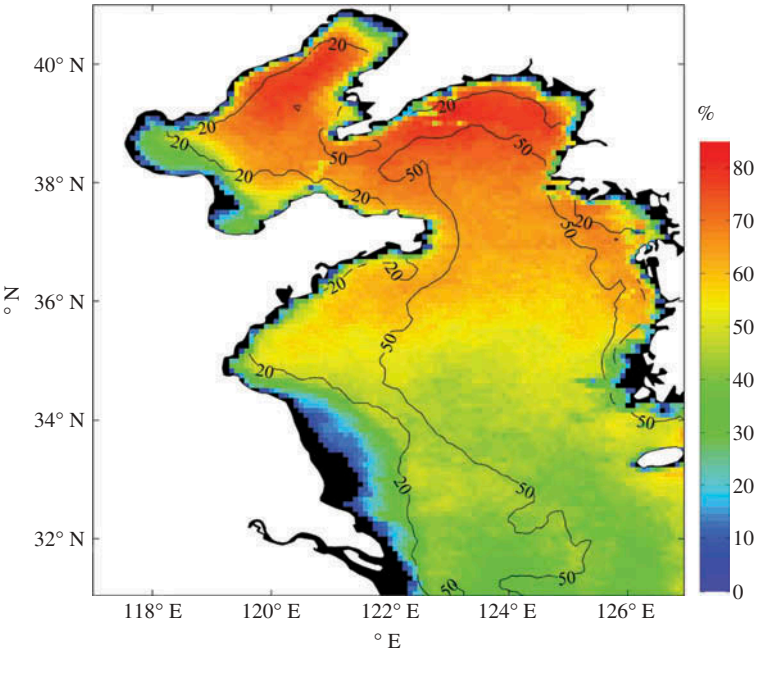

Figure 2. Mean percentage data coverage of the original chl-*a* data set. White denotes land and black denotes no data, and the solid line denotes bathymetry of 20 and 50 m, respectively.

Table 1. Characteristics of the whole data set and three subsets.

Data set	Average depth (m)	Matrix dimensions	Data coverage (%)	Average	Standard deviation	Skewness	Kurtosis
Whole area	46.7	7441 × 562	50.2	0.27	0.31	−0.2952	2.7388
<20 m	12.6	1447 × 562	37.0	0.59	0.14	0.5205	7.5913
20–50 m	28.6	2752 × 562	55.2	0.39	0.22	−0.2720	4.4942
>50 m	68.1	3242 × 562	51.8	0.05	0.27	0.1925	3.1932

average chl-*a* values and low chl-*a* variability, and deep waters present low average chl-*a* value and high chl-*a* variability. So our subdivision can also efficiently test the relationship between data-distributing features and the accuracy of DINEOF reconstruction. In the following section, the DINEOF reconstruction was performed on the whole area, 0–20 m area, 20–50 m area, and >50 m area individually, to test the improvement in DINEOF reconstructed accuracy from this depth subdivision scheme.

3.2. *DINEOF reconstruction*

The ordinary DINEOF reconstruction was directly performed on each data set (<20 m area, 20–50 m area, >50 m area, and whole area), respectively, predefined in Section 3.1. After individual DINEOF reconstruction, the reconstructed fields of the three subsets were merged into the entire data set (S-DINEOF) for comparison with the result of the single whole data set reconstruction (O-DINEOF). SNR, *r*, RMSE, and MAD from the reconstructed and original values for the existing points were employed as a measurement of reconstructed accuracy.

Figure 3 shows the cross-validation RMSE for the reconstruction of the four chl-*a* data sets and Table 2 summarizes the four validation parameters of the O-DINEOF and S-DINEOF methods; note that the reconstructed results of the three subsets are also shown in Table 2. The O-DINEOF method obtained 48 EOFs as the optimal number for DINEOF reconstruction (cross-validation RMSE = 0.1094), and the individual subset data reconstruction obtained 30 (cross-validation RMSE = 0.0821), 44 (cross-validation RMSE = 0.0997), and 52 (cross-validation RMSE = 0.1068) EOFs as the optimal number in the 0–20, 20–50, and >50 m data sets, respectively. From the four reconstructions, the total computation time of the three subsets reconstruction was relatively marginally higher (by about 12%) than the single whole data set reconstruction. This may be due to more total iterative times in the three subsets reconstruction (1359) than in the whole data set reconstruction (859). The accuracy of each reconstruction (Table 2) shows that SNR and *r* present a significant improvement in reconstructed accuracy with the subdivision scheme, and both individual and total errors (RMSE = 0.0721, MAD = 0.0520) for the three subsets reconstruction were all lower than the errors (RMSE = 0.0814, MAD = 0.0591) of the whole data set reconstruction. This result indicates that the depth subdivision scheme can significantly improve the accuracy of ordinary DINEOF reconstruction, but with less computational efficiency due to more total iterations in the S-DINEOF method.

The parameters SNR and *r* are more suitable parameters than RMSE and MAD for comparison of reconstructed accuracy with different data number and data amplitude. Table 2 shows that SNR and *r* were higher in the >50 m area than in other areas, associated with the data distribution property of each subset described in Section 3.1,

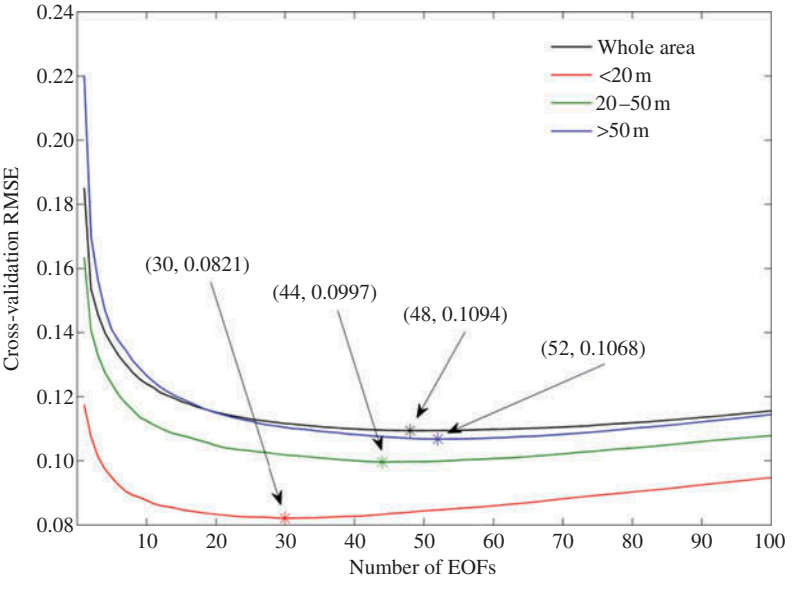

Figure 3. RMSE obtained by cross-validation for the reconstruction of the four chl-*a* data sets. The marker (*) indicates the convergence point and the text shows the optimal number of EOFs and corresponding cross-validation RMSE.

Table 2. Reconstructed results using O-DINEOF and S-DINEOF methods, including subset reconstruction.

	SNR	*r*	RMSE	MAD
O-DINEOF	3.7226	0.9658	0.0814	0.0591
S-DINEOF	4.2363	0.9733	0.0721	0.0520
<20 m	2.1636	0.9091	0.0602	0.0440
20–50 m	2.9396	0.9470	0.0720	0.0520
>50 m	3.4424	0.9604	0.0755	0.0546

which provides proof that the DINEOF method can obtain a better reconstructed accuracy for data sets with higher variability and gradient than for homogeneous data sets.

3.3. *Outlier detection in DINEOF reconstruction*

The ordinary DINEOF reconstruction was first applied to the original chl-*a* data set over the whole area. We then applied outlier detection and the removal programme introduced in Section 2.3, and made an evaluation of its improvement in the accuracy of DINEOF reconstruction.

Figure 4 shows original chl-*a* values and the truncated EOFs reconstructed values for 12–19 August 2000 over the existing data points. As will be seen in the detailed rectangle (Figure 4, bottom panels), there are some 'hot spot' pixels (denoted as outliers) whose values deviate markedly from their surrounding observations in the original image (Figure 4(*a*)), but the reconstructed pixels (Figure 4(*b*)) show reasonable values over the corresponding data points.

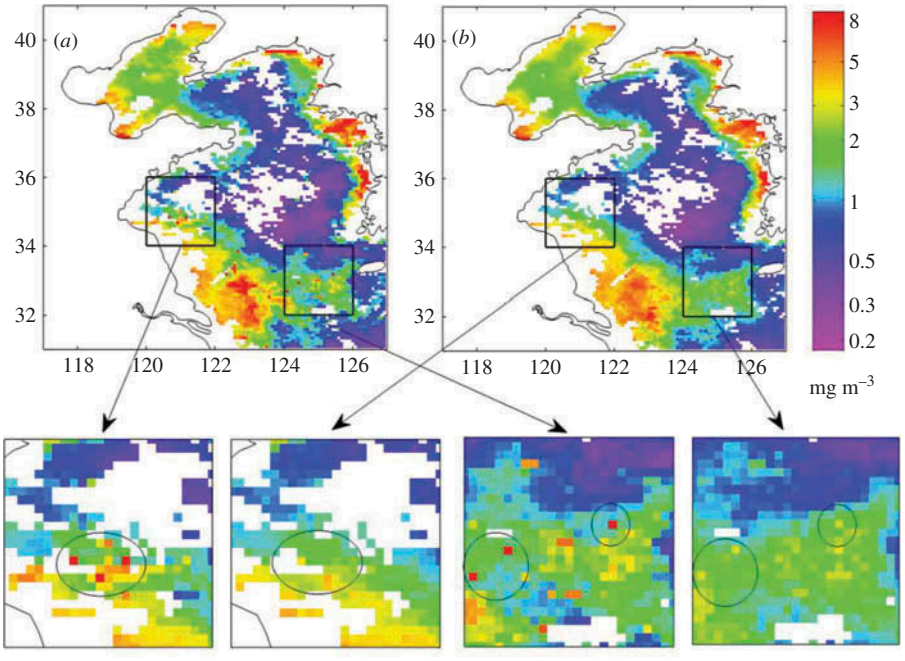

Figure 4. Snapshot comparison of (*a*) original chl-*a* values and (*b*) truncated EOFs reconstructed values over the existing data points. The four bottom panels show the detail in the respective rectangles; the circles represent more outlier details.

Figure 5 shows a comparison of DINEOF reconstructed chl-*a* values and original chl-*a* values over all existing data points (Figure 5(*a*)), and the reasonable data points (Figure 5(*b*)) and 'outlier' points (Figure 5(*c*)) in the original data set. The result of DINEOF reconstruction showed a well-reconstructed accuracy ($r = 0.97$, $p < 0.0001$) as shown in Table 2. There are a total of 2,099,466 data points (all existing data points) in the original data set and 107,201 points are detected as outliers (accounting for 5.1%). That is to say, 94.9% of the reconstructed values are the same as the original values.

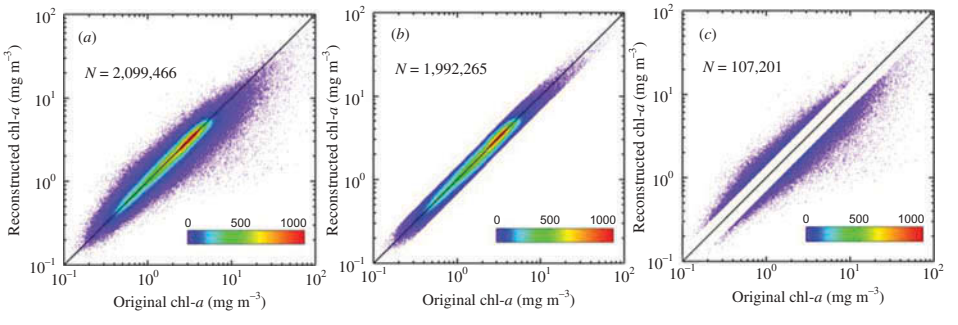

Figure 5. Scatter plot comparisons of DINEOF reconstructed chl-*a* values based on truncated EOFs and original chl-*a* values: (*a*) all existing data points, (*b*) 'reasonable' points, and (*c*) 'outlier' points. The black solid line denotes the 1:1 ratio, and the colour scale indicates the relative density function (unit: number of points per bin).

Table 3. Reconstructed results using the O-DINEOF, C-DINEOF, and D-DINEOF methods.

Method	EOFs	Cross-validation RMSE	Existing data validation (without outliers)			
			SNR	r	RMSE	MAD
O-DINEOF	48	0.1094	3.7226	0.9658	0.0629	0.0503
C-DINEOF	49	0.0870	3.7645	0.9665	0.0634	0.0503
D-DINEOF	56	0.0766	5.4027	0.9833	0.0552	0.0434

After outlier detection, the 'outlier' data points were eliminated and we performed another DINEOF reconstruction on the remaining reasonable data set (D-DINEOF). The reconstructed result was then compared with the reconstructed result in the absence of outlier detection (O-DINEOF). Considering that the outliers in the cross-validation data set might influence the DINEOF reconstruction process and confuse the comparison between the O-DINEOF and D-DINEOF methods, we designed a third DINEOF reconstruction procedure (C-DINEOF) similar to O-DINEOF, but used the cross-validation data set without outliers detected in the D-DINEOF method. The reconstructed results of the three methods are shown in Table 3; the three reconstructed validations were calculated from the same existing data points without outliers, and the C-DINEOF and D-DINEOF methods used the same cross-validation data set without outliers.

Comparison between the O-DINEOF method (cross-validation RMSE = 0.1094) and the C-DINEOF method (cross-validation RMSE = 0.0870) in Table 3 shows that cross-validation RMSE is very sensitive to outliers in the cross-validation data set, but the reconstructed process (number of EOFs = 48 vs 49) and accuracy (existing data validation) are similar. Comparison between O-DINEOF accuracy (SNR = 3.7226, r = 0.9658, RMSE = 0.0629, MAD = 0.0503) and D-DINEOF accuracy (SNR = 5.4027, r = 0.9833, RMSE = 0.0552, MAD = 0.0434) shows that outlier detection and elimination from the original data can effectively improve the accuracy of ordinary DINEOF reconstruction.

Figure 6 shows the whole DINEOF reconstructed procedure with outlier detection (D-DINEOF). The validation from a snapshot image of 24–31 October 1997 also shows that the DINEOF method with outlier detection exhibits higher accuracy (SNR = 5.0511, r = 0.9802, RMSE = 0.0541, MAD = 0.0420) than that of the ordinary DINEOF method (SNR = 3.0947, r = 0.9464, RMSE = 0.0610, MAD = 0.0480).

4. Discussion

Because of a lack of *in situ* measurements, the measurement of reconstructed accuracy is based on cross-validation and statistical description between original values and reconstructed values. Previous studies (Shi and Wang 2012; Yamaguchi et al. 2012) showed that most areas of the Bohai and Yellow seas are characterized by optically complex case II coastal waters and that satellite-derived chl-*a* has a high level of error using global chlorophyll-*a* algorithms (IOCCG 2000; Gregg and Casey 2004; Siswanto et al. 2011). The RMSEs of the reconstructed validation in this study are much smaller in magnitude than the chl-*a* algorithm 'error' (Tan et al. 2011), even smaller than the global algorithm 'noise' (RMSE = 0.22) (O'Reilly et al. 1998). Thus, the results obtained in this study indicate that the DINEOF method is a reasonable interpolation technique for chl-*a* reconstruction over this complex marine area.

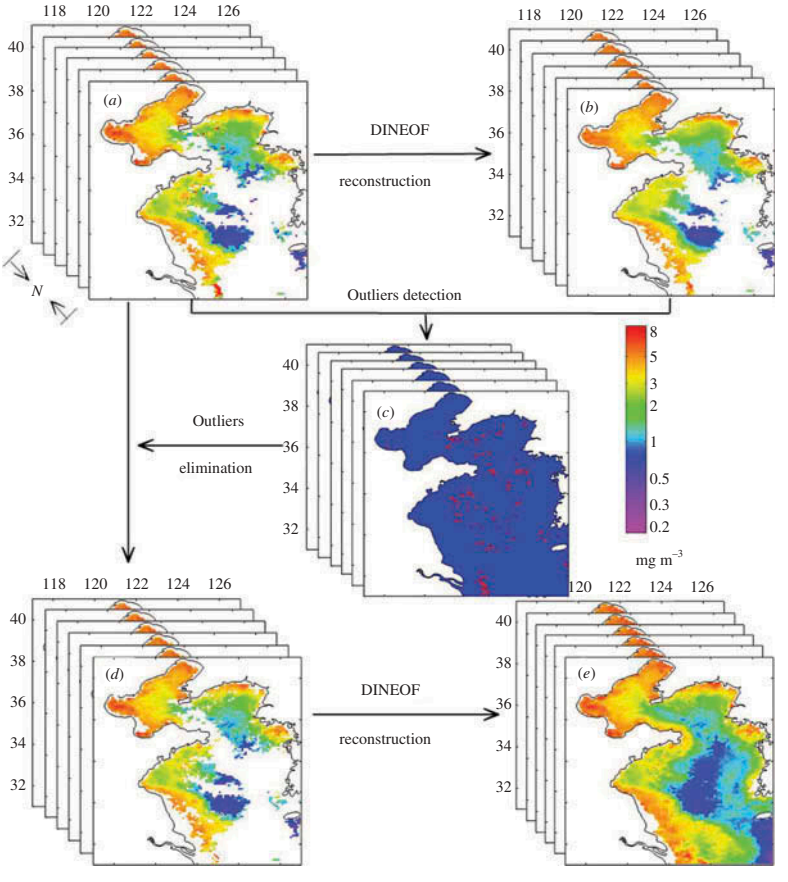

Figure 6. Procedure of DINEOF reconstruction with outlier detection. (*a*) Original images, (*b*) existing data reconstructed by truncated EOFs after the first DINEOF reconstruction, (*c*) outlier images with red points indicating outliers, (*d*) original images with outlier elimination, and (*e*) final reconstructed images after the second DINEOF reconstruction based on the original data set without outliers. The snapshot example image is from the period 24–31 October 1997.

In this study, improvement in the ordinary DINEOF method using depth subdivision based on 20 and 50 m isobaths was tested. Comparison between the entire data set reconstruction and those using depth-subdivided subsets shows greatly improved accuracy but at the cost of increased computation time. However it will be possible to parallelize the computational load to speed up the reconstruction process substantially in future studies. From Tables 1 and 2, the three subsets exhibit different numerical characteristics according to water depth, with higher chl-*a* values and lower variability in shallow than deep waters, as shown in previous research (Shi and Wang 2012; Yamaguchi et al. 2012). Different reconstruction accuracy of the three subsets indicates that the DINEOF method can obtain a better result over regions of high variability and gradients, relative to regions of more homogeneous values, similar to the findings of Sirjacobs et al. (2011) testing reconstruction accuracy in artificially clouded areas.

The difficulty in outlier detection in the satellite data set lies in the fact that there is no absolutely unique definition of the outlier. The definition might vary depending on the data set property, the research topic, and the basis of the detection algorithm (Hu, Carder,

and Muller-Karger 2000, 2001; Park, Chae, and Park 2013). In this study, we assume that all true chl-*a* properties in the data set can be summarized by the truncated EOFs obtained from DINEOF reconstruction and that these truncated EOFs were a robust fit to the real chl-*a* values, so any pixels with extreme original values that were inconsistent with the truncated EOF reconstructed values could be defined as 'outliers'. This definition is determined only on a statistical basis and contains no direct information on physical laws. For detection of these statistical outliers, we applied the standardized residuals method proposed by Rousseeuw and Leroy (1987). In general, this study did not consider spatially coherent structures, and only single pixels or small zones (several pixels) were diagnosed by this method. In addition, we also made a comparison with the method of Alvera-Azcárate et al. (2012) (data not included) using a threshold of three (82,516 outliers and only contain the EOF test). They present similar results, but our method is simpler and less computationally demanding. The reason might be that the data sets used here were 8 day composite images of low spatial resolution (9 km) and the study area contains fewer pixels than in other experiments using the DINEOF method (Nechad et al. 2011; Sirjacobs et al. 2011; Alvera-Azcárate et al. 2012), which can weaken the spatially coherent property of the chl-*a* data set. From visual verification of the snapshot image and validation of the reconstruction with and without the outlier detection procedure, the modified DINEOF method with outlier detection can significantly improve the accuracy of reconstruction.

There are also some restrictions in regard to the modified DINEOF method: as with the ordinary DINEOF method, the improved method cannot reconstruct coverage images of high missing data (>95%), resulting in discontinuity in time scale; DINEOF reconstruction (especially with outlier detection) has a smoothing effect and some extreme events accounting for little variance may not be effectively present in the reconstructed data set; and DINEOF reconstruction with the subdivision scheme may result in an unreasonable jumping effect at the borders of subregions. These unresolved issues will be addressed in our future research.

5. Conclusions

The present work illustrates the successful application of the self-consistent DINEOF method to reconstruct 13 years' (the lifetime of the SeaWiFS sensor) satellite-derived chl-*a* data sets in the Bohai and Yellow Seas. This is the first time the DINEOF method has been applied to such a long-term chl-*a* data set. The depth subdivision scheme used in DINEOF reconstruction made a significant improvement to the accuracy of the reconstruction, but at a cost of lower computational efficiency. A new outlier detection method based on standardized residuals theory can efficiently detect and eliminate spurious values, and significantly improve the accuracy of DINEOF reconstruction. The examples above show that the DINEOF method can be successfully applied to variables with very different characteristics, but specific modification according to the property of the data set can improve the reconstructed accuracy.

Acknowledgement

The authors would like to thank the Ocean Biology Processing Group of NASA for providing the SeaWiFS chlorophyll-*a* concentration data set (http://oceancolor.gsfc.nasa.gov/).

Funding

This study was funded by the Natural Science Foundation of China [grant number 41376121] and the Strategic Priority Research Programme of the Chinese Academy of Sciences [grant number XDA05130703], [grant number XDA11020405].

References

Alvera-Azcarate, A., A. Barth, J. M. Beckers, and R. H. Weisberg. 2007. "Multivariate Reconstruction of Missing Data in Sea Surface Temperature, Chlorophyll, and Wind Satellite Fields." *Journal of Geophysical Research* 112 (C3): C03008.

Alvera-Azcárate, A., A. Barth, M. Rixen, and J. M. Beckers. 2005. "Reconstruction of Incomplete Oceanographic Data Sets Using Empirical Orthogonal Functions: Application to the Adriatic Sea Surface Temperature." *Ocean Modelling* 9: 325–346.

Alvera-Azcárate, A., D. Sirjacobs, A. Barth, and J. M. Beckers. 2012. "Outlier Detection in Satellite Data Using Spatial Coherence." *Remote Sensing of Environment* 119: 84–91.

Beckers, J. M., and M. Rixen. 2003. "EOF Calculations and Data Filling from Incomplete Oceanographic Datasets." *Journal of Atmospheric and Oceanic Technology* 20: 1839–1856.

Borzelli, G., G. Manzella, S. Marullo, and R. Santoleri. 1999. "Observations of Coastal Filaments in the Adriatic Sea." *Journal of Marine Systems* 20: 187–203.

Campbell, J. W. 1995. "The Lognormal Distribution as a Model for Bio-Optical Variability in the Sea." *Journal of Geophysical Research* 100 (C7): 13237–13254.

Everson, R., P. Cornillon, L. Sirovich, and A. Webber. 1997. "An Empirical Eigenfunction Analysis of Sea Surface Temperatures in the Western North Atlantic." *Journal of Physical Oceanography* 27: 468–479.

Fieguth, P., D. Menemenlis, T. Ho, A. Willsky, and C. Wunsch. 1998. "Mapping Mediterranean Altimeter Data with a Multiresolution Optimal Interpolation Algorithm." *Journal of Atmospheric and Oceanic Technology* 15: 535–546.

Gregg, W. W., and N. W. Casey. 2004. "Global and Regional Evaluation of the SeaWiFS Chlorophyll Data Set." *Remote Sensing of Environment* 93: 463–479.

Groeneveld, R. A., and G. Meeden. 1984. "Measuring Skewness and Kurtosis." *The Statistician* 33: 391–399.

He, R., R. H. Weisberg, H. Zhang, F. E. Muller-Karger, and R. W. Helber. 2003. "A Cloud-Free, Satellite-Derived, Sea Surface Temperature Analysis for the West Florida Shelf." *Geophysical Research Letters* 30 (15): 1811.

Hoer, J. L., and J. She. 2007. "Optimal Interpolation of Sea Surface Temperature for the North Sea and Baltic Sea." *Journal of Marine Systems* 65: 176–189.

Hu, C., K. L. Carder, and F. E. Muller-Karger. 2000. "Atmospheric Correction of SeaWiFS Imagery over Turbid Coastal Waters: A Practical Method." *Remote Sensing of Environment* 74: 195–206.

Hu, C., K. L. Carder, and F. E. Muller-Karger. 2001. "How Precise Are SeaWiFS Ocean Color Estimates? Implications of Digitization-Noise Errors." *Remote Sensing of Environment* 76: 239–249.

IOCCG. 2000. "Remote Sensing of Ocean Colour in Coastal and Other Optically Complex Waters, Reports of the International Ocean-Colour Coordination Group, no. 3." In *Reports of the International Ocean Color Coordinating Group*, edited by S. Sathyendranath, 140. Dartmouth: IOCCG.

Mardia, K. V. 1970. "Measures of Multivariate Skewness and Kurtosis with Applications." *Biometrika* 57: 519–530.

Nechad, B., A. Alvera-Azcaràte, K. Ruddick, and N. Greenwood. 2011. "Reconstruction of MODIS Total Suspended Matter Time Series Maps by DINEOF and Validation with Autonomous Platform Data." *Ocean Dynamics* 61: 1205–1214.

O'Reilly, J. E., S. Maritorena, B. G. Mitchell, D. A. Siegel, K. L. Carder, S. A. Garver, M. Kahru, and C. McClain. 1998. "Ocean Color Chlorophyll Algorithms for SeaWiFS." *Journal of Geophysical Research* 103 (C11): 24937–24953.

O'Reilly, J. E., S. Maritorena, D. A. Siegel, M. C. O'Brien, D. Toole, B. G. Miithell, M. Kahru, F. P. Chavez, P. Strutton, G. Cota, S. B. Hooker, C. R. Mcclain, K. L. Carder, F. Muller-Karger, L. Harding, A. Magnuson, D. Phinney, G. F. Moore, J. Aiken, K. R. Arrigo, R. Letelier, and M. Culver. 2000. "Ocean Color Chlorophyll a Algorithms for SeaWiFS, OC2, and OC4: Version

4." In *SeaWiFS Postlaunch Calibration and Validation Analyses, Part 3*, edited by S. B. Hooker, and E. R. Firestone, NASA Technical Memorandum 2000–206892 11, 9–24. Greenbelt, MD: NASA Goddard Space Flight Center.

Park, K., H. Chae, and J. Park. 2013. "Characteristics of Satellite Chlorophyll-*a* Concentration Speckles and a Removal Method in a Composite Process in the East/Japan Sea." *International Journal of Remote Sensing* 34: 4610–4635.

Pukhtyar, L., S. Stanichny, and I. Timchenko. 2009. "Optimal Interpolation of the Data of Remote Sensing of the Sea Surface." *Physical Oceanography* 19: 225–239.

Rousseeuw, P. J., and A. M. Leroy. 1987. *Robust Regression and Outlier Detection.* New York: Wiley.

Ruddick, K. G., F. Ovidio, and M. Rijkeboer. 2000. "Atmospheric Correction of SeaWiFS Imagery for Turbid Coastal and Inland Waters." *Applied Optics* 39: 897–912.

Saulquin, B., F. Gohin, and R. Garrello. 2011. "Regional Objective Analysis for Merging High-Resolution MERIS, MODIS/Aqua, and SeaWiFS Chlorophyll- *a* Data from 1998 to 2008 on the European Atlantic Shelf." *IEEE Transactions on Geoscience and Remote Sensing* 49: 143–154.

Shi, W., and M. Wang. 2012. "Satellite Views of the Bohai Sea, Yellow Sea, and East China Sea." *Progress in Oceanography* 104: 30–45.

Sirjacobs, D., A. Alvera-Azcárate, A. Barth, G. Lacroix, Y. Park, B. Nechad, K. Ruddick, and J. Beckers. 2011. "Cloud Filling of Ocean Colour and Sea Surface Temperature Remote Sensing Products over the Southern North Sea by the Data Interpolating Empirical Orthogonal Functions Methodology." *Journal of Sea Research* 65: 114–130.

Siswanto, E., J. Tang, H. Yamaguchi, Y. Ahn, J. Ishizaka, S. Yoo, S. Kim, Y. Kiyomoto, K. Yamada, C. Chiang, and H. Kawamura. 2011. "Empirical Ocean-Color Algorithms to Retrieve Chlorophyll-*a*, Total Suspended Matter, and Colored Dissolved Organic Matter Absorption Coefficient in the Yellow and East China Seas." *Journal of Oceanography* 67: 627–650.

Sorjamaa, A., A. Lendasse, Y. Cornet, and E. Deleersnijder. 2010. "An Improved Methodology for Filling Missing Values in Spatiotemporal Climate Data Set." *Computational Geosciences* 14: 55–64.

Sun, L., M. Guo, and X. Wang. 2010. "Ocean Color Products Retrieval and Validation Around China Coast with MODIS." *Acta Oceanologica Sinica* 29: 21–27.

Swardika, I. K., T. Tanaka, and H. Ishida. 2012. "Study on the Characteristics of the Indonesian Seas Using Satellite Remote-Sensing Data for 1998–2007." *International Journal of Remote Sensing* 33: 2378–2394.

Tan, S., G. Shi, J. Shi, H. Gao, and X. Yao. 2011. "Correlation of Asian Dust with Chlorophyll and Primary Productivity in the Coastal Seas of China During the Period from 1998 to 2008." *Journal of Geophysical Research* 116 (G2): G02029.

Vantrepotte, V., and F. Mélin. 2011. "Inter-Annual Variations in the SeaWiFS Global Chlorophyll a Concentration (1997–2007)." *Deep Sea Research Part I: Oceanographic Research Papers* 58: 429–441.

Walczak, B., and D. L. Massart. 1995. "Robust Principal Components Regression as a Detection Tool for Outliers." *Chemometrics and Intelligent Laboratory Systems* 27: 41–54.

Yamaguchi, H., H. Kim, Y. B. Son, S. W. Kim, K. Okamura, Y. Kiyomoto, and J. Ishizaka. 2012. "Seasonal and Summer Interannual Variations of SeaWiFS Chlorophyll *a* in the Yellow Sea and East China Sea." *Progress in Oceanography* 105: 22–29.