On the sensitivity of field reconstruction and prediction using Empirical Orthogonal Functions derived from gappy data

Marc H. Taylor[1,2] *, Martin Losch[1], Manfred Wenzel[1], Jens Schröter[1]

[1] *Alfred Wegener Institute for Polar and Marine Research, Bremerhaven, Germany*

[2] *present address: Leibniz Center for Tropical Marine Ecology, Bremen, Germany*

* *Corresponding author address:* Marc H. Taylor, Leibniz Center for Tropical Marine Ecology, Fahrenheitstrasse 6, D-28359 Bremen, Germany

E-mail: marchtaylor@yahoo.com

ABSTRACT

Empirical Orthogonal Function (EOF) Analysis is commonly used in the climate sciences and elsewhere to describe, reconstruct, and predict highly dimensional data fields. When data contain a high percentage of missing values (i.e. "gappy"), alternate approaches must be used in order to correctly derive EOFs. The aims of this paper are to assess the accuracy of several EOF approaches in the reconstruction and prediction of gappy data fields, using the Galapagos Archipelago as a case study example. EOF approaches included least-squares estimation via a covariance matrix decomposition (LSEOF), "Data Interpolating Empirical Orthogonal Functions" (DINEOF), and a novel approach called "Recursively-Subtracted Empirical Orthogonal Functions" (RSEOF). Model-derived data of historical surface Chlorophyll *a* concentrations and sea surface temperature, combined with a mask of gaps from historical remote sensing estimates, allowed for the creation of "true" and "observed" fields by which to gauge the performance of EOF approaches. Only DINEOF and RSEOF were found to be appropriate for gappy data reconstruction and prediction. DINEOF proved to be the superior approach in terms of accuracy, especially for noisy data with a high estimation error, although RSEOF may be preferred for larger data fields due to its relatively faster computation time.

1. Introduction

Empirical Orthogonal Function (EOF) Analysis, or *Principal Component Analysis* (PCA) in other disciplines, is commonly used in climate research as a tool to analyze meteorological fields with high spatio-temporal dimensionality. The leading EOF modes will typically describe large scale dynamical features in the field, and reconstruction of the field using a truncated subset of EOFs can filter out small scale features or noise. Furthermore, EOF truncation may be useful for further statistical analysis by reducing the dimensionality of the data. For example, EOF coefficients have been used in Canonical Correlation Analysis (CCA) for the identification of patterns in coupled fields (Barnett and Preisendorfer 1987). Other techniques like principal oscillation analysis (POP) or principal interaction patterns (PIP) aim at the approximation of complex dynamical systems by a simple dynamical model. Usually EOF techniques are applied in this reduction (Hasselmann 1988). The approach by Kaplan et al. (2000), in their work "Reduced Space Optimal Interpolation of Historical Marine Sea Level Pressure", has goals similar to our presentation. We will augment their work by comparing a suite of numerical techniques designed for this task.

a. Basic EOF Approaches

EOF analysis is typically conducted via two main approaches; either by direct Singular Value Decomposition (SVD) of the observed data matrix or by an Eigenvalue decomposition of a covariance matrix. When fields are complete (i.e. no gaps with missing values), EOFs can be calculated in either way to achieve the same outcome.

For all presented approaches, we will consider a data matrix $\mathbf{X} = x_{ij}$, where i is the time index (length M) and j is the space index (length N). Each sample time series (columns) is centered (mean-subtracted) so that the EOFs describe patterns of temporal covariance.

1) Direct Data Matrix Decomposition

The direct approach via SVD is as follows:

$$\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathrm{T}}, \qquad x_{ij} = \sum_{k=l,N} u_{ik}\,\sigma_k\,v_{kj} \tag{1}$$

where $\mathbf{X}$ is an $M \times N$ data matrix, $\mathbf{V}$ is an $N \times N$ matrix containing the EOF patterns, $\mathbf{U}$ is an $M \times N$ matrix of the EOF coefficients, $\boldsymbol{\Sigma}$ is an $N \times N$ matrix containing the singular values on the diagonal, and k is the EOF mode index (length N). Only EOFs $\leq \min(M, N)$ will carry information. The explained variance of each mode is calculated as the square of each ${\sigma_k}^2$, which is typically presented as a percent:

$$\%\text{ explained variance} = \frac{{\sigma_k}^2 * 100}{\sum_{k=1}^{N} {\sigma_k}^2}. \tag{2}$$

2) Covariance Matrix Decomposition

The covariance matrix decomposition approach requires a square matrix. One first constructs a covariance matrix $\mathbf{C}$,

$$\mathbf{C} = \frac{1}{M}\mathbf{X}^{\mathrm{T}}\mathbf{X}, \qquad c_{jj'} = \frac{\sum_{i=1}^{M} x_{ji}\,x_{ij'}}{M} \tag{3}$$

where $\mathbf{C}$ is an $N \times N$ matrix containing the covariance values between columns $\mathbf{x}_j$ of $\mathbf{X}$. This is subsequently decomposed via Eigenvalue decomposition,

$$\mathbf{C} = \mathbf{E}\boldsymbol{\Lambda}\mathbf{E}^{\mathrm{T}}, \qquad c_{jj'} = \sum_{k=1}^{N} e_{jk}\,\lambda\,e_{kj'} \tag{4}$$

where $\mathbf{E}$ is an $N \times N$ matrix of the EOF patterns, and $\boldsymbol{\Lambda}$ is an $N \times N$ matrix containing the eigenvalues on the diagonal. Again, only EOFs $\leq \min(M, N)$ will carry information. $\mathbf{X}$ is then projected onto $\mathbf{E}$ to derive the EOF coefficients (sometimes referred to as the "principal components"),

$$\mathbf{A} = \mathbf{X}\mathbf{E}, \qquad a_{ik} = \sum_{j}^{N} x_{ij}\,e_{jk} \tag{5}$$

where $\mathbf{A}$ is an $M \times N$ matrix of the EOF coefficients. Due to the projection, $\mathbf{A}$ carries the magnitude of $\mathbf{\Lambda}$. In order to create a normalized version of the EOF coefficients, $\mathbf{A}^+$, each EOF coefficient a_k must be divided by the square-root of their corresponding $\mathbf{\Lambda}$ values λ_k ,

$$\mathbf{A}^+ = \mathbf{A}\mathbf{\Lambda}^{-\frac{1}{2}}, \qquad {a_{ik}}^+ = \frac{1}{\sqrt{\lambda_k}} a_{ik}. \tag{6}$$

Explained variance of each EOF mode k is calculated as follows:

$$\%\text{ explained variance} = \frac{\lambda_k * 100}{\sum_{k=1}^{N} \lambda_k}. \tag{7}$$

Following normalization, the two basic approaches are related as follows: $\mathbf{V} = \mathbf{E}$, $\mathbf{A}^+ = \mathbf{U}$ and $\mathbf{\Sigma}^2 = \mathbf{\Lambda}$.

b. Gappy Data EOF Approaches

Gappiness in data fields can be due to instrument limitations (coverage), or errors in measurement. When gappiness is extreme, interpolation becomes impractical and EOF reconstruction can provide a more accurate alternative.

1) Covariance Matrix Decomposition / Least-squares estimation of coefficients - LSEOF

Due to the inability to decompose a matrix containing missing values, a direct data matrix decomposition via SVD is not possible. The approach via covariance matrix decomposition is possible; however, due to the missing values, one must adopt a least-squares approach that takes into account the number of paired observations between samples. In this work, we will refer to this approach as “Least-Squares Empirical Orthogonal Functions” (LSEOF). In LSEOF, the above covariance matrix calculation (Eq. 3) must be scaled by the number of shared, non-missing values between samples (von Storch and Zwiers 1999; Kaplan et al. 1997; Boyd et al. 1994),

$$c_{jj'} = \frac{\sum_{i \in I_{jj'}} x_{ji}\, x_{ij'}}{\dim(I_{jj'})} \tag{8}$$

where $I_{jj'}$ is the set of valid pairs $(x_{ji}, x_{ij'})$ ($i = M$ when there are no gaps).

Following the decomposition of $\mathbf{C}$ to obtain the EOFs $\mathbf{E}$ (Eq. 4), the EOF coefficients $\mathbf{A}$ can be estimated via a least-squares approximation,

$$\mathbf{X} = \mathbf{A}\mathbf{E}^{\mathrm{T}} + \epsilon, \qquad \phi = \epsilon^{\mathrm{T}}\epsilon = (\mathbf{X} - \mathbf{A}\mathbf{E}^{\mathrm{T}})^{\mathrm{T}}(\mathbf{X} - \mathbf{A}\mathbf{E}^{\mathrm{T}}) \tag{9}$$

where ϵ is the error and ϕ is the objective function with the solution

$$a_{ik} = \frac{\sum_{j \in J_i} x_{ij}\, e_{jk}}{\sum_{j \in J_i} |{e_{jk}}^2|} \tag{10}$$

where J_i is the set of non-missing values at time i. Note that the denominator reduces to 1 when there are no missing values; thus, equaling the scalar product for $\mathbf{A}$ shown above (Eq. 5).

Several issues have been identified with the use of this approach. First and foremost is the problem that the calculation of a covariance matrix derived from gappy data is not necessarily positive definite, and decomposition via LSEOF can contain negative λ values. Since the variance of the data set is contained in the trace of the covariance matrix $\mathbf{C}$ and, subsequently, equal to the sum of $\mathbf{\Lambda}$, having negative values will mean that other EOFs e_k will have higher λ_k than in reality; thus, overestimating their amplitude and the amount of explained variance contained therein (Beckers and Rixen 2003; Björnsson and Venegas 1997).

λ amplification also has consequences for the assessment of EOF "significance" – i.e. differentiation between EOFs that describe large-scale patterns from those associated with small-scale features and noise. This is likely to equally affect both subjective methods, such as truncation based on visual inspection, e.g. Scree plots, and objective methods, e.g. *North's Rule of Thumb* (North et al. 1982).

A second problem is that the decomposition of a non-positive definite covariance matrix is a loss of orthogonality between EOFs (Björnsson and Venegas 1997), which makes their use in predictive models less attractive. For example, Barnett and Preisendorfer (1987) describe a method of Canonical Correlation Analysis (CCA) based on EOF coefficients, which is useful in determining the correlation between coupled fields. When correlations are high, issues associated with multi-collinearity can affect the predictive ability of the model.

2) Data Interpolating Empirical Orthogonal Functions - DINEOF

An alternate approach, DINEOF (Beckers and Rixen 2003; Alvera-Azcárate et al. 2005), interpolates missing values via an iterative SVD algorithm. DINEOF has similarities with approaches aimed at iterative estimation of the covariance matrix (e.g. Bien and Tibshirani 2011), although DINEOF directly iterates values in the data matrix itself.

Missing values are initially filled by an unbiased guess (zero in the typical case of mean-subtracted data). In addition, some non-missing values (the authors recommend a small percentage of the data points or at least 30 points) are also treated as gaps (e.g. zero-substituted) while their original values are retained separately for assessing the root mean square error (RMS) of the interpolated values.

The DINEOF algorithm subsequently decomposes the data matrix via SVD and a reconstruction is calculated using a single, leading EOF mode. The interpolated values for the missing locations are then substituted in the original matrix. Subsequent SVD iterations, and their resulting EOF reconstructions, will continually modify the values in the gaps until convergence of the RMS. Following convergence, a second EOF is then added to the reconstruction and again interpolated until convergence using two EOFs. This procedure continues with an increasing number of EOFs until the RMS converges (see Beckers and Rixen (2003) and Alvera-Azcárate et al. (2005) for further description of the algorithm). The resulting interpolated matrix will no longer contain gaps, thus overcoming the drawbacks of the previous approach.

3) Recursively-Subtracted Empirical Orthogonal Functions - RSEOF

A third approach, RSEOF, is proposed in this work. It is an adaptation of LSEOF (Sect. 1.b.1) in that it uses the same basic methodology of decomposition of a covariance matrix with least squares expansion of EOF coefficients (Eqs. 8, 10); however, the procedure is done in a recursive fashion by solving for one EOF at a time. In each iteration, the leading EOF mode is used to reconstruct a truncated approximation of the data field, which is subsequently subtracted from the remaining data in the field. In principle, the procedure should better preserve orthogonality among EOFs and prevent λ amplification.

The approach is as follows:

i. The observed data matrix $\mathbf{X}^{\mathrm{O}}$ is (optionally) centered and/or scaled prior to the decomposition, and is renamed as $\mathbf{X}^i$ for the first iteration, $i = 1$.

ii. A covariance matrix $\mathbf{C}^i$ is calculated from $\mathbf{X}^i$ (Eq. 8).

iii. $\mathbf{C}^i$ is subjected to Eigenvalue decomposition giving $\mathbf{E}^i$ and $\mathbf{\Lambda}^i$ (Eq. 4).

iv. $\mathbf{A}^i$ is computed using the least-squares approach (Eq. 10)

v. A truncated version of the data is reconstructed using the leading EOF mode, e_1^i and a_1^i, resulting in $\mathbf{X}^{recon,i}$.

vi. This field is then subtracted from the data to give a new field for iteration $i + 1$; $\mathbf{X}^{i+1} = \mathbf{X}^i - \mathbf{X}^{recon,i}$

vii. Steps ii-vi are then iterated until a given criterion (e.g. for $i \rightarrow N$; remaining % variance level, as calculated by $\sum \mathrm{tr}(\mathbf{C}^i)$; minimization of reconstruction error, e.g. MAE, RMS).

c. Data Reconstruction

Reconstruction of the data field can simply be calculated as the scalar product of the EOFs and their coefficients. For the approaches involving and Eigenvalue decomposition of a covariance matrix (e.g. LSEOF, RSEOF), this operation is as follows,

$$\mathbf{X} = \mathbf{A}\mathbf{E}^{\mathrm{T}}, \qquad x_{ij} = \sum_{k=1}^{N} a_{ik}\, e_{kj} \tag{11}$$

where x_{ij} is the reconstructed data field. Under cases of non-gappy data, when the full set of EOFs N is used, the reconstruction is said to be complete and exact. If $k < N$ (e.g. truncated to include only the leading EOFs with largest λ values), then the reconstruction is approximate (Wilks 2006). Reconstruction from EOFs derived via SVD (Eq. 1) or DINEOF require that $\mathbf{\Sigma}$ is included in the scalar product, since neither the EOFs $\mathbf{V}$ nor the EOF coefficients $\mathbf{U}$ carry the units of the field in the way that $\mathbf{A}$ does.

d. Summary of Gappy Approaches and Aims of the Present Work

We have outlined three main approaches for calculating EOFs with gappy data; including: 1. Decomposition of a covariance matrix followed by a least-squares estimate of EOF coefficients (LSEOF); 2. Filling of gaps via iterative SVD interpolation (DINEOF); 3. Recursive subtraction of EOFs from the data field (RSEOF). The first approach is known to have drawbacks associated with λ amplification, while the latter two approaches attempt to remedy this issue by either attempting to better preserve orthogonality of trailing EOFs (RSEOF) or by eliminating the problems associated with the decomposition of a non-positive definite matrix via an optimal interpolation algorithm (DINEOF).

In order to illustrate these issues in a simple example, we can observe the performance of each approach in reconstructing a gappy field containing a single temporal sine-wave signal:

$$x_{ij} = \sin(t_i)\, s_j \tag{12}$$

where $t_i = i2\pi/M$, $s_j = j$, $M = 200$ and $N = 100$. Differing levels of gappiness (20, 40, 60 and 80%) are randomly distributed throughout the field. The leading λ values are nearly identical for all approaches although trailing λ's are amplified substantially in the LSEOF approach. This amplification increases with the degree of gappiness in the observed field (Fig. 1, top panels). Statistics relating to field reconstruction can be seen in the middle and bottom panels of Fig. 1. The effect of λ amplification in the LSEOF approach is evident in the variance of the reconstructed field relative to true non-gappy field. Reconstructions using EOFs derived from RSEOF and DINEOF do not exceed a relative variance of 100%. Another statistic describing the fit of the reconstruction is that of the mean absolute error (MAE), which is calculated as follows

$$MAE = \frac{1}{n}\sum_{k=1}^{n} |\mathrm{pred}_k - \mathrm{obs}_k| \tag{13}$$

where $(\mathrm{pred}_k, \mathrm{obs}_k)$ is the k^{th} of n pairs of predictions and observations. The MAE is the arithmetic average of the absolute error (Wilks 2006) and is of practical use for inter-comparisons given that it presents the magnitude of average model-performance error in the same units as the field (Willmott and Matsuura 2005). Again, LSEOF amplifies the error of the reconstruction using trailing EOFs while RSEOF and DINEOF continue to decrease MAE before it flattens out. In this example, DINEOF outperforms RSEOF in terms of MAE under all degrees of gappiness.

The aims of the present work are to further evaluate the performance of these EOF approaches in the reconstruction and prediction of gappy data fields. Towards this aim, we consider a more realistic example using modeled surface Chlorophyll a (Chl$a$) concentrations that have been masked by historical cloud cover.

2. Experiments

a. Case Study Description

In order to examine the performance of the EOF approaches on a more realistic data field, we use the example of remotely-sensed surface Chl$a$ concentration. Estimates of Chl$a$ have become a valuable source of information regarding the biological productivity and variability of aquatic systems ever since the regular availability of data, coinciding with start of the operation of SeaWiFS (Sea-viewing Wide Field-of-view Sensor) in 1997. Since then, additional satellite sensors (e.g. MODIS, MERIS) have been implemented to complement and improve upon its estimation from ocean color. Despite improvements in coverage, and the availability of merged products (e.g. Globcolour Project - `http://www.globcolour.info`), cloud coverage continues to make the use of daily resolution data impractical for many analyses due to the high degree of missing values.

We have chosen to use the example of the Galapagos Archipelago as an interesting test case due to the known variability in the ecosystem at both seasonal and inter-annual scales via the El Niño Southern Oscillation (ENSO). The Galapagos lie in the heart of the Equatorial Upwelling (EU) region of the eastern tropical Pacific. Nutrients are supplied to the photic zone by equatorial upwelling and mixing, and by topographic upwelling of the Equatorial Undercurrent (EUC) on the western side of the archipelago (Chavez and Brusca 1991). In particular, cold, nutrient-rich waters of the EUC are brought to the surface following contact with the western side of the archipelago. As a result, the Galapagos are able to support at least twice the phytoplankton biomass and primary production as the remainder of the EU or any of the open-ocean regions of the eastern tropical Pacific (Pennington et al. 2006).

Under ENSO-neutral or negative (La Niña) conditions, tradewinds drive surface waters to the western tropical Pacific and create a basin-wide slope, where sea surface is about 1/2 meter higher at Indonesia than at Ecuador, effectively pushing down surface waters in the west. In the eastern tropical Pacific, the thermocline is closer to the surface, which facilitates the availability of nutrients to primary producers via upwelling. By contrast, ENSO-positive (El Niño) conditions are a result of weakened tradewinds, causing surface waters to relax back to the east, which lowers the thermocline and the EUC. As a result, the availability of cool nutrient-rich waters to upwelling is decreased and primary production is dramatically reduced.

Remote sensing Chl$a$ data (Globcolour GSM merged product, 4.63 km resolution) of the region reveals that missing values show a distinct spatio-temporal pattern as related to cloud coverage. Highest gappiness is observed in the warmer oceanic waters north of the archipelago and during the austral winter months, while lowest gappiness is associated with the colder upwelling centers west of the archipelago (Fig. 2).

b. Synthetic Data Set

In order to obtain full, non-gappy data fields, we use model-derived data. The model consisted of a biogeochemical model, REcoM (Regulated Ecosystem Model) (Schartau et al. 2007), coupled to a global general circulation model, MITgcm (Massachusetts Institute of Technology General Circulation Model) (Marshall et al. 1997; MITgcm Group 2012). The model had a mean horizontal resolution of 18 km and a vertical resolution of 10 m near the surface. The simulation spanned the years 1992 through 2007 (for additional details, see Taylor et al. 2013).

Daily 4.63 km resolution Globcolour chlorophyll data were used to create a cloud mask for the modeled data fields. When no valid data values were recorded within each larger grid of the model, the matrix location was classified as a missing value. In this way, we were able to obtain both the "true" non-gappy field and an "observed" gappy data field masked primarily by clouds. We examined the region between $93\,°W$ – $88\,°W$ and $1\,°N$ – $2\,°S$ for the period coinciding with remote-sensing estimates (1 September 1997 – 31 December 2007). Additionally, modelled sea surface surface temperature (SST) fields were used for the construction of a predictive CCA model. Both Chl$a$ and SST data were transformed to anomalies by subtracting the long-term monthly means from the time series of each grid. The resulting dimensions of the data matrices were 3774 × 608 (day × grid).

c. Analyses of Performance

EOF was used to decompose true (i.e. non-gappy) and observed (i.e. gappy) Chl$a$ and true SST fields. All three gappy approaches (LSEOF, RSEOF, and DINEOF) were used on the observed Chl$a$ field. For the DINEOF approach, we interpolated the missing values according to the methodology described earlier in Sect. 1.b.2. 10000 observed Chl$a$ values (approximately 1% of the known values) were used as the independent measure of RMS fit. The threshhold for convergence was set at $\delta\mathrm{RMS} \leq 1e^{-5}\,[\mathrm{mg\,Chl}a\,\mathrm{m}^{-3}]$. Following convergence, these values were restored to their original values in the interpolated matrix and a final EOF decomposition was performed on the interpolated data field.

1) EOF Reconstruction

The Chl$a$ fields were reconstructed using variable degrees of EOF truncation ($k = 1 \rightarrow 20$). Error of the reconstructed field was measured against the true Chl$a$ field via MAE.

2) EOF/CCA Prediction

Significant SST EOF modes were identified via North's Rule of Thumb (North et al. 1982). A Canonical Correlation Analysis (CCA) was performed using these SST EOF coefficients as the predictor and a variable number of Chl$a$ EOF coefficients as the predictand ($k = 1 \rightarrow 20$). The use of a truncated number of EOF coefficients in a CCA model was demonstrated by Barnett and Preisendorfer (1987) and has been shown to be an effective way of identifying coupled patterns between fields (Bretherton et al. 1992). The resulting model was used to predict Chl$a$ EOF coefficients, which were subsequently used to reconstruct the Chl$a$ field. Error of the reconstructed field was measured against the true Chl$a$

field via MAE.

3) Influence of Noise

The influence of noise in a given gappy dataset on the accuracy of EOF reconstruction was explored for each of the approaches. In the case of remote sensing estimates of chlorophyll, estimation error is typically given as percent difference, implying that error increases proportionally with concentration. Error from SeaWiFS is usually within ±35% for Case I waters, but can reach ±60% (Hu et al. 2001). Estimated error from Globcolour is of a similar magnitude (Globcolour Project 2007). In order to simulate estimation error, normally distributed random numbers of mean = 0 and variable standard deviation (∼0.1–0.5) were added to the log-transformed true Chl$a$ field, which translated to a median percent error of ∼10–30%. EOFs derived from these noisy data fields were used to reconstruct the field using variable degrees of truncation ($k = 1 \rightarrow 50$). Error of the reconstructed field was measured against the true Chl$a$ field via MAE.

3. Results

a. EOF Modes

The top three EOF modes for SST anomaly and Chl$a$ anomaly fields are presented in Fig. 3. All fields show a signal resembling inter-annual ENSO variability in the leading EOF mode. The strong El Niño event of 1997/98 is seen in the corresponding EOF coefficients of the leading mode, with opposing signs for SST and Chl$a$. Such a relationship is to be expected; warm El Niño conditions are a result of a relaxation of trade winds and subsequent lowering of the thermocline, which in turn prevents upwelling of nutrient-rich, cold waters to the euphotic zone where they are used by primary producers. The second EOF mode relates to variations in the main upwelling center west of the archipelago, while the third EOF mode appears related to the shifting inter-tropical convergence zone. All three gappy approaches produced similar spatial EOF patterns as compared to the true Chl$a$ field; however, the LSEOF approach resulted in noisier EOF coefficients as well as much higher λ values, which amplified the variance of the reconstruction relative to the true field. RSEOF and DINEOF produced similar EOF coefficients, both in magnitude and pattern, as compared to those of the true field.

Fig. 4 shows the correlation between EOF coefficients produced by the three approaches. A high loss of orthogonality is evident in the LSEOF approach. Some loss of orthogonality occurs in the RSEOF approach, although all off-diagonal correlations were low ($|R| < 0.2$). There was no loss in orthogonality with DINEOF as the EOFs are ultimately derived from an interpolated, non-gappy matrix.

b. EOF Reconstruction

Examples of daily field reconstructions using the top 20 EOF are presented in Fig. 5. RSEOF and DINEOF generally result in lower daily MAE, but this is not consistent for all days presented. The degree of gappiness and the location of gaps appear to have an effect on how well the EOFs are able to predict the missing values. LSEOF overestimates negative anomalies in the upwelling zone to the west of the archipelago in the July and October maps.

The effect of truncation level on MAE in the reconstruction can be seen in Fig. 6 (left plot). The MAE of the reconstruction using the EOFs of the true field is provided as reference. MAE increases with truncation level when using EOFs derived by LSEOF, while those derived with RSEOF and DINEOF progressively decrease MAE. EOFs derived by the DINEOF approach provided the best fit as evaluated against the true Chl$a$ field.

c. EOF/CCA Prediction

Fig. 6 (right plot) shows the MAE of the predicted Chl$a$ field using the CCA model of SST and Chl$a$ EOF coefficients as predictor and predictand. All models show similar trends in that increasing EOF truncation does not greatly improve MAE. This is due to the fact that the leading EOF coefficients received the highest CCA loadings and carry the highest amount of variance (i.e. λ values) of the observed Chl$a$ field. Subsequent EOF coefficients are down-weighted by the CCA model and contribute little to the prediction. EOF coefficients derived by the DINEOF approach provided the best prediction as evaluated against the true Chl$a$ field.

d. Influence of Noise

The accuracy of reconstruction with LSEOF-derived EOFs was even poorer with noisy fields and, thus, only results for RSEOF and DINEOF are shown. The addition of noise to the data affected the optimal level of truncation and accuracy of the reconstruction of both the RSEOF and DINEOF approaches (Fig. 7). As expected, MAE increases with increasing observation error, while the optimal truncation level decreases. For all levels of error, DINEOF outperformed RSEOF in terms of the MAE of the reconstruction, and was able to incorporate a higher number of EOFs before MAE increased.

4. Discussion

a. EOF Reconstruction and Prediction

Of the gappy EOF approaches evaluated, DINEOF is shown to be superior as indicated by its accuracy in the reconstruction and prediction of data fields. The RSEOF approach was also successful in providing reliable results, yet with a slightly lower accuracy, while the more traditional LSEOF approach was not appropriate for reconstruction. The LSEOF approach provided similar output in terms of spatial EOF patterns, but corresponding EOF coefficients showed increased noise and amplified λ values leading to increased variance (Fig. 3) and, subsequently, error in the reconstruction (Fig. 6). This approach should be discouraged, as it has been shown here to be deficient in cases where gappiness is high.

We find that the error of the reconstruction (MAE) is positively related to the degree of gappiness in the data. Fig. 8 shows the relationship of increasing MAE with gappiness for daily maps using each of the approaches. RSEOF and DINEOF both dramatically reduce the MAE over that of LSEOF. A slightly lower slope is found for DINEOF as compared to RSEOF, again showing it to be the superior approach.

Field prediction based on the EOF/CCA model also shows the best accuracy for the DINEOF approach. The same issue of increasing MAE with truncation level was not found with the predictive CCA model using the LSEOF-derived EOF coefficients. This is in part due to the fact that the main link between the SST and Chl$a$ anomaly fields is through the leading EOF, whereas later truncation only provide small improvements. Furthermore, the leading EOF is less affected by the problems associated with subsequent EOFs mentioned in Sect.1.b.1. Even when these higher EOF modes are included, the CCA model is able to filter out this noise and prevents a rise in MAE with increasing truncation. Thus, the use of LSEOF-derived EOFs in CCA predictive models appears to be less problematic than in field reconstruction, especially in cases where the strongest correlation is via a dominant leading EOF mode.

DINEOF is also shown to deal better with data fields containing a high degree of noise. In addition to producing more accurate leading EOFs, a larger number of trailing EOFs can be used in the truncated reconstruction (as compared to RSEOF) before error begins to increase (Fig. 7). Thus, DINEOF is better able to determine both leading, large-scale EOFs, as well as higher EOFs, which correspond to small-scale features.

b. Computational Considerations

This work has focused on the accuracy of gappy EOF approaches rather than their respective computational speed since we believe that, for most cases, missing data is more likely to be the limiting factor for many analyses. Nevertheless, it is important to mention the differences between the RSEOF and DINEOF approaches, which may be of interest to larger analyses. Users will need to evaluate whether improvements in EOF accuracy merit the additional computational costs of the DINEOF approach.

The DINEOF approach required ∼400 iterations (i.e. individual SVD operations) to converge on an optimized interpolation using 70 EOFs, while RSEOF provided nearly as good a fit, yet at a fraction of the computational time. As suggested by one of the reviewers, the speed of DINEOF can be increased through the adoption of less strict RMS convergence criteria for earlier EOF modes, while maintaining more strict convergence criteria in later iterations. Furthermore, RSEOF may be used in combination with DINEOF by providing a better first guess estimate of missing values and help reduce the number of iterations needed for convergence.

For very large matrices, the computational speed of both DINEOF and RSEOF can be increased through combination with a Lanczos bidiagonalization, which derives a smaller subset of EOF patterns through partial SVD. The Lanczos solver is included in the UNIX distribution of DINEOF but will need to be implemented for use in other programming languages (e.g. R package *irlba*, Baglama and Reichel 2012).

5. Conclusions

EOFs derived from gappy data by means of a covariance matrix decomposition and subsequent least-squares estimate of EOF coefficients (LSEOF) is demonstrated to be deficient for use in data field reconstruction and prediction. At the heart of this deficiency is the decomposition of a non-positive definite covariance matrix, which results in amplified λ values and EOF coefficients that are not strictly orthogonal. As a consequence, the variance of the reconstructed field is also amplified.

The DINEOF and RSEOF approaches are able to successfully remedy these shortcomings through, respectively, optimal EOF interpolation of missing values or preservation of EOF orthogonality by recursive EOF subtraction. The DINEOF approach is shown to be the superior approach, and is especially useful in deriving smaller-scale features in noisy fields. The RSEOF approach, introduced here, provides an reliable alternative, which may be attractive in exploratory analyses of large data fields or as a means of providing an initial estimate of missing values preceding a more refined interpolation with DINEOF.

Acknowledgments.

We would like to thank the three anonymous reviewers for their helpful critique of this work. The authors are grateful to the German Research Foundation for funding (ID: LO-1143/6). All calculations and figures were done with R (R Core Team 2012).

REFERENCES

Alvera-Azcárate, A., A. Barth, M. Rixen, and J. Beckers, 2005: Reconstruction of incomplete oceanographic data sets using Empirical Orthogonal Functions: Application to the Adriatic sea surface temperature. *Ocean Modelling*, **9 (4)**, 325–346.

Baglama, J. and L. Reichel, 2012: *irlba: Fast partial SVD by implicitly-restarted Lanczos bidiagonalization.* URL `http://CRAN.R-project.org/package=irlba`, R package version 1.0.2.

Barnett, T. P. and R. Preisendorfer, 1987: Origins and levels of monthly and seasonal forecast skill for United States surface air temperatures determined by canonical correlation-analysis. *Monthly Weather Review*, **115 (9)**, 1825–1850.

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

Bien, J. and R. J. Tibshirani, 2011: Sparse estimation of a covariance matrix. *Biometrika*, **98 (4)**, 807–820.

Björnsson, H. and S. Venegas, 1997: A manual for EOF and SVD analyses of climate data. Tech. rep., Department of Atmospheric and Oceanic Sciences and Centre for Climate and Global Change Research, McGill University.

Boyd, J. D., E. P. Kennelly, and P. Pistek, 1994: Estimation of eof expansion coefficients from incomplete data. *Deep-Sea Research Part I-Oceanographic Research Papers*, **41 (10)**, 1479–1488.

Bretherton, C. S., C. Smith, and J. M. Wallace, 1992: An intercomparison of methods for finding coupled patterns in climate data. *Journal of Climate*, **5 (6)**, 541–560.

Chavez, F. P. and R. Brusca, 1991: *The Galapagos Islands and their relation to oceanographic processes in the tropical Pacific*, 933. Plenum Press, New York.

Globcolour Project, 2007: Full Validation Report. Online documentation, ACRI-ST/LOV, Sophia-Antipolis Cedex, France. `http://www.globcolour.info/validation/report/GlobCOLOUR_FVR_v1.1.pdf`.

Hasselmann, K., 1988: PIPs and POPs: The reduction of complex dynamical systems using principal interaction and oscillation patterns. *J. Geophys. Res*, **93 (11)**, 015–11.

Hu, C., K. Carder, and F. Muller-Karger, 2001: How precise are SeaWiFS ocean color estimates? Implications of digitization-noise errors. *Remote Sensing of Environment*, **76 (2)**, 239–249.

Kaplan, A., Y. Kushnir, and M. Cane, 2000: Reduced space optimal interpolation of historical marine sea level pressure: 1854-1992*. *Journal of Climate*, **13 (16)**, 2987–3002.

Kaplan, A., Y. Kushnir, M. A. Cane, and M. B. Blumenthal, 1997: Reduced space optimal analysis for historical data sets: 136 years of Atlantic sea surface temperatures. *Journal Of Geophysical Research-Oceans*, **102 (C13)**, 27 835–27 860.

Marshall, J., A. Adcroft, C. Hill, L. Perelman, and C. Heisey, 1997: A finite-volume, incompressible Navier Stokes model for studies of the ocean on parallel computers. *Journal Of Geophysical Research-Oceans*, **102 (C3)**, 5753–5766.

MITgcm Group, 2012: MITgcm User Manual. Online documentation, MIT/EAPS, Cambridge, MA 02139, USA. `http://mitgcm.org/public/r2_manual/latest/online_documents`.

North, G., T. Bell, R. Cahalan, and F. Moeng, 1982: Sampling errors in the estimation of Empirical Orthogonal Functions. *Mon. Wea. Rev.*, **110**, 699–706.

Pennington, J., K. Mahoney, V. Kuwahara, D. Kolber, R. Calienes, and F. Chavez, 2006: Primary production in the eastern tropical Pacific: A review. *Progress in Oceanography*, **69 (2)**, 285–317.

R Core Team, 2012: *R: A Language and Environment for Statistical Computing*. Vienna, Austria, R Foundation for Statistical Computing, URL `http://www.R-project.org/`, ISBN 3-900051-07-0.

Schartau, M., A. Engel, J. Schroter, S. Thoms, C. Volker, and D. Wolf-Gladrow, 2007: Modelling carbon overconsumption and the formation of extracellular particulate organic carbon. *Biogeosciences*, **4 (4)**, 433–454, 1726-4170.

Taylor, M. H., M. Losch, and A. Bracher, 2013: On the drivers of phytoplankton blooms in the antarctic marginal ice zone: A modeling approach. *Journal of Geophysical Research-Oceans*, **118**, 63–75.

von Storch, H. and F. W. Zwiers, 1999: *Statistical analysis in climate research*. Cambridge University Press, Cambridge ; New York, 484 pp.

Wilks, D. S., 2006: *Statistical methods in the atmospheric sciences*. 2d ed., International geophysics series, Academic Press, Amsterdam ; Boston, 627 pp.

Willmott, C. J. and K. Matsuura, 2005: Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, **30 (1)**, 79.

List of Figures

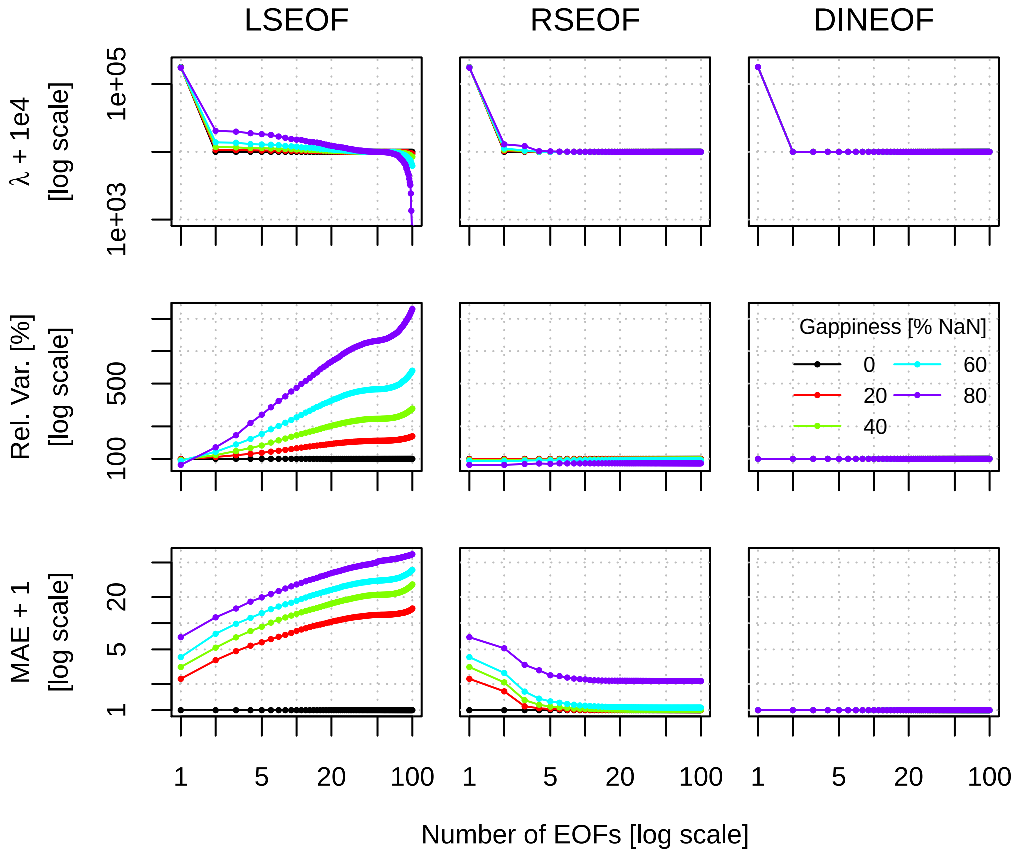

FIG. 1. Comparison of gappy EOF approaches in the accuracy of field reconstruction under variable levels EOF truncations. The gappy field contains a single signal with differing levels of gappiness. λ is determined directly from the EOF analysis. Relative variance compares the reconstructed field's variance to that of the observed gappy field. Mean absolute error (MAE) is calculated between the reconstructed field and the true non-gappy field. The amplified λ values calculated by LSEOF result in EOFs that carry a higher degree of variance and, thus, increased error (MAE) in the reconstruction. Plots for DINEOF are nearly identical for all levels of gappiness, preventing the visualization of all lines.

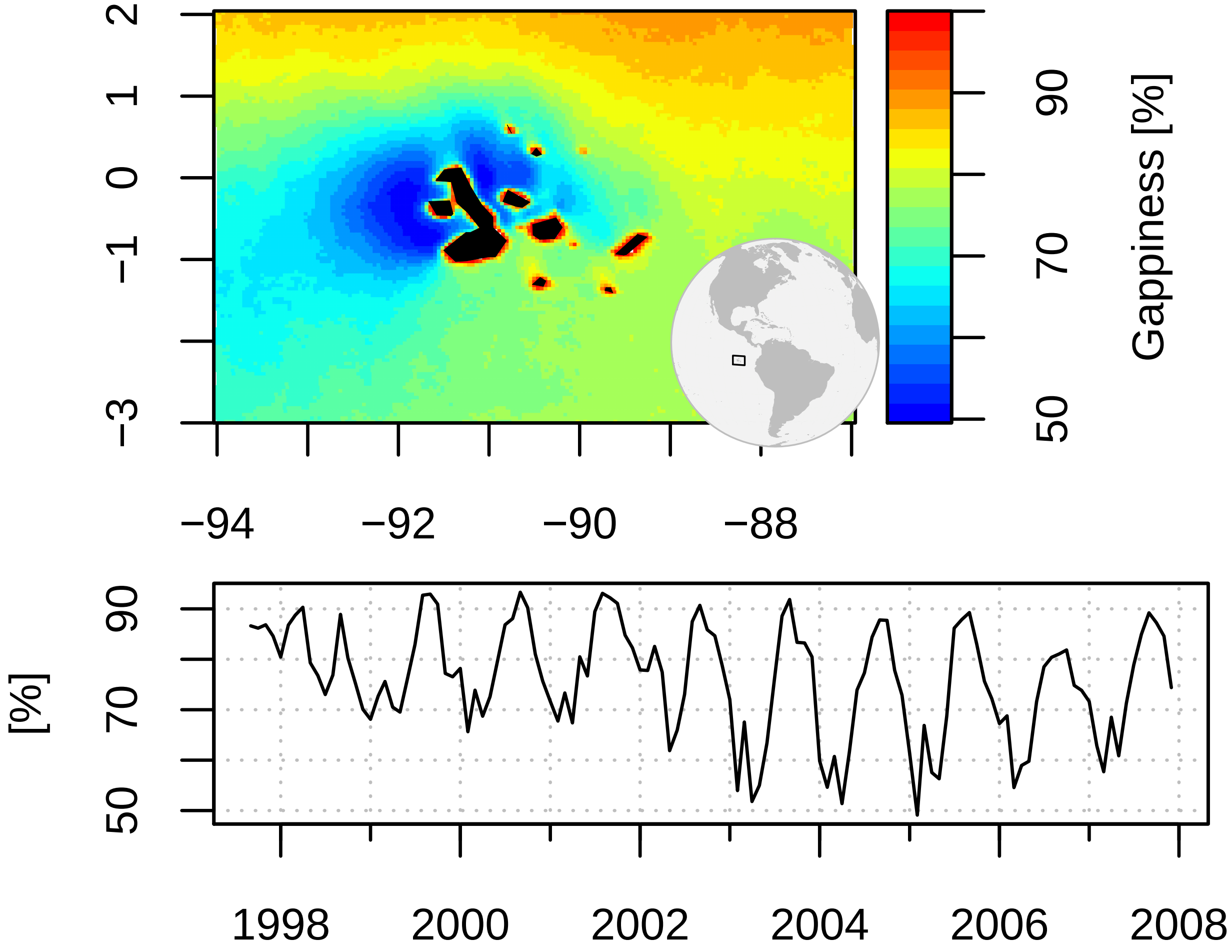

FIG. 2. Gappiness of remote sensing Globcolour Project (`http://www.globcolour.info`) chlorophyll data for the Galapagos Archipelago. For the period of 1997-2007, average daily mean gappiness is shown in the map, while the time series of monthly mean gappiness for the mapped area is shown below. Time axis ticks indicate the beginning of each year (Jan 1st).

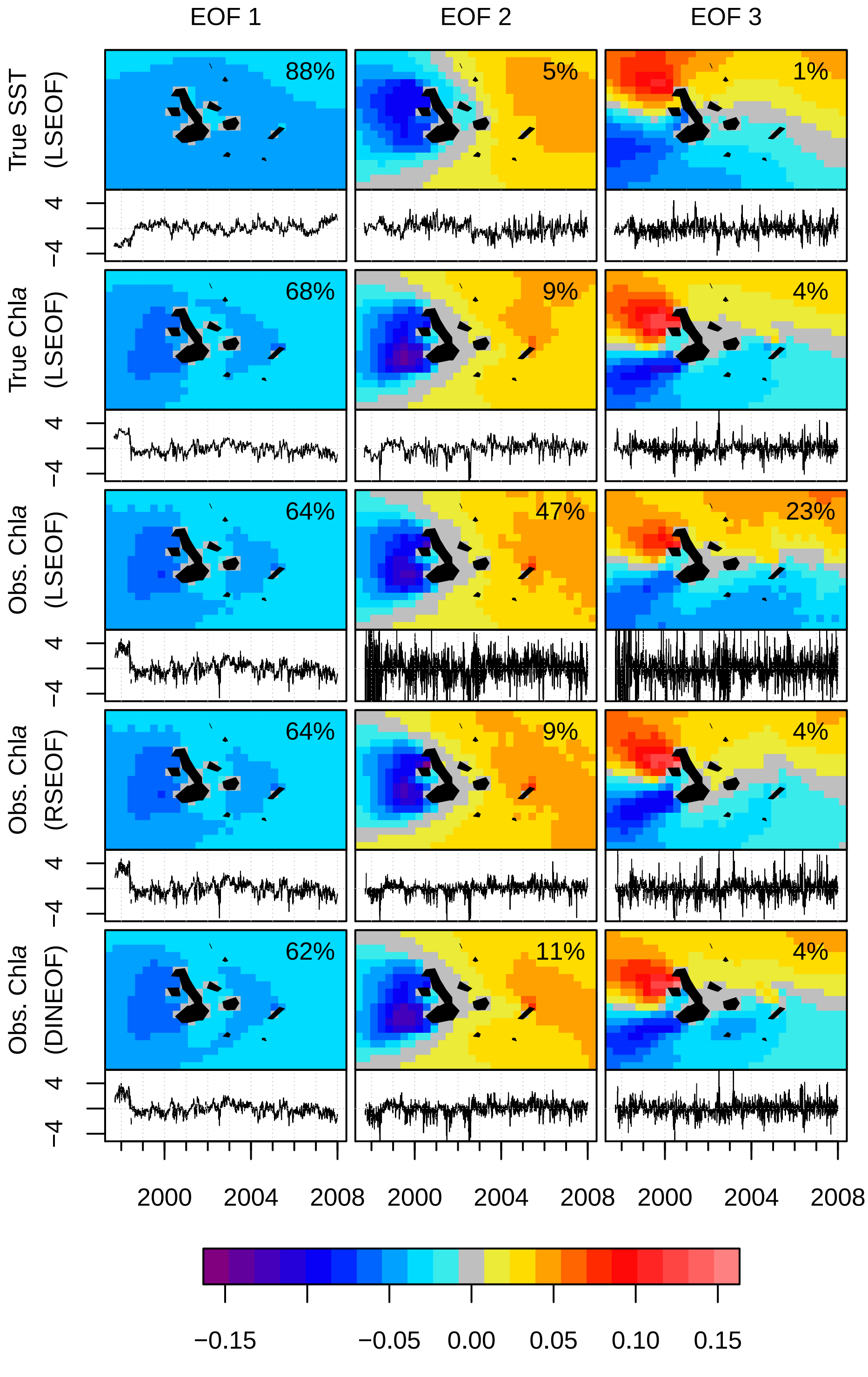

Fig. 3. The top three EOF modes derived from true SST anomaly, true Chl$a$ anomaly, and observed (i.e. gappy) Chl$a$ anomaly fields. Observed Chl$a$ anomaly fields were subjected to the three gappy EOF approaches (bottom three rows). Relative explained variance of each EOF mode as compared to the variance of the observed Chl$a$ anomaly field is displayed in the upper right corner of each map. Time axis ticks indicate the beginning of each year (Jan 1st)

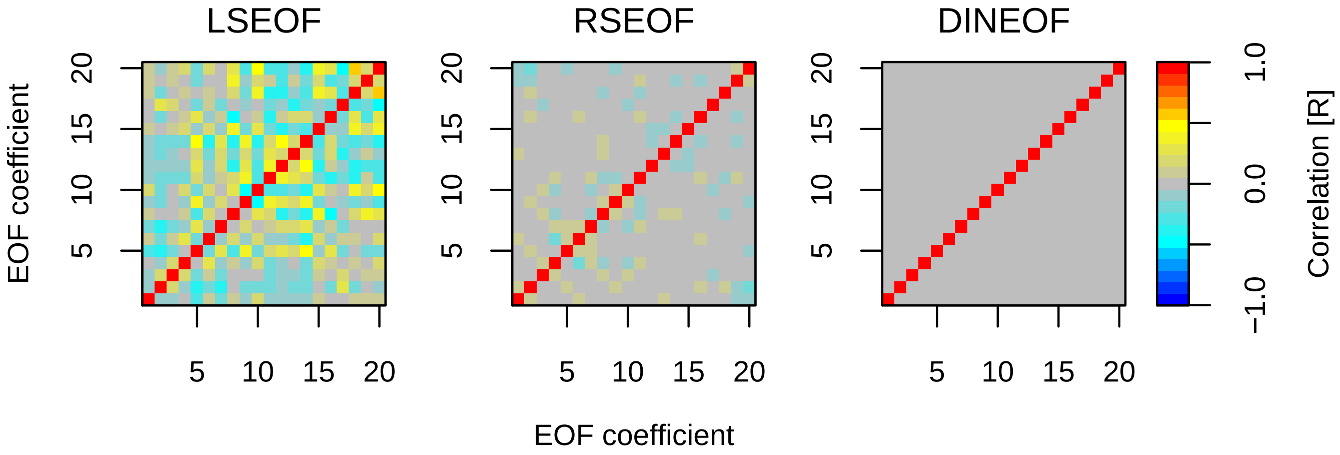

FIG. 4. Correlation of top 20 EOF coefficients from the observed (i.e. gappy) Chl$a$ anomaly field as derived from the three EOF approaches.

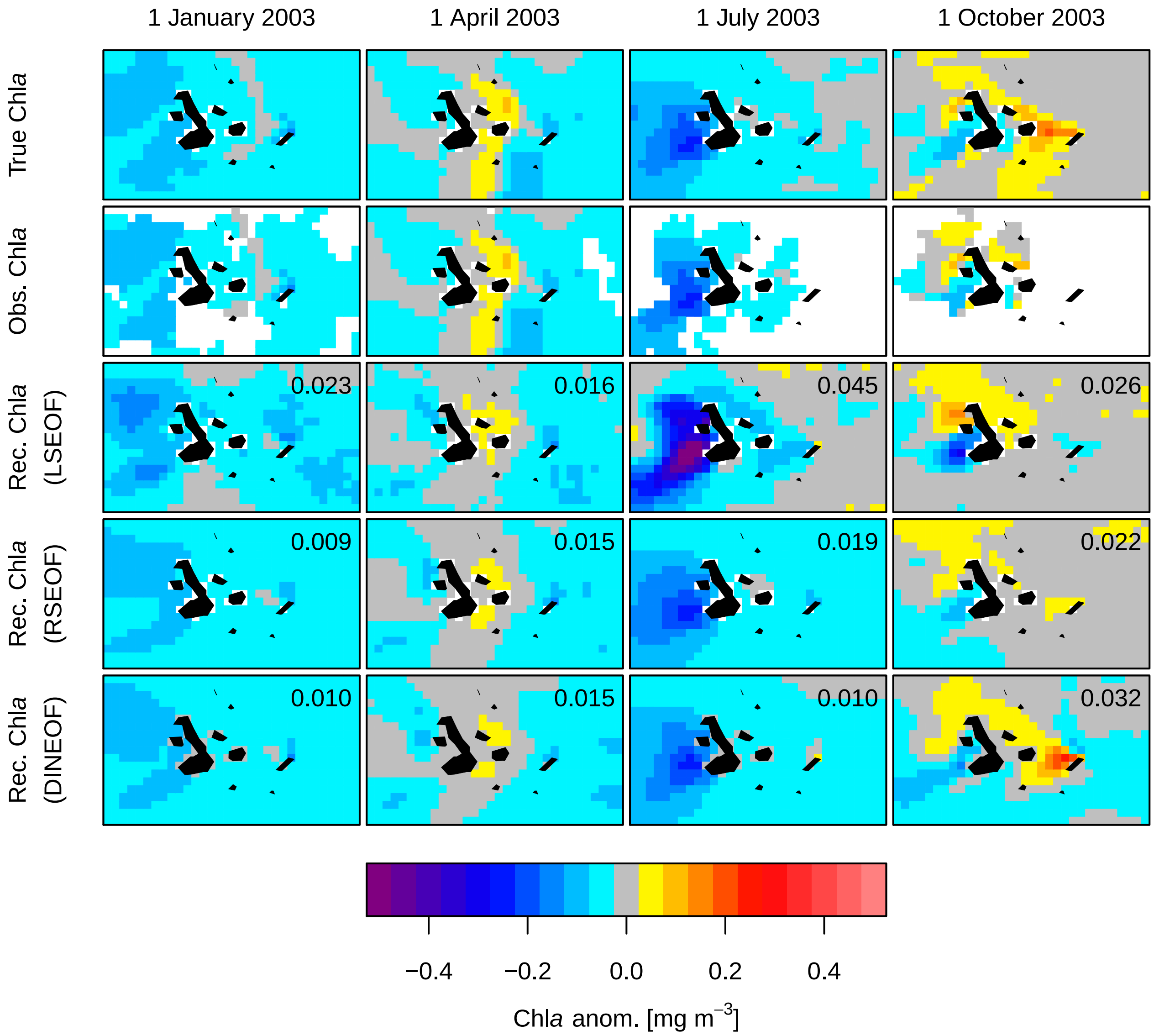

FIG. 5. Examples of reconstructed Chl$a$ anomalies for several dates using the top 20 EOFs derived from the three gappy EOF approaches. Maps of the true data are in the top row while the observed (i.e. gappy) data are shown in the second row. Grids with missing values are white in color. Reconstructions using the gappy approaches are in the lower three rows. The mean absolute error (MAE) of each day's reconstruction, as compared to the true non-gappy data, is displayed in the upper right corner of the maps.

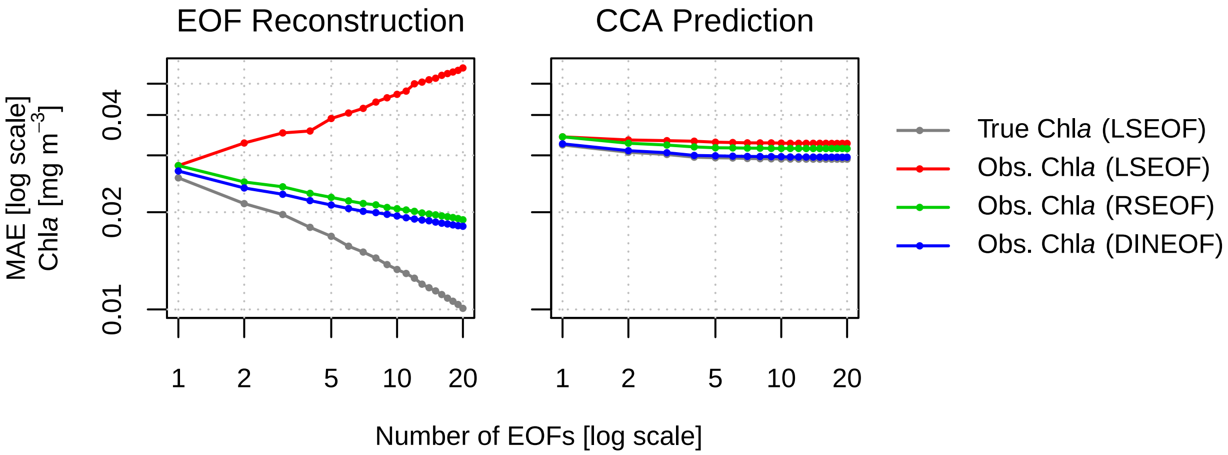

FIG. 6. Mean Absolute Error (MAE) of EOF reconstructed (left) and CCA predicted (right) fields of Chl$a$ anomalies. EOFs were derived from the either the true or observed (i.e. gappy) Chl$a$ anomaly fields and error was gauged against true Chl$a$ anomaly field. The CCA model uses normalized EOF coefficients from true SST anomaly ($n = 6$) and observed Chl$a$ anomaly (variable n) fields as predictor and predictand, respectively. The MAE of the true Chl$a$ field (grey line) is provided as a reference for a perfect reconstruction/prediction.

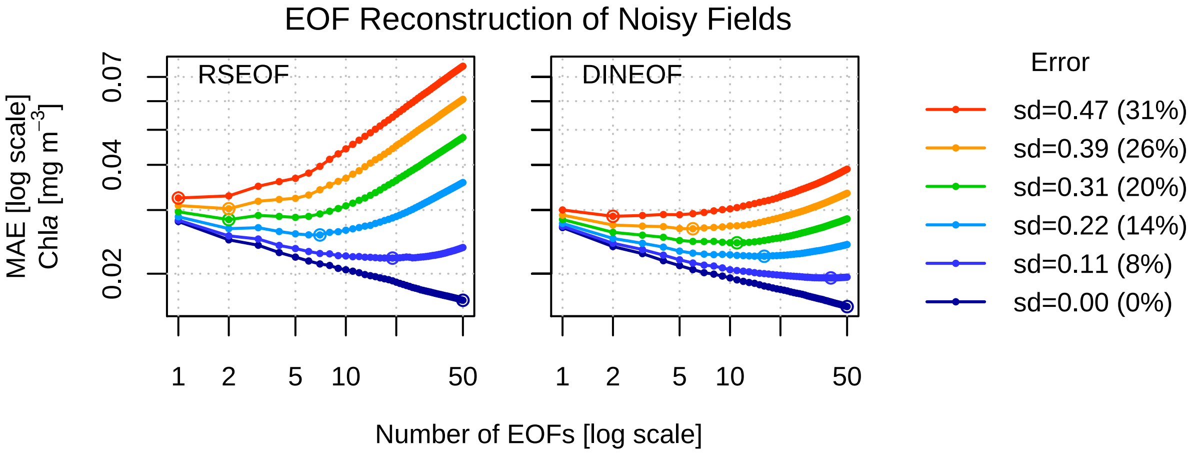

FIG. 7. Mean absolute error (MAE) of EOF reconstructions for the observed (i.e. gappy) Chl$a$ anomaly field with variable error (i.e. noise) added to the true signal. Error levels are given as standard deviation of log-transformed Chl$a$, with corresponding median percent error given in parentheses. Open circle symbols designate the truncation level of lowest MAE.

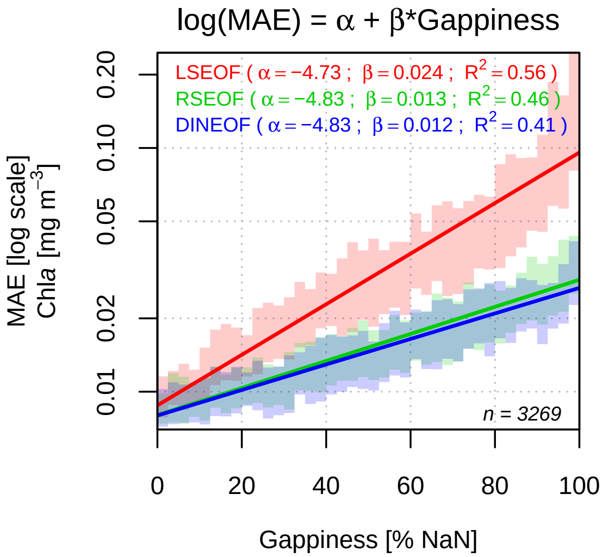

FIG. 8. Linear regressions of daily spatial gappiness versus log-transformed MAE of the EOF reconstructed Chl$a$ anomaly fields (using the top 20 EOFs) for each gappy EOF approach. MAE is calculated against the true field. Shaded areas show the 25% and 75% quartiles for gappiness intervals by approach. Fitted regressions are shown as solid lines. Regression coefficients and R^2 values are displayed at the top of the plot area. All regressions are based on $n = 3269$ data points and are significantly different from each other at the level $p < 0.001$ (F-test).