

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

3     MARC H. TAYLOR<sup>1,2</sup> \*, MARTIN LOSCH<sup>1</sup>, MANFRED WENZEL<sup>1</sup>, JENS SCHRÖTER<sup>1</sup>

<sup>1</sup>*Alfred Wegener Institute for Polar and Marine Research, Bremerhaven, Germany*

<sup>2</sup>*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](mailto: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.

# 21 1. Introduction

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

## 36 a. Basic EOF Approaches

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

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

44 1) DIRECT DATA MATRIX DECOMPOSITION

45 The direct approach via SVD is as follows:

46 
$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T, \quad x_{ij} = \sum_{k=1,N} u_{ik} \sigma_k v_{kj} \quad (1)$$

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

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

52 2) COVARIANCE MATRIX DECOMPOSITION

53 The covariance matrix decomposition approach requires a square matrix. One first con-  
 54 structs a covariance matrix  $\mathbf{C}$ ,

$$\mathbf{C} = \frac{1}{M} \mathbf{X}^T \mathbf{X}, \quad c_{jj'} = \frac{\sum_{i=1}^M x_{ji} x_{ij'}}{M} \quad (3)$$

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

$$\mathbf{C} = \mathbf{E}\mathbf{\Lambda}\mathbf{E}^T, \quad c_{jj'} = \sum_{k=1}^N e_{jk} \lambda_{ekj'} \quad (4)$$

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

$$\mathbf{A} = \mathbf{X}\mathbf{E}, \quad a_{ik} = \sum_j x_{ij} e_{jk} \quad (5)$$

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

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

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

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

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

### 67 *b. Gappy Data EOF Approaches*

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

#### 71 1) COVARIANCE MATRIX DECOMPOSITION / LEAST-SQUARES ESTIMATION OF COEF- 72 FICIENTS - LSEOF

73 Due to the inability to decompose a matrix containing missing values, a direct data matrix  
 74 decomposition via SVD is not possible. The approach via covariance matrix decomposition  
 75 is possible; however, due to the missing values, one must adopt a least-squares approach  
 76 that takes into account the number of paired observations between samples. In this work,  
 77 we will refer to this approach as “Least-Squares Empirical Orthogonal Functions” (LSEOF).  
 78 In LSEOF, the above covariance matrix calculation (Eq. 3) must be scaled by the number  
 79 of shared, non-missing values between samples (von Storch and Zwiers 1999; Kaplan et al.  
 80 1997; Boyd et al. 1994),

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

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

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

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

84 where  $\epsilon$  is the error and  $\phi$  is the objective function with the solution

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

85 where  $J_i$  is the set of non-missing values at time  $i$ . Note that the denominator reduces to  
 86 1 when there are no missing values; thus, equaling the scalar product for  $\mathbf{A}$  shown above  
 87 (Eq. 5).

88 Several issues have been identified with the use of this approach. First and foremost  
 89 is the problem that the calculation of a covariance matrix derived from gappy data is not  
 90 necessarily positive definite, and decomposition via LSEOF can contain negative  $\lambda$  values.  
 91 Since the variance of the data set is contained in the trace of the covariance matrix  $\mathbf{C}$  and,  
 92 subsequently, equal to the sum of  $\mathbf{\Lambda}$ , having negative values will mean that other EOFs  $e_k$   
 93 will have higher  $\lambda_k$  than in reality; thus, overestimating their amplitude and the amount  
 94 of explained variance contained therein (Beckers and Rixen 2003; Björnsson and Venegas  
 95 1997).

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

101 A second problem is that the decomposition of a non-positive definite covariance matrix is  
 102 a loss of orthogonality between EOFs (Björnsson and Venegas 1997), which makes their use  
 103 in predictive models less attractive. For example, Barnett and Preisendorfer (1987) describe  
 104 a method of Canonical Correlation Analysis (CCA) based on EOF coefficients, which is

105 useful in determining the correlation between coupled fields. When correlations are high,  
106 issues associated with multi-collinearity can affect the predictive ability of the model.

## 107 2) DATA INTERPOLATING EMPIRICAL ORTHOGONAL FUNCTIONS - DINEOF

108 An alternate approach, DINEOF (Beckers and Rixen 2003; Alvera-Azcárate et al. 2005),  
109 interpolates missing values via an iterative SVD algorithm. DINEOF has similarities with  
110 approaches aimed at iterative estimation of the covariance matrix (e.g. Bien and Tibshirani  
111 2011), although DINEOF directly iterates values in the data matrix itself.

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

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

127 3) RECURSIVELY-SUBTRACTED EMPIRICAL ORTHOGONAL FUNCTIONS - RSEOF

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

135 The approach is as follows:

- 136 i. The observed data matrix  $\mathbf{X}^O$  is (optionally) centered and/or scaled prior to the de-  
 137 composition, and is renamed as  $\mathbf{X}^i$  for the first iteration,  $i = 1$ .
- 138 ii. A covariance matrix  $\mathbf{C}^i$  is calculated from  $\mathbf{X}^i$  (Eq. 8).
- 139 iii.  $\mathbf{C}^i$  is subjected to Eigenvalue decomposition giving  $\mathbf{E}^i$  and  $\mathbf{\Lambda}^i$  (Eq. 4).
- 140 iv.  $\mathbf{A}^i$  is computed using the least-squares approach (Eq. 10)
- 141 v. A truncated version of the data is reconstructed using the leading EOF mode,  $e_1^i$  and  
 142  $a_1^i$ , resulting in  $\mathbf{X}^{recon,i}$ .
- 143 vi. This field is then subtracted from the data to give a new field for iteration  $i + 1$ ;  
 144  $\mathbf{X}^{i+1} = \mathbf{X}^i - \mathbf{X}^{recon,i}$
- 145 vii. Steps ii-vi are then iterated until a given criterion (e.g. for  $i \rightarrow N$ ; remaining %  
 146 variance level, as calculated by  $\sum \text{tr}(\mathbf{C}^i)$ ; minimization of reconstruction error, e.g.  
 147 MAE, RMS).



148 *c. Data Reconstruction*

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

$$\mathbf{X} = \mathbf{A}\mathbf{E}^T, \quad x_{ij} = \sum_{k=1}^N a_{ik} e_{kj} \quad (11)$$

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

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

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

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

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

169 where  $t_i = i2\pi/M$ ,  $s_j = j$ ,  $M = 200$  and  $N = 100$ . Differing levels of gappiness (20, 40, 60

170 and 80%) are randomly distributed throughout the field. The leading  $\lambda$  values are nearly  
 171 identical for all approaches although trailing  $\lambda$ 's are amplified substantially in the LSEOF  
 172 approach. This amplification increases with the degree of gappiness in the observed field  
 173 (Fig. 1, top panels). Statistics relating to field reconstruction can be seen in the middle and  
 174 bottom panels of Fig. 1. The effect of  $\lambda$  amplification in the LSEOF approach is evident  
 175 in the variance of the reconstructed field relative to true non-gappy field. Reconstructions  
 176 using EOFs derived from RSEOF and DINEOF do not exceed a relative variance of 100%.  
 177 Another statistic describing the fit of the reconstruction is that of the mean absolute error  
 178 (MAE), which is calculated as follows

$$MAE = \frac{1}{n} \sum_{k=1}^n |\text{pred}_k - \text{obs}_k| \quad (13)$$

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

186 The aims of the present work are to further evaluate the performance of these EOF  
 187 approaches in the reconstruction and prediction of gappy data fields. Towards this aim, we  
 188 consider a more realistic example using modeled surface Chlorophyll  $a$  (Chl $a$ ) concentrations  
 189 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

216 the availability of nutrients to primary producers via upwelling. By contrast, ENSO-positive  
217 (El Niño) conditions are a result of weakened tradewinds, causing surface waters to relax  
218 back to the east, which lowers the thermocline and the EUC. As a result, the availability of  
219 cool nutrient-rich waters to upwelling is decreased and primary production is dramatically  
220 reduced.

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

#### 226 *b. Synthetic Data Set*

227 In order to obtain full, non-gappy data fields, we use model-derived data. The model  
228 consisted of a biogeochemical model, REcoM (Regulated Ecosystem Model) (Schartau et al.  
229 2007), coupled to a global general circulation model, MITgcm (Massachusetts Institute of  
230 Technology General Circulation Model) (Marshall et al. 1997; MITgcm Group 2012). The  
231 model had a mean horizontal resolution of 18 km and a vertical resolution of 10 m near the  
232 surface. The simulation spanned the years 1992 through 2007 (for additional details, see  
233 Taylor et al. 2013).

234 Daily 4.63 km resolution Globcolour chlorophyll data were used to create a cloud mask  
235 for the modeled data fields. When no valid data values were recorded within each larger  
236 grid of the model, the matrix location was classified as a missing value. In this way, we were  
237 able to obtain both the “true” non-gappy field and an “observed” gappy data field masked  
238 primarily by clouds. We examined the region between  $93^{\circ}W - 88^{\circ}W$  and  $1^{\circ}N - 2^{\circ}S$   
239 for the period coinciding with remote-sensing estimates (1 September 1997 – 31 December  
240 2007). Additionally, modelled sea surface surface temperature (SST) fields were used for  
241 the construction of a predictive CCA model. Both Chl $a$  and SST data were transformed to

242 anomalies by subtracting the long-term monthly means from the time series of each grid.  
243 The resulting dimensions of the data matrices were  $3774 \times 608$  (day  $\times$  grid).

### 244 *c. Analyses of Performance*

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

#### 253 1) EOF RECONSTRUCTION

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

#### 256 2) EOF/CCA PREDICTION

257 Significant SST EOF modes were identified via North's Rule of Thumb (North et al.  
258 1982). A Canonical Correlation Analysis (CCA) was performed using these SST EOF coef-  
259 ficients as the predictor and a variable number of Chl $a$  EOF coefficients as the predictand  
260 ( $k = 1 \rightarrow 20$ ). The use of a truncated number of EOF coefficients in a CCA model was  
261 demonstrated by Barnett and Preisendorfer (1987) and has been shown to be an effective  
262 way of identifying coupled patterns between fields (Bretherton et al. 1992). The resulting  
263 model was used to predict Chl $a$  EOF coefficients, which were subsequently used to recon-  
264 struct the Chl $a$  field. Error of the reconstructed field was measured against the true Chl $a$

265 field via MAE.

### 266 3) INFLUENCE OF NOISE

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

## 278 3. Results

### 279 a. EOF Modes

280 The top three EOF modes for SST anomaly and Chl $a$  anomaly fields are presented in  
281 Fig. 3. All fields show a signal resembling inter-annual ENSO variability in the leading EOF  
282 mode. The strong El Niño event of 1997/98 is seen in the corresponding EOF coefficients  
283 of the leading mode, with opposing signs for SST and Chl $a$ . Such a relationship is to be  
284 expected; warm El Niño conditions are a result of a relaxation of trade winds and subsequent  
285 lowering of the thermocline, which in turn prevents upwelling of nutrient-rich, cold waters to  
286 the euphotic zone where they are used by primary producers. The second EOF mode relates  
287 to variations in the main upwelling center west of the archipelago, while the third EOF mode

288 appears related to the shifting inter-tropical convergence zone. All three gappy approaches  
289 produced similar spatial EOF patterns as compared to the true *Chl $a$*  field; however, the  
290 LSEOF approach resulted in noisier EOF coefficients as well as much higher  $\lambda$  values, which  
291 amplified the variance of the reconstruction relative to the true field. RSEOF and DINEOF  
292 produced similar EOF coefficients, both in magnitude and pattern, as compared to those of  
293 the true field.

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

#### 299 *b. EOF Reconstruction*

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

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

310 *c. EOF/CCA Prediction*

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

319 *d. Influence of Noise*

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

327 **4. Discussion**

328 *a. EOF Reconstruction and Prediction*

329 Of the gappy EOF approaches evaluated, DINEOF is shown to be superior as indicated  
330 by its accuracy in the reconstruction and prediction of data fields. The RSEOF approach was  
331 also successful in providing reliable results, yet with a slightly lower accuracy, while the more  
332 traditional LSEOF approach was not appropriate for reconstruction. The LSEOF approach



333 provided similar output in terms of spatial EOF patterns, but corresponding EOF coefficients  
334 showed increased noise and amplified  $\lambda$  values leading to increased variance (Fig. 3) and,  
335 subsequently, error in the reconstruction (Fig. 6). This approach should be discouraged, as  
336 it has been shown here to be deficient in cases where gappiness is high.

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

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

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

358 *b. Computational Considerations*

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

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

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

378 **5. Conclusions**

379 EOFs derived from gappy data by means of a covariance matrix decomposition and sub-  
380 sequent least-squares estimate of EOF coefficients (LSEOF) is demonstrated to be deficient  
381 for use in data field reconstruction and prediction. At the heart of this deficiency is the de-  
382 composition of a non-positive definite covariance matrix, which results in amplified  $\lambda$  values

383 and EOF coefficients that are not strictly orthogonal. As a consequence, the variance of the  
384 reconstructed field is also amplified.

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

392 *Acknowledgments.*

393 We would like to thank the three anonymous reviewers for their helpful critique of this  
394 work. The authors are grateful to the German Research Foundation for funding (ID: LO-  
395 1143/6). All calculations and figures were done with R (R Core Team 2012).

## REFERENCES

- 398 Alvera-Azcárate, A., A. Barth, M. Rixen, and J. Beckers, 2005: Reconstruction of incom-  
399 plete oceanographic data sets using Empirical Orthogonal Functions: Application to the  
400 Adriatic sea surface temperature. *Ocean Modelling*, **9** (4), 325–346.
- 401 Baglama, J. and L. Reichel, 2012: *irlba: Fast partial SVD by implicitly-restarted Lanczos*  
402 *bidiagonalization*. URL <http://CRAN.R-project.org/package=irlba>, R package version  
403 1.0.2.
- 404 Barnett, T. P. and R. Preisendorfer, 1987: Origins and levels of monthly and seasonal fore-  
405 cast skill for United States surface air temperatures determined by canonical correlation-  
406 analysis. *Monthly Weather Review*, **115** (9), 1825–1850.
- 407 Beckers, J. M. and M. Rixen, 2003: EOF calculations and data filling from incomplete  
408 oceanographic datasets. *Journal of Atmospheric and Oceanic Technology*, **20** (12), 1839–  
409 1856.
- 410 Bien, J. and R. J. Tibshirani, 2011: Sparse estimation of a covariance matrix. *Biometrika*,  
411 **98** (4), 807–820.
- 412 Björnsson, H. and S. Venegas, 1997: A manual for EOF and SVD analyses of climate data.  
413 Tech. rep., Department of Atmospheric and Oceanic Sciences and Centre for Climate and  
414 Global Change Research, McGill University.
- 415 Boyd, J. D., E. P. Kennelly, and P. Pistek, 1994: Estimation of eof expansion coefficients  
416 from incomplete data. *Deep-Sea Research Part I-Oceanographic Research Papers*, **41** (10),  
417 1479–1488.

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

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

422 Globcolour Project, 2007: Full Validation Report. Online documentation, ACRI-ST/LOV,  
423 Sophia-Antipolis Cedex, France. [http://www.globcolour.info/validation/report/](http://www.globcolour.info/validation/report/GlobCOLOUR_FVR_v1.1.pdf)  
424 [GlobCOLOUR\\_FVR\\_v1.1.pdf](http://www.globcolour.info/validation/report/GlobCOLOUR_FVR_v1.1.pdf).

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

427 Hu, C., K. Carder, and F. Muller-Karger, 2001: How precise are SeaWiFS ocean color esti-  
428 mates? Implications of digitization-noise errors. *Remote Sensing of Environment*, **76 (2)**,  
429 239–249.

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

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

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

438 MITgcm Group, 2012: MITgcm User Manual. Online documentation, MIT/EAPS,  
439 Cambridge, MA 02139, USA. [http://mitgcm.org/public/r2\\_manual/latest/online\\_](http://mitgcm.org/public/r2_manual/latest/online_documents)  
440 [documents](http://mitgcm.org/public/r2_manual/latest/online_documents).

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

443 Pennington, J., K. Mahoney, V. Kuwahara, D. Kolber, R. Calienes, and F. Chavez, 2006:  
444 Primary production in the eastern tropical Pacific: A review. *Progress in Oceanography*,  
445 **69 (2)**, 285–317.

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

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

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

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

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

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

## 462 List of Figures

- 463 1 Comparison of gappy EOF approaches in the accuracy of field reconstruction  
464 under variable levels EOF truncations. The gappy field contains a single  
465 signal with differing levels of gappiness.  $\lambda$  is determined directly from the  
466 EOF analysis. Relative variance compares the reconstructed field's variance  
467 to that of the observed gappy field. Mean absolute error (MAE) is calculated  
468 between the reconstructed field and the true non-gappy field. The amplified  
469  $\lambda$  values calculated by LSEOF result in EOFs that carry a higher degree  
470 of variance and, thus, increased error (MAE) in the reconstruction. Plots  
471 for DINEOF are nearly identical for all levels of gappiness, preventing the  
472 visualization of all lines. 24
- 473 2 Gappiness of remote sensing Globcolour Project ([http://www.globcolour.](http://www.globcolour.info)  
474 [info](http://www.globcolour.info)) chlorophyll data for the Galapagos Archipelago. For the period of 1997-  
475 2007, average daily mean gappiness is shown in the map, while the time series  
476 of monthly mean gappiness for the mapped area is shown below. Time axis  
477 ticks indicate the beginning of each year (Jan 1<sup>st</sup>). 25
- 478 3 The top three EOF modes derived from true SST anomaly, true Chl*a* anomaly,  
479 and observed (i.e. gappy) Chl*a* anomaly fields. Observed Chl*a* anomaly  
480 fields were subjected to the three gappy EOF approaches (bottom three rows).  
481 Relative explained variance of each EOF mode as compared to the variance  
482 of the observed Chl*a* anomaly field is displayed in the upper right corner of  
483 each map. Time axis ticks indicate the beginning of each year (Jan 1<sup>st</sup>) 26
- 484 4 Correlation of top 20 EOF coefficients from the observed (i.e. gappy) Chl*a*  
485 anomaly field as derived from the three EOF approaches. 27

- 486 5 Examples of reconstructed *Chla* anomalies for several dates using the top 20  
487 EOFs derived from the three gappy EOF approaches. Maps of the true data  
488 are in the top row while the observed (i.e. gappy) data are shown in the  
489 second row. Grids with missing values are white in color. Reconstructions  
490 using the gappy approaches are in the lower three rows. The mean absolute  
491 error (MAE) of each day's reconstruction, as compared to the true non-gappy  
492 data, is displayed in the upper right corner of the maps. 28
- 493 6 Mean Absolute Error (MAE) of EOF reconstructed (left) and CCA predicted  
494 (right) fields of *Chla* anomalies. EOFs were derived from the either the true  
495 or observed (i.e. gappy) *Chla* anomaly fields and error was gauged against  
496 true *Chla* anomaly field. The CCA model uses normalized EOF coefficients  
497 from true SST anomaly ( $n = 6$ ) and observed *Chla* anomaly (variable  $n$ ) fields  
498 as predictor and predictand, respectively. The MAE of the true *Chla* field  
499 (grey line) is provided as a reference for a perfect reconstruction/prediction. 29
- 500 7 Mean absolute error (MAE) of EOF reconstructions for the observed (i.e.  
501 gappy) *Chla* anomaly field with variable error (i.e. noise) added to the true  
502 signal. Error levels are given as standard deviation of log-transformed *Chla*,  
503 with corresponding median percent error given in parentheses. Open circle  
504 symbols designate the truncation level of lowest MAE. 30
- 505 8 Linear regressions of daily spatial gappiness versus log-transformed MAE of  
506 the EOF reconstructed *Chla* anomaly fields (using the top 20 EOFs) for each  
507 gappy EOF approach. MAE is calculated against the true field. Shaded  
508 areas show the 25% and 75% quartiles for gappiness intervals by approach.  
509 Fitted regressions are shown as solid lines. Regression coefficients and  $R^2$   
510 values are displayed at the top of the plot area. All regressions are based on  
511  $n = 3269$  data points and are significantly different from each other at the  
512 level  $p < 0.001$  (F-test). 31



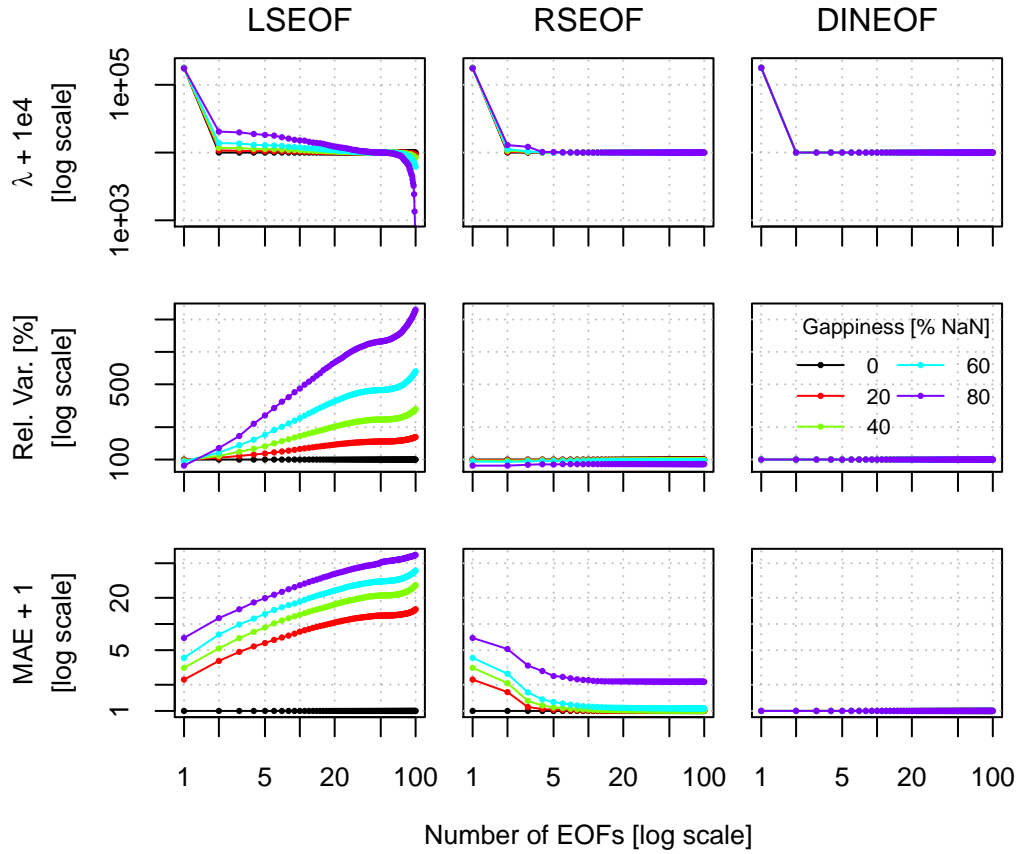


FIG. 1. Comparison of gappy EOF approaches in the accuracy of field reconstruction under variable levels EOF truncations. The gappy field contains a single signal with differing levels of gappiness.  $\lambda$  is determined directly from the EOF analysis. Relative variance compares the reconstructed 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  $\lambda$  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.

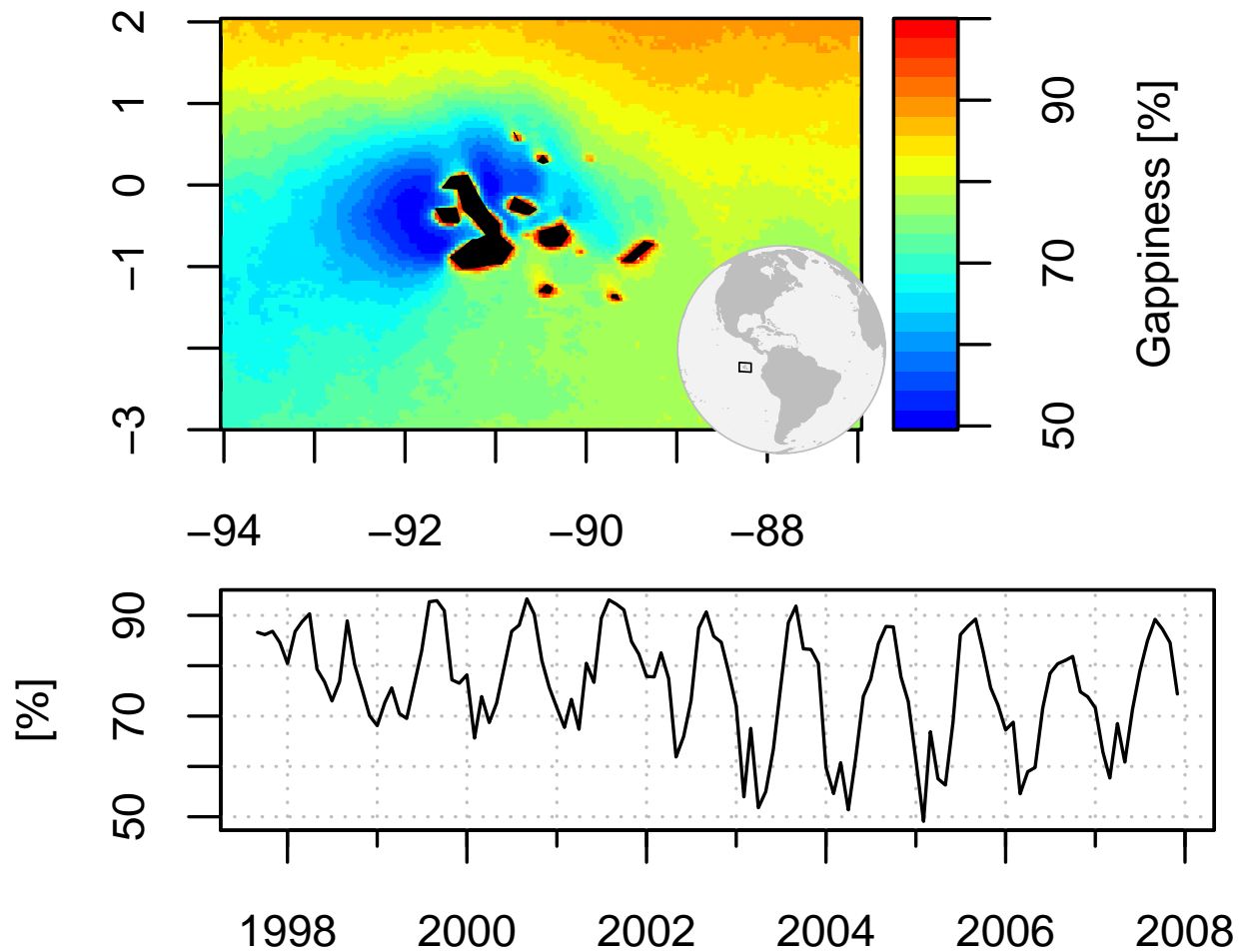


FIG. 2. Gappiness of remote sensing Globcolour Project (<http://www.globcolour.info>) chlorophyll data for the Galapagos Archipelago. For the period of 1997-2007, average daily mean gappiness is shown in the map, while the time series of monthly mean gappiness for the mapped area is shown below. Time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

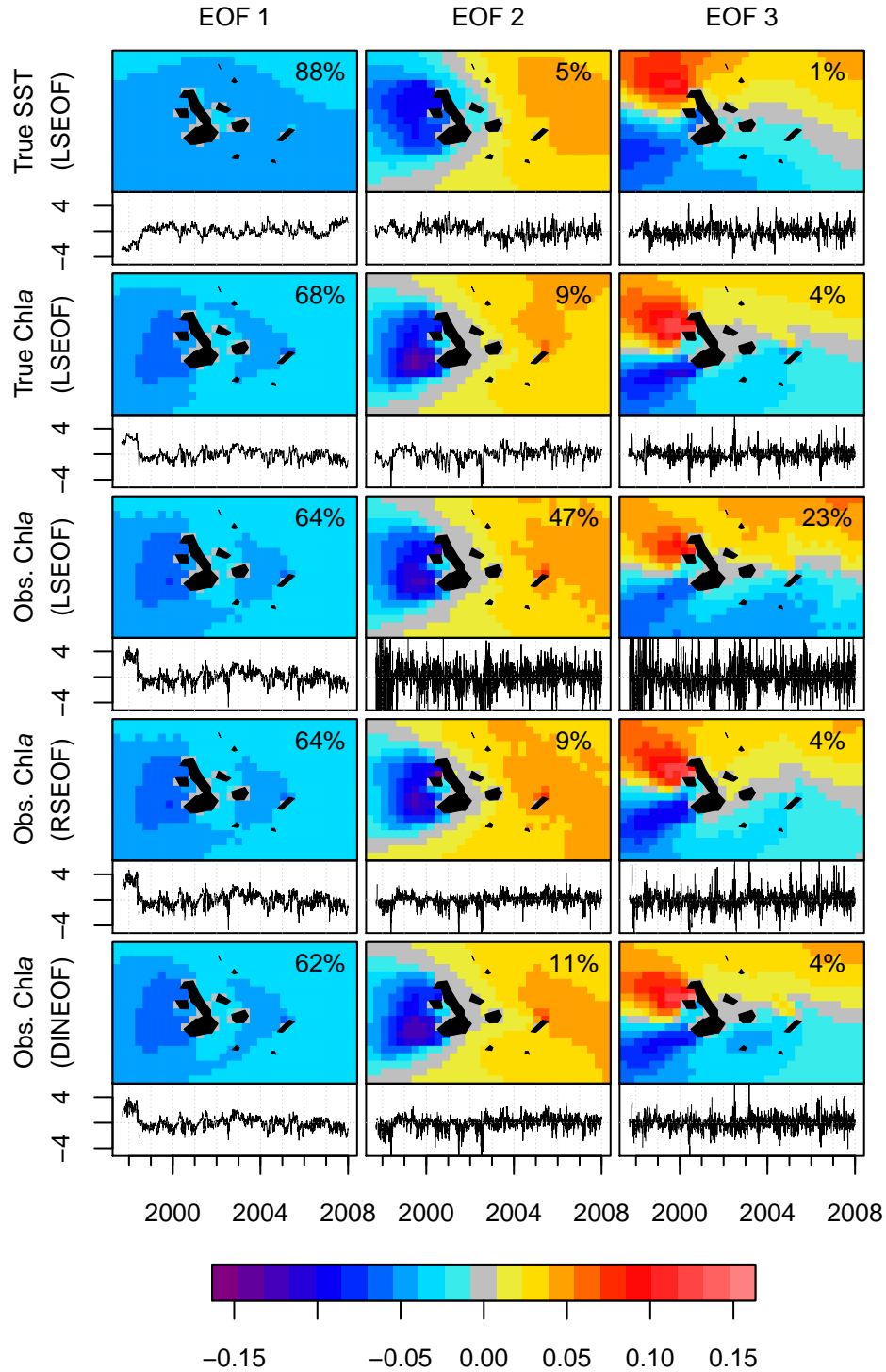


FIG. 3. The top three EOF modes derived from true SST anomaly, true Chla anomaly, and observed (i.e. gappy) Chla anomaly fields. Observed Chla anomaly fields were subjected to the three gappy EOF approaches (bottom three rows). Relative explained variance of each EOF mode as compared to the variance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

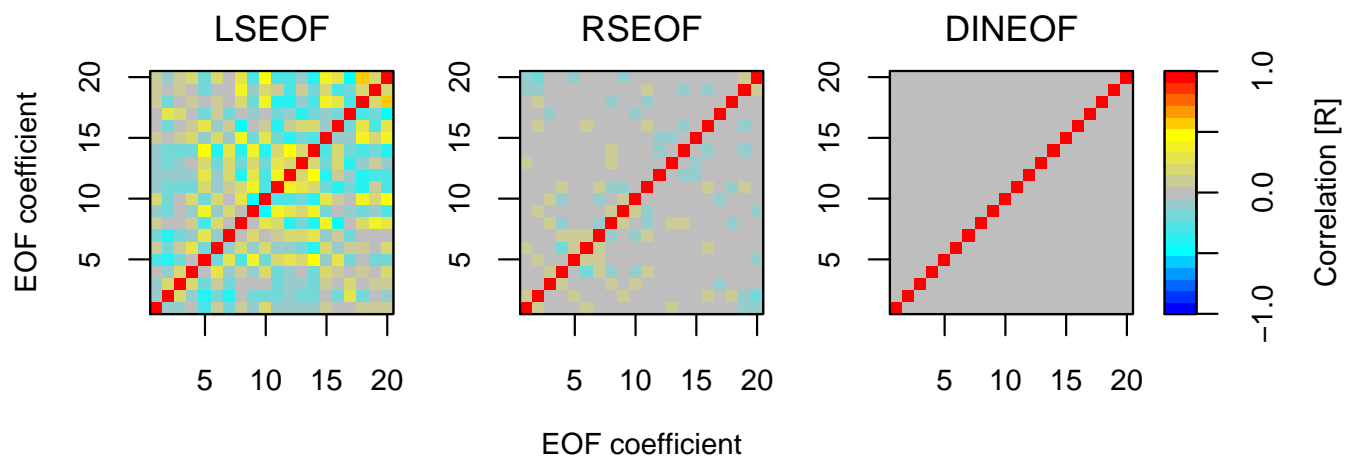


FIG. 4. Correlation of top 20 EOF coefficients from the observed (i.e. gappy) Chl $a$  anomaly field as derived from the three EOF approaches.

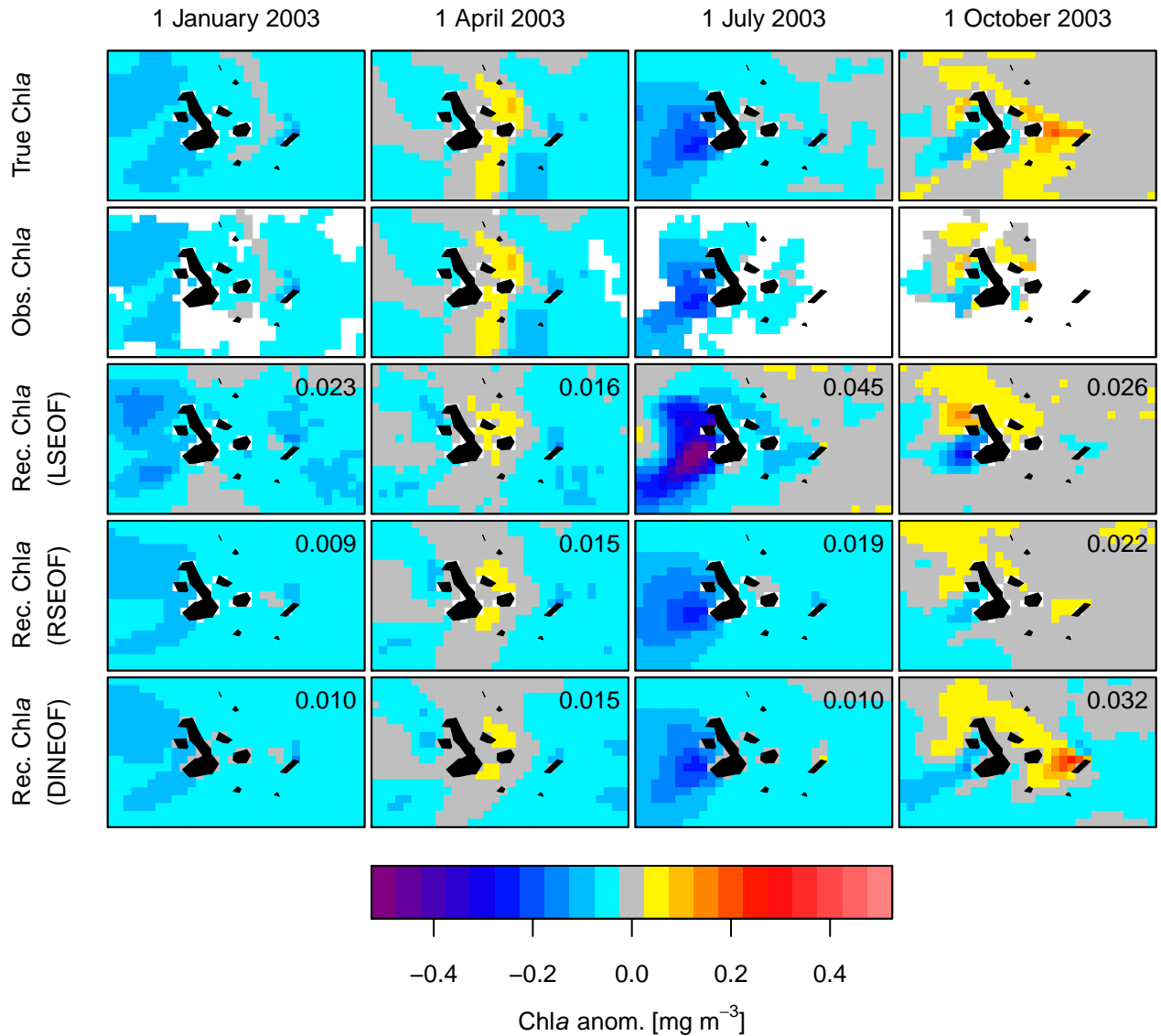


FIG. 5. Examples of reconstructed Chl *a* anomalies for several dates using the top 20 EOFs derived from the three gappy EOF approaches. Maps of the true data are in the top row while the observed (i.e. gappy) data are shown in the second row. Grids with missing values are white in color. Reconstructions 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.

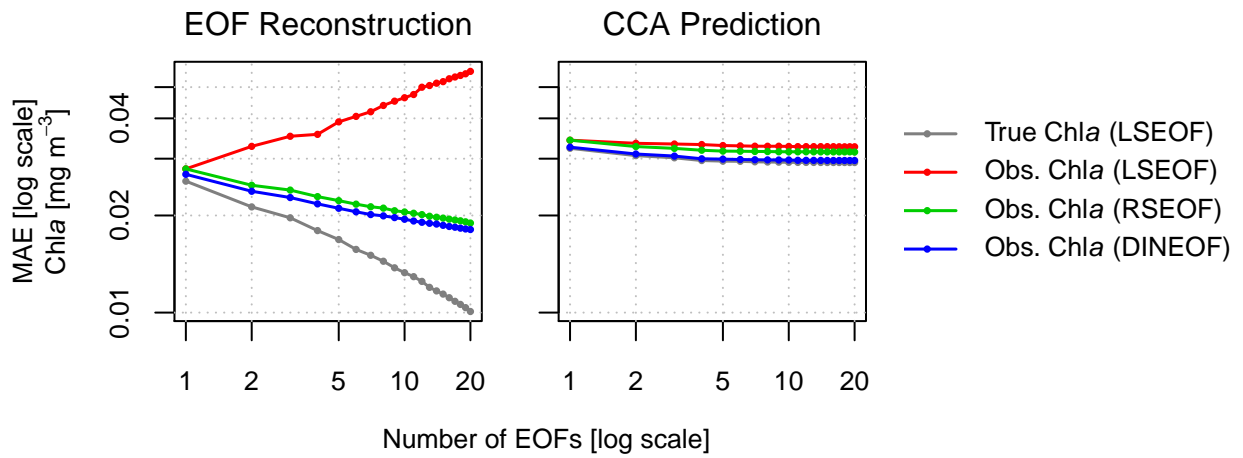


FIG. 6. Mean Absolute Error (MAE) of EOF reconstructed (left) and CCA predicted (right) fields of *Chla* anomalies. EOFs were derived from either the true or observed (i.e. gappy) *Chla* anomaly fields and error was gauged against true *Chla* anomaly field. The CCA model uses normalized EOF coefficients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

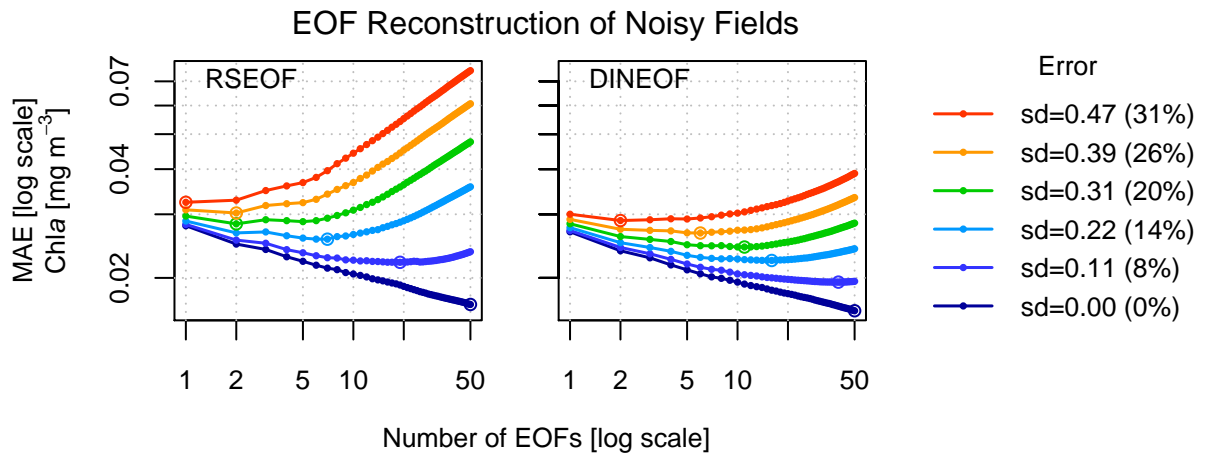


FIG. 7. Mean absolute error (MAE) of EOF reconstructions for the observed (i.e. gappy) *Chla* anomaly field with variable error (i.e. noise) added to the true signal. Error levels are given as standard deviation of log-transformed *Chla*, with corresponding median percent error given in parentheses.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

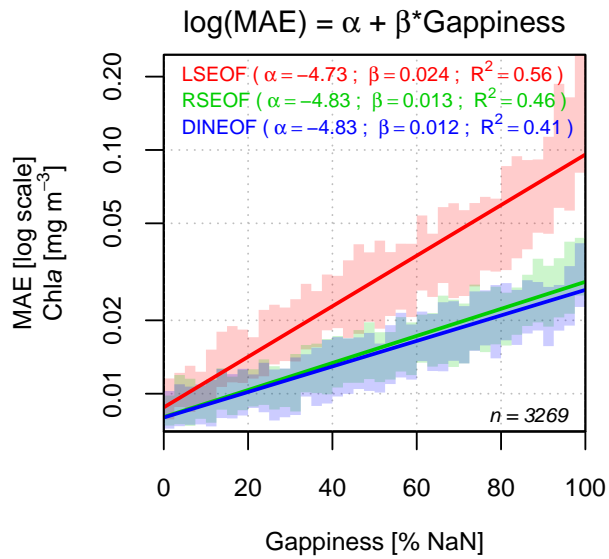


FIG. 8. Linear regressions of daily spatial gappiness versus log-transformed MAE of the EOF reconstructed Chla anomaly fields (using the top 20 EOFs) for each gappy EOF approach. MAE is calculated against the true field. Shaded areas show the 25% and 75% quartiles for gappiness intervals by approach. 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).