Filtering remotely sensed chlorophyll concentrations in the Red Sea using a space-time covariance model and a Kalman Filter

Denis Dreano[a], Bani Mallick[b], Ibrahim Hoteit[a,*]

[a] *Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology*
[b] *Department of Statistics, Texas A&M University*

Abstract

A statistical model is proposed to filter satellite-derived chlorophyll concentration from the Red Sea, and to predict future chlorophyll concentrations. The seasonal trend is first estimated after filling missing chlorophyll data using an Empirical Orthogonal Function (EOF)-based algorithm (Data Interpolation EOF). The anomalies are then modeled as a stationary Gaussian process. A method proposed by Gneiting [1] is used to construct positive-definite space-time covariance models for this process. After choosing an appropriate statistical model and identifying its parameters, Kriging is applied in the space-time domain to make a one step-ahead prediction of the anomalies. The latter serves as the prediction model of a reduced-order Kalman filter, which is applied to assimilate and predict future observations of chlorophyll concentrations. The proposed method decreases the root mean square

*Corresponding author
Email address: ibrahim.hoteit@kaust.edu.sa (Ibrahim Hoteit)

(RMS) prediction error by about 11% compared with the seasonal prediction.

1. Introduction

The Red Sea is an elongated basin situated between the Asian and the African shelves, connected to the Mediterranean Sea in the north through the Suez Canal, and to the Gulf of Aden in the south through the Strait of Bab el Mandeb [2]. It is one of the warmest and most saline seas in the world, with a rich ecosystem that has adapted to these extreme conditions [3]. This unique natural resource is however threatened by an abrupt increase of temperature since 1994 [4].

Phytoplankton are small, unicellular, photosynthetic algae. They are the primary producers for marine ecosystems, and at the base of the marine food chain. Phytoplankton concentration is therefore important for fisheries [5]. By fixing atmospheric CO2 and sinking to form sediment at the bottom of the sea, phytoplankton also acts as a biological pump. This phenomenon is crucial to understanding climate change [6]. Thus modeling and predicting changes in chlorphyll concentrations have many purposes. In the Red Sea, phytoplankton is particularly important to the extensive coral reefs along its shores.

The Red Sea is generally deficient in major inorganic nutrients, and its productivity is relatively low [7, 8]. The high productivity observed in the southern Red Sea is attributed to the intrusion of nutrient-rich waters from the Gulf of Aden [7, 8]. Red Sea chlorophyll concentrations follow seasonal

patterns, with a winter bloom following a weak summer productivity. Considerable interannual variability in chlorophyll concentrations has also been observed [9]. However, the Red Sea ecosystem has not yet been fully explored, and very few in- itu measurements have been conducted in its basin [10], increasing the need for remotely sensed data. Satellite observations of chlorophyll concentrations have been shown to be reliable datasets to study the primary productivity of the oceans [11] and they constitute the basis of several studies in the Red Sea [7, 10, 9].

Modeling and predicting changes in phytoplankton concentrations is challenging. It requires the coupling of an ecological model with a hydrological model, which provides the physical forcing that influences the ecology. A broad range of models has been developed by the marine ecosystem research community, from the simple NPZ model with only nutrients, phytoplankton and zooplankton as state variables, to much more complex models. NPZ models have been implemented with various degrees of success on very different regions [12]. However, there are cases when distinguishing between specific groups of phytoplankton can be useful. Examples are the study of export, sinking or climate feedback [12]. The European Regional Seas Ecosystem Model (ERSEM) [13] is an example of a much more sophisticated ecosystem model, initially designed for simulations of the North Sea. It distinguishes functional phytoplankton and zooplankton groups, and models the complete cycling of different nutrient groups and O_2 and CO_2, and includes the effect of higher trophic groups. It has recently been implemented in the

Red Sea [8]. Configuring ERSEM and coupling it with a physical ocean model is, however, delicate and requires considerable efforts and expertise [14]. A more important issue with this class of models is the number of empirical equations that governs its dynamics and the number of involved ecological variables and parameters (over 50 for ERSEM [15]). This makes such models difficult to calibrate and validate, since there are usually not enough observations to constrain the large number of parameters [12].

An alternative approach is to follow a statistical framework to model the space-time evolution of chlorophyll concentrations. Statistical methods have not been widely used in this field, which has previously relied on time-series observations. Artificial neural networks were applied to forecasting algal blooms in freshwater and marine systems [16, 17], and generalized additive models have been use in finding explanatory variables for the chlorophyll concentrations in the Pagasitikos Gulf and the subartic North Atlantic [18, 19].

Geostatistical spatio-temporal models are extensions of the spatial classical geostatistical methods [20]. These methods consider space-time data as the realization of a Gaussian process, from which a mean and a covariance function can be estimated. Although, in most applications, such a stochastic modeling approach has no rigorous scientific basis, geostatistical methods may capture some patterns in the data and avoid the difficulties of dynamical models [21, 20]. These methods are widely employed in meteorology to model the surface temperature over land and oceans [22, 23], in an ecological

context to study moth populations [24], and to characterize soil and pollution [25, 26].

The first proposed spatio-temporal geostatistical model was based on separable covariance functions [20]. Such functions can either be the product, or the sum, of a purely temporal covariance model and a purely spatial covariance model. These covariance models are convenient but have non-physical properties, limiting their use in many situations [20]. As a result, significant research has been recently conducted to construct nonseparable families of covariance functions. In [27, 1], the authors proposed methods to construct families from known temporal and spatial covariance functions. The method proposed in [1] is appealing because of its modularity and interpretability. We adopte it here to model the anomaly fields of chlorophyll concentration in the Red Sea.

Once the mean and covariance functions have been estimated from available satellite data after missing data have been filled with an Empirical Orthogonal Function (EOF)-based method (DINEOF [28, 29]), it is possible to condition on present observations to improve the forecast of future chlorophyll concentration measurements via Kriging [20]. However, this requires the inversion of the spatial observation autocovariance matrix with a 0 temporal. In the context of remotely sensed two-dimensional (2D) fields, the number of observations may be close to the number of variables. For a large area and/or high resolution, this inversion can be computationally prohibitive. Moreover, in the case of large regions of missing observations, as

in the southern Red Sea, the prediction will be biased at these locations [30]. This problem is tackled here by identifying the Kriging operator as an evolution matrix in a state-space context and then using a reduced-order Kalman filter [31] for the filtering and prediction of the observations. This method is computationally efficient; it implicitly uses past observations in the estimation process; and it provides a prediction of the entire state at any point in time. The reduced-order Kalman filter significantly reduces the computational burden of the traditional Kalman filter by projecting the filter covariance matrices on a fixed low-rank basis. Since the anomalies are modeled with a space-time Gaussian process, they are correlated in time. The formulation of the reduced-order Kalman filter is therefore expanded to take into account a colored model noise [32].

The paper is organized as follows. Section 2 describes the satellite data. Section 3 discusses the methodology for the data filling and geostatistical modeling. Section 4 derives the state-space formulation and introduces the space-time Kalman filtering problem and its solution. The experimental setup and numerical results are presented and discussed in section 4. Concluding remarks are provided in section 5.

2. Data and preprocessing

Satellite data provide chlorophyll (CHL) concentrations with a spatial and temporal resolution not achievable with in situ observations, making them particularly relevant to the Red Sea, where very few in situ data collection

are conducted.

Level-3 mapped data from the NASA SeaWiFS (Sea-Viewing Wide Field-of-View Sensor) satellite sensor are used in this study. The dataset is publicly available at `http://oceancolor.gsfc.nasa.gov`. In this study, we use the 9km resolution mapped weekly averages from January 1998 to December 2007 (460 time steps). At each time step, a 133×188 pixel map is available for a domain extending from longitudes between 33°E and 44°E and latitudes between 12°N and 28°N, of which 5635 pixels correspond to actual Red Sea surface (see Figure 1(a)). A log-transformation was applied in order to obtain an approximately Gaussian distribution [33]. Pixels with too few observations were discarded, and a control quality check was applied to remove outliers [34].

Remotely sensed CHL may have missing data because of cloud coverage. The cloud variability in the Red Sea follows a seasonal cycle. Figure 1(c) shows that the cloud coverage is particularly pronounced during summers because of the monsoon and it is sparse during winters. The cloud coverage is, however, not homogenous over the Red Sea. It is much more pronounced in the south (figure 1(b)). In this region, almost no data are available during summers.

The dataset is separated into a training set, composed of the first seven years of data, and a validation set, containing the remaining three years. The computations for the data filling and for the covariance model selection and estimation are based only on the training dataset. The testing dataset

is used to validate the model predictions outside the training period.

3. Data filling and modeling

3.1. Data filling

The DINEOF (Data Interpolating Empirical Orthogonal Function) is an EOF-based, recursive method for the reconstruction of data matrices with missing values [28, 29]. It estimates the values of the missing data by successive singular values decompositions (SVD) of a given data matrix and truncated reconstructions. The advantage of this method is that it does not require any a priori information about the data. It has been successfully used for reconstruction of incomplete chlorophyll datasets in different regions of the ocean [35, 36, 37].

Let $\mathbf{X}$ be an $m \times k$ centered data matrix with missing values initially filled with 0s. Then, until the missing values have converged, the following steps are repeated. An SVD is first applied to the data matrix: $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, with $\mathbf{U}$ an $m \times m$ unitary matrix, $\mathbf{\Sigma}$ an $m \times k$ diagonal matrix and $\mathbf{V}$ a $k \times k$ unitary matrix. The missing values are then replaced by the truncated reconstruction order n of the data matrix: $\{\mathbf{X}\}_{i,j} = \{\mathbf{U}^{(n)}\mathbf{\Sigma}^{(n)}(\mathbf{V}^{(n)})^T)\}_{i,j}$, for i, j indices of the missing values, with $\mathbf{U}^{(n)}$ the $m \times n$ matrix composed of the n first columns of $\mathbf{U}$, $\mathbf{V}^{(n)}$ the $k \times n$ matrix composed of the n first columns of $\mathbf{V}$, and $\mathbf{\Sigma}^{(n)}$ the $n \times n$ diagonal matrix with the n largest eigenvalues on its diagonal. It is assumed that the eigenvalues and eigenvector are sorted by decreasing order of eigenvalues. In [29], the authors introduced the filtering

of the temporal covariance matrix as a way of reducing spurious oscillations that may appear when the data are sparsely sampled in time. This filtering is controlled by the parameter of the Laplacian filter and the number of times the filter is applied.

The values of the DINEOF parameters are determined following the method outlined in [29]. The smoothing parameter of the Laplacian filter is set to 0.005. The number of modes in the truncation and the number of times the filter is applied are chosen following a cross-validation technique. A random subset of observed values is taken from X and assumed to be missing before the DINEOF is applied. The algorithm is then run with different numbers of iterations (1, 3, 10, 30, 100) and orders of truncation (from 2 to 50). The set of parameters minimizing the RMS error over the cross-validation data is chosen as the best number of iterations and order of truncation. The approach of [38] is followed to select a cross-validation dataset. Instead of selecting it by sampling the dataset point by point, contiguous regions are set aside. These regions correspond to regions of missing data from the original dataset and are selected randomly until 3% of the data have been extracted.

3.2. Geostatiscal modeling

The chlorophyll concentration data are modeled as a space-time random Gaussian process:

$$Z(\mathbf{s};t), (\mathbf{s};t) \in \mathbb{R}^2 \times \mathbb{R}. \tag{1}$$

Such a process is entirely characterized by the mean function $\mu(\mathbf{s};t) = E(Z(\mathbf{s};t))$, and the covariance function $K(\mathbf{s},\mathbf{r};t,q) = \text{cov}(Z(\mathbf{s};t), Z(\mathbf{r};q))$. The process is further assumed to be stationary, such that the covariance function can be written as $K(\mathbf{s},\mathbf{r};t,q) = \text{cov}(Z(\mathbf{s};t), Z(\mathbf{r};q))$.

Covariance functions should be positive-definite, meaning that for any ensemble of space-time coordinates, $\{(\mathbf{s}_i;t_i)\}_{i=1,...,k}$, and real coefficients, $d_1, ..., d_k$, the following condition should hold:

$$\sum_{i=1}^{k}\sum_{j=1}^{k} d_i d_j C(\mathbf{s}_i - \mathbf{s}_j; t_i - t_j) \geq 0. \tag{2}$$

One way to enforce positive-definiteness in practice is to assume that the covariance function belongs to a parametric family of covariance functions, denoted by $C(\mathbf{h};u|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the vector of parameters to be estimated [39].

3.3. Mean function estimation and anomalies

The data exhibit a distinct seasonal signal accounting for roughly 50% of the total variability as can be seen in Figure 2. It is therefore reasonable to model it in the mean function $\mu(\mathbf{s};t)$. Once the data are filled using the DINEOF algorithm, the seasonal signal is estimated for each week with the weekly average computed from the training data.

The anomalies are then computed by substracting the weekly averages from the data. The CHL anomalies of the first 8 weeks of data are displayed in Figure 3. There are large regions of similar colors indicating spatial corre-

lations. Moreover, two maps adjacent in time display similar patterns, suggesting that the dataset is also correlated in time and justifying the use of space-time covariance functions to model CHL anomalies. From the anomaly time-series at three locations, plotted in Figure 4, one may conclude that the data have been successfully detrended. Since no specific pattern appears, the assumption of stationarity for the covariance model seems to be appropriate.

3.4. Construction and fitting of the covariance model

For space-time processes, families of covariance functions are typically built by combining known spatial and temporal covariance models. Separable models are the simplest example of this approach, taking the form $C(\mathbf{h}; u|\boldsymbol{\theta}) = C^1(\mathbf{h}|\boldsymbol{\theta}_1)C^2(t|\boldsymbol{\theta}_2)$, where C^1 and C^2 are respectively pure spatial and temporal covariance models parametrized by $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. However, realizations from such families of covariance models were shown to exhibit non-physical behaviors [1, 27]. Research has been conducted to develop methods for constructing non-separable covariance functions to overcome these limitations [27, 1].

In this work, families of space-time covariance functions are constructed following an approach proposed by Gneiting [1]. Given any completely monotonous function, $\varphi(t), t \geq 0$, and any positive function with a completely monotonous derivative, $\psi(t), t \geq 0$, the following space-time covari-

ance function is valid (i.e., positive-definite):

$$C(\mathbf{h};u) = \frac{\sigma^2}{\psi(|u|^2)}\varphi\left(\frac{||\mathbf{h}||^2}{\psi(|u|^2)}\right), (\mathbf{h},u) \in \mathbb{R}^2 \times \mathbb{R}. \tag{3}$$

In general, φ is defined as a temporal covariance function, and ψ as a spatial covariance model, e.g., the exponential model or the Matérn model. Tables 1 and 2 present some examples of these functions that were introduced in [1]. To obtain more specialized families, a nugget term can be added.

The typical approach for estimating the parameters of the covariance function is by fitting it to the empirical space-time covariance matrix [39]. The empirical covariance function is first estimated using the formula:

$$2\hat{C}(\mathbf{h}(l);u) = \frac{1}{|N(\mathbf{h}(l);u)|}\sum_{(i,j,t,t')\in N(\mathbf{h}(l);u)} a(\mathbf{s}_i;t)a(\mathbf{s}_j;t'), \tag{4}$$

with $N(\mathbf{h}(l);u) = \{(i,j,t,t') \text{ such that } : |t-t'| = u, \mathbf{s}_i - \mathbf{s}_j \in \text{Bin}(\mathbf{h}(l))\}$, where $\mathbb{R}^2$ has been divided into a finite number of bins, $\text{Bin}(\mathbf{h}(l))$, each of which has a representative lag, $\mathbf{h}(l)$. Here, bins are considered of the form $\mathbf{k} \in \text{Bin}(\mathbf{h}(l))$ if $(l-1) * \delta h \leq ||\mathbf{k}|| \leq l\delta h, l = 0, ..., L$, with δ a fixed interval size and L the number of retained intervals. The parameters are then estimated by fitting the covariance model using the Weighted Least Squares (WLS) approach. This method is popular due to its simplicity and proven efficiency [40]. In practice, the estimator of $\boldsymbol{\theta}$ minimizes the weighted sum of the squares of

errors:

$$S(\boldsymbol{\theta}) = \sum_{l} \sum_{u=0}^{3} w_{l,u} \left(\hat{C}(\mathbf{h}(l); u) - C(\mathbf{h}(l); u|\boldsymbol{\theta}) \right)^2, \tag{5}$$

with weights $w_{l,u} = \frac{|N(\mathbf{h}(l);u)|}{(1-C(\mathbf{h}(l);u|\boldsymbol{\theta}))^2}$. ϕ and ψ are chosen from the functions presented in Tables 1 and 2, by separately fitting the covariance function in time and space for all possible choices of ϕ and ψ. The choice of the functions is then made based on the quality of the fit.

4. State-space formulation and filtering

So far, the CHL data are modeled as $Z(\mathbf{s};t) = \mu(\mathbf{s};t) + a(\mathbf{s};t)$, where μ is a deterministic function representing the seasonality and a is a zero-mean space-time Gaussian process representing the weekly anomalies, with the covariance function estimated as described in the previous section. However, a state-space model needs to be formulated to use the Kalman filter. This is done by identifying a CHL underlying process, $Y(\mathbf{s};t)$, distinct from the data process $Z(\mathbf{s};t)$. These processes are then discretized in space and time to obtain the state-space formulation, on which a reduced-order variant of the Kalman filter is applied.

227 *4.1. State-space modeling*

To derive the state-space model, we start from the following model:

$$Y(\mathbf{s};t) = \mu(\mathbf{s};t) + a(\mathbf{s};t), \text{ where } \mathbf{s}, t \text{ are observable,} \tag{6}$$

$$Z(\mathbf{s};t) = Y(\mathbf{s};t) + \varepsilon(\mathbf{s};t), \text{ where } \mathbf{s}, t \text{ are observed.} \tag{7}$$

228 In this formulation, $\varepsilon(\mathbf{s};t)$ is a white noise process representing the mea-
229 surement errors, $a(\mathbf{s};t)$ is the anomaly process with the covariance matrix
230 estimated as described in the previous section, $\mu(\mathbf{s};t)$ is the seasonal mean
231 function, $Y(\mathbf{s};t)$ is the underlying CHL log-concentration process, and $Z(\mathbf{s};t)$
232 the data process.

To build a state-space model for filtering and forecasting the spatial variability of CHL in time, once the covariance function is estimated, we resort to Kriging. In classical geostatistics, Kriging interpolates available observations to provide the best linear unbiased estimate of a spatial process at unobserved locations [39]. This technique easily generalizes to space-time Gaussian processes, particularly for forecasting. The equations are obtained by conditioning the anomalies at time t by the anomalies at time $t-1$. Using the vectorial notations $\mathbf{a}_t = \{a(\mathbf{s}_{i,t})\}_{i=1,\ldots,k}$, where $\mathbf{s}_{i,t}$ is the spatial location of the i-th observation at time t, $[\mathbf{a}_t^T, \mathbf{a}_{t-1}^T]^T$ is a Gaussian vector. Therefore,

by conditioning

$$\mathbf{a}_t|\mathbf{a}_{t-1} \sim N(\mathbf{M}\mathbf{a}_{t-1}, \mathbf{Q}), \tag{8}$$

$$\tag{9}$$

where

$$\mathbf{M} = \mathbf{C}_1\mathbf{C}_0^{-1}, \tag{10}$$

$$\mathbf{Q} = \mathbf{C}_0 - \mathbf{C}_1\mathbf{C}_0^{-1}\mathbf{C}_1, \tag{11}$$

with $\mathbf{C}_0 = E[\mathbf{a}_t\mathbf{a}_t^T] = \{C(\mathbf{s}_{i,t}-\mathbf{s}_{j,t}), 0\}_{i,j=1,\ldots,k}$ and $\mathbf{C}_1 = E[\mathbf{a}_t\mathbf{a}_{t-1}^T] = \{C(\mathbf{s}_{i,t}-\mathbf{s}_{j,t-1}), 1\}_{i,j=1,\ldots,k}$, the following recursive state-space model can be derived:

$$\mathbf{y}_t = \boldsymbol{\mu}_t + \mathbf{M}(\mathbf{y}_{t-1} - \boldsymbol{\mu}_{t-1}) + \boldsymbol{\eta}_t, \tag{12}$$

$$\mathbf{z}_t = \mathbf{H}_t\mathbf{y}_t + \boldsymbol{\varepsilon}_t, \tag{13}$$

and $\mathbf{y}_t = \mathbf{a}_t + \boldsymbol{\mu}_t$ is the CHL process discretized in space, $\boldsymbol{\mu}_t$ is the vector seasonal component, H_t is the observation operator that returns the CHL concentration at the observed location, $\boldsymbol{\eta}_t$ represents the model error, and $\boldsymbol{\varepsilon}_t$ represents the measurement error. $\mathbf{y}_t$ is a vector of fixed size that represents the whole model domain (Red Sea), whereas $\mathbf{z}_t$ has a variable size equal to the number of available observations at time t.

The model can be reformulated such that only the anomalies are filtered

and a_t is the state vector:

$$\mathbf{a}_t = \mathbf{M}\mathbf{a}_{t-1} + \boldsymbol{\eta}_t, \tag{14}$$

$$\mathbf{z}_t = \mathbf{y}_t - \mathbf{H}_t\mathbf{s}_t = \mathbf{H}_t\mathbf{a}_t + \boldsymbol{\varepsilon}_t. \tag{15}$$

This system is equivalent to the preceding one, but is more practical to implement. Given $\boldsymbol{\eta}_t \sim N(0, \mathbf{Q})$ and assuming $\boldsymbol{\varepsilon}_t \sim N(0, \sigma^2_{\text{obs}}\mathbf{I})$, idependent and identically distributed (i.i.d.), with σ^2_{obs} a fixed constant that is tuned by minimizing the RMS prediction error.

4.2. Low-rank Kalman filter

Two issues need to be tackled before using the time evolution model for predicting the anomalies. First, the consequent amount of missing data in the CHL satellite observations makes it difficult to obtain frequent initial CHL concentrations to integrate the model forward in time for forecasting, and second, the model is linear with eigenvalues smaller than one[1] in absolute values, making it inappropriate for long-term predictions. The latter means that CHL model forecasts would decrease exponentially in time, quickly becoming close to 0. The Kalman filter solves both problems by recursively assimilating the observations and providing an optimal estimate, in the mean-square sense, of the anomalies to start a new forecast cycle. The filter operates in two steps to compute the best linear estimate of the state

[1]This was verified numerically in the present work. We are not aware of a general result.

of a linear dynamical model given past observations [41].

- *Forecast step:* Starting from the best available estimate of the state, a^a_{t-1}, and the associated error covariance matrix, P^a_{t-1}, at a given time, $t-1$, the forecast state, a^f_t, and its error covariance matrix, P^f_t, are obtained by integrating the model forward to the time of the next available observation:

$$\mathbf{a}^f_t = \mathbf{M}\mathbf{a}^a_{t-1}, \tag{16}$$

$$\mathbf{P}^f_t = \mathbf{M}\mathbf{P}^a_t\mathbf{M}^T + \mathbf{Q}. \tag{17}$$

- *Update step:* Once a new observation, $\mathbf{z}_t$, is available, the forecast state, $\mathbf{a}^f_t$, and its error covariance, $\mathbf{P}^f_t$, are updated to their analysis counterparts, $\mathbf{a}^a_t$, and, $\mathbf{P}^a_t$, as:

$$\mathbf{a}^a_t = \mathbf{a}^f_t + \mathbf{K}_t(\mathbf{z}_t - \mathbf{H}_t\mathbf{a}^f_t), \tag{18}$$

$$\mathbf{P}^a_t = (\mathbf{I} - \mathbf{K}_t\mathbf{H}_t)\mathbf{P}^f_t, \tag{19}$$

$$\mathbf{K}_t = \mathbf{P}^f_t\mathbf{H}^T_t(\mathbf{H}_t\mathbf{P}^f_t\mathbf{H}^T_t + \sigma^2_{\text{obs}}\mathbf{I})^{-1} \tag{20}$$

where K_t is known as the Kalman gain.

The inversion in the computation of the Kalman gain is computationally quite demanding because the number of observations is as large as the size of the state. To speed up this computation, the reduced-order Kalman filter [31]

is used. This filter approximates the covariance matrices as $\mathbf{P}_t^f = \mathbf{L}\mathbf{U}_t^f\mathbf{L}^T$ and $\mathbf{P}_t^a = \mathbf{L}\mathbf{U}_t^a\mathbf{L}^T$, where $\mathbf{L}$ is a $n \times r$ matrix whose columns are the r leading EOFs (computed here with the DINEOF algorithm), and $\mathbf{U}_t^f$ and $\mathbf{U}_t^a$ are $r \times r$ matrices. The inversion is then applied on the $r \times r$ matrices $\mathbf{U}_t^f$ and $\mathbf{U}_t^a$, which considerably reduces the computational burden since $r \ll n$ in practice [31]. Defining

$$\mathbf{V} = \mathbf{L}^T\mathbf{Q}\mathbf{L}, \tag{21}$$

$$\mathbf{W} = \mathbf{L}^T\mathbf{M}\mathbf{L}, \tag{22}$$

which respectively represent the projection of the model dynamics and the model error on the leading EOFs, the reduced-order Kalman filter equations for the forecast and update of the covariance matrix can be simplified by only updating $\mathbf{U}_t^f$ and $\mathbf{U}_t^a$ as follows:

$$\mathbf{U}_t^f = \mathbf{W}\mathbf{U}_{t-1}^a\mathbf{W}^T + \mathbf{V}, \tag{23}$$

$$(\mathbf{U}_t^a)^{-1} = (\mathbf{U}_t^f)^{-1} + \mathbf{L}^T\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\mathbf{L}. \tag{24}$$

The Kalman gain can be then computed as

$$\mathbf{K}_t = \sigma_{\text{obs}}^{-2}\mathbf{L}\mathbf{U}_t^a\mathbf{L}^T\mathbf{H}^T. \tag{25}$$

Otherwise, the forecast and update steps are identical to those of the Kalman filter.

4.3. Low-rank Kalman filter with colored noise

One issue is that the model noise is correlated in time, whereas the Kalman filter assumes it to be white. Indeed, $\boldsymbol{\eta}_t = \mathbf{a}_t - \mathbf{M}\mathbf{a}_{t-1}$ is a space-time Gaussian process. One can show that $[\boldsymbol{\eta}_t^T, \boldsymbol{\eta}_{t-1}^T]^T$ is a Gaussian vector of mean 0 and covariance matrix $\begin{bmatrix} \mathbf{A}_0 & \mathbf{A}_1 \\ \mathbf{A}_1^T & \mathbf{A}_0 \end{bmatrix}$, with $\mathbf{A}_0 = \mathbf{C}_0 - \mathbf{C}_1\mathbf{C}_0^{-1}\mathbf{C}_1^T$, $\mathbf{A}_1 = \mathbf{C}_1\mathbf{C}_0^{-1}\mathbf{C}_1\mathbf{C}_0^{-1}\mathbf{C}_1 - \mathbf{C}_2\mathbf{C}_0^{-1}\mathbf{C}_1$. We can then predict $\boldsymbol{\eta}_t$ by conditioning on the previous equation and derive the model $\boldsymbol{\eta}_t = \mathbf{N}\boldsymbol{\eta}_{t-1} + \boldsymbol{\xi}_t$, with $\mathbf{N} = \mathbf{A}_1\mathbf{A}_0^{-1}$ and $\boldsymbol{\xi}_t \sim \mathcal{N}(0, \mathbf{A}_0 - \mathbf{A}_1\mathbf{A}_0^{-1}\mathbf{A}_1)$.

To take into account this correlation in the Kalman filter, the state-space system is enlarged by including $\boldsymbol{\eta}_t$ in the state vector [32], so the new state-space system becomes:

$$\hat{\mathbf{a}}_t = \boldsymbol{\Gamma}\hat{\mathbf{a}}_{t-1} + \hat{\boldsymbol{\eta}}_t, \tag{26}$$

$$\mathbf{y}_t = \hat{\mathbf{H}}\hat{\mathbf{a}}_{t-1} + \boldsymbol{\varepsilon}_t, \tag{27}$$

with $\hat{\mathbf{a}}_t = [\mathbf{a}_t^T, \boldsymbol{\eta}_t^T]^T$, $\boldsymbol{\Gamma} = \begin{bmatrix} \mathbf{M} & \mathbf{I} \\ \mathbf{0} & \mathbf{N} \end{bmatrix}$, $\hat{\boldsymbol{\eta}}_t = [\mathbf{0}, \boldsymbol{\xi}_t^T]^T$ and $\hat{\mathbf{H}}_t = [\mathbf{H}_t, \mathbf{0}]$, on which the Kalman filter is applied. Of course, $\boldsymbol{\xi}_t$ is again colored, but the procedure can be iterated until no significant correlation is left to exploit in the noise.

5. Results

Figure 5 summarizes the workflow of the experimental setup. The chlorophyll concentration is the sum of a seasonal component and an anomaly. The former is estimated using data filled by the DINEOF algorithm over a learning period (first seven years of CHL data). The anomalies are assumed to be a stationary Gaussian process whose space-time covariance matrix has been estimated. A reduced-order Kalman filter is then applied to perform one-step ahead predictions. These are compared with the observations over the remaining three years of validation data to validate the system's performance and results. The results of the DINEOF data filling algorithm are presented before a covariance model for the data is chosen and fitted. The results of the filtering are finally examined and analyzed.

5.1. DINEOF Analysis

Figure 6 plots the (RMS) error for the cross-validation dataset and for different values of the number of smoothing iterations and EOFs. A lower RMS error is achieved with 24 EOFs and 30 iterations.

Figures 7(a) and 7(c) show the first spatial and associated temporal modes resulting from the DINEOF analysis. Figure 7(c) exhibits a regular peak during winters and a secondary peak of varying size during summers. These can be associated with the winter bloom and the secondary summer bloom, both described in [9]. Figure 7(a) shows that the bloom is relatively homogeneous over the Red Sea, except in the southwest corner where the variation

is strongly pronounced.

Figures 7(b) and 7(d) plot the second spatial and temporal modes respectively. The spatial mode shows a north-south contrast. The time series displays a large peak around summer 2000, which corresponds to a large positive anomaly taking place during this period (Figure 7(e)). Except for the first two modes, all other modes tend to explain a local feature of the data. This makes the interpretation of the EOF analysis difficult and the convergence of the spectrum very slow as shown in Figure 7(f).

5.2. Covariance model estimation

Figure 8a plots the empirical space covariance function $\hat{C}(||\mathbf{h}||, 0)$. Among the functions in Table 1, the ones that best fit this curve are selected. A WLS minimization is applied to fit the parameters c, γ and ν, as well as a the noise variance $\sigma^2_{\text{spatial}}$ while ignoring the observations for $||\mathbf{h}|| = 0$. This allows a nugget effect to be taken account. The results are superimposed on Figure 8a. The Matérn model clearly fits poorly. By examining the residuals, we can obviously eliminate φ_4. Finally, between the two remaining candidates, both of which appear to be plausible, φ_1 is chosen since it involves fewer degrees of freedom.

Figure 8b plots the empirical time covariance function $\hat{C}(0; u)$. ψ_2 defines a time covariance that decreases very slowly. The covariance with ψ_3 converges to a constant as the time-lag goes to infinity, which is not realistic in the case of anomalies. The first-time covariance model is filled with a nugget

effect for $u = 0$, using WLS, and the result is displayed in Figure 8b, where a nugget effect can be clearly identified.

To build the space-time covariance function, we use the reparametrization in [1] (see example 1), along with the building blocks in time and space that have been derived in Section 3. A time-only nugget effect is also added, leading to the following covariance model:

$$C(\mathbf{h}; u) = \begin{cases} \sigma^2 \exp(-c||\mathbf{h}||^{2\gamma}) & \text{, if } u = 0, \\ \frac{\sigma^2 \tau}{a|u|^{2\alpha}+1} \exp(-\frac{c||\mathbf{h}||^{2\gamma}}{(a|u|^{2\alpha}+1)^{\beta\gamma}}) & \text{, otherwise,} \end{cases} \tag{28}$$

with $0 < \tau \leq 1$.

The initial values for the WLS fitting are given by the results of the preceding purely spatial and purely temporal regressions. The results are shown in Table 3. A contour plot of the resulting covariance function is shown in Figure 9. Our fitted correlation model has level curves very close to those of the empirical covariance model.

5.3. Filtering

σ^2_{obs} is first determined empirically by trial and error, chosen as the value that leads to the minimal averaged RMS error (RMSE). $\sigma^2_{\text{obs}} = 1$ is found to be a reasonable choice.

The anomaly model with a space-time covariance function helps to decrease the RMS error by nearly 11% over the test period. As Figure 10a indicates, the improvement is most noticible during periods with large anomalies.

The boxplot in Figure 10b shows that the variability of errors is reduced with the proposed filtering approach. Since the modeling is purely statistical and is not based on physical quantities, it was not able to anticipate the start of a bloom. However, it successfully extrapolates the current estimate of the anomaly in space and time and improves the prediction compared to the seasonal component.

In Figures 11 (a to c), the predictions of the model are compared with the observations and the seasonal signal for a fall week in 2006. The model captures some differences with the seasonal regime, such as a larger northern region with a low CHL concentration extending south below 22°N, and a more intense bloom in the south. The usual seasonal dynamic is altered with an extension of the stratified regime in the north and a larger intrusion of the nutrient rich waters of the Gulf of Aden [9].

Figure 11 (d to f) shows a similar comparaison for a winter week in 2006. The model captures a weak winter bloom in the northern half of the Red Sea. A similar pattern has been described in [7] for winter 1999, and it seems to be a common feature of El Nina years. The model also successfully captures a high CHL concentration in the south, and a lower than usual CHL concentration in the center [9].

Figure 12 plots the error variance as predicted by the filter and the actual prediction RMSE, computed as the difference between the model forecast and data, averaged over the three year training period at every point of the grid. The predicted RMSE corresponds to the diagonals of the prediction

covariance matrix as estimated by the reduced-order Kalman filter. The prediction RMSE is computed from the error between the model prediction and the observation. We can see that both values are close in the northern half of the Red Sea, but the RMSE is much larger in the south. This is caused by the lack of data and the fact that the dynamics in the South is different from that in the north, making the process nonstationary. Indeed, as shown in Figure 3, the anomalies in the south clearly exhibit smaller spatial and temporal correlation length scales compared with those in the north.

One way to evaluate the filter's behavior is to examine the distribution of the innovations and the increments [42]. The innovation corresponds to the difference: $\mathbf{y}_t - \mathbf{H}_t\mathbf{a}_t^f$, whereas the increment corresponds to the difference: $\mathbf{H}_t(\mathbf{a}_t^a - \mathbf{a}_t^f)$. Figure 13(c) shows that the innovation seems to be approximately normally distributed, as expected for a properly tuned Kalman filter. Figure 13(b) shows that the averaged innovation size decreases to a value close to zero as the filter assimilates the data over time, which also suggests that the filter is properly working. However, Figure 13(a) indicates that the increments of the filter tend to be positively biased in some regions. This indicates that the model tends to underestimate the amount of CHL, which may be associated with the statistical model's inability to forecast CHL blooms.

Figures 14 (a), (b) and (c) plot the spatially averaged corelations between the observations and the seasonal predictions, the Kalman filter forecasts and the analyses, respectively. Compared with the seasonal correlations, we can see that the model improves the prediction skill over the entire Red Sea,

particularly in its central part, with correlations ranging between 0.28 and 0.92. The filter further improves the prediction-data correlation, over the entire domain with correlations up to 0.96.

6. Discussion

Here, we considered the filtering problem of satellite-derived chlorophyll (CHL) concentration in the Red Sea using a data-driven approach in which the CHL spatio-temporal evolution is modeled as a space-time Gaussian process. The DINEOF data-filling algorithm [38] was applied to compute an estimate of the seasonal signal in the data, which was used as the mean function of the process. To model the residual anomalies, the method proposed by Gneiting in [1] was applied to construct an appropriate family of covariance functions.

From the anomalies, a family of covariance functions is constructed and then fitted to the data. Based on a space-time Kriging formulation, a linear model was derived to capture changes in the chlorophyll concentration. A reduced-order variant of the Kalman filter was then applied to forecast and filter the CHL concentration. The results of our experiments suggest that the proposed system works reasonably well, reducing the RMS error in CHL concentration prediction by about 11% as compared with the seasonal mean.

The proposed approach is not difficult to apply, but requires some coding

efforts. For the DINEOF, a standalone package is freely available online[2]. To the authors' knowledge, there is currently no R package or library for fitting custom space-time covariance models. The reduced-order Kalman filter is straightforward to implement.

The proposed method requires the estimation of a very few parameters, which may prevent overfitting. Another advantage is its stability, as the anomaly prediction cannot grow over time. One problem with Kriging is that in the case of missing data over large areas, its prediction will be the mean function [30]. In this study, this problem is alleviated using the Kalman filter which always provides a prediction over the whole domain.

Forecasting and filtering CHL concentration in the Red Sea using remotely sensed data is challenging. Because of cloud coverage over the southern Red Sea during the summer, large areas remain unobserved. Moreover, the Red Sea is a heterogenous environment with different ecological and physical dynamics from north to south. We can therefore expect the anomalies to be non-stationary. Fortunately, Kriging is robust to non-stationarity [30]. A problem with this correlation-based modeling is its inability to forecast CHL blooms. Our linear model tends to underestimate the amount of CHL concentration and to miss blooms in some regions, before the Kalman filter corrects this error when new observations are available.

The proposed method can be further generalized by considering more

[2]http://modb.oce.ulg.ac.be/mediawiki/index.php/DINEOF

sophisticated mean functions. For example, one can use a linear model with additional covariates, such as sea surface temperature, sea surface height, thermocline depth or wind speed. Another way to improve the model is to use non-stationary covariance functions, or to use non-Gaussian models to predict blooms. These tasks will be considered in future studies.

Acknowledgment

The research reported in this publication was supported by King Abdullah University of Science and Technology (KAUST).

7. Bibliography

[1] T. Gneiting, Nonseparable, Stationary Covariance Functions for Space-Time Data, Journal of the American Statistical Association 97 (458) (2002) 590–600. doi:10.1198/016214502760047113.

[2] F. Yao, I. Hoteit, L. J. Pratt, A. S. Bower, P. Zhai, A. Köhl, G. Gopalakrishnan, Seasonal overturning circulation in the Red Sea: part 1. Model validation and summer circulation, Journal of Geophysical Research: Oceans (2014) 2238–2262doi:10.1002/2013JC009331.Key.

[3] S. Heileman, N. Mistafa, III-6 Red Sea: LME# 33, lme.noaa.gov. URL `http://www.lme.noaa.gov/lmeweb/LME_Report/lme_33.pdf`

[4] D. E. Raitsos, I. Hoteit, P. K. Prihartato, T. Chronis, G. Triantafyllou,

Y. Abualnaja, Abrupt warming of the Red Sea, Geophysical Research Letters 38. doi:10.1029/2011GL047984.

[5] A. Lo-Yat, S. D. Simpson, M. Meekan, D. Lecchini, E. Martinez, R. Galzin, Extreme climatic events reduce ocean productivity and larval supply in a tropical reef ecosystem, Global Change Biology 17 (2011) 1695–1702. doi:10.1111/j.1365-2486.2010.02355.x.
URL http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2486.2010.02355.x/full

[6] K. H. Mann, J. R. N. Lazier, Dynamics of Marine Ecosystems, 2006.

[7] J. Acker, G. Leptoukh, S. Shen, T. Zhu, S. Kempler, Remotely-sensed chlorophyll a observations of the northern Red Sea indicate seasonal variability and influence of coastal reefs, Journal of Marine Systems 69 (2008) 191–204. doi:10.1016/j.jmarsys.2005.12.006.
URL http://linkinghub.elsevier.com/retrieve/pii/S0924796307000589

[8] G. Triantafyllou, F. Yao, G. Petihakis, K. Tsiaras, D. E. Raitsos, I. Hoteit, Exploring the Red Sea seasonal ecosystem functioning using a three-dimensional biophysical model, Journal of Geophysical Research: Oceans (2014) 1791–1811doi:10.1002/2013JC009641.Received.

[9] D. E. Raitsos, Y. Pradhan, R. J. W. Brewin, G. Stenchikov, I. Hoteit, Remote Sensing the Phytoplankton Seasonal Succession of the Red Sea, PLoS ONE 8 (6). doi:10.1371/journal.pone.0064909.

[10] R. J. W. Brewin, D. E. Raitsos, Y. Pradhan, I. Hoteit, Comparison of chlorophyll in the Red Sea derived from MODIS-Aqua and in vivo fluorescence, Remote Sensing of Environment 136 (2013) 218–224. doi:10.1016/j.rse.2013.04.018.
URL http://dx.doi.org/10.1016/j.rse.2013.04.018

[11] C. R. McClain, A Decade of Satellite Ocean Color Observations, Annual Review of Marine Science (2009) 19–42doi:10.1146/annurev.marine.010908.163650.

[12] T. R. Anderson, Plankton functional type modelling: running before we can walk?, Journal of Plankton Research 27 (11). doi:10.1093/plankt/fbi076.
URL http://www.plankt.oxfordjournals.org/cgi/doi/10.1093/plankt/fbi076

[13] J. Baretta, W. Ebenhöh, P. Ruardij, The European regional seas ecosystem model, a complex marine ecosystem model, Netherlands Journal of Sea Research 33 (3/4) (1995) 233–246. doi:10.1016/0077-7579(95)90047-0.
URL http://linkinghub.elsevier.com/retrieve/pii/0077757995900470

[14] G. Petihakis, G. Triantafyllou, I. J. Allen, I. Hoteit, C. Dounas, Modelling the spatial and temporal variability of the Cretan Sea ecosystem, Journal of Marine Systems 36 (2002) 173–196.

[15] J. C. Blackford, P. J. Radford, A Structure and methodology for ma-

rine ecosystem modelling, Netherlands Journal of Sea Research 33 (3/4) (1995) 247–260.

[16] J. H. W. Lee, Y. Huang, M. Dickman, A. W. Jayawardena, Neural network modelling of coastal algal blooms 159 (2003) 179–201.

[17] F. Recknagel, Artifical neural network approach for modelling and prediction of algal blooms, Ecological Modelling 96 (1997) 11–28.

[18] D. Raitsos, G. Korres, G. Triantafyllou, G. Petihakis, M. Pantazi, K. Tsiaras, A. Pollani, Assessing chlorophyll variability in relation to the environmental regime in Pagasitikos Gulf, Greece, Journal of Marine Systems 94 (2012) S16–S22. doi:10.1016/j.jmarsys.2011.11.003. URL http://linkinghub.elsevier.com/retrieve/pii/S0924796311002703

[19] D. E. Raitsos, S. J. Lavender, Y. Pradhan, T. Tyrrell, P. C. Reid, M. Edwards, Coccolithophore bloom size variation in response to the regional environment of the subarctic North Atlantic, Limnology and Oceanography 51 (5) (2006) 2122–2130. doi:10.4319/lo.2006.51.5.2122. URL http://www.aslo.org/lo/toc/vol_51/issue_5/2122.html

[20] P. C. Kyriakidis, A. G. Journel, Geostatistical Space-Time Models: A Review, Mathematical Geology 31 (6) (1999) 651–684.

[21] N. Cressie, C. K. Wikle, Statistics for Spatio-Temporal Data, 2011.

[22] M. S. Handcock, J. R. Wallis, An Approach to Statistical Spatial-

Temporal Modeling of Meteorological Fields, Journal of the American Statistical Association 89 (426) (1994) 368–378.

[23] G. R. North, J. Wang, M. G. Genton, Correlation Models for Temperature Fields, Journal of Climate 24 (2011) 5850–5862. doi:10.1175/2011JCLI4199.1.

[24] M. E. Hohn, A. M. Liebhold, L. S. Gribko, Geostatistical Model for Forecasting Spatial Dynamics of Defoliation Caused by the Gypsy Moth (Lepidoptera: Lymantriidae), Environmental Entomology 22 (5) (1993) 1066–1075.

[25] R. Lark, Towards soil geostatistics, Spatial Statistics 1 (2012) 92–99.

[26] A. Keaney, J. McKinley, C. Graham, M. Robinson, A. Ruffell, Spatial statistics to estimate peat thickness using airborne radiometric data, Spatial Statistics 5 (2013) 3–24.

[27] N. Cressie, H.-C. Huang, Classes of Nonseparable, Spatio-Temporal Stationary Covariance Functions, Journal of the American Statistical Association 94 (448) (1999) 1330–1340.

[28] J.-M. Beckers, M. Rixen, EOF Calculations and Data Filling from Incomplete Oceanographic Datasets, Journal of Atmospheric and Oceanic Technology 20 (2003) 1839–1856.

[29] A. Alvera-Azcárate, A. Barth, D. Sirjacobs, J.-M. Beckers, Enhancing

temporal correlations in EOF expansions for the reconstruction of missing data using DINEOF, Ocean Science 5 (2009) 475–485.

[30] P. Monestiez, L. Dubroca, E. Bonnin, J.-P. Durbec, C. Guinet, Geostatistical modelling of spatial distribution of Balaenoptera physalus in the Northwestern Mediterranean Sea from sparse count data and heterogeneous observation efforts, Ecological Modelling 193 (3-4) (2006) 615–628. doi:10.1016/j.ecolmodel.2005.08.042.
URL http://linkinghub.elsevier.com/retrieve/pii/S0304380005004436

[31] I. Hoteit, D.-T. Pham, J. Blum, A simplified reduced order Kalman filtering and application to altimetric data assimilation in Tropical Pacific, Journal of Marine Systems 36 (2002) 101–127.

[32] C. K. Chui, G. Chen, Kalman Filtering, with Real-Time Applications, Fourth Edition, 2009.

[33] J. W. Campbell, The lognormal distribution as a model for bio-optical variability in the sea, Journal of Geophysical Research 100 (13) (1995) 237–254.

[34] J. K. Willis, Interannual variability in upper ocean heat content, temperature, and thermosteric expansion on global scales, Journal of Geophysical Research 109. doi:10.1029/2003JC002260.
URL http://doi.wiley.com/10.1029/2003JC002260

[35] T. N. Miles, R. He, Temporal and spatial variability of Chl-a and SST on the South Atlantic Bight: Revisiting with cloud-free reconstructions of MODIS satellite imagery, Continental Shelf Research 30 (2010) 1951–1962. doi:10.1016/j.csr.2010.08.016.
URL http://dx.doi.org/10.1016/j.csr.2010.08.016

[36] D. Sirjacobs, A. Alvera-Azcárate, A. Barth, G. Lacroix, Y. Park, B. Nechad, K. Ruddick, J.-M. Beckers, Cloud filling of ocean colour and sea surface temperature remote sensing products over the Southern North Sea by the Data Interpolating Empirical Orthogonal Functions methodology, Journal of Sea Research 65 (2011) 114–130. doi:10.1016/j.seares.2010.08.002.
URL http://linkinghub.elsevier.com/retrieve/pii/S1385110110001036

[37] J. N. Waite, F. J. Mueter, Spatial and temporal variability of chlorophyll-a concentrations in the coastal Gulf of Alaska, 1998-2011, using cloud-free reconstructions of SeaWiFS and MODIS-Aqua data, Progress in Oceanography 116 (2013) 179–192. doi:10.1016/j.pocean.2013.07.006.
URL http://dx.doi.org/10.1016/j.pocean.2013.07.006

[38] J.-M. Beckers, A. Barth, A. Alvera-Azcárate, DINEOF reconstruction of clouded images including error maps-application to the Sea-Surface Temperature around Corsican Island, Ocean Science 2 (2006) 183–199.

561 doi:10.5194/osd-3-735-2006.
562 URL http://www.ocean-sci-discuss.net/3/735/2006/

563 [39] D. L. Zimmerman, M. Stein, Classical Geostatistical Methods,
564 in: A. Gelfand, P. Diggle, M. Fuentes, P. Guttorp (Eds.),
565 Handbook of Spatial Statistics, Vol. 20103158 of Chapman &
566 Hall/CRC Handbooks of Modern Statistical Methods, 2010, pp. 29–44.
567 doi:10.1201/9781420072884.

568 [40] D. L. Zimmerman, M. B. Zimmerman, A Comparison of Spatial Semi-
569 variogram Estimators and Corresponding Ordinary Kriging Predictors,
570 Technometrics 33 (1) (1991) 77–91.

571 [41] R. H. Shumway, S. S. David, Time Series Analysis and Its Applications,
572 2011.

573 [42] P. Brasseur, Ocean data assimilation using sequential methods based
574 on the Kalman filter, in: Ocean Weather Forecasting, springer Edition,
575 2006, pp. 271–316.

List of Tables

1 Spatial building blocks proposed by [1] 36
2 Temporal building blocks proposed by [1] 37
3 Parameters of the covariance model given in equation (28) estimated by WLS. 38

Table 1: Spatial building blocks proposed by [1]

Function	Parameters
$\varphi_1(t) = \exp(-ct^\gamma)$	$c > 0, 0 < \gamma \leq 1$
$\varphi_2(t) = (2^{\nu-1}\Gamma(\nu))^{-1}(ct^{1/2})^\nu K_\nu(ct^{1/2})$	$c > 0, \nu > 0$
$\varphi_3(t) = (1 + ct^\gamma)^{-\nu}$	$c > 0, 0 < \gamma \leq 1, \nu > 0$
$\varphi_4(t) = 2^\nu(\exp(ct^{1/2}) + \exp(-ct^{1/2}))^{-\nu}$	$c > 0, \nu > 0$

Table 2: Temporal building blocks proposed by [1]

Function	Parameters
$\psi_1(t) = (at^\alpha + 1)^\beta$	$a > 0, 0 < \alpha \leq 1, 0 \leq \beta \leq 1$
$\psi_2(t) = \ln(at^\alpha + b)/\ln(b)$	$a > 0, b > 1, 0 < \alpha \leq 1$
$\psi_3(t) = (at^\alpha + b)/(b(at^\alpha + 1))$	$a > 0, 0 < b \leq 1, 0 < \alpha \leq 1$

Table 3: Parameters of the covariance model given in equation (28) estimated by WLS.

σ^2	τ	c	γ	a	α	β
0.08	0.82	0.45	0.10	0.1	1.0	0.92

List of Figures

1 Raw data plots: (a) map of average log-concentrations of CHL between 1998 and 2004, (b) map of average percentage of missing data for each location between 1998 and 2006, and (c) time-series of the percentage of missing data over the Red Sea between 1998 and 2004. 41

2 Log-CHL concentration time-series for every pixel of the domain (blue curves). The red curve plots the spatially averaged log-concentrations and the black curve plots the spatial average of the seasonal component. Both are computed from the data filled with DINEOF. 42

3 Anomalies estimated by DINEOF for the first 8 weeks of 1998. 43

4 Anomaly times series (panels b to d) at three locations as indicated in panel (a) 44

5 Description of the experimental setup. 45

6 RMS error over the cross-validation period for a varying number of smoothing iterations and number of EOFs. Crosses indicate the minimal error for a given number of iterations. . . 46

7 EOF modes after the DINEOF data filling: (a) first spatial EOF mode, (b) second spatial EOF mode, (c) first temporal EOF mode, (d) second temporal EOF mode, (e) spatial average of the seasonal anomalies computed from the DINEOF filled data, (f) cumulative sum of the percentage of variance explained for the modes computed with DINEOF. 47

8 Fitting the covariance model to the space-time empirical covariance matrix: (b) empirical space covariance function and fitted space covariance functions, (c) empirical time covariance function and fitted time covariance functions. 48

9 Contour plot of the empirical covariance function (dashed curves) and of the fitted space-time covariance function (solid curves) 49

10 Results of the Kalman filter: (a) time series of RMS errors of forecasts for the covariance model compared with a pure seasonal prediction, (b) distribution of the prediction RMS error over the three-year cross-validation period. 50

11 Predictions for the period between 24 October 2006 and 31 October 2006 (a, b, c), and between 6 March 2006 and 14 March 2006 (d, e, f): (a, d) plot the seasonal prediction, (b, e) the model prediction with the Kalman filter, and (c, f) the observations. 51
12 Spatial averages over the validation period of: (a) the variances of the prediction error resulting from by the reduced-order Kalman filter, and (b) the RMS errors. 52
13 Diagnostic statistics of the filter: (a) increment distribution in space averaged over the validation period, (b) cumulative averaged innovation over time, (c) distribution of the innovation values for the validation period. 53
14 Average correlation maps over the validation period between the observations and (a) the seasonal forecast, (b) the CHL forecasts (including the seasonal component), and (c) the filter analyses (including the seasonal component). 54

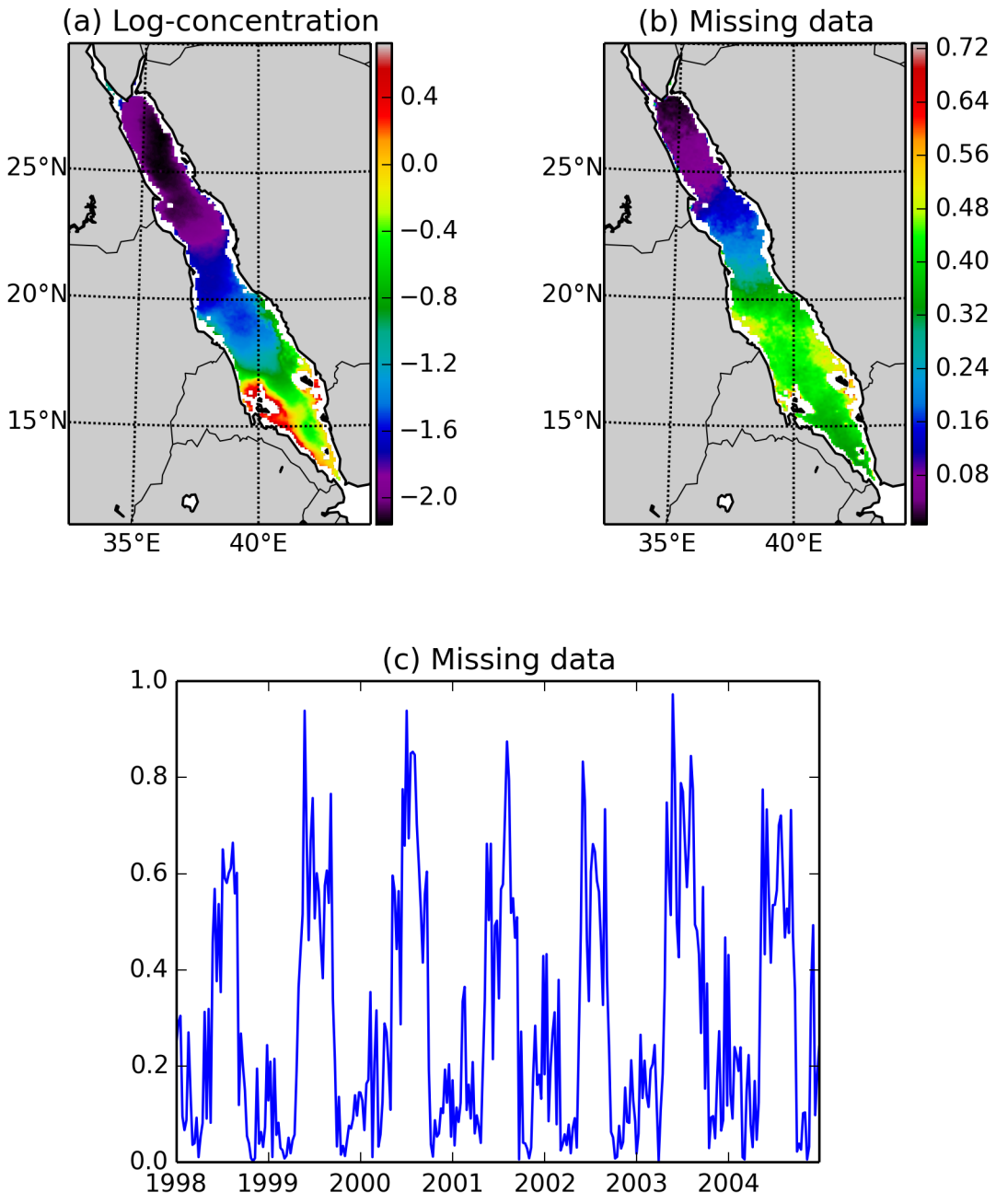

Figure 1: Raw data plots: (a) map of average log-concentrations of CHL between 1998 and 2004, (b) map of average percentage of missing data for each location between 1998 and 2006, and (c) time-series of the percentage of missing data over the Red Sea between 1998 and 2004.

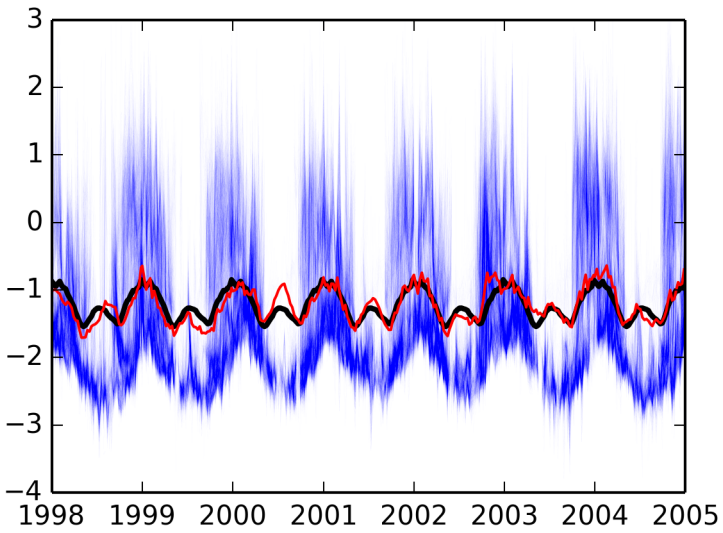

Figure 2: Log-CHL concentration time-series for every pixel of the domain (blue curves). The red curve plots the spatially averaged log-concentrations and the black curve plots the spatial average of the seasonal component. Both are computed from the data filled with DINEOF.

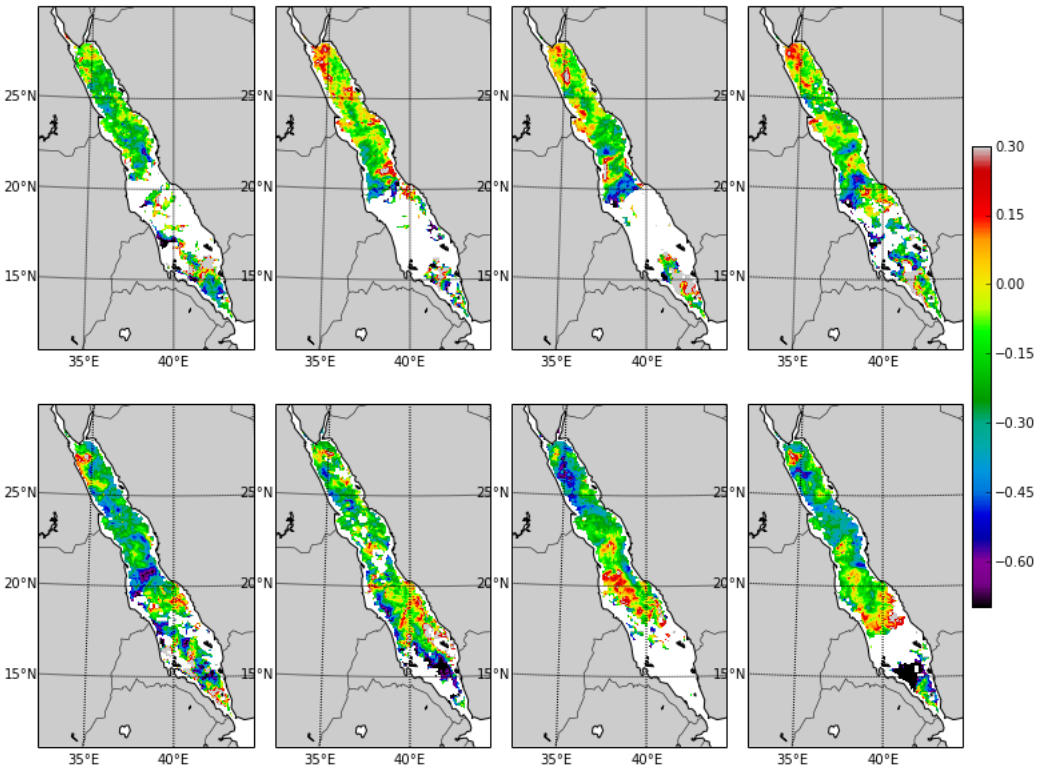

Figure 3: Anomalies estimated by DINEOF for the first 8 weeks of 1998.

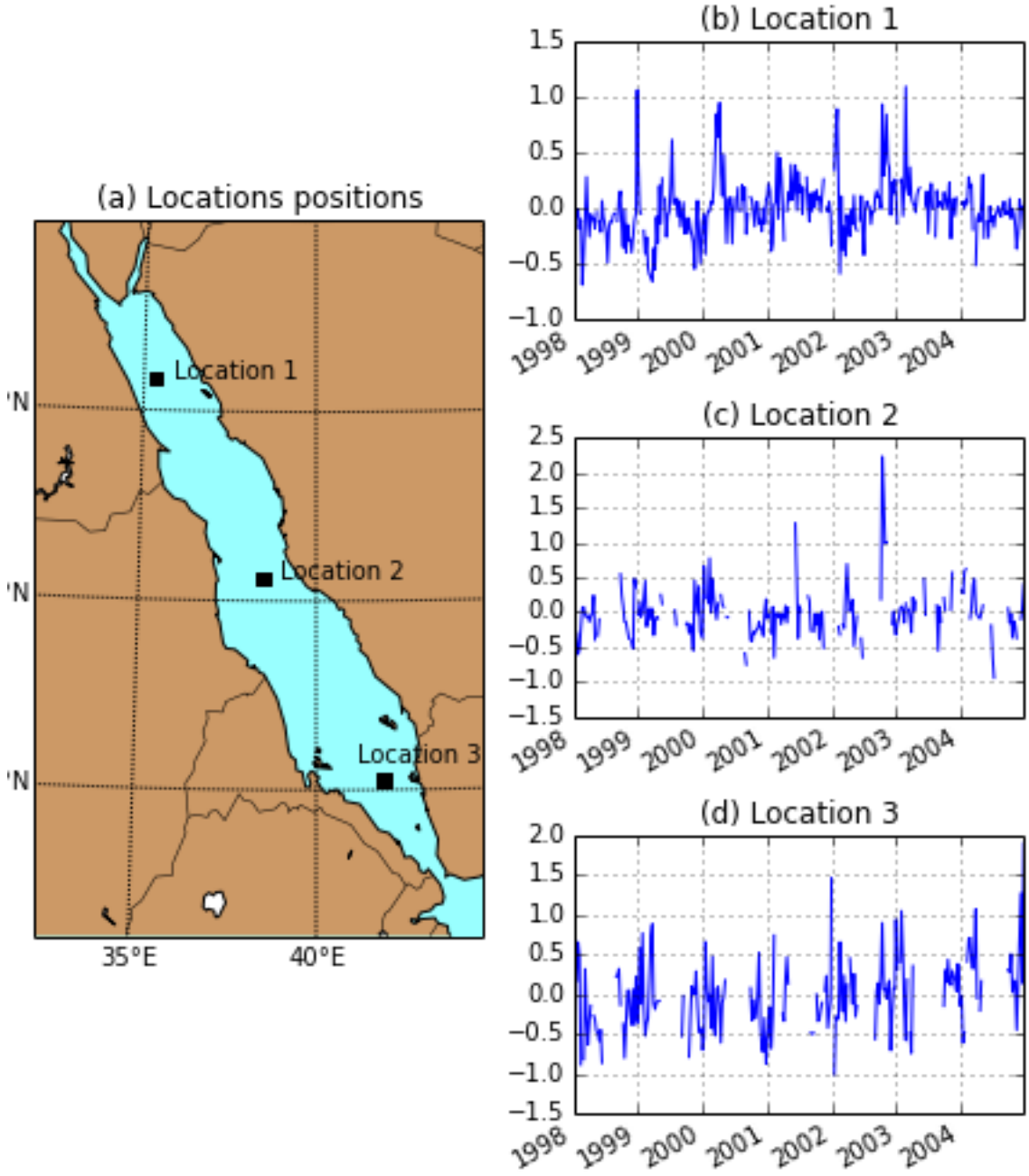

Figure 4: Anomaly times series (panels b to d) at three locations as indicated in panel (a)

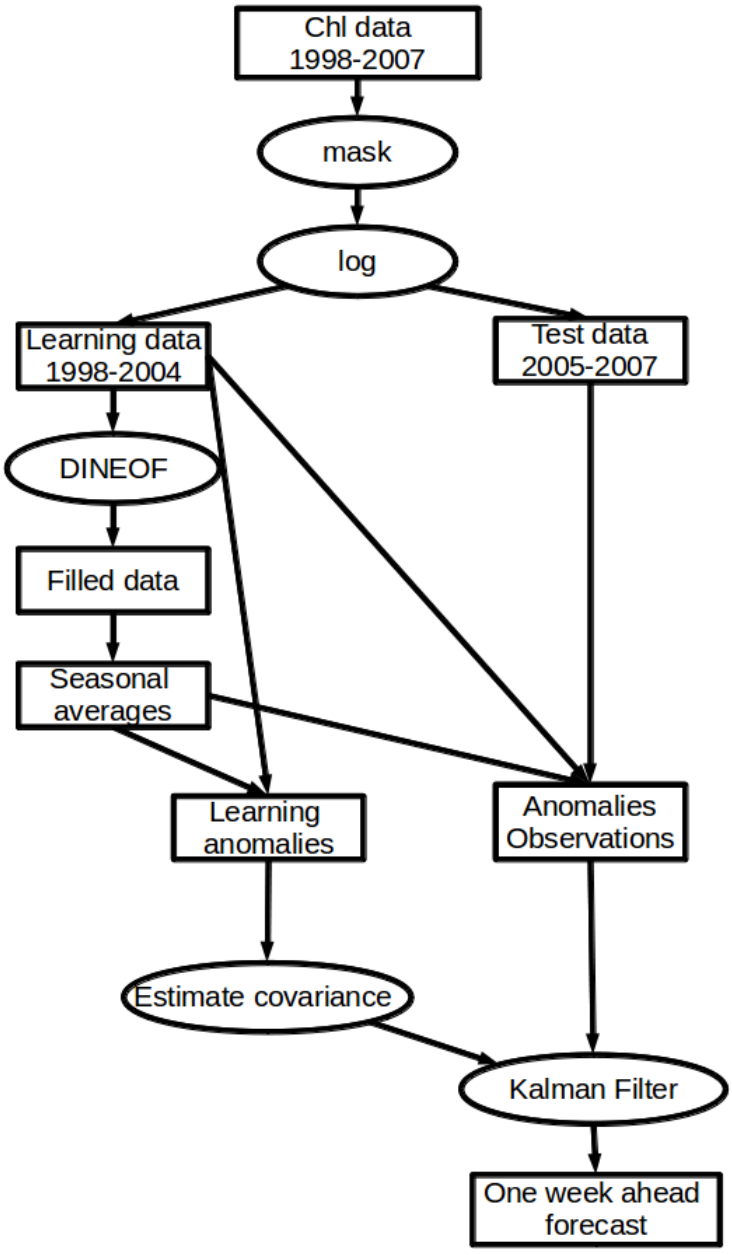

Figure 5: Description of the experimental setup.

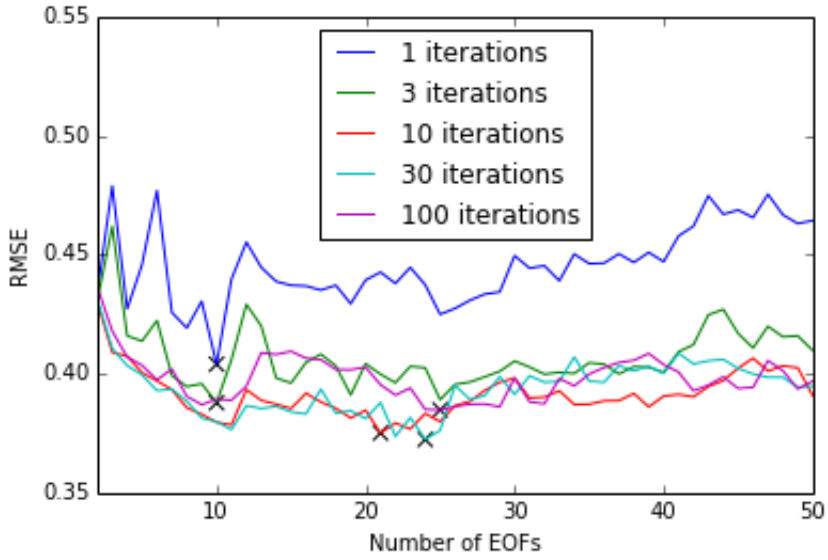

Figure 6: RMS error over the cross-validation period for a varying number of smoothing iterations and number of EOFs. Crosses indicate the minimal error for a given number of iterations.

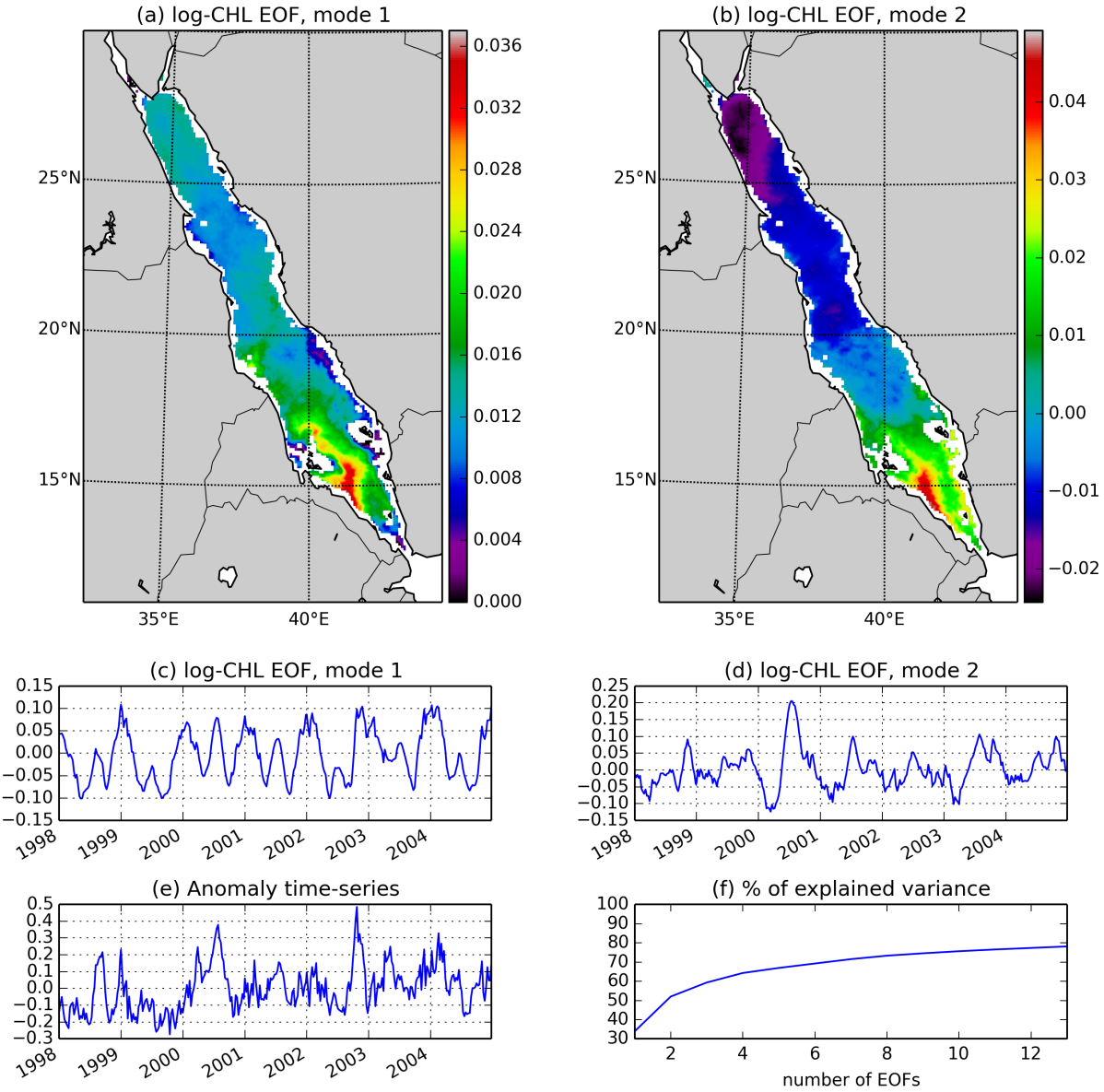

Figure 7: EOF modes after the DINEOF data filling: (a) first spatial EOF mode, (b) second spatial EOF mode, (c) first temporal EOF mode, (d) second temporal EOF mode, (e) spatial average of the seasonal anomalies computed from the DINEOF filled data, (f) cumulative sum of the percentage of variance explained for the modes computed with DINEOF.

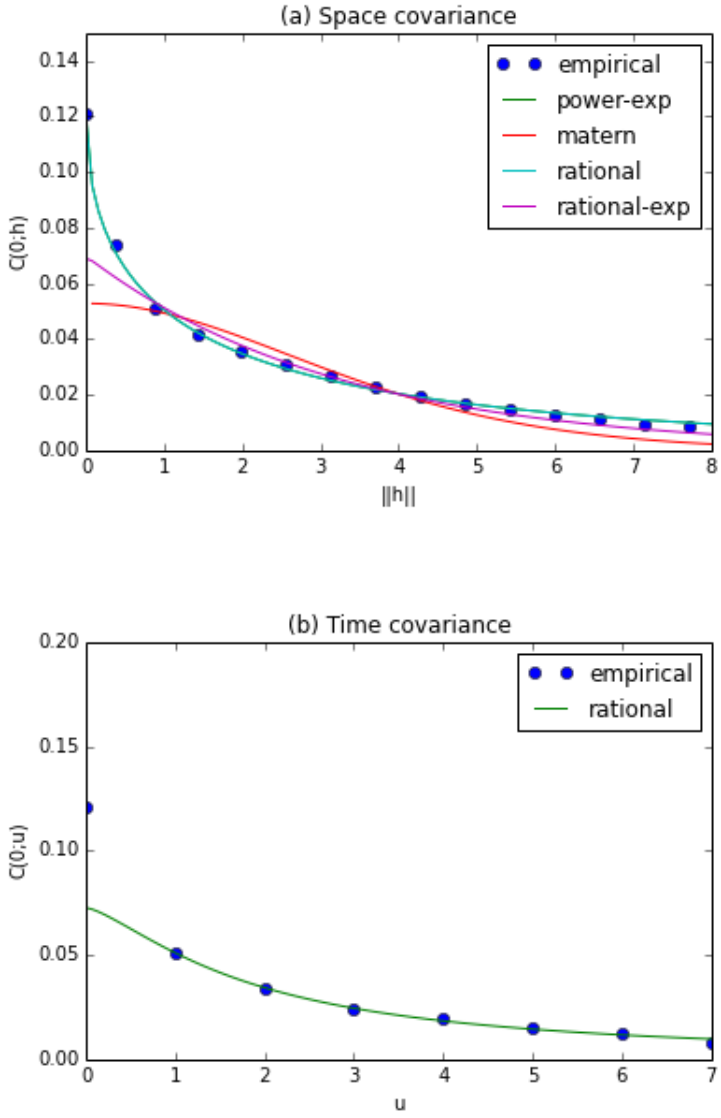

Figure 8: Fitting the covariance model to the space-time empirical covariance matrix: (b) empirical space covariance function and fitted space covariance functions, (c) empirical time covariance function and fitted time covariance functions.

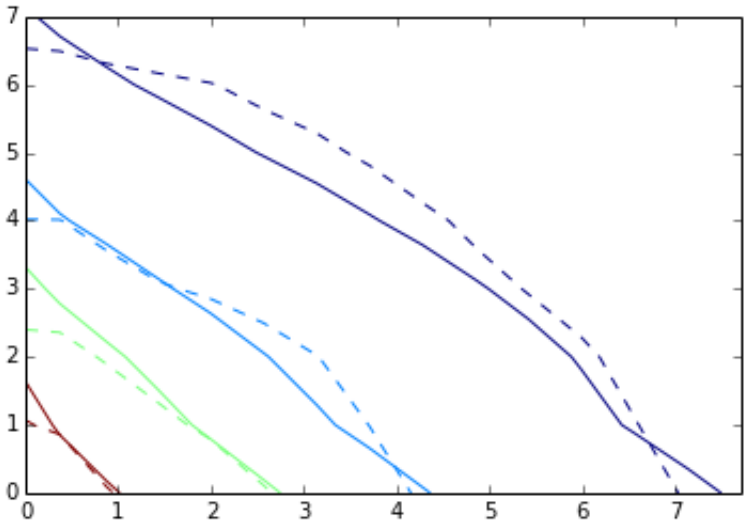

Figure 9: Contour plot of the empirical covariance function (dashed curves) and of the fitted space-time covariance function (solid curves)

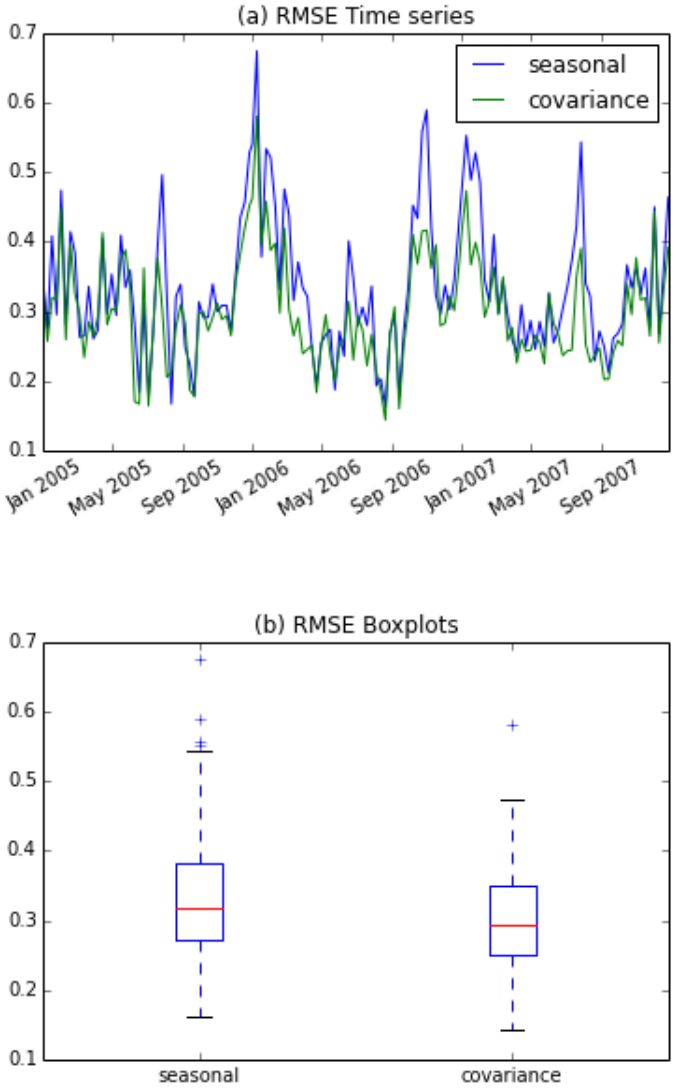

Figure 10: Results of the Kalman filter: (a) time series of RMS errors of forecasts for the covariance model compared with a pure seasonal prediction, (b) distribution of the prediction RMS error over the three-year cross-validation period.

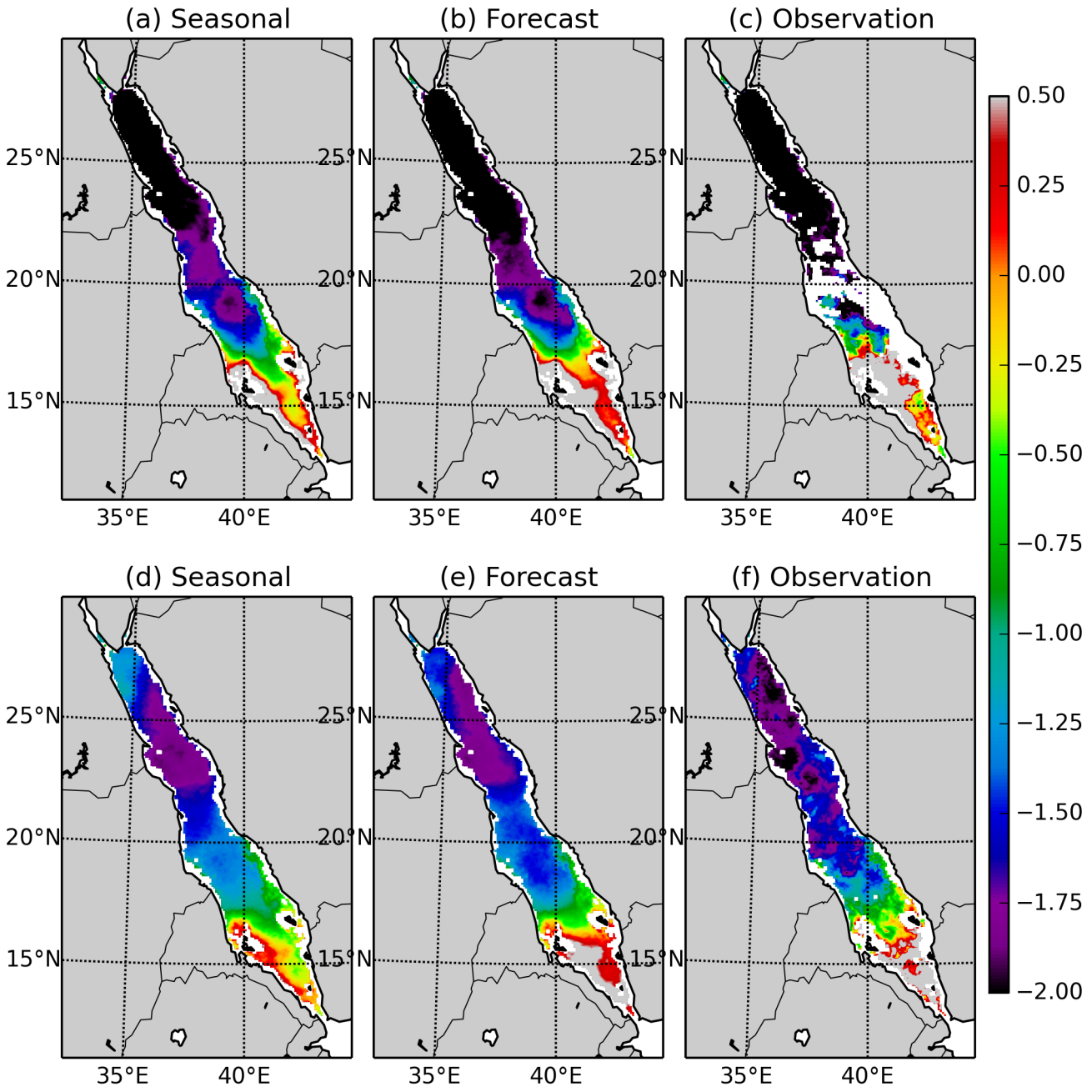

Figure 11: Predictions for the period between 24 October 2006 and 31 October 2006 (a, b, c), and between 6 March 2006 and 14 March 2006 (d, e, f): (a, d) plot the seasonal prediction, (b, e) the model prediction with the Kalman filter, and (c, f) the observations.

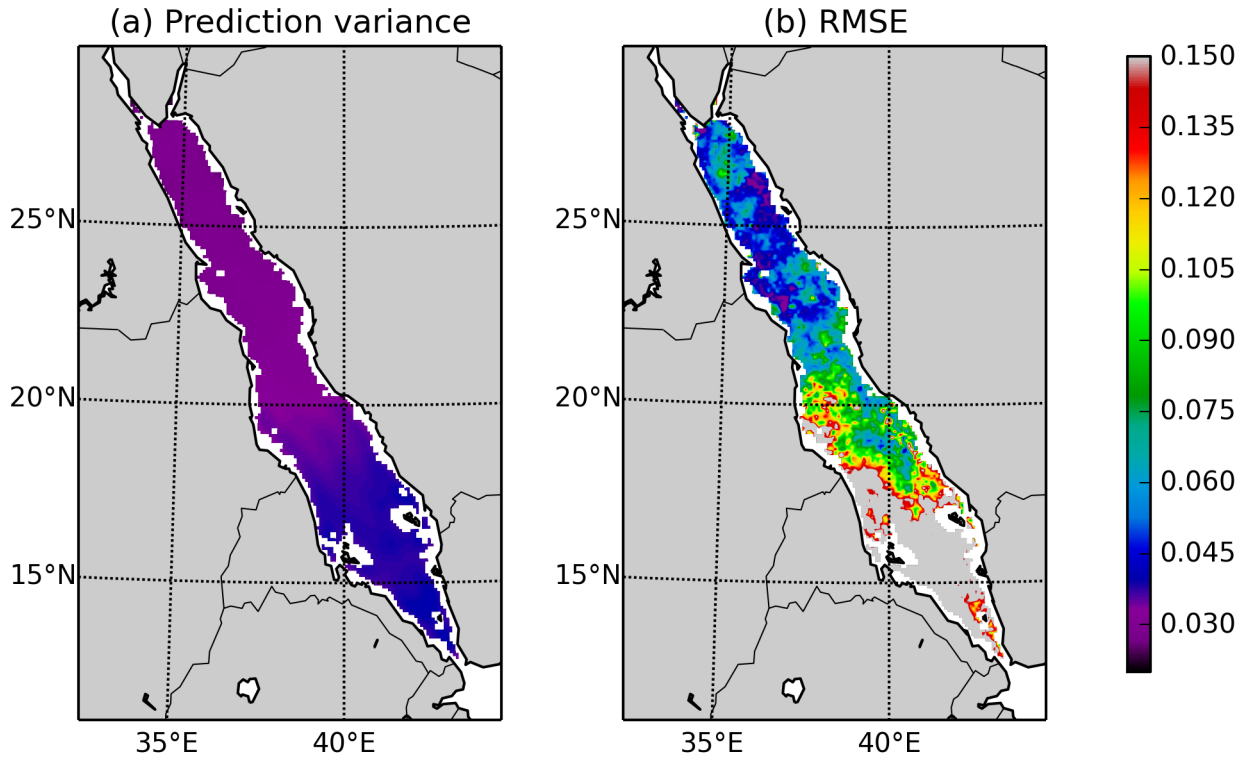

Figure 12: Spatial averages over the validation period of: (a) the variances of the prediction error resulting from by the reduced-order Kalman filter, and (b) the RMS errors.

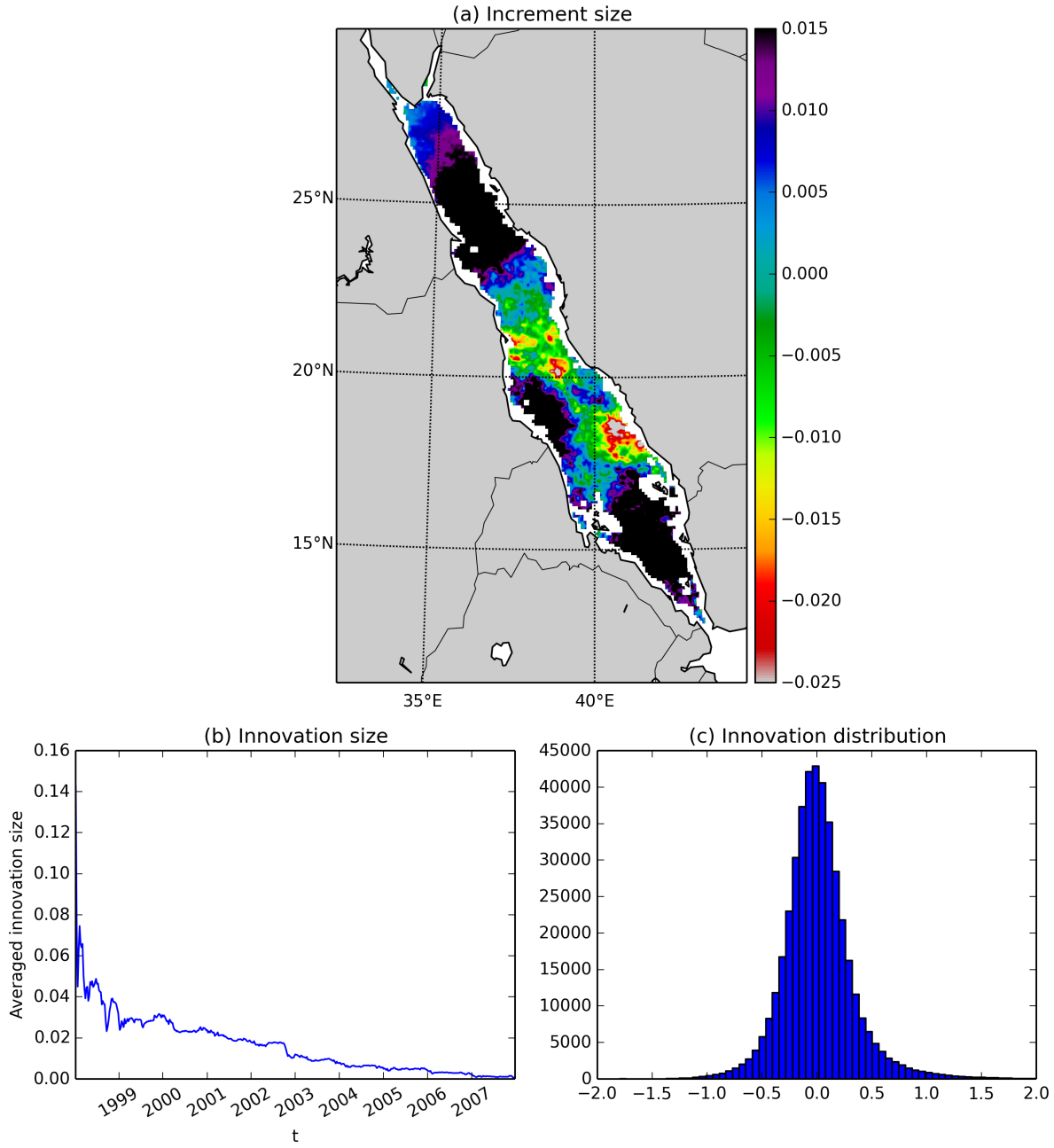

Figure 13: Diagnostic statistics of the filter: (a) increment distribution in space averaged over the validation period, (b) cumulative averaged innovation over time, (c) distribution of the innovation values for the validation period.

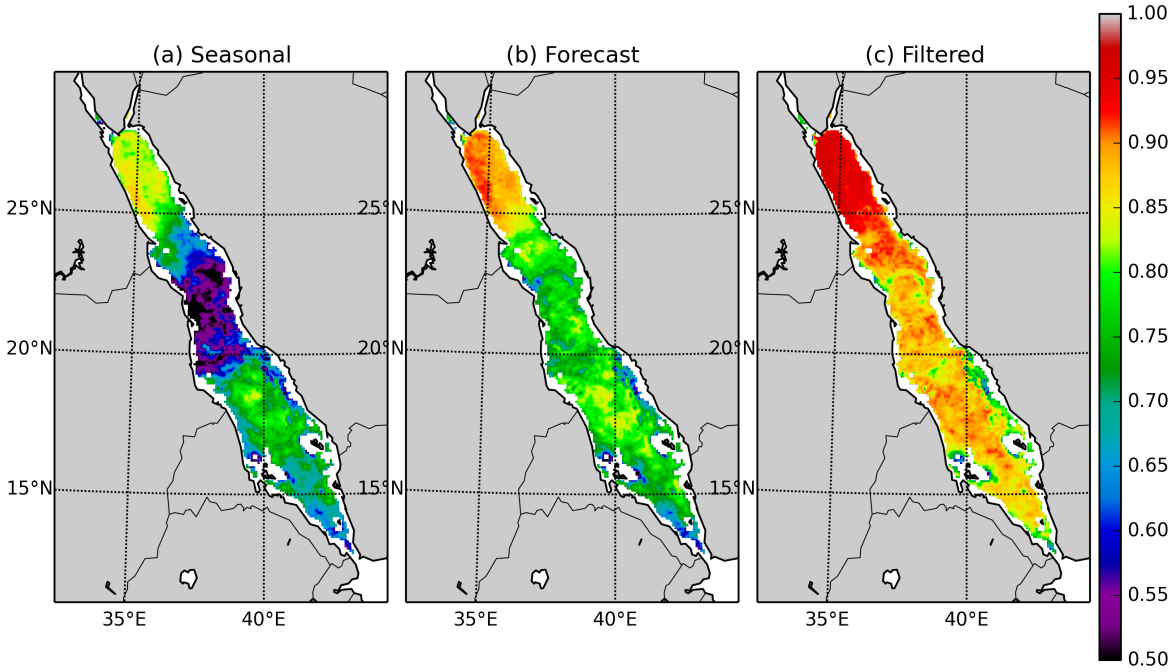

Figure 14: Average correlation maps over the validation period between the observations and (a) the seasonal forecast, (b) the CHL forecasts (including the seasonal component), and (c) the filter analyses (including the seasonal component).