Accepted Manuscript

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

Denis Dreano, Bani Mallick, Ibrahim Hoteit

PII: S2211-6753(15)00026-3

DOI: <http://dx.doi.org/10.1016/j.spasta.2015.04.002>

Reference: SPASTA 94

To appear in: *Spatial Statistics*

Received date: 2 October 2014

Accepted date: 7 April 2015

Please cite this article as: Dreano, D., Mallick, B., Hoteit, I., Filtering remotely sensed chlorophyll concentrations in the Red Sea using a space-time covariance model and a Kalman filter. *Spatial Statistics* (2015), <http://dx.doi.org/10.1016/j.spasta.2015.04.002>

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

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 chlorophyll concentrations. The proposed method decreases the root mean square (RMS)

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

prediction error by about 11% compared with the seasonal average.

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]. 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-situ measurements have been collected 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, up to this point, they constitute the basis of several studies in the Red Sea [7, 10, 9]. Here we are interested in forecasting chlorophyll concentrations in the Red Sea, and we therefore need a dynamical model to mimic its evolution.

There are two approaches to chlorophyll modeling: deterministic or data-driven (statistical). A broad range of deterministic models has been developed by the marine ecosystem research community, from the very simple NPZ model with only nutrients, phytoplankton and zooplankton as state variables, to much more complex models, such as the European Regional Seas Ecosystem Model (ERSEM) [12]. The latter distinguishes functional phytoplankton and zooplankton groups, and models the complete cycling of different nutrient groups and O_2 and CO_2 , including the effect of higher trophic groups. An example of a 3D ERSEM coupled model has recently been implemented in the Red Sea [8]. Simulating the ecosystem with such a model can be however very challenging, as it requires the coupling to an ocean circulation model, which provides the physical forcing. Configuring these models further requires considerable efforts and expertise [13], because of the high number of parameters (over 50 for ERSEM [14]). This makes such models difficult to calibrate and validate, since there are usually not

enough observations to constrain the parameters [15].

An alternative approach is to follow a statistical framework to model the space-time evolution of chlorophyll concentrations. Space-time statistical methods have not been used yet in this field, which has so far 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 used for finding explanatory variables for the chlorophyll concentrations in the Pagasitikos Gulf and the subarctic 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 is not based on a dynamical framework, geostatistical methods may capture some patterns in the data and avoid the difficulties of developing dynamical models [21, 20]. These methods have been 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 in environment applications 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. [27, 1] proposed methods to construct such families from known temporal and spatial covariance functions. The method proposed in [1] is appealing because of its modularity and interpretability.

We adopt the aforementioned method to model the anomaly fields of chlorophyll concentration in the Red Sea. To the best of our knowledge, this is the first application of a space-time geostatistical modeling approach to model and predict remotely-sensed chlorophyll data. After filling the data with an Empirical Orthogonal Function (EOF)-based method (DINEOF [28, 29]), we model it as the realization of a spatio-temporal Gaussian process and estimate the mean and covariance functions. Based on these, one may forecast chlorophyll concentration using Kriging by conditioning on past observations [20]. However, this approach is not computationally efficient, as it requires inverting a large matrix at every time step when many observations are available, which is the case here. We therefore propose a new approach, in which we reformulate the problem withing a state-space framework and use an efficient reduced-order Kalman filter [30] to filter and forecast chlorophyll. The state-space system is further expanded to take into account the colored model noise [31]. In contrast with the space-time hierarchical models proposed in [32, 33], the linear dynamical model is derived from a space-time geostatistical model, avoiding the need to directly estimate the transition matrix. Furthermore, the dimension reduction is not part of the modeling,

but is implemented at the filtering level through a reduced-order Kalman filter.

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 at high with spatial and temporal resolutions, making them particularly valuable to the Red Sea, where very few in situ data are available.

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 is applied in order to obtain an approximately Gaussian distribution [34]. Pixels with too few observations are discarded, and a control quality check is applied to remove

113 outliers [35].

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

121 The dataset is separated into a training set, composed of the first seven
 122 years of data, and a validation set, containing the remaining three years.
 123 Both datasets were preprocessed and log-transformed in the same way. The
 124 computations for the data filling and for the covariance model selection and
 125 estimation are based only on the training dataset. The testing dataset is
 126 used to validate the model predictions outside the training period.

127 **3. Data filling and modeling**

128 *3.1. Data filling*

129 The DINEOF (Data Interpolating Empirical Orthogonal Function) is an
 130 EOF-based, recursive method for the reconstruction of data matrices with
 131 missing values [28, 29]. It estimates the values of the missing data by suc-
 132 cessive singular values decompositions (SVD) of a given data matrix and
 133 truncated reconstructions. The advantage of this method is that it does not
 134 require any a priori information about the data. It has been successfully

used reconstructing incomplete chlorophyll datasets in various regions of the ocean [36, 37, 38].

Let $\mathbf{X}$ be an $m \times k$ data matrix, where each row is an observation, and each column is a variable. The average of each variable is removed from the corresponding observation in order to center the data matrix. The missing values are 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]. As in the former, the smoothing parameter of the Laplacian filter is set to $0.01d^2$, where d is the laps of time between two successive observations. 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 $\mathbf{X}$ 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 [39] is followed to select a cross-validation dataset. The idea is to apply cross-validation over contiguous regions, instead of applying it to randomly selected points. The shape of the cross-validation regions are randomly selected from the shape of missing data regions in the dataset. We set aside the cross-validation data by randomly applying these shapes to the data until 3% of the total number of data points has been extracted.

3.2. Geostatistical modeling

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

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

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 $K(\mathbf{s}, \mathbf{r}; t, q) = C(\mathbf{r} - \mathbf{s}, q - t)$.

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

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

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 $\mathbf{h}$ is the spatial lag, u is the temporal lag, and $\boldsymbol{\theta}$ is the vector of parameters to be estimated [40].

3.3. Mean function estimation and anomalies

As can be seen in Figure 2, the Red Sea chlorophyll data exhibit a strong seasonality. The climatological means explain 50% of the variability in the data. We thus choose to model the mean function $\mu(\mathbf{s}; t)$ as the climatological average. 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 subtracting 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 correlations. 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 increasing pattern appears, we conclude, as in [27], that the assumption of stationarity for the covariance model is 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 in [1]. We used this method because of its flexibility and simplicity, since no transformation of the covariance function is involved. 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 covariance function is valid (i.e., positive-definite):

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

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 [40]. The empirical covariance function is first estimated using the formula:

$$\hat{C}(\mathbf{h}(l); u) = \frac{1}{2|N(\mathbf{h}(l); u)|} \times \quad (4)$$

$$\sum_{(i,j,t,t') \in N(\mathbf{h}(l); u)} (Z(\mathbf{s}_i; t) - \hat{\mu}(\mathbf{s}_i; t))(Z(\mathbf{s}_j; t') - \hat{\mu}(\mathbf{s}_j; t')),$$

$$N(\mathbf{h}(l); u) = \{(i, j, t, t') \text{ such that : } |t - t'| = u, \mathbf{s}_i - \mathbf{s}_j \in \text{Bin}(\mathbf{h}(l))\}, \quad (5)$$

where $(\mathbf{s}_i, t)$ and $(\mathbf{s}_j, t')$ are spatio-temporal locations at which Z is observed, and $\hat{\mu}$ is an estimate of the mean function. $\mathbb{R}^2$ has been divided into a finite number N_s 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 $(l - 1) * \delta h \leq \|\mathbf{k}\| \leq l\delta h, l = 0, \dots, L$, 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 [41]. In practice, the estimator of $\boldsymbol{\theta}$ minimizes the weighted sum of the squares of errors:

$$S(\boldsymbol{\theta}) = \sum_{l=1}^{N_s} \sum_{u=0}^T w_{l,u} \left(\hat{C}(\mathbf{h}(l); u) - C(\mathbf{h}(l); u | \boldsymbol{\theta}) \right)^2, \quad (6)$$

with weights $w_{l,u} = \frac{|N(\mathbf{h}(l);u)|}{(1-C(\mathbf{h}(l);u|\boldsymbol{\theta}))^2}$. T is the maximum temporal lag considered. ϕ and ψ are chosen from the functions presented in Tables 1 and 2, by separately fitting the covariance function in space and time 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 apply 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.

4.1. State-space modeling

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

$$Y(\mathbf{s};t) = \mu(\mathbf{s};t) + a(\mathbf{s};t), \text{ where } \mathbf{s}, t \text{ are observable locations,} \quad (7)$$

$$Z(\mathbf{s};t) = Y(\mathbf{s};t) + \varepsilon(\mathbf{s};t), \text{ where } \mathbf{s}, t \text{ are observed locations.} \quad (8)$$

In this formulation, $\varepsilon(\mathbf{s}; t)$ is a white noise process representing the measurement errors, $a(\mathbf{s}; t)$ is the anomaly process with the covariance matrix estimated as described in the previous section, $\mu(\mathbf{s}; t)$ is the seasonal mean function, $Y(\mathbf{s}; t)$ is the underlying CHL log-concentration process, and $Z(\mathbf{s}; t)$ 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 apply Kriging. In classical geostatistics, Kriging interpolates available observations to provide the best linear unbiased estimate of a spatial process at unobserved locations [40]. This technique 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 $\mathbf{a}_t = \{a(\mathbf{s}_{i,t})\}_{i=1,\dots,k}$, 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

$$\mathbf{a}_t | \mathbf{a}_{t-1} \sim N(\mathbf{M}\mathbf{a}_{t-1}, \mathbf{Q}), \quad (9)$$

$$(10)$$

where

$$\mathbf{M} = \mathbf{C}_1 \mathbf{C}_0^{-1}, \quad (11)$$

$$\mathbf{Q} = \mathbf{C}_0 - \mathbf{C}_1 \mathbf{C}_0^{-1} \mathbf{C}_1, \quad (12)$$

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

$$\mathbf{y}_t = \boldsymbol{\mu}_t + \mathbf{M}(\mathbf{y}_{t-1} - \boldsymbol{\mu}_{t-1}) + \boldsymbol{\eta}_t, \quad (13)$$

$$\tilde{\mathbf{z}}_t = \mathbf{H}_t \mathbf{y}_t + \boldsymbol{\varepsilon}_t, \quad (14)$$

with $\mathbf{y}_t = \mathbf{a}_t + \boldsymbol{\mu}_t$ the CHL process discretized in space, $\boldsymbol{\mu}_t$ is the vector seasonal component, $\mathbf{H}_t$ 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 $\tilde{\mathbf{z}}_t$ 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 $\mathbf{a}_t$ is the state vector:

$$\mathbf{a}_t = \mathbf{M} \mathbf{a}_{t-1} + \boldsymbol{\eta}_t, \quad (15)$$

$$\mathbf{z}_t = \tilde{\mathbf{z}}_t - \mathbf{H}_t \boldsymbol{\mu}_t = \mathbf{H}_t \mathbf{a}_t + \boldsymbol{\varepsilon}_t. \quad (16)$$

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 $\boldsymbol{\varepsilon}_t \sim N(0, \sigma_{\text{obs}}^2 \mathbf{I})$, independent and identically distributed (i.i.d.), with σ_{obs}^2 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 one¹ 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 of a linear dynamical model given past observations [42].

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

$$\mathbf{a}_t^f = \mathbf{M}\mathbf{a}_{t-1}^a, \quad (17)$$

$$\mathbf{P}_t^f = \mathbf{M}\mathbf{P}_{t-1}^a\mathbf{M}^T + \mathbf{Q}. \quad (18)$$

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

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

$$\mathbf{a}_t^a = \mathbf{a}_t^f + \mathbf{K}_t(\mathbf{z}_t - \mathbf{H}_t\mathbf{a}_t^f), \quad (19)$$

$$\mathbf{P}_t^a = (\mathbf{I} - \mathbf{K}_t\mathbf{H}_t)\mathbf{P}_t^f, \quad (20)$$

$$\mathbf{K}_t = \mathbf{P}_t^f\mathbf{H}_t^T(\mathbf{H}_t\mathbf{P}_t^f\mathbf{H}_t^T + \sigma_{\text{obs}}^2\mathbf{I})^{-1}. \quad (21)$$

255 where $\mathbf{K}_t$ is known as the Kalman gain.

The inversion in the calculation of the Kalman gain is computationally quite demanding because the number of observations can be as large as the size of the state. To speed up this computation, the reduced-order Kalman filter [30] 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 [30]. Defining

$$\mathbf{V} = \mathbf{L}^T\mathbf{Q}\mathbf{L}, \quad (22)$$

$$\mathbf{W} = \mathbf{L}^T\mathbf{M}\mathbf{L}, \quad (23)$$

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:

$$\mathbf{U}_t^f = \mathbf{W}\mathbf{U}_{t-1}^a\mathbf{W}^T + \mathbf{V}, \quad (24)$$

$$(\mathbf{U}_t^a)^{-1} = (\mathbf{U}_t^f)^{-1} + \mathbf{L}^T\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\mathbf{L}. \quad (25)$$

The Kalman gain can be then computed as

$$\mathbf{K}_t = \sigma_{\text{obs}}^{-2}\mathbf{L}\mathbf{U}_t^a\mathbf{L}^T\mathbf{H}^T. \quad (26)$$

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

4.3. Low-rank Kalman filter with colored noise

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 [31], so the new state-space system becomes:

$$\hat{\mathbf{a}}_t = \mathbf{\Gamma} \hat{\mathbf{a}}_{t-1} + \hat{\boldsymbol{\eta}}_t, \quad (27)$$

$$\mathbf{z}_t = \hat{\mathbf{H}} \hat{\mathbf{a}}_{t-1} + \boldsymbol{\varepsilon}_t, \quad (28)$$

with $\hat{\mathbf{a}}_t = [\mathbf{a}_t^T, \boldsymbol{\eta}_t^T]^T$, $\mathbf{\Gamma} = \begin{bmatrix} \mathbf{M} & \mathbf{I} \\ \mathbf{0} & \mathbf{N} \end{bmatrix}$, $\hat{\boldsymbol{\eta}}_t = [\mathbf{0}, \boldsymbol{\xi}_t^T]^T$ 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 three years of testing data to validate the system's performance and results. The results of the DINEOF data filling algorithm are presented before a covari-

ance 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. The lowest RMS error is achieved with 24 EOFs and 30 iterations. These parameters were therefore selected in DINEOF to fill the missing values in the training dataset.

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]. The winter bloom has been linked to convective overturning the northern Red Sea, and enhanced water intrusion from the Gulf of Aden in the South, whereas the causes of summer bloom are still not completely understood yet [9]. Figure 7(a) shows that the bloom is relatively homogeneous over the Red Sea, except in the southwest corner where the variation is more 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 $\sigma_{\text{spatial}}^2$ while ignoring the observations for $\|\mathbf{h}\| = 0$. This allows a nugget effect to be taken into 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 $\hat{C}(0; u)$. Among the functions in Table 2, we eliminate ψ_2 , which defines a temporal covariance function that decreases very slowly, which is not realistic to model anomalies. We also eliminate ψ_3 which defines a temporal covariance function that do not tend to 0 for increasing time lags. We therefore use the temporal covariance function defined by ψ_1 , to which we add a temporal nugget effect. We fit the parameters using WLS and display the results in Figure 8b, where the 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. The reparametrization is used to separate the variables that control the temporal covariance function and those controlling the spatial covariance function. A single parameter β controls the space-time interaction strength. A time-only nugget effect is also added, leading to the following covariance model:

$$C(\mathbf{h}; u) = \begin{cases} \sigma^2 \exp(-c\|\mathbf{h}\|^{2\gamma}) & , \text{ if } u = 0, \\ \frac{\sigma^2 \tau}{a|u|^{2\alpha+1}} \exp(-\frac{c\|\mathbf{h}\|^{2\gamma}}{(a|u|^{2\alpha+1})^{\beta\gamma}}) & , \text{ otherwise,} \end{cases} \quad (29)$$

with $0 < \tau \leq 1$.

The initial values for the WLS fitting are given by the results of the preceding purely spatial and 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 exhibits level curves very close to those of the empirical covariance model. It does not, however, reproduce their curvature.

5.3. Filtering

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

The anomaly model with a space-time covariance function helps decreasing the RMS error by nearly 11% over the test period ($p < 0.01$ using a

permutation test). As Figure 10a indicates, the improvement is most noticeable 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) plots a similar comparison 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 RMS error, 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 RMS error corresponds to the diagonals of the prediction covariance matrix as estimated by the reduced-order Kalman filter. The prediction RMS error is computed from the error between the model prediction and the observation. One can see that both values are close in the northern half of the Red Sea, but the RMS error is much larger in the south. This is caused by the lack of data and the fact that the dynamics in the south are different from those in the north, making the process non-stationary. 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.

Another way to evaluate the filter's behavior is to examine the distribution of the innovations and the increments [43]. The innovation corresponds to the difference: $\mathbf{z}_t - \mathbf{H}_t \mathbf{a}_t^f$, 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 from 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 are positively biased in some regions. This suggests 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 correlations between the observations and the seasonal predictions, the Kalman filter forecasts and

the analyses, respectively. Compared with the seasonal correlations, one 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

We considered the filtering problem of satellite-derived chlorophyll (CHL) concentration in the Red Sea. This is a difficult problem, as ecological system are highly nonlinear, and difficult to forecast. Moreover, the CHL dataset has a large number of missing data, particularly in the southern Red Sea, during summer. We proposed a data-driven approach in which the CHL spatio-temporal evolution is modeled as a space-time Gaussian process. The DINEOF data-filling algorithm [39] 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.

Based on a space-time Kriging formulation, a linear model was derived to model 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 average computed from the

data. The reduction is more important during large anomalies, in particular in winter (Figure 10a). In general the model reduces the spread of the anomalies (Figure 10b). The RMS error reduction is however not spatially uniform and is more pronounced in the southern Red Sea (Figure 12). This can be caused by the number of missing data in this regions, as well as the fact that the anomaly process is not stationary over the whole domain.

The proposed approach is not difficult to implement, but requires some coding efforts. For the DINEOF, a standalone package is freely available online². 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 [44]. 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 phys-

²<http://modb.oce.ulg.ac.be/mediawiki/index.php/DINEOF>

ical dynamics from north to south. We can therefore expect the anomalies to be non-stationary. Fortunately, Kriging is robust to non-stationarity [44]. 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 reduces this error when new observations are available.

The proposed method can be further generalized by considering more 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.
- [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.
- [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.
- [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,

- 469 Remote Sensing the Phytoplankton Seasonal Succession of the Red Sea,
470 PLoS ONE 8 (6). doi:10.1371/journal.pone.0064909.
- 471 [10] R. J. W. Brewin, D. E. Raitsos, Y. Pradhan, I. Hoteit, Compari-
472 son of chlorophyll in the Red Sea derived from MODIS-Aqua and in
473 vivo fluorescence, *Remote Sensing of Environment* 136 (2013) 218–224.
474 doi:10.1016/j.rse.2013.04.018.
- 475 [11] C. R. McClain, A Decade of Satellite Ocean Color Ob-
476 servations, *Annual Review of Marine Science* (2009) 19–
477 42doi:10.1146/annurev.marine.010908.163650.
- 478 [12] J. Baretta, W. Ebenhöf, P. Ruardij, The European regional seas ecosys-
479 tem model, a complex marine ecosystem model, *Netherlands Journal of*
480 *Sea Research* 33 (3/4) (1995) 233–246. doi:10.1016/0077-7579(95)90047-
481 0.
- 482 [13] G. Petihakis, G. Triantafyllou, I. J. Allen, I. Hoteit, C. Dounas, Mod-
483 elling the spatial and temporal variability of the Cretan Sea ecosystem,
484 *Journal of Marine Systems* 36 (2002) 173–196.
- 485 [14] J. C. Blackford, P. J. Radford, A Structure and methodology for ma-
486 rine ecosystem modelling, *Netherlands Journal of Sea Research* 33 (3/4)
487 (1995) 247–260.
- 488 [15] T. R. Anderson, Plankton functional type modelling: running

- before we can walk?, *Journal of Plankton Research* 27 (11).
doi:10.1093/plankt/fbi076.
- [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, Artificial neural network approach for modelling and pre-
diction 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 Ma-
rine Systems* 94 (2012) S16–S22. doi:10.1016/j.jmarsys.2011.11.003.
- [19] D. E. Raitsos, S. J. Lavender, Y. Pradhan, T. Tyrrell, P. C. Reid, M. Ed-
wards, Coccolithophore bloom size variation in response to the regional
environment of the subarctic North Atlantic, *Limnology and Oceanog-
raphy* 51 (5) (2006) 2122–2130. doi:10.4319/lo.2006.51.5.2122.
- [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] 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.
- [31] C. K. Chui, G. Chen, *Kalman Filtering, with Real-Time Applications*, Fourth Edition, 2009.
- [32] C. K. Wikle, N. Cressie, A dimension-reduced approach to space-time kalman filtering, *Biometrika* 86 (4) (1999) 815–829.
- [33] N. Cressie, T. Shi, E. L. Kang, Fixed rank filtering for spatio-temporal data, *Journal of Computational and Graphical Statistics* 19 (3) (2010) 724–745, 658JF Times Cited:33 Cited References Count:37. doi:Doi 10.1198/Jcgs.2010.09051.
- [34] 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.
- [35] 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.
- [36] 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.

- [37] 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.
- [38] 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.
- [39] 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. doi:10.5194/osd-3-735-2006.
- [40] D. L. Zimmerman, M. Stein, Classical Geostatistical Methods, in: A. Gelfand, P. Diggle, M. Fuentes, P. Guttorp (Eds.), *Handbook of Spatial Statistics*, Vol. 20103158 of Chapman & Hall/CRC Handbooks of Modern Statistical Methods, 2010, pp. 29–44. doi:10.1201/9781420072884.
- [41] D. L. Zimmerman, M. B. Zimmerman, A Comparison of Spatial Semi-

- 1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
- 571 variogram Estimators and Corresponding Ordinary Kriging Predictors,
572 Technometrics 33 (1) (1991) 77–91.
- 573 [42] R. H. Shumway, S. S. David, Time Series Analysis and Its Applications,
574 2011.
- 575 [43] P. Brasseur, Ocean data assimilation using sequential methods based
576 on the Kalman filter, in: Ocean Weather Forecasting, springer Edition,
577 2006, pp. 271–316.
- 578 [44] P. Monestiez, L. Dubroca, E. Bonnin, J.-P. Durbec, C. Guinet, Geo-
579 statistical modelling of spatial distribution of Balaenoptera physalus in
580 the Northwestern Mediterranean Sea from sparse count data and het-
581 erogeneous observation efforts, Ecological Modelling 193 (3-4) (2006)
582 615–628. doi:10.1016/j.ecolmodel.2005.08.042.

583 **List of Tables**

584	1	Spatial building blocks proposed by [1]	36
585	2	Temporal building blocks proposed by [1]	37
586	3	Parameters of the covariance model given in equation (29)	
587		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 (29) 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

588 **List of Figures**

589	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
590			
591	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
592			
593	3	Anomalies estimated by DINEOF for the first 8 weeks of 1998.	43
594			
595	4	Anomaly times series (panels b to d) at three locations as indicated in panel (a)	44
596			
597	5	Description of the experimental setup.	45
598			
599	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
600			
601	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
602			
603	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
604			
605	9	Contour plot of the empirical covariance function (dashed curves) and of the fitted space-time covariance function (solid curves)	49
606			
607	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
608			
609			
610			
611			
612			
613			
614			
615			
616			
617			
618			
619			
620			
621			

622	11	Predictions for the period between 24 October 2006 and 31	
623		October 2006 (a, b, c), and between 6 March 2006 and 14	
624		March 2006 (d, e, f): (a, d) plot the seasonal prediction, (b,	
625		e) the model prediction with the Kalman filter, and (c, f) the	
626		observations.	51
627	12	Spatial averages over the validation period of: (a) the vari-	
628		ances of the prediction error resulting from by the reduced-	
629		order Kalman filter, and (b) the RMS errors.	52
630	13	Diagnostic statistics of the filter: (a) increment distribution in	
631		space averaged over the validation period, (b) cumulative av-	
632		eraged innovation over time, (c) distribution of the innovation	
633		values for the validation period.	53
634	14	Average correlation maps over the validation period between	
635		the observations and (a) the seasonal forecast, (b) the CHL	
636		forecasts (including the seasonal component), and (c) the filter	
637		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).