Data-Driven and Dynamics-Driven Approaches for Forecasting the Red Sea Chlorophyll

Thesis by

Denis Dreano

In Partial Fulfillment of the Requirements

For the Degree of

Doctor of Philosophy

King Abdullah University of Science and Technology, Thuwal,

Kingdom of Saudi Arabia

Insert Date (Month, Year)

The thesis of Your Full Name is approved by the examination committee

Committee Chairperson: Your advisor's name

Committee Member: Second name

Committee Member: Third name

TABLE OF CONTENTS

Introduction and Motivation

Phytoplankton and the Red Sea Ecology: Significance, Large-Scale Features, and Applications

The Importance of Phytoplankton

Phytoplankton are unicellular, free-floating, photosynthetic algae that live in the upper layers of bodies of water (ocean, lakes, rivers or ponds). There exists a wide diversity of phytoplankton species. Up to date, about 5000 of them have been identified (*Tett and Barton*, 1995). Phytoplankton are also highly variable in sizes, ranging from 0.2m for cyanobacteria to 200m for the largest species of diatom (*Pal*, 2014). In the oceans, phytoplankton live in the surface layer where there is enough sunlight for photosynthesis.

Phytoplankton play a fundamental role for the ocean ecology. They are at the basis of the marine food web and trap most of the energy used by pelagic ecosystems (*Pal*, 2014). Zooplankton graze phytoplankton, which are in turn consumed by higher trophic levels. It has been estimated that nearly 98% of the ocean primary productivity comes from phytoplankton (*Pal*, 2014). Phytoplankton are also responsible for maintaining the dissolved oxygen level necessary for other species to survive. However, high phytoplankton concentration may also impact their environment by

creating dead zones. When they die and sink, the bacteria that decompose them can consume all the available oxygen (*Pal*, 2014), and this may cause massive mortality in the fauna. Because of its short life cycle, phytoplankton respond very well to changes in its environment, making it a key parameter to monitor water quality (*Wu et al.*, 2014).

Phytoplankton place at the bottom of the marine food chain makes it an important factor for fisheries. Productive fishing zones such as the regions in the Arabian seas, Californian coast, north-west African coast and Chilean coast, are due to the upwelling of cold nutrient rich water favourable to phytoplankton growth. As such, remotely-sensed chlorophyll data have been routinely used since the last decade to help fisheries predict the timing of phytoplankton blooms (*Robinson*, 2010). On the other hand, the El-Nino phenomenon creates less favourable conditions for phytoplankton in the Eastern Pacific, resulting in a dramatic reduction of fish catches of fisheries in the western coast of South America (*Robinson*, 2010). In contrast, in the Red Sea, the MEI (Multivariate ENSO Index) has been found to positively correlate with chlorophyll concentration, a fact that could be of importance for regional fisheries (*Raitsos et al.*, 2015).

Phytoplankton also plays the role of a biological CO2 pump and strongly impact the Earth climate. During photosynthesis, phytoplankton captures carbon and releases oxygen. A part of this organic material stays in the food web, either transmitted to higher trophic level, or degraded by bacteria. Another part, sinks to the bottom of the ocean to form sediments. It is estimated that phytoplankton accounts for 48% of Earth carbon fixation (*Pal*, 2014).

Red Sea Large-Scale Phytoplankton Dynamics

Typical tropical seas (TTS), like the Red Sea, are characterized by a highly stratified structure, where warm nutrient-depleted surface water is separated from the cold nutrient-rich deep water by a steep gradient of temperature zone called pycnocline. The pycnocline acts as barrier that limits the upward nutrients flow (*Mann and Lazier*, 2006). As a result, TTS are oligotrophic and have low chlorophyll concentrations. Until recently, marine biologists believed that tropical and subtropical seas have therefore a very low productivity. However, recent investigations have contested this idea, suggesting that different upwelling mechanisms (winter deep mixing, storms, eddies, etc) exist, which bring new nutrients to the surface water (*Mann and Lazier*, 2006).

Despite being an oligotrophic and challenging environment for marine life, the Red Sea presents a surprisingly rich and diverse ecosystem (*Raitsos et al.*, 2011), and a very well developed coral reef system (*Racault et al.*, 2015). The source of nutrient for sustaining such a developed ecosystem is not well understood yet, but the exchange with the open ocean, the atmospheric depositions and transport through the mesoscale eddies are believed to play an important role (*Raitsos et al.*, 2013; *Zhan et al.*, 2014).

Although the Red Sea environment is still relatively well preserved, it is increasingly stressed by human activities. The continuous urbanization and fishing activity contribute to the fragilization of this unique ecosystem (*Acker et al.*, 2008). An abrupt increase of temperature has further occurred in the last decade, which may threaten the fragile coral reef system (*Raitsos et al.*, 2011).

Because of the lack of in-situ data, the large-scale phytoplankton dynamics of the Red Sea remain largely unknown (*Raitsos et al.*, 2013; *Triantafyllou et al.*, 2014).

However, in recent studies, remotely-sensed data and computer simulations have been used to improve our knowledge of the ecology of this region. The Red Sea is deficient in major nutrients (*Weikert*, 1987), and the only significant input of water comes from the Gulf of Aden (*Yao and Hoteit*, 2015). This explains a general pattern of chlorophyll concentration increase from north to south (*Raitsos et al.*, 2013), with the lowest concentration found in the northern central Red Sea. The Red Sea also displays a distinct seasonality, with a peak in concentration during the winter. A weak summer peak is also observed around July, everywhere except in the northernmost region (*Raitsos et al.*, 2013). Despite this regularity, a strong interannual variability is observed, with blooms that can reach mesotrophic concentration levels (*Raitsos et al.*, 2013). According to *Triantafyllou et al.* (2014), the variations in the Red Sea ecology are mainly driven by physical circulation. In the rest of this section, we explore some of the mechanisms that have been linked to the major features of chlorophyll concentration.

The exchange of water with the nutrient-rich Gulf of Aden is a major driving mechanism for the whole Red Sea (*Triantafyllou et al.*, 2014). It is the most important source of nutrients. The maximum chlorophyll concentration observed in the southern Red Sea during winter is attributed to wind-driven water intrusion (*Raitsos et al.*, 2013). In Summer, this exchange of water is believed to be the only significant source of nutrients for the whole Red Sea. The influence of the water intrusion weakens as the latitude increases, explaining the low concentration in the northern half of the Red Sea (*Raitsos et al.*, 2013).

Deep convection also plays an important role in allowing nutrient-rich deep water to mix with water of the euphotic zone. Vertical mixing is the most vigorous in the northern extremity of the Red Sea during the winter. This explains its higher chlorophyll concentration compared to the north-central Red Sea, a region of weak

mixing (*Raitsos et al.*, 2013). The northern Red Sea mixing is believed to be driven by wind (*Raitsos et al.*, 2013).

The Red Sea circulation is strongly influenced by mesoscale eddies (*Yao et al.*, 2014a,b; *Zhan et al.*, 2014) that could impact primary production (*Zhai and Bower*, 2013). In particular, the anti-cyclonic eddy in the central Red Sea is believed to control the June concentration peak and the summer productivity of this region, by transporting nutrients and/or phytoplankton from the adjacent coral reefs (*Raitsos et al.*, 2013). In the northern Red Sea, a cold-core eddy plays a role in enhancing the vertical mixing in this region (*Raitsos et al.*, 2013).

Aerial depositions of dust could also be an important input of nutrients for the Red Sea, but it has been largely left unexplored (*Triantafyllou et al.*, 2014). *Raitsos et al.* (2013) noticed for example that sand storms in the Red Sea most frequently happen in June and July, which coincides with the summer chlorophyll peak. Finally, climate mode indices have been shown to be strongly correlated with air-sea heat exchanges in the Red Sea (*Abualnaja et al.*, 2015), and might therefore influence its biology. This has been recently confirmed by *Raitsos et al.* (2015), who have shown that El Nino has a positive impact on the chlorophyll concentration, by strenghtening the wind transporting nutrients into the Red Sea from the Gulf of Aden.

Remotely-Sensed Chlorophyll Data: Relevance and Challenges for the Red Sea

Measuring Chlorophyll Concentration

Chlorophyll is a molecule present in algae, phytoplankton and plants that is critical for photosynthesis. It is a poor absorber of green light, and is responsible for the

coloration of plants (*Pal*, 2014). When phytoplankton are present in high concentrations, the water also takes a detectable green coloration (it can also take a red or blue coloration depending on the type of dominating phytoplankton) (*Robinson*, 2010). This offers an efficient way to observe the phytoplankton concentration from space.

In-situ measurement of chlorophyll concentration can be gathered through scientific cruises, buoy stations or gliders (unmanned submarines). These methods are expensive to deploy and therefore generally have limited temporal and spatial coverage (*Robinson*, 2010). Political issues, as in the Red Sea, as well as security issues, as in the Arabian Sea, set also barriers to in-situ measurements.

Satellite measurements of chlorophyll provide excellent proxies for phytoplankton concentrations with a good temporal and spatial coverage (*Robinson*, 2010). The SeaWIFS, MODIS and MERIS missions have provided an uninterrupted coverage of the world since 1997. High-resolution maps of daily chlorophyll concentration are freely accessible to the scientific community (*McClain*, 2009). Despite some limitations, like missing data due to cloud coverage and sunglint, or problematic values in coastal areas, remotely-sensed chlorophyll concentrations are intensively used by the scientific community. In regions, like in the Red Sea, where little in-situ measurements are available (*Raitsos et al.*, 2013; *Brewin et al.*, 2013), these constitute the most important data source.

Limitation of Remotely-Sensed Chlorophyll Data in the Red Sea

The quality of remotely-sensed chlorophyll data products such as MODIS and SeaWiFS in the Red Sea is comparable with that of the rest of the world for case I waters (open sea) (*Brewin et al.*, 2013). However, the data contains a large amount of miss-

ing values because of persistent clouds, sun-glint and sensor saturation (*Racault et al.*, 2015). This problem is particularly acute during the summer in the southern Red Sea where the data coverage is almost null (*Racault et al.*, 2015).

Chlorophyll concentration estimation in optically complex case II waters is a recurrent problem in this remotely-sensed data that particularly affects the southern Red Sea. In this region, the remotely sensed chlorophyll data could be overestimated (*Raitsos et al.*, 2013). However, all high values are not necessarily bad, as highly productive coral reefs are also present in this region (*Raitsos et al.*, 2013). However, these values have not been validated yet, due to the lack of in situ measurements (*Raitsos et al.*, 2013).

One solution to missing and bad values is to use a data filling algorithm, of which one of the most popular is the Date INterpolating Empirical Orthogonal Functions (DINEOF). It is an EOF based data filling approach introduced by *Beckers and Rixen* (2003). In *Sirjacobs et al.* (2011), it has been employed to fill chlorophyll data with 70% of missing values. *Taylor et al.* (2013) has compared DINEOF with other EOF-based reconstruction algorithms, suggesting that the former is the best method for data filling. DINEOF has been employed in several other chlorophyll studies (*Miles and He*, 2010; *Waite and Mueter*, 2013). It has also been used for multivariate reconstruction of SST fields using chlorophyll data in *Alvera-Azcarate et al.* (2007).

The OC-CCI is a new chlorophyll data product that considerably increases the Red Sea coverage. It merges the data from sensors SeaWiFS, MODIS and MERIS. Overall, it achieves a 75-80% coverage in the entire Red Sea basin against 50-65% for a single sensor (*Racault et al.*, 2015). This is mostly due to the use of the POLYMER algorithm (*Steinmetz et al.*, 2011) that allows to exploit MERIS data collected during hazy conditions. However, this new dataset has not been fully explored to revisit the assumptions made on the large-scale Red Sea phytoplankton productivity.

Modeling and Forecasting Chlorophyll: Data-Driven and Physics-Driven Approaches, and Applications

Why Modeling Chlorophyll?

Models could be useful to identify causes behind the chlorophyll patterns we observe in the Red Sea. Many hypotheses have been made about the drivers of chlorophyll concentration in this regions, but some of them have not been yet investigated through models. The role played by the exchange of water with the Gulf of Aden and winter overturning in the northern Red Sea have been successfully modeled with a 3D coupled ecological model (*Triantafyllou et al.*, 2014). However, the interaction between the open sea and coral reefs, and the role of atmospheric depositions have not been investigated yet. Models, can also be helpful for understanding governing dynamics affecting the chlorophyll concentration. In particular, the interaction between the productivity level of the different regions of the Red Sea is yet to be explored.

Model predictions of chlorophyll concentration also have practical applications. Phytoplankton blooms can be harmful to humans and marine life and are closely monitored in many regions of the world (*Pettersson and Pozdniakov*, 2013). In the Red Sea, where tourism and aquaculture are developing, it is likely to become a concern too. Phytoplankton is also directly, and indirectly through zooplankton, the cause of microfouling that affects desalination plants. In 2008-2009, a red tide forced the shutdown of desalination plants along the Gulf of Oman and the Arabian Gulf (*Richlen et al.*, 2010).

Deterministic Models

Ecological Models

There is a rich literature on the modeling of marine ecosystem using differential equations (see *Fennel and Neumann* (2004) for an introduction In these models the interactions of complex physical, chemical and biological processes are modeled by differential equations that represent the flow of carbon, nitrogen, phosphate and silicon. The biota is divided into trophic levels, and can be further divided by feeding methods and size classes (*Triantafyllou et al.*, 2014).

Ecological deterministic models in use vary widely in diversity depending on on the number of state parameters and interactions represented. They can be as simple as the nutrient-phytoplankton-zooplankton (NPZ) model (*Anderson*, 2005) that only has three variables representing two trophic levels and nitrate, or as complex as the European regional seas ecosystem model (ERSEM) that has dozens of variables (*Baretta et al.*, 1995). NPZ models are extensively used because of its simplicity and its capacity to model the the large-scale features of marine ecosystems (*Anderson*, 2005). ERSEM has been use in many studies. It has recently been coupled to the MITgcm circulation model used to simulate the Red Sea ecology (*Triantafyllou et al.*, 2014). However the complexity of these models makes them difficult to parametrize if not enough data are available, which is usually the case (*Anderson*, 2005).

Data Assimilation

Data assimilation is used to improve the simulations of ecological dynamical models and enhance their forecasting capabilities by constraining their predictions with available observations *Edwards et al.* (2015). Such prediction capabilities are deployed in operational expert systems, for example to study the impact of human activities on

the ecosystem of the Gulf of Pagasitikos (*Korres et al.*, 2012). The deployment of a similar forecasting system in the Red Sea is currently under development (*Triantafyllou et al.*, 2014). Hindcasting, the estimation of unobserved variables, is another application of assimilation scheme. *Ciavatta et al.* (2011) showed that they could improve the seasonal and annual hindcast of non-assimilated biogeochemical properties in the shelf area of Western English Channel. Data assimilation can also be used for reanalysis, to provide estimates of past years biogeochemical variables (*Fontana et al.*, 2013).

In the marine ecology modeling community, three assimilation schemes have been widely used: the Ensemble Kalman filters (EnKF), the Singular Evolutive Extended Kalman filter (SEEK), and its ensemble variant, the Singular Evolutive Interpolated Kalman filter (SEIK). The stochastic EnKF, a Monte-Carlo approximation of the Kalman Filter, has been used in *Ciavatta et al.* (2011, 2014). This scheme may however suffer from sampling errors when the ensemble size is smaller than the number of observations, as is usually the case when assimilating remotely-sensed data. SEEK is a reduced-rank variant of the Extended Kalman filter (EK). It was introduced to run efficiently when the state dimension is very large, as is the case in ocean applications. It is based on the projection of the error covariance onto a low dimensional space. SEEK has a long history in data assimilation for marine ecology models and is still extensively used in recent studies (*Fontana et al.*, 2013; *Korres et al.*, 2012; *Butenschon and Zavatarelli*, 2012). SEIK is an ensemble variant of the SEEK and a deterministic version of the EnKF that do not suffer from observations sampling errors, as it updates the filter and forecast exactly as in the Kalman filter, but requires a resampling step to generate a new ensemble for the next forecast step. SEIK has been used by (*Triantafyllou et al.*, 2013; *Korres et al.*, 2012). *Korres et al.* (2012) shows that SEIK and SEEK are both comparably robust methods for highly nonlinear

systems. *Hoteit et al.* (2005) has shown that SEIK outperforms SEEK when using a high-resolution non linear model.

Ecological models are challenging applications for state of the art data assimilation schemes (*Edwards et al.*, 2015). First, biogeochemical variables are usually positive concentration, whereas Kalman filters expect Gaussian variables, and log-transformation may fail at avoiding this issue (*Ciavatta et al.*, 2011). In an attempt to mitigate this problem, *Fontana et al.* (2013) has introduced Gaussian anamorphosis transformations. Second, ecological blooms are intermittent and highly nonlinear, conditions that are challenging for Kalman filter-based assimilation schemes (*Triantafyllou et al.*, 2013; *Korres et al.*, 2012). Third, SEIK, EnKF and SEEK project the error covariance onto some subspace, resulting in an underestimation of the estimation error. *Butenschon and Zavatarelli* (2012) studied different ways to propagate the error covariance in order to alleviate this issue. Finally, the model error statistics are required by Kalman-derived filters, but are difficult to estimate. *Triantafyllou et al.* (2013) proposed to use the H_∞ method with SEIK in order alleviate this requirement.

Particle filters represent a class of data assimilation scheme that are not derived from the Kalman filter, and do not make any linearity or Gaussianity assumption (*Edwards et al.*, 2015). They have been studied in the case of 0D and 1D ecological models (*Edwards et al.*, 2015). Their application to 3D model is an active field of research (*Edwards et al.*, 2015).

Data-Driven Approaches

Compared to data assimilative ecological models, data-driven statistical models are relatively simpler to apply. They are relevant when the phenomenon producing the

data is dynamically complex to model, or simply poorly understood (*James et al.*, 2013). They have been applied to predict chlorophyll concentration, mostly in small regions with complex dynamics (see references below). Some statistical models, such as linear regression, Gaussian additive models, or tree regression have the advantage of being easy to interpret (*James et al.*, 2013), and could be used to understand the dynamics driving the chlorophyll concentration (*Raitsos et al.*, 2012).

Statistical Models

Statistical and machine learning models have been used for estimation and classification problems related to phytoplankton concentrations. One application is the detection of harmful algal bloom from spatio-temporal satellite dataset, that has been addressed by *Gokaraju et al.* (2011) in the Gulf of Mexico using support vector machines. Another application is the estimation of chlorophyll concentration in case II coastal water using satellite radiance data. This problem has been considered by *Kim et al.* (2014) on the west coast of South Korea, and by *Camps-Valls et al.* (2006) using a global dataset of in situ measurements. The former used the support vector regression algorithm, while the latter used the random forest algorithm.

Machine learning algorithms, in particular Artificial Neural Networks, have been very popular for forecasting regional chlorophyll concentration in regions with very complex dynamics, where deterministic ecological models usually perform poorly. Neural networks have been widely used for forecasting chlorophyll concentration in fresh as well as in coastal water systems. In *Jeong et al.* (2006), temporal recurrent recursive neural network have been used and found superior to traditional time-series models for daily forecasts of chlorophyll concentration. *Wang and Yang* (2013) also used recurrent neural networks for daily chlorophyll forecasting in Lake Taihu, China. *Mulia et al.* (2013) combined Neural Network and genetic algorithm for nowcasting

and forecasting of the chlorophyll concentration up to 14 days ahead, in the tidal dominated coast of Singapore. Finally, *Lee et al.* (2003) used neural networks for the forecasting of algal bloom with one or two weeks lags in the coastal waters of Hong-Kong.

Geostatistics

Phenomena such as propagation and diffusion play a key role in the chlorophyll spatial concentration, but are difficult to represent without spatial modeling. There is also a difference in the chlorophyll patterns of different regions of the Red Sea, in particular between the nutrient rich southern Red Sea and the oligotrophic northern Red Sea, and between the open ocean and the coastal waters (*Raitsos et al.*, 2013). One should also expect the different regions of the Red Sea to interact.

Contrary to the statistical models presented abose geostatistical models are a natural way to model spatial data. They model spatial data as the realization of a two-dimensional Gaussian process Geostatistics can be easily extended to spatio-temporal datasets. Many flexible ways of constructing space-time covariance functions for these models have been recently proposed (*Gneiting*, 2002; *Cressie and Huang*, 1999; *Stein*, 2005). Space-time geostatistics have been applied in many environmental studies, but not yet to chlorophyll data.

The theory of space-time geostatistics is closely related to that of spatial statistics. In fact, the time dimension is an additional dimension. However, the relationship between these two dimensions derive from a dynamical process, that must be taken into account in the definition of the covariance funciton (*Gneiting and Guttorp*, 2010). Some space-time covariance models can actually be derived from a physical formulation, such as the frozen fields (*Gneiting and Guttorp*, 2010), or stochastic differential equations (*Brown et al.*, 2000; *North et al.*, 2011).

Physically-derived space-time covariance functions have been little used (*Gneiting and Guttorp*, 2010). More popular, are covariance functions built from simple building blocks. One of the most simple types are separable covariance functions, that are the products of a spatial covariance function and temporal covariance function. They are computationally efficient, but are not suitable to represent space-time interactions (*Cressie and Huang*, 1999; *Stein*, 2005), making them of limited use for modeling physical systems. The Cressie, Huang spectral characterization theorem of space-time covariance functions has opened the door to wider ways of constructing them. For example, *Gneiting* (2002) presented a simple criterion that allows their construction from a very large class of models.

Space-time geostatistical models have been used in a variety of applications. *Hohn et al.* (1993) used it for forecasting the outbreaks of an invasive specie. These methods have also been used in meteorology to model temperature fields (*Handcock and Wallis*, 1994; *North et al.*, 2011) or wind (*Cressie and Huang*, 1999; *Gneiting*, 2002), and in environmental studies for ground-level ozone concentration. *Gneiting et al.* (2007); *Gneiting and Guttorp* (2010) present recent more details on the theory of space-time geostatistics and its applications.

Thesis Objectives

The goal of this thesis is to develop novel models for chlorophyll forecasting in the Red Sea. Two approaches will be considered: a data-driven driven approach that derives a model from data using statistical models, and a dynamical approach that derives a model from first principles. The merits of both approaches will be compared for the first time in the same region using the same data. We will also propose ways to combine both methods for improving chlorophyll forecasting.

There is a crucial need for improving the modeling of large-scale features of marine ecosystems. The ocean physical, chemical and ecological processes are complex and poorly understood. Even though ocean color remote-sensing data has revolutionized our understanding of marine ecology, it only gives information about the phytoplankton at the sea surface. Our knowledge is even more limited in the Red Sea, due to the paucity of in situ data.

We will develop efficient approaches to model the Red Sea chlorophyll, explore the possibility of improving the state of the art data assimilative ecosystem modeling, and introduce the use of geostatistics to the field of marine ecology.

Currently, both deterministic and statistical approaches have been used to predict chlorophyll. However, these approaches have never been compared, nor combined in the same problem. A thorough comparison is very much needed. It could help modelers to have a foundation on which to base their modeling strategy. Combining them may improve the forecasting skill for chlorophyll.

This thesis will also help increase our understanding of the Red Sea ecology. In particular, it will identify possibles drivers for the chlorophyll seasonality and interannual variability. We will also identify and characterize the different Red Sea eco-regions. This work has practical applications for the region. In particular, better forecasting phytoplankton blooms may be extremely helpful of the operation of desalination plants.

Research Plan

The research plan including the different chapters of the thesis is presented hereafter. To tackle the objectives, I have organized the thesis in six chapters composed of the first four chapters in which the data is collected and analyzed (chapter 1), and used to construct statistical (chapter 2) and geostatistical models (chapter 3 and 4) that are used for chlorophyll forecasting. Data assimilative ecological models are then introduced in chapter 5. Innovative methods for enhancing the state of the art ensemble methods will be explored in chapter 6.

For each chapter, I will first explain its importance for reaching the thesis objectives. Then, a list of open scientific questions will be presented, that the chapter will contribute to. An outline of the methodology will be introduce. Finally a list of outcomes, will be given against which the chapter can be evaluated.

Chapter 1 presents the dataset that will be used in the following chapters. In chapter 2, we will divide the Red Sea into eco-regions using clustering algorithms. Then, we will fit statistical models for forecasting chlorophyll in these regions. In chapter 3, we will construct a spatio-temporal geostatistical model for the global Red Sea chlorophyll, and use it for forecasting. In chapter 4, this geostatistical model will be refined by using several regional geostatistical models. In chapter 5, chlorophyll will be forecasted using assimilated 1D regional deterministic ecological models, and the results will be compared to the statistical models. Chapter 6, will develop a

method to combine statistical and deterministic models for improving chlorophyll forecasting in the Red Sea.

Chapter 1: Dataset Building and Exploration

A preliminary task to data modeling, is the gathering, cleaning and exploration of the data. Given the complexity and the size (40 GB) of the dataset, this is not an easy task. This first data analysis, will reveal if enough data has been gathered to make meaningful forecast, and what accuracy we can expect from the models. This step will also provide information that will help in designing statistical models: most significant variables, differences between regions, relevant data transformation, etc. Finally, this step will identify patterns in the data that will be useful to qualitatively evaluate predictive models.

Open Questions

- Can we efficiently identify outliers in the chlorophyll values?
- Is there a way to efficiently fill the missing values in the chlorophyll dataset?
- Can the data help understanding the mechanisms behind extreme blooms in the Red Sea?
- Will the data support the hypothesizes about the dynamics behind the chlorophyll seasonal cycle in the Red Sea?

Methods

1. Identify data sources and gather the data.
2. Clean the data and fill in missing values (DINEOF).
3. Align and format the data to build a cleaned dataset.
4. Explore the dataset.

- Study the correlation between chlorophyll and other variables (Linear Regression, GAM, data transformations) in order to know which ones will be important in the model.
- Do variable selection (Lasso, single variable regression, multistep regression): as too many variables may introduce noise in the models or make the fitting slower.
- Study the regional aggregation (ACF), to characterize the clusters.
- Explore spatiotemporal correlations (hovmoller plots, PCA, variograms) to get an understanding on the spatial features of the data.

Expected Outcomes

- A cleaned dataset that can be used in the following tasks.
- A comprehensive exploration of the available data to study chlorophyll variability in the Red Sea.
- A preliminary variable selection.
- A clear picture of the major spatio-temporal patterns in the data.
- A critical evaluation of the current hypotheses about the chlorophyll dynamics in the Red Sea.

Chapter 2: Forecasting Chlorophyll Concentration in Regional Aggregates

The complexity of marine ecosystem is reflected in that of chlorophyll data. It is therefore useful to first simplify it by aggregating it spatially. The space-time dynamics of the chlorophyll data reflects the highly nonlinear dynamics of the underlying physical, chemical and biological phenomenons. As shown by the north-south gradient and the seasonal behavior, the resulting space-time process is nonstationary in time and in space. The high-dimensionality in space can be reduced by considering a regional aggregation of the data. This would allow us to focus on the global scale phenomenons: such as the interactions between neighboring regions, the time-scale of large events and the difference in the physical variables affecting the chlorophyll concentration in each region. In the following tasks, these simple predictive models will also be a reference for evaluating more complex ones.

Open Questions

- Is the biological aggregation of the Red Sea proposed in *Raitsos et al.* (2013) statistically meaningful?

- Can clustering methods be used to identify marine ecological zones based on chlorophyll data?

- Would a simple forecasting model allow us to identify the causes of chlorophyll blooms?

Methods

1. Define the datasets on which the models will be trained and tested, as well as the cross-validation methodology.

2. Apply unsupervised learning algorithms to the dataset to divide the Red Sea into clusters corresponding to different environmental conditions.

3. Build regional statistical models to forecast chlorophyll concentration.

4. Evaluate the models on the prediction of extreme blooms.

Expected Outcomes

- A division of the Red Sea into environmentaly distinct regions that has been quantitatively evaluated.
- A critical evaluation of current hypotheses about the chlorophyll dynamics in the Red Sea.
- A lower bound for the performance of a more sophisticated model.
- An assessment of the limitation of aggregate methods for Chlorophyll data.
- An understanding of how the treatment of spatial correlations may improve cholorphyll forecasting.

Chapter 3: Global Geostatistical Model for Chlorophyll Forecasting

Geostatistical methods can be used to construct dynamical models for forecasting the chlorophyll concentrations. Geostatistics is a robust method to model spatial data. Recently there has been increasing interest for expanding it to model spatio-temporal data. Through the use of Kriging interpolation, these models provide powerful tools for spatio-temporal prediction. As a particular case of Kriging, by predicting the spatial future field given the observation of the present field, we can derive a linear dynamical model for forecasting the data. This linear model may be employed in a filtering setting like the Kalman filter. This is a desirable framework, and is similar to the way deterministic models are employed to make forecasts given past observations.

Open Questions

- Can we construct a global geostatistical model that fits well enought chlorophyll data in the Red Sea?
- How non stationary the chlorophyll data is in time and space?
- Which spatio-temporal covariance functions best fit the chlorophyll data?
- Can geostatistical methods be employed in a filtering setup?

Method

Most of the work proposed in this task has already been accomplished and has been the object of a submission for publication. It remains to compare this model to the models from the previous chapters.

Expected Outcomes

- A new methodology to employ a geostatistical model in a filtering setup.
- A characterization of the space-time non stationarity of the data.
- A framework for using spatial aggregation and geostatistical models in the same model.

Chapter 4: Local Geostatistical Model for Chlorophyll Forecasting

This part will bring together the results of the two preceding tasks to develop a geostatistical model that takes into account the large-scale dynamics and the regional spatio-temporal dynamics. In task 2, a predictive model is built, that represents the large scales behaviour of the Red Sea, but the spatial dimension inside each region is not addressed. We expect local features to play a role, such as the proximity to the coast, the bathymetry, proximity to other regions or major cities, etc. In task 3, we proposed a methodology to use a geostatistical model in a dynamic fashion to do pixel-scale forecast. In this task, each regions will be modeled separately by a local geostatistical model that can provide local prediction. These models will have access to aggregate covariates from neighboring regions to represent the global scale behaviours.

Open Questions

- What are the most adapted space-time covariance models for chlorophyll data?
- How to use global covariates in a geostatistical model?
- What are the differences in the spatio-temporal correlation of chlorophyll in each region?
- Can the regional behaviour of phytoplankton be accurately predicted?
- What are the spatial features that are important for regional chlorophyll dynamics in the Red Sea?

Methods

- Extract local datasets from previous tasks on which we will train regional geostatistical models.
- Design the training and test datasets, and the cross-validation method, as in chapter 2.
- Estimate the chlorophyll mean functions for each region in function of the past covariates, to make the model useful for forecasting.
- Fit local geostatistical models to the residuals, to model them as space-time Gaussian processes.
- Evaluate the model predictions and compare the results with chapters 2 and 3.

Expected Outcomes

- A methodology to aggregate local geostatistical models.
- An improvements in the prediction skills over the models of chapters 2 and 3.
- An understanding of the differences between the regions that have been identified in chapter 2.
- A critical evaluation of the space-time covariance models for fitting chlorophyll data.
- A better characterization of the regional chlorophyll dynamics.

Chapter 5: Assimilation of Regional 1D Ecological Models and Comparison to Statistical Models

The three previous chapters focus on constructing increasingly sophisticated statistical predictive models for the chlorophyll concentration in the Red Sea. In this chapter these models will be evaluated against a 1D ecological model (ERSEM). The goal of this part will be to identify the merits of each modeling approach, and propose ways in which they can complement each other. The models will be run in each of the regions found in chapter 2. Available data will also be assimilated to the model through an ensemble Kalman-based smoothing assimilation scheme that we will further equip with an efficient expectation-maximization algorithm for parameters estimation.

Open Questions

- Are statistical methods competitive for forecasting chlorophyll concentrations?
- How may statistical and deterministic models complement each other?
- May statistical methods forecast interesting dynamical features?
- Are there significant regional differences in the relative performances of both approaches?
- How best to estimate the parameters of ecological models?

Methods

1. Calibrate the ERSEM model on each of the regions to have useful regional models running.

2. Define an assimilation scheme and the data that will be assimilated by each regional model.

3. Implement the assimilation scheme.

4. Run the simulation and aggregate the results to allow for comparison with the statistical models introduced in the previous chapters.

5. Define the metrics for comparison between statistical and deterministic models.

6. Conduct the comparison with the statistical models and summarize the results.

Expected Outcomes

- A complete set of measures of the prediction skills of each approach.
- A method to estimate parameters in an assimilative ecological model.
- A set of case studies of the behaviours of each method for forecasting blooming events.
- An understanding of the limitations of geostatistical models to predict nonlinear dynamics.
- Propositions on how the statistical and dynamical approaches can complement each other.

Chapter 6: Combining Statistical and Data Assimilative Predictive Models for Improved Chlorophyll Forecasting

In the previous chapters, we proposed to evaluate the forecasts of the ecological ERSEM model with that of the statistical models we proposed from chapters 2 to 4. In this chapter we will study how these two approaches can be complementary. Specifically, we will study the use of statistical forecasts model to improve the forecasts of the ERSEM ecological model. The forecasts of the statistical models will be treated as observations, that can be assimilated by the filtering scheme used with the ERSEM model, which will hopefully lead to improved forecasts. On the other hand, real observations will be assimilated sequentially when they become available. This method will allow the different ERSEM models on each cluster to communicate indirectly their states to one another.

Open Questions

- Can statistical predictive models be used to communicate information between deterministic models?
- Would the access to information about other regions improve the model forecasts?
- What are the global patterns of ecological dynamics in the Red Sea?

Methods

1. Develop a new assimilation scheme to assimilate statistical predictions and improve the 1D regional deterministic models.

2. Define metrics to measure model improvement over the results of chapter 5.

3. Prepare training and test datasets, as in previous chapters.

4. Train statistical models to predict chlorophyll as in chapter 2.

5. Run simulations with assimilation of statistical and real observations.

6. Compare with results of chapter 5 to see if the method improves over the deterministic model alone.

Expected Outcomes

- An improvement in the prediction skills of the dynamically-driven approach.
- A methodology to couple assimilative dynamically-driven regional ecological models through statistical models.
- Insights on the global ecological dynamics of the Red Sea.

Summary of Progress

Data Preparation

All of the datasets necessary for the analysis has been gathered. Almost all of it has been downloaded from public sources on the Internet, using Shell and R scripts. The wind dataset has been provided by Yesubabu Viswanadhapalli, and is the result of the assimilation of QuickScat satellite and in situ data to the Weather Research and Forecasting (WRF) regional Red Sea model. There are additional datasets that will be useful for the analysis, but they are mainly climate mode time-series like IMI and EAWR, that are easy to download. The downloaded data is listed in table 3.1.

So far, the MODIS and CCI data have been cropped over the region of interest, cleaned and exported to the format TIFF, which can be read easily by most software, R in particular. Each dataset has then been aggregated in a single file in the native R raster format. Applying this processing to the remaining raster data should be straightforward. Then, the data will need to be aligned and aggregated on the same temporal and spatial resolution, before aggregating it in table format.

Variable	Description	Mission/Source	Unit	Temporal Resolution	Period	Spatial Resolution
CHL	Chlorophyll	CCI	mg/m^3	8-days	1997-2012	4km
SST	Sea Surface Temperature	MODIS	Celsius	8-days	2002-2015	4km
AOD	Aerosol Optical Depth	MODIS	N/A	daily	2000-2015	28km
CDOM	Colored Dissolved Organic Matter	MODIS	N/A	8-days	2002-2015	4km
NAO	North Atlantic Oscillation Index	NOAA	N/A	daily	1979-2015	N/A
PAR	Photosynthetically Available Radiations	MODIS	$Einstein/m^2.Day$	8-days	2002-2015	4km
POC	Particulate Organic Carbon	MODIS	mg/m^3	8-days	2002-2015	4km
RAIN	Precipitations	TRMM	mm/h	daily	1998-2015	28km
SOI	Southern Oscillation Index	Australian Bureau of Meteorology	N/A	daily	1999-2015	N/A
SLA	Sea Level Anomaly	Aviso	cm	daily	1993-2013	28km
WIND	Vector Wind Field	Simulated	m/s	3-hourly	2000-2014	10km

Table 3.1: Downloaded data

Red Sea Chlorophyll Data Analysis

The MODIS and CCI chlorophyll data products have been explored and compared. The MODIS data include an important amount of missing data which may reach 100% during summer in the southern Red Sea, increasing the complexity of any analysis in this region. The CCI data product solves this issue by merging three sources of remotely-sensed chlorophyll data (MODIS, MERIS and SeaWiFS) and using a new algorithm for retrieving chlorophyll values when the cloud cover is important. This increases the coverage to 70% during summer months in the southern Red Sea.

The SeaWiFS chlorophyll data has been used in the submitted companion article of chapter 3 (see Appendice B). The seasonal signal in the data is strong and has been shown to account for 50% of the variability. The seasonal anomalies display a strong spatio-temporal correlation: the temporal correlation of anomalies is 40%, whereas two locations at 0.5 degrees apart are nearly 60% correlated. Not shown in this article, I also compared the SST and chlorophyll data, and found an important negative correlation. However, when looking at the anomalies, the correlation vanished, suggesting that the causes of seasonal and interannual variability are distinct.

Red Sea Regional Clustering

I used clustering algorithms in order to derive the Red Sea eco-regions. These were applied to monthly log-concentration of chlorophyll. I used SeaWiFS data, that has been filled using DINEOF. I used the popular K-means, and the Gaussian Mixture Model (GMM) clustering algorithms.

I found that GMM provides more robust results. With any number of clusters, we obtain a division of the Red Sea into regions of comparable sizes. With 5 clusters,

the regions are very similar to those identified by *Raitsos et al.* (2013). Contrary to the purely latitudinal division proposed by the former, we observe that the separation between clusters is curved at the position of major Red Sea eddies. The fact that the curvature is oriented toward the south suggests that most nutrients propagate northward from the Gulf of Aden.

In Chapter 2, I plan to use the dataset constructed in Chapter 1. By using CCI chlorophyll data instead of SeaWiFS, the need for data filling is minimized. This is desirable, as data filling can introduce biases. It will also be possible to use additional variables. For example, we can expect the temperature and the bathymetry to have a large impact on the Red Sea phytoplankton biology. Sea level anomaly can be useful in that it indicates the presence of mesoscale eddies. Finally, alternative clustering algorithms will be tested.

Global Geostatistical Model

The research described in chapter 3 of the thesis has mostly been done and, is the object of an article currently in review. The current version of the manuscript can be found in appendice. We include the abstract below:

> 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* (2002) 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 (RMS) prediction error by about 11% compared with the seasonal prediction.

Regional 1D Ecological Models

The 1D regional ecological models used for this thesis have been configured and are operational. Three models will be used: for the northern, central and southern Red Sea. The extreme south of the Red Sea is not modeled, as its dynamics is poorly understand and we miss in situ data. The ecology is modeled with ERSEM, and the hydrodynamics is modeled with the MITgcm.

The results of the MITgcm are those from *Yao et al.* (2014a,b), in which a simulation of the Red Sea and part of the Gulf of Aden circulation was run over 50 years. The NCEP data were used for atmospheric forcing, and the ocean ECCO data for the open boundary conditions in the Gulf of Aden. The output of the 50 years run are used for the temperature and vertical circulation at the modeled points.

ERSEM simulates the complete water column with the pelagic and benthic ecosystems, as well a their coupling. The equations model the flow of carbon, nitrogen, phosphorus and silicon in the ecosystem. Living organisms are modeled in terms of population processes (growth and mortality) and physiological processes (ingestion, respiration, excretion, and egestion). The biota is divided into functional groups according to their trophic levels: producers (phytoplankton), consumers (zooplankton) and decomposers (bacteria), and further subdivided according to their sizes (*Baretta*

et al., 1995).

The ecological models are initialized with the results of a 3D ecological simulation of the Red Sea (*Triantafyllou et al.*, 2014). The nutrient concentrations are initialized using values from the World Ocean Atlas 2005 (WOA 2005).

Data Assimilation into the 1D Ecological Models

The assimilation scheme for the ecological models has been implemented and is operational. The chosen scheme is the hybrid-SEIK, described in *Hamill and Snyder* (2000). It can be seen as a variant of the 3DVAR variational assimilation scheme. 3DVAR assumes that the error forecast covariance is fixed in time. In the case of the hybrid, the covariance is a linear combination of the 3DVAR covariance and the time-evolving SEIK covariance matrix.

The problem of optimal filtering can be solved exactly by the Kalman Filter for linear systems. For nonlinear models, one can use the Extended Kalman (EK) filter, in which the model is linearized by computing the error covariance function. However, when the state is large, as is often the case for oceanographic applications, the EK is intractable. In that case, SEEK can be used, where the error covariance function is projected into a smaller subspace. This subspace evolves to ensure that most of the error is represented and filtered out. SEIK can be viewed as an ensemble variant of the SEEK, where the error covariance function is represented exactly by an ensemble of states. This avoids the computation of model gradients, and allows the assimilation scheme to perform better when this model is strongly non-linear. SEIK has been shown to be efficient for large-scale 3D ecosystem simulations (*Triantafyllou et al.*, 2003).

The Expectation-Maximization scheme to estimate the filter parameters has also

been derived. It is similar to that proposed by *Tandeo et al.* (2014), except that the model is non linear. The scheme will be used to improve the estimates of the observation and model covariance errors.

REFERENCES

Abualnaja, Y., V. P. Papadopoulos, S. A. Josey, I. Hoteit, H. Kontoyiannis, and D. E. Raitsos (2015), Impacts of climate modes on air-sea heat exchange in the Red Sea, *Journal of Climate*, p. 150106132132005, doi:10.1175/jcli-d-14-00379.1, in press.

Acker, J., G. Leptoukh, S. Shen, T. Zhu, and S. Kempler (2008), Remotely-sensed chlorophyll a observations of the northern red sea indicate seasonal variability and influence of coastal reefs, *Journal of Marine Systems*, *69*(3-4), 191–204, doi:10.1016/j.jmarsys.2005.12.006.

Alvera-Azcarate, A., A. Barth, J. M. Beckers, and R. H. Weisberg (2007), Multivariate reconstruction of missing data in sea surface temperature, chlorophyll, and wind satellite fields, *Journal of Geophysical Research-Oceans*, *112*(C3), doi:ArtnC03008Doi10.1029/2006jc003660.

Anderson, T. R. (2005), Plankton functional type modelling: running before we can walk?, *Journal of Plankton Research*, doi:10.1093/plankt/fbi076.

Baretta, J. W., W. Ebenhh, and P. Ruardij (1995), The european regional seas ecosystem model, a complex marine ecosystem model, *Netherlands Journal of Sea Research*, *33*(3/4), 233–246, doi:10.1016/0077-7579(95)90047-0.

Beckers, J. M., and M. Rixen (2003), EOF calculations and data filling from incomplete oceanographic datasets, *Journal of Atmospheric and Oceanic Technology*, *20*(12), 1839–1856, doi:Doi10.1175/1520-0426(2003)020⟨1839:Ecadff⟩2.0.Co;2.

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

Brown, P. E., K. F. Karesen, G. O. Roberts, and S. Tonellato (2000), Blur-generated non-separable space-time models, *Journal of the Royal Statistical Society Series B-Statistical Methodology*, *62*, 847–860, doi:Doi10.1111/1467-9868.00269.

Butenschon, M., and M. Zavatarelli (2012), A comparison of different versions of the SEEK Filter for assimilation of biogeochemical data in numerical models of marine ecosystem dynamics, *Ocean Modelling*, *54-55*, 37–54, doi:Doi10.1016/J.Ocemod.2012.06.003.

Camps-Valls, G., L. Gomez-Chova, J. Munoz-Mari, J. Vila-Frances, J. Amoros-Lopez, and J. Calpe-Maravilla (2006), Retrieval of oceanic chlorophyll concentration with relevance vector machines, *Remote Sensing of Environment*, *105*(1), 23–33, doi:Doi10.1016/J.Rse.2006.06.004.

Ciavatta, S., R. Torres, S. Saux-Picart, and J. I. Allen (2011), Can ocean color assimilation improve biogeochemical hindcasts in shelf seas?, *Journal of Geophysical Research-Oceans*, *116*, doi:ArtnC12043Doi10.1029/2011jc007219.

Ciavatta, S., R. Torres, V. Martinez-Vicente, T. Smyth, G. Dall'Olmo, L. Polimene, and J. I. Allen (2014), Assimilation of remotely-sensed optical properties to improve marine biogeochemistry modelling, *Progress in Oceanography*, *127*, 74–95, doi:Doi10.1016/J.Pocean.2014.06.002.

Cressie, N., and H.-C. Huang (1999), Classes of nonseparable, spatio-temporal stationary covariance functions, *Journal of the American Statistical Association*, *94*(448), 1330–1339, doi:10.1080/01621459.1999.10473885.

Edwards, C. A., A. M. Moore, I. Hoteit, and B. D. Cornuelle (2015), Regional ocean data assimilation, *Ann Rev Mar Sci*, *7*, 21–42, doi:10.1146/annurev-marine-010814-015821.

Fennel, W., and T. Neumann (2004), *Introduction to the Modelling of Marine Ecosystems*.

Fontana, C., P. Brasseur, and J. M. Brankart (2013), Toward a multivariate reanalysis of the North Atlantic Ocean biogeochemistry during 1998-2006 based on the assimilation of SeaWiFS chlorophyll data, *Ocean Science*, *9*(1), 37–56, doi:Doi10.5194/Os-9-37-2013.

Gneiting, T. (2002), Nonseparable, stationary covariance functions for spacetime data, *Journal of the American Statistical Association*, *97*(458), 590–600, doi:10.1198/016214502760047113.

Gneiting, T., and P. Guttorp (2010), Continuous parameter spatio-temporal processes, *Handbook of Spatial Statistics*, *97*, 427–436.

Gneiting, T., M. G. Genton, and P. Guttorp (2007), Geostatistical space-time models, stationarity, separability, and full symmetry, *Statistical Methods for Spatio-Temporal Systems*, *107*, 151–175.

Gokaraju, B., S. S. Durbha, R. L. King, and N. H. Younan (2011), A machine learning based spatio-temporal data mining approach for detection of harmful algal blooms in the Gulf of Mexico, *Ieee Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *4*(3), 710–720, doi:Doi10.1109/Jstars.2010.2103927.

Hamill, T. M., and C. Snyder (2000), A hybrid ensemble kalman filter-3d variational analysis scheme, *Monthly Weather Review*, *128*(8), 2905–2919, doi:Doi10.1175/1520-0493(2000)128⟨2905:Ahekfv⟩2.0.Co;2.

Handcock, M. S., and J. R. Wallis (1994), An approach to statistical spatial-temporal modeling of meteorological fields, *Journal of the American Statistical Association*, *89*(426), 368–378.

Hohn, M. E., A. M. Liebhold, and L. S. Gribko (1993), Geostatistical model for forecasting spatial dynamics of defoliation caused by the gypsy moth (lepidoptera: Lymantriidae), *Environmental Entomology*, *22*(5), 1066–1075.

Hoteit, I., G. Korres, and G. Triantafyllou (2005), Comparison of extended and ensemble based kalman filters with low and high resolution primitive equation ocean models, *Nonlinear Processes in Geophysics*, *12*(5), 755–765.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2013), *An Introduction to Statistical Learning: with Applications in R*.

Jeong, K. S., D. K. Kim, and G. J. Joo (2006), River phytoplankton prediction model by artificial neural network: Model performance and selection of input variables to predict time-series phytoplankton proliferations in a regulated river system, *Ecological Informatics*, *1*(3), 235–245, doi:Doi10.1016/J.Ecoinf.2006.04.001.

Kim, Y. H., J. Im, H. K. Ha, J. K. Choi, and S. Ha (2014), Machine learning approaches to coastal water quality monitoring using GOCI satellite data, *Giscience & Remote Sensing*, *51*(2), 158–174, doi:Doi10.1080/15481603.2014.900983.

Korres, G., G. Triantafyllou, G. Petihakis, D. E. Raitsos, I. Hoteit, A. Pollani, S. Colella, and K. Tsiaras (2012), A data assimilation tool for the Pagasitikos Gulf ecosystem dynamics: Methods and benefits, *Journal of Marine Systems*, *94*, S102–S117, doi:Doi10.1016/J.Jmarsys.2011.11.004.

Lee, J. H. W., Y. Huang, M. Dickman, and A. W. Jayawardena (2003), Neural network modelling of coastal algal blooms, *Ecological Modelling*, *159*(2-3), 179–201, doi:PiiS0304-3800(02)00281-8Doi10.1016/S0304-3800(02)00281-8.

Mann, K. H., and J. R. N. Lazier (2006), *Dynamics of marine ecosystems: Biological-Physical Interactions in the Oceans*, Blackwell Publishing.

McClain, C. R. (2009), A decade of satellite ocean color observations, *Annual Review of Marine Science*, *1*, 19–42, doi:Doi10.1146/Annurev.Marine.010908.163650.

Miles, T. N., and R. He (2010), 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*(18), 1951–1962, doi:10.1016/j.csr.2010.08.016.

Mulia, I. E., H. Tay, K. Roopsekhar, and P. Tkalich (2013), Hybrid ANN-GA model for predicting turbidity and chlorophyll-a concentrations, *Journal of Hydro-Environment Research*, *7*(4), 279–299, doi:Doi10.1016/J.Jher.2013.04.003.

North, G. R., J. Wang, and M. G. Genton (2011), Correlation models for temperature fields, *Journal of Climate*, *24*(22), 5850–5862, doi:10.1175/2011jcli4199.1.

Pal, R. (2014), *An introduction to phytoplanktons : diversity and ecology*, pages cm pp., Springer, New York.

Pettersson, L. H., and D. V. Pozdniakov (2013), *Monitoring of harmful algal blooms*, Springer-Praxis books in geophysical sciences, Springer, published in association with Praxis Publishing, Chichester, UK.

Racault, M. F., D. E. Raitsos, M. L. Berumen, R. J. Brewin, T. Platt, S. Sathyendranath, and I. Hoteit (2015), Phytoplankton phenology indices in coral reef ecosystems: application to ocean-colour observations in the red sea, *Submitted*.

Raitsos, D. E., I. Hoteit, P. K. Prihartato, T. Chronis, G. Triantafyllou, and Y. Abualnaja (2011), Abrupt warming of the red sea, *Geophysical Research Letters*, *38*(14), doi:ArtnL14601Doi10.1029/2011gl047984.

Raitsos, D. E., G. Korres, G. Triantafyllou, G. Petihakis, M. Pantazi, K. Tsiaras, and A. Pollani (2012), Assessing chlorophyll variability in relation to the environmental regime in pagasitikos gulf, greece, *Journal of Marine Systems*, *94*, S16–S22, doi:10.1016/j.jmarsys.2011.11.003.

Raitsos, D. E., Y. Pradhan, R. J. Brewin, G. Stenchikov, and I. Hoteit (2013), Remote sensing the phytoplankton seasonal succession of the Red Sea, *PLoS One*, *8*(6), doi: 10.1371/journal.pone.0064909.

Raitsos, D. E., X. Yi, T. Platt, M.-F. Racault, R. J. W. Brewin, Y. Pradhan, V. P. Papadopoulos, S. Sathyendranath, and I. Hoteit (2015), Monsoon oscillations regulate fertility of the red sea, *Geophysical Research Letters*, p. 2014GL062882, doi: 10.1002/2014GL062882.

Richlen, M. L., S. L. Morton, E. A. Jamali, A. Rajan, and D. M. Anderson (2010), The catastrophic 2008-2009 red tide in the arabian gulf region, with observations on the identification and phylogeny of the fish-killing dinoflagellate cochlodinium polykrikoides, *Harmful Algae*, *9*(2), 163–172, doi:Doi10.1016/J.Hal.2009.08.013.

Robinson, I. S. (2010), *Discovering the ocean from space : the unique applications of satellite oceanography*, Springer praxis series geophysical sciences 4110, 1st ed., xivi, 638 p. pp., Springer, New York.

Sirjacobs, D., A. Alvera-Azcrate, A. Barth, G. Lacroix, Y. Park, B. Nechad, K. Ruddick, and J.-M. Beckers (2011), Cloud filling of ocean colour and sea surface temperature remote sensing products over the southern north sea by the data interpolating empirical orthogonal functions methodology, *Journal of Sea Research*, *65*(1), 114–130, doi:10.1016/j.seares.2010.08.002.

Stein, M. L. (2005), Spacetime covariance functions, *Journal of the American Statistical Association*, *100*(469), 310–321, doi:10.1198/016214504000000854.

Steinmetz, F., P. Y. Deschamps, and D. Ramon (2011), Atmospheric correction in presence of sun glint: application to meris, *Optics Express*, *19*(10), 9783–9800, doi:Doi10.1364/Oe.19.009783.

Tandeo, P., M. Pulido, and F. Lott (2014), Offline parameter estimation using enkf and maximum likelihood error covariance estimates: Application to a subgrid-scale orography parametrization, *Quarterly Journal of the Royal Meteorological Society*, doi:10.1002/qj.2357.

Taylor, M. H., M. Losch, M. Wenzel, and J. Schroter (2013), On the sensitivity of field reconstruction and prediction using empirical orthogonal functions derived from gappy data, *Journal of Climate*, *26*(22), 9194–9205, doi:Doi10.1175/Jcli-D-13-00089.1.

Tett, P., and E. D. Barton (1995), Why are there about 5000 species of phytoplankton in the sea, *Journal of Plankton Research*, *17*(8), 1693–1704, doi: Doi10.1093/Plankt/17.8.1693.

Triantafyllou, G., I. Hoteit, and G. Petihakis (2003), A singular evolutive interpolated kalman filter for efficient data assimilation in a 3-d complex physical-biogeochemical model of the cretan sea, *Journal of Marine Systems*, *40*, 213–231, doi:Doi10.1016/S0924-7963(03)00019-8.

Triantafyllou, G., I. Hoteit, X. Luo, K. Tsiaras, and G. Petihakis (2013), Assessing a robust ensemble-based Kalman filter for efficient ecosystem data assimilation of the Cretan Sea, *Journal of Marine Systems*, *125*, 90–100, doi:Doi10.1016/J.Jmarsys.2012.12.006.

Triantafyllou, G., F. Yao, G. Petihakis, K. P. Tsiaras, D. E. Raitsos, and I. Hoteit (2014), Exploring the red sea seasonal ecosystem functioning using a three-dimensional biophysical model, *Journal of Geophysical Research: Oceans*, *119*(3), 1791–1811, doi:10.1002/2013jc009641.

Waite, J. N., and F. J. Mueter (2013), 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*, 179–192, doi:10.1016/j.pocean.2013.07.006.

Wang, H., and X. Yang (2013), Prediction and elucidation of algal dynamic variation in Gonghu Bay by using artificial neural networks and canonical correlation analysis.

Weikert, H. (1987), *Plankton and the pelagic environment*, pp. 90–111, Pergamon Press, Oxford, doi:http://dx.doi.org/10.1016/B978-0-08-028873-4.50010-4.

Wu, N. C., J. C. Huang, B. Schmalz, and N. Fohrer (2014), Modeling daily chlorophyll a dynamics in a german lowland river using artificial neural networks and multiple linear regression approaches, *Limnology*, *15*(1), 47–56, doi: Doi10.1007/S10201-013-0412-1.

Yao, F., and I. Hoteit (2015), Thermocline regulated seasonal evolution of surface chlorophyll in the gulf of aden, *PLoS One*.

Yao, F. C., I. Hoteit, L. J. Pratt, A. S. Bower, A. Kohl, G. Gopalakrishnan, and D. Rivas (2014a), Seasonal overturning circulation in the red sea: 2. winter circulation, *Journal of Geophysical Research-Oceans*, *119*(4), 2263–2289, doi: Doi10.1002/2013jc009331.

Yao, F. C., I. Hoteit, L. J. Pratt, A. S. Bower, P. Zhai, A. Kohl, and G. Gopalakrishnan (2014b), Seasonal overturning circulation in the red sea: 1. model validation and summer circulation, *Journal of Geophysical Research-Oceans*, *119*(4), 2238–2262.

Zhai, P., and A. Bower (2013), The response of the red sea to a strong wind jet near the tokar gap in summer, *Journal of Geophysical Research-Oceans*, *118*(1), 422–434, doi:Doi10.1029/2012jc008444.

Zhan, P., A. C. Subramanian, F. C. Yao, and I. Hoteit (2014), Eddies in the red sea: A statistical and dynamical study, *Journal of Geophysical Research-Oceans*, *119*(6), 3909–3925, doi:Doi10.1002/2013jc009563.

APPENDICES

Table of Data

Detailed experimental procedures, data tables, computer programs, etc. may be placed in appendices. This may be particularly appropriate if the dissertation or thesis includes several published papers.

Paper 1: Filtering remotely sensed chlorophyll concentrations in the Red Sea using a space-time covariance model and a Kalman Filter

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

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

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

175 *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-

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

199 *3.4. Construction and fitting of the covariance model*

200 For space-time processes, families of covariance functions are typically
201 built by combining known spatial and temporal covariance models. Sep-
202 arable models are the simplest example of this approach, taking the form
203 $C(\mathbf{h}; u|\boldsymbol{\theta}) = C^1(\mathbf{h}|\boldsymbol{\theta}_1)C^2(t|\boldsymbol{\theta}_2)$, where C^1 and C^2 are respectively pure spa-
204 tial and temporal covariance models parametrized by $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. However,
205 realizations from such families of covariance models were shown to exhibit
206 non-physical behaviors [1, 27]. Research has been conducted to develop meth-
207 ods for constructing non-separable covariance functions to overcome these
208 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.

255 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:

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

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

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

256 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

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

257 Otherwise, the forecast and update steps are identical to those of the Kalman
258 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,

379 particularly in its central part, with correlations ranging between 0.28 and
380 0.92. The filter further improves the prediction-data correlation, over the
381 entire domain with correlations up to 0.96.

382 6. Discussion

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

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

398 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.

459 [10] R. J. W. Brewin, D. E. Raitsos, Y. Pradhan, I. Hoteit, Compari-
460 son of chlorophyll in the Red Sea derived from MODIS-Aqua and in
461 vivo fluorescence, Remote Sensing of Environment 136 (2013) 218–224.
462 doi:10.1016/j.rse.2013.04.018.
463 URL http://dx.doi.org/10.1016/j.rse.2013.04.018

464 [11] C. R. McClain, A Decade of Satellite Ocean Color Ob-
465 servations, Annual Review of Marine Science (2009) 19–
466 42doi:10.1146/annurev.marine.010908.163650.

467 [12] T. R. Anderson, Plankton functional type modelling: running
468 before we can walk?, Journal of Plankton Research 27 (11).
469 doi:10.1093/plankt/fbi076.
470 URL http://www.plankt.oxfordjournals.org/cgi/doi/10.1093/plankt/fbi076

471 [13] J. Baretta, W. Ebenhöh, P. Ruardij, The European regional seas
472 ecosystem model, a complex marine ecosystem model, Netherlands
473 Journal of Sea Research 33 (3/4) (1995) 233–246. doi:10.1016/0077-
474 7579(95)90047-0.
475 URL http://linkinghub.elsevier.com/retrieve/pii/0077757995900470

476 [14] G. Petihakis, G. Triantafyllou, I. J. Allen, I. Hoteit, C. Dounas, Mod-
477 elling the spatial and temporal variability of the Cretan Sea ecosystem,
478 Journal of Marine Systems 36 (2002) 173–196.

479 [15] J. C. Blackford, P. J. Radford, A Structure and methodology for ma-

480 rine ecosystem modelling, Netherlands Journal of Sea Research 33 (3/4)
481 (1995) 247–260.

482 [16] J. H. W. Lee, Y. Huang, M. Dickman, A. W. Jayawardena, Neural
483 network modelling of coastal algal blooms 159 (2003) 179–201.

484 [17] F. Recknagel, Artifical neural network approach for modelling and pre-
485 diction of algal blooms, Ecological Modelling 96 (1997) 11–28.

486 [18] D. Raitsos, G. Korres, G. Triantafyllou, G. Petihakis, M. Pantazi,
487 K. Tsiaras, A. Pollani, Assessing chlorophyll variability in relation
488 to the environmental regime in Pagasitikos Gulf, Greece, Journal of
489 Marine Systems 94 (2012) S16–S22. doi:10.1016/j.jmarsys.2011.11.003.
490 URL http://linkinghub.elsevier.com/retrieve/pii/S0924796311002703

491 [19] D. E. Raitsos, S. J. Lavender, Y. Pradhan, T. Tyrrell, P. C. Reid, M. Ed-
492 wards, Coccolithophore bloom size variation in response to the regional
493 environment of the subarctic North Atlantic, Limnology and Oceanog-
494 raphy 51 (5) (2006) 2122–2130. doi:10.4319/lo.2006.51.5.2122.
495 URL http://www.aslo.org/lo/toc/vol_51/issue_5/2122.html

496 [20] P. C. Kyriakidis, A. G. Journel, Geostatistical Space-Time Models: A
497 Review, Mathematical Geology 31 (6) (1999) 651–684.

498 [21] N. Cressie, C. K. Wikle, Statistics for Spatio-Temporal Data, 2011.

499 [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

520 temporal correlations in EOF expansions for the reconstruction of miss-
521 ing data using DINEOF, Ocean Science 5 (2009) 475–485.

522 [30] P. Monestiez, L. Dubroca, E. Bonnin, J.-P. Durbec, C. Guinet, Geo-
523 statistical modelling of spatial distribution of Balaenoptera physalus
524 in the Northwestern Mediterranean Sea from sparse count data and
525 heterogeneous observation efforts, Ecological Modelling 193 (3-4) (2006)
526 615–628. doi:10.1016/j.ecolmodel.2005.08.042.
527 URL http://linkinghub.elsevier.com/retrieve/pii/S0304380005004436

528 [31] I. Hoteit, D.-T. Pham, J. Blum, A simplified reduced order Kalman fil-
529 tering and application to altimetric data assimilation in Tropical Pacific,
530 Journal of Marine Systems 36 (2002) 101–127.

531 [32] C. K. Chui, G. Chen, Kalman Filtering, with Real-Time Applications,
532 Fourth Edition, 2009.

533 [33] J. W. Campbell, The lognormal distribution as a model for bio-optical
534 variability in the sea, Journal of Geophysical Research 100 (13) (1995)
535 237–254.

536 [34] J. K. Willis, Interannual variability in upper ocean heat content, tem-
537 perature, and thermosteric expansion on global scales, Journal of Geo-
538 physical Research 109. doi:10.1029/2003JC002260.
539 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

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

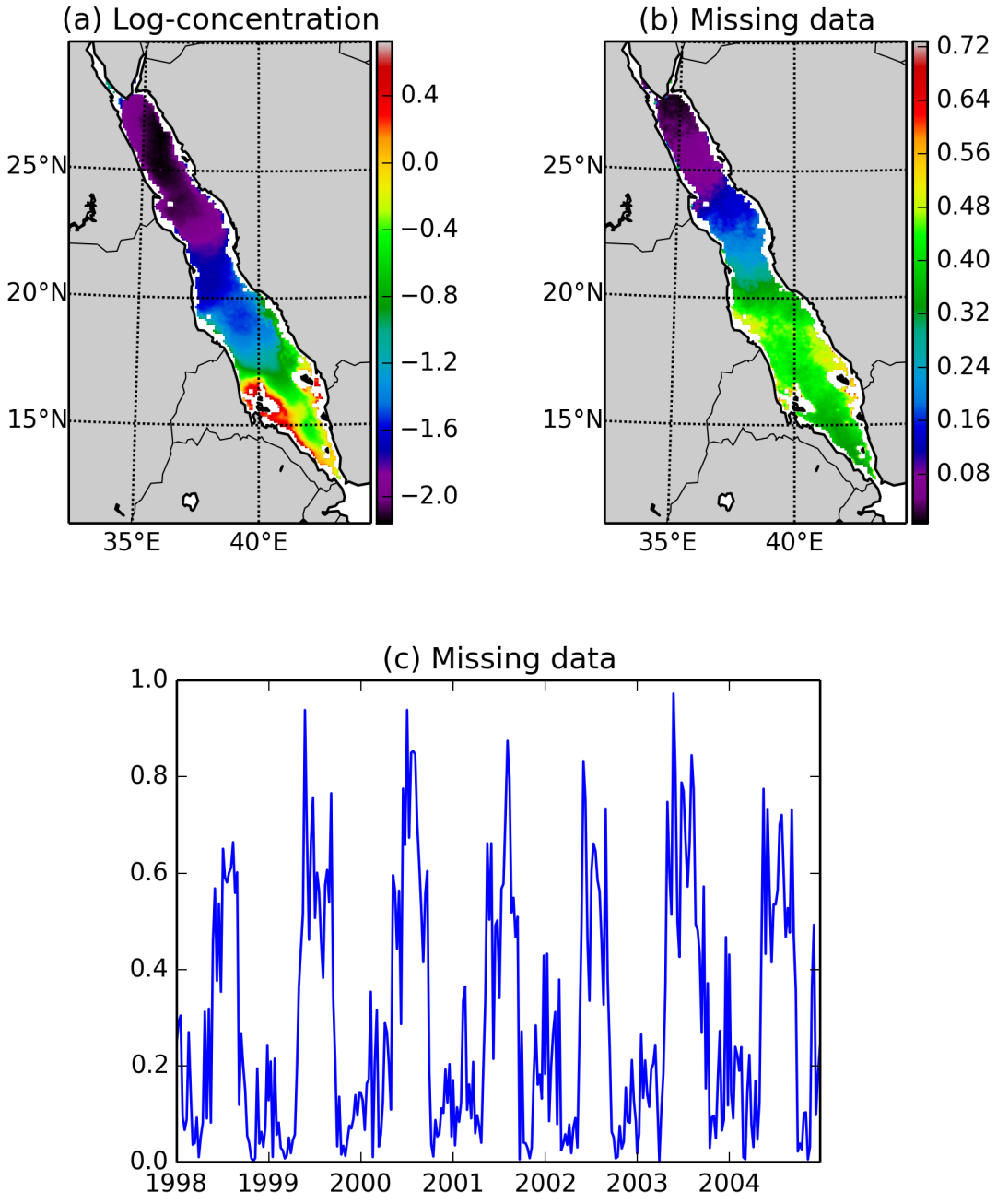

Figure 1: Raw data plots: (a) map of average log-concentrations of CHL between 1998 and 2004, (b) map of average percentage of missing data for each location between 1998 and 2006, and (c) time-series of the percentage of missing data over the Red Sea between 1998 and 2004.

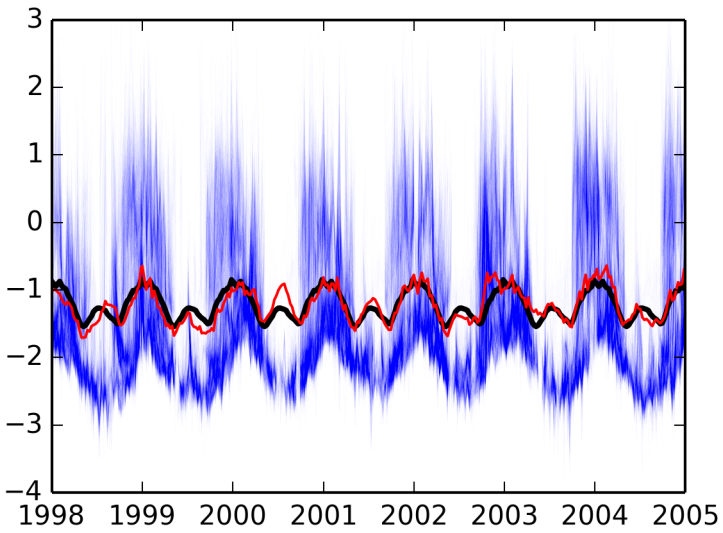

Figure 2: Log-CHL concentration time-series for every pixel of the domain (blue curves). The red curve plots the spatially averaged log-concentrations and the black curve plots the spatial average of the seasonal component. Both are computed from the data filled with DINEOF.

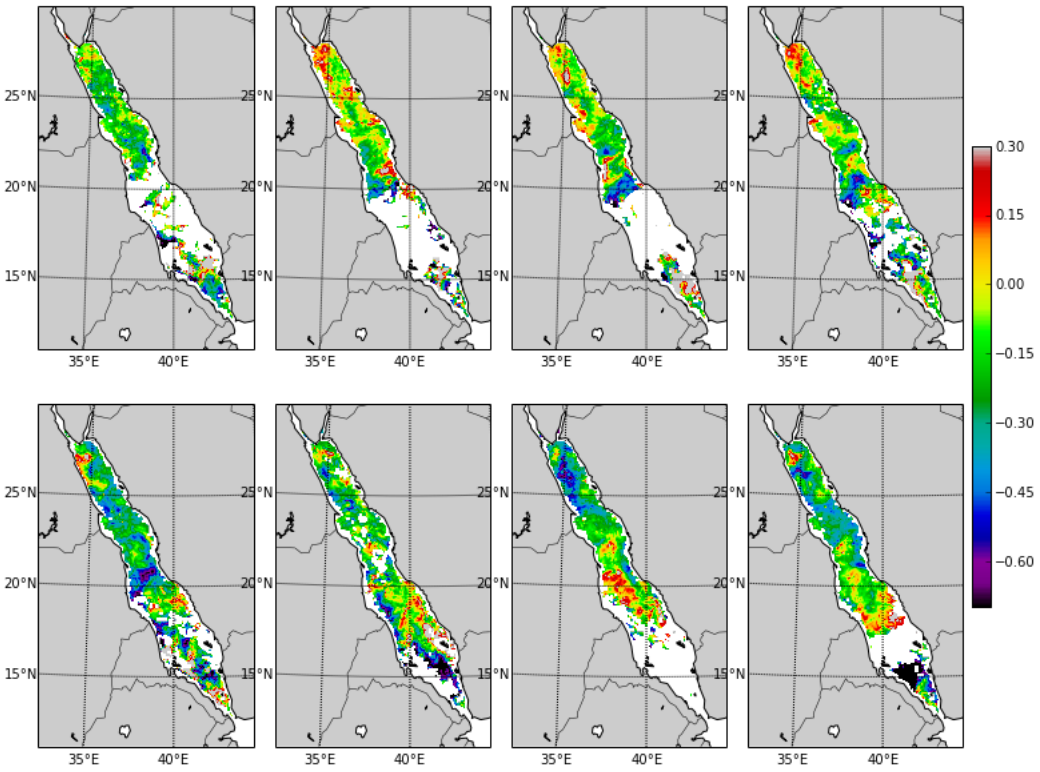

Figure 3: Anomalies estimated by DINEOF for the first 8 weeks of 1998.

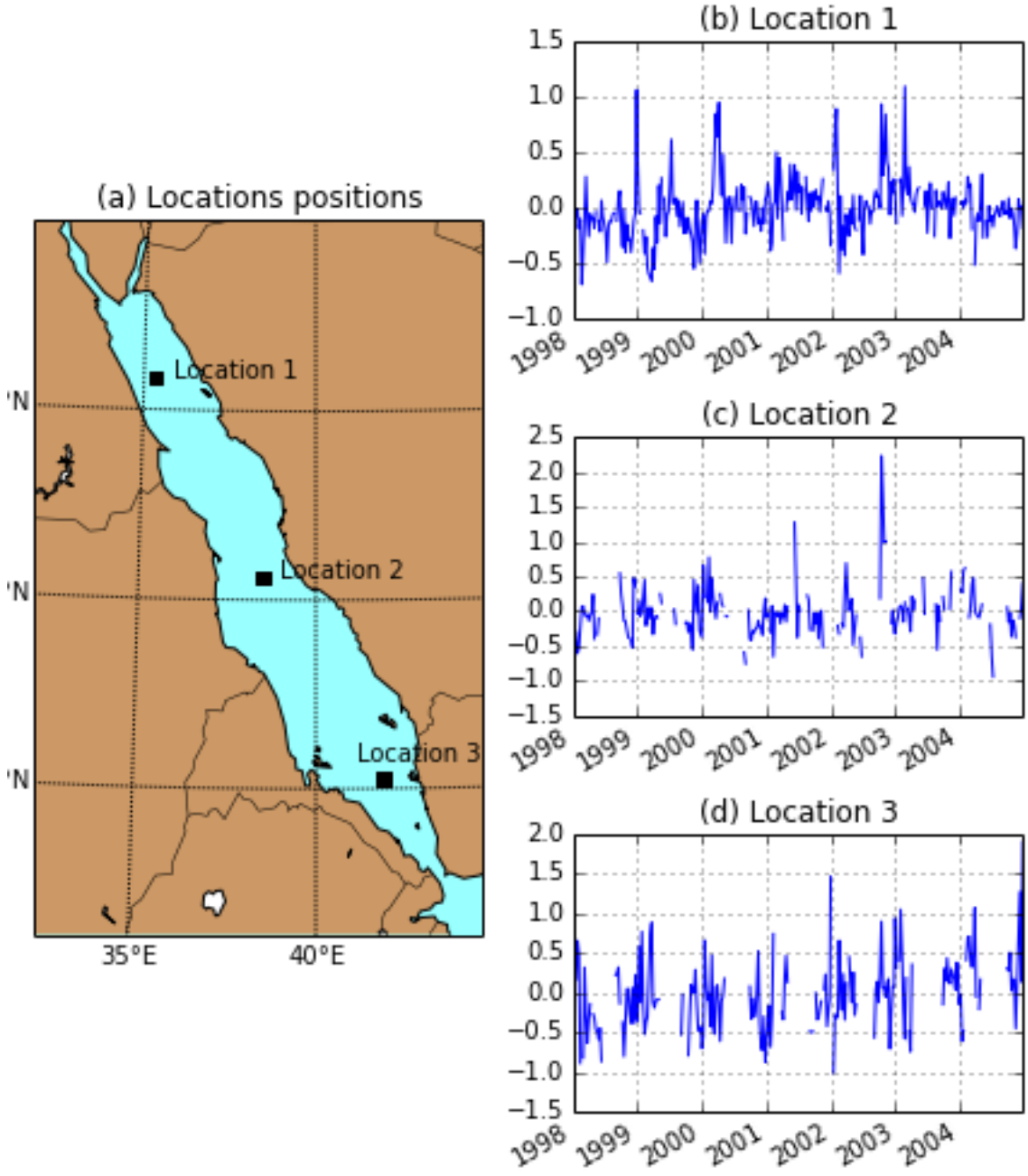

Figure 4: Anomaly times series (panels b to d) at three locations as indicated in panel (a)

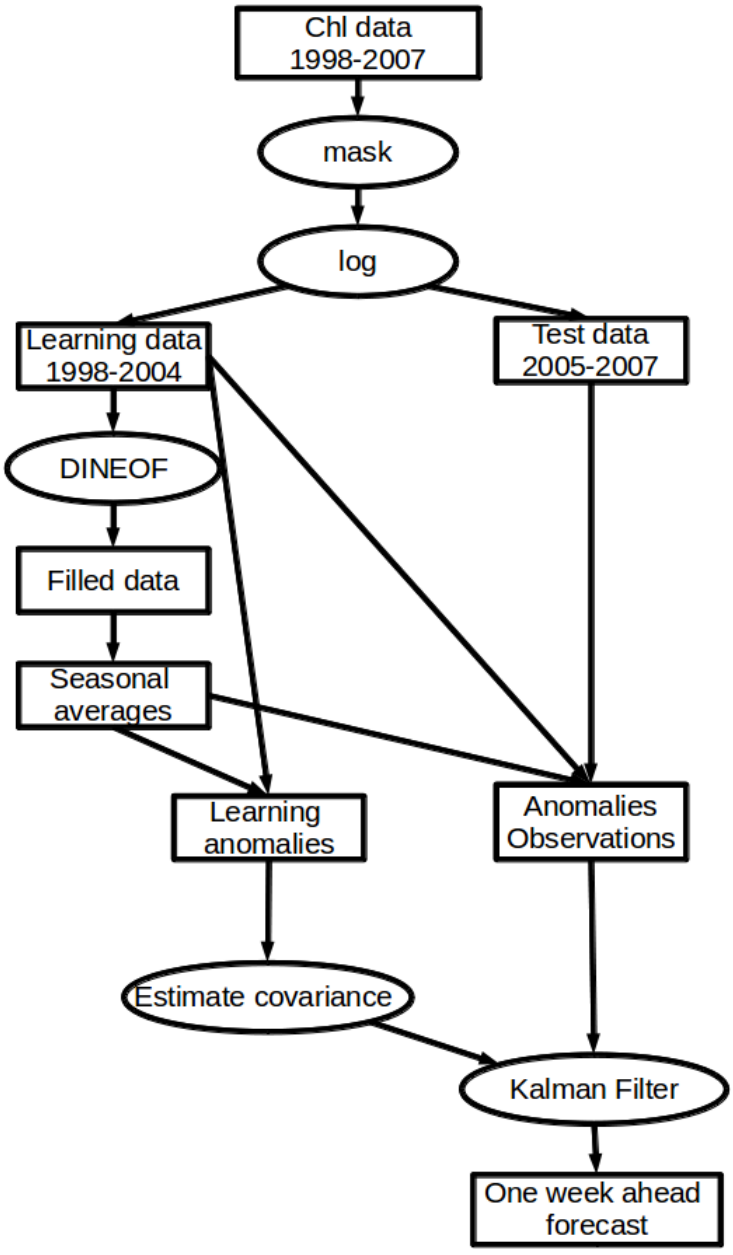

Figure 5: Description of the experimental setup.

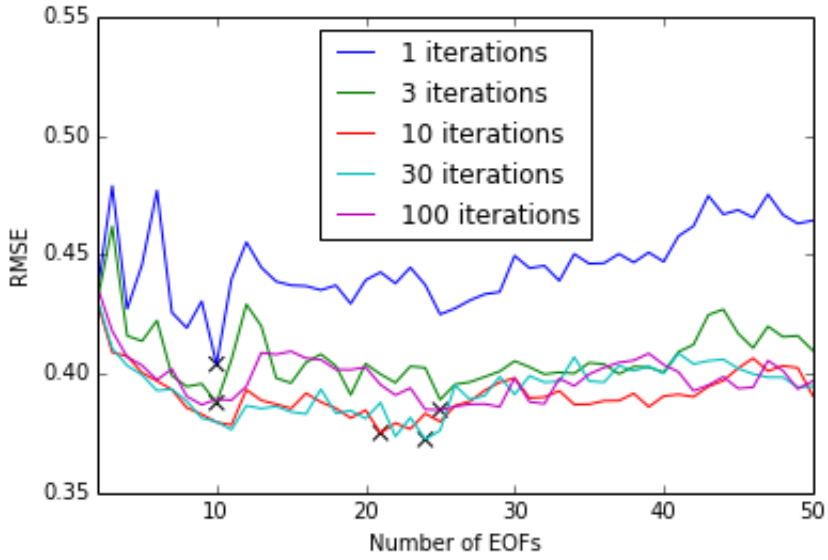

Figure 6: RMS error over the cross-validation period for a varying number of smoothing iterations and number of EOFs. Crosses indicate the minimal error for a given number of iterations.

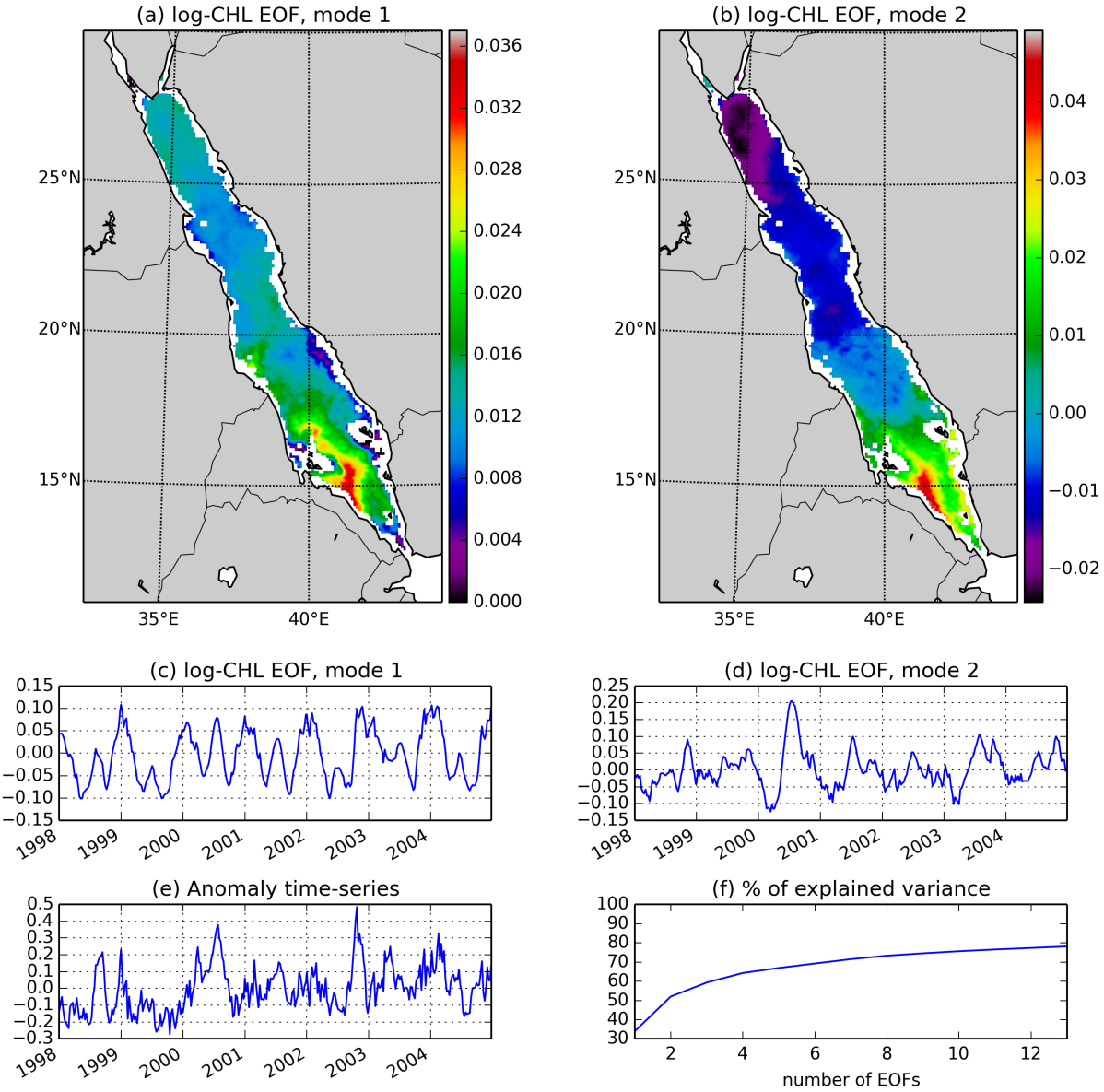

Figure 7: EOF modes after the DINEOF data filling: (a) first spatial EOF mode, (b) second spatial EOF mode, (c) first temporal EOF mode, (d) second temporal EOF mode, (e) spatial average of the seasonal anomalies computed from the DINEOF filled data, (f) cumulative sum of the percentage of variance explained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

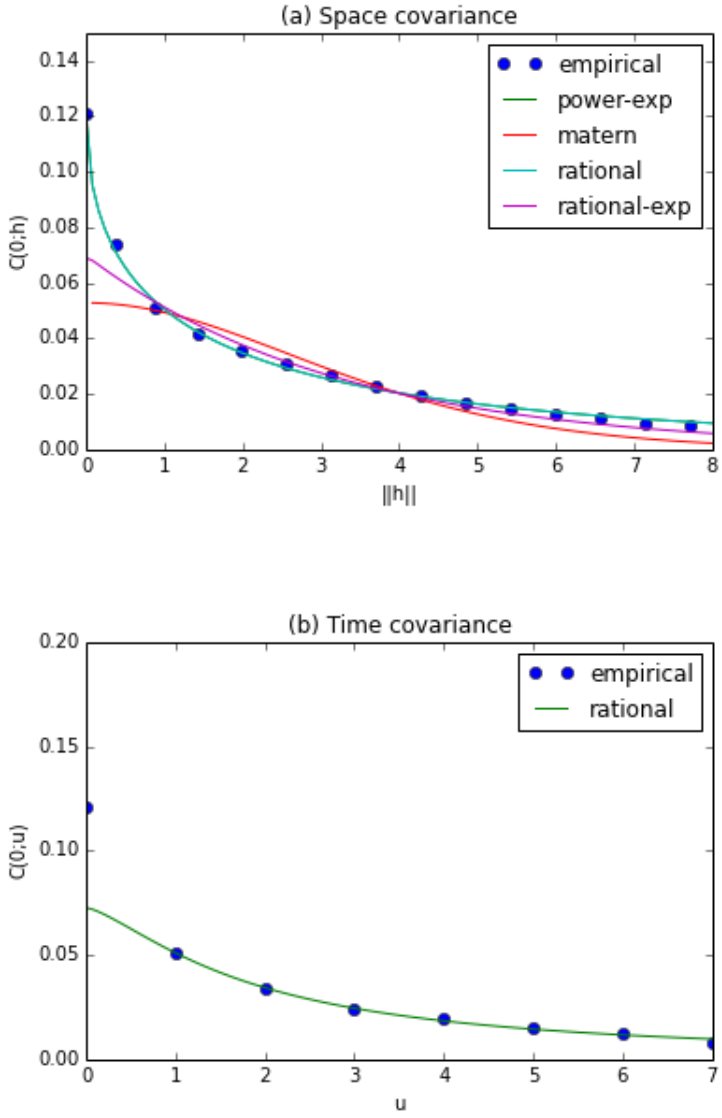

Figure 8: Fitting the covariance model to the space-time empirical covariance matrix: (b) empirical space covariance function and fitted space covariance functions, (c) empirical time covariance function and fitted time covariance functions.

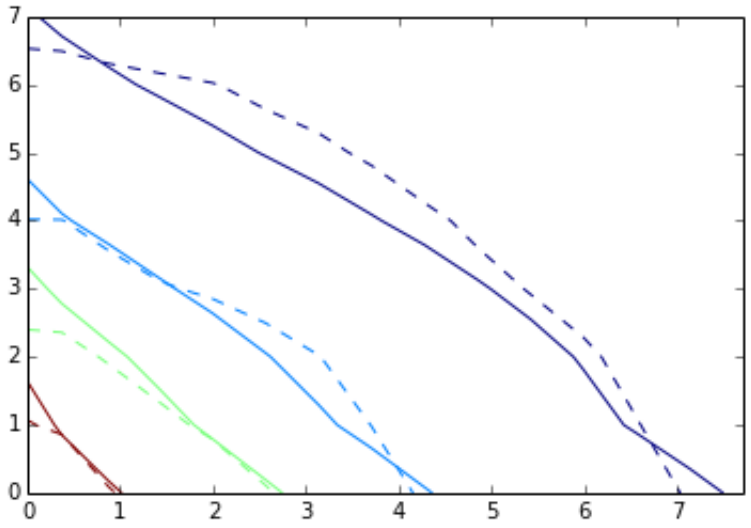

Figure 9: Contour plot of the empirical covariance function (dashed curves) and of the fitted space-time covariance function (solid curves)

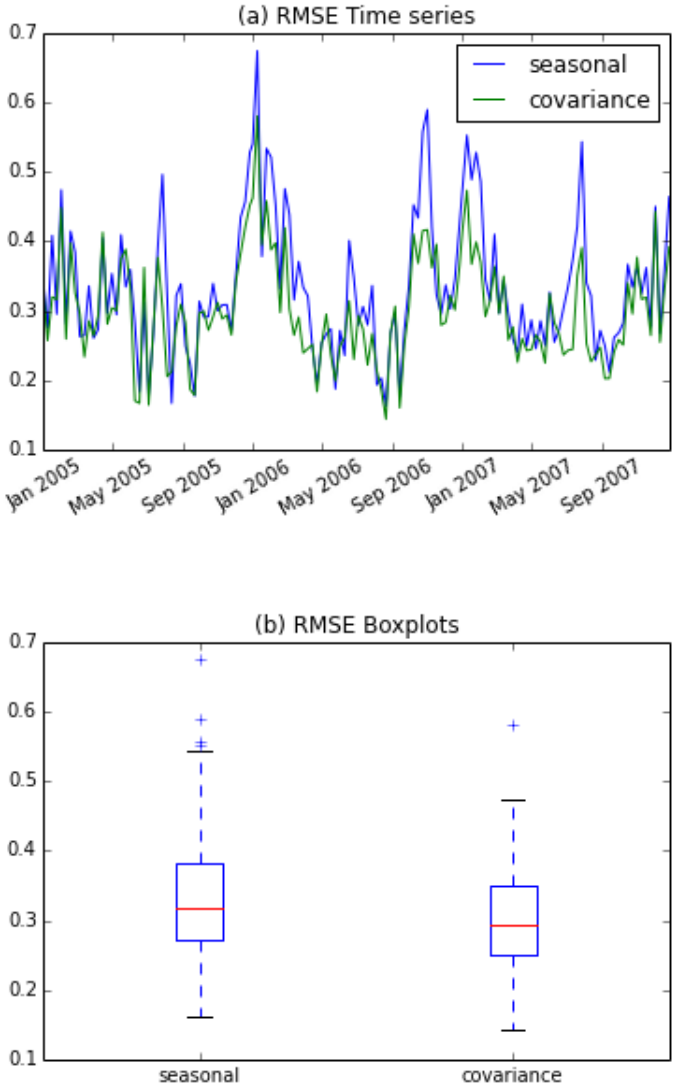

Figure 10: Results of the Kalman filter: (a) time series of RMS errors of forecasts for the covariance model compared with a pure seasonal prediction, (b) distribution of the prediction RMS error over the three-year cross-validation period.

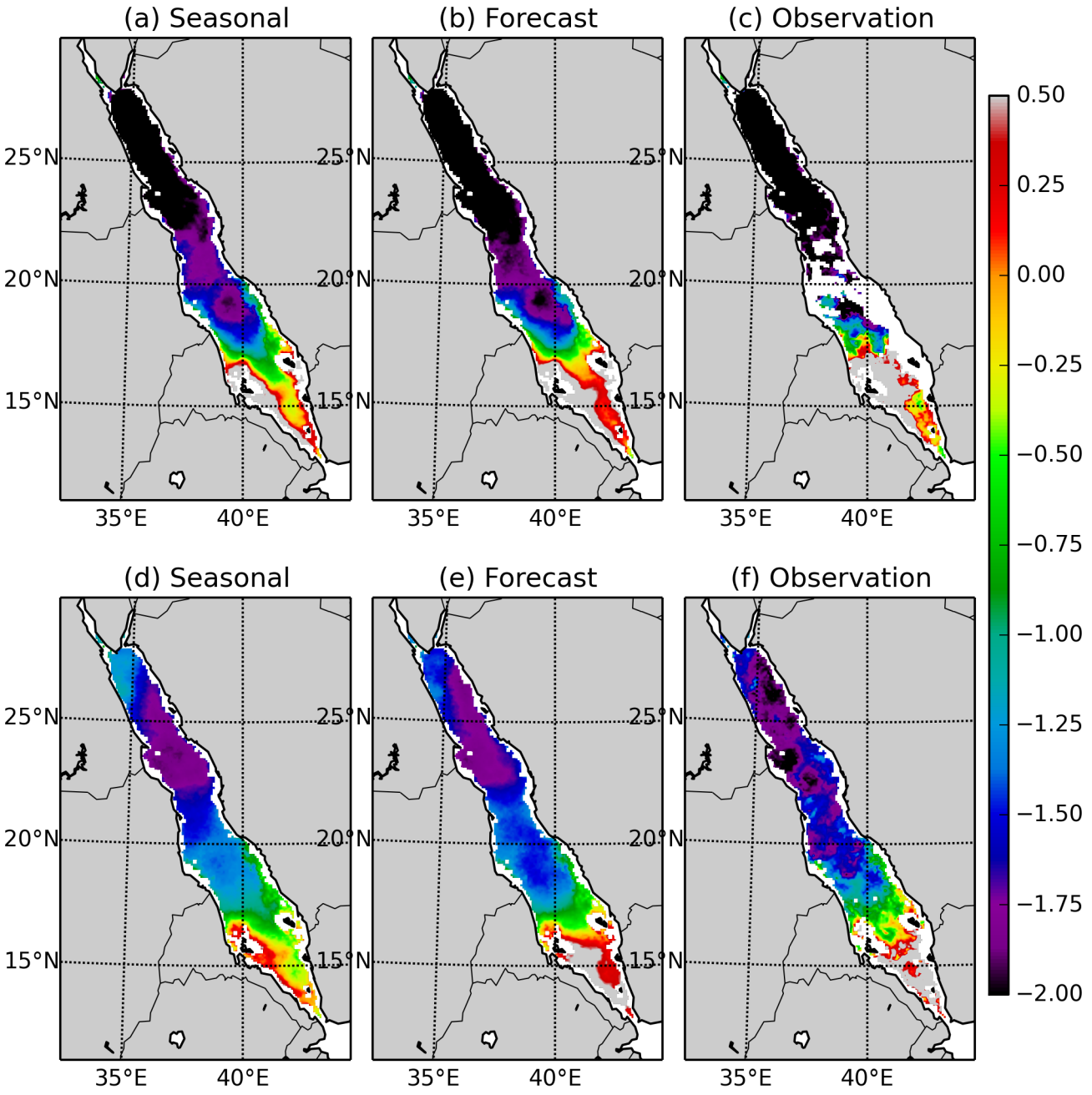

Figure 11: Predictions for the period between 24 October 2006 and 31 October 2006 (a, b, c), and between 6 March 2006 and 14 March 2006 (d, e, f): (a, d) plot the seasonal prediction, (b, e) the model prediction with the Kalman filter, and (c, f) the observations.

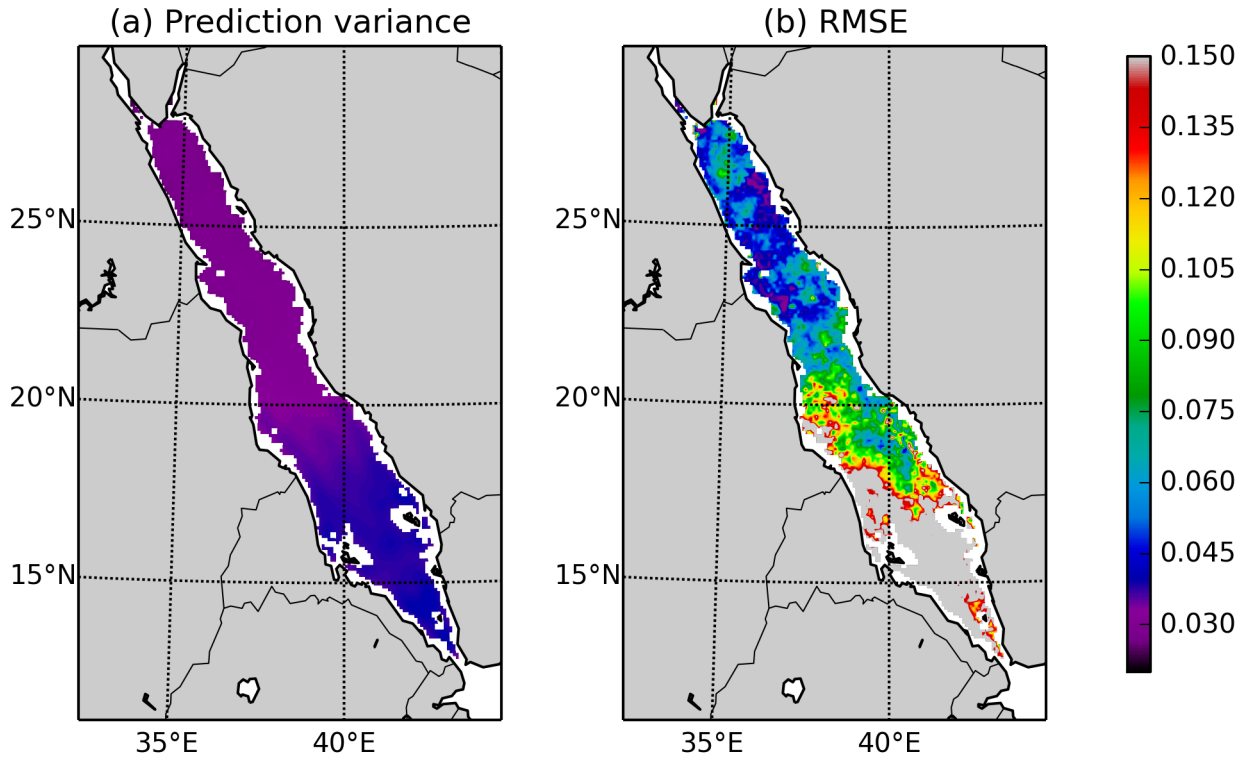

Figure 12: Spatial averages over the validation period of: (a) the variances of the prediction error resulting from by the reduced-order Kalman filter, and (b) the RMS errors.

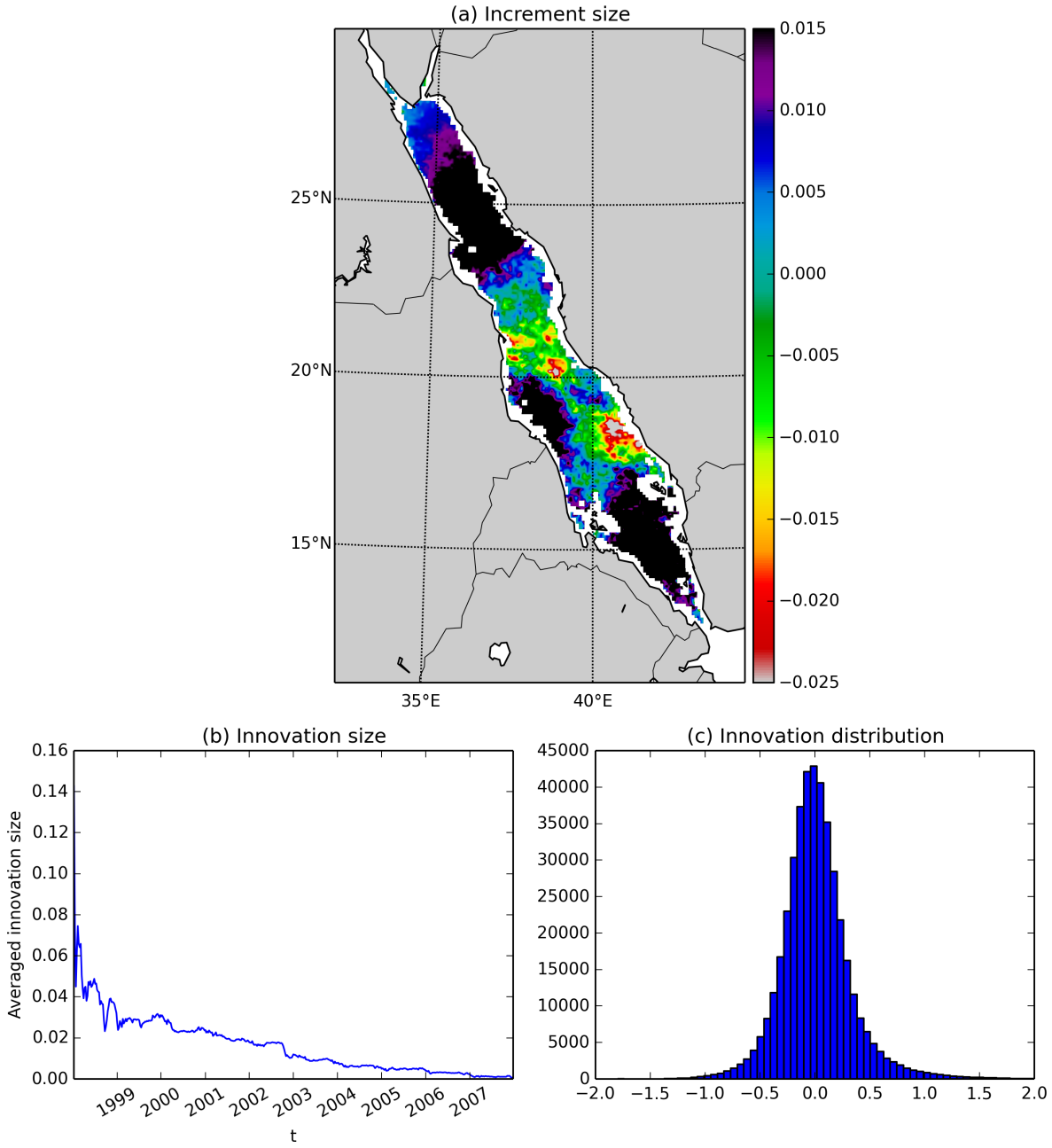

Figure 13: Diagnostic statistics of the filter: (a) increment distribution in space averaged over the validation period, (b) cumulative averaged innovation over time, (c) distribution of the innovation values for the validation period.

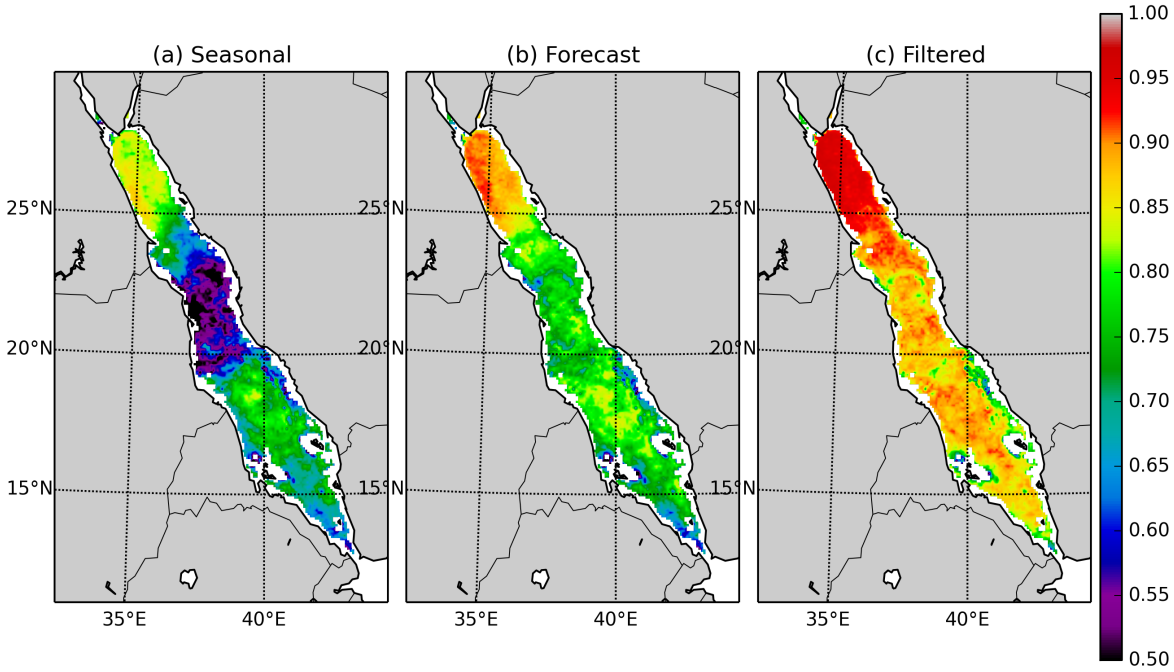

Figure 14: Average correlation maps over the validation period between the observations and (a) the seasonal forecast, (b) the CHL forecasts (including the seasonal component), and (c) the filter analyses (including the seasonal component).