Statistical Modeling of the Red Sea Chlorophyll Concentration and Application to the ERSEM Ecological Model

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: Importance, Impact, 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 plays a fundamental role for the ocean ecology. It is the basis of the marine food web and traps most of the energy used by pelagic ecosystems [Pal, 2014]. Zooplankton graze phytoplankton which are then 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 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], causing massive mortality in the fauna. Due to the rapid growth of phytoplankton, it responds 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 is an important factor for fisheries. Productive fishing zones like the regions in the Arabian seas, Californian coast, north-west African coast and Chilean coast are explained by the upwelling of cold nutrient rich water favourable to phytoplankton growth. 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]. Remotely-sensed chlorophyll data have been routinely used since the last decade to help fisheries predict the timing of phytoplankton blooms [Robinson, 2010]. 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, however, sinks to the bottom of the ocean and 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 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 interaction with the open sea through the mesoscale eddies is believed to play an important role [Raitsos et al., 2013, Zhan et al., 2014].

Although the Red Sea environment is relatively preserved, it is stressed by human activities. The increasing 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 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 biology of this region. The Red Sea is deficient

in the major nutrients [Weikert, 1987], and the only significant input of water comes from the Gulf of Aden. This explains a general increase of chlorophyll concentration from north to south [Raitsos et al., 2013]. The lowest concentration is 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 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 nutrient. 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. The 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., 2014,

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 that region [Raitsos et al., 2013].

Aerial depositions of dust could also be an important input of nutrient 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.

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. 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 a way to estimate the chlorophyll concentration of the water.

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 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, are also a barrier 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 concentration are used intensively 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

The performance 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 present a huge amount of missing 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 0% [Racault et al., 2015].

Chlorophyll concentration estimation in optically complex case II waters is a recurrent problem in this 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 data [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 DINEOF. It is an EOF based data filling approach introduced by [Beckers and Rixen, 2003]. It has been used for multivariate reconstruction of SST fields using chlorophyll data in [Alvera-Azcarate et al., 2007]. 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, showing 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].

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 in a 3D coupled ecological model [Triantafyllou et al., 2014]. However, the interaction between the open sea and coral reefs, and the role of sand storms 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 for 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 Persian Gulf [Richlen et al., 2010].

Deterministic Models

Ecological Models

Ecological ordinary differential equation (ODE) deterministic models are a popular way to model marine ecology. Such models can be as simple as the nutrient-phytoplankton-zooplankton (NPZ) model that only has three variables representing two trophic levels, or as complex as the European regional seas ecosystem model (ERSEM) that has dozens of variables and represents many ecological, biological and chemical interactions. Such a model has 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 correctly if not enough data is available, which is usually the case [Anderson, 2005].

Data Assimilation

Data assimilation is used to improve the simulations of ecological dynamics models and enhance their forecasting capabilities by correcting their predictions with observations. 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 study [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 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, Ciavatta et al., 2014]. However, it suffers 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 order version of the Extended Kalman filter (EK), that in intractable in high-dimensions. Like SEIK, it projects the error covariance to a low dimensional space. SEEK has a long history in data assimilation for marine ecology

models and is still used in recent studies [Fontana et al., 2013, Korres et al., 2012, Butenschon and Zavatarelli, 2012]. SEIK is a deterministic version of the EnKF that do not suffer from sampling problems, as it projects the propagated error in a low-dimensional subspace. 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 non linear systems.

Ecological models are challenging applications for data assimilation schemes. First, biogeochemical variable are usually positive concentration, whereas Kalman filters expect Gaussian variables, and log-transformation can fail at solving this issue [Ciavatta et al., 2011]. However, [Fontana et al., 2013] has successfully 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, EnKS and SEEK both project the error covariance in a 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] proposes to use the H_∞ method with SEIK in order alleviate this requirement.

Data-Driven Approaches

On the other hand, data-driven statistical models are relatively easier to apply. They are relevant when the phenomenon producing the data is dynamically complex to model or simply poorly understood. They have been applied to predict chlorophyll concentration, mostly in small regions that have complex dynamics (see **??**). Some statistical models, such as linear regression, Gaussian additive models, or tree regres-

sion have the advantage of being easy to interpret [James et al., 2013], and can be used to understand the dynamics driving the chlorophyll concentration.

Machine Learning Algorithms

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 in [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 addressed 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 also 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. In such regions, deterministic ecological models are usually too complex and less efficient than data-driven approaches. 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 model 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]. There is however no clear cut division between regions with different patterns, making it difficult to divide the Red Sea into regions. Finally we can expect the different regions of the Red Sea to interact.

Classical geostatistics is the most widely used spatial statistical model. It models spatial data as the realization of a two-dimensional Gaussian process, of which one can estimate the parameters. Geostatistics can be easily extended to spatio-temporal datasets. Many flexible ways of constructing space-time covariance functions for these models have been proposed recently [Gneiting, 2002, Cressie and Huang, 1999, Stein, 2005]. Space-time geostatistics has been applied in many environmental studies, but not to chlorophyll data yet.

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 time and space interactions derive from physical interaction, and must be taken into account in the definition of the covariance function [Gneiting and Guttorp, 2010]. Some space-time can actually be derived from a physical formulation, such as the frozen fields [Gneiting and Guttorp, 2010], or SDEs [Brown et al., 2000, North et al., 2011].

Despite their theoretical interest, physically-derived space-time covariance function 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 sep-

arable covariance functions, that are the product of a spatial covariance function and temporal covariance function. They are computationally efficient, but are enable 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 use in a variety of applications. [Hohn et al., 1993] used it for forecast the outbreaks of an invasive specie. They have 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 demonstrate that statistical predictive models can be used to help efficiently forecasting chlorophyll concentration in the Red Sea. Statistical methods can be more robust and computationally efficient when the underlying dynamics are very complex and observations are limited. The study of chlorophyll concentration is such a case. The dissertation will compare the 8-days prediction skill of increasingly sophisticated predictive models to the highly sophisticated ecological model ERSEM. We will explore the possibilities of combining statistical and deterministic models to improve the chlorophyll forecasts.

Efficient statistical models may be used as alternatives to the much more complex

deterministic models in other seas. They can help researchers to gain insight into the dynamics of the phytoplankton. They can also help coastal communities to mitigate the effects of harmful phytoplankton blooms on public health and their economy. Moreover this study will help confirming hypothesizes that have been made about the interaction of Red Sea phytoplankton with the regional and global circulation.

Research Plan

Write some introductory text here

Chapter 1: Dataset Building and Exploration

Duration: 2 months (by December 2014)

Submission: Journal of Marine Systems

Collaborator: Dionysios Raitsos

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 data, 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?
- Can the hypothesizes about the dynamics behind the chlorophyll seasonal cycle be confirmed by the data?
- Are there more blooms in the past years?

Method and Work Done

1. Identify data sources and load the data.
2. Clean the data and fill missing values (DINEOF).
3. Align and format the data to build a unique dataset.
4. Explore the dataset.
    - Study the correlation between chlorophyll and other variables (Linear Regression, GAM, data transformations).
    - Select variables (Lasso, single variable regression, multistep regression).
    - Study the regional aggregation (ACF).
    - Explore spatiotemporal correlations (hovmoller plots, PCA, variograms).
    - Estimate the Bayes factor.

Expected Outcomes

- A cleaned dataset that can be used in the following tasks
- A comprehensive exploration of the available data for chlorophyll study 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 hypothesis about the chlorophyll dynamics in the Red Sea

Work Accomplished and Prelimnary results

Chapter 2: Forecasting Chlorophyll Concentration in Regional Aggregates

Duration: 2 months (by February 2015)

Submission: Progress in Oceanography

Collaborator: Dionysios Raitsos

Chlorophyll data is very complex. 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 results. 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?
- Can a simple forecasting model allow us to understand the causes of chlorophyll blooms?

- Can the current hypothesizes about the seasonal chlorophyll dynamics be validated?

Method

1. Define datasets (training and test datasets, cross-validation)
2. Variable selection (Lasso, L1 regression, single-variable linear regression)
3. Define regional aggregations (unsupervised learning, Hierarchical clustering, K-means)
4. Forecasts chlorophyll concentration (linear regression, GAM models, diagnostic, k-nearest neighbors)
5. Predicting future extreme blooms (nearest-neighbours, logistic regression, decision trees)

Expected Outcomes

- A regional division of the Red Sea that has been quantitatively evaluated.
- A critical evaluation of current hypothesis about the chlorophyll dynamics in the Red Sea.
- A lower bound on the performance of a more sophisticated model.
- An assessment of the limitation of aggregate methods for Chlorophyll data.
- An understanding on how the treatment of spatial correlations can improve the results.

Work Accomplished and Prelimnary results

Chapter 3: Global Geostatistical Model for Chlorophyll Forecasting

Duration: 1 month (by March 2015)

Submission: Spatial Statistics

Geostatistical methods can be used to construct dynamical models for forecasting the chlorophyll concentrations that we can compare to deterministic models. Geostatistics is a robust method to model spatio-temporal data. Recently there has been a lot of interest for expanding it to model spatio-temporal data. Through the use of Kriging interpolation, these models provide powerful tools for do 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. This linear model can 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 a global geostatistical model fit chlorophyll data?
- How non stationary is the data in time and space?
- What spatiotemporal covariance functions best fit the chlorophyll data?
- Can geostatistical methods be employed in a filtering setup?

Method This task has already been started and had been the object of a submission for publication. The remaining work includes:

- Use the new dataset and the new covariates

- Compare the results to the ones of with the regional aggregates

Expected Outcomes

- A methodology to employ a geostatistical model in a filtering problem.
- A characterization of the space-time non stationarity of the data, and the interaction of the temporal and spatial dimensions.
- An understanding of how spatial aggregation and geostatistical models can be used in the same model.

Work Accomplished and Preliminary Results

Chapter 4: Local Geostatistical Model for Chlorophyll Forecasting

Duration: 3 months (by June 2015)

Submission: Journal of the American Statistical Association (Case Study)

Collaborator: Raphael Huser

This part will bring together the results of the two preceding tasks to develop a predictive 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 fine-scale dynamics of chlorophyll in each region?
- Can the fine scale behaviour of phytoplankton be accurately predicted?
- What are the spatial features that are important for the chlorophyll dynamics?

Method

- Extract local dataset from previous tasks.
- Design the training and test datasets, and the cross-validation method.
- Design and evaluate the mean function given the past covariates.
- Fit the local geostatistical model to the residuals.
- Evaluate the model predictions and compare the results with task 2 and 3.

Expected Outcomes

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

Work Accomplished and Preliminary Results

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

Duration: 3 months (by September 2015)

Submission: Journal of Geophysical Research

Collaborator: George Triantafyllou / Boujeema

The three previous tasks focus on constructing increasingly sophisticated statistical predictive models for the chlorophyll concentration in the Red Sea. In this part these models will be compared to a 1D ecological model (ERSEM). This model is well detailed and very complex. 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. To allow for comparison, the model will be run in each of the regions found in task 2. Available data will also be assimilated to the model through an ensemble Kalman-based smoothing assimilation scheme that will use an expectation-maximization algorithm for parameters estimation.

Open Questions

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

Method

1. Define the metrics for comparison.
2. Calibrate the ERSEM model on each of the regions.
3. Define an assimilation scheme and the data for the ERSEM model.
4. Implement the assimilation scheme.
5. Run the simulation and aggregate the results.
6. Conduct the comparisons with the statistical models.

Expected Outcomes

- A complete set of measures of the prediction skills of each approach.
- A method to estimate parameters in an assimilated ecological model.
- A set of case studies of the behaviours of each method for forecasting interesting 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.

Work Accomplished and Preliminary Results

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

Duration: 3 months (by September 2015)

Submission: Journal of Geophysical Research

Collaborator: George Triantafyllou / Boujeema

In the previous task, we proposed to compare the forecasts of the ecological ERSEM model to that of the statistical models we proposed from tasks 2 to 4. In this task 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, and hopefully leading to improved forecast. When real observations will be available, they will be assimilated sequentially. 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 model?
- Would the access to information about other regions improve the model forecasts?
- What are the global patterns of ecological dynamics in the Red Sea?

Method

1. Define new assimilation scheme

2. Define metrics to measure model improvement
3. Prepare training and test datasets
4. Train statistical model
5. Run simulation with assimilation of statistical observation
6. Compare with results of task 5

Expected Outcomes

- An improvement in the prediction skills of the deterministic approach
- A methodology to couple deterministic ecological models through statistical models
- Insights on the global ecological dynamics of the Red Sea

Work Accomplished and Preliminary Results

Preliminary Results

Data Loading

Most of the data necessary for the analysis has been gathered. Almost all of it is freely available on Internet. Shell and R scripts have been written and used for loading the following data all the data excepted the wind. The latter has been provided by Yesubabu Viswanadhapalli, and is the result of the assimilation of QuickScat satellite and in situ data to the WRF regional wind model. There are additional datasets that could be interesting to use in the analysis, but they are mainly climate mode time-series like IMI and EAWR, that are easy to download. More details about the data can be found in table xxx in the appendice.

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 of these variables 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.

Red Sea Chlorophyll Data Exploration

The MODIS and CCI chlorophyll data products have been explored and compared. The MODIS data presents an important amount of missing data that can reach 100% during summer in the southern Red Sea, making any analysis in this region very difficult. The CCI data product solve 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. The increases the coverage to 70% during summer month.

In a previous study (in appendice), the SeaWiFS chlorophyll data has been used. The seasonal signal is 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 anomlaly at the same location from one week to the other is correlated at 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 disappeared, suggesting that the causes of seasonal and interannual variability are distinct.

Red Sea Ecoregion Clustering

Global Geostatistical Model

Regional 1D Assimilated Ecological Model

REFERENCES

[Abualnaja et al., 2015] Abualnaja, Y., Papadopoulos, V. P., Josey, S. A., Hoteit, I., Kontoyiannis, H., and Raitsos, D. E. (2015). Impacts of climate modes on air-sea heat exchange in the red sea. *Journal of Climate*, page 150106132132005.

[Acker et al., 2008] Acker, J., Leptoukh, G., Shen, S., Zhu, T., and Kempler, S. (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.

[Alvera-Azcarate et al., 2007] Alvera-Azcarate, A., Barth, A., Beckers, J. M., and Weisberg, R. H. (2007). Multivariate reconstruction of missing data in sea surface temperature, chlorophyll, and wind satellite fields. *Journal of Geophysical Research-Oceans*, 112(C3).

[Anderson, 2005] Anderson, T. R. (2005). Plankton functional type modelling: running before we can walk? *Journal of Plankton Research*.

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

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

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

[Butenschon and Zavatarelli, 2012] Butenschon, M. and Zavatarelli, M. (2012). A comparison of different versions of the SEEK Filter for assimilation of biogeochem-

ical data in numerical models of marine ecosystem dynamics. *Ocean Modelling*, 54-55:37–54.

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

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

[Ciavatta et al., 2011] Ciavatta, S., Torres, R., Saux-Picart, S., and Allen, J. I. (2011). Can ocean color assimilation improve biogeochemical hindcasts in shelf seas? *Journal of Geophysical Research-Oceans*, 116.

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

[Fontana et al., 2013] Fontana, C., Brasseur, P., and Brankart, J. M. (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.

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

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

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

[Gokaraju et al., 2011] Gokaraju, B., Durbha, S. S., King, R. L., and Younan, N. H. (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.

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

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

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

[Jeong et al., 2006] Jeong, K. S., Kim, D. K., and Joo, G. J. (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.

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

[Korres et al., 2012] Korres, G., Triantafyllou, G., Petihakis, G., Raitsos, D. E., Hoteit, I., Pollani, A., Colella, S., and Tsiaras, K. (2012). A data assimilation tool for the Pagasitikos Gulf ecosystem dynamics: Methods and benefits. *Journal of Marine Systems*, 94:S102–S117.

[Lee et al., 2003] Lee, J. H. W., Huang, Y., Dickman, M., and Jayawardena, A. W. (2003). Neural network modelling of coastal algal blooms. *Ecological Modelling*, 159(2-3):179–201.

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

[McClain, 2009] McClain, C. R. (2009). A decade of satellite ocean color observations. *Annual Review of Marine Science*, 1:19–42.

[Miles and He, 2010] Miles, T. N. and He, R. (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.

[Mulia et al., 2013] Mulia, I. E., Tay, H., Roopsekhar, K., and Tkalich, P. (2013). Hybrid ANN-GA model for predicting turbidity and chlorophyll-a concentrations. *Journal of Hydro-Environment Research*, 7(4):279–299.

[North et al., 2011] North, G. R., Wang, J., and Genton, M. G. (2011). Correlation models for temperature fields. *Journal of Climate*, 24(22):5850–5862.

[Pal, 2014] Pal, R. (2014). *An introduction to phytoplanktons : diversity and ecology*. Springer, New York.

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

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

[Raitsos et al., 2011] Raitsos, D. E., Hoteit, I., Prihartato, P. K., Chronis, T., Triantafyllou, G., and Abualnaja, Y. (2011). Abrupt warming of the red sea. *Geophysical Research Letters*, 38(14).

[Raitsos et al., 2013] Raitsos, D. E., Pradhan, Y., Brewin, R. J., Stenchikov, G., and Hoteit, I. (2013). Remote sensing the phytoplankton seasonal succession of the Red Sea. *PLoS One*, 8(6).

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

[Richlen et al., 2010] Richlen, M. L., Morton, S. L., Jamali, E. A., Rajan, A., and Anderson, D. M. (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.

[Robinson, 2010] Robinson, I. S. (2010). *Discovering the ocean from space : the unique applications of satellite oceanography*. Springer praxis series geophysical sciences 4110. Springer, New York, 1st edition.

[Sirjacobs et al., 2011] Sirjacobs, D., Alvera-Azcrate, A., Barth, A., Lacroix, G., Park, Y., Nechad, B., Ruddick, K., and Beckers, J.-M. (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.

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

[Steinmetz et al., 2011] Steinmetz, F., Deschamps, P. Y., and Ramon, D. (2011). Atmospheric correction in presence of sun glint: application to meris. *Optics Express*, 19(10):9783–9800.

[Taylor et al., 2013] Taylor, M. H., Losch, M., Wenzel, M., and Schroter, J. (2013). On the sensitivity of field reconstruction and prediction using empirical orthogonal functions derived from gappy data. *Journal of Climate*, 26(22):9194–9205.

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

[Triantafyllou et al., 2013] Triantafyllou, G., Hoteit, I., Luo, X., Tsiaras, K., and Petihakis, G. (2013). Assessing a robust ensemble-based Kalman filter for efficient ecosystem data assimilation of the Cretan Sea. *Journal of Marine Systems*, 125:90–100.

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

[Waite and Mueter, 2013] Waite, J. N. and Mueter, F. J. (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.

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

[Weikert, 1987] Weikert, H. (1987). *Plankton and the pelagic environment*, pages 90–111. Pergamon Press, Oxford.

[Wu et al., 2014] Wu, N. C., Huang, J. C., Schmalz, B., and Fohrer, N. (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.

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

[Zhai and Bower, 2013] Zhai, P. and Bower, A. (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.

[Zhan et al., 2014] Zhan, P., Subramanian, A. C., Yao, F. C., and Hoteit, I. (2014). Eddies in the red sea: A statistical and dynamical study. *Journal of Geophysical Research-Oceans*, 119(6):3909–3925.

APPENDICES

A Appendix A Title

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.

B Appendix B Title

Your content goes here.

C Papers Submitted and Under Preparation

• Author 1 Name, Author 2 Name, and Author 3 Name, "Article Title", *Submitted to Conference/Journal Name*, further attributes.

• Author 1 Name, Author 2 Name, and Author 3 Name, "Article Title", *Submitted to Conference/Journal Name*, Mon. Year.