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

Examination Committee Approval **2**

Copyright **3**

List of Figures **7**

List of Tables **8**

1 Motivation **9**

1.1 Importance of phytoplankton 9
1.2 Measuring Chlorophyll Concentration 10
1.3 Primary productivity in the Red Sea 11
1.4 Chlorophyll Concentration Prediction 12
1.5 Modeling Chlorophyll Concentration 13

2 Plan **15**

2.1 Task 1: Dataset building and Exploration 15
2.1.1 Motivation 15
2.1.2 Open Questions 16
2.1.3 Method and Work Done 16
2.1.4 Expected Outcomes 17
2.2 Task 2: Forecasting Chlorophyll Concentration in Regional Aggregates 18
2.2.1 Motivation 18
2.2.2 Open Questions 18
2.2.3 Method 19
2.2.4 Expected Outcomes 19
2.3 Task 3: Global geostatistical model for forecasting 20
2.3.1 Motivation 20
2.3.2 Open Questions 20
2.3.3 Method 21
2.4 Task 4: Local geostatistical model for forecasting 22

2.4.1 Motivation . . . 22
2.4.2 Open Questions . . . 22
2.4.3 Method . . . 23
2.4.4 Expected Outcomes . . . 23
2.5 Task 5: Assimilation of 1D ecological models and comparison to statistical models . . . 24
2.5.1 Motivation . . . 24
2.5.2 Open Questions . . . 24
2.5.3 Method . . . 25
2.5.4 Expected Outcomes . . . 25
2.6 Task 6: Improving an Ecological Model Data Assimilation Scheme through Statistical Predictive Models . . . 26
2.7 Motivation . . . 26
2.7.1 Open Questions . . . 26
2.7.2 Method . . . 27
2.7.3 Expected Outcomes . . . 27

3 Miscellaneous 28
3.1 Red Sea large-scale phytoplankton dynamics . . . 28
3.2 Challenges with remotely-sensed chlorophyll data . . . 30
3.3 Data Assimilation for marine ecology models . . . 31
3.4 Statistical models for chlorophyll concentration . . . 32
3.5 Space-time geostatistical models for forecasting (Genton) . . . 33

4 Creating the List of Abbreviations and List of Symbols 34
4.1 The Problem is the dataset can be handled . . . 34
4.2 The Red Sea can be divided into regions with very different dynamics 34
4.3 There is enough correlation in the dataset to build predictive models 35
4.4 Geostatistical methods can used for forecast . . . 35

References 36

Appendices 37

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Motivation

1.1 Importance of phytoplankton

Phytoplankton are unicellular, photosynthetic algae that live in the upper layers of bodies of water (ocean, lakes, rivers or ponds). There a three main types of phytoplankton species: diatom (5-200m), dinoflagellates (5-200m) and cyanobacteria (¡5m). Diatoms and dinoflagellates are found in nutrient rich environments and multiply rapidly when conditions are favorable. On the other hand, cyanobacteria are capable of surviving in very oligotrophic (nutrient poor) environments like the Red Sea. 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 as it is at the basis of the marine food web. Zooplankton graze phytoplankton which are consumed by larger species. Higher concentrations of phytoplankton will therefore result in larger stock of predator fishes. High phytoplankton concentration also impact their environment by creating dead zone when they die and are decomposed by bacteria. Due to the rapid growth of phytoplankton, it responds very well to changes in its environment, making it a key parameter to monitor the water quality.

Due to phytoplankton place at the bottom of the marine food chain, it is a key factor for fisheries and the marine ecology. 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, which in turn feeds higher trophic levels. On the other hand, during 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.

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. Research is currently done to evaluate the way this biological pump works and how it affects climate.

1.2 Measuring Chlorophyll Concentration

Chlorophyll is a molecule that is food in algae, phytoplankton and plants that is critical for photosynthesis. Phytoplankton is a poor absorber of green light, and is responsible for the coloration of plants. When phytoplankton are present in high concentrations they change the water also takes a detectable green coloration (it can also take a red or blue coloration depending on the type of phytoplankton dominating). This offers a way to estimate the chlorophyll concentration of the water, which is a good proxy for phytoplankton concentration.

In-situ measurements of chlorophyll are however expensive and have limited temporal and spatial coverage. 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 the coverage is limited. Political issues, like in the Red Sea, can also be a practical barrier to in-situ measurements.

Satellite measurements of chlorophyll provide excellent proxies for phytoplankton concentrations with a good temporal and spatial coverage. 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. Despite some limitations, in particular of 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, it is often the most important data source.

1.3 Primary productivity in the Red Sea

Typical tropical seas (TTS), like the Red Sea, are characterized by a highly stratified structure, where warm nutrient-depleted surface water is separated by cold nutrient-rich by a steep gradient of temperature zone called pycnocline. The pycnocline acts as barrier that prevents nutrients to reach the surface water. As a result, TTS are oligotrophic and have low chlorophyll concentrations. Until recently, marine biologists have thought that TTS had therefore a very low productivity. However, recent investigations have contested this idea, that different upwelling mechanisms exist that bring new nutrients to the surface water.

Despite being an oligotrophic and challenging environments for marine life, the Red Sea presents a surprisingly rich and diverse ecosystem. Most of it lives in the very developed coral reef system. 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.

Remotely-sensed chlorophyll data show an important seasonality of the Red Sea primary productivity, that has been linked to winter deep mixing, and the inversion of the wind direction in the southern Red Sea, enhancing intrusion of nutrient rich Gulf of Aden water. Despite this strong seasonality, there is a large interannual variability caused by the unpredictable occurrence of large phytoplanktonic blooms. Diverse causes have been hypothesized for these blooms such as wind-induced mixing, eddies or dust storms carrying nutrients.

Although the Red Sea environment is relatively preserved, it is under increasing pressure due to human activities. An abrupt increase of temperature has occurred in the last decade that threatens the fragile coral reef system. Moreover, the increasing urbanization and fishing activity contribute to the fragilization of this unique ecosystem.

1.4 Chlorophyll Concentration Prediction

Models can be useful to identify causes behind the chlorophyll patterns we observe in the Red Sea. Many hypotheses have be 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 circulation and ecological models. However, the interaction between the open sea and coral reefs and the role of sand storms has not been investigated yet. Models, can also be helpful in discovering new dynamics affecting the chlorophyll concentration. In particular, the interaction between the productivity level of the different regions of the Red Sea

has not been studied yet.

Model predictions for chlorophyll concentration can also have practical applications. Phytoplankton blooms can be harmful to humans and marine life and are closely monitored in many regions of the world. 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. Anticipating a phytoplanktonic bloom might therefore be helpful in taking preventive actions. Finally, due to their short life-cycle, phytoplankton concentration reacts quickly to changes the environment, making it a key variable in water quality monitoring.

1.5 Modeling Chlorophyll Concentration

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 level, or as complex as the European regional seas ecosystem model (ERSEM) that has dozens of variables and represent many ecological, biological and chemical interactions. Such a model has been couple to the MITgcm circulation model used to simulate the Red Sea ecology. However the complexity of these models makes them difficult to deploy and interpret their results.

On the other hand, data-driven statistical models are relatively easier to apply. They are relevant when the phenomenon producing the data is very complex or poorly known. They have been applied to predict chlorophyll concentration, mostly in small regions that have complex dynamics. Some statistical models, such as linear regression, GAM or tree regression have the advantage of being easy to interpret, and can

be used to understand the dynamics driving the chlorophyll concentration.

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. There is however no clear cut division between regions with different pattern, making it difficult to divide the Red Sea into regions. Finally the different regions of the Red Sea are believed to interact. A model is therefore needed to account for the spatial and temporal interaction of the chlorophyll.

Classical geostatistics is the most widely used spatial statistical models. 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. Space-time geostatistics has been applied to many environment studies, but not to chlorophyll data yet.

Chapter 2

Plan

2.1 Task 1: Dataset building and Exploration

Duration: 2 months (by December 2014)

Submission: Journal of Marine Systems

Collaborator: Dionysios Raitsos

2.1.1 Motivation

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.

2.1.2 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?

2.1.3 Method and Work Done

1. Identify data sources and load the data60
2. Clean the data and fill missing values (DINEOF).....50
3. Align and format the data in order to have a unique dataset...0
4. Explore the dataset..20
    - 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/

2.1.4 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

2.2 Task 2: Forecasting Chlorophyll Concentration in Regional Aggregates

Duration: 2 months (by February 2015)

Submission: Progress in Oceanography

Collaborator: Dionysios Raitsos

2.2.1 Motivation

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.

2.2.2 Open Questions

- Is the biological aggregation of the Red Sea proposed by (Raitsos 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?

2.2.3 Method

1. Define datasets (training and test datasets, cross-validation).....0
2. Variable selection (Lasso, L1 regression, single-variable linear regression).....0
3. Define regional aggregations (unsupervised learning, Hierarchical clustering, K-means).....50
4. Forecasts chlorophyll concentration (linear regression, GAM models, diagnostic, k-nearest neighbors).....0
5. Predicting future extreme blooms (nearest-neighbours, logistic regression, decision trees).....0

2.2.4 Expected Outcomes

- A regional division of the Red Sea that has been quantitatively evaluated.
- A critic 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.

2.3 Task 3: Global geostatistical model for forecasting

Duration: 1 month (by March 2015)

Submission: Spatial Statistics

2.3.1 Motivation

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 research on expanding it to model spatio-temporal data. With Kriging, these models a are powerful ways to 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 setup, as it is similar to the way deterministic models are employed to do forecasts given past observations.

2.3.2 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?

2.3.3 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.

2.4 Task 4: Local geostatistical model for forecasting

Duration: 3 months (by June 2015)

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

Collaborator: Raphael Huser

2.4.1 Motivation

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 developed 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 do local prediction. These models will have access to aggregate covariates from neighboring regions in represent the global scale behaviours.

2.4.2 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 predicted accurately?
- What are the spatial features that are important for the chlorophyll dynamics?

2.4.3 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.

2.4.4 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.

2.5 Task 5: Assimilation of 1D ecological models and comparison to statistical models

Duration: 3 months (by September 2015)

Submission: Journal of Geophysical Research

Collaborator: George Triantafyllou / Boujeema

2.5.1 Motivation

The three previous tasks focus on constructing increasingly sophisticated 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 a smoothing assimilation scheme that will use an expectation-maximization algorithm for parameters estimation.

2.5.2 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?

2.5.3 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. Do the comparisons with the statistical models

2.5.4 Expected Outcomes

- A complete set of measures of the prediction skills of each approach.
- A method to estimate assimilation and model 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 two approaches can complement each other.

2.6 Task 6: Improving an Ecological Model Data Assimilation Scheme through Statistical Predictive Models

Duration: 3 months (by September 2015)

Submission: Journal of Geophysical Research

Collaborator: George Triantafyllou / Boujeema

2.7 Motivation

In the previous task, we compared the forecasts of the ecological ERSEM model to that of the statistical models we developed 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 will give an 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.

2.7.1 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?

2.7.2 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

2.7.3 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

Chapter 3

Miscellaneous

3.1 Red Sea large-scale phytoplankton dynamics

Despite an increasing number of study for the last two decades, the large-scale phytoplankton dynamics of the Red Sea remains largely unknown [Raitsos 2013, Triant 2014] [1]. The Red Sea is deficient in the major nutrients [Weikert 1987], and the only significant input of fresh water comes from the Gulf of Aden. This explains a general increase of chlorophyll concentration from north to south with the NCRS having the lowest concentration [raitsos 2013]. The Red Sea also displays a distinct seasonality, with the highest chlorophyll concentration in winter. A minor summer peak is also observed around July, everywhere except in the NRS [Raitsos]. Despite this regularity, a strong interannual variability is observed, with blooms that can reach mesotrophic concentration levels [raitsos 2013]. According to [Triant 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 chlorophyll concentration.

The exchange of water with the nutrient-rich Gulf of Aden is a major driving mechanism for the whole Red Sea [Triant. 2014]. It is the most important source

of fresh water and nutrients. The maximum chlorophyll concentration observed in the SRS during the summer is due to the wind-driven water intrusion [Raitsos 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 in as the latitude increases, explaining the lowest concentration in the northern half of the Red Sea [Raitsos 2013].

Stratification and deep convection also play an important role in allowing nutrient-rich deep waters to mix with waters of the euphotic zone. The vertical mixing is the most vigorous in the NRS during the winter, explaining why its chlorophyll concentration is higher than in the NCRS, where there is a lack of mixing [Raitsos 2013]. The NRS mixing is believed to be driven by wind [Raitsos 2013].

The Red Sea circulation is strongly influenced by mesoscale eddies that impact primary production [KIM 2011]. In particular, the anti-cycloninc eddy in the CRS is believed to control the June concentration peak and summer productivity in this region, by transporting nutrients and/or phytoplankton from the adjacent coral reefs [Raitsos 2013]. The cold core eddy in the NRS also plays a role in enhancing the vertical mixing in that region.

Aerial depositions could also be an important input of nutrient for the Red Sea, but it has been largely left unexplored [Triant 2014]. [Raitsos 2013] noticed for example that the Sand storms in the Red Sea most frequently happen in June and July, coinciding with the summer chlorophyll peak.

3.2 Challenges with remotely-sensed chlorophyll data

Chlorophyll remotely-sensed data in the Red Sea suffer from many problems that need to taken into account before using them. Until recently, the data coverage in the southern Red Sea was almost 0% during the summer due to sunglint, clouds and aerosols [Racault]. However the OC-CCI data product that merges different chlorophyll data sources considerably increases the Red Sea coverage. However, this new dataset has not been used to revisit the assumption made in the large-scale Red Sea phytoplankton productivity. If for case I open water, the data in the Red Sea has been shown to be of quality comparable with that the rest of the world [Brewin 2013], case II waters remain a challenge, especially in the SRS where the water is particularly shallow and the coral reef extended [Triant 2014, Raitsos 2013]. Theses problems generally lead to an overestimation of the chlorophyll level, but not necessarily to erroneous values [Raitsos 2013].

One of the most popular algorithm for the data filing of remotely-sensed data is DINEOF. It is an EOF based data filling approach introduced by [Beckers]. It has been used for multivariate reconstruction of SST fields using chlorophyll data in [Alvera 2007]. In [Sicarjobs 2011], it has been employed to fill chlorophyll data with 70% of missing values. [Taylors 2013] has compared DINEOF with other EOF-based reconstruction algorithms and shown that the former is the best method for data filling. DINEOF has been employed in several other chlorophyll studies [miles 2007, Waite]

3.3 Data Assimilation for marine ecology models

Data assimilation schemes are used to improve the simulations of ecological dynamics models by correcting their predictions with observations. The most common use of data assimilation is to improve the forecast of an ecological model, by providing it with an estimated initial state. 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 2012]. The deployment of such a forecasting system in the Red Sea is currently under study [Triant. 2014]. Hindcasting, the estimation of unobserved variables, is another application of assimilation scheme. [Ciavatta 2011] showed that he could improve the seasonal and annual hindcast of non assimilated biogeochemical properties in a shelf area (Wester English Channel). Finally, data assimilation can be used for reanalysis, to provide estimates of past years biogeochemical variables [Fontana 2013].

In the marine ecology modeling community, two assimilation schemes flavours have been widely used: Ensemble Kalman filters (EnKF) and the Singular Evolutive Extended Kalman filter (SEEK). The Stochastic EnKF, a Monte-Carlo approximation of the Kalman Filter, has been used in [Ciavatta 2011 and 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. The Singular Evolutive Interpolated Kalman filter (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 by [Trian 2012 and Korres 2012]. 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 in a low dimensional space. SEEK has a long history in data assimilation for marine ecology models and

is still used in recent studies [Fontana 2013, Korres 2012, Butenschon 2012]. [Korres 2012] shows that SEIK and SEEK are both comparably robust methods for highly non linear systems.

Current assimilation schemes are however affected by problems that have been addressed in the past years. First, biogeochemical variable are usually positive concentration, whereas Kalman filters expect Gaussian variables, and log-transformation can fail at solving this issue [Ciavatta 2011]. However, [Fontana 2013] has successfully introduced Gaussian anamorphosis transformations to solve this issue. Second, ecological blooms are intermittent and highly nonlinear, conditions that are challenging for assimilation schemes [Triant 2012, Korres 2012]. Third, SEIK and SEEK both project the error covariance in a subspace, resulting in an underestimation of the estimation error. [Butenschon 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. [Triant 2012] proposes to use the H_∞ method with SEIK in order remove this requirement.

3.4 Statistical models for chlorophyll concentration

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 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 2014]

on the west coast of South Korea, and by [Camps-Valls 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 complicated to use 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 2007], temporal recurrent recursive neural network have been used and found superior to traditional time-series model for daily forecasts of chlorophyll concentration. [Wang 2013] also used recurrent neural networks for daily chlorophyll forecasting in Lake Taihu, China. [Mulia 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 2013] used neural networks for the forecasting of algal bloom with one and two weeks lags in the coastal waters of Hong-Kong.

3.5 Space-time geostatistical models for forecasting (Genton)

Chapter 4

Creating the List of Abbreviations and List of Symbols

4.1 The Problem is the dataset can be handled

- CHL Outliers can be removed
- Data can be filled with DINEOF
- Data can be aligned and aggregated

4.2 The Red Sea can be divided into regions with very different dynamics

- North/South mean
- North/Souch difference in seasonality
- Blooms are localized

4.3 There is enough correlation in the dataset to build predictive models

- Correlation study between CHL and individual predictors
- Percentage of variance explained in CHL
- Estimating the Bayes Factor (Nearest-neighbor)

4.4 Geostatistical methods can used for forecast

- Demonstrated in Paper 1

REFERENCES

[1] D. E. Raitsos, Y. Pradhan, R. J. Brewin, G. Stenchikov, and I. Hoteit, "Remote sensing the phytoplankton seasonal succession of the red sea," *PLoS One*, vol. 8, no. 6, p. e64909, 2013, raitsos, Dionysios E Pradhan, Yaswant Brewin, Robert J W Stenchikov, Georgiy Hoteit, Ibrahim eng Research Support, Non-U.S. Gov't 2013/06/12 06:00 PLoS One. 2013 Jun 5;8(6):e64909. doi: 10.1371/journal.pone.0064909. Print 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23755161

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.