Climate change induced uncertainties in future coastal ecosystem state

Mészáros, L.

DOI
10.4233/uuid:ba1e9415-8339-4a3c-8034-67db565be793

Publication date
2023

Document Version
Final published version

Citation (APA)
Mészáros, L. (2023). *Climate change induced uncertainties in future coastal ecosystem state*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:ba1e9415-8339-4a3c-8034-67db565be793

CLIMATE CHANGE INDUCED UNCERTAINTIES IN FUTURE COASTAL ECOSYSTEM STATE

CLIMATE CHANGE INDUCED UNCERTAINTIES IN FUTURE COASTAL ECOSYSTEM STATE

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus prof. dr. ir. T.H.J.J. van der Hagen
chair of the Board for Doctorates,
to be defended publicly on Friday 30 June 2023 at 12:30 o'clock

by

Lőrinc MÉSZÁROS

Master of Science in Euro Hydro-informatics and Water Management,
Erasmus Mundus - Euroaquae Joint Master Program,
born in Nyíregyháza, Hungary.

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus	chairperson
Prof. dr. ir. G. Jongbloed	Delft University of Technology, promotor
Prof. dr. ir. F. H. van der Meulen	Vrije Universiteit Amsterdam, promotor
Dr. ir. G. Y. H. El Serafy	Delft University of Technology, copromotor
Independent members:	
Prof. dr. A. Taramelli	Università di Pavia, Italy
Prof. dr. E.A. Cator	Radboud University Nijmegen, The Netherlands
Prof. dr. ir. H.W.J Russchenberg	Delft University of Technology
Prof. dr. ir. A.W. Heemink	Delft University of Technology
Prof. dr. H.M. Schuttelaars	Delft University of Technology, reserve member

Keywords: climate change, uncertainty quantification, coastal phytoplankton phenology, Bayesian models, data fusion, multivariate analysis

Printed by: Ipskamp Printing

Front & Back: L. Mészáros using generative AI

ISBN 978-94-6366-700-5

An electronic version of this dissertation is available at
http://repository.tudelft.nl/.

As far as the laws of mathematics refer to reality,
they are not certain;
and as far as they are certain,
they do not refer to reality.

Albert Einstein

CONTENTS

Summary ix

Samenvatting xi

1 Introduction 1
- 1.1 Uncertainty as an inherent component of future predictions . . . 2
- 1.2 Objectives and Outline . . . 5

2 The value of probabilistic predictions for coastal ecosystems 11
- 2.1 Introduction . . . 12
- 2.2 Modelling instrument . . . 13
- 2.3 Data sources . . . 15
- 2.4 Ensemble forecasting Methodology . . . 16
- 2.5 Selection of important parameters . . . 18
- 2.6 Sampling procedure . . . 21
- 2.7 Forecast verification and Performance analysis . . . 27
- 2.8 Results and Discussion . . . 27
- 2.9 Conclusions . . . 32

3 Statistical underpinning of atmospheric variables selection 35
- 3.1 Introduction . . . 37
- 3.2 Materials and Methods . . . 39
  - 3.2.1 Dataset . . . 40
  - 3.2.2 Two-way and multi-way methods . . . 44
  - 3.2.3 Dynamic Factor Analysis . . . 50
  - 3.2.4 Functional PCA . . . 52
- 3.3 Results . . . 54
  - 3.3.1 Comparing two-way and multi-way methods . . . 54
  - 3.3.2 Dynamic Factors . . . 55
  - 3.3.3 Functional Principal Components . . . 56
- 3.4 Discussion . . . 60
- 3.5 Conclusions . . . 63

4 Bayesian stochastic climate generator 65
- 4.1 Introduction . . . 66
- 4.2 Dataset . . . 68
- 4.3 Stochastic generator methodology . . . 70
  - 4.3.1 Bias correction . . . 71
  - 4.3.2 Temporal evolution . . . 71
  - 4.3.3 Time series model definition . . . 73

4.3.4 Prior specification . . . 79
4.3.5 Gibbs sampler for drawing from the posterior distribution . . . 80
4.4 Propagation of uncertainty - Demonstration case . . . 81
4.5 Results . . . 84
4.5.1 Results of stochastic generator . . . 84
4.5.2 Results of probabilistic water quality simulation . . . 86
4.6 Conclusions and Recommendations . . . 90

5 Climate Change Induced Trends and Uncertainties in Phytoplankton Spring Bloom Dynamics 93
5.1 Introduction . . . 94
5.2 Materials and Methods . . . 98
5.2.1 Data sources . . . 99
5.2.2 Data fusion of chlorophyll-a measurements . . . 104
5.2.3 Long term projection using Bayesian structural time series models . 107
5.2.4 Tracking phytoplankton spring bloom dynamics . . . 111
5.3 Results . . . 114
5.3.1 Fused chlorophyll-a concentration signal . . . 114
5.3.2 Long term chlorophyll-a projection . . . 114
5.3.3 Changes in phytoplankton bloom dynamics . . . 118
5.4 Discussion . . . 124

6 Conclusions and discussion 129
6.1 Overarching conclusions . . . 129
6.2 Main limitations . . . 133

Acknowledgements 137

Bibliography 138

Appendix 163

Curriculum Vitæ 165

List of Publications 167

SUMMARY

This thesis presents a doctoral research where statistical concepts and techniques are applied to problems at the interface of marine and atmospheric processes. The research was conducted at the Statistics group of the Delft Institute of Applied Mathematics (TU Delft) and the Marine and Coastal unit of Deltares. The main objective of the work is to provide statistical tools to understand multi-dimensional climate and marine environmental datasets, as well as to offer ways for quantifying the uncertainties in the coastal ecological response that are driven by the climatic variation. Statistical quantification of uncertainties in data, models and predictions is therefore the central topic of the thesis.

The research is built on open source data (in-situ and satellite measured as well as numerically modelled) from the Copernicus Marine Environment Monitoring Service, the Dutch Directorate-General for Public Works and Water Management (Rijkswaterstaat), the Royal Netherlands Meteorological Institute, and the Euro-CORDEX regional climate modelling experiment. It also uses the open source numerical modelling software Delft3D from Deltares. All other statistical models and algorithms developed during the research are published and available open source.

The thesis starts by demonstrating the value of probabilistic predictions and uncertainty quantification for coastal ecosystems. That is done by constructing an ensemble modelling framework where certain chosen numerical model inputs and model process parameters are perturbed, to which the simulated coastal chlorophyll-a concentration is sensitive. The model perturbation was implemented using Latin Hypercube Sampling with Dependence (LHSD), and more than 150 ensemble members were produced using the Delft3D model. This ensemble prediction system is then compared to the deterministic model setup. A range of verification metrics that describe the goodness-of-fit, accuracy, reliability, and discrimination properties of both modelling experiments were computed. Apart from the verification metrics, the value of probabilistic predictions was also showcased by evaluating the benefit of having temporal and spatial estimates of uncertainty by producing ensemble band, predictive uncertainty intervals and standard deviations maps.

In Chapter 3 of the thesis, we work towards the quantification of climate change induced uncertainties in coastal phytoplankton response. The first necessary step is a comprehensive data exploration and dimension reduction, which also provides a statistical underpinning of atmospheric variable selection for the climate impact studies conducted later in the thesis. Here a range of existing dimension reduction techniques are described and applied to seven atmospheric variables (air temperature, solar radiation, eastward wind, northward wind, air pressure, relative humidity, and total cloud cover) and the chlorophyll-a data at hand. These techniques are applied in a structured way to include spatial and temporal correlation, as well as functional features in the multi-dimensional data. The applied methods include Principal Component Analysis (PCA), Principal Component Regression (PCR), Partial Least Squares (PLS) Regression,

multi-way models (PARAFAC, Tucker and N-PLS), Dynamic Factor Analysis (DFA), and Functional PCA. Room for dimension reduction in the atmospheric data was identified, underlying temporal patterns in the chlorophyll-a signal at different locations were revealed, structural similarities (characterized by a mean function and functional variation) in the Euro-CORDEX climate projections were found, and the most influential atmospheric variables (solar radiation and air temperature) were chosen.

Building on these findings, we propose a way to quantify uncertainties in the climate scenarios that are used for the climate impact studies. The basis of this research step is the development of a stochastic climate generator, which is first tested on the solar radiation variable. This climate generator takes the existing Euro-CORDEX scenarios (a combination of Representative Concentration Pathways and Generic Circulation Model forcings) and enriches them by generating numerous new synthetic scenarios around them. These new generated scenarios are representative of the original ones due to the way the stochastic climate generator is constructed. The basis of the climate generator is a Bayesian multi-layered (hierarchical) model. In this model there are model parameters representing variation in the long term trend, seasonal amplitude, time shift, and additive residual. The generator estimates the distribution of each model parameter with Bayesian inference, and using data from all scenarios. Then, when sampling from the parameter distributions, numerous climate trajectories can be constructed. The climate generator is successfully tested on the solar radiation variable and the generated synthetic radiation projections are used in a demonstration study where uncertainties are further propagated to chlorophyll-a concentrations using the Delft3D numerical model.

In the final research step of the thesis, this Bayesian stochastic generator is extended to air temperature. This way we have numerous (>100) radiation and temperature projections available to propagate climate induced uncertainties to coastal chlorophyll-a response once again, this time covering the entire 21^{st} century. In order to translate the climate signal into chlorophyll-a response, we make use of a Bayesian structural time series model. This model follows a piecewise linear trend and continues to repeat its multi-seasonal behavior, learnt from the past data, and most importantly also includes linear effects of the two climate variables. For the training of this time series model, we construct a historical chlorophyll-a signal by fusing in-situ and satellite measurements. This fused signal helps us to take advantage of the more frequent satellite measurements while correcting them with the more accurate in-situ measurements that are also available for a longer historical period. The Bayesian structural time series model is then trained on the fused chlorophyll-a signal and used for long term projection, taking the generated radiation and temperature scenarios as regressors. Since our main interest is the phytoplankton spring bloom dynamics, as a last step we extract yearly spring bloom cardinal dates (beginning, peak, end) from the long-term chlorophyll-a projections using a non-parametric shape constrained method (log-concave regression). The final result is therefore the estimation of climate change induced uncertainty in the coastal phytoplankton spring bloom dynamics.

SAMENVATTING

Dit proefschrift presenteert een promotieonderzoek waarin statistische concepten en technieken worden toegepast op problemen op het grensvlak van marine en atmosferische processen. Het onderzoek is uitgevoerd bij de sectie Statistiek van het Delft Institute of Applied Mathematics (TU Delft) en de unit Marine en Kustsystemen van Deltares. Het hoofddoel van het werk is om statistische hulpmiddelen te bieden om multidimensionale klimaat- en marine milieudatasets te begrijpen, en om manieren te bieden voor het kwantificeren van de onzekerheden in de ecologische respons die worden aangedreven door de klimaatvariatie. Het statistisch kwantificeren van onzekerheden in data, modellen en voorspellingen is daarom het centrale onderwerp van het proefschrift.

Het onderzoek is gebaseerd op open source data (in-situ en satelliet gemeten en numeriek gemodelleerd) van de Copernicus Marine Environment Monitoring Service, Rijkswaterstaat, het Koninklijk Nederlands Meteorologisch Instituut, en het Euro-CORDEX regionale klimaatmodelleringsexperiment. Het maakt ook gebruik van de open source numerieke modelleringssoftware Delft3D van Deltares. Alle andere statistische modellen en algoritmen die tijdens het onderzoek zijn ontwikkeld, zijn gepubliceerd en open source beschikbaar.

Het proefschrift begint met het aantonen van de waarde van probabilistische voorspellingen en onzekerheidskwantificering voor kustecosystemen. Dat wordt gedaan door een ensemble-modelleringskader te construeren waarin bepaalde gekozen numerieke modelinvoer en modelprocesparameters worden verstoord, waarvoor de gesimuleerde chlorofyl-a-concentratie aan de kust gevoelig is. De modelverstoring is geïmplementeerd met behulp van Latin Hypercube Sampling with Dependence (LHSD), en meer dan 150 ensembleleden werden geproduceerd met behulp van het Delft3D-model. Dit ensemble-voorspellingssysteem wordt vervolgens vergeleken met de deterministische modelopstelling. Een reeks verificatiestatistieken die de goodness-of-fit, nauwkeurigheid, betrouwbaarheid en discriminatie-eigenschappen van beide modelleringsexperimenten beschrijven, werden berekend. Afgezien van de verificatiestatistieken, werd de waarde van probabilistische voorspellingen ook aangetoond door het voordeel te evalueren van het hebben van temporele en ruimtelijke schattingen van onzekerheid door kaarten te produceren, ensembleband, voorspellende onzekerheidsintervallen en standaarddeviaties.

In hoofdstuk 3 gaat het over de kwantificering van door klimaatverandering veroorzaakte onzekerheden in de respons van kustfytoplankton. De eerste noodzakelijke stap is een uitgebreide gegevensverkenning en dimensiereductie, die ook een statistische onderbouwing biedt van de selectie van atmosferische variabelen voor de klimaatimpactstudies die later in het proefschrift worden uitgevoerd. Hier wordt een reeks bestaande dimensiereductietechnieken beschreven en toegepast op zeven atmosferische variabelen (luchttemperatuur, zonnestraling, oostelijke wind, noordelijke wind, luchtdruk, relatieve vochtigheid en totale bewolking) en de beschikbare chlorofyl-a-gegevens. Deze

technieken worden op een gestructureerde manier toegepast om zowel ruimtelijke en temporele correlatie als functionele kenmerken in de multidimensionale gegevens op te nemen. De toegepaste methoden omvatten Principal Component Analysis (PCA), Principal Component Regression (PCR), Partial Least Square (PLS) Regression, multi-way modellen (PARAFAC, Tucker en N-PLS), Dynamic Factor Analysis (DFA) en Functional PCA. Mogelijkheden tot dimensiereductie in de atmosferische data werden geïdentificeerd, onderliggende temporele patronen in het chlorofyl-a signaal op verschillende locaties werden onthuld, structurele overeenkomsten (gekenmerkt door een gemiddelde functie en functionele variatie) in de klimaatprojecties (Euro-CORDEX) werden gevonden, en de meest invloedrijke atmosferische variabelen (zonnestraling en luchttemperatuur) werden gekozen.

Voortbouwend op deze bevindingen wordt een manier voorgesteld om onzekerheden te kwantificeren in de klimaatscenario's die worden gebruikt voor de klimaatimpactstudies. De basis van deze onderzoeksstap is de ontwikkeling van een stochastische klimaatgenerator, die eerst wordt getest op de variabele zonnestraling. Deze klimaatgenerator neemt de bestaande Euro-CORDEX-scenario's (een combinatie van Representatieve Concentratieroutes en Generic Circulation Model-forceringen) en verrijkt deze door talloze nieuwe synthetische scenario's eromheen te genereren. Deze nieuw gegenereerde scenario's zijn representatief voor de originele vanwege de manier waarop de stochastische klimaatgenerator is geconstrueerd. De basis van de klimaatgenerator is een Bayesiaans hiërarchisch model. In dit model zijn er modelparameters die variatie in de langetermijntrend, seizoensamplitude, tijdverschuiving en additief residu vertegenwoordigen. De generator schat de verdeling van elke modelparameter via een Bayesiaanse aanpak en gebruikt gegevens uit alle scenario's. Vervolgens kunnen bij het nemen van steekproeven uit de parameterverdelingen talrijke klimaattrajecten worden geconstrueerd. De klimaatgenerator is met succes getest op de zonnestralingsvariabele en de gegenereerde synthetische stralingsprojecties worden gebruikt in een demonstratiestudie waar onzekerheden verder worden gepropageerd naar chlorofyl-a-concentraties met behulp van het Delft3D-numerieke model.

In het laatste hoofdstuk wordt deze Bayesiaanse stochastische generator uitgebreid tot luchttemperatuur. Op deze manier hebben we tal van (> 100) stralings- en temperatuurprojecties beschikbaar om klimaatgeïnduceerde onzekerheden voor chlorofyl aan de kust te verspreiden - een projectie die deze keer de hele 21^{st} eeuw beslaat. Om het klimaatsignaal te vertalen naar chlorofyl-a-respons, maken we gebruik van een Bayesiaans structureel tijdreeksmodel. Dit model volgt een stuksgewijs lineaire trend en blijft zijn seizoensgedrag herhalen, geleerd van de gegevens uit het verleden, en het belangrijkste is dat het ook lineaire effecten van de twee klimaatvariabelen omvat. Voor de training van dit tijdreeksmodel construeren we een historisch chlorofyl-a-signaal door in-situ- en satellietmetingen te fuseren. Dit gefuseerde signaal helpt ons om te profiteren van de frequentere satellietmetingen en deze te corrigeren met de meer nauwkeurige in-situ metingen die ook beschikbaar zijn voor een langere historische periode. Het Bayesiaanse structurele tijdreeksmodel wordt vervolgens getraind op het gefuseerde chlorofyl-a-signaal en gebruikt voor langetermijnprojectie, waarbij de gegenereerde stralings- en temperatuurscenario's als regressoren worden genomen. Aangezien de belangrijkste interesse de dynamiek van de lentebloei van fytoplankton is, extraheren we als laatste stap

de zogenaamde kardinale data van de lentebloei (begin, piek, einde) uit de langetermijn chlorofyl-a-projecties met behulp van een niet-parametrische vormgelimiteerde methode (log-concave regressie). Het uiteindelijke resultaat is daarom de schatting van de door klimaatverandering veroorzaakte onzekerheid in de dynamiek van de voorjaarsbloei van fytoplankton langs de kust.

1

INTRODUCTION

1

Predicting coastal ecosystem state, and specifically phytoplankton biomass that is approximated by chlorophyll-a concentration, is essential in keeping coastal ecosystems healthy and safeguarding their benefits to human societies. On the one hand, phytoplankton and their seasonally occurring blooms are vital to marine ecosystems as they are a major source of energy input for higher trophic levels, providing approximately half of the global primary productivity [70]. On the other hand, phytoplankton blooms may also be harmful and cause mortality of other marine organisms (eutrophication). In addition to the ecosystem impacts, the socio-economic consequences of phytoplankton blooms are also important. They may pose human health issues and affect coastal activities, including fisheries, aquacultures, tourism, and ports [50]. Consequently, accurate and reliable chlorophyll-a concentration prediction is required for both ecosystem and economic benefits.

Changing climatic conditions only add to the complexity of predicting phytoplankton biomass, especially in the long run. It is known that long term climate impacts on phytoplankton are manifested as shifts in seasonal dynamics, species composition, and population size structure [217]. However, as the direct and indirect impacts of changing climatic conditions on phytoplankton are not fully understood, their deterministic prediction with process based models remains difficult. For this reason, the scientific community needs ways to quantify these climate change induced uncertainties to better understand and anticipate the variability in ecosystem response. The present thesis addresses this scientific need by offering a range of statistical tools (together with applications) for the quantification of climate change induced uncertainties in phytoplankton biomass.

1.1. UNCERTAINTY AS AN INHERENT COMPONENT OF FUTURE PREDICTIONS

Identifying, quantifying and reducing uncertainty is a crucial issue in any modelling application. The ramifications of uncertainty yield significant impacts on ecosystem response estimates, and therefore uncertainty is relevant to the concept of environmental impact studies. Consequently, the acknowledgement, quantification, and the subsequent management of uncertainty will influence the optimal decision pathways. Rational decision making therefore requires the quantification of the uncertainty attached to the predicted variable of interest. With this in mind, we must also accept that uncertainty in prediction of future events persists and it cannot be eliminated, only quantified and possibly reduced (with better model formulations and increased spatio-temporal resolutions). This thesis aims to contribute to the scientific community by proposing statistical techniques to quantify climate change induced uncertainties for the future coastal ecosystem state.

According to Krzysztofowicz [120], "uncertainty about future events is the reason for forecasting", indicating that a forecast has to provide an estimate of the uncertain future event and that forecasting will always come with uncertainties. This statement is even more important if we consider long-term predictions simulating climate impacts. Ideally, uncertainty quantification techniques characterize the predictive distribution of our variable of interest. Todini [199] also suggested that the ultimate goal in uncertainty

quantification is the description of the uncertainty of the predicted value (predictive uncertainty), rather than the uncertainty of the prediction model (model uncertainty). This brings us to the question: Where do the uncertainties originate from? What uncertainties can and should be quantified?

In order to answer these questions first one has to study the uncertainty sources, even though from the decision makers' point of view uncertainty can simply be defined as the lack of knowledge, irrespective to the cause of this deficiency [166]. From the mathematical point of view, however, we must understand those causes of deficiencies. Diving into the process of mathematical modelling, we will soon recognize that plenty of assumptions and compromises need to be made so that we can arrive at a reasonable (and hopefully optimal) representation of real physical processes. Walking through the steps we need to make to arrive at predictions we recognize that the sources of uncertainty may be numerous. The uncertainty could originate from inadequate information on the physics, incorrect assumptions, or simply from the variability of natural processes, to name a few. When looking at the literature related to uncertainty estimation in the domain of water management and environmental modelling, various classifications can be found [167, 166, 134, 205]. Here a more generic classification of uncertainties is presented following a classification proposed by Loucks and van Beek [134]. This classification discerns three types of uncertainties: natural variability, knowledge uncertainty and decision uncertainty, as depicted in Figure 1.1. A short description of these uncertainty types is given below in order to aid the understanding of the scope, focus, and limitations of this research.

Natural variability

Natural variability refers to the variability in physical processes, such as meteorological and environmental processes. These stochastic processes have temporal and/or spatial variability. Natural variability is usually quantified through performing statistical assessment of historical data but the question remains whether the statistics of the historical data will accurately represent the future, especially including the impact of long term processes such as climate change. Because of this, natural variability is one of the hardest sources of uncertainty to be dealt with. This source of uncertainty cannot be reduced by developing more sophisticated models, performing more calibration or increasing the temporal and/or spatial model resolution. It will persist regardless of the improvements in our prediction model. In this thesis we address the temporal (natural) variability in stochastic processes at hand (atmospheric processes and marine water quality) by considering temporal variability (seasonal shifts) in climate scenarios and in the chlorophyll-a concentration signal. These identified temporal features have been used to construct new synthetic climate trajectories in the stochastic climate generator (see Chapter 4), which can reproduce the observed variability in this way. Moreover the variation in chlorophyll-a concentration seasonality, as proxy for phytoplankton phenology, is addressed in Chapter 5.

Knowledge uncertainty

Another type of uncertainty is knowledge uncertainty. The name originates from the fact that our knowledge of the physical processes, and how mathematical tools approximate

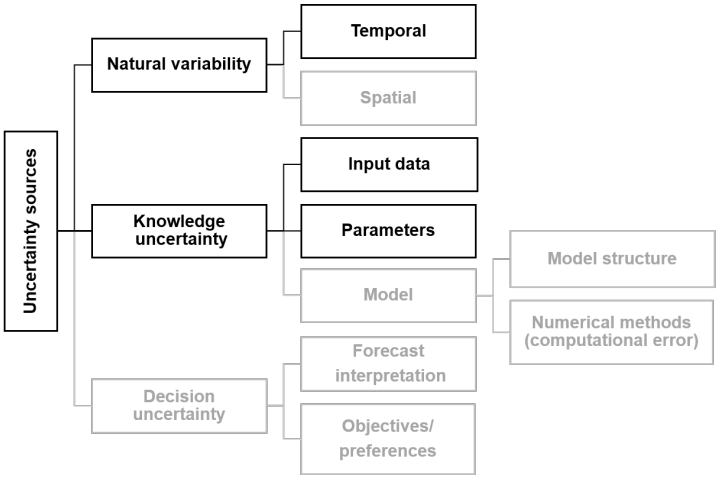


Figure 1.1: Classification of uncertainty sources (adapted from [134]). Uncertainty sources addressed in this thesis are highlighted.

these are uncertain. The author believes that the quote from Einstein at the beginning of this thesis refers to knowledge uncertainty. Knowledge uncertainty is the most researched source of uncertainty and most mathematicians or numerical modellers focus on this aspect. Probably, due to the fact that we have the biggest chance to properly quantify and reduce this type of uncertainty. One part of the knowledge uncertainty is caused by our imperfect knowledge of the input data, boundary conditions and/or model parameters (parameter uncertainty). The other part of the knowledge uncertainty results from our imperfect knowledge on system functions and processes, which in turn causes an imperfect representation of the real world. Since complex systems have to be simplified to a certain degree, this simplification introduces model structural errors. In addition, the choice of numerical methods and numerical schemes can determine the computational errors (e.g. truncation error) of the model. When addressing model related uncertainties, it is likely that increasing the model complexity would reduce the model uncertainty but it could also raise the number of possible error sources and parameter uncertainties. This trade-off should always be considered when seeking for an optimal model formulation.

This thesis primarily deals with knowledge uncertainties where statistical quantification is the most helpful. More specifically, the focus is on the input data and model parameter uncertainties. Climate impact studies, such as the ones in this thesis (Chapter 4 and 5), rely on climate inputs that are driving the ecological processes. Climate input uncertainties are identified in Chapter 3 and addressed through the stochastic climate generator in Chapter 4. Moreover parameter uncertainties in the proposed statistical models are considered by applying Bayesian models, which result in distributions of the inferred model parameters. Uncertainties in the model process parameters of the numerical model Deltf3D-WAQ have also been considered by perturbing their values and producing ensemble simulations in Chapter 2. In Chapter 2 other input data uncer-

tainties are also considered, namely the suspended particular matter field and the river discharges and nutrient loads. By quantifying the above mentioned uncertainty sources we are able to derive predictive uncertainties in the modelled quantities, chlorophyll-a concentration and phytoplankton spring bloom cardinal dates.

The atmospheric and environmental measurements (in-situ and satellite observed) used in this thesis are themselves subject to error. This error can be simply described as the difference between the theoretical true value and the measured value. The measurement errors can be classified as random or systematic, where a random error is commonly known as noise and systematic error is called bias. The measurement uncertainty describes the magnitude of the error and characterizes the distribution of error. Therefore, an estimate of measurement uncertainty is necessary to appropriately use the information conveyed by the measurements. In this thesis measurement errors and measurement uncertainties are considered in two ways. Firstly, systematic errors of the climate scenarios are corrected by applying quantile mapping bias corrections. Secondly, uncertainties in the satellite measured chlorophyll-a concentrations were addressed by fusing them with in-situ measurements and thus creating a reconstructed chlorophyll-a signal.

Decision uncertainty

The last group of uncertainty sources in this classification is the decision uncertainty. This group refers to the uncertainty sources that are not related to the variability of nature, the modeller (imperfect knowledge) or the model (imperfect model), still they influence the decisions made based on them. Instead, decision uncertainty is an acknowledgement that the interpretation of forecasts is subjective, moreover that future decisions, objectives and preferences of individuals and organizations are unpredictable, since these are dynamically evolving. Consequently, our predictions (and in general the supplied scientific evidence) will most likely not result in the same decisions at all times. In Chapter 2 we use probabilistic performance metrics that require a certain exceedance probability as threshold, which must be chosen by decision makers. In that context the scientific background, risk taker or avoiding behaviour will influence how the provided probabilistic prediction turns into decisions. Although it is a very interesting branch of science, dealing with decision uncertainty is out of the scope of this thesis.

1.2. OBJECTIVES AND OUTLINE

After placing this work in the larger context of dealing with uncertainty sources, the research objectives and the outline are presented in brief.

The aim of the thesis is to understand the available multi-sourced and multi-dimensional climate and environmental datasets, connect the atmospheric and the ecological realms in the context of coastal water quality, and offer ways for quantifying the uncertainties in the ecological response that are driven by the climatic variation.

The emphasis is placed on the quantification of uncertainty using statistical techniques, rather than on the accurate portrayal of ecological implications of climate change. This main objective can be broken down into four sub-objectives according to the main chapters of the thesis.

1. Objective 1: Showcase that predictive uncertainty estimates on water quality indicators are beneficial for expressing confidence in the simulated environmental changes and that probabilistic predictions are more informative than deterministic predictions (Chapter 2)

2. Objective 2: Provide a statistical underpinning of climate variables selection for coastal ecological impact studies (Chapter 3)

3. Objective 3: Propose a way for enriching existing climate scenarios whose outputs can be used for probabilistic climate impact studies (Chapter 4)

4. Objective 4: Quantify climate change induced uncertainties in coastal phytoplankton spring bloom dynamics (Chapter 5)

The conceptual flow of the thesis and the connections between the chapters are depicted in Figure 1.2. First of all, the datasets (indicated as grey boxes) that serve as the basis for this research should be mentioned. From the atmospheric side, the applied data sources cover the same atmospheric variables (air temperature, solar radiation, eastward wind, northward wind, air pressure, relative humidity, and total cloud cover) but originate from various sources and cover different time intervals:

- short term (a single year) outputs of a numerical weather prediction model, that is used to drive the physical model simulating water quality processes;
- long-term climate change data covering the entire 21^{st} century produced by a regional climate modelling experiment;
- measured solar radiation and air temperature data from ground stations (in-situ) available for the historical period (from 1970s until today).

From the environmental side, chlorophyll-a concentrations are obtained from:

- a physical model simulating water quality processes in short term (a single year),
- satellite observations for a shorter historical period (from end of 1990s until today),
- in-situ observations for the historical period (from 1970s until today).

Again, these three types of chlorophyll-a data cover different time intervals and have varying temporal resolutions (from 6-hourly to monthly). This multitude of atmospheric and environmental data is used for different purposes along the thesis to train and validate the applied techniques.

In the first step motivation is provided for the rest of the thesis by showcasing the potential of adding predictive uncertainties to ecological model outputs and by concluding that atmospheric uncertainties should also be considered. Then coastal environmental and atmospheric data reduction takes place in support of the subsequent ecological impact studies. These findings guide the formulation of a stochastic generator that is aimed at enriching the existing eight Euro-CORDEX climate scenarios. The stochastic

generator is first introduced for solar radiation but later extended to air temperature. Finally, the generated synthetic climate projections of both solar radiation and air temperature are used to simulate long-term chlorophyll-a concentrations with a statistical model (Bayesian structural time series model), and cardinal dates (beginning, peak, end) of the phytoplankton spring bloom dynamics are extracted. The final output is therefore the quantified uncertainty around the spring bloom dynamics (beginning, peak, end) and spring bloom peak magnitude by the end of the century.

The dissertation outline and the description of each chapter can be found below:

Chapter 2 - The value of probabilistic predictions for coastal ecosystems

This chapter describes the benefits of probabilistic prediction of chlorophyll-a concentrations over deterministic prediction. It therefore gives a practical motivation why we should consider uncertainty quantification in environmental systems. In this chapter an input ensemble is generated from perturbed water quality model process parameters and external forcings and the simulation is performed with a numerical model from Deltares. The advantage of ensemble prediction over a deterministic forecast is assessed using several forecast verification metrics that can describe the forecast accuracy, reliability and discrimination.

Chapter 3 - Statistical underpinning of atmospheric variables selection

This chapter provides climate variable selection for predicting coastal chlorophyll-a concentrations. The outcome of this selection is the choice of two variables, solar radiation and air temperature, to study the long term changes and uncertainties in coastal ecological response. The chapter introduces and applies a selection of dimension reduction models for discrete data: Principal Component Analysis (PCA), Principal Component Regression (PCR), Partial Least Squares (PLS) Regression, multi-way PLS, Dynamic Factor Analysis (DFA), and for functional data: Functional PCA. These methods are applied to seven atmospheric variables and chlorophyll-a data. The dimension reduction methods are then compared and findings are presented on underlying latent factors, similarities between Euro-CORDEX climate change scenarios, and the most influential climate variables driving changes in chlorophyll-a concentrations.

Chapter 4 - Bayesian stochastic climate generator

This chapter focuses on the most influential climate variable (solar radiation), selected in the previous chapter, and explains how to generate additional climate scenarios (enriching the existing Euro-CORDEX scenarios) for the 21^{st} century to guarantee a better characterization of climate change induced uncertainties. A description is given on the stochastic climate generator methodology, which uses the EURO-CORDEX climate change projections as input and produces new synthetic scenarios as output. A probabilistic simulation (using a numerical water quality model) is executed to showcase the advantages of taking the generated synthetic climate data as model input and deriving the predictive uncertainties of the chlorophyll-a signal.

1

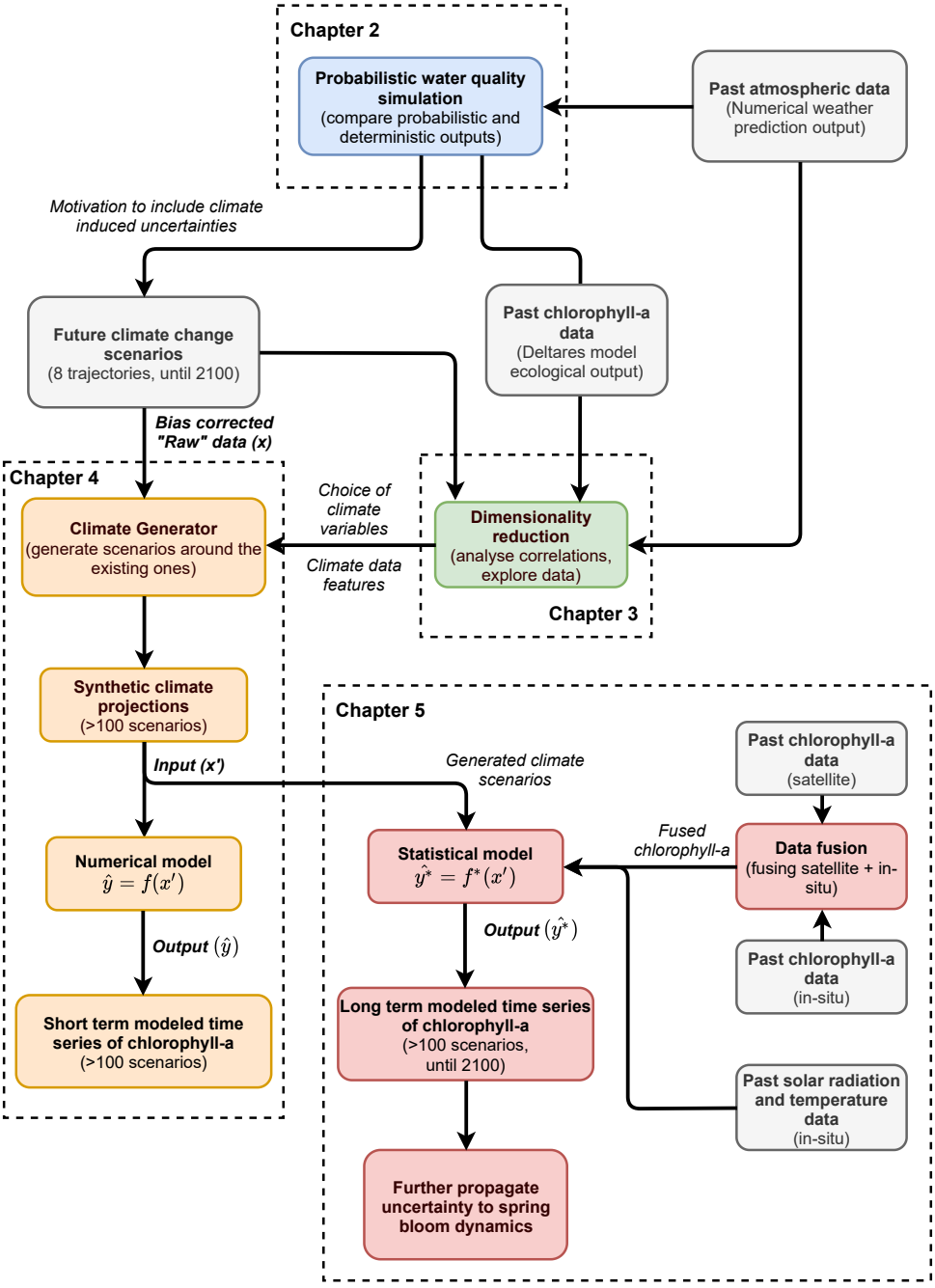


Figure 1.2: Thesis structure and connections between chapters. Grey boxes represent the datasets. Colored boxes represent methodological steps within the thesis chapters.

Chapter 5 - Climate Change Induced Trends and Uncertainties in Phytoplankton Spring Bloom Dynamics

The last methodological chapter builds upon the generated radiation and temperature scenarios, produced in Chapter 4. This chapter projects trends and most importantly uncertainty estimates of coastal phytoplankton spring bloom dynamics for the 21^{st} century. The three main methodological steps to achieve this goal include (1) developing a data fusion model to interlace coastal in-situ measurements and satellite chlorophyll-a observations into a single long-term (>40 years) signal; (2) applying a Bayesian structural time series forecasting model to produce long-term prediction of chlorophyll-a concentrations; and (3) developing a feature extraction method to derive the cardinal dates (beginning, peak, end) of the spring bloom to track the historical and the projected evolution of its dynamics.

Chapter 6 - Conclusions and discussion

This final chapter summarizes the research findings of all chapters, and overarching conclusions are drawn that lead to final statements on research objectives. Limitations are highlighted and recommendations for improvements are formulated incorporating the experience gained through the completion of the research.

2

THE VALUE OF PROBABILISTIC PREDICTIONS FOR COASTAL ECOSYSTEMS

Prediction systems, such as the coastal ecosystem models, often incorporate complex non-linear ecological processes. There is an increasing interest in the use of probabilistic predictions instead of deterministic predictions in cases where the inherent uncertainties in the prediction system are important. The primary goal of this chapter is to showcase an ensemble prediction system for the simulation of chlorophyll-a concentration in coastal waters, using the Generic Ecological Model (GEM). The input ensemble is generated from perturbed model process parameters and external forcings through Latin Hypercube Sampling with Dependence (LHSD). The performance of the ensemble prediction is assessed using several verification metrics that can describe the prediction accuracy, reliability and discrimination. The verification is performed against in-situ measurements and remote sensing data. The ensemble prediction moderately out-performs the deterministic prediction in the coastal waters. Using probabilistic simulations and quantifying predictive uncertainties is therefore valuable for an enhanced description of coastal ecosystem state.

Parts of this chapter have been published in Journal of Hydroinformatics **20**, 4 (2018) [147]

2.1. INTRODUCTION

Water quality is a crucial factor for both coastal ecosystems and human societies. Phytoplankton blooms in the North Sea may cause mortality of mussels and other benthic organisms. Furthermore, fisheries and aquacultures are influenced by the algal primary production since it is the base of the food web. Consequently, accurate real-time phytoplankton concentration prediction is required for ecosystem and economic benefits. The timely information about water quality allows for early warning and adequate response such as mitigation measures and targeted monitoring.

Existing hybrid ecosystem models are powerful tools for modelling water quality; however, their reliability highly depends on the uncertainty stemming from different sources. Uncertainty originating from external forcings is further propagated and complicated by the non-linear ecological processes, with numerous intercorrelated parameters incorporated in the water quality model. Considering the high level of uncertainty in the coastal water quality forecasting process, it is assumed that a single-valued deterministic forecast may not be sufficiently reliable for decision making. Thus, a strong need arises for an ensemble forecasting system that could potentially account for the uncertainty associated with the driving forces, the model simplifications or the parameterization. Ensemble forecasting of water quality might require considerably more effort than the deterministic approach, though in return gives added value to the forecast. Deterministic forecasts only provide a point estimate, whereas probabilistic forecasts determine the probability density function of the predictand. By performing a statistical analysis of the acquired probability density function various measures can be derived such as the mean and standard deviation of the model output distribution, or the likelihood of a specific output value. Through these measures the uncertainty in the model output can be quantified, which serves decision support functions.

The application of ensemble prediction systems is well-known in various fields, especially in numerical weather prediction, yet few examples [69, 201, 103] show their applicability in water quality forecasting. A recent study by [4] investigated the possibility to develop a water quality forecasting system for riverine ecosystems with ensemble streamflow prediction method. In that study a historical rainfall and temperature ensemble was applied as a forcing condition in order to produce a probabilistic water quality prediction. Based on those findings it can be concluded that implementing an ensemble system for water quality forecasts by addressing the uncertainty in the input variables is advantageous. Another research in ensemble forecast accuracy in the southern North Sea was conducted by [99]. The ensemble forecasting system was set up using perturbed meteorological forcings and the Delft3D-FLOW model. Despite expectations, the results showed that the impact of the meteorological ensemble input is not sufficient to estimate the total uncertainty in the forecasted parameters (salinity and temperature). In order to complement those findings this paper suggests the application of an extended input ensemble with the aim to provide further uncertainty estimates.

This paper aims to set up the framework for an operational water quality ensemble forecasting system, even though the proposed method is only applied in a hindcasting case study and further steps are required to reach the operational stage. The model output of interest is the chlorophyll-a concentration which is a commonly used indicator for water quality. The simulation is carried out using the Delft3D-WAQ software package,

more specifically using its sub-component the Generic Ecological Model (GEM) with the advance algal speciation module-BLOOM.

The chapter starts with a description of GEM, its application to the North Sea, and the different data sources used for the forecast verification. The research methodology is presented in the following sections including the significant parameter selection, the Latin Hypercube Sampling with Dependence (LHSD) steps, and the applied forecast verification method. The performance analysis of the ensemble forecasting system applied to the southern North Sea is given together with the validation and spatial results. The conclusions and recommendations for further research are presented in the final section of the paper.

2.2. MODELLING INSTRUMENT

The modelling instrument is a comprehensive hybrid ecological model combining a three dimensional hydrodynamic model (Delft3D-FLOW) and the GEM biogeochemical model. The GEM model includes an array of modules reproducing water quality processes that are then combined with the transport model. Most importantly, the model computes primary production and chlorophyll-a concentration while integrating dynamic process modules for dissolved oxygen and nutrient concentration calculation. Furthermore, the GEM model includes a phytoplankton module (BLOOM) that simulates the growth, respiration and mortality of phytoplankton. Using this module the species competition and their adaptation to limiting nutrients or light can be simulated. The model offers flexibility in the processes selection and provides general applicability in diverse case studies [30]. In recent years GEM has already been applied to the southern North Sea. A three dimensional GEM application was presented by [133], while a two dimensional application was done by Salacinska et al. (2010).

Model calibration and validation of the 3 dimensional GEM model for the southern North Sea was done in previous studies by [133] using in-situ dataset for the year 1989, in [113] for the year 2007, and in [14] for years 2009-2010 also using the in-situ dataset. In general, all the reports mention acceptable chlorophyll-a prediction accuracy, however, the results are not homogeneous across the stations and highly dependent on proper description of the Suspended Particulate Matter field. Furthermore, it should be noted that in the Dutch coastal zone, the observed gradients of algal biomass are very steep and there is considerable natural variability in the chlorophyll-a concentration. The peak spring bloom could be shifted ±1 month and its magnitude could vary ±80% [113].

Transport of substances

The core of the water quality model is a mass balance for the simulated state variables necessary to describe the problem at hand. This mass balance is described by the advection-dispersion equation (three dimensional case), see Eq. (2.1). The equation is solved by the hydrodynamic model (Delft3D-FLOW), where a broad selection of numerical schemes is available to compute the transport part of the equation. The selection of numerical solutions helps to cope with the 1D/2D/3D model discretization and even complex, irregular geometries for both steady and unsteady cases. This advection-dispersion equation is responsible for the change in concentration of the substances in time due to the

advective transport (transport of substances via the fluid movement) and the diffusive and/or dispersive transport (spreading of mass from highly concentrated areas to less concentrated areas), along with the sources and sinks (direct inputs (discharges and waste loads), and/or mortality (for bacteria), decay, sedimentation (for solid particles)):

$$\frac{\delta C}{\delta t} = -u\frac{\delta C}{\delta x} - v\frac{\delta C}{\delta y} - w\frac{\delta C}{\delta z} + \frac{\delta}{\delta x}\left(D_x\frac{\delta C}{\delta x}\right) + \frac{\delta}{\delta y}\left(D_y\frac{\delta C}{\delta y}\right) + \frac{\delta}{\delta z}\left(D_z\frac{\delta C}{\delta z}\right) + S + P \quad (2.1)$$

where C is the concentration of the state variables $[gm^{-3}]$; u, v, w are velocity vector components $[ms^{-1}]$; D_x, D_y, D_z are dispersion tensor components $[m^2s^{-1}]$; x, y, z are coordinates $[m]$; S is the source and sink term of mass due to loads and boundaries; P is the source and sink term of mass due to processes; and t is time $[s]$. In the equation $\frac{\delta C}{\delta t}$ represents the change in concentration of the modelled substances in time, $-u\frac{\delta C}{\delta x} - v\frac{\delta C}{\delta y} - w\frac{\delta C}{\delta z}$ is the advective transport, and $\frac{\delta}{\delta x}\left(D_x\frac{\delta C}{\delta x}\right) + \frac{\delta}{\delta y}\left(D_y\frac{\delta C}{\delta y}\right) + \frac{\delta}{\delta z}\left(D_z\frac{\delta C}{\delta z}\right)$ is the diffusive and/or dispersive transport.

Ecological processes

The concentrations of 30 state variables including algae concentrations, nutrients, and salinity are calculated through various ecological processes. The most important processes are related to nutrient cycles, oxygen dynamics, energy availability and phytoplankton processes. The GEM ecological processes are considered to be moderately complex by [133] based on the fact that GEM excludes microbial loop, explicit grazing and higher trophic levels; also the benthic processes are relatively simple. On the other hand, full nutrient cycles and complex phytoplankton kinetics were implemented. Figure 2.1 contains the schematic overview of all possible variables and processes in the GEM model. The reader may refer to Blauw et al. (2009) and [133] for further description of the ecological processes.

GEM makes use of more than 400 parameters for the processes calculations. These parameters are related to algae's characteristics such as nutrient-to-carbon ratios and growth rates. Some describe light availability through extinction coefficients and settling velocities and others are connected to the nutrient cycles. GEM is part of an integrated modelling system which contains separate modules for hydrodynamics, waves and sediment transport calculation. The water quality simulation therefore requires external forcings such as meteorological conditions, hydrodynamics, Suspended Particulate Matter (SPM) concentration field as well as nutrient loadings from atmospheric deposition and riverine loads.

Spatial discretization

The domain decomposition of the Southern North Sea is a three dimensional curvilinear grid with finer resolution along the coast as shown in Figure 2.2. The grid contains 12 sigma layers with unequal thicknesses (see Table 2.1). The distribution of the layers was designed to provide higher resolution on the top and the bottom of the water column to enable more accurate suspended matter calculation for light attenuation on the top

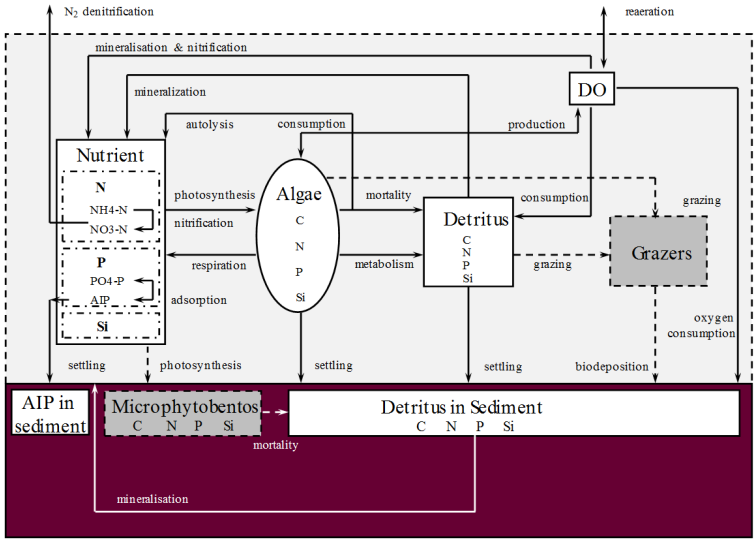


Figure 2.1: State variables and processes in GEM. State variables in grey and processes indicated by dashed lines have not been included in the North Sea modelling applications. Source: [133].

Table 2.1: Relative thickness of the sigma layers in the three dimensional grid

Layer no.	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*Total*
Relative thickness (%)	4	5,6	7,8	10,8	10,9	10,9	10,9	10,9	10,9	7,8	5,6	4	100

and more detailed resuspension calculation near the bed. The resolution of the coarse segments varies from 6-by-5 km to 20-by-30 km and the resolution of the fine grid ranges from 1-by-2 km to 2.5-by-3 km.

2.3. DATA SOURCES

In order to validate the water quality model in this paper three types of data sources are used: a set of historical surface in-situ measurements, and remote sensing images together with their gap-filled reconstruction (see Figure 2.3). Fifteen stations were selected with available chlorophyll-a in-situ measurements for the given years 2007 and 2009 in the focus area provided by Rijkswaterstaat (Dutch Ministry of Infrastructure and Environment). It should be noted that the samples are taken close to the water surface (usually the upper 3-5 metres of the water column) and therefore they can only be used to validate the first layer of the model.

The second type of dataset is remote sensing which gives a more comprehensive view of the model's success in capturing seasonal variability since the spatial and temporal resolution of the in-situ dataset is low. The remotely sensed sea colour images applied in this study are retrieved from the MEdium Resolution Imaging Spectrometer instru-

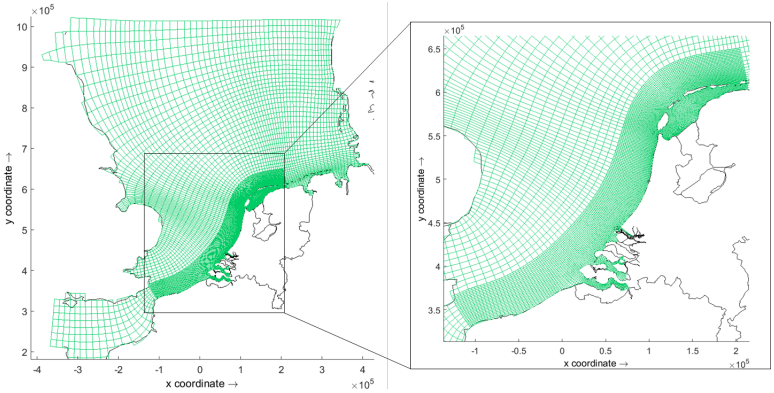


Figure 2.2: Spatial discretization of the GEM model in the North Sea.

ment (MERIS), which was on the board of the European Space Agency ENVironmental SATellite (ENVISAT) spacecraft. Since the ENVISAT ended its mission in 2012, the application of MERIS data in this paper is considered as a preparation for the new Copernicus Sentinel dataset usage which has recently stepped into the fully operational stage.

The recorded optical reflectance of the water surface is transformed into chlorophyll-a or SPM concentration using the HYDROPT algorithm given in [206]. The algorithm not only computes the concentration of the substances but also provides a measure of error in the estimates [182]. The retrieved pixel data is then interpolated onto the Southern North Sea Domain Decomposition (ZUNO-DD) grid taking into account the provided error estimate, and resulting in the so-called gridded data [28]. It should be noted that the raw pixel data was first interpolated using nearest-to-the-centre interpolation onto the ZUNO-COARSE gridding and then it was subsequently transposed onto the ZUNO-DD domain gridding.

Most forecast verification metrics are based on the differences between the prediction and observation; however, these differences can only be calculated if both values are available for the same time step at the same location. Considering the incomplete temporal and spatial coverage of the samples from the MERIS measurements due to clouds, the number of matchups is fairly limited. As an attempt to tackle this issue a third data set is used for the model validation, which is the gap-filled version of the gridded MERIS images. The gap filling is done using a data interpolating algorithm called DINEOF. The DINEOF algorithm determines the major spatial and temporal patterns of variation in the MERIS dataset and produces the gap-filled reconstruction at all segments and all time steps [28]. The gap-filled data applied in this study was constructed using the MERIS data only from the years of interest; 2007 and 2009.

2.4. ENSEMBLE FORECASTING METHODOLOGY

The main objective of the paper is to set up a viable and generally applicable water quality ensemble forecasting system for coastal ecosystems. The proposed methodology (see Figure 2.4) includes the following important steps:

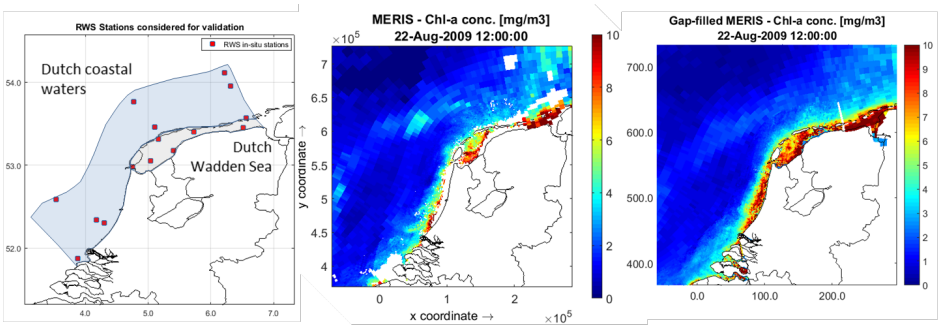


Figure 2.3: Validation data sources: in-situ measurements (left), MERIS (middle) and gap-filled MERIS (right) remote sensing data.

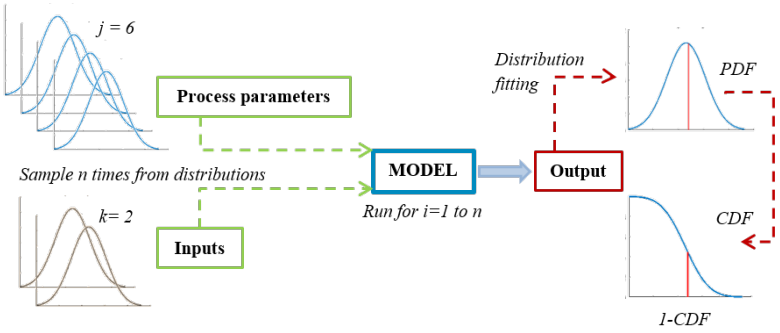


Figure 2.4: Schematization of the ensemble forecasting methodology.

- Selection of the significant model process parameters and external forcings responsible for the uncertainty in the chlorophyll-a concentration prediction.
- Statistically efficient sampling of the input ensemble (stratified sample with parameter dependency) using the selected significant parameters and forcings.
- Post-processing of the ensemble model run by fitting probability density functions to the model outputs at every time step and all locations.
- Forecast verification to assess and compare the performance of the deterministic and ensemble predictions.

Note that during the post-processing step continuous probability distributions were fitted and ranked by the Bayesian Information Criterion (BIC) in order to obtain the best fit. This fitting procedure does not allow multimodal fitting meaning that the PDF cannot have multiple peaks; this simplification might result in inaccuracies.

Table 2.2: Selected significant GEM model process parameters for ensemble generation

No.	*Parameter*	*Description*	*Constraint*
1	$e_{s,IM}$	Specific Extinction coeff. of suspended inorganic matter	*Energy / growth*
2	$kgp^{0}_{Di,E}$	Maximum Growth Rate of Diatoms type E at 0°C	*Growth*
3	$ktpg_{Di,E}$	Temperature coefficient for growth of Diatoms type E	*Growth*
4	$R_{den,sed}$	Denitrification rate in the sediment	*Nutrient*
5	$R_{den,wat}$	Denitrification rate in the water column	*Nutrient*
6	b_{S1}	Burial rate for layer S1	*Nutrient*

2.5. SELECTION OF IMPORTANT PARAMETERS

The previously introduced ecological processes of GEM make use of more than 400 model process parameters in total. Incorporating all these parameters into the ensemble generation is not feasible and would not yield in a better uncertainty estimate, only in overlapping simulation results since most of them do not affect the chlorophyll-a concentration.

A previous study investigated the sensitivity of the chlorophyll-a concentration to process parameters in the North Sea application of GEM [174] and identified 20 parameters that obtained the highest rank. These parameters consist of growth rates, extinction coefficients and nutrient-to-carbon ratios (e.g. P:C, N:C) of the different algal species. Furthermore, according to [125] primary production mainly depends on the specific rates of growth, mortality, and maintenance respiration as well as the temperature coefficient because these rates are calculated as a function of the temperature. In addition, [54] conducted a sensitivity analysis of the nutrient concentrations influenced by the different processes such as denitrification, nitrification, mineralization and burial. Those findings suggest that in coastal regions nitrate mainly sensitive to the denitrification rate in the top sediment layer (S1), while phosphate and all other nutrients are mainly sensitive to the burial rate. This extensive list of significant parameters was narrowed down to 6 parameters (see Table 2.2) considering the dominating algal type (Diatoms type E) and the coastal environment in the Southern North Sea. Further description of the significant parameters can be found below.

Model process parameters

The chlorophyll-a concentration C_{chlfa} is function of the algal biomass concentration $C_{\text{alg,i}}$ of each algal type i and the stoichiometry of chlorophyll-a in each algae type $S_{\text{chlfa,i}}$ (Eq. (2.2)), thus any parameter that alters the algal biomass will also change the chlorophyll-a concentration.

$$C_{\text{chlfa}} = \sum_{i=1}^{n} \left(S_{\text{chlfa,i}} \times C_{\text{alg,i}} \right) \tag{2.2}$$

The algal biomass concentration is influenced by energy-, growth-, nutrient- and mortality constraints. This paper focuses on the peak algae bloom prediction which is not limited by the mortality constraint and for this reason it is not considered. Therefore, the selected model process parameters can be divided into three main groups depending on whether they affect the energy-, growth- or nutrient constraints as shown in Table 2.2.

The specific extinction coefficient $e_{\mathrm{s,IM}}$ and concentration of suspended inorganic matter C_{IM} (external forcing) affect the light regime in the water which determines the energy availability and growth of the phytoplankton species. In shallow coastal ecosystems like the Wadden Sea the concentration of suspended inorganic matter is relatively high and dynamically varying. The extinction of light is due to the particulate and dissolved light absorbing substances in the water such as the algae biomass, detritus and the suspended inorganic matter. Therefore, the total extinction coefficient (et) is calculated as the sum of the partial extinction coefficients of these substances (e.g. partial extinction coefficient of inorganic matter $e_{\mathrm{st,IM}}$) and the background extinction of the water $e_{\mathrm{background}}$ (Eq. (2.3). The partial extinction of these substances is expressed with their specific extinction coefficient and their concentration (Eq. (2.4)).

$$et = e_{\mathrm{st,IM}} + e_{\mathrm{st,other}} + e_{\mathrm{background}} \tag{2.3}$$

$$e_{\mathrm{st,IM}} = e_{\mathrm{s,IM}} \times C_{\mathrm{IM}} \tag{2.4}$$

The visible light intensity I is then described by an exponential attenuation, namely the Lambert-Beer law in Eq. (2.5), which is a function of the total extinction coefficient et and the water depth H:

$$I_{\mathrm{b}} = I_{\mathrm{t}} \times \exp^{-et \times H} \tag{2.5}$$

denoting the light intensity at the bottom of the water column I_{b} and at the top of the water column I_{t}. Finally, the actual visible light intensity is converted to a light efficiency factor $E_{\mathrm{f,i}}$ which is used in the growth constraint of the GEM model, see Eq. (2.6), to determine the maximum concentration of the algae type i ($C_{\mathrm{algmax,i}}$) for each time-step Δt.

$$C_{\mathrm{algmax,i}} = C_{\mathrm{alg,i}} \times \exp^{(kgp_{\mathrm{i}} \times E_{\mathrm{f,i}} - krsp_{\mathrm{i}}) \times \Delta t} \tag{2.6}$$

In the growth constraint equation, kgp_{i} stands for specific growth rate of algae type i and $krsp_{\mathrm{i}}$ is the specific maintenance respiration rate of algae type i. This specific growth is calculated using the maximum growth rate at $0°C$ kgp^0_{i} together with the water temperature T and temperature coefficient for growth $ktpg_{\mathrm{i}}$ in Eq. (2.7). Consequently, these parameters also have a direct influence on the algae growth. It should be noted that in the southern North Sea the spring bloom of phytoplankton is dominated by diatoms [29], and the beginning of the peak algal bloom is mainly influenced by the energy limited type (E-type) algae species. The maximum growth rate and the temperature coefficient for growth parameters are therefore selected for the Diatoms type E ($i = Di, E$).

$$kgp_{\mathrm{i}} = kgp^0_{\mathrm{i}} \times \left(T - ktpg_{\mathrm{i}}\right) \tag{2.7}$$

The last group of parameters considered for the input ensemble affects the availability of nitrate, phosphate and silicate that are essential nutrients for the algae growth. Thus perturbation on these nutrient concentrations provides further uncertainty estimation of the chlorophyll-a concentration. The denitrification rate and burial rate influence the nutrient availability in the water column and in the sediment for the algae species

through the denitrification (Eq. (2.8)) and burial processes. Denitrification removes nitrate NO_3^- from the water and generates elemental nitrogen N_2, which can leave the water phase and escape into the air:

$$4NO_3^- + 4H_3O^+ \Rightarrow 2N_2 + 6H_2O \tag{2.8}$$

In the GEM model the denitrification process depends on the denitrification rate R_{den}, the nitrate concentration C_{NO_3}, and denitrification process temperature $f_{\text{T,den}}$. It is important to mention that the denitrification process does not take place under a critical temperature T_{c}. Nutrient availability for the living organisms is also influenced by the burial process. The importance of the burial process in the model is to remove the dead particulate organic matter POX_{S} from the active top sediment layer $S1$ to the deeper sediment layers. However, the burial rate represents more than just the actual burial process; it also stands for uncertainties in the nutrient loadings and unknowns in the nutrient mass balance. In the model the burial is a function of the burial rate b_{S1}, the total concentration of the dead particulate organic matter and the water depth H, see Eq. (2.9).

$$Burial_{\text{POX}_\text{S}} = b_{\text{S1}} \times \frac{POX_{\text{S}}}{H} \tag{2.9}$$

Model forcings

In addition to the model process parameters certain model forcings are selected to be perturbed based on previous research findings focusing on the Southern North Sea coastal ecosystem. The importance of the concentration of suspended inorganic matter forcing, often referenced as Suspended Particulate Matter (SPM), was previously demonstrated in Eq. (2.4). In the continental coastal waters 25 to 75 % of the light extinction is caused by suspended particulate matter [131]. Furthermore, [159] investigated the effect of using different types of SPM sources on the chlorophyll-a concentration in the southern North Sea and the results suggested considerable impact. [65] also indicated that the SPM data sources in the North Sea, such as the field measurements and the Delft3D-WAQ-SPM model results, are rather uncertain. Based on these findings it was decided to include the SPM concentration field in the ensemble generation. By reason of simplicity the spatial correlation in the SPM concentration field is not considered, and only a simple error range was applied to the input segment function that contains the suspended particulate matter concentration for each segment at all time-steps. Spatial correlation, however, was previously found to be important in [110] since the seasonal variation in fine sediment dynamics forms several ellipsoidal shaped sedimentation traps along the Dutch coast.

Further model forcing affecting the chlorophyll-a concentration, especially in coastal ecosystems, is the riverine nutrient load. Nutrient loads from rivers provide a significant portion of the total nitrogen load (12 – 17%) and total phosphorus load (8 – 11%) in the North Sea [54]. Since the information about the river discharges and nutrient loads is often estimated, and the daily discharge values are interpolated from the less frequently available flow data, the riverine nutrient input is also included in the ensemble generation. For practical reasons, instead of directly perturbing the nutrient concentrations,

Table 2.3: Chlorophyll-a concentration's sensitivity to the selected parameters and forcings (Correlation patterns)

	Correlation		
Parameter	***During spring bloom***	***After spring bloom***	***Time lag in peak***
$e_{s,IM}$	(--)	(-)	Yes (→)
SPM	(--)	(-)	Yes (→)
$kgp^{0}_{Di,E}$	(++)	(+)	Yes (←)
$ktpg_{Di,E}$	(--)	0	No
$R_{den,sed}$	(-)	(-)	No
$R_{den,wat}$	(-)	(-)	No
b_{S1}	(--)	(--)	No
	Energy limited period	***Nutrient limited period***	
Riverine nutrient load*	0	(++)	No
Q_river (river discharge)*	(--)	(++)	No

Correlation: (-) Negative, (--) Strong negative; (+) Positive, (++) Strong positive; 0 - No correlation
Time lag in peak: (→) Peak later, (←) Peak earlier
** Large spatial variability, depending on distance to river mouth*

the river discharges are altered, which linearly affect the nutrient concentrations. In this paper only those nine river discharges are perturbed that are identified to be influential in the study area based on the nutrient composition matrix derived by [132].

Sensitivity analysis

A simplified sensitivity analysis was carried out in order to comprehend the chlorophyll-a concentration's response to the parameter value modification and to verify if previous findings can be confirmed in the present case. The sensitivity analysis was a one-factor-at-a-time method. This method allows us to gain information about the model's sensitivity to a specific parameter; however, it does not account for the dependencies between variables since the parameters interact in a non-linear way. In the sensitivity test, model runs were executed using the baseline, lower-, and upper bound values of all identified parameters as in Table 2.4. The changes in residuals between the baseline and the scenarios were analysed at one station over one year with particular attention given to the peak concentration and timing. A brief summary of the observed correlation patterns between the selected parameters and the chlorophyll-a concentration can be found in Table 2.3. The table demonstrates the type and strength of the correlation between the investigated parameters and the output parameter, also it indicates whether the time of the peak concentration was affected or not. While the results confirm previous findings, it should be noted that the denitrification rates were found to be less significant compared to other parameters, hence, these might be negligible in deeper coastal waters.

2.6. SAMPLING PROCEDURE

Latin Hypercube Sampling with Dependence (LHSD)

The ensemble simulation requires a sample generated based on the above listed variables with two specific requirements to fulfil: (1) the dependence structure between the variables must be represented and (2) a good coverage of the parameter space should be achieved with only a limited sample size. The water quality model makes use of numerous model parameters which are in many cases correlated. If the sampling technique ignores those correlations, the simulation could result in unrealistic outputs. One way to describe dependency between ecological process variables is to use copulas [108]. In addition to the parameter dependency, choosing a sampling technique with variance reduction gives a greater precision of the output random variables for a given number of iterations, this way providing a statistically efficient sampling process. A well-known variance reduction technique is stratified sampling. Stratification divides the range of the input variable into specific subsets, so-called strata, possibly based on the cumulative distribution function. The samples are then (randomly) drawn from the intervals representing values from each stratum.

Latin Hypercube Sampling with Dependence (LHSD) fulfils these requirements as it accounts for parameter dependence and employs stratification. A demonstrative two dimensional example of LHSD is presented in Figure 2.5. In this figure an original sample linked with the Gaussian copula ($\rho = 0.8$) is compared to its corresponding LHSD sample. We can observe that the two coordinates of the LHSD sample are uniformly spread over the unit interval, in this specific case each univariate sample is located in the middle of its stratum ($\eta^j_{\mathrm{i,n}} = 0.5$), thereby achieving variance reduction. While the original dependence structure is broken, the error between the original copula and the copula of the LHSD samples is small (with increasing sample size the error is decreasing) [157]. LHSD and its preliminaries (stratified sampling and Latin Hypercube Sampling) are introduced below (coordinate-wise).

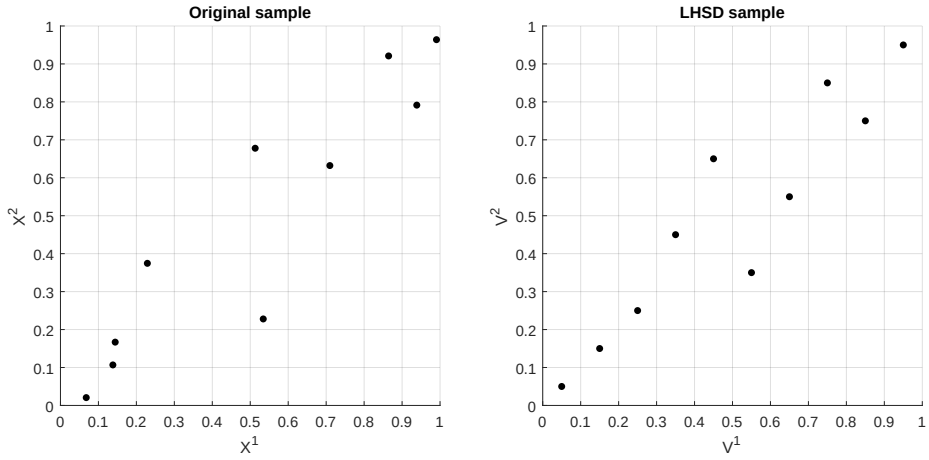
Stratified sampling constrains the fraction of samples drawn from specific subsets (so-called strata) by drawing independent $U(0,1)$ samples (i.e., a uniform random variable on $[0,1]$), $U_1, \ldots, U_{\mathrm{n}}$:

$$V_{\mathrm{i}} = \frac{i-1}{n} + \frac{U_{\mathrm{i}}}{n},\ i = 1, \ldots, n \tag{2.10}$$

Latin Hypercube Sampling (LHS) extends stratified sampling to independent d dimensional random vectors:

$$\overrightarrow{X_1}, \overrightarrow{X}_2, \ldots, \overrightarrow{X_n} \in [0,1]^d$$
$$\{X_i^j : 1 \le i \le n; 1 \le j \le d\},\ d \times n.$$

This is done by generating n independent samples and d independent permutations $(\pi^1, \ldots, \pi^d$ of $\{1, \ldots, n\})$ of each dimension drawn in such a way that the permutations are equiprobable. As a result the Latin hypercube sample is as follows (two dimensional example, $d = 2$):



(a) Original sample

(b) LHSD sample

Figure 2.5: Demonstrative example of Latin Hypercube Sampling with Dependence (LHSD). Left: original sample $(X_1^1, X_1^2), \ldots, (X_{10}^1, X_{10}^2)$ linked with a Gaussian copula with correlation $\rho = 0.8$ and right: corresponding LHSD sample $(V_1^1, V_1^2), \ldots, (V_{10}^1, V_{10}^2)$ with $n = 10$, and $\eta_{i,10}^j = 0.5$. Permutations are $\pi^1 = \{2,5,7,8,9,6,1,4,3,10\}$ and $\pi^2 = \{2,7,6,9,8,4,1,5,3,10\}$.

$$\text{First coordinates: } (X_1^1, \ldots, X_n^1) \to \pi_i^1 (1 \le i \le n) \to V_i^1 = \frac{\pi_i^1 - 1}{n} + \frac{U_i^1}{n}$$

$$\text{Second coordinates: } (X_1^2, \ldots, X_n^2) \to \pi_i^2 (1 \le i \le n) \to V_i^2 = \frac{\pi_i^2 - 1}{n} + \frac{U_i^2}{n}$$

In general form:

$$V_i^j = \frac{\pi_i^j - 1}{n} + \frac{U_i^j}{n}, \; i = 1, \ldots, n; \; j = 1, \ldots, d \tag{2.11}$$

where π_i^j is the value to which i is mapped by the j-th permutation. Note that for LHS the independence of the random vector components is fundamental. Applying LHS to random vectors with dependence will destroy the dependence because of the random and independent permutations in each dimension. For this reason, Packham and Schmidt [157] proposed an extension of the Latin Hypercube Sampling from independent random vectors to random vectors with dependence, see Eq. (2.12). This extension allows us to make use of the important variance reduction property of the LHS, while taking into account the dependence structure between the random vector components. The main idea behind the modification is to choose a specific permutation in each dimension instead of randomly drawing a permutation. The specific permutation depends on the rank of the random variables. As in our example (Figure 2.5), the components of the random vectors are linked by a copula (e.g. Gaussian copula) and the sample generated

by LHSD is given as:

$$V_{\mathrm{i,n}}^{j} = \frac{r_{\mathrm{i,n}}^{j} - 1}{n} + \frac{\eta_{\mathrm{i,n}}^{j}}{n},\ i = 1,\ldots,n;\ j = 1,\ldots,d \tag{2.12}$$

where $r_{\mathrm{i,n}}^{j}$ is the rank of the original random vector (X_i^j in set $\{X_k^j : 1 \le k \le n\}$) and $\eta_{\mathrm{i,n}}^{j}$ is a possibly random value in $[0,1]$ (e.g. $\eta_{\mathrm{i,n}}^{j} = 0.5$ means that each sample is in the middle of its stratum, a computationally efficient choice). For a thorough description of LHSD the reader should refer to [157].

Latin Hypercube Sampling with Dependence (LHSD) steps (see Figure 2.6):

1. **Random vectors are generated from a Gaussian copula:** creating random vectors from the multivariate distribution by specifying the mean vector (equal to zero), the sample size and the covariance matrix (based on the rank correlation matrix). These random vectors are then transformed into uniform marginal with the probability integral transform using the centered normal cumulative distribution functions with appropriate variance.

2. **Copula vectors are redistributed to form a Latin hypercube:** Sampling random vectors from the multivariate distribution does not ensure desired spread of the samples. This can be improved by redistributing the copula vector to form a Latin hypercube, in this way providing good value coverage of the parameter space. The first step is to compute the rank of the vectors and constraining the samples with the rank. Applying this method we can get spread out marginal values on the unit intervals and as a result the random vectors are evenly distributed. The benefit of this step is visible in Figure 2.5).

3. **Creating sample distributions by reverse steps:** The redistributed copula random values are transformed back to the desired marginal distributions by applying reverse steps (using inverse CDF), maintaining their ranks (together with their rank correlations) from the original random sample. One advantage of this method is that different inverse CDFs can be chosen for each parameter, which gives us the possibility to deal with the parameters separately. In this study this is an important feature since three of the parameters (denitrification rate in water column, denitrification rate in sediment and burial rate) are transformed to exponential distribution, while the rest of the parameters are transformed to normal distribution. The resulting sample distributions can be found in Figure 2.7.

The sample size $n = 160$ was chosen based on [108] who conducted a simple analysis to identify the optimal sample size. Applying the LHSD, evenly stratified samples could be generated (see Figure 2.7) while maintaining the parameter dependency. Samples that fall outside of the predefined range were removed in order to avoid values with non-physical meaning (truncation). Truncation could have been avoided by using a distribution restricted to the given range (conditioning). The parameter statistics and distributions are described below.

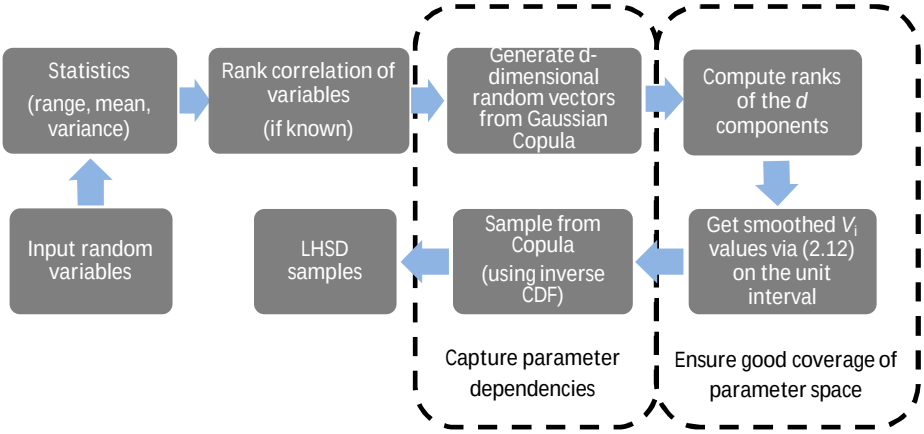


Figure 2.6: Latin Hypercube Sampling with Dependence (LHSD) steps.

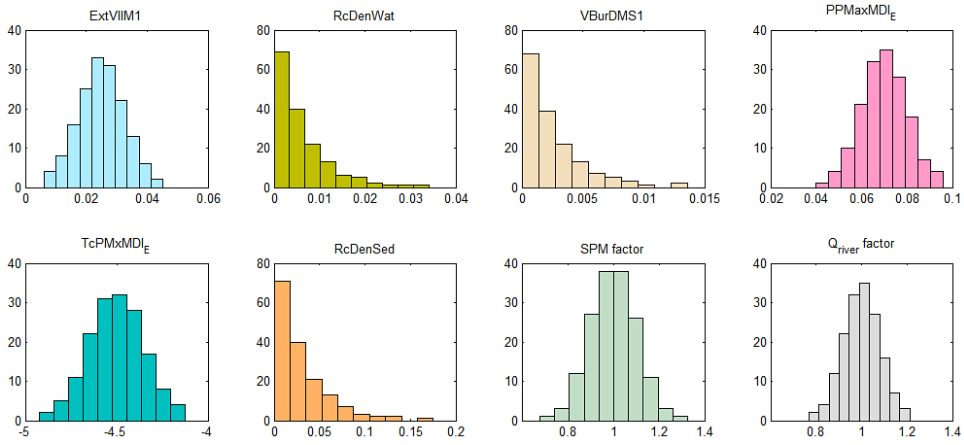


Figure 2.7: Sample distribution of the selected significant parameters from Latin Hypercube Sampling with Dependence, $n = 160$.

Table 2.4: Parameter statistics for sampling

No	*Parameter*	*Range*	*Mean*	*Standard deviation*
1	$e_{s,IM}$	0.01-0.05	0.025	0.0075
2	$R_{den\ ,wat}$	0.00-0.20	0.006	
3	b_{S1}	0.00-0.25	0.0025	
4	$kgp^{0}_{Di,E}$	0.05-0.15	0.07	0.01
5	$ktpg_{Di,E}$	(-5.00) - (-1.75)	-4.5	0.15
6	$R_{den\ ,sed}$	0.00-0.20	0	
7	SPM	0.7-1.3*	1	0.1*
8	Q_river	0.75-1.25*	1	0.08*

**based on the estimated uncertainty level*

Parameter statistics

The parameter statistics (possible value range, mean and standard deviations) are given in Table 2.4. The mean values are assumed to be the calibrated baseline values of the deterministic model setup. The parameter ranges for process parameters were mainly found in [174] and [30], whereas for the model inputs value ranges are set according to the estimated uncertainty level. The uncertainty level in the river discharges was assumed as 25% based on a study conducted by [128] on discharge and nutrient uncertainty in streams in the UK, which reported observational discharge uncertainties ranging from ±2 to 25%. The standard deviations for few process parameters are used as described in [108]. For the remaining process parameters, and for the model inputs, the standard deviations are assumed in such a way that the resulting sampled values fall between the specified value ranges, given that for normal distribution 99.7% of the data are within 3 standard deviations of the mean. Information about the parameter dependency was available for 4 process parameters from [108] as rank correlation coefficient (Spermann's rho), while having no information about dependency for the rest of the parameters they are assumed to be independent (rank correlation coefficient equal to zero).

Parameter distributions

In this case study, the selected parameters are not measured/observed quantities but model process parameters governing the process formulations in the numerical model. For this reason, empirical distributions could not be obtained and expert elicitation was required to gain further information on them. Consequently, the theoretical distributions were assumed according to the literature [108], as the best available knowledge. Based on the literature [108], exponential distributions were assumed for some parameters, whereas normal distributions were applied to others. Note that the exponential distribution is parameterized in terms of the scale parameter $\beta = \frac{1}{\lambda}$, which is the mean. Moreover, these distributions were adopted partly due to the fact that in the sampling procedure a Gaussian Copula model is used, which has the underlying assumption that the joint distributions of the random variables are symmetric. Another possibility could have been to assume uniform distributions. [72] investigated the influence of the choice of prior distributions in Bayesian uncertainty estimation methods for water quality mod-

elling. It was demonstrated for a specific case study that in cases when information about the parameters is not available (or weak), choosing uniform prior distributions might be more appropriate. The reasoning was that applying normal distribution (without evidence that the parameters are indeed normally distributed) may result in wrong estimation of uncertainty as it leads to a narrower uncertainty band as compared to the uniform distribution [61, 72]. This may create excessive confidence in the model results especially in water quality modelling applications where the inherent uncertainties are high.

2.7. FORECAST VERIFICATION AND PERFORMANCE ANALYSIS

With the help of multiple verification metrics, different forecast attributes can be assessed and quantified. In this paper the Mean Absolute Error (MAE), Root Mean Square Error (RMSE) [57] and Percent Bias (PBIAS) [90] are applied to quantify the error between the observation and forecast. In addition, further accuracy measures such as the Brier Score (BS) [33] and the Continuous Ranked Probability Score (CRPS) [39] are included in the verification. The CRPS can be calculated for single-valued and probabilistic predictions, allowing a direct comparison between the two types of forecasts. The forecast reliability is assessed with the False Alarm Ratio (FAR), whereas the discrimination attribute is evaluated with the Probability Of Detection (POD) [142]. Finally, the model's goodness-of-fit was determined by the Index of Agreement (IoA), Coefficient of determination (R2) and Nash-Sutcliffe (NS) model efficiency [57]. All the metrics are collected in Table 2.5 together with their perfect scores. The computed verification metrics are then used to assess the improvement in forecast performance. Table 2.6 shows the relationship between the changes in verification metrics, grouped by forecast attributes, and the corresponding changes in forecast performance.

Some of the above mentioned metrics such as the POD and FAR use the comparison of the measured and forecast occurrence frequency of a predefined event. In this study an event is defined when the chlorophyll-a concentration in the top layer of the water column exceeds the elevated assessment levels. The mean elevated assessment level during the growth season (March-September) is 7.5 mg/cm^3 in the Dutch coastal waters and 12 mg/cm^3 in the Dutch Wadden Sea. These levels were previously specified by [16] for the application of the OSPAR Comprehensive procedure.

2.8. RESULTS AND DISCUSSION

The above presented ensemble method was tested for hindcasting the chlorophyll-a concentration in the southern North Sea for different hydrodynamic years, first for year 2009 and then for year 2007. The ensemble method's performance is analysed separately in the Dutch Wadden Sea and Dutch coastal waters using three observation types with varying temporal coverage and accuracy. In-situ measurements are considered as the most reliable data but their temporal coverage is low, only 15 measurements per year on average. In comparison, the remote sensing data has increased temporal coverage, 30 MERIS observations per year and 200 gap-filled MERIS observations per year on average, allowing better verification of the model but introducing more uncertainty. Ensemble forecasts are often assessed by verifying how well the ensemble band could capture

Table 2.5: Verification metrics used for performance analysis. The bar indicates average as in $\overline{(f-o)^2}$.

Verification metrics	*Formulas*	*Perfect score*
Index of Agreement	$d = 1 - \frac{\sum_{i=1}^{n}(o_i - f_i)^2}{\sum_{i=1}^{n}(\lvert f_i - \bar{o}\rvert + \lvert o_i - \bar{o}\rvert)^2}$	1
Coefficient of determination	$r^2 = \frac{[cov(f,o)]^2}{[var(f)][var(o)]}$	1
Nash-Sutcliffe	$E = 1 - \frac{\sum_{i=1}^{n}(o_i - f_i)^2}{\sum_{i=1}^{n}(o_i - \bar{o})^2}$	1
Mean Absolute Error	$MAE = \overline{(\lvert f - o\rvert)}$	0
Root Mean Square Error	$RMSE = \sqrt{\overline{(f-o)^2}}$	0
Percent Bias	$PB = 100 \times \left[\sum_{i=1}^{n}(f_i - o_i) \Big/ \sum_{i=1}^{n} o_i\right]$	0
False Alarm Ratio (FAR)	$FAR = \frac{False\ alarms}{Hits + False\ alarms}$	0
Probability of Detection (POD)	$POD = \frac{Hits}{Hits + Misses}$	1
Continuous Ranked Probability Score (CRPS)	$\int_{x=-\infty}^{x=\infty}\left(F_i^f(x) - F_i^o(x)\right)^2 dx$ F_i^f - CDF of forecast, F_i^o - CDF of observation	0
Brier Score (BS)	$BS = \overline{(f-o)^2}$, where $f = [0,1], o = 0\ or\ 1$	0

Table 2.6: Relationship between changes in verification metrics and forecast performance

Performance improvement	*Verification metrics*
Improved goodness-of-fit (regression line of the forecast better approximates the observations)	(↑) Index of Agreement (IoA) (↑) Coefficient of determination (R^2) (↑) Nash-Sutcliffe (NS)
Improved accuracy (reduced error measures)	(↓) Mean Absolute Error (MAE) (↓) Root Mean Squared Error (RMSE) (↓) Percent Bias (PBIAS) (↓) Continuous Ranked Probability Score (CRPS) (↓) Brier Score (BS)
Improved reliability	(↓) False Alarm Ratio (FAR)
Improved discrimination attribute	(↑) Probability of Detection (POD)

(↑) - Increase, (↓) - decrease

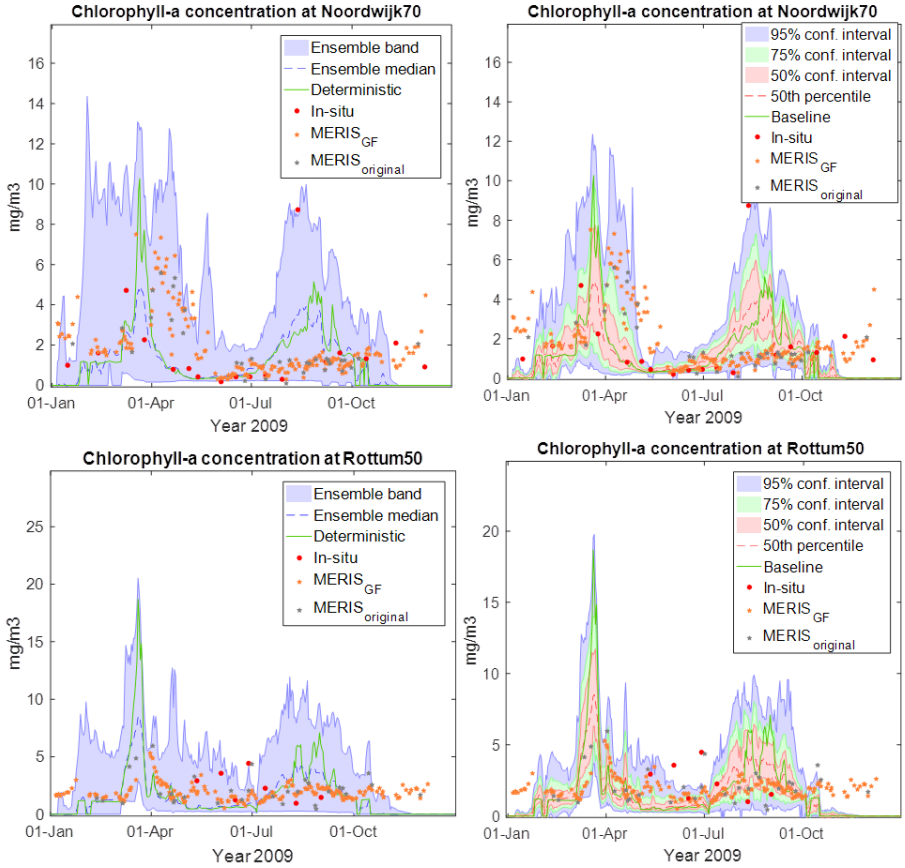


Figure 2.8: Ensemble band (left) and confidence intervals (right) with the measurements at station Noordwijk 70 and Rottum 50, year 2009.

the observations. Figure 2.8 and Table 2.7 show the percentage of measurements captured by the ensemble band and the confidence intervals (50%, 75%, 95%) at the in-situ stations. Overall, it can be concluded that the percentage of measurements lying within the ensemble band is relatively low, 60% on average considering all stations and both years. The percentage results for the 95% and 75% confidence intervals are even lower, but the accuracy of the former is still acceptable for the research. Nevertheless, particular differences can be observed in the ensemble forecast's skill once the two areas are analysed separately. In the Wadden Sea only 35 to 57% of the measurements could be captured in the ensemble band while in the coastal waters this percentage is as high as 67 to 80%.

In this paper processed remote sensing images are also applied as verification dataset. Given the high degree of uncertainty in this dataset, it is advised to provide an error estimate together with the results. The MERIS data is compared with the in-situ measurements to obtain the error estimate at the stations where sufficient number of matchups

Table 2.7: Percentage of measurements captured in the ensemble band at the stations, and average results for the 95%, 75% confidence intervals, years 2009 and 2007

No.	*Station*	*In-situ [%]*	*MERIS [%]*	*MERIS_GF [%]*	*In-situ [%]*	*MERIS [%]*	*MERIS_GF [%]*
			2009			*2007*	
				Dutch Wadden Sea			
1	Marsdiep	45.0	28.6	45.9	25.0	58.8	56.1
2	DoovBW	33.3	70.4	61.3	n/a	n/a	n/a
3	Vlies	58.3	51.6	51.9	36.4	50.0	44.9
4	Dantzig	19.0	50.0	41.4	26.3	30.8	30.7
5	ZuidOlWot	42.9	41.7	21.0	57.9	70.6	75.6
6	Harling	11.1	0.0	16.6	50.0	74.1	74.1
Average (ensemble band)		**34.9**	**40.4**	**39.7**	**39.1**	**56.9**	**56.3**
Average (95% conf. int.)		**23.8**	**35.5**	**30.3**	**24.8**	**45.3**	**43.4**
Average (75% conf. int.)		**13.9**	**21.7**	**16.6**	**16.3**	**29.4**	**27.0**
				Dutch coastal waters			
7	Rottum3	66.7	80.0	70.7	70.0	76.1	87.3
8	Rottum50	85.7	90.6	86.7	100.0	74.5	82.9
9	Rottum70	57.1	75.0	76.2	100.0	72.3	80.0
10	Terschelling10	61.1	87.5	81.8	64.3	64.5	62.4
11	Terschelling50	82.4	88.6	84.0	n/a	n/a	n/a
12	Noordwijk10	58.6	72.2	69.1	27.6	35.5	42.9
13	Noordwijk20	47.1	70.8	56.9	38.9	44.7	40.0
14	Noordwijk70	77.8	90.9	87.8	78.9	76.0	85.9
15	Goeree6	66.7	65.2	69.1	63.6	66.7	78.5
Average (ensemble band)		**67.0**	**80.1**	**75.8**	**67.9**	**63.8**	**70.0**
Average (95% conf. int.)		**54.1**	**68.2**	**62.8**	**51.3**	**53.6**	**57.5**
Average (75% conf. int.)		**33.6**	**43.7**	**42.2**	**38.8**	**32.0**	**36.0**

are available. Considering all stations, the MERIS dataset has an average Mean Absolute Error (MAE) of 41% in 2009 and 48% in 2007, whereas the DINEOF gap-filled MERIS data's average error is 52% and 48% for the same years. Thus, applying averaged error bands for the remote sensing dataset instead of the uncertain single point values might provide a better picture of the ensemble forecast's performance. The reader may refer to [57] for further information about considering measurement uncertainty in the evaluation of goodness-of-fit metrics in water quality modelling.

Taking into account the measurement uncertainty in the remote sensing data, the percentage of captured measurements would increase considerably up to 84% in the Wadden Sea, and up to 89% in the coastal waters, however, 100% coverage could not be achieved. A possible reason for this might be that the coastal zone and the Wadden Sea are shallow, dynamically varying ecosystems with high turbidity and therefore the deterministic simulation, which serves as the base for the ensemble forecast, should be recalibrated here to achieve improved results. This model uncertainty should be considered when evaluating the results. Moreover, the fact that not all uncertainty sources were taken into account may explain the additional uncertainty stemming from the model structure and from the meteorological- and hydrodynamic inputs. Finally, the presumed level of uncertainty in the inputs (e.g. SPM concentration field), might be underestimated or their assumed sample distribution may not be appropriate.

Table 2.8: Averaged % improvement in the groups of verification metrics if the ensemble median prediction is used instead of deterministic forecast

	*% improvement**			
	2009		*2007*	
Group of verification metrics	*Dutch coastal waters*	*Dutch Wadden Sea*	*Dutch coastal waters*	*Dutch Wadden Sea*
Goodness-of-fit	28%	4%	18%	12%
Accuracy	16%	8%	12%	7%
Reliability	7%	-13%	0%	-4%
Discrimination	6%	-26%	1%	-12%

** Group average using the three validation datasets. The metrics within the groups are not weighted*

Forecast verification metrics

The direct comparison of the deterministic and ensemble forecast was achieved through calculating a selection of verification metrics, (introduced above) for both types of predictions. In order to be able to compute deterministic scores for the ensemble forecast, the ensemble median (50^th percentile) prediction was selected in this study, although further options are available such as the ensemble mean or any other percentile. The ensemble median (or ensemble mean) usually verifies better than the deterministic forecast by most verification scores, because it presents the most predictable elements of the forecast and smoothes out the extreme unpredictable elements. It indicates the future values of the model output variable that can be predicted with confidence, but it will rarely capture the extreme events, and therefore should not be relied upon on its own.

The percentage improvement in the groups of verification metrics if the ensemble median prediction is used instead of the deterministic forecast is presented in Table 2.8. It is important to note that all metrics are equally weighted within the groups. Nevertheless, it might be possible that in other studies the metrics would be weighted or some specific metrics would be omitted according to the forecasters' need. The group members are shown in Table 2.6. The verification metrics are averaged values of 6 stations in the Dutch Wadden Sea and 9 stations in the coastal waters. Furthermore, the percentage improvements in the metrics are computed as an average of the results using all measurement types.

The goodness-of-fit and accuracy metrics show moderate improvement in both years and both areas but the magnitude of the improvement is markedly higher in the coastal waters nonetheless. The Brier Score is only calculated for the ensemble forecast and thus it cannot be compared to the deterministic prediction. Considering all measurement types the average Brier Score in the Wadden Sea is 0.13 in 2009 and 0.12 in 2007, whereas at the Dutch Wadden Sea stations the Brier Scores for the same years are 0.05 and 0.09. Reliability and discrimination measures only experience a minor improvement in the coastal waters and no improvement at all in the Wadden Sea. The verification results confirm the previous finding that the ensemble forecast performs better in the Dutch coastal waters than in the Wadden Sea. In the problematic Wadden Sea area the ensemble method's efficiency is less convincing; on the other hand in the coastal waters the preliminary results are promising.

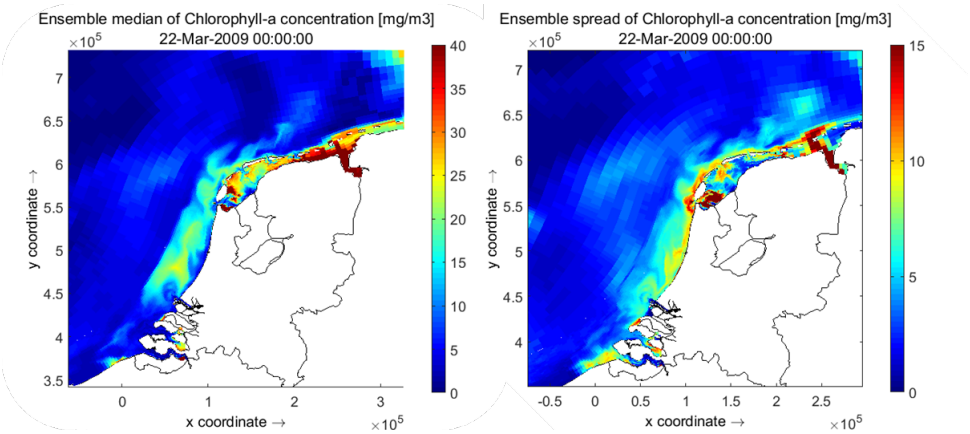


Figure 2.9: Ensemble median prediction (left) and corresponding ensemble spread (right) of chlorophyll-a concentration along the Dutch coast during the peak spring algal bloom in 2009.

Spatial results

Making use of the three dimensional water quality model the forecast can be visualized as chlorophyll-a concentration map. As mentioned above, the ensemble median prediction should not be relied upon on its own but with the indication of the degree of certitude. The ensemble spread can provide a measure of the level of uncertainty in the output parameter; hence, it is appropriate to complement the ensemble median. Figure 2.9 illustrates the ensemble median prediction map at a specific time step together with the ensemble spread map calculated as the standard deviation of the chlorophyll-a concentration.

Observing the ensemble spread map allows us to draw conclusions on the prediction's spatial uncertainty. In late March, during the peak spring algal bloom, the prediction in most off-shore areas has low standard deviation, while significant spread can be observed in the near shore, shallow areas of the North Sea and in the Dutch Wadden Sea. This might be partly due to the direct effect of perturbations on the river loads in the focus area and the high spread may indicate the rivers' zone of reach. Nonetheless, it could be also explained with the specific regional system dynamics, since in the near shore shallower zones the water is low-dynamic and the primary production levels are already elevated.

2.9. CONCLUSIONS

This paper contributes to existing knowledge in probabilistic forecasting by providing an application to water quality prediction with a three dimensional ecosystem model. Our results indicate that ensemble prediction techniques can produce enhanced forecast of water quality indicators due to their ability to account for the error in the model output variable. Nevertheless, the potential of the ensemble forecasting system ultimately depends on the input ensemble. Moreover, the sampling technique that is used to generate the input ensemble should be tailored to the specific requirements of the application.

In the presented case study the ensemble forecast moderately out performs the deterministic forecast at the coastal stations, which might be advantageous in decision making if the underlying baseline (deterministic) model is sufficiently well calibrated and validated. Moreover, the degree of certitude in the forecast and the likelihood of a predicted event, which could be expressed through ensemble forecasting, provide opportunity to set risk-based criteria for the response measures for decision makers. In other words, the uncertainty estimate produced by the proposed ensemble forecast promotes rational decision making, and offers potential for additional ecosystem and economic benefits.

Recommendations

Due to their fundamental importance in the methodology the identified important parameters should be revised based on a comprehensive sensitivity analysis. In addition, the parameter value ranges and distributions should be re-evaluated for future applications. Furthermore, the uncertainty from the atmospheric inputs should also be included in the ensemble generation in order to better estimate the model input uncertainty. Even so, if the aim is to express climate induced uncertainties. The spatial correlation of the suspended particulate matter concentration field should also be considered to achieve a more realistic perturbation on that model input.

3

STATISTICAL UNDERPINNING OF ATMOSPHERIC VARIABLES SELECTION

Coastal climate impact studies make increasing use of multi-source and multi-dimensional atmospheric and environmental datasets to investigate relationships between climate signals and the ecological response. The large quantity of numerically simulated data may, however, include redundancy, multi-colinearity and excess information not relevant to the studied processes. In such cases techniques for feature extraction and identification of latent processes prove useful. Using dimensionality reduction techniques this chapter provides a statistical underpinning of variable selection to study the impacts of atmospheric processes on coastal chlorophyll-a concentrations, taking the Dutch Wadden Sea as case study. Dimension reduction techniques are applied to environmental data simulated by the Delft3D coastal water quality model, the HIRLAM numerical weather prediction model and the Euro-CORDEX climate modelling experiment. The dimension reduction techniques were selected for their ability to incorporate (1) spatial correlation via multi-way methods, (2) temporal correlation through Dynamic Factor Analysis, and (3) functional variability using Functional Data Analysis. The data reduction potential and explanatory value of these methods are showcased and important atmospheric variables affecting the chlorophyll-a concentration are identified.

Our results indicate room for dimensionality reduction in the atmospheric variables (2 principle components can explain the majority of variance instead of 7 variables), in the chlorophyll-a time series at different locations (two characteristic patterns can describe the 10 locations), and in the climate projection scenarios of solar radiation and air temperature variables (a single principle component function explains 77% of the variation for solar radiation and 57% of the variation for air temperature). It was also found that solar radiation followed by air temperature are the most important atmospheric variables related to coastal chlorophyll-a concentration, noting that regional differences exist, for

instance the importance of air temperature is greater in the Eastern Dutch Wadden Sea at Dantziggat than in the Western Dutch Wadden Sea at Marsdiep Noord. Common trends and different regional system characteristics have also been identified through dynamic factor analysis between the deeper channels and the shallower intertidal zones, where the onset of spring blooms occurs earlier. The functional analysis of climate data showed clusters of atmospheric variables with similar functional features. Moreover, functional components of Euro-CORDEX climate scenarios have been identified for radiation and temperature variables, which provide information on the dominant mode (pattern) of variation and its uncertainties. The findings suggest that radiation and temperature projections of different Euro-CORDEX scenarios share similar characteristics and mainly differ in their amplitudes and seasonal patterns, offering opportunities to construct statistical models that do not assume independence between climate scenarios but instead borrow information ("borrow strength") from the larger pool of climate scenarios. The presented results were used in follow up studies to construct a Bayesian stochastic generator to complement existing Euro-CORDEX climate change scenarios (see Chapter 4) and to quantify climate change induced trends and uncertainties in phytoplankton spring bloom dynamics in the Dutch Wadden Sea (see Chapter 5).

Parts of this chapter have been published in Frontiers in Marine Science - Coastal Environmental and Ecological Data Analysis (2022)

3.1. INTRODUCTION

Motivation - *The present study is part of an overarching research investigating possibilities for statistical quantification of climate change induced uncertainties in future coastal ecosystem state. The research builds on a multitude of data sources, prominently using numerical models. As the research focuses on statistical methods to quantify and propagate uncertainties, a proper understanding of the multivariate input data, its redundancy, and most importantly the identification of latent variables and extraction of features is a natural first step in the analysis. A host of methods for dealing with these issues is available in the literature but scattered over various disciplines, such as chemometrics, econometrics and mathematics. This chapter investigates how these methods can be applied to achieve the higher level objectives: (1) providing statistical underpinning for atmospheric variables selection to study chlorophyll-a response, and (2) identifying important features of the climate projections for further statistical models, for instance the Bayesian stochastic generator implemented in [148]. More specifically, in this chapter a case study (Dutch Wadden Sea) is presented, first introduce the main idea of selected statistical methods, subsequently applying them to a particular dataset (consisting of coastal biogeochemical model, numerical weather prediction model and climate model outputs) and interpret the results. While the applied statistical methods are separately well documented in the literature (in their own fields), structured and combined use of them for the multivariate analysis of air-sea interactions to informing ecological impact studies is a novelty to the marine scientific community.*

Scientists aiming to study the air-sea interactions either in (operational) short term or (climate) long term scale often make use of numerical models, which produce approximate solutions to the underlying physical phenomena. The role of these physics-based models is even more prominent with the increasing (cloud) computing capabilities [208] that facilitate further refined spatial scales and improved process parametrizations. Using these models, gap-free (in space) and high frequency (in time) fields of atmospheric and environmental datasets can be produced. Such multi-dimensional numerical model simulated dataset often includes several variables at many locations (e.g. three dimensional spatial discretization) over long periods of time and covering different model scenarios (e.g. various model boundary conditions and model initializations). While the increasing volume of marine data contains abundant information and insights into the physical processes (also their interconnections and long term evolution), it must be noted that the processes underlying the variations in these simulated data are complex, the data might be noisy, and not all modelled variables are relevant to the studied processes. Consequently, latent variables can be useful for exploring and reducing the data. Traditionally, dimension reduction methods are used for such purposes.

Dimension reduction is an approach often used in multivariate data analysis and it is implemented for several reasons. Firstly, using dimension reduction techniques high-dimensional data can often be transformed to a lower dimensional space without significant loss of statistical information (preserving accuracy). Secondly, dimension reduction techniques help in the removal of multi-colinearity in the dataset. The multicolinearity problem is present if two or more variables are highly correlated, and therefore one can be accurately linearly predicted from the others. This is an unwanted prop-

erty as it increases the variance in estimates of regression parameters [139] and makes interpretation difficult. A further advantage of dimension reduction is that it facilitates the interpretation and visualization of high dimensional data as it is reduced to lower dimensions. Additionally, transforming data into lower dimensions decreases the required processing time and storage, and therefore makes analysis algorithms more efficient.

Various dimension reduction methods exist, some use linear combinations of variables to reduce dimensions (linear methods), whereas others use non-linear functions of variables (non-linear methods). A collection of non-linear dimension reduction methods can be found in [95]. The most widely used linear dimension reduction techniques are the Principal Component Analysis (PCA), an unsupervised technique, and the Partial Least Squares (PLS) [139], a supervised technique. These are useful dimension reduction methods in regression problems due to the following features. Firstly, applying the transformed principal components instead of the original predictive variables tackles the problem of multi-colinearity since the covariance of principal components is zero. Secondly, the principal components successively capture the maximum variance of the predictor matrix, and therefore it is natural to use the first few components as predictive variables for regression. In most cases the majority of the variance is captured by them.

While their concept offers clear advantages, a practical limitation of these standard dimension reduction methods is that they work with "2-way" matrices. The 2-way structure usually contains the observations as rows and the variables as columns. A third way of the matrix, that could be the temporal or spatial dimension for instance, cannot be explicitly included. Multi-way analysis can help to resolve this issue. Multi-way analysis techniques also project variables to low dimensional spaces, therefore they can be called dimension reduction methods, but they are also able to work with multi-way ($N > 2$) data structures. Similarly to the other dimension reduction techniques, multi-way analysis can create latent variables by transforming the original variables, it can reduce noise, and it can explain which original variables are most important to the latent variables [185]. Further purpose of applying multi-way methods is data exploration, which includes finding patterns and interrelations (e.g. temporal and spatial behaviour of the different variables), or summarizing the data through decomposition.

Another missing feature in standard dimension reduction techniques that is quite essential in atmospheric and environmental time series is temporal correlation. For this reason, temporal correlation is included in this research through Dynamic Factor Analysis (DFA). Moreover, in this study the discrete-time data are also investigated using Functional Data Analysis (FDA), after transforming them to functional data through a basis function expansion. This is also motivated by the fact that certain variables display 'strong periodic behaviour', such as the sinusoidal shape of air temperature or solar radiation. Similarly to the dimension reduction techniques on discrete-time data, Functional Data Analysis also aims to find common patterns and underlying functions that can describe the general shape of the curves and explain their variability.

In this chapter the above described statistical models are applied to atmospheric and environmental datasets in the Dutch Wadden Sea to investigate the relationships between atmospheric signals and the ecological response. Due to the complex interactions of atmospheric forcing with biological processes, the phytoplankton response is not trivial to understand, especially in our case study area. Considering the system dy-

namics, the southern North Sea is a tidally mixed region [129] but in our study area other shallow water, coastal, and estuarine fronts are also prominent. This makes it possible that certain regions are seasonally stratified while others are permanently mixed [122]. Consequently, in the offshore areas surface mixing and convective cooling have a greater impact on phytoplankton biomass [31], while in the highly dynamic coastal systems tidal mixing is more dominant.

The relationship between physical factors (atmospheric and oceanic) and the selected ecological response variable (chlorophyll-a) is well documented in the literature, nevertheless, debates still exist between scientists. In general, chlorophyll-a concentration (a proxy for phytoplankton biomass) is coupled to thermal stratification, resource and energy dynamics, as well as predator-prey interactions [22]. Based on a cross correlation analysis conducted by [31] in the North Sea (at a site with dynamics similar to our study area), the highest correlations were found with solar radiation, air temperature, turbidity, and tidal mixing. This study considered a range of physical factors (tidal mixing, wind mixing, solar radiation, air temperature, SST, salinity, turbidity) and chlorophyll-a. [143] found that inter-annual variability in phytoplankton dynamics in North Atlantic coastal waters were related to solar radiation, sea surface temperature, as well as Si availability. On the other hand, in the offshore regions it was mainly regulated by temperature, Atlantic inflow, wind stress and North Atlantic Oscillation (NAO). Moreover, in his study describing interannual changes in phytoplankton seasonality due to climate forcing, [84] used the following variables: sea surface temperature that impacts the physiological and ecological processes and is a tracer of vertical mixing; solar radiation that limits phytoplankton growth rates or increases pigment cell levels; wind that is responsible for surface mixing and turbulence; and ocean current variability impacting stratification. [111] also found that atmospheric variability are associated with chlorophyll-a concentration changes but the study considered large-scale modes of atmospheric variability. A shortcoming of our study is that it focuses on a small-scale coastal area, therefore large scale processes cannot be revealed.

3.2. MATERIALS AND METHODS

This research aims to support ecological impact studies in coastal ecosystems by providing a statistical framework for investigating latent processes and selecting important atmospheric variables. This statistical framework contains three types of dimension reduction techniques (Figure 3.1). Firstly, discrete-time data is considered and temporal correlation is neglected. Supervised and unsupervised techniques are compared and spatial correlation is included through multi-way methods. Secondly, temporal correlation is incorporated by applying dynamic factor models. Lastly, the discrete-time climate data is transformed into functional data representation, by smoothing them with basis function expansion (e.g. Fourier basis expansion), and subsequently study the functional variation with Functional PCA. While discrete-time data is a set of discretely measured values $y_{i1}, \ldots, y_{in}$, functional data is when these values are converted to a function x_i with values $x_i(t)$ computable for any desired time t [164].

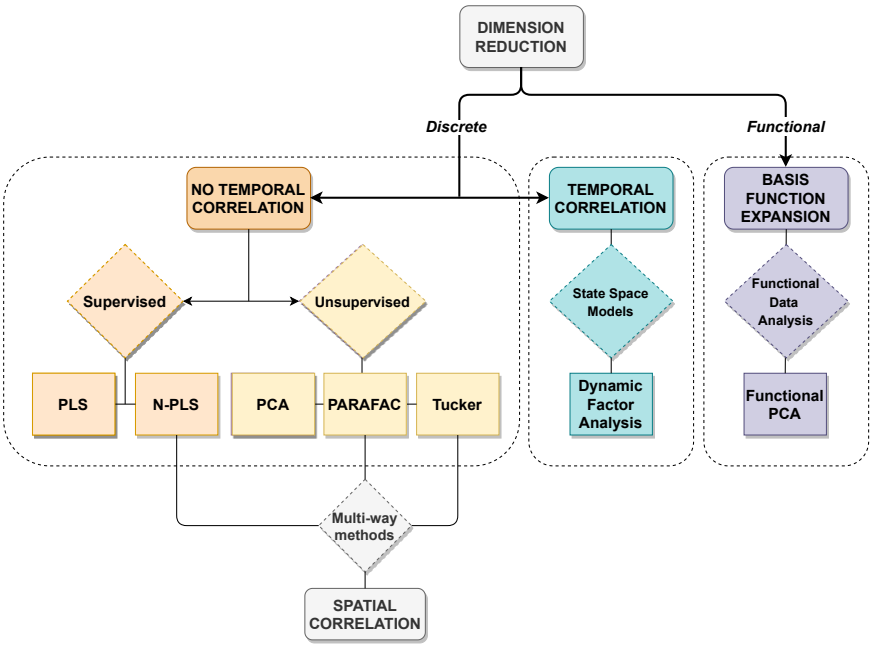


Figure 3.1: Overview of the applied statistical techniques for discrete-time and functional data, including temporal and spatial correlation.

3.2.1. DATASET

Our study is based on data from various numerical models (see Figure 3.2): a coastal water quality model, a numerical weather prediction model, and a climate model. The ecological indicator variable is chlorophyll-a concentration, a proxy for algal biomass, while the atmospheric variables are air temperature, solar radiation, eastern and northern wind components, air pressure, relative humidity, and total cloud cover. These are standard atmospheric variables simulated by most modelling systems for both operational purposes and climate experiments.

CHLOROPHYLL-A CONCENTRATION DATA

The chlorophyll-a concentration data is obtained from the water quality sub-module of the Delft3D integrated modelling system, Delft3D-WAQ (https://www.deltares.nl/en/software/delft3d-4-suite/) [30]. In this research an existing model setup is used, which has been previously calibrated and validated for the location of our study area [133]. The spatial domain of the physical model covers the Southern North Sea with coarser horizontal resolution offshore and finer resolution along the Dutch coast, as shown in Figure 3.3. The model comprises of twelve vertical layers, making it a three dimensional physical model. The horizontal resolution of the water quality model in the Dutch Wadden Sea ranges from 1-by-2 km to 2.5-by-3 km on a curvilinear grid.

Delft3D-WAQ is a comprehensive hybrid ecological model including an array of modules reproducing water quality processes that are then combined with a transport mod-

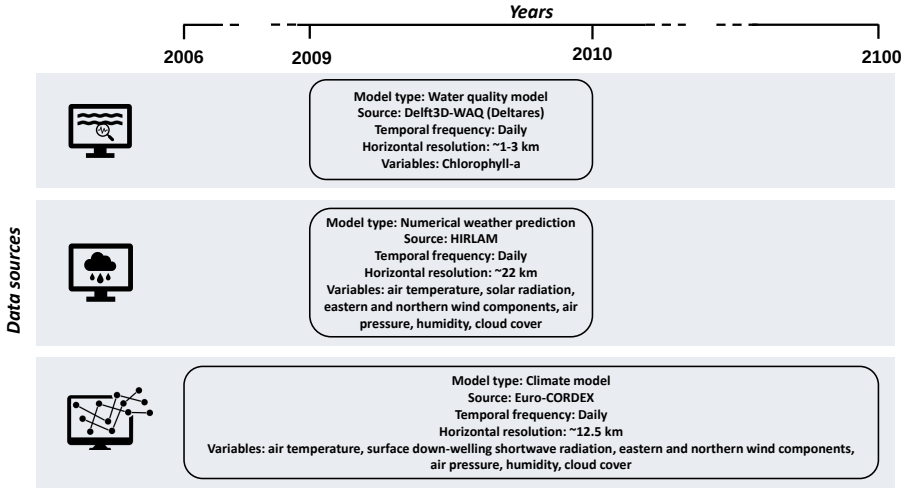


Figure 3.2: Overview of the data used in the study (model type, source, temporal frequency and variables) for the marine water quality model (top), numerical weather prediction model (middle), and climate model (bottom).

Figure 3.3: Case study area: Dutch Wadden Sea. Delft3D-WAQ model domain in the Southern North Sea and along the Dutch coast (left panel, source: [148]). Location of the stations where time series data was extracted (right panel).

ule to calculate advection and dispersion. The model most importantly calculates primary production and chlorophyll-a concentration while integrating dynamic process modules for dissolved oxygen, nutrient availability and phytoplankton species. This Delft3D-WAQ setup includes the phytoplankton module (BLOOM) that simulates the growth, respiration and mortality of phytoplankton. Using this module the species competition and their adaptation to limiting nutrients or light are simulated [133].

ATMOSPHERIC DATA

Two sources of atmospheric data are used in this study: (1) outputs of an operational numerical weather prediction model, and (2) results of a regional climate modelling experiment. First, the High Resolution Limited Area Model (HIRLAM) model [145] output is used, which was applied as atmospheric forcing for the Delft3D-WAQ model setup to compute chlorophyll-a concentration. HIRLAM is a Numerical Weather Prediction (NWP) system developed by the international HIRLAM programme (http://hirlam.org/) [204]. Since it is the Delft3D-WAQ input data that drives the processes, it allows the exploration of the correlations between atmospheric forcing and numerically computed ecological response. The data for this study are obtained from the 22 km grid resolution HIRLAM model and include near-surface air temperature, solar radiation, eastern and northern near-surface wind components, surface pressure, near-surface relative humidity, and total cloud cover. All HIRLAM model output variables were used in the Delft3D-WAQ model as temporally and spatially variable forcing fields except solar radiation, which is an area average, therefore the same for the entire domain.

Additionally, simulated values of climate variables are acquired from the high resolution 0.11 degree (~ 12.5 km) EURO-CORDEX Coordinated Regional Downscaling Experiment (`https://www.euro-cordex.net/`) [106], which uses the Swedish Meteorological and Hydrological Institute Rossby Centre regional atmospheric model (SMHI-RCA4) [175]. In order to produce various regionally downscaled scenarios, EURO-CORDEX applies a range of General Circulation Models (GCMs) to drive the above mentioned Regional Climate Model (RCM). The four driving GCMs in this study are the National Centre for Meteorological Research general circulation model (CNRM-CM5) [212], the global climate model system from the European EC-Earth consortium (EC-EARTH) [97], the Institut Pierre Simon Laplace Climate Model at medium resolution (IPSL-CM5A-MR) [62], and the Max-Planck-Institute Earth System Model at base resolution (MPI-ESM-LR) [79]. In addition to the driving models, further scenarios are obtained by considering different socio-economic changes described in the Representative Concentration Pathways (RCPs). RCPs are labeled according to their specific radiative forcing pathway in 2100 relative to pre-industrial values. This study includes RCP8.5 (high), and RCP4.5 (medium-low) [214] and four driving GCMs for the projection period between 2006-2100. Together the four different driving GCMs and two RCPs provide us with an ensemble of eight trajectories per climate variable. The climate variables included in the analysis are near-surface air temperature, surface downwelling shortwave radiation, eastern and northern near-surface wind components, surface pressure, near-surface relative humidity, and total cloud cover. For this dataset, near-surface means at a height between 1.5 to 10.0 m.

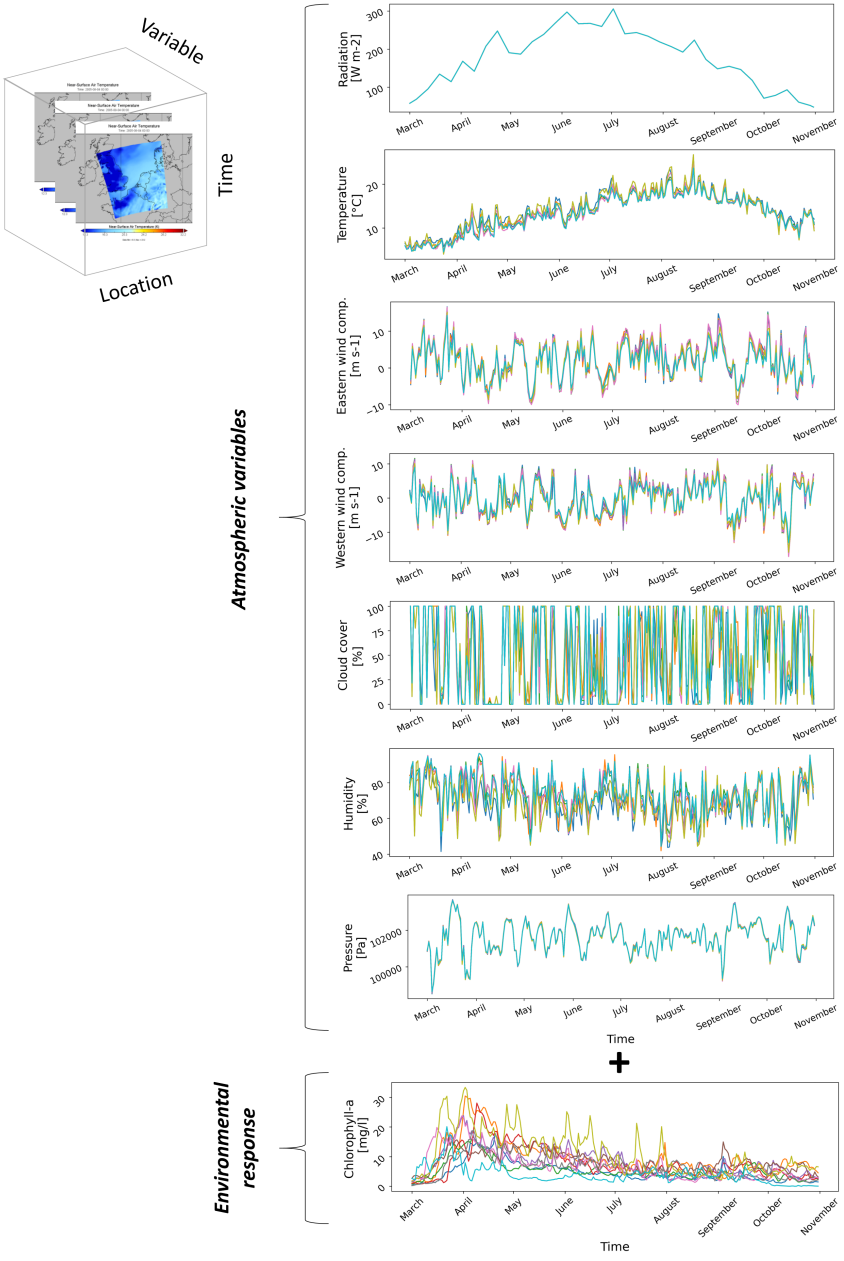


Figure 3.4: Illustration of the atmospheric and environmental variables used in the study

DATA PROCESSING

The above introduced datasets are temporally varying multivariate fields covering large domains. For the purpose of this study, time series data were extracted at ten locations of Rijkswaterstaat monitoring stations in the Dutch Wadden Sea (see Figure 4.9). Both the atmospheric variables and the chlorophyll-a concentration were provided as 6-hourly datasets. The longer and higher frequency data were sub-sampled to the period between 1st of March and 1st of November, daily at 12:00 (245 time steps). The model simulation year (2009) was chosen based on the fact that a detailed study was conducted (at Deltares) for that year with high resolution information on the suspended matter fields which are crucial for water quality computation in the shallow Wadden Sea. The reason for selecting a reduced time period (9 months) is to concentrate on the season of high phytoplankton productivity and to eliminate near zero chlorophyll-a values during winter. Moreover, the daily time step at 12:00 was selected to eliminate zero radiation values during the night. All variables were then centered to their mean and divided by their standard deviation to eliminate the problem of different measurement units. Finally, the right skewed chlorophyll-a concentration was log transformed to achieve a more symmetrical distribution that may improve the performance of statistical models used in the study. It is a standard practice to log transform chlorophyll-a as it is approximately lognormally distributed in marine waters [41]). The distribution of chlorophyll-a concentration (all locations and all time steps) before and after log transformation are shown in Figure S1 (Supplementary Material). The pair plot of all variables with kernel density estimation is displayed in Figure S2 (Supplementary Material).

Figure 3.5 shows the Spearman's rank correlation coefficient of all variables after scaling (data taken from all stations). The same plot using Pearson correlation coefficient can be found in Figure S3 (Supplementary Material). It can be observed that solar radiation and air temperature have the highest correlation with chlorophyll-a. Moreover, cross-correlation between the atmospheric data an also be identified, e.g. pressure and northern wind component or humidity and air temperature. It is important to note that while air temperature and solar radiation are positively correlated, they have different impact on chlorophyll-a concentration: air temperature has negative correlation, whereas solar radiation has positive correlation with chlorophyll-a. Since in the North Sea the correlation between solar radiation/air temperature and chlorophyll-a concentration highly depends on the region (offshore or coastal) and the temporal scales (short, seasonal, long) there could be various reasons. In our case, it might be attributed to the phenomena reported by [31], who found that the thermal mixing of phytoplankton cells (from the deep chlorophyll maximum) into the surface layer is the dominant process explaining the negative correlation between sea surface temperature and the chlorophyll concentration in the daily time series (in the Southern North Sea).

3.2.2. TWO-WAY AND MULTI-WAY METHODS

FROM PCA TO N-PLS

This section briefly introduces the steps to extend the two-way component methods to mutli-way regression methods. For convenience, Principal Component Analysis (PCA), Principal Component Regression (PCR) and ordinary PLS regression are introduced briefly, because the N-PLS regression is based on these algorithms. Assuming that $X \in \mathbb{R}^{I \times J}$ and

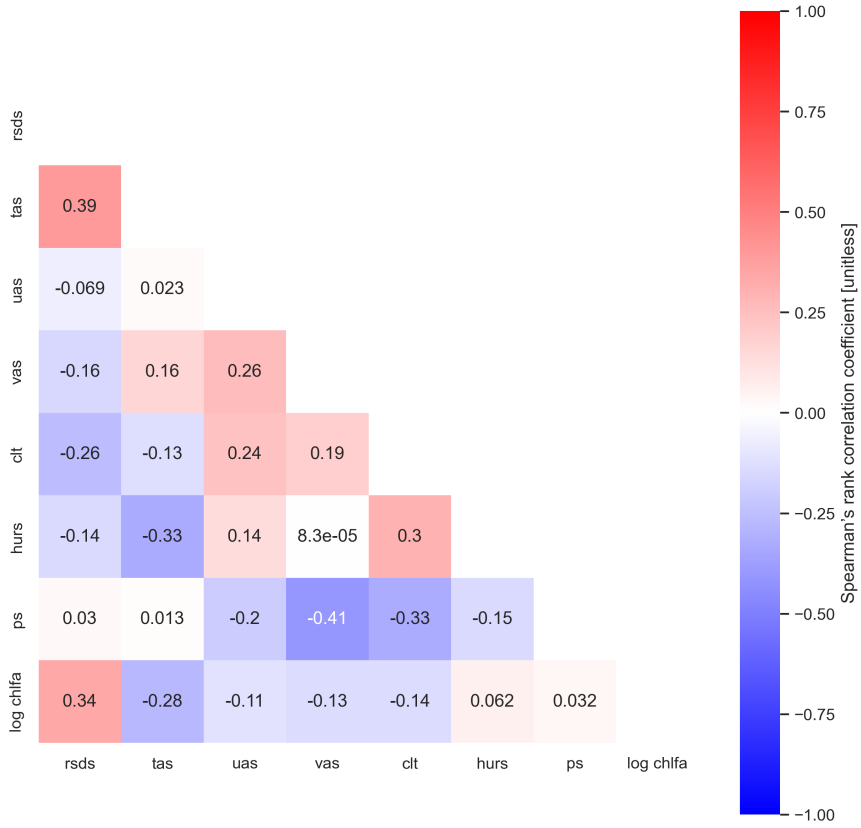


Figure 3.5: Heatmap with Spearman's rank correlation coefficient. Dark red indicates strong positive, while dark blue indicates strong negative correlations. Data from all time series. Abbreviations: solar radiation (rsds), air temperature (tas), eastern (uas) and northern (vas) wind components, cloud cover (clt), humidity (hurs), air pressure (ps), chlorophyll-a (chlfa).

$\underline{y} \in \mathbb{R}^I$ are column centred and scaled matrices, the predictor matrix X and response $\underline{y}$ are decomposed as follows:

$$X = TP' + E_X \tag{3.1}$$

$$\underline{y} = T\underline{q} + e_Y \tag{3.2}$$

where T is a matrix of scores ($T = XP$); P' is a matrix of X-loadings, q is a matrix of $\underline{y}$ loadings, whereas E_X and e_Y are the residuals. PCA focuses only on the predictor matrix projecting each data point onto the principal components while preserving as much of the data's variation as possible. PCA finds R components such that they maximize the variance of the projected data in X. The description below is written for $R = 1$. To calculate the 1^{st} PCA component, $\hat{x}_{ij}$ is approximated with t_i score and w_j loading:

$$\hat{x}_{ij} = t_i w_j \tag{3.3}$$

where $\underline{t} \in \mathbb{R}^I, \underline{w} \in \mathbb{R}^J, i \in \{1,\dots,I\}, j \in \{1,\dots,J\}$, and $||w|| = 1$. Then the score vector and loading vector can be obtained as follows:

$$\underline{t}(w) = \underset{\underline{t}}{\operatorname{argmin}} \sum_{i=1}^{I}\sum_{j=1}^{J} \left(x_{ij} - t_i w_j\right)^2 \tag{3.4}$$

$$\underline{w}^* = \underset{\underline{w}:||w||=1}{\operatorname{argmax}} \operatorname{var} \underline{t}(\underline{w}) \tag{3.5}$$

$$\Rightarrow \underline{t}^* = \underline{t}(w^*) \tag{3.6}$$

Then the approximation of $\hat{X}$ can be rewritten as:

$$\hat{X} = TP' \tag{3.7}$$

with $T = \underline{t}$, and $P' = \underline{w}$. Finally, the decomposition of X (for the 1^{st} PC situation) is obtained as:

$$X = \hat{X} + E_X \tag{3.8}$$

The PCR algorithm is similar to the PCA algorithm except that it is extended with response $\underline{y}$ using Eq. (3.2). In other words, PCR constructs R components the same way as PCA, but adds a regression step to it. Consequently, the regression coefficient $\underline{q}$ is obtained from regressing $\underline{y}$ on T:

$$\underline{q}^* = \underset{\underline{q}}{\operatorname{argmin}} ||\underline{y} - T\underline{q}||^2 = \left(T'T\right)^{-1} T'\underline{y}. \tag{3.9}$$

The PLS regression differs from PCR, due to its supervised nature, as it finds R components from both X and $\underline{y}$ such that covariance between the score vector $t(w)$ and $\underline{y}$ is maximized:

$$w^* = \underset{\underline{w}:||w||=1}{\operatorname{argmax}} \operatorname{cov}\left(\underline{t}(\underline{w}), \underline{y}\right) \tag{3.10}$$

$$\Rightarrow \underline{t}^* = \underline{t}(w^*) \tag{3.11}$$

Again, rewrite the approximation as Eq.(3.7) with $T = \underline{t}$ and $P = \underline{w}$. Then obtain the decomposition as in Eq. (3.8). Subsequently from Eq. (3.2) the regression coefficient q is obtained as in Eq. (3.9). As a consequence, PLS finds loading w that leads to a least squares solution to Eq. (3.3). Moreover, the PLS score vector has maximal covariance with $\underline{y}$.

In general, both PCA and PLS achieve dimension reduction by converting highly correlated variables to a set of uncorrelated variables through linear transformation. The difference is that PCA, as an unsupervised technique, captures maximum variance only in the predictor matrix without considering how each predictive variable may be related to the response variable. On the other hand, PLS combines information about the variances of both the predictors and the responses, while also considering the correlations among them (supervised dimension reduction). PLS is considered useful in particular if there are more independent (predictor) variables than dependent (response) variables, and if there is multi-colinearity in the predictors. Since in this study several correlated atmospheric variables are used to estimate one ecological response variable, the use of supervised dimension reduction techniques is preferable.

The N-PLS regression algorithm is an extension of the PLS regression algorithm to multi-way data, where essentially the bilinear model of X is replaced with a multilinear model of X. In case the data is three-way, as in this study, then an appropriate model of X is a trilinear decomposition, as depicted in Eq. (3.16). The model of x_{ij} in ordinary PLS is shown in Eq. (3.3), whereas in three-way PLS the approximation of x_{ijk} is given by the following equation:

$$\hat{x}_{ijk} = t_i w_j^J w_k^K \tag{3.12}$$

where $\underline{t} \in \mathbb{R}^I, \underline{w}^J \in \mathbb{R}^J, \underline{w}^K \in \mathbb{R}^K$. In this case the three-way decomposition is defined by:

$$\underline{t}\left(w^J, w^K\right) = \underset{\underline{t}}{\operatorname{argmin}} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \left(x_{ijk} - t_i w_j^J w_k^K\right)^2 \tag{3.13}$$

$$\left(w^{*J}, w^{*K}\right) = \underset{||w^J||=1, ||w^K||=1}{\operatorname{argmax}} \operatorname{cov}\left(\underline{t}\left(w^J, w^K\right), \underline{y}\right) \tag{3.14}$$

$$\Rightarrow \underline{t}^* = \underline{t}(w^{J*}, w^{K*}) \tag{3.15}$$

where $||w^J|| = 1$ and $||w^K|| = 1$. The regression coefficient q is obtained by regressing $\underline{y}$ on T as in Eq. (3.9), rewriting the approximation as above in Eq. (3.7) with $T = [t]$ and $P = [w]$, subsequently obtaining the decomposition as in Eq. (3.8). Similar to ordinary PLS the resulting score vector has maximal covariance with $\underline{y}$ and the loadings (w_j^J and w_k^K) lead to a least square solution. For $R > 1$ further components can be obtained as follows.

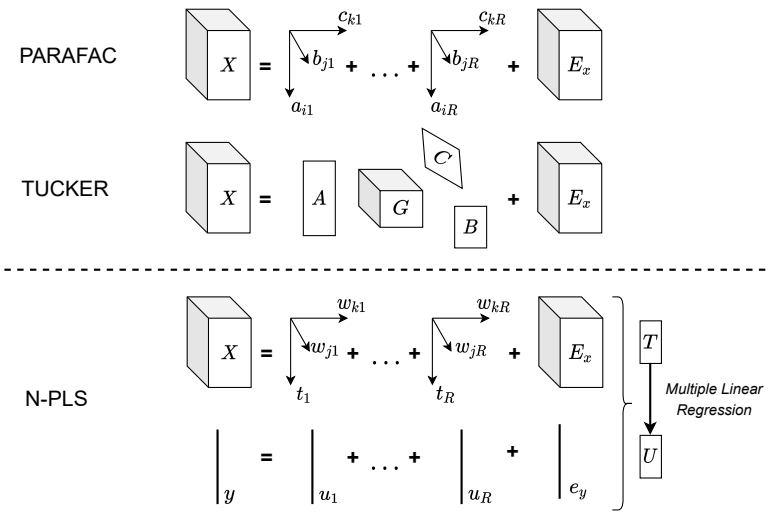


Figure 3.6: Schematization of multi-way models. The cubes (cuboids) represent three dimensional arrays (X denoting the data array, G the core-array for the Tucker model and E_x the residual array for all models), the three arrows represent orthogonal vectors (trilinear factors a_1, b_1, c_1, and loading vectors w_k, w_j with score vector t), while the lines represent vectors (response vector y, score vector u, and residual vector e_y), and the rectangles represent loading matrices for the Tucker model (A, B, C). Adopted from [185].

Rewrite Eq. (3.7) with $T = \left[\underline{t_1}, \ldots, \underline{t_R}\right]$, $P = \left[\underline{w_1}, \ldots, \underline{w_R}\right]$. Finally, decomposition of X as in Eq. (3.8), and subsequently from Eq. (3.2) the regression coefficient $\underline{q}$ is obtained as in Eq. (3.9).

In summary, the N-PLS model first extracts the important features from the predictor dataset into the loading array P, then estimates the regression coefficient vector $\underline{q}$ using least squares. For a more detailed description of the N-PLS algorithm the reader is referred to [36, 184, 35, 109, 34, 185].

COMPARISON OF MULTI-WAY METHODS

Atmospheric datasets are often multi-dimensional due to the fact that they contain several variables, which are not only varying over time but also over space. Moreover, often additional dimensions are present such as different climate projection scenarios, or model ensembles, which simulate the same information but use different assumptions or initial conditions. Three-way data that contain information on different variables, over time and space can be organized in a three-way array $\underline{\mathbf{X}} = X_{i,j,k}$. In our case the first dimension (mode 1 or index i) of the three-way array $\underline{\mathbf{X}}$ corresponds to time, the second dimension (mode 2 or index j) corresponds to different atmospheric variables, and the third dimension (mode 3 or index k) corresponds to location. Consequently, each frontal slice X_k represents a location with variables j sampled over time i.

The distinction between component and regression models should also be noted. The typical purpose of component models on one block of data is exploring the patterns and interrelations using latent variables (principal components), while regression mod-

els are aimed at predicting a block of data (response) using another block of data (predictors) through a prediction model. Consequently, component models require one block data, while regression models need multi block data. The above mentioned dimension reduction methods (PCA and PLS) are two-way component and regression models that cannot be directly applied to multi-way data. The traditional approach to deal with multi-way data is to use unfold methods (sliced analysis) such as the one introduced by [219]. Unfold methods first unfold the multi-way array to a two-way matrix and then perform ordinary PCA and PLS analysis. However, as Bro [36] has pointed out, the unfolding methods are not favourable since they do not make use of the multi-way structure in the data, they are often complex (using many parameters) and more difficult to interpret compared to the multi-way methods that do not use unfolding.

More appropriate models have been developed for handling multi-way data, which are the so-called multi-way component and regression models, schematized in Figure 3.6. Multi-way component models are basically generalizations of the two-way solutions to higher order arrays. One generalization of PCA to higher orders is Parallel factor analysis (PARAFAC), also known as trilinear decomposition, with general equation given by:

$$x_{ijk} = \sum_{r=1}^{R} a_{ir} b_{jr} c_{kr} + e_{ijk} \tag{3.16}$$

where R is the number of components used to fit the model; a_{ir}, b_{jr}, c_{kr} are 'triads' (trilinear factors) and e_{ijk} is the residual (see Figure 3.6). Note that here $R > 1$ is explicitly possible, compare PCA and PLS descriptions above. Another generalization is the Tucker decomposition, also called N-mode Principal Component Analysis [37]. For the three-way case, Smilde et al. [185] describe the Tucker3 model with the following equation:

$$x_{ijk} = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} a_{ip} b_{jq} c_{kr} g_{pqr} + e_{ijk} \tag{3.17}$$

where a_{ip}, b_{jq}, c_{kr} are elements of the loading matrices A, B, C; g_{pqr} is an element of the core-array G and e_{ijk} is the residual element in E, as depicted in Figure 3.6.

Similarly, the two-way partial least squares regression was also extended to multi-way data as described in Section 3.2.2. The N-way Partial Least Squares (N-PLS) method was developed by [36] and further elaborated by [184, 35, 109, 34]. A pictorial representation of N-PLS model is shown in Figure 3.6. Due to its desirable properties, as compared to the unfolding methods, the N-PLS method has been applied in a range of areas such as chemometrics, neuroscience and environmental analysis [38], food industry [67], organic pollutants in the environment [141] or most recently in agriculture [130].

Moreover, recently another generalized multilinear regression method, the Higher Order Partial Least Squares (HOPLS), was introduced by [225]. HOPLS differs substantially from N-PLS in that it uses the Tucker tensor decomposition (see Eq. (3.17)) instead of the trilinear decomposition (see Eq. (3.16)), hence, it benefits from the advantages of Tucker over PARAFAC. Zhao et al. [225] found that HOPLS could outperform N-PLS and PLS in case of small sample sizes and higher order ($N > 3$) response data ($\underline{y}$). While HOPLS appears to be a promising method in those cases, it should be noted that in this

study sufficient number of samples is available and the response $\underline{y}$ dataset is not high dimensional ($N \leq 3$).

The substantial differences between the above mentioned multi-way methods can be seen from their schematic representation (Figure 3.6). A comprehensive review of other dimension reduction methods for multidimensional data via Multilinear Subspace Learning (MSL) can be found in [135]. In this study the PCA, PLS, PARAFAC and Tucker algorithms were implemented using open source Python packages such as scikit-learn and TensoLy, whereas for the N-PLS algorithm the N-way Toolbox [12] was used in Matlab.

In order to showcase the differences between the various two-way (PCA, PLS) and multi-way (PARAFAC, TUCKER, N-PLS) dimension reduction methods, they were applied on the atmospheric and environmental data (from Section 4.2) for prediction. Their prediction errors were analysed from 10-fold cross-validation. K-fold cross-validation, briefly described in [95], uses a subset of the available data as a training set to fit the model and a different subset as a test set, where the full dataset is split into K equal-sized parts, in this case $K = 10$. For the prediction of every $k-th$ subset the model is fitted to the remaining $K-1$ subsets of the data and the prediction error of the fitted model is calculated. This process is repeated for $k = 1, 2, \ldots, K$ and the K estimates of prediction error are averaged. First the Mean Squared Error (MSE) with only the intercept (no principal components in regression) was calculated, and later on the MSE is computed using 10-fold cross-validation for the principal components, adding one component at the time in increasing order. The error measures of the unsupervised methods were obtained by extracting their computed model factors (with different number of components) which were then used to fit linear regression. The results of estimated mean squared errors of predicting y from 10-fold cross-validation are shown in Figure 3.7 (Section 3.3.1).

Apart from the prediction accuracy, it is also investigated how strongly each component (latent variable) in the two component N-PLS model (the best performing multi-way model) depends on the original variables (see Figure 3.8).

3.2.3. DYNAMIC FACTOR ANALYSIS

The previously presented dimension reduction techniques are able to identify unobserved factors that influence a substantial portion of the variation in a larger number of observed variables, and able to summarize the dataset through decomposition. None of these techniques, however, is designed for time series analysis as temporal correlation is neglected. Dynamic Factor Analysis (DFA) is a factor model that explicitly models the transition dynamics of the unobserved factors; hence, it is a dimension reduction technique that is designed for time series data. In fact, DFA is a multivariate time-series analysis technique that estimates underlying common trends in multivariate time series [93, 150, 138]. The time series are modelled using a linear combination of common trends, explanatory variables, and a noise component [226].

Given N time series, these could be analysed by univariate models by treating them as N separate univariate time series. However, this would result in N estimated trends without considering the interactions between them. DFA aims to overcome this disadvantage by reducing the N univariate trends to M common trends, where $1 \leq M < N$. The main objectives of DFA on environmental time series are therefore identifying un-

derlying common trends (unobserved factors) in the input time series, identifying interactions between the time series, and analysing the effects of explanatory variables.

The basic concept of DFA is to decompose the multivariate data into trends, explanatory variables and noise. Supposing that y_t is a univariate response variable measured in time t, where $t = 1, \ldots, T$, one of the simplest univariate time series models is given as follows:

$$y_t = \gamma \alpha_t + \epsilon_t \tag{3.18}$$

$$\alpha_t = \alpha_{t-1} + \eta_t \tag{3.19}$$

where α_t represents the factor (unknown trend) at time t, while ϵ_t and η_t are error components (noise). This model is called the random walk trend plus noise model. A formulation for the DFA with N time series (N rows) and M common trends (M columns) can be written as:

$$\begin{bmatrix} y_{1t} \\ \vdots \\ y_{Nt} \end{bmatrix} = \begin{bmatrix} \gamma_{11} & \cdots & \gamma_{1M} \\ \vdots & & \vdots \\ \gamma_{N1} & \cdots & \gamma_{NM} \end{bmatrix} \begin{bmatrix} \alpha_{1t} \\ \vdots \\ \alpha_{Mt} \end{bmatrix} + \epsilon_t \tag{3.20}$$

$$\begin{bmatrix} \alpha_{1t} \\ \vdots \\ \alpha_{Mt} \end{bmatrix} = \begin{bmatrix} \alpha_{1,t-1} \\ \vdots \\ \alpha_{M,t-1} \end{bmatrix} + \begin{bmatrix} \eta_{1t} \\ \vdots \\ \eta_{Mt} \end{bmatrix} \tag{3.21}$$

or in generic form:

$$y_t = \Gamma \alpha_t + \epsilon_t \tag{3.22}$$

$$\alpha_t = \alpha_{t-1} + \eta_t \tag{3.23}$$

where Γ is a factor loading matrix with dimension $N \times M$ and contains the unknown factor loadings, which are multiplication factors that determine the linear combination of the original variables; and α_t is a vector of the M common trends at time t with dimension $M \times 1$. It is generally assumed that the error terms are independent, normally distributed with mean 0 and an unknown diagonal or symmetric/non-diagonal covariance matrix: $\epsilon_t \sim N(0, H)$, $\eta \sim N(0, Q)$ and $\alpha_0 \sim N(\alpha_0, V_0)$ where H, Q, V_0 are covariance matrices [227]. Based on these parameters the covariance matrix of y_t can be written as:

$$\text{cov}(y_t) = \Gamma \, \text{var}(\alpha_t) \Gamma' + H. \tag{3.24}$$

In order to include K explanatory variables in the DFA, equations (3.22)–(3.23) can be extended to the following model:

$$y_t = \Gamma \alpha_t + D x_t + \epsilon_t \tag{3.25}$$

$$\alpha_t = \alpha_{t-1} + \eta_t \tag{3.26}$$

where D is an $N \times K$ matrix containing the partial (standardized) regression coefficients, and x_t is a $K \times 1$ vector containing the values of the K explanatory variables at time t. The effects of explanatory variables are modelled as in linear regression, and therefore it depends on the same underlying assumptions, such as normality, independence, and homogeneity of residuals [226].

Equations (3.25)-(3.26) can be cast into state space form, and the unknown trends can be estimated via the Kalman filter. The likelihood is then evaluated based on the filtering recursions, and maximum likelihood estimation is used to estimate the parameters. The Kalman Filter and smoother algorithm for the model in equations (3.25)-(3.26) can be found in [227].

The dynamic factor model was applied to the ten chlorophyll-a time series. The main objective was to identify underlying common trends and further analyse the effects of atmospheric variables on chlorophyll-a concentrations, this time considering temporal correlation. Since standard dynamic factor models are not designed for multi-way data, such as N-PLS, the atmospheric data is averaged over the locations. In order to verify that the underlying assumptions of the dynamic factor model are not violated, several tests were conducted. These tests include plotting the standardized residuals over time, checking the normality of the residuals and plotting the correlogram (see Figure S4 in the Supplementary Material). It was verified that residuals are uncorrelated (since the autocorrelations are near zero of all time-lag separations), and normally distributed with mean zero. Thus, underlying assumptions are valid.

3.2.4. FUNCTIONAL PCA

So far we have investigated the features and relationships between short term (1 year long) meteorological data and environmental response. These datasets offered us the opportunity to apply supervised techniques since the environmental response was computed with the meteorological data as input. Moreover, we could consider temporal dependence and compute unobserved factors in the time series due to the reasonable number of time steps that allow us to apply computationally intensive state space models. However, apart from the analysis on short term data, we are also interested in investigating the features of the long term (climate scale) atmospheric projections and potential for data reduction. In order to achieve this, firstly we use Euro-CORDEX climate projections (covering the entire 21^{st} century) instead of numerical weather prediction model outputs. Secondly, we analyse the discretely computed (in time) atmospheric data in the functional data space. This allows us to apply functional data analysis and study functional variation, which is more logical for climate projections that are long time series of modelled variables and are not meant to study short term changes and daily variability. Naturally, an interesting feature of the climate projections is their long term trends. Conclusions on their seasonal variability and the similarities between climate scenarios are less often drawn, however. We aim to reach such conclusions through Functional Data Analysis. By treating these long term climate projections as functional data our objective is to find an underlying function that can characterize the general shape of the time series, explain their variability (functional variation), reduce data complexity, and to aid the interpretation of the underlying variability sources [164]. The findings of the previous analyses and the Functional Data Analysis can be jointly used for

climate impact assessment by aiding the atmospheric variables selection for studying chlorophyll-a climate response, as well as the identification of important features of the climate projections for further statistical models.

Functional data representation is commonly done by smoothing the discrete-time data with basis expansion (e.g. constant, polynomial, polygonal, B-splines, power, exponential, Fourier) as a pre-processing step. In our study, a Fourier basis expansion is applied, which has good computational properties especially when the data points are equally spaced. Moreover, Fourier bases are natural for describing periodic data, such as atmospheric variables, and therefore it is commonly used in this domain. The functional basis components can be then estimated through Functional Principal Component Analysis (FPCA).

The underlying idea is that a function $x_i(t)$ can be expressed as a basis expansion:

$$x_i(t) = \bar{x}(t) + \sum_{j=1}^{\infty} f_{ij}\varphi_j(t) \tag{3.27}$$

and

$$f_{ij} = \int \varphi_j(t)\,[x_i(t) - \bar{x}(t)]\,dt \tag{3.28}$$

where $\bar{x}$(t) is the functional mean (zero if the data is mean centered), $\varphi_j(t)$ are the orthonormal eigenfunctions and f_{ij} are the Functional Principal Component Scores. The first few eigenfunctions and eigenvalues can be used for data reduction and feature extraction, while the Functional Principal Component Scores can be used to describe, cluster and classify the curves [180]. The Functional Principal component analysis in this research uses an open source Matlab toolbox [165].

While the other above mentioned methods (multi-way methods and dynamic factor model) are used for identifying the most important atmospheric variables affecting chlorophyll-a concentrations in a shorter time interval, in this research Functional Principal Component Analysis is used to investigate different features of the long term climate projections spanning the 21^{st} century (from 2006 to 2100). Functional Principal Component Analysis was therefore applied to the Euro-CORDEX climate projections to compare the functional variation of climate variables, and to describe, cluster and classify the climate scenarios for the two most important variables (radiation and temperature).

The discrete-time data points are first transformed to functional data using a Fourier basis expansion. The left panel of Figure 3.11 shows the atmospheric variables as functional data for an arbitrarily selected year within the 95 year interval. The well distinguishable sinusoidal shapes of solar radiation and temperature can be seen in the figure. Functional Principal Component Analysis with two principal components is then performed on the functional data and the scores of the first two components are plotted to analyse similarities between the variables (right panel of Figure 3.11). Moreover, as a second experiment, using Functional Principal Component Analysis the aim is to classify and cluster the climate scenarios (Representative Concentration Pathways and driving General Circulation Models) for the two important climate variables (radiation and temperature), see Figure 3.12 and 3.13.

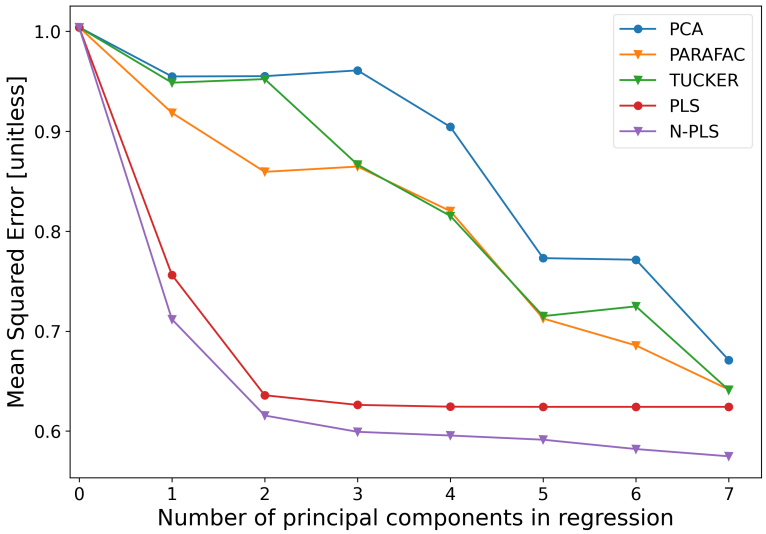


Figure 3.7: Comparison of two-way (point markers) and multi-way (triangle markers), unsupervised (PCA, PARAFAC, TUCKER) and supervised (PLS, N-PLS) dimension reduction models. Prediction errors (MSE) from 10-fold cross-validation with increasing number of components. Prediction is done at Marsdiep Noord station.

3.3. RESULTS

3.3.1. COMPARING TWO-WAY AND MULTI-WAY METHODS

First the results obtained from the comparison between the various two-way (PCA, PLS) and multi-way (PARAFAC, TUCKER, N-PLS) dimension reduction methods are presented. In Figure 3.7 one can observe that unsupervised (PCR, PARAFAC, Tucker) and supervised (PLS, N-PLS) methods form two distinct clusters where the unsupervised techniques have higher mean squared errors if not all components are included in the regression ($n < 7$). The cross-validation results also show that supervised techniques perform the same with only 2 components as unsupervised ones with 7 components, indicating that there is room for dimensionality reduction. PLS and N-PLS models are therefore more parsimonious and explain the majority of the variance in the data with a lower number of predictor variables, due to the fact that they include correlation to the response. Another observation from Figure 3.7 is that, for this dataset, multi-way methods outperformed their two-way counterparts for both supervised and unsupervised groups by including the third dimension (spatial correlation). When modeling the data by a two-way model the assumption is made that the latent phenomenon (atmospheric or environmental processes) at a certain location is completely independent of the phenomenon at another location. This is not correct in most cases as usually the phenomena will be similar at different locations. For multi-way models the assumption is that the phenomenon describing the variation at one location is the same as the phenomenon at another location, although its magnitude is different. This assumption should be closer to the truth than neglecting or "unfolding" the spatial dimension [35].

Another observation is the performance difference between PCA, PARAFAC and Tucker

models. Both PARAFAC and TUCKER are generalizations of PCA to a higher order, with the important difference that the PARAFAC model has the attractive feature of providing unique solutions (there is no problem with rotational freedom). If the data are approximately trilinear, the true underlying phenomena can be found if the right number of components is used and the signal-to-noise ratio is appropriate [35]. The Tucker model is, however, more flexible and has rotational freedom. It is not structurally unique as PARAFAC. This makes the Tucker model complex and might explain why it has lower performane for this specific example. A restricted Tucker model version exists where domain knowledge is used to restrict the core elements, forcing individual elements to take specific values. This way it is possible to define models that uniquely estimate certain properties. This could be seen as a structural model tailored to a specific problem. In this research restricted Tucker models were not used.

In Figure 3.8 the loadings of the first two components of the N-PLS model (the best performing multi-way model) are given for two different locations. By identifying the original predictor variables that weight most heavily one can draw conclusions on the underlying physical processes. Moreover, less important predictors could be excluded from the dataset in order to reduce the number of variables. In Figure 3.8 it can be observed that at Marsdiep Noord, a location of a deeper tidal inlet, the highest loadings are given to radiation in the first component and to temperature in the second component. On the other hand, at Dantziggat, located in the shallow inter-tidal area, the opposite can be observed: the highest loadings are given to temperature in the first component and to radiation in the second component. Moreover, apart from temperature and radiation which have the highest loadings, northward-wind also has high loading in the second component at Datziggat. The factor loadings indicate the differences in the physical systems between the two locations. In deeper areas (Marsdiep Noord) solar radiation is the primary driver of the onset of phytoplankton blooms, while in shallower areas (Dantzigat) radiation intensity is slightly less limiting and light availability in the water column heavily depend on wind, which influences turbidity due to the mixing of layers and suspension. This could explain the greater importance of wind speed at Dantziggat, especially that northerly winds cause the highest surges of sea water along the Dutch coast [115] that leads to enhanced mixing. In addition, thermal stratification and vertical mixing conditions are different at the two locations, Marsdiep Noord being intermittently stratified and Dantziggat being permanently mixed [122]. This influences nutrient availability in the mixed layer depth as well as phytoplankton composition and therefore could be responsible for the greater importance of air temperature at Dantziggat. Moreover, top-down phytoplankton governing factors (e.g. grazing, filter-feeding) are also different at the two locations. For instance the density of filter-feeders is much higher near Dantziggat [71].

3.3.2. DYNAMIC FACTORS

Choosing the optimal number of unobserved factors is crucial to find a model that identifies common trends in the dataset without significant loss of statistical information. In order to find the optimal number of factors, the Akaike's Information Criterion (AIC) for each model setup (different number of factors, error covariance matrix diagonal or unstructured) was calculated and the model containing the lowest AIC value was selected

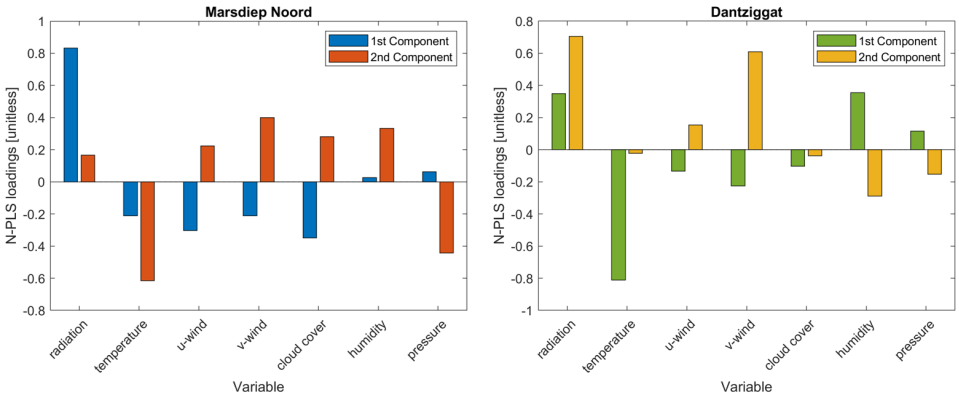


Figure 3.8: Importance of predictor variables in the N-PLS model. Loadings of atmospheric variables in the 1st and 2nd N-PLS components for locations Mardiep Noord (left) and Dantziggat (right). Eastern and northern wind components are denoted by u-wind and v-wind, respectively.

as optimal. The selected model contains two factors if the error covariance matrix is set to diagonal. The identified two unobserved factors can be seen in the middle panel of Figure 3.9. It should be noted that the second factor has negative factor loadings, and for demonstration purposes, it was plotted with negative sign. The results indicate two well distinguished trends. The first factor represents the trends of those locations where the chlorophyll-a concentration peak (spring bloom) occurs earlier, such as Dantziggat. This is confirmed by the factor loadings (Γ). On the other hand, the second factor shows the pattern of the locations where the occurrence of the peak is delayed. The identified temporal shift between locations in the onset of the spring bloom can be explained by the different system dynamics of the areas, for instance shallower intertidal zones and the proximity from river or tidal inlets.

The partial standardized regression coefficients of the dynamic factor model for the two representative stations Marsdiep Noord and Dantziggat (see Figure 3.10) are in agreement with the findings of N-PLS loadings and confirm that radiation and temperature are the most important atmospheric variables. It is also confirmed that at Dantziggat air temperature has significantly larger impact than at Marsdiep Noord. As mentioned above, this might be related to the differences in thermal stratification, mixing conditions and trophic interactions between the two locations. Nevertheless, considering temporal correlation the relative impact of solar radiation (compared to the other variables) seems to be even more important, especially at station Marsdiep Noord. This finding could be explained by the fact that phytoplankton biomass onset in this coastal ecosystem highly depends on the timing of increased energy from solar radiation during spring [188]. In fact, it was reported by [188] that the (external) light regime appears to play a more important role in the initiation of spring blooms than temperature.

3.3.3. FUNCTIONAL PRINCIPAL COMPONENTS

An important aspect of FPCA is the examination of the scores of each curve (variable) on each component (here we display the first two). Figure 3.11 (right panel) shows the

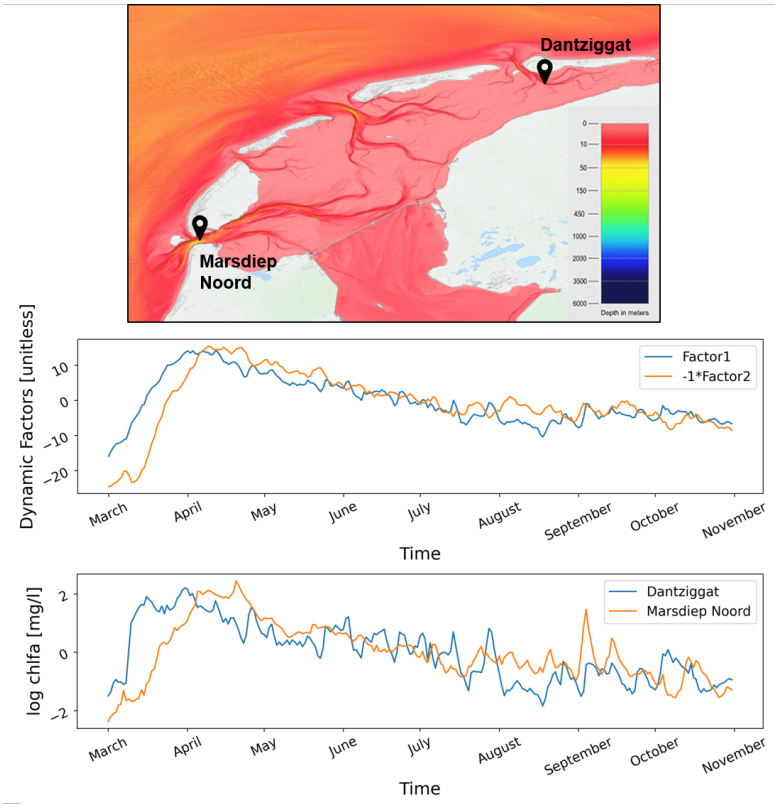


Figure 3.9: Underlying trends in chlorophyll-a time series. First and second factors of the Dynamic Factor model simulating chlorophyll-a trends with atmospheric variables as exogenous variables (middle). Log chlorophyll-a time series at representative stations (bottom). Bathymetry map showing the location of the representative stations (top). Map source is https://portal.emodnet-bathymetry.eu/.

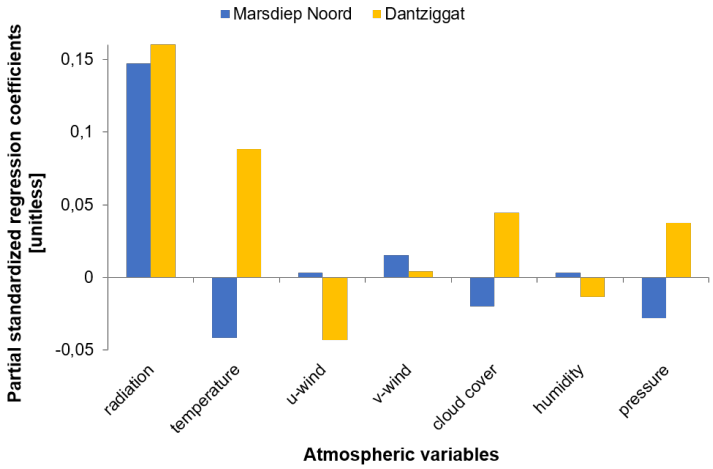


Figure 3.10: Dynamic Factor model partial standardized regression coefficients for Marsdiep Noord and Dantziggat stations demonstrating the effects of atmospheric variables as exogenous variables to model chlorophyll-a trends.

scores of the first two components of the Functional Principal Component Analysis applied to Euro-CORDEX climate variables. In order to draw conclusions from this figure, one must take into consideration the inverse correlation between two group of variables: cloud cover and northerly wind on one hand and radiation, temperature, and pressure on the other hand. These variables have relatively similar FPCA function scores but the scores of second group have negative signs (expressing the inverse correlation). Known examples are the anticorrelation of atmospheric pressure and cloud cover (high pressure meaning lower cloudiness), or cloud cover and solar radiation (high cloudiness meaning lower surface downward solar radiation).

After accounting for the sign of the FPCA scores, a single main cluster can be distinguished that group variables (their functional representations) with similar characteristics and two variables, eastern wind and humidity, that are relatively separated. In general the correlation between cloud cover and wind speed is documented [66] but the reason for eastern wind to be separated could be explained by the fact that at this specific location the maritime air mass is mainly brought by the northerly wind from the North Sea to replace the dry continental air mass [115] causing cloud formation. The fact that radiation and temperature are positively correlated and lie near each other is expected, due to their similar sinusoidal functional shapes. The relationship of variations in air temperature to changes in air pressure was also reported in literature [3] based on the analysis of long historical records. They concluded that changes in atmospheric circulation (influenced by air pressure) has a key role in air temperature variation, acknowledging that the relationship is seasonally dependent and impacted by the regional topography. Indirect links between air pressure and solar radiation were also discovered by [115]. They argue that high pressure systems impact air quality, which in turn affects solar irradiance [224, 222, 82]. However, as the considered data are outputs of a climate

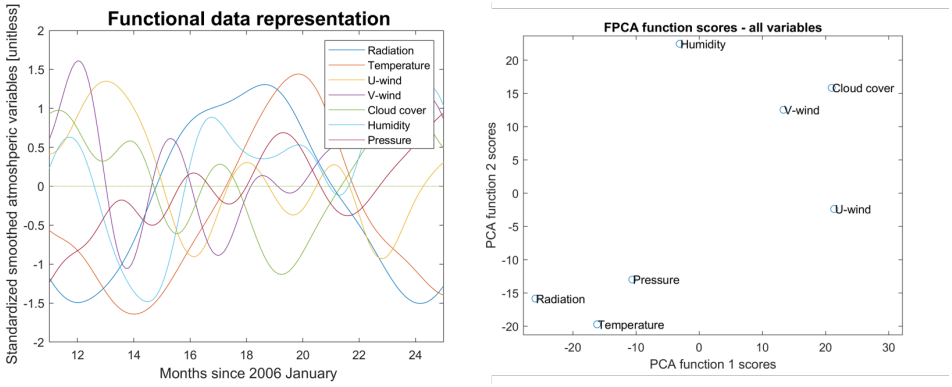


Figure 3.11: Functional Principal Component scores for the Euro-CORDEX climate variables. Atmospheric variables transformed to functional data (for an arbitrarily selected year) on the left panel, and clustering of variables based on the FPCA function scores on the right panel.

model, which does not include air quality processes, this could not have been captured in our dataset.

Considering the Functional Principal Component Analysis results for the removal of multi-collinearity, one could expect that using only radiation or temperature might be sufficient without significant loss of statistical information. For climate impact studies at this location it should be considered, however, that solar radiation and temperature display different long term trends in this region and influence the phytoplankton dynamics differently. Based on the Euro-CORDEX projections, long term trends of radiation is constant or slightly decreasing, whereas air temperature trends are increasing.

Figure 3.12 depicts the functional representation of the eight climate scenarios for solar radiation and the first FPCA function with ± two standard deviation. Most of the variability (77%) can be explained by the first FPCA function, which suggest that the scenarios are largely similar. Nevertheless, varying amplitudes and time shifts are observed between scenarios. These deviations from the mean function are depicted in the lower left panel of Figure 3.12. Furthermore, when comparing the component scores it can be clearly identified that the climate scenarios are clustered based on the driving GCMs, and the two RCP scenarios (RCP4.5 and RCP 8.5) per driving GCM have similar characteristics. This is in line with previous finding that uncertainty in Regional Climate Model projections are primarily influenced by the driving GCMs while the impact of RCPs is less dominant [151]. The results also suggest that the CNRM and ICHEC driving GCMs are very similar to each other, whereas the IPSL driving GCM is divergent from the other driving GCMs. This was also reported by [148] based on an in-dpeth analysis of the characteristics of Euro-CORDEX climate projections.

The same exercise was performed for air temperature and the results are shown in Figure 3.13. Similarly to the radiation scenarios, the air temperature scenarios also differ in their amplitudes and seasonality (temporal shift). The uncertainty around the mean function (first FPCA component) clearly illustrates this phenomenon. In this case, the variability explained by the first FPCA function is smaller (57%), indicating that temper-

ature scenarios are less similar, perhaps due to the long term trends (moderately increasing for RCP4.5 but more sharply increasing for RCP8.5). Surprisingly, the FPCA component scores show a different picture from the results of the radiation variable. While the IPSL driving GCM is still farthest from the others and ICHEC RCP4.5 and CNRM RCP8.5 remain similar, the other scenarios are not clustered by driving GCMs anymore.

These findings indicate that the time series of Euro-CORDEX climate scenarios (for both solar radiation and air temperature) show structural differences across driving GCMs but full independence between the scenarios cannot be assumed as their functional features are similar. In fact, they can be described with a mean function and varying amplitude plus phase shift. This feature should be incorporated in any statistical model that is aimed at generating new representative climate scenarios similar to the existing Euro-CORDEX projection scenarios. While the results of the Functional Principle Component Analysis do not allow us to draw conclusions about shifting seasonality of radiation scenarios on the long-term, but it does express the strength of the mean signal (77% and 57% variance explained by the first FPCA for radiation and temperature respectively) and highlights the source of the variability around the mean signal.

In a related study [148] a deeper analysis of the same Euro-CORDEX climate dataset has been performed that reached conclusions on the long term characteristics. In this analysis the radiation projections have been modelled by a structural time series model that has various components accounting for long term trend, seasonal shape with varying amplitude and time shift, and an additive residual term. The parameters of these time series model components have been estimated through Bayesian parameter inference based on the eight Euro-CORDEX climate projection scenarios over the 21^{st} century. The seasonal shift was represented by the deviations in the (yearly) seasonal cycle lengths. It was observed that the deviations are centered around zero (deviations were maximum around 14 days) and have a negative lag 1 autocorrelation meaning that most positive deviations tend to be followed by negative deviations and vice versa. In this way the yearly cycle lengths remain close to the ideal cycle length (one calendar year) throughout the entire time series. Therefore, no consistent shift in seasonality was identified. Regarding the trend slope, the general expectation that RCP8.5 has steeper slope than RCP4.5 was confirmed for the temperature variable and also for solar radiation but much less pronounced. Finally, regarding the amplitude of the seasonal shape, deviations of up to around 20% were observed but without consistent trend.

3.4. DISCUSSION

It must be emphasized once again that all statistical techniques applied in this study are well documented in the literature. Consequently, the added value of our research to the marine scientific community is not the development of novel techniques but the application of carefully selected dimension reduction techniques (originating from various domains) to marine and climate big data, in order to provide statistical underpinning for climate variable selection and data reduction to support subsequent ecological impact studies. In addition, our study also offers a framework for the structured application of these dimension reduction techniques to specifically cover three features in marine and climate datasets: (1) spatial correlation, (2) temporal correlation, and (3) functional variability. The chapter therefore offers a “dimension reduction tool kit” that goes beyond

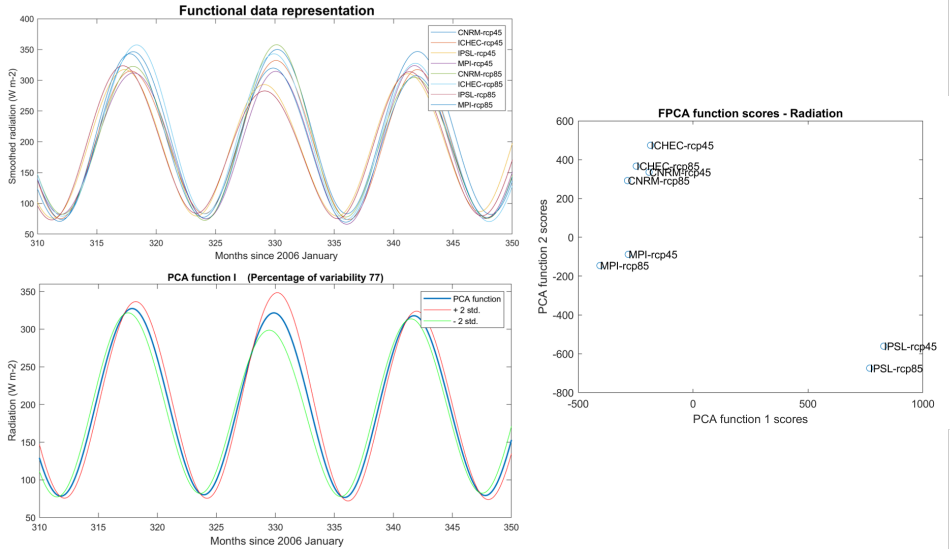


Figure 3.12: Functional Principal Component Analysis for the Euro-CORDEX solar radiation scenarios. Eight solar radiation scenarios transformed to functional data on the upper left panel, the first PCA component with ± two standard deviations on the bottom left panel, and clustering of scenarios based on the PCA function scores on the right panel.

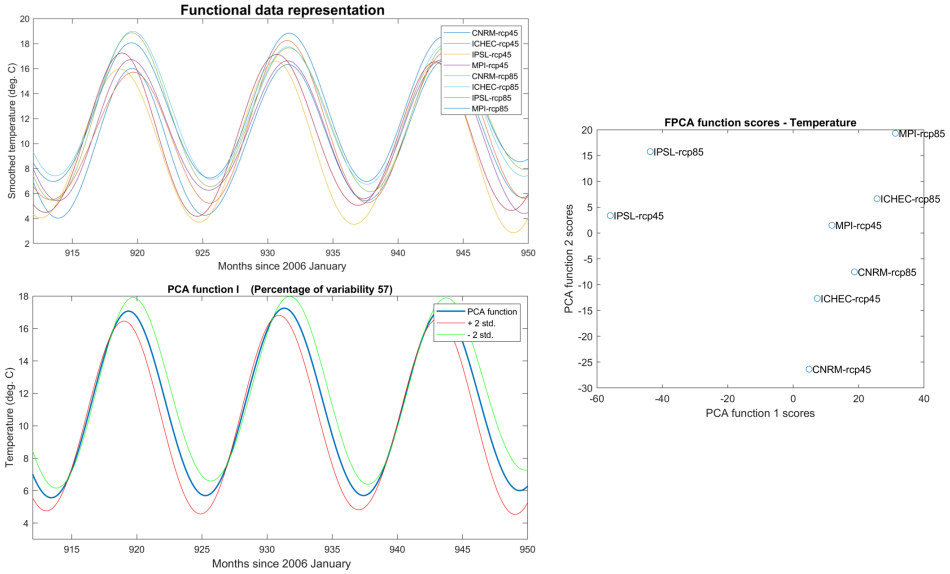


Figure 3.13: Functional Principal Component Analysis for Euro-CORDEX air temperature scenarios. Eight air temperature scenarios transformed to functional data on the upper left panel, the first PCA component with ± two standard deviations on the bottom left panel, and clustering of scenarios based on the PCA function scores on the right panel.

the standard practice and is suitable to jointly study marine and climate datasets.

For instance, N-PLS was developed in the domain of chemometrics, and while several applications in other domains were reported [24, 38, 141, 67, 130], it has not been applied in coastal ecological impact studies, to the best of the author's knowledge. An N-PLS application particularly relevant to our research is the study of [24] in the field of applied climatology that used N-PLS as an empirical downscaling tool for predicting climate variables. That study employed N-PLS regression using average monthly near-ground air temperature, specific humidity and sea-level pressures from Global Climate Models as predictors for downscaled average monthly air temperature, dew temperature, and precipitation. The results of the N-PLS regression were then compared to the ones form Principal Component Regression (PCR). It was concluded that in general N-PLS regression outperforms the commonly used PCR, and therefore presents a promising alternative. While that study presented comparison to PCR, our study extends the comparison of the N-PLS results to a range two-way and multi-way methods. Moreover, the application of N-PLS is also extended by including ecological response apart from the climate data. This provides further evidence on the benefits of N-PLS in the fields of marine and climate sciences.

As opposed to N-PLS, Dynamic Factor Analysis has been more widely used in environmental studies [75, 47, 121], including marine ecosystem studies [227, 226, 173], also considering the impact of climate change [119] to identify general patterns in multivariate time series, interactions between the time series, and the correlation between the time series and explanatory variables. Nevertheless, our application can complement the studies of [227, 226] that focused on macro zoobenthos and fisheries, as our study describes phytoplankton biomass (via chlorophyll-a as proxy), which has different role in the marine food web.

Our study also advances scientific knowledge related to the analysis (and data reduction) of climate scenarios. In the past, PCA has been applied to various climate multi-model ensembles to reduce the larger ensemble sizes into smaller subsets. [176] applied PCA to define a measure of similarity between models in the Coupled Model Intercomparison Project (CMIP5) [195]. [146] also used PCA to find common climate change patterns within a multi-model ensemble (ENSEMBLES regional multi-model ensemble), combined with cluster analysis detecting model similarities. Furthermore, [55] presented a methodology using PCA for reducing the climate projection ensemble size of EURO-CORDEX for subsequent impact studies. There are important differences between our research and these existing studies, however. Firstly, the motivation for those studies to use PCA was to select a subset of scenarios from a larger ensemble while keeping the characteristics representative, whereas our goal is more than just the clustering of climate scenarios. Our study identified features of the radiation and temperature functions such as the sources of variability (e.g. time lag and amplitude shift). These identified properties allow us to construct synthetic realizations of the climate projection scenarios in subsequent studies (using climate generators). Thus, the objectives are in sharp contrast, the former aiming to support scenario studies (based on a reduced number of representative ensemble members) and the latter supporting probabilistic studies (based on numerous synthetic realizations). Secondly, all of these studies used ordinary PCA, not Functional Principal Component analysis. By considering climate data to be

functional data, although computed at discrete time intervals, Functional Data Analysis allowed us to represent the entire measured function on a continuum interval. This paradigm shift from discrete-time to functional data presents an alternative approach to the conventional statistical methods, since it provides additional information on the underlying functions. Of course, Functional Data Analysis itself is also not new and has been previously applied in various fields, such as hydrology [193, 44, 192, 6, 91], climatology [32, 191], water quality [98, 83], and others [203]. [191] already documented the benefits of using Functional Data Analysis to study temporal features of climate data, although in that study Functional PCA was applied to historical data, namely the El Niño Southern Oscillation. In our research Functional PCA is applied to an ensemble of future climate projections.

3.5. CONCLUSIONS

In this chapter a variety of statistical methods for the multivariate analysis of air-sea interactions are applied in order to aid the understanding of complex multi-dimensional datasets and to support ecological impact studies. The selected dimension reduction methods were chosen to account for spatial correlation, temporal correlation, and functional variability. The presented methods were found to be useful in exploring the datasets, identifying latent processes, removing multi-collinearity and selecting atmospheric variables that are the most important when predicting chlorophyll-a response. A comparison of standard two-way (PCA, PLS) and less frequently used multi-way methods (PARAFAC, Tucker, N-PLS) showcased the potential of multi-way methods to construct parsimonious data reduction models. The results allow us to conclude that there is room for dimension reduction in the atmospheric dataset since in most cases low prediction errors could be achieved with as few as 2 principal components. Further conclusions could be drawn on the predictors that affect the coastal chlorophyll-a concentration the most. All used methods indicate solar radiation to be the most important influencing factor, followed by air temperature and wind in shallow zones. The dynamic factor model proved to be an appropriate tool to acquire information about underlying common trends in chlorophyll-a time series across stations, and to investigate the effects of atmospheric explanatory variables with the inclusion of temporal structure when constructing unobserved factors. The difference in phytoplankton bloom onset at different parts of the Dutch Wadden Sea was revealed by the dynamic factor model and solar radiation was re-confirmed to be the most dominant atmospheric variable when temporal correlation is considered. Finally, using Functional Principal Component analysis further insights into the Euro-CORDEX regional climate data were gained by identifying features of the climate projection scenarios.

Overall, our findings support the use of solar radiation as the primary driving atmospheric variable to simulate climate impacts on coastal chlorophyll-a concentrations in the Dutch Wadden Sea. Moreover, structural patterns of Euro-CORDEX climate scenarios for solar radiation and air temperature have been determined, which provide information on the mean functions and their uncertainties. In ecological impact studies, uncertainties stemming from the climate scenarios are often only represented by picking few climate ensemble members (some of the driving GCMs and RCPs). Instead of such scenario studies it is advised to use the presented uncertainty intervals in the func-

tional variation of the Euro-CORDEX climate scenarios and perform a fully probabilistic assessment for proper climate uncertainty propagation. In this context, the findings can also inform studies in which climate generators are proposed to produce numerous synthetic realizations of solar radiation and air temperature projections. The underlying structural time series models of such climate generators should incorporate the two identified features: varying amplitudes and time lag (shift) in seasonality. Moreover, due to the identified shared characteristics, climate scenarios seem exchangeable rather than independent, hence, the pooling of scenarios is recommended in hierarchical models to borrow strength and make statistical models more optimal.

4

BAYESIAN STOCHASTIC CLIMATE GENERATOR

Available climate change projections, which can be used for quantifying future changes in marine and coastal ecosystems, usually consist of a few scenarios. Studies addressing ecological impacts of climate change often make use of a low- (RCP2.6), moderate- (RCP4.5) or high climate scenario (RCP8.5), without taking into account further uncertainties in these scenarios. In this chapter a methodology is proposed to generate further synthetic scenarios, based on existing datasets, for a better representation of climate change induced uncertainties. The methodology builds on Regional Climate Model scenarios provided by the EURO-CORDEX experiment.

In order to generate new realizations of climate variables, such as radiation or temperature, a hierarchical Bayesian model is developed. In addition, a parameterized time series model is introduced, which includes a linear trend component, a seasonal shape with varying amplitude and time shift, and an additive residual term. The seasonal shape is derived with the non-parametric Locally Weighted Scatterplot Smoothing (LOWESS), and the residual term includes the smoothed variance of residuals and independent and identically distributed noise. The distributions of the time series model parameters are estimated through Bayesian parameter inference with Markov chain Monte Carlo sampling (Gibbs sampler). By sampling from the predictive distribution numerous new statistically representative synthetic scenarios can be generated including uncertainty estimates.

As a demonstration case, utilizing these generated synthetic scenarios and a physically based ecological model (Delft3D-WAQ) that relates climate variables to ecosystem variables, a probabilistic simulation is conducted to further propagate the climate change induced uncertainties to marine and coastal ecosystem indicators.

Parts of this chapter have been published in Stochastic Environmental Research and Risk Assessment (2021) [148]

4.1. INTRODUCTION

It is widely accepted that long term changes in climatic variables will cause shifts (phenological and biogeographic shifts) in species distributions, but the extent of these shifts is not yet well understood and any prediction will have a high level of associated uncertainty [80]. Climate change data in ecosystem assessments are used as forcing conditions for the numerous non-linear ecological processes. These ecological processes are influenced by changes in extreme values, or shifts in distributions and peaks of the climate forcings. Applicable methodologies for estimating ranges and expected changes in statistical properties of the climate scenarios are therefore essential for subsequent ecological impact assessment.

The uncertainty accumulated throughout the climate modelling chain, such as initial conditions, boundary conditions, parametric and model structure of both General Circulation Models (GCMs) and Regional Climate Models (RCMs) may further propagate and influence ecological impact estimates. Yet in most impact studies climate change induced uncertainty is only characterized by different GCM and Representative Concentration Pathway (RCP) configurations in a small ensemble of climate scenarios to anticipate potential trajectories [211], but without a fully probabilistic uncertainty quantification.

If available time series of the climate variables are not sufficient to serve as stochastic input variables for ecological, agro-meteorological or hydro-meteorological assessment studies, one way to obtain better uncertainty estimates is to generate multiple realizations of the climate input variables. Numerous studies exist on generating new datasets of meteorological variables using probabilistic models. These models are often referred to as stochastic weather generators. Some well known examples of stochastic weather generators are LARS-WG [181], WeaGETS [45], or CLIMAK [56]. These widely used stochastic weather generators have been compared in various studies to assess their validity for long-term climate data simulation [144], performance in different climatic regions [213], adequacy for water resources systems risk assessment [8], or to quantify uncertainty due to the choice of the weather generator [211]. In short the aim of all stochastic weather generators is to simulate new synthetic sets of meteorological time series with statistical properties similar to the historical data or models [27]. The expected impact of such methods on the relevant scientific community is to facilitate studying the effect of long-term changes in mean climate variables, climatic variability, and the frequency of extreme events [211].

In the above mentioned weather generators the primary variable of interest is precipitation and the simulation of other variables, such as temperature and solar radiation, is conditioned on the occurrence of rainfall (wet or dry days). Thus, most of these stochastic weather generators are of Richardson type [170]. The concept of these types of generators is that solar radiation and temperature are modeled jointly as a bivariate stochastic process with the daily means and standard deviations conditioned on the wet or dry state. First a 'residual' time series is obtained by removing a periodic trend. This residual time series is assumed stationary and normally distributed, and the autocorrelation and cross-correlation coefficients are estimated using the residuals of the maximum temperature, minimum temperature, and solar radiation variables. Finally, the removed means and standard deviations are reintroduced to produce the generated daily values.

Recently, for the simulation of temperature improvements have been made to the Richardson type weather generators. One of the major improvements is simulating non-stationary temperature time series directly instead of simulating standardized residuals first and then adding them to the periodic mean and standard deviation [186]. The proposed approach called Stochastic Harmonic Autoregressive Parametric weather generator (SHArP) allows for trends and seasonality in the temperature generation. Another extension, the Seasonal Functional Heteroscedastic Autoregressive (SFHAR) generator [53], uses a decomposition of the temperature signal into trends and seasonality in the mean and the standard deviation, and a stochastic part. This was later applied to generate a long trajectory of past and near future (up to 2040) temperature by also incorporating GCM simulations [158]. This is an innovative feature considering that the commonly used weather generators focus on historical periods with observed climate characteristics and allow the inclusion of future climate projections only through change factors. Those change factors are then used to alter the observed statistics to account for the offset in the future projections and recalibrate the weather generators.

The lack of proper treatment of parameter uncertainty in previous weather generators gave rise to studies which employed Bayesian methods. These methods have a clear advantage as they better capture uncertainty by providing the full distribution of model parameters instead of a single best estimate. This enhanced parameter uncertainty characterization allows us to represent the full range of plausible climate scenarios and subsequently the full range of impacts, once climate input is propagated through process-based models [210]. For these reasons, hierarchical Bayesian frameworks have been increasingly applied for a range of purposes in the field of weather generators. Applications have primarily focused on precipitation and temperature modelling, such as spatial modelling of extreme precipitation [168], spatial modelling of daily precipitation and temperature [210], statistical downscaling of precipitation [94], or to quantify future temperature and precipitation uncertainties from multiple climate models [198, 197, 152, 112]. Even though all these applications benefited from the hierarchical or multi-layered Bayesian model structure, their exact model formulations are not transferable to our case, in which the parameters of our proposed time series model are to be inferred in order to simulate long term traces of radiation, and making use of an ensemble of RCM simulations.

While our proposed method shares the primary objective of existing stochastic weather generators, in that we also aim to generate numerous gap-free time series of atmospheric variables using available climate data and with statistical characteristics similar to these, there are few important differences in the main concept. Firstly, we aim to directly simulate trajectories with long-term trend, avoiding the common practice of simulating residuals which are then added to climatology (historical or climate change adjusted). Moreover, simulating a very long future projection until the end of the century is not common in existing studies. Secondly, we use a high-resolution 0.11 degree (or 12.5 km) RCM ensemble from Euro-CORDEX as calibration data for our generator to quantify the temporal evolution of future uncertainty in regional climate change radiation projections, as opposed to most previous studies using GCMs and focusing on precipitation and temperature. Since we propose a single site generator, we do not make full use of the high spatial resolution of the RCM ensemble, on the other hand, we can argue that high-

resolution RCMs describe regional and local processes more accurately than GCMs. In this regard, the novelty is not that the generator can create spatial fields, but rather that it is using input data from a climate modelling experiment that describes local processes the best, which was often not the case in existing stochastic climate generators. Thirdly, our hierarchical Bayesian model consist of a new time series model formulation and derived Gibbs parameter update formulas for the parameter inference. The proposed multi-layered Bayesian structure combines different climate scenarios into one model (rather than separately treating them), making the estimates statistically more robust. Lastly, we apply the generator to simulate marine water quality indicators, whereas previous weather generators were mainly focusing on land based impacts (hydrology, agriculture, ecosystem changes). While these conceptual elements separately exist in the field of stochastic weather generators or more broadly in the field of climate sciences and/or environmental sciences, the combination of these features can be considered innovative.

In summary, this study presents a Bayesian approach to simulate climate variables in analogy with stochastic weather generators extended to a larger temporal scale. The generated ensemble of future radiation projections is used to characterize climate model uncertainties and to assess ecological response in marine and coastal ecosystems through a physically-based impact model.

4.2. DATASET

Numerous General Circulation Models (GCMs) and Regional Climate Models (RCMs) exist that produce long term predictions of climate variables. In this study the Surface Downwelling Shortwave Radiation dataset, hereafter referred to as radiation, was obtained from the high resolution 0.11 degree (or 12.5 km) EURO-CORDEX (Coordinated Regional Downscaling Experiment) [106] which uses the Swedish Meteorological and Hydrological Institute Rossby Centre regional atmospheric model (SMHI-RCA4). Radiation was chosen for the demonstration case due to its high influence on ecological processes. In short, radiation is the measure of solar radiation energy received on a given surface area. Radiation influences light and energy availability for living organisms in the water column and therefore controls their growth and mortality among others, such as nutrient availability and temperature.

In order to produce various regionally downscaled scenarios, EURO-CORDEX applies a range of GCMs to drive the above mentioned RCM. The four driving GCMs in this study are the National Centre for Meteorological Research general circulation model (CNRM-CM5), the global climate model system from the European EC-Earth consortium (EC-EARTH), the Institut Pierre Simon Laplace Climate Model at medium resolution (IPSL-CM5A-MR), and the Max-Planck-Institute Earth System Model at base resolution (MPI-ESM-LR). These GCMs have been previously used in recent studies describing the impacts of climate change on ecosystem state and biodiversity [80]. In addition to the driving models, further scenarios are obtained by considering different socio-economic changes described in the Representative Concentration Pathways (RCPs). RCPs are labeled according to their specific radiative forcing pathway in 2100 relative to pre-industrial values. In this study we include RCP8.5 (high), and RCP4.5 (medium-low) [214]. Together the four different driving GCMs and two RCPs provide

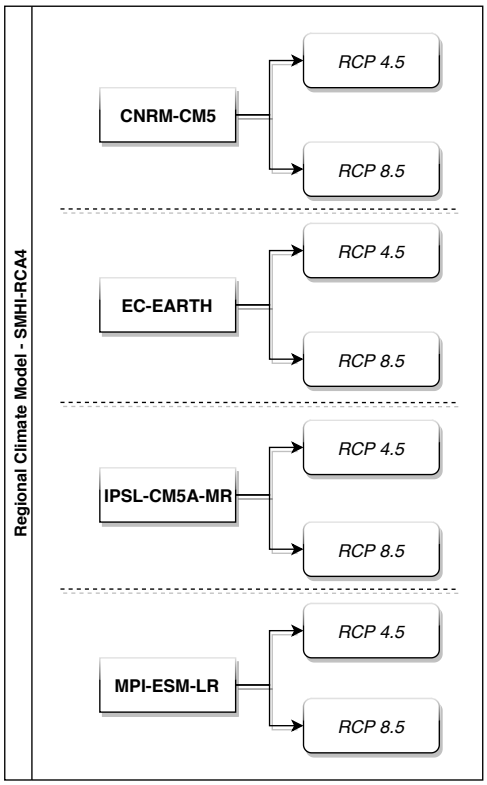


Figure 4.1: Overview of the eight EURO-CORDEX climate scenarios used in the study derived from four driving general circulation models, and two socio-economic scenarios.

us with an ensemble of eight future radiation scenarios (see Figure 4.1). We make use of these driving GCMs and RCP scenarios as they were previously selected and post-processed for climate change assessments for this study area by the Institute of Atmospheric Sciences and Climate (ISAC-CNR) within the EU H2020 ECOPOTENTIAL project.

Daily field observations of solar radiation energy were obtained from the Royal Netherlands Meteorological Institute (KNMI) at the closest weather station, De Kooy, from 1970 until August 2020. This time interval covers the entire Euro-CORDEX reference period (1970-2005) and more than 14 years of the projection period (2006 - mid 2020). These observations were used for the bias correction of the RCM scenarios and for validation of the generated scenarios.

While this ensemble of $GCM \times RCP$ combinations already encompasses a certain degree of uncertainty, the number of ensemble members might not be sufficient and information on the likelihood of its members is difficult to obtain. This is due to the fact that RCP scenarios have not been assigned a formal likelihood and it is generally assumed that each climate model is independent and of equal ability [96].

Previously, attempts have been made to assess the likelihood of the different climate change pathways [42] and a number of studies use model weighting based on past per-

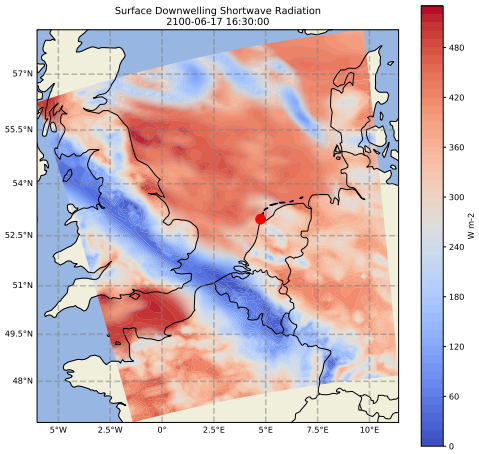


Figure 4.2: Surface downwelling shortwave Radiation from RCM SMHI-RCA4 driven by GCM CNRM-CM5 with RCP8.5. Cooler colors indicate lower solar radiation energy received at the surface area, while warmer colors indicate higher radiation energy. The location of the study site is shown by the red dot.

formance, yet technological, economic, political and climatic factors underlying RCP scenarios remain largely uncertain and model weights based on historical performance might not be adequate for other regions, variables and for future projections [116]. For this reason, in this study the given ensemble of scenarios is enriched to be used for comprehensive uncertainty quantification studies and for fully probabilistic simulations in assessment studies.

The used subset of the EURO-CORDEX dataset covers a domain between 2.0W-10E longitudes and 48-57N latitudes, as depicted in Figure 4.2, with a resolution of 0.11 degree on curvilinear grid (cca. 12x10 km). For the purpose of this study, time series were extracted at a given location in the Dutch Wadden Sea (see red dot in Figure 4.2) to reduce the data dimension and to be used as input for the single site stochastic generator. The original radiation time series is a high frequency dataset, with three-hourly time step, which was aggregated to daily averages excluding zero radiation values during night time. Data aggregation was done to match the daily time step of the validation data and to reduce unnecessary noise (sub-daily variations) in the dataset as the sub-daily scale of the processes are not relevant for the purpose of this study. In other cases where smaller temporal scales are important the data aggregation step could be excluded.

4.3. STOCHASTIC GENERATOR METHODOLOGY

The methodological workflow depicted in Figure 4.3 starts with the pre-processing of the time series of radiation data. This step includes bias correction of the Euro-CORDEX RCM scenarios, as well as extracting the seasonal shape φ^S, the seasonality in the variance of residuals φ^V, and deviations from the seasonal cycle (d_j). The stochastic generator uses a parameterized time series model as in equation (4.1) below which consists of a linear component (eq. (4.2)), seasonal component (eq. (4.3)), and a variance compo-

nent (eq. (4.4)). Consequently, the model contains the following parameters: intercept α, slope β, amplitudes of seasonal shapes $\{A_j^S\}$, amplitudes of seasonality in the variance of residuals $\{A_j^V\}$) and a variance parameter (σ^2). We will endow all parameters with a prior distribution and perform inference within the Bayesian setup. Sampling from the posterior distribution is done using the Gibbs sampler. Finally, the posterior samples can be used to sample from the predictive distribution and generate synthetic radiation signals. The temporal evolution of these generated synthetic scenarios is regular, meaning that the lengths of yearly cycles are always equal. Since we observe in the pre-processing step that in reality seasonal cycles should not be equally long, temporal deregularization is done using a time change function $\tau(t_k)$ (see eq. (4.5)). The end result is numerous generated radiation scenarios, which are representative of the input Euro-CORDEX scnearios and have varying lengths of seasonal cycles.

4.3.1. BIAS CORRECTION

The RCM simulations are subject to climate model structural error and boundary errors from the driving GCMs [153], hence, they should be bias corrected before applying them in impact studies [136]. These systematic biases present in climate models are most commonly addressed using standard bias correction techniques, such as mean adjustment or quantile mapping. Nevertheless, due to the known problems with these bias correction techniques [140], one can confidently apply them only if the relevant processes are reasonably well captured by the chosen climate models, since fundamental model biases cannot be corrected by the bias correction approaches. While a comprehensive validation of the RCM simulations was not conducted in this study, sufficient credibility of the future projections in representing local and large-scale processes is assumed since they are originated from a high-resolution regional downscaling experiment adhered to a coordinated model evaluation framework.

Based on this assumption, quantile mapping bias correction [11] was applied using the RCM simulations for reference period (1976-2005) and daily historical radiation field measurements from KNMI for the same period. The quantile-quantile mapping transfer functions were established for the reference period and separately for each RCM simulation. The transfer functions were then applied for the bias correction of each future projections separately. Figure 4.4 depicts the histogram of observations together with the uncorrected and bias corrected RCM simulations in the projection time interval when field measurements are still available (2006-2020). While dissimilarities exist between modelled and observed distributions, these are not major, indicating that key processes are not misrepresented by this RCM [140].

4.3.2. TEMPORAL EVOLUTION

It was verified that significant differences in the temporal evolution of the selected RCM scenarios during the projection interval (2006-2100), which could be reflected in differences in trends, do not exist. Nevertheless, pre-processing steps were applied to remove identified minor differences in time evolution. Since it was observed that not all years had the same number of data points (within RCM scenarios and across scenarios), the time evolution was regularized by interpolation using nearest neighbor method. As the differences in the number of yearly data points were minor, the interpolation had lim-

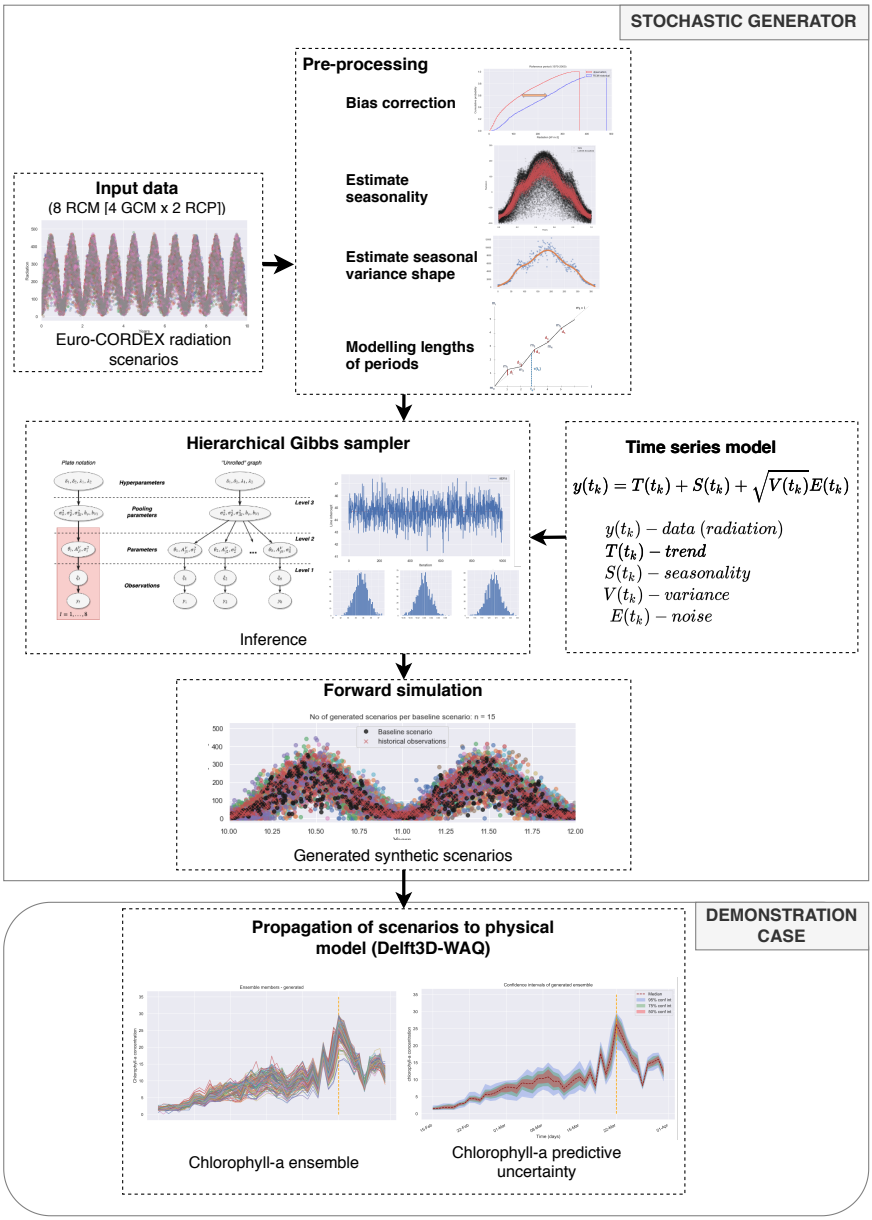


Figure 4.3: Schematization of the stochastic generator methodology and demonstration case

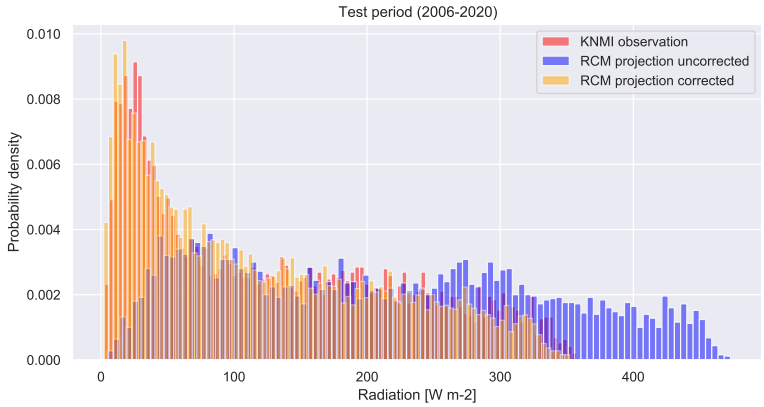


Figure 4.4: Quantile mapping bias correction. Comparison of the histograms of the KNMI observations (red) with the uncorrected (blue) and corrected versions (yellow) of the RCM projection for the test period (2006-2020). Example for one RCM projection.

ited impact on the dataset. After this regularization step all scenarios had uniform time evolution. Further considerations regarding the lengths of the yearly cycles is described in Section 4.3.3.

4.3.3. TIME SERIES MODEL DEFINITION

For simplicity we first consider only one scenario in this section and later extend the time series model to all scenarios in its full form (Section 4.3.3). Suppose $t_0 < t_1 < \cdots t_K$ are observation times (in years) and $j \in \{1, \ldots, J\}$ indexes years (used later in equations). Let $y(t_k)$ denote the radiation measurement at time t_k. We assume

$$y(t_k) = T(t_k) + S(t_k) + \sqrt{V(t_k)}E(t_k) \tag{4.1}$$

with $T(t_k)$ the trend component, $S(t_k)$ the seasonal component, $V(t_k)$ the variance of noise and $E(t_k)$ the noise component.

SEASONALITY

As a first step, we assume a linear trend in the data:

$$T(t_k) = \alpha + \beta t_k, \tag{4.2}$$

where α and β are the intercept and slope parameters respectively. The detrended time series has a noisy but clearly distinguishable yearly cyclic pattern. Our goal is to estimate this cyclic behaviour and define a seasonal shape function φ^S that represents the seasonality. In order to achieve this, the following steps are performed. After removing the trend $T(t_k)$, the time series is smoothed using a non-parametric smoother LOWESS (Locally Weighted Scatterplot Smoothing) [48] 'to remove noise'. Then the local minima points (m_j) of the smooth time series are identified, see the upper plot in Figure 4.5. Based on the identified local minima points the original detrended time series is split

into yearly curves and the LOWESS smoother is applied again to estimate the seasonal shape (center plot in Figure 4.5) separately for each scenarios.

In the LOWESS smoothing, the time window was chosen intuitively by plotting various LOWESS curves and comparing the fits graphically to avoid over- or under-fitting. The aim was to find a time window which allows us to obtain a seasonal curve with sufficient details to describe characteristic features, such as the two "shoulders" in the seasonal curve, but at the same time removing all noise. The fraction value found to be most appropriate for the yearly seasonality was 0.1 (10% of the yearly data points) meaning that the time window is roughly one month. Another option could have been to choose a time window that optimizes the fit of the LOWESS curve through bias-corrected Akaike information criterion (AIC) method, Generalized Cross-Validation (GCV) method or similar. While these options might be more robust in other cases, we think that intuitively choosing the time window for the seasonal shape is preferable for the stochastic generator methodology as it is more flexible and allows us to incorporate domain knowledge and preferences. The averaged LOWESS smoothed seasonal shape (φ^S) is depicted in the lower plot in Figure 4.5.

Considering a time series with seasonal cycles (years) the seasonality $S(t_k)$ in the time series model is defined by

$$S(t_k) = \sum_{j=1}^{J} A_j^S \varphi^S(t_k - j + 1)\, \mathbb{1}_{[j-1,j)}(t_k), \tag{4.3}$$

where A_j^S is a scaling factor for year j, and the seasonal shape $\varphi^S : [0,1] \to \mathbb{R}$. As an example if $t_k = 1.5$ then $S(1.5) = A_2^S \varphi^S(0.5)$ since it is in the second year. Note that $\mathbb{1}_{[j,j+1)}(t_k)$ is an indicator function which is 1 for all elements within the interval $[j, j+1)$ and 0 otherwise.

RESIDUALS AND SEASONAL SHAPE IN THE VARIANCE φ^V

Apart from the linear- and seasonal trends there is an additive residual term $\sqrt{V(t_k)}E(t_k)$ in the time series model. In this residual term the noise variables $E(t_k)$, where $0 \le k \le K$, are assumed to be independent and identically distributed (i.i.d) $N(0,\sigma^2)$ and the variance term $V(t_k)$ is defined similarly to the seasonal component in equation ((4.3)):

$$V(t_k) = \sum_{j=1}^{J} A_j^V \varphi^V(t_k - j + 1)\, \mathbb{1}_{[j-1,j)}(t_k) \tag{4.4}$$

where A_j^V is a scaling factor for year j and $\varphi^V : [0,1] \to \mathbb{R}$ is the LOWESS smoothed variance of residuals. The seasonal shape of the variance φ^V depends on the specific scenario, same as for the seasonal shape φ^S. The seasonal variance shape is depicted in Figure 4.6. The lower panel of Figure 4.6 shows the comparison of the observed (in blue) and modeled (in red) residual, which show good agreement, meaning that the time series model is capable of representing the input signal. The time series model refinement process stops at this point when residuals are properly modeled.

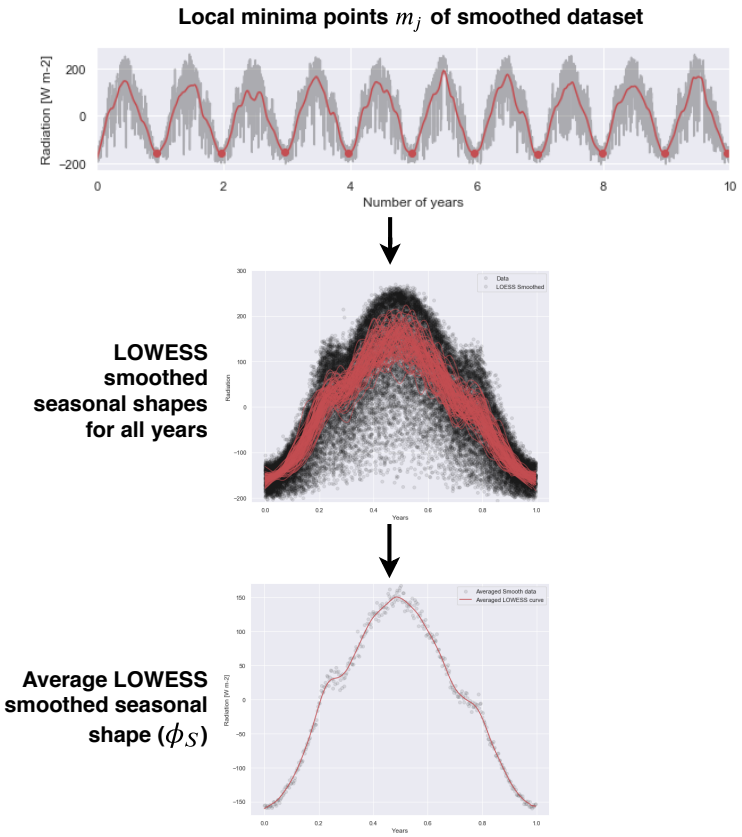
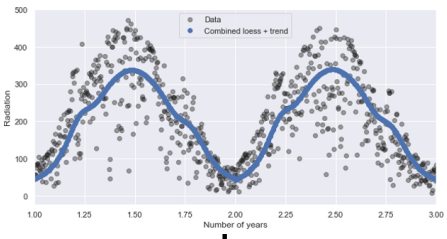
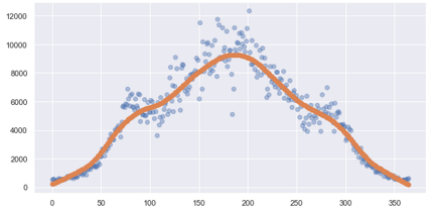


Figure 4.5: Deriving seasonal shape φ^S. Local minima points and smoothed dataset (top), comparison of yearly data and smoothed curves (middle), average smothed curve as seasonal shape (bottom).

Seasonal signal and orginal dataset



Average LOWESS smoothed seasonal variance shape (ϕ_V)



Residual term

$N(0, V(\tau(t_k))\sigma_E^2)$

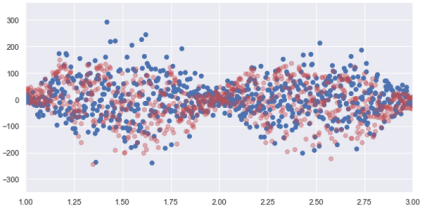


Figure 4.6: Residuals and deriving seasonal shape in the variance φ^V. Seasonal signal and data (top), smoothed shape of the variance (middle), residual term including variance and independent and identically distributed noise (bottom). In the bottom panel the blue color represents the observed residuals while the red color represents the modeled residual.

MODELLING LENGTHS OF PERIODS

In the radiation dataset the local minima points (upper panel of Figure 4.5) are not equidistant, indicating that the seasonal cycle lengths have slight deviations over the years. Since the variation in the length of seasonal periods is an important feature, it should be incorporated in the stochastic generator. The deviations from the calendar year d_j are defined as:

$$d_j = j - m_j$$

where m_j is the j-th local minimum location (in years). The upper panel of Figure 4.7 shows the deviations from the calendar year and their autocorrelation, while the middle panel shows the distribution of these deviations. It can be observed that the deviations are centered around zero and have a negative lag 1 autocorrelation meaning that most positive deviations tend to be followed by negative deviations and vice versa. In this way the yearly cycle lengths remain close to the ideal cycle length (one calendar year) throughout the time series.

In order to account for these non-uniform cycle lengths when generating new synthetic scenarios, a time change function $\tau(t_k)$ is introduced. For a visual representation of $\tau(t_k)$ see the bottom panel of Figure 4.7.

When a new synthetic scenario is generated, for each year j a deviation value d_j is produced by sampling from the observed deviation distribution of that scenario. Then, knowing the deviation for each year, we calculate the location of the end of the j-th period m_j. Plotting m_j against the year j will result in a piecewise function as shown in the bottom panel of Figure 4.7. When introducing $\tau(t_k)$ we essentially create a piecewise linear function which provides for every time instance t_k the new continuous time value $\tau(t_k)$ assuming that between $[m_j, m_{j+1}]$ the time is linearly increasing. Mathematically, the piecewise linear time change function is described as follows:

$$\tau(t_k) = \Big(m_j + (m_{j+1} - m_j)(t_k - j)\Big)\,\mathbb{1}_{[j,j+1)}(t_k) \tag{4.5}$$

where $m_j = j - d_j$, and the sequence $\{d_j\}$ is modeled as independent and identically distributed random variables $\mathrm{N}\left(0, \sigma_d^2\right)$.

FULL TIME SERIES MODEL

In this section we write the introduced time series model in full form, extended to all scenarios, without time change $\tau(t_k)$.

Recall that j indexes years and k indexes days. Let $\ell \in \{1, \ldots, L\}$ index the scenarios (we have $L = 8$). For scenario ℓ we have measurements

$$y_\ell = \begin{bmatrix} y_\ell(t_0) & \cdots & y_\ell(t_K) \end{bmatrix}^T.$$

Define

$$\varphi^u_{\ell,j,k} = \varphi^u_\ell(t_k - j + 1)\,\mathbb{1}_{[j-1,j)}(t_k), \qquad u \in \{S, V\}$$

and set

$$\Phi^S_\ell = \begin{bmatrix} 1 & t_0 & \varphi^S_{\ell,1,0} & \cdots & \varphi^S_{\ell,1,K} \\ \vdots & \vdots & \ddots & \cdots & \vdots \\ 1 & t_K & \varphi^S_{\ell,J,0} & \cdots & \varphi^S_{\ell,J,K} \end{bmatrix}$$

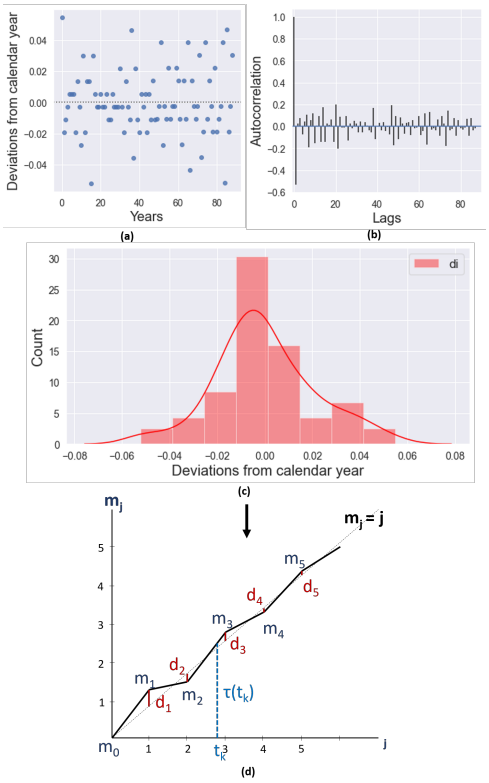


Figure 4.7: Deviations from the calendar year (a), together with their autocorrelation plot (b), and histogram (c). Sub-figure (d) depicts the $\tau(t_k)$ time change piecewise linear function. This function is represented by the j-th local minimum location (in years) m_j on the y-axis, against the number of years j on the x-axis. Differences between m_j and the diagonal line are the deviations from the calendar years d_j. Example shown for one Euro-CORDEX scenario.

along with

$$V_\ell = \sigma_\ell^2 \operatorname{diag}\left(\sum_{j=1}^{J} \varphi_{\ell,j,0}^V A_{j\ell}^V, \ldots, \sum_{j=1}^{J} \varphi_{\ell,j,K}^V A_{j\ell}^V\right)$$

and

$$A_\ell^u = \begin{pmatrix} A_{1\ell}^u & \cdots & A_{J\ell}^u \end{pmatrix}, \qquad u \in \{S, V\}$$

The conditional distribution for the observation vector for scenario ℓ is given by

$$y_\ell \mid \xi_\ell \sim \mathrm{N}\left(\Phi_\ell^S \theta_\ell, \sigma_\ell^2 V_\ell\right),$$

where

$$\xi_\ell = \left(\theta_\ell, A_\ell^V, \sigma_\ell^2\right), \qquad \text{with} \qquad \theta_\ell = \left(\alpha_\ell, \beta_\ell, A_\ell^S\right)$$

denoting the vector obtained by stacking all components of its elements.

4.3.4. PRIOR SPECIFICATION

We choose partially conjugate priors to simplify MCMC-sampling with the Gibbs sampler. We denote the Normal, Inverse Gamma and Gamma distributions by N, IG and G respectively. Moreover, we denote by $\mathrm{N}(x; \mu, \sigma)$ the density of $\mathrm{N}(\mu, \sigma)$-distribution, evaluated at x. Similar notation is used for the Gamma- and InverseGamma distributions. In the Appendix we specify the densities of these distributions to clarify the parametrisations used in their definitions. We take the following prior for ξ_ℓ

$$\begin{aligned}
\{\alpha_\ell\} \mid \sigma_\alpha^2 &\overset{\text{iid}}{\sim} \mathrm{N}\left(0, \sigma_\alpha^2\right) \\
\{\beta_\ell\} \mid \sigma_\alpha^2 &\overset{\text{iid}}{\sim} \mathrm{N}\left(0, \sigma_\beta^2\right) \\
\{A_\ell^S\} \mid \{\sigma_{S\ell}^2\} &\overset{\text{ind}}{\sim} \bigotimes_{j=1}^{J} \mathrm{N}\left(0, \sigma_{S\ell}^2\right) \\
\{A_\ell^V\} \mid \{b_{V\ell}\} &\overset{\text{ind}}{\sim} \bigotimes_{j=1}^{J} \mathrm{IG}\left(a_V, b_{V\ell}\right) \\
\{\sigma_\ell^2\} \mid b_\sigma &\sim \mathrm{IG}\left(a_\sigma, b_\sigma\right)
\end{aligned}$$

for hyperparameters a_V and a_σ. To tie together the laws for different scenarios, we complete the prior specification by another layer

$$\begin{aligned}
\sigma_\alpha^2, \sigma_\beta^2, \{\sigma_{S\ell}^2\} &\overset{\text{iid}}{\sim} \mathrm{IG}\left(\delta_1, \delta_2\right) \\
\{b_{V\ell}\}, b_\sigma &\overset{\text{iid}}{\sim} \mathrm{G}\left(\lambda_1, \lambda_2\right)
\end{aligned}$$

for hyperparameters $\delta_1, \delta_2, \lambda_1, \lambda_2$.

Since climate scenarios originate from a common genealogy (e.g. similar computational schemes, description of similar physical processes) [190], our underlying idea is that scenarios can be assumed exchangeable rather than independent. This induced the well known phenomenom of "borrowing strength" where estimates for parameters over different scenarios are combined ("pooled"), see Figure 4.8. This can correct outlier-like behaviour and makes the estimates statistically more robust [76, 78].

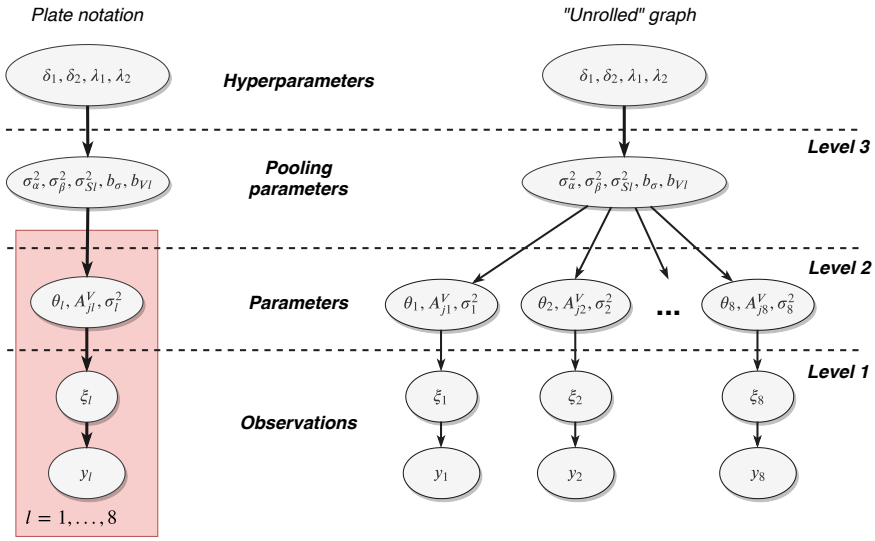


Figure 4.8: Illustration of the hierarchical Bayesian model structure. Scenarios are indexed by l.

In our case the reason to opt for a hierarchical setup is to enhance the Bayesian model by using all the data (all scenarios) to perform inferences for each group (scenario). This provides a trade off between the noisy within-group estimate, where parameters are estimated independently from the other groups, and an oversimplified parameter estimate that uses all data and ignores the presence of groups [78].

4.3.5. GIBBS SAMPLER FOR DRAWING FROM THE POSTERIOR DISTRIBUTION

Sampling from the posterior can be done using a blocked Gibbs sampler where for the parameters in the second layer we sample from the following full conditionals

- $\theta_\ell \sim \mathrm{N}\left(\Sigma(\Phi^S_\ell)^T \sigma_\ell^{-2} V_\ell^{-1} y_\ell, \Upsilon_\ell\right)$, where

$$\Upsilon_\ell^{-1} = \Sigma_\ell^{-1} + (\Phi^S_\ell)^T \sigma_\ell^{-2} V_\ell^{-1} \Phi^S_\ell, \text{with } V_\ell = \mathrm{diag}(\sigma^2_\alpha, \sigma^2_\beta, \sigma^2_{SL}, \ldots, \sigma^2_{S\ell}).$$

- $A^V_{j\ell} \sim \mathrm{IG}\left(a_v + |I_j|/2, b_{V\ell} + (2\sigma^2_\ell \varphi^V_{\ell,j,k})^{-1} \sum_{k \in I_j} (y_\ell(t_k) - \mu_{\ell,k})^2\right)$,
with

$$I_j = \{k : t_k \in [j, j+1)\} \text{ and } \mu_{\ell,k} = \mathrm{row}_k(\Phi^S_\ell)\theta_\ell.$$

- $\sigma^2_\ell \sim \mathrm{IG}\left(a_\sigma + K/2, b_\sigma + \frac{1}{2}(y_\ell - \Phi^S_\ell \theta_\ell)^T V_\ell^{-1} (y_\ell - \Phi^S_\ell \theta_\ell)\right)$

Here, $\ell = 1, \ldots, L$, and $j = 1, \ldots, J$. For the third layer we sample from

- $\sigma^2_\alpha \sim \mathrm{IG}\left(\delta_1 + L/2, \delta_2 + \frac{1}{2}\sum_{\ell=1}^L \alpha^2_\ell\right)$
- $\sigma^2_\beta \sim \mathrm{IG}\left(\delta_1 + L/2, \delta_2 + \frac{1}{2}\sum_{\ell=1}^L \beta^2_\ell\right)$
- $\sigma^2_{S\ell} \sim \mathrm{IG}\left(\delta_1 + J/2, \frac{1}{2}\sum_{j=1}^J (A^S_{j\ell})^2\right)$

- $b_{V\ell} \sim \mathrm{G}\left(\lambda_1 + Ja_v, \lambda_2 + \sum_{j=1}^{J}\left(A_{j\ell}^{V}\right)^{-1}\right)$
- $b_\sigma \sim \mathrm{G}\left(\lambda_1 + La_\sigma, \lambda_2 + \sum_{\ell=1}^{L}\sigma_\ell^{-2}\right)$

Values of the hyperparameters were set to $\delta_1 = \lambda_1 = 2$, $\delta_2 = 1$, and $\lambda_2 = 0.01$.

Derivation of these updates is straightforward as the hierarchical model implies that the posterior satisfies

$$
\begin{aligned}
& p\left(\{\xi_\ell\}, \sigma_\alpha^2, \sigma_\beta^2, \{\sigma_{S\ell}^2\}, \{b_{V\ell}\}, b_\sigma \mid \{y_\ell\}\right) \\
& \quad \propto \prod_{\ell=1}^{L}\left\{\mathrm{N}\left(y_\ell; \Phi_\ell^S \theta_\ell, V_\ell \sigma_\ell^2\right)\mathrm{N}\left(\alpha_\ell; 0, \sigma_\alpha^2\right)\mathrm{N}\left(\beta_\ell; 0, \sigma_\beta^2\right)\right. \\
& \quad \left.\times \mathrm{IG}\left(\sigma_\ell^2; a_\sigma, b_\sigma\right)\prod_{j=1}^{J}\left[\mathrm{N}\left(A_{j\ell}^S; 0, \sigma_{S\ell}^2\right)\mathrm{IG}\left(A_{j\ell}^V; a_V, b_{V\ell}\right)\right]\right\} \\
& \quad \times \mathrm{IG}\left(\sigma_\alpha^2; \delta_1, \delta_2\right)\mathrm{IG}\left(\sigma_\beta^2; \delta_1, \delta_2\right)\mathrm{G}\left(b_\sigma; \lambda_1, \lambda_2\right) \\
& \quad \prod_{\ell=1}^{L}\left\{\mathrm{IG}\left(\sigma_{S\ell}^2; \delta_1, \delta_2\right)\mathrm{G}\left(b_{V\ell}; \lambda_1, \lambda_2\right)\right\}.
\end{aligned}
$$

Only derivation of the update for A_ℓ^V is slightly tedious and requires bookkeeping that any time t_k is only in one year (indexed by j). Note that the priors are chosen such that all update steps in the Gibbs sampler are partially conjugate. Due to the random error, generated values may fall below zero when radiation is low. In order to avoid this, results are truncated at zero to comply with physics, as solar radiation cannot be negative. For this reason, the above introduced model is an approximation. In the model formulation we neglect the impact of truncation.

4.4. PROPAGATION OF UNCERTAINTY - DEMONSTRATION CASE

In this demonstration case, the generated radiation scenarios are used to drive a physically based model to investigate the effects on water quality (ecology) and further propagate climate related uncertainties to better characterize the response of the ecological system. The optimal number of stochastic generator realizations for environmental applications has been previously investigated [89, 7]. These studies assessed the impact of output size of weather generators on statistical characteristics and indices as compared to historical data and try to reach a predefined accuracy. Apart from accuracy, for probabilistic impact studies one should also consider the impact of ensemble size on how well the predictive distribution of a weather-related variable, such as radiation, can be estimated [123]. Based on these studies the authors conclude that an ensemble size of around 100 members is optimal for the demonstration case, while also computationally feasible.

For the impact modelling, the water quality sub-module of the Delft3D integrated modelling system, Delft3D-WAQ, is used with an existing model setup which has been previously calibrated and validated for the location of our demonstration case [133]. The

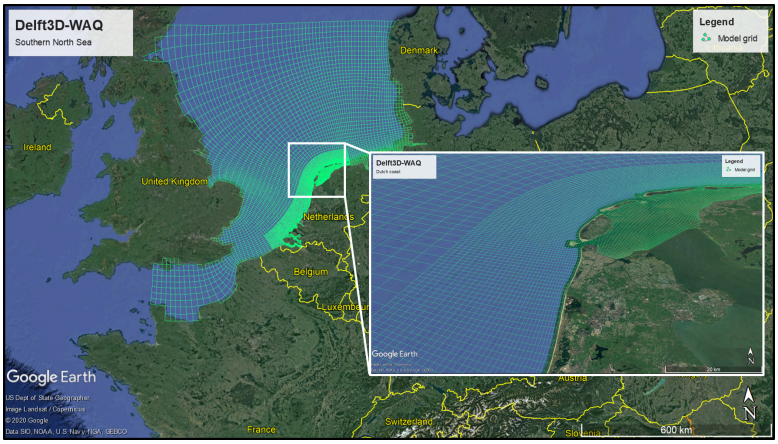


Figure 4.9: Demonstration case: Delft3D-WAQ model domain in the Southern North Sea and along the Dutch coast.

spatial domain of the physical model covers the Southern North Sea with coarser horizontal resolution offshore and finer resolution along the Dutch coast, as shown in Figure 4.9. The model comprises of twelve vertical layers, making it a three dimensional physical model.

Delft3D-WAQ is a comprehensive hybrid ecological model including an array of modules reproducing water quality processes that are then combined with a transport module to calculate advection and dispersion. The model most importantly calculates primary production and chlorophyll-a concentration while integrating dynamic process modules for dissolved oxygen, nutrient availability and phytoplankton species. Delft3D-WAQ can include a phytoplankton module (BLOOM) that simulates the growth, respiration and mortality of phytoplankton. Using this module the species competition and their adaptation to limiting nutrients or light can be simulated [133]. A graphical overview of the modelled ecological processes can be seen in Figure 4.10. Without describing in details the formulation of these ecological processes we briefly introduce how our variable of interest, chlorophyll-a, is calculated and how solar radiation influences its concentration.

The chlorophyll-a content of algae is species specific. The total chlorophyll-a concentration is equal to the sum of the contributions of all algae species:

$$C_{\text{chlfa}} = \sum_{i=1}^{n} C_{\text{alg,i}} \tag{4.6}$$

where C_{chlfa} is the total chlorophyll-a concentration and $C_{\text{alg,i}}$ is the biomass concentration for algae species type i. The mass balances for algae types are based on growth, respiration and mortality which are influenced by factors such as nutrients (nitrogen, phosphorus, silicon, carbon) and light in the water column. BLOOM uses linear optimisation to calculate the species competition and the optimum distribution of biomass over all algae types. The goal of the optimization process is to maximize the net growth

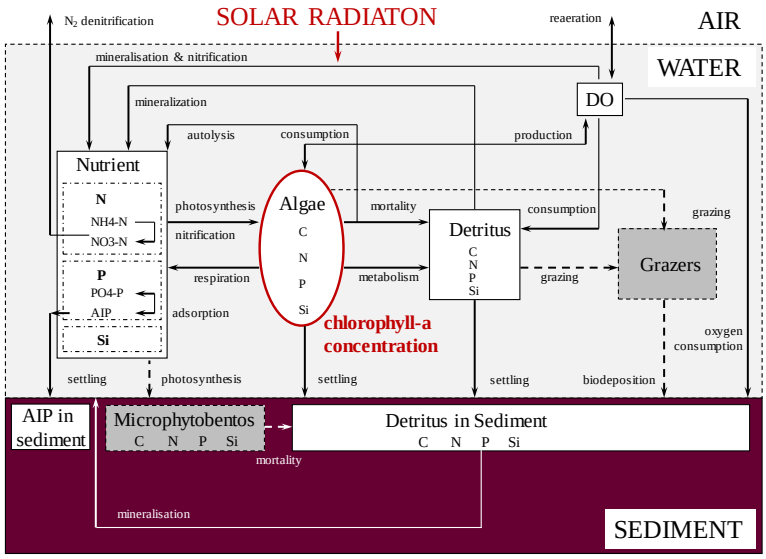


Figure 4.10: Delft3D-WAQ processes. State variables in grey and processes indicated by dashed lines have not been included in the North Sea model application. AIP is absorbed inorganic phosphorus.

rate of the total of all algae types under nutrient availability, light availability, maximum growth rate, and maximum mortality rate constraints.

Light availability, therefore, is an important driving factor for phytoplankton processes, and this light availability is a function of solar radiation energy provided by the RCMs and the stochastic generator. More specifically, the available light at a particular water depth is calculated as a function of solar irradiation on the top layer and the light attenuation in the water column caused by extinction (scattering and absorption). This light extinction is modelled by the Lambert-Beer law (eq. (4.7)) which states that the light intensity in the water layers is exponentially decreasing with the water depth:

$$I_b = I_t e^{(-KH)} \tag{4.7}$$

where I_b is the light intensity at bottom of the water column, I_t is the light intensity at the top of the water column (solar radiation forcing), K is the light extinction coefficient and H is the water depth. The extinction coefficient K is the sum of the background extinction and the extinction of all other light absorbing suspended organic or inorganic matter (the self-shading of phytoplankton, extinction of total suspended matter and the dissolved humic substances).

Consequently, projected change in solar radiation at the water surface, which translates into light intensity in the water column, is an influential factor to determine future changes in chlorophyll-a concentration. The demonstration case aims to showcase this cause-effect relationship and quantify the associated uncertainties.

4.5. RESULTS

4.5.1. RESULTS OF STOCHASTIC GENERATOR

Initial states and hyperparameters were specified, and the Gibbs sampler was run for over a thousand iterations. Samples drawn from the posterior distributions for all scenarios are summarized in Figure 4.11 as violins plots. For the interpretation of the results the reader is reminded that scenarios one to four are from different driving GCMs with RCP4.5, and scenarios five to eight are from the same GCMs as the first four scenarios, respectively, but with RCP8.5. It can be seen that the intercept and slope parameters of all scenarios are similar and their ranges are overlapping, even though, scenario four shows a slightly different behaviour. GCM number three (IPSL-CMSA-MR) and 4 (MPI-ESM-LR) have higher variances than the other GCMs as indicated by the σ^2 plot (see third plot in Figure 4.11). It should be mentioned that before pooling was applied, through an additional layer in the hierarchical model, scenario four showed stronger outlier behaviour. This behaviour is reduced in the hierarchical scheme as estimates get pulled towards the overall mean of the various scenarios. We can also observe in the third panel of Figure 4.11, where the σ^2 estimates are shown for each scenario, that dissimilarities between scenarios are dominated by driving GCMs. This is in line with previous finding that uncertainty in the RCM projection scenarios are primarily influenced by the driving GCMs while the impact of RCPs is less dominant [151].

Regarding the trend slope, it is a general expectation that RCP8.5 (scenarios 5-8) has steeper slope than RCP4.5 (scenarios 1-4). This expectation was confirmed for the temperature variable. While trend slopes for solar radiation under RCP8.5 are also slightly steeper, with an average difference of 0.014, it is less pronounced. This unexpected feature could be explained by the complexity of projecting solar radiation for this region, which has been previously discussed in literature. A study by [17] found remarkable discrepancy between RCMs and their driving GCMs, since GCMs consistently indicated increase in solar radiation over Europe until the end of the century, while most RCMs detected general decrease. Moreover, the difficulty of projecting cloud cover and solar radiation changes in coastal areas with sea-land-atmosphere boundaries, such as the study site, has also been highlighted.

By sampling from the predictive distribution new synthetic scenarios are generated. Posterior predictive checks [78] have been done visually by comparing the original and sampled data, together with observations, as shown in Figure 4.14. It can be observed that the seasonal shape is well reproduced and the ensemble band of the new scenarios around the baseline scenario suggests the presence of uncertainties both in peak concentrations and phase shifts. This indicates benefits of using a larger ensemble of scenarios as input for ecological studies. Further validation of the stochastic generator, and especially its ability to accurately represent long-term trends, has been done by fitting it with the observations for the period between 1970 and 2020. The time series of the observations and generated scenarios have been decomposed and their trend, seasonal and residual components have been compared in Figure 4.12. We can observe that the time series model performs as intended and able to closely reproduce the long-term trend, seasonal and residual signals of the observations. Consequently, we can conclude that the stochastic generator produces valid outputs including correct representation of the climate change signal.

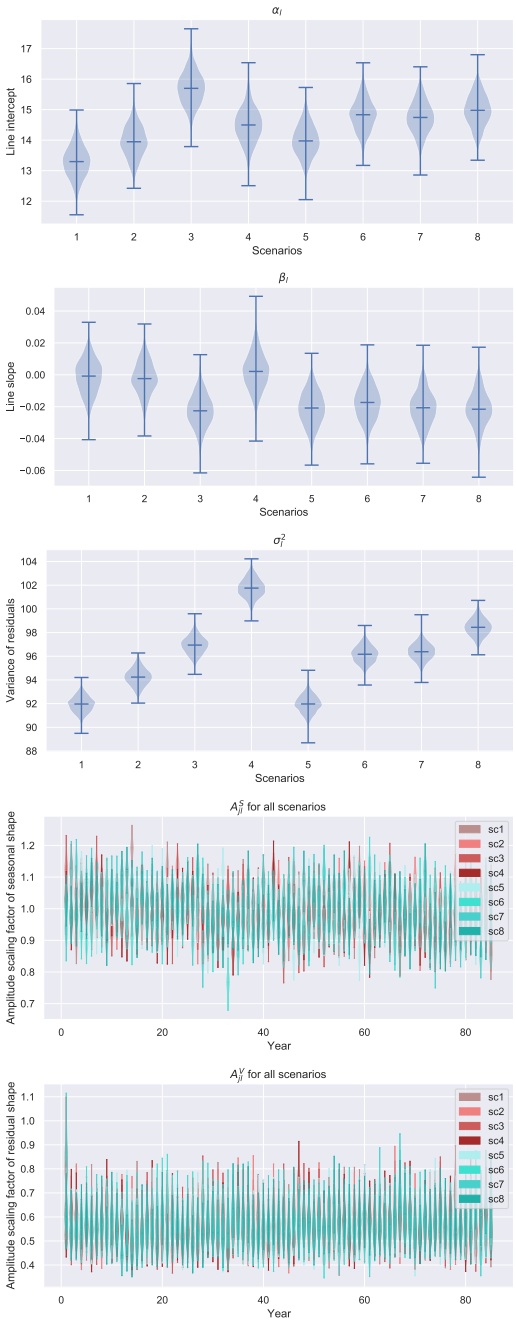


Figure 4.11: Violin plots of parameter samples for the eight baseline scenarios. The top three plots show constant values for the α, β, and σ_E^2 parameters, while the bottom two plots display varying values over 85 years for the A_j^S and A_j^V parameters.

For the demonstration case an ensemble of 120 new scenarios were generated by equally drawing from each baseline scenario (15 new scenarios of each type). The generated scenarios have been verified by comparing their empirical quantiles graphically to the baseline scenarios for the entire projection period (2006-2090), depicted in Figure 4.13. The quantile-quantile plots of three example generated scenarios approximately lie on the diagonal line and there are no obvious discrepancies, except for the tale, which can be explained by the fact that we take normally distributed noise which is symmetric.

Finally, Figure 4.15 shows boxplots of the generated scenarios for each month, with the corresponding monthly mean statistics of the baseline scenarios as solid lines. Since the temporal evolution of baseline and generated scenarios are similar, there is no problem with the long-term linear trend differences and therefore the figure covers the entire projection period (2006-2090). We can observe that each RCM climatological mean for all the months are well captured by the generated synthetic scenarios, as they fall within the interquartile range of the boxplots.

4.5.2. RESULTS OF PROBABILISTIC WATER QUALITY SIMULATION

The eight baseline Euro-CORDEX radiation scenarios and the 120 generated ones are used as input for the Delft3D-WAQ numerical model to drive ecological processes which calculate chlorophyll-a concentration, among others. The objective of this demonstration case is to illustrate the benefit of using a larger ensemble of radiation inputs and to assess the impacts of different radiation intensities towards the end of this century, during the early spring season when (solar) energy is the limiting factor to biomass growth. Consequently, further analysis focuses on the early spring months. In order to simulate ecological variables for the spring season the baseline Euro-CORDEX radiation scenarios and the outputs of the stochastic generator were post-processed. Seasonal averages of the first (2006-2015) and last (2081-2090) simulated decade were derived, thus obtaining a single year signal for each baseline and generated scenarios. These processed radiation signals were then used to force the deterministic physical model. The simulated chlorophyll-a concentrations therefore indicate the characteristic spring peak of the beginning and the end of the century, not a single event.

A subset of the simulation results are shown in Figure 4.16 focusing on the spring peak. The figure depicts the chlorophyll-a concentration ensemble members and their medians derived from the baseline and generated scenarios for the first and last simulated decade, as well as the prediction intervals that can be computed using the generated scenarios. The figure aims at comparing the evolution and peak of the characteristic spring blooms. In the upper panel, it is visible that most of the baseline scenarios are close to each other and only one or two scenarios behave slightly differently. In the stochastic generator the parameter inference process favors the majority behavior as the data drives the process. Therefore, when generating new radiation traces it is more likely produce scenarios which are similar to the majority of the baseline scenarios rather than the one(s) with outlier behavior. For this reason, it may happen that a baseline ensemble member is outside of the generated ensemble band at few time steps. In addition to this, it should be noted that the more synthetic scenarios we produce with the stochastic generator, the larger the ensemble band will become since it can better cover the parameter space. Despite these facts, the generated ensemble has an uncertainty band which

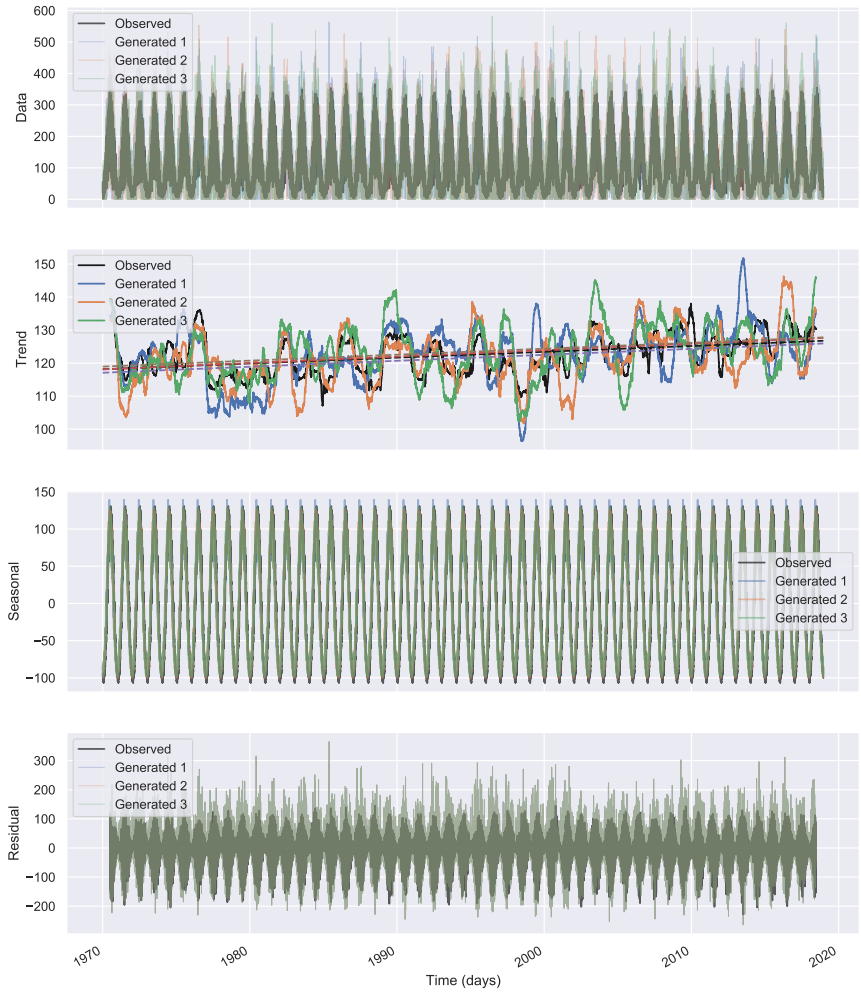


Figure 4.12: Time series decomposition of observations (black), and three example generated scenarios (colored) for the time interval 1970 - mid 2020. The first panel depicts the time series, while the panels below show their trend, seasonal and residual components, respectively.

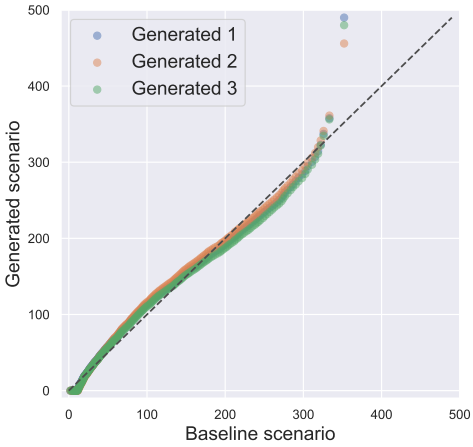


Figure 4.13: Quantile-quantile plots of three example generated scenarios compared to their baseline scenario (2006-2090).

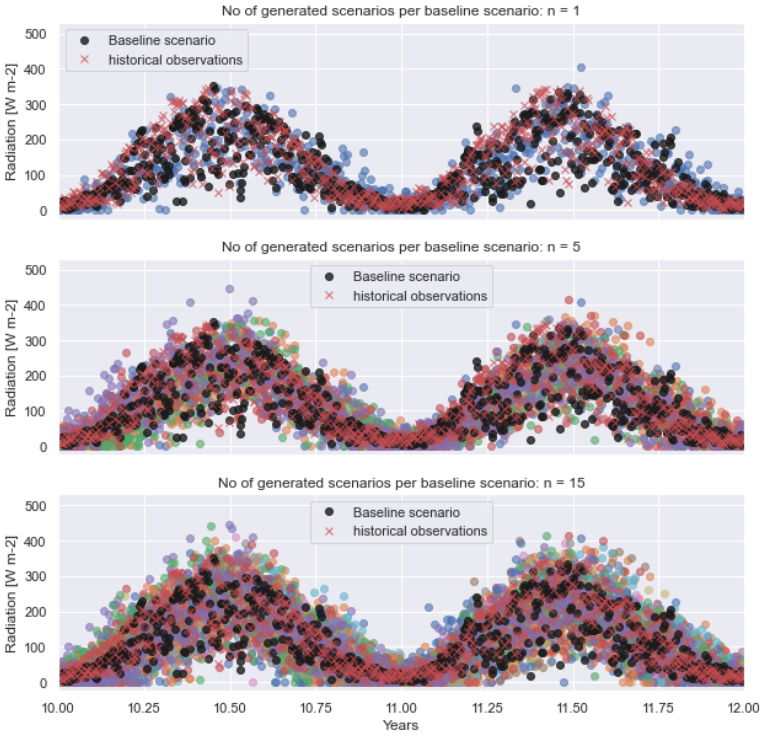


Figure 4.14: Comparison of one baseline scenario (black), observations (red x) and the new generated scenarios (colored). The number of generated scenarios is 1, 5, 15 respectively.

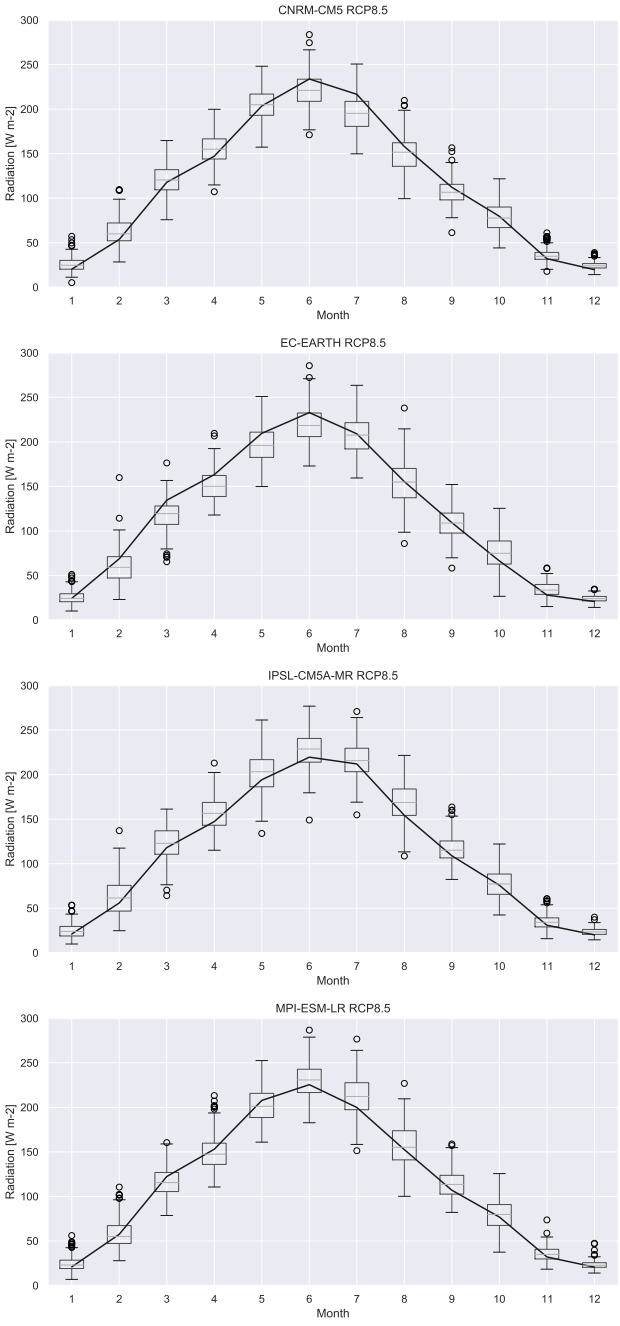


Figure 4.15: Boxplots of generated scenarios with RCM climatological mean (solid black line) per calendar month for the entire projection period (2006-2090). Only RCP 8.5 is shown for the four driving GCMs.

covers well the baseline ensemble members.

Regarding climate change impacts, we can observe from the baseline scenarios that the characteristic spring bloom of the end of the century (2081-2090) is consistently lower than the one representing the beginning of the century (2006-2015). Generated scenarios accurately reproduce this phenomena. This finding is in line with physical expectations since radiation projections show mild negative long-term trend for almost all scenarios (second panel of Figure 4.11), and during the energy limited spring period radiation positively correlates with the chlorophyll-a concentration. It should be noted, however, that in this experiment we only consider the effect of radiation and assume all other climate forcing, such as temperature, unchanged. Consequently, this demonstration does not replace comprehensive climate change impact studies but rather showcases a possible use of the radiation generator.

In order to demonstrate the benefit of a larger ensemble, Figure 4.17 depicts the histogram of the pointwise predictive distribution at the time of the characteristic spring peak concentration. One can argue that the eight baseline ensemble members may be used to derive an ensemble mean and width of the ensemble band (or spread), but not for full uncertainty characterization which also includes the predictive distribution. Looking at the basic uncertainty metrics the baseline ensemble has a mean of 25.42 mg/m^3, standard deviation of 3.7 mg/m^3, and 11.33 mg/m^3 wide uncertainty band. The generated ensemble has comparable metrics with 25.58 mg/m^3 mean, 3.13 mg/m^3 standard deviation, and 14.13 mg/m^3 uncertainty band. While the basic metrics remain similar the added value is that the larger ensemble permits us to derive predictive distribution and better express confidence in the predictions. Having only few ensemble members reduces the ability to resolve the unknown probability distribution that one tries to estimate, hence, higher number of ensemble members providing sufficient resolution in terms of probabilities is required [123].

4.6. CONCLUSIONS AND RECOMMENDATIONS

This chapter presents an approach to complement existing regional climate projections by generating new synthetic scenarios with similar statistical properties. Due to the Bayesian hierarchical (multi-level) setup the proposed method offers flexibility and allows full characterisation of uncertainties. Thus, the main value of the proposed methodology is that we can compute predictive uncertainty conditional on all the data (considering all scenarios).

Moreover, the pre-processing step allows adaptability to other climate variables, such as temperature, or potentially to other environmental variables, noting that adjustments to the model formulation might be necessary as the current time series model was defined for time series with seasonality. The underlying parameterized time series model formulation therefore needs to be adjusted for non-seasonal signals with substantially different characteristics.

In addition, there is a practical limitation to the number of generated scenarios in cases when probabilistic simulations are performed using computationally expensive physical models. The three dimensional physical model, used in the demonstration case, covers a large spatial domain, hence, simulation times are long (approx. 12 hours for one year simulation on a medium performance baremetal Linux computer cluster).

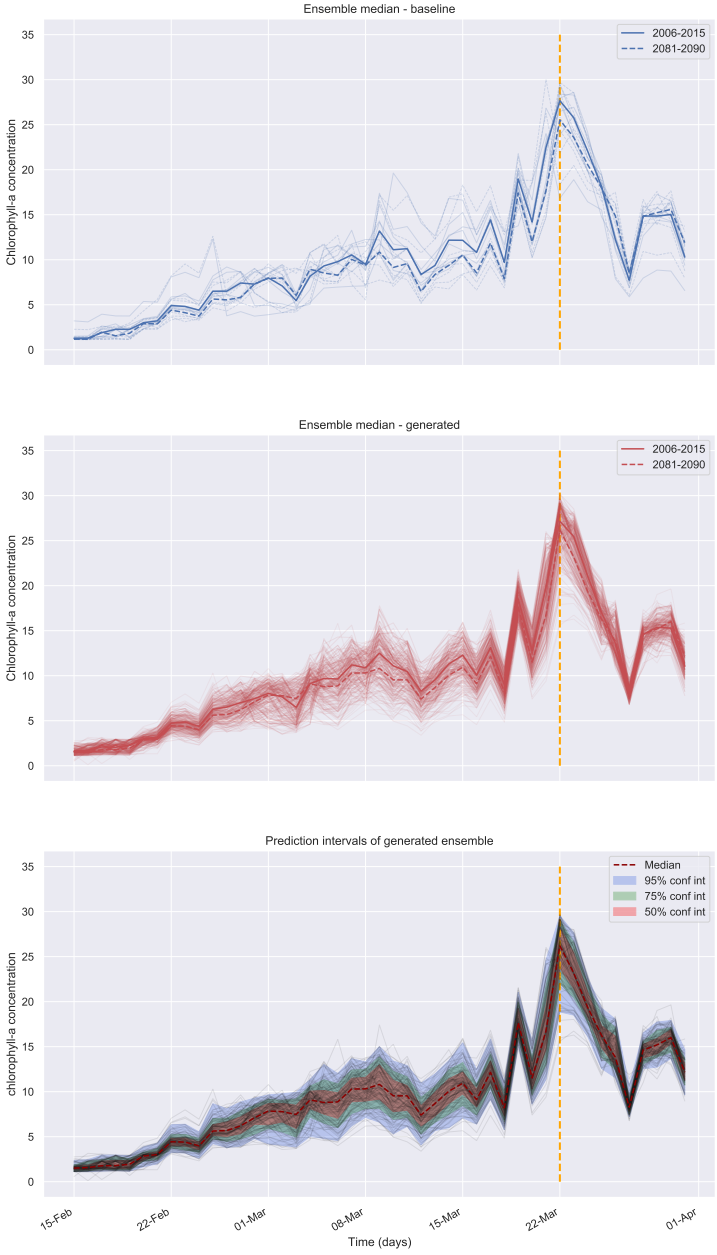


Figure 4.16: Simulated chlorophyll-a concentrations. Ensemble members and median of the baseline scenarios for the first and last simulated decades (top), ensemble members and median of the generated scenarios for the first and last simulated decades (middle), and pointwise prediction intervals derived from the generated ensemble (bottom). Orange dashed line indicates the time of the spring peak concentration.

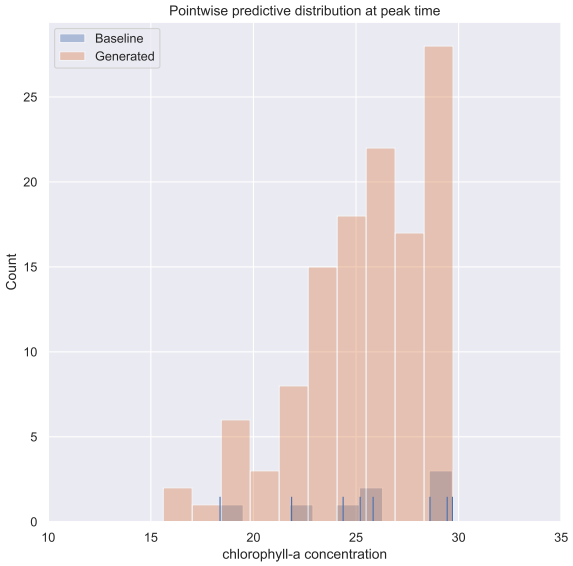


Figure 4.17: Histogram of baseline ensemble members (blue), and pointwise predictive distribution of generated ensembles (orange) at the peak time. Baseline ensemble members are also represented by blue marks (zero-width bins) along the x-axis.

Based on this we can conclude that while the stochastic generator itself has no computational time limitations, the subsequent model that is utilized to forward propagate uncertainties may present such limitations. On the other hand, if the synthetic radiation scenarios are used as input to surrogate models, this limitation on the number of scenarios is reduced.

Moreover, for future research the authors recommend to extend the current Bayesian hierarchical model to include spatial correlation (multi-site stochastic generator) and to incorporate other climate variables, since currently only one location and variable is considered. Extending the stochastic generator in this way would allow us to make use of the multi-dimensional data structure.

Finally, we conclude that the demonstration case, in which the generated synthetic radiation scenarios were utilized for probabilistic water quality simulation, could showcase the potential of the presented approach to express future likelihoods of predicted chlorophyll-a concentrations via pointwise predictive distributions. Since with smaller ensembles one may only derive ensemble mean and spread as a proxy of uncertainty, it is this added feature of simulating numerous chlorophyll-a concentration scenarios and subsequently deriving the pointwise predictive distribution, which helps to achieve better characterization of uncertainties. This enhanced uncertainty estimate in turn supports better informed and rational decision making which often brings socio-economic and monetary benefits.

5

CLIMATE CHANGE INDUCED TRENDS AND UNCERTAINTIES IN PHYTOPLANKTON SPRING BLOOM DYNAMICS

This chapter aims to study the evolution of phytoplankton bloom dynamics in the Dutch coastal waters, using a variety of historical data and projected future solar radiation and air temperature trajectories from regional climate models as driving forces covering the 21^{st} *century. The main objective is to quantify climate induced uncertainties in future coastal phytoplankton phenology stemming from important climate variables. The three main methodological steps to achieve this goal include (1) developing a data fusion model to interlace coastal in-situ measurements and satellite chlorophyll-a observations into a single multi-decadal signal; (2) applying a Bayesian structural time series forecasting model to produce long-term prediction of chlorophyll-a concentrations over the* 21^{st} *century; and (3) developing a feature extraction method to derive the cardinal dates (beginning, peak, end) of the spring bloom to track the historical and the projected evolution of its dynamics. Research findings indicate that at the study site location the spring bloom characteristics are impacted by the changing climatic conditions. Towards the end of the* 21^{st} *century climatic factors may shift spring blooms slightly earlier, resulting in longer spring bloom duration. Moreover higher chlorophyll-a concentration peaks can be expected. Based on the ensemble simulation the largest uncertainty lies in the timing of the spring bloom beginning and -end timing, while the peak timing has less variation.*

Parts of this chapter have been published in Frontiers in Marine Science - Global Change and the Future Ocean (2021) [149]

5.1. INTRODUCTION

Phytoplankton and their seasonally occurring blooms are vital to marine ecosystems as they are a major source of energy input for higher trophic levels [183]. Phytoplankton blooms are natural phenomena occurring when phytoplankton growth exceeds the losses (mortality, respiration, feeding, sinking and dispersive losses) and rapid accumulation takes place when optimal abiotic and biotic conditions are present for the growth. An early account of the bloom phenomenom is given by [194]. Phytoplankton blooms can be identified through chlorophyll-a concentration, which is an indicator for algal biomass, though concerns were raised [10] about using chlorophyll as phytoplankton biomass proxy in the North Sea. In the Dutch coastal zone, phytoplankton mass seasonality is described by a prominent spring bloom (diatom dominated) and a less pronounced late summer bloom. This is partly driven by increased riverine nutrient loads (melting snow and spring rains) and intensified mixing by seasonal winds blowing over the shallow shelf sea. The onset of spring blooms is usually initiated by correlated changes in water temperature and the light availability [217] but coupled to and controlled by thermal stratification, resource dynamics (e.g. nutrient availability) and predator–prey interactions (e.g. grazing) [22]. Temperate marine environments, such as the Dutch coastal waters, are particularly sensitive to changes in spring bloom initiation due to the fact that higher trophic levels are greatly dependent on synchronized planktonic production [64].

When studying the functioning of continental shelf ecosystems, such as the southern North Sea, one should consider various influencing elements. Regarding the hydrodynamics, the southern North Sea is a tidally mixed region where tidal fronts occur across the English Channel. The variability in the tidal fronts influence stratification and mixing regimes and have ecological consequences, or may even be the driving force of regime shifts in the North Sea ecosystem [129]. In addition to tidal fronts, along the Dutch coast, other shallow water (e.g. Wadden Sea), coastal, and estuarine fronts are impacting the system dynamics. These fronts are characterized by turbidity and salinity gradients. Since the study location is situated at the boundary of the North Sea and the shallower Wadden Sea, in the Mardiep tidal inlet, the coastal influence is an important factor. In the Dutch coastal zone the observed gradients of phytoplankton biomass are very steep and there is considerable natural variability in the chlorophyll-a concentration. In these shallower coastal waters the concentration of suspended inorganic matter, which influences the extinction of light, is relatively high and dynamically varying. According to [131] in Dutch coastal waters 25 to 75% of the light extinction is caused by suspended matter. Further coastal influencing factor affecting the spring bloom is the riverine nutrient loads. In the North Sea rivers provide a significant portion of the total nitrogen and phosphorus load [132]. Although the study site is not situated at a river outflow, there are nine major rivers that affect the Dutch coastal waters based on the nutrient composition matrix derived by [132]. The plumes of these major effluents, especially the Rhine, are significant influencing factors to phytoplankton dynamics.

Available climate models offer us a range of (atmospheric) climate variables that could be considered as external drivers influencing phytoplankton seasonality. The climate variables include air temperature, precipitation, solar radiation, eastward and northward wind, air pressure, humidity, and cloud cover. In this study we focus on air temper-

ature and solar radiation that were found [148] to be the most influential atmospheric variables affecting coastal chlorophyll-a concentrations in the Dutch coastal waters, along with wind speed (in shallow systems). This conclusion was reached by applying various statistical techniques to explore temporal, spatial, and functional correlations from the historical atmospheric and chlorophyll-a time series at this location.

In its recent comprehensive study of the Wadden Sea eutrophication trends, [25] lists the phytoplankton governing factors, both bottom-up (light, nutrient) and top-down (grazing, filter feeding). Through the review of various studies, it was concluded that light is the dominating limiting factor, which is present all year long, while nutrient limitation occurs during summer and toward the end of the growth season. Moreover, a cross correlation analysis was conducted by [31] in the North Sea between environmental variables (tidal mixing, wind mixing, solar radiation, air temperature, SST, salinity, turbidity) and chlorophyll-a hourly time series, including various lags. At the site with dynamics similar to our study area, the highest correlations were found with solar radiation, air temperature, turbidity and tidal mixing. Additionally, [105] reports that sea surface temperature is the best predictor of chlorophyll-a concentration in the North Atlantic. In their climate impact study, [169] also opted to use only mean annual sea surface temperature as an environmental driver since it acts as a useful proxy for other physical processes and influences seasonal and regional changes in vertical stratification, nutrients, and winds. We should also note that there is relationship between air temperature, solar radiation and mixing. [31] indicated that in the North Sea air temperature and solar radiation influences phytoplankton biomass through diurnal variation in convective mixing and diurnal vertical migration of motile phytoplankton. Supporting this, [207] reported that the diurnal variation in convective mixing is attributed to the sinking of phytoplankton during daytime (thermal micro-stratification) and resuspension at night (surface cooling). [105] also confirmed that temperature is correlated with stratification, mixed layer depth and nutrient availability and their temporal changes.

The thermal structure of the North Sea as a whole is characterized by a well-developed thermocline during summer and well-mixed water column during winter [85]. Nevertheless, there are important regional differences. In the central North Sea the water column can be strongly stratified and the tidal-induced mixing is less important. In these regions wind-driven mixing and convective cooling have a greater impact on phytoplankton biomass [31]. This seasonally stratified condition is in stark contrast with the highly dynamic coastal systems where tidal mixing is the most dominant physical factor. [143] also documented important differences between the offshore and coastal North Sea regarding the impact of climatic conditions and nutrient availability. It was found that inter-annual variability in phytoplankton dynamics of the offshore regions was mainly regulated by temperature, Atlantic inflow, as well as co-varying wind stress and North Atlantic Oscillation (NAO). Contrarily, in coastal waters solar radiation and sea surface temperature, as well as Si availability was dominant [143]. In addition to the regional differences, the influence of environmental drivers of phytoplankton biomass also differs at different temporal scales [31]. At short time scales, the physical transport of phytoplankton cells by wind-driven or tidal mixing is the dominant. On the other hand, focusing on the seasonal time scales it is solar radiation and air temperature, together with associated changes in thermal stratification, nutrient availability and grazing, that

dominate phytoplankton dynamics [194, 189, 31]. Finally, at longer inter-annual and decadal time scales climatic variation and long-term human impacts on the eutrophication status will become influential [169, 31]. Consequently, we acknowledge that in other regions physical processes play a dominant role in coastal chlorophyll-a concentrations, especially through the mixing (e.g. wind-driven) of nutrients into the euphotic layer during stratified conditions. Although this is particularly important in oligotrophic regions where solar energy is abundant and phytoplankton dynamics is mainly limited by nutrient availability [223], it is less influential in our case.

Our study is motivated by the fact that climate-induced regime shifts reportedly took place in the North Sea [9, 18]. Consequently, seasonal variability of phytoplankton biomass in relation to light and temperature is particularly important aspect in the North West Shelf Seas [127, 202]. The interactive effects of temperature and solar irradiance on phytoplankton have been extensively studied without clear consensus. This may be partly due to the fact that phytoplankton response to temperature change greatly varies between individual and aggregate level. Considering the individual level phytoplankton responses to temperature are exponentially or linearly increasing until the optimum, and declining above that [63]. On the other hand, looking at the aggregate level, species can replace one another along a temperature gradient via competition resulting in monotonically increasing growth rates. However, temperature also influences predator-prey interactions, not only phytoplankton growth. The intensity of grazing (or zooplankton ingestion) is partly determined by temperature, along with the available phytoplankton biomass and the zooplankton biomass [200].

Due to the complex interactions of physical forcing conditions with food web processes, phenological responses of phytoplankton to climate change are not trivial to estimate. Nevertheless, according to [172], focusing on the spring season may help to reduce the complexity. It was suggested that in temperate marine systems the impact of physical environment and the response of the biological system can be best studied in spring. During spring, the physical limiting factors like temperature, light availability, and mixing are more prominent than the non-physical ones, such as trophic interactions (e.g. grazing). While in the spring period trophic interactions may not be limiting, later on in the year, they become more important and may dominate over the physical factors [187, 189]. Thus, we acknowledge the complexity of physical and trophic interactions and do not dismiss their influence on the phytoplankton phenology. Nevertheless, this study aims to focus on the physical drivers, or more precisely on the climatic ones. Consequently, to limit the masking effect of trophic interactions, as far as this may be possible, we focus on the spring phytoplankton bloom to study the impact of changing climatic conditions in the Dutch coastal zone.

Changing climatic conditions directly affect the photosynthetic metabolism of phytoplankton, but also indirectly impact them by modifying their physical environment [52]. Climate change impacts on phytoplankton are manifested as shifts in seasonal dynamics, species composition, and population size structure [217]. Since in the current study we only use chlorophyll-a concentration as response variable, we can only draw conclusions on the seasonal dynamics of the aggregate level, not on species composition or population structure. As an indicator of climate change impacts on seasonal phytoplankton dynamics, we selected the long term changes in spring bloom dynam-

ics. There is, however, no single definition of phytoplankton blooms in the literature or in policies, for instance based on the rate of change or the threshold of concentration, as this is highly dependent on the type of ecosystems (e.g. inland or marine, local species, climate, bathymetry). In this study we describe the spring bloom dynamics by their cardinal dates (bloom initiation, -peak and -ending) using log-concave regression. Alternatives methods of deriving cardinal dates and the benefits of using log-concave regression are presented in the Section 5.2.4.

A range of studies investigating climate change induced shifts in phytoplankton bloom dynamics in the North Sea already exist. Most of these studies derive their findings from historical chlorophyll-a data, measured either by in-situ sensors or remote sensing [59, 100, 161, 64, 73], or from laboratory experiments [124, 218]. Climate impact studies which focus on future developments of phytoplankton bloom dynamics generally use few climate change scenarios from global or regional climate models and traditionally use physically-based models [74, 101, 102, 163, 179]. We acknowledge that previous papers already introduced ways to characterize phytoplankton blooms [215, 161, 100, 172, 124]. Nevertheless, uncertainty quantification in the shift of phytoplankton dynamics in these studies is not a central topic.

There are, however, existing studies that address uncertainty in bloom detection. [49] investigates the impact of missing data on phytoplankton phenology metrics (threshold-based definition) using satellite observed chlorophyll-a; [68] compares the accuracy and precision of three bloom metrics (biomass-based threshold method, cumulative biomass-based threshold method, rate of change) on biogeochemical model outputs and satellite observed chlorophyll-a; while [84] performs probabilistic phytoplankton phenology characterisation using Bayesian harmonic regression and a threshold-based definition of bloom metrics based on satellite observed chlorophyll-a. Major advantage of these studies is the quantification of errors or uncertainties in the computation of the bloom metrics. Our research deviates from these studies in that we do not focus on historical data but aim to quantify future projected uncertainties in spring bloom dynamics. In fact, in our analysis the bloom detection algorithm is the only step where "model uncertainties" are not quantified and instead all other steps involve uncertainty estimates. The reason for this is that in future climate change studies the main source of uncertainty does not arise from the derivation of the bloom metrics but from the climate forcings and from the projection of the chlorophyll-a signal. Our method does provide uncertainty ranges for the bloom metrics but that is derived from the ensemble of generated chlorophyll-a projections. The benefit of reconstructing a range (> 100) of full seasonal cycles is therefore to obtain predictive uncertainty estimates on bloom metrics from the input data rather than from the bloom detection itself.

Considering the above, the novelty of our work lies in the following features. In our research we make use of both in-situ and satellite observations jointly by applying a data fusion algorithm to get a more complete, more accurate and longer data record. While a range of possibilities already exist to describe phytoplankton blooms, in our research we propose a new way of extracting the cardinal dates of the phytoplankton spring blooms. We use non-parametric shape constrained (log-concave) regression, which provides a flexible formulation without tuning parameters and assumptions on the distribution patterns and can be directly applied on the annual bi-modal time series without any

pre-processing. Consequently, our proposed method is less sensitive to bloom amplitude, missing data, and observational noise.

Moreover, we augment existing climate change scenarios with synthetically generated ones, thus supplying numerous (> 100) trajectories for air temperature and solar radiation development. In addition to this, our proposed method complements the computationally expensive numerical models for chlorophyll-a simulation with a data driven approach, using a Bayesian structural time series model. Complementing physically-based prediction models with statistical ones allows us to compute a large number of simulations and achieve better characterization of predictive uncertainties. These methodological advances enable the combination of different chlorophyll-a data sources, the incorporation of climate covariates and the propagation of uncertainty from observations to nonlinear estimates of projected changes in spring bloom metrics under an enriched number of climate change scenarios (associated to future development and emission pathways).

5.2. MATERIALS AND METHODS

In this chapter we describe the data sources and introduce the main methods that were developed and/or applied within the framework of this study. When new methods are proposed, such as the data fusion model and the shape constraint model to derive bloom metrics, we aim to sufficiently document those to allow replication studies.

Figure 5.1 presents the methodological framework and summarizes the connections between elements. Our research aims to study changes in phytoplankton phenology based on historical data and future climate projections. Given the historical records of chlorophyll-a concentrations obtained from various data sources, one can extract the cardinal dates of the spring bloom for the past decades using the proposed feature extraction technique. Furthermore, changes in the spring blooms may be projected for the future by utilizing the correlation between climatic factors, represented by air temperature and solar radiation, and the ecological response, indicated by the chlorophyll-a concentration. This correlation can be inferred from past records since air temperature and solar radiation were measured by field sensors for the past decades. Though future chlorophyll-a concentrations are not available to us, we attempt to make projections using the trends and seasonality from historical observations and taking into account the correlations with projected air temperature and solar radiation, produced by regional climate models. While this methodological framework allows us to investigate past and projected spring bloom dynamics, we note that there are several sources of uncertainties, both data and model related ones, which are propagated through the steps. These uncertainty sources (+/- U) are marked in Figure 5.1. In order to address this issue, we aim to use transparent statistical approaches that allow us to quantify intrinsic uncertainties. Noting that the projected trends in bloom metrics constitute the main findings of the research, the importance of the uncertainty quantification framework should also be emphasized, which should always go hand-in-hand with climate change impact studies.

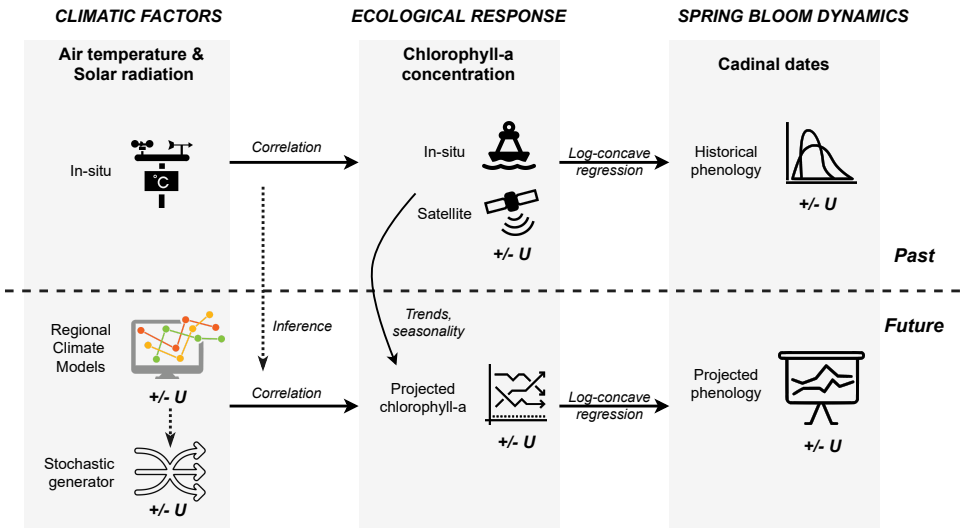


Figure 5.1: Methodological framework including three main elements with causal and temporal relations: (1) climatic factors, (2) ecological response, and (3) spring bloom dynamics

5

5.2.1. DATA SOURCES

This research is based on a multitude of data sources from sensors and numerical models of various types. The environmental and climate variables in this study are chlorophyll-a concentration, air temperature, and solar radiation. In order to investigate past trends and obtain the correlation between these variables, we make use of historical measurements, whereas to anticipate future climate change impacts, climate model outputs are used.

CHLOROPHYLL-A CONCENTRATION MEASUREMENTS

Available historical chlorophyll-a data includes field observations at Marsdiep Noord station (see Figure 5.2), from the Dutch Directorate-General for Public Works and Water Management (Rijkswaterstaat), covering more than forty years from 1976 to 2018, but measured rather sparsely. To complement these field measurements, processed and validated satellite observed chlorophyll-a concentration (extracted at the same location) was used from the Copernicus Marine Environment Monitoring Service (CMEMS) from 1997 to 2019. We should note that satellite observation of phytoplankton biomass in the Dutch coastal waters is complex since the chlorophyll-a signal may be mixed with the relative distribution of suspended matter and CDOM instead of phytoplankton biomass [129].

The specific product in use is the North Atlantic Chlorophyll-a, daily interpolated and reprocessed product with one km spatial resolution (OCEANCOLOUR_ATL_CHL_L4_REP_OBSERVATIONS_009_098). The satellite product is limited to the surface depth. This chlorophyll-a product is produced using multiple sensors (multi-sensor product), multiple chlorophyll-a algorithms and a daily space-time interpolation scheme [178]. The interpolation scheme includes a combination of a water-typed merge of chlorophyll-

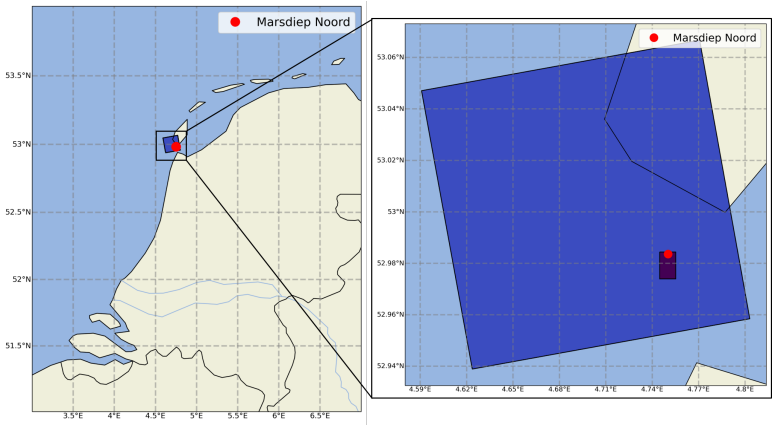


Figure 5.2: Location of the study area and the monitoring point together with the pixels of the matching Euro-CORDEX climate model output and CMEMS satellite measured chlorophyll-a

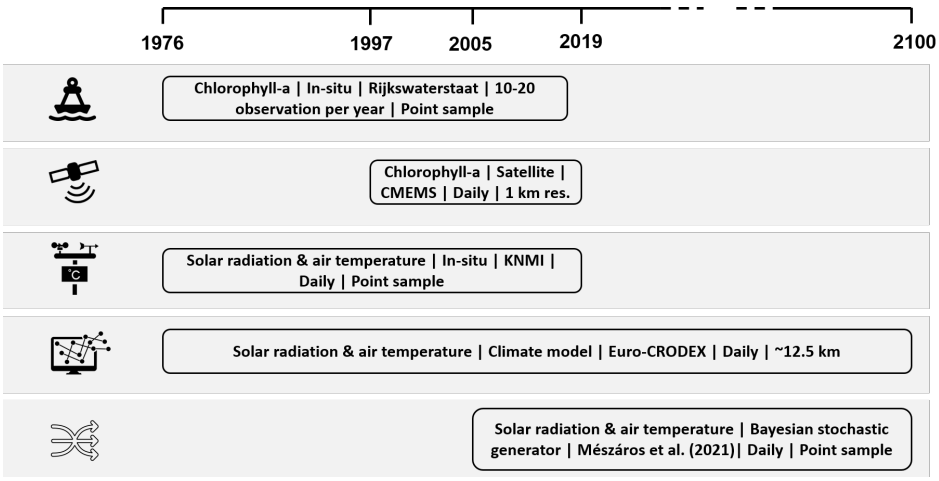


Figure 5.3: Overview of data sources. The description includes variable name, data type, data source, data frequency, and spatial resolution.

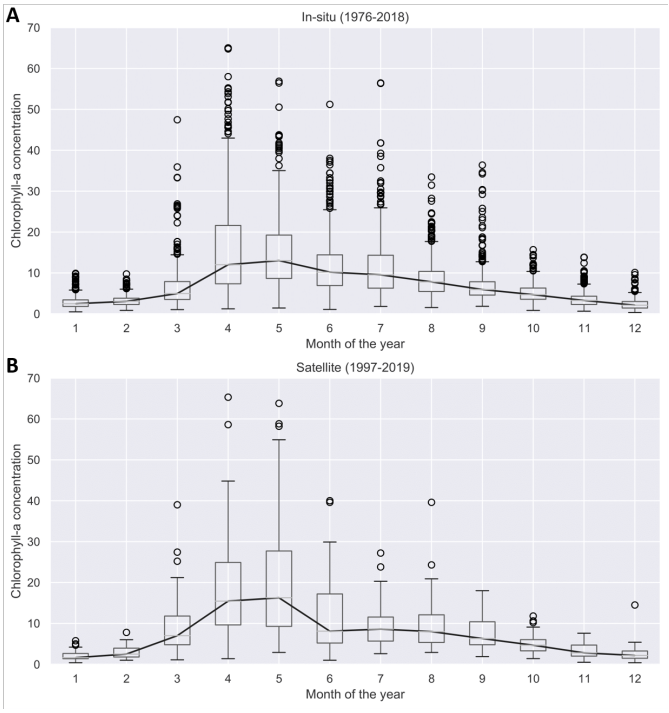


Figure 5.4: Historical chlorophyll-a concentrations measured in the Dutch Wadden Sea using in-situ data between 1976-2018 **(A)** and satellite images between 1997-2019 **(B)**. Climatological median (solid black line) per calendar is also shown.

a estimates and kriging interpolation method with regional anisotropic covariance models at the shore, as described in [178]. This product uses the Copernicus-GlobColour processor and it is obtained by merging the following sensors: SeaWIFS, MODIS Aqua, MODIS Terra, MERIS, VIIRS NPP, VIIRS-JPSS1 OLCIS3A and S3B. For coastal waters the product uses the standard OC3-OC4 [13, 156, 155] and OC5 [81] algorithms. The latest product validation results against in-situ measurements show an r^2 of 0.73 with N = 11502 data points [77]. For a more in-depth description of this satellite product the reader is referred to the QUality Information Document (QUID) [77].

The chlorophyll-a concentration seasonality from in-situ observation is shown in Figure 5.4A, and from satellite observations in Figure 5.4B. Naturally these data sources have different sampling methods and associated uncertainties. The in-situ observations are point samples taken by the Dutch national in-situ monitoring programme (MWTL) `https://waterinfo-extra.rws.nl/monitoring/`. It should be noted that the samples are taken close to the water surface, usually in the upper 3-5 meters of the water column. These observations are often considered as ground truth and are the most reliable, however, in the case of chlorophyll-a concentration the temporal frequency of the observations is relatively low, around 10-20 observations per year. This amount of field observations poses a limitation to assess annual phytoplankton bloom cycles [216]. Thus,

the more frequently sampled satellite images are also used to complement the in-situ measurements for a better assessments of bloom characteristics. This complementary data source is used noting that satellite derived chlorophyll-a is only available at the water surface (lack of vertical resolution), has a coarse one km resolution and suffers from algorithmic and interpolation errors, consequently having a higher level of associated uncertainty.

Since the two types of chlorophyll-a measurements describe the same underlying process, we propose a data fusion model to combine them. This data fusion model interlaces the in-situ and satellite observations into a single chlorophyll-a concentration signal, which is more complete then the individual observations and covers a longer time period. The data fusion model is described in Section 5.2.2.

SOLAR RADIATION AND AIR TEMPERATURE MEASUREMENTS

The historical daily solar radiation and air temperature records are obtained at the nearest weather station (De Kooy) from the Royal Netherlands Meteorological Institute (KNMI) for the matching period (1976-2019). Apart from historical data, future projected values of air temperature and solar radiation are acquired from the high resolution 0.11 degree (~ 12.5 km) EURO-CORDEX Coordinated Regional Downscaling Experiment [106], which uses the Swedish Meteorological and Hydrological Institute Rossby Centre regional atmospheric model (SMHI-RCA4). In order to produce various regionally downscaled scenarios, EURO-CORDEX applies a range of General Circulation Models (GCMs) to drive the above mentioned Regional Climate Model (RCM). In addition to the driving models, further scenarios are obtained by considering different socio-economic changes described in the Representative Concentration Pathways (RCPs). RCPs are labeled according to their specific radiative forcing pathway in 2100 relative to pre-industrial values. The EURO-CORDEX scenario simulations use the RCPs defined for the Fifth Assessment Report of the IPCC. In this study we include RCP8.5 (high), and RCP4.5 (medium-low) [214] and four driving GCMs.

In the upcoming Sixth Assessment Report new scenarios and pathways will also be included, which are called Shared Socioeconomic Pathways (SSPs) [2]. SSPs describe five alternative socioeconomic pathways (SSP1 to SSP5) for future society enhancing the existing RCPs with socioeconomic challenges to adaptation and mitigation. Such socioeconomic challenges are population, economic growth, urbanisation or technological development for instance [154]. It should be emphasized that SSPs are not replacing but complementing RCPs. In the Sixth Assessment Report the RCP-based climate projections and SSP-based socioeconomic scenarios are combined to achieve an integrative framework for climate impact and policy analysis [2]. From the SSP scenarios SSP5-8.5 corresponds to RCP8.5 and represents the high end of the range of future forcing pathways, while SSP2-4.5 represents the medium part and corresponds to RCP4.5 [1].

Together the four different driving GCMs and two RCPs that are applied in this study provide us with an ensemble of eight future solar radiation and temperature trajectories. Since the RCM simulations are subject to climate model structural error and boundary errors from the driving GCMs [153], they should be bias corrected before applying them in impact studies [136]. For this reason, quantile mapping bias correction [11] was applied using the RCM simulations for the reference period (1976-2005) and daily historical field measurements from KNMI for the same period, as described in [148]. The

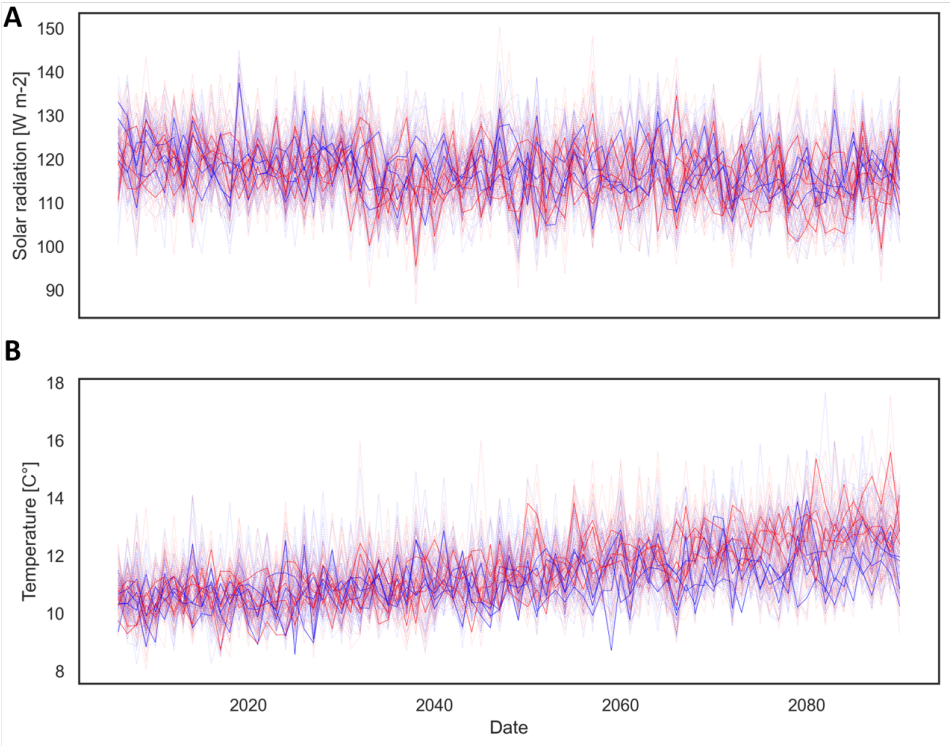


Figure 5.5: Eight EURO-CORDEX (darker solid line) and 120 generated synthetic (shaded dashed line) climate change projections for solar radiation **(A)** and air temperature **(B)**, grouped by RCP scenarios (blue - RCP4.5, red - RCP8.5). Plot of the yearly averages based on the daily data.

quantile-quantile mapping transfer functions were established for the reference period and separately for each RCM simulation. The transfer functions were then applied for the bias correction of each future projections (2006-2100) separately.

This ensemble of climate trajectories is used to simulate a range of possible phytoplankton seasonality shifts and the associated uncertainty described by the predictive distribution of the phytoplankton bloom cardinal dates. It should be noted that applying only eight climate projections reduces the ability to adequately resolve the unknown predictive distribution that one tries to estimate, hence, higher number of climate trajectories providing sufficient resolution in terms of probabilities is required [123]. Consequently, to better characterize uncertainties, an enriched set of climate change projections is employed. This set of air temperature and solar radiation projections was produced using a Bayesian stochastic generator [148], which builds on the above mentioned Regional Climate Model scenarios provided by the EURO-CORDEX experiment and generates further synthetic scenarios using a hierarchical Bayesian model. The generated ensemble of air temperature and solar radiation projections include 120 members and their statistical properties are similar to the input projections. Both the EURO-CORDEX and synthetic projections are shown for air temperature in Figure 5.5A and for solar ra-

diation in Figure 5.5B. At this specific location we can observe a consistently increasing temperature trend over the $21st$ century and a slightly decreasing solar radiation trend. While increasing air temperatures are in line with expectations, decreasing solar radiation trends may need further explanation. The main cause of this negative trend is the fact that total cloud cover at this site is projected by EURO-CORDEX to increase, hence, limiting surface downwelling shortwave radiation. This is a region specific feature, and the difficulty of projecting cloud cover and solar radiation changes in coastal areas with sea-land-atmosphere boundaries, such as the study site, has been previously highlighted by [17], along with discrepancy between RCMs and their driving GCMs in their solar radiation projections over Europe.

5.2.2. DATA FUSION OF CHLOROPHYLL-A MEASUREMENTS

STATISTICAL MODEL

In order to describe the chlorophyll-a concentration, we assume that there is a continuously evolving latent signal $(X_t,\, t \in [0,T])$ that satisfies the stochastic differential equation (sde)

$$\mathrm{d}X_t = -\alpha(X_t - \mu(t))\,\mathrm{d}t + \sigma\,\mathrm{d}W_t. \tag{5.1}$$

The underlying idea is to model a stochastic process that is mean reverting (with strength α) towards the deterministic signal $t \mapsto \mu(t)$. We will take μ to be periodic with period 1. We start off from a continuous time description as in-situ measurements are not collected at regular times. Observations can be of three types

1. $Y_i \sim N(X_{t_i}, \psi_1)$;
2. $Y_i \sim N(X_{t_i}, \psi_2)$;
3. $Y_i \sim N_2\left(\begin{bmatrix}1\\1\end{bmatrix} X_{t_i}, \begin{bmatrix}\psi_1 & 0\\ 0 & \psi_2\end{bmatrix}\right)$.

This reflects having two types of measurements (in-situ and satellite) with different accuracies. Sometimes one measurement is obtained, sometimes the other one, and sometimes both are available. We take Y_i to be the log of the measured concentration (componentwise) to ensure the model only predicts non-negative concentrations. While we acknowledge that there are other mapping functions to achieve non-negativity, taking the log of chlorophyll-a concentration is often used in practice [41].

Assuming successive observations are obtained closely in time, i.e. $\Delta_i := t_i - t_{i-1}$ being small for all i, we have

$$X_{t_i} \approx X_{t_{i-1}} - \alpha(X_{t_{i-1}} - \mu(t_{i-1}))\Delta_i + \sigma\sqrt{\Delta_i}\epsilon_i,$$

where $\{\epsilon_i\}_i$ is a sequence of independent standard Normal random variables. Ignoring discretisation error, the resulting equation can be rewritten and combined with the observation scheme:

$$\begin{aligned} X_i &= (1-\alpha\Delta_i)X_{i-1} + \alpha\mu(t_{i-1})\Delta_i + \sigma\sqrt{\Delta_i}\epsilon_i \\ Y_i &= N(L_i X_i, \Upsilon_i), \end{aligned}$$

where $X_i \equiv X_{t_i}$. For numerical stability, it is better to discretise (5.1) using an implicit scheme on the deterministic part. This leads to the dynamical system

$$X_i = \frac{X_{i-1} + \alpha\mu(t_i)\Delta_i}{1+\alpha\Delta_i} + \sigma\sqrt{\Delta_i}\epsilon_i$$
$$Y_i = N(H_i X_i, R_i),$$

We write the model in state-space form, sticking to the notation in Särkka [177]

$$\begin{aligned} X_i &= A_{i-1}X_{i-1} + a_{i-1} + N(0, Q_{i-1}) \\ Y_i &= H_i X_i + N(0, R_i) \end{aligned} \tag{5.2}$$

Here

$$A_{i-1} = (1+\alpha\Delta_i)^{-1} \qquad a_{i-1} = \frac{\alpha\Delta_i}{1+\alpha\Delta_i}\mu(t_i) \qquad Q_{i-1} = \sigma^2\Delta_i,$$

$$R_i = \begin{cases} \psi_1 & \text{if only in-situ measurement} \\ \psi_2 & \text{if only satellite measurement} \\ \begin{bmatrix} \psi_1 & 0 \\ 0 & \psi_2 \end{bmatrix} & \text{both in-situ and satellite measurements} \end{cases}$$

and

$$H_i = \begin{cases} \begin{bmatrix} 1 \end{bmatrix} & \text{if only 1 measurement is available at time } t_i \\ \begin{bmatrix} 1 & 1 \end{bmatrix}' & \text{if both measurements are available at time } t_i \end{cases}.$$

Note that (5.2) specifies a linear Gaussian state-space model. The equation for Y is the observation equation, that for X the state-equation. We will parametrise ψ_1, ψ_2 by taking

$$\psi_1 = \eta\bar{\psi}\psi \qquad \psi_2 = \psi,$$

where $\eta \in (0,1)$ is fixed and $\bar{\psi}$ will get assigned a prior distribution supported on $(0,1)$. This reflects apriori knowledge that the in-situ measurements are believed to be more accurate. The in-situ chlorophyll-a observations are obtained from sampling campaigns (bucket water samples from a sampling jetty) and therefore considered as the true values (ground truth). While the satellite product is calibrated with many in-situ observations in the North Sea, it does not produce perfect match with the in-situ observations at the study location. Moreover, the number of satellite observations is much higher than the in-situ observations. This over-representation is counter balanced by the fusion model otherwise the reconstruction would be mostly determined by the satellite measurements.

We model the mean trend using the series expansion of the form

$$\mu(x) = \sum_{k=1}^{K} \xi_k \varphi_k(x),$$

where K is fixed, and $\boldsymbol{\xi} := (\xi_1, \ldots, \xi_K) \sim N_K(0, \sigma_\xi^2 I)$. This term allows us to account for a varying shape of the seasonal cycle. The functions φ_k are taken as follows: $\varphi_1 = \mathbf{1}_{[0,1]}$ and for $j \in \{1, \ldots, J\}$

$$\varphi_{jk}(x) = j^{-1}\varphi_0(2^{j-1}x - k), \quad \text{with } k \in \{0, \ldots, 2^{j-1} - 1\}.$$

5

We take

$$\varphi_0(x) = \frac{9}{2}x^2\mathbf{1}_{[0,1/3]}(x) + \left(\frac{3}{4} - 9(x-1/2)^2\right)\mathbf{1}_{[1/3,2/3]}(x) + \frac{9}{2}(1-x)^2\mathbf{1}_{[2/3,1]}(x),$$

which is the quadratic B-spline function scaled to have support $[0,1]$. Note that φ_0 is continuously differentiable. The hierarchical structure of the basis is exactly like the Schauder basis, but uses a smoother basic element than the traditional "hat"-function.

INFERENCE

Let $\theta = (\alpha, \xi, \sigma^2, \psi, \bar{\psi})$. Inference can be carried out by initialising θ and iterating the following steps [171]:

1. conditional on $\theta, Y_1, \ldots, Y_n$, run the Forward Filtering Backwards Sampling (FFBS)-algorithm (see Appendix) to reconstruct $X_1, \ldots, X_n$;

2. draw from the posterior of θ, conditional on $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ (note that the likelihood is simple, once we know the latent path $X_1, \ldots, X_n$).

For updating parameters we use Gibbs sampling. Note that the updates for $\bar{\psi}$ and ψ only depend on $Y_1, \ldots, Y_n$ and updates for all other parameters only depend on $X_1, \ldots, X_n$.

- The updates steps for σ^2 and ψ are trivial when using independent InverseGamma distributions as prior due to partial conjugacy.

- For $\bar{\psi}$ we assume the $Unif(0,1)$-prior. A Metropolis-Hastings step is implemented where we use random-walk type proposals [171] of the form

$$\log\frac{\bar{\psi}^\circ}{1-\bar{\psi}^\circ} := \log\frac{\bar{\psi}}{1-\bar{\psi}} + N(0, \tau^2_{\bar{\psi}}),$$

  which implies that the proposal ratio equals

$$\frac{q(\bar{\psi} \mid \bar{\psi}^\circ)}{q(\bar{\psi}^\circ \mid \bar{\psi})} = \frac{\bar{\psi}^\circ(1-\bar{\psi}^\circ)}{\bar{\psi}(1-\bar{\psi})}.$$

  Note that $\bar{\psi}^\circ = \bar{\psi}/(\bar{\psi} + (1-\bar{\psi})\tau_{\bar{\psi}}Z)$, where $Z \sim N(0,1)$.

- For updating α we use a Metropolis-Hastings step of the form $\log\alpha^\circ := \log\alpha + N(0, \tau^2_\alpha)$.

- The "full" conditional density for $\boldsymbol{\xi}$ is proportional to

$$\exp\left(-\frac{1}{2\sigma^2_\xi}\|\boldsymbol{\xi}\|^2 - \frac{1}{2\sigma^2}\sum_{i=2}^{n}\Delta_i^{-1}\left(X_i - A_{i-1}X_{i-1} - \frac{\alpha\Delta_i}{1+\alpha\Delta_i}\sum_{k=1}^{K}\xi_k\varphi_k(t_i)\right)^2\right)$$
$$= \exp\left(-\frac{1}{2\sigma^2_\xi}\|\boldsymbol{\xi}\|^2 - \frac{1}{2\sigma^2}\sum_{i=2}^{n}\left(U_i - \bar{\alpha}_i\sum_{k=1}^{K}\xi_k\varphi_k(t_i)\right)^2\right),$$

where

$$U_i = \Delta_i^{-1/2}(X_i - A_{i-1}X_{i-1}) \qquad \bar{\alpha}_i = \frac{\alpha\sqrt{\Delta_i}}{1+\alpha\Delta_i}.$$

This is proportional to

$$\exp\left(\left(-\frac{1}{2}\boldsymbol{\xi}'(\sigma^{-2}V + \sigma_\xi^{-2}I_K)\boldsymbol{\xi} + \sigma^{-2}\boldsymbol{\nu}'\boldsymbol{\xi}\right)\right)$$

with

$$\boldsymbol{\nu}_k = \sum_{i=2}^{n} U_i\bar{\alpha}_i\varphi_k(t_i) \qquad V_{k\ell} = \sum_{i=2}^{n}\bar{\alpha}_i^2\varphi_k(t_i)\varphi_\ell(t_i).$$

Hence, the update step for $\boldsymbol{\xi}$ boils down to sampling from a multivariate normal distribution with precision $\sigma^{-2}V + \sigma_\xi^{-2}I_K$ and potential vector $\sigma^{-2}\boldsymbol{\nu}$ (the potential vector is the product of the precision matrix with the mean vector).

Details on the prior specification: for both σ^2 and ψ we took (independently) InverseGamma priors, parameterized with shape and scale, with both parameters equal to 0.1. For α we took the Exponential distribution with mean 10. We took $\sigma_\xi^2 = 10$ and tuned the step-sizes τ_ψ and τ_α such that the corresponding random-walk Metropolis-Hastings steps were accepted with probability in between 25% and 50%. In the series expansion we took a fixed value for K=5. We took $\eta = 658/8005$, which is the ratio of the in-situ and satellite measurements.

5.2.3. LONG TERM PROJECTION USING BAYESIAN STRUCTURAL TIME SERIES MODELS

After the fused historical chlorophyll-a concentration signal has been derived, it is used to train the time series model for scenario analysis. It was previously argued that variability in the spring bloom dynamics occur due to changing environmental conditions. Consequently, apart from historical trends and seasonality in the observed chlorophyll-a concentration time series, projected solar radiation and air temperature are also used to drive future chlorophyll-a concentration trajectories. These simulated trajectories are then utilized to extract the bloom characteristics applying the feature extraction methodology described in section 5.2.4.

In this study an existing Bayesian structural time series modelling framework is customized to our purpose, which is the Prophet forecasting model [196]. This is a decomposable time series model with trend, seasonality and additional regressor component, as well as error term as the main model components:

$$y(t) = g(t) + l(t) + \epsilon(t).$$

where, at time t, $y(t)$ is the response variable (chlorophyll-a concentration), $g(t)$ is a piecewise linear trend model, $l(t)$ is a linear component representing seasonality and additional regressors, and $\epsilon(t)$ is the error term (independent and identically distributed noise). In order to avoid negatively predicted values, the natural logarithm of the response variable was taken in the model, and the prediction was then transformed back to its original scale by using the exponential function. An advantage of the Prophet model

is that it can handle irregular intervals, which is important as our fused chlorophyll-a observations are not regularly spaced. Prohpet is similar to other decomposition based approaches to time-series forecasting except that it uses generalized additive models instead of a state-space representation to describe each component. Using state space models would offer a more generic model formulation, whereas this approach explicitly models features common to the chlorophyll-a time series at hand, such as multi-period seasonality. The structural time series model could alternatively be put into state-space format, but rewriting it into that form would not alter the results.

Bayesian structural time series models possess further key features for modelling time series data that are favorable for long-term chlorophyll-a scenario analysis studies. The main feature is uncertainty quantification, as they allows us to quantify the posterior uncertainty of the individual components, control the variance of the components, and impose prior beliefs on the model. This is crucial as uncertainties increase over time in the future, especially in long-term projections. The second key feature is transparency, since the model is decomposed into simple time series components, which can be visually inspected. Moreover, they do not rely on differencing or moving averages, which make them more transparent than other autoregressive moving average models. The third key feature is the ability to incorporate regressors (covariates) as explanatory variables in the model. This feature is beneficial to include climate change impacts on chlorophyll-a trajectories from solar radiation and air temperature.

Here we briefly introduce the model without aiming completeness; for the full model formulation the reader is referred to [196]. We use a piecewise linear model with a constant rate of growth and change points. Suppose there are S change points, over a history of T points, at times $s_j, j = 1,\ldots,S$. We define a vector of rate adjustments $\delta \in \mathbb{R}^S$, where δ_j is the change in rate that occurs at time s_j. The rate at any time t is then the base rate k, plus all of the adjustments up to that point, which is represented by a vector $\boldsymbol{a}(t) \in \{0,1\}^S$ such that

$$a_j(t) = \begin{cases} 1, & \text{if } t \geq s_j, \\ 0, & \text{otherwise.} \end{cases}$$

The piecewise linear trend model with change points is then

$$g(t) = \left(k + \boldsymbol{a}(t)^T \boldsymbol{\delta}\right) t + \left(m + \boldsymbol{a}(t)^T \boldsymbol{\gamma}\right)$$

where k is the growth rate, $\boldsymbol{a}(t)$ is a change point indicator as defined above, $\boldsymbol{\delta}$ is the vector of rate adjustments, m is the offset parameter, and to make the function continuous, γ_j is set to $-s_j\delta_j$. We employ the following prior on $\boldsymbol{\delta} = (\delta_1,\ldots,\delta_S)$.

$$\delta_j \sim \text{Laplace}\,(0,\tau)$$

where τ controls the flexibility of the model in alternating its rate. While the model automatically detects change points and allows the trend to adapt appropriately, we have control over the trend flexibility by adjusting the strength of the sparse prior using the change point prior scale τ. In this application trend flexibility is significantly reduced by decreasing the change point prior scale to one fifth of its default value. The value

was fined tuned by balancing between the training error (which is lower with more flexibility) and the prediction error, while keeping the width of the projected uncertainty interval reasonable.

When the model is used for forecasting, the trend has constant rate and the uncertainty in the forecast trend is estimated. Future rate changes are simulated that emulate those of the past. In a fully Bayesian framework this can be done with a hierarchical prior on τ to obtain its posterior. In long-term projections, which is our purpose, one of the most influential factors is the uncertainty in the future trend. In this model, the uncertainty in the forecast trend is estimated by assuming that in the future the same average frequency and magnitude of rate changes will occur as observed in the past:

$$\text{for all j} > T, \begin{cases} \delta_j = 0 & \text{with probability } \frac{T-S}{T} \\ \delta_j \sim \text{Laplace}(0,\lambda) & \text{with probability } \frac{S}{T}. \end{cases}$$

Once λ has been inferred from the data, we use this model to simulate possible future trends and to compute uncertainty intervals. Due to the assumptions in the trend forecasting (matching historical frequency and magnitude) the trend intervals may not be exact, nevertheless they provide an indication of the level of uncertainty and also reveals trend model overfitting.

In the seasonality model we approximate seasonal effects with a standard Fourier series expansion with chosen periodicity P, and Fourier order n. The seasonality model is:

$$s(t) = \sum_{n=1}^{N} \left(a_n \cos\left(\frac{2\pi n t}{P}\right) + b_n \sin\left(\frac{2\pi n t}{P}\right)\right).$$

In this model the following periods are used, $P = 3652.5$ for decadal periodicity, $P = 365.25$ for yearly periodicity, $P = 182.625$ for half-yearly periodicity, and $P = 91.3125$ for quarterly periodicity (in days). The Fourier order was chosen as $N = 10$ after tuning such that under-fitting and over-fitting is avoided by minimizing the test error. The linear component then becomes

$$l(t) = X(t)\boldsymbol{\beta}$$

where $X(t) = \left[\cos\left(\frac{2\pi 1t}{P}\right), \sin\left(\frac{2\pi 1t}{P}\right), \ldots, \cos\left(\frac{2\pi Nt}{P}\right), \sin\left(\frac{2\pi Nt}{P}\right), R_1(t), \ldots, R_J(t)\right]$ is a matrix of seasonal components $s(t)$ and additional vectors of regressors, while $\boldsymbol{\beta} = \left[a_1, b_1, \ldots, a_N, b_N, r_1, \ldots, r_J\right]^T$ includes the $2N$ parameters of the Fourier series expansion and the R regression coefficients of the additional explanatory variables. The following $\boldsymbol{\beta} \sim N(0, \sigma^2)$ prior is imposed independently on each component of β. By default the linear component of the model only contains features for modeling seasonality but through specifying covariates ("regressors") we can include additional arbitrary vectors to $X(t)$ whose regression coefficients will be inferred. Combining the trend, seasonality and error components the final model becomes:

$$y(t) \big| m, \delta, \beta, \sigma \sim N\big(g(t) + l(t), \sigma\big)$$

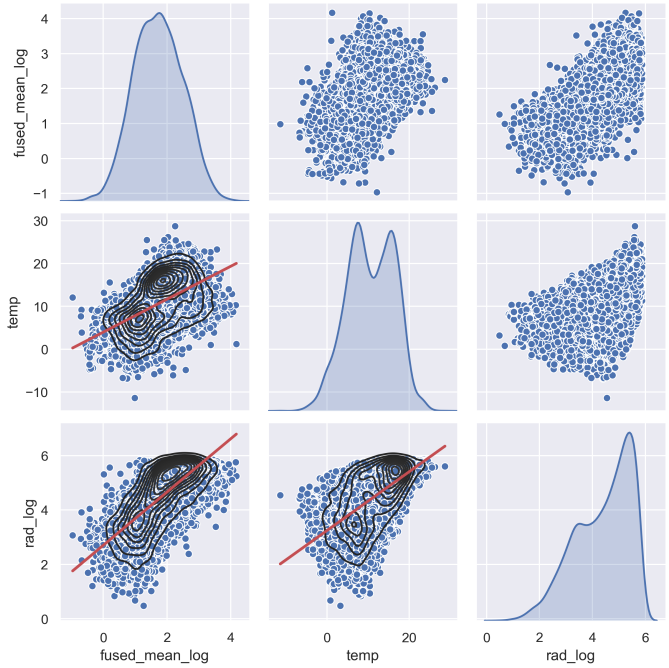


Figure 5.6: Pair plots of the log transformed response variable (fused chlorophyll-a), and the explanatory variables (log transformed radiation and temperature). Scatter plots are shown together with Kernel Density Estimates (black) and linear regression (red).

In order to construct an appropriate structural time series model, the selection of model components was facilitated by exploratory analysis steps, such as seasonal shape extraction, investigating the correlation of explanatory and response variables (Figure 5.6), produce periodogram and wavelet analysis to explore periodicity, and perform time series decomposition. Apart from chlorophyll-a, the solar radiation regressor data is also log transformed, since that produces a correlation structure to log chlorophyll, which is closer to linearity (see Figure 5.6). The temperature data could not be log transformed as it contains negative values. The continuous wavelet power spectrum revealed a persistent 12-month periodicity, which explained the largest amount of variability over the sampling period, while the rest of the variability is attributed to 6-month and 3-month periodicity. This is in line with previous research findings of wavelet analysis for the same observation station [216].

In the current structural time series model implementation the following components are used. Linear trend with change points (change point prior scale is defined), multi-period seasonality: decadal, yearly, half-yearly, and quarterly (periodicity, Fourier order, and prior scale are defined), as well as four additional regressors (air temperature, solar radiation, and their lag1). It should be noted, that adding more than lag1 of the regressors did not improve the prediction further. The parameter inference can be either done by optimization, using Limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm

(L-BFGS) to find a maximum a posteriori estimate, or through full posterior inference to include model parameter uncertainty in the forecast uncertainty.

5.2.4. TRACKING PHYTOPLANKTON SPRING BLOOM DYNAMICS

In order to track phytoplankton spring bloom dynamics, the last step of the methodological framework focuses on deriving spring bloom metrics obtained from the chlorophyll-a concentration time series. We must emphasize that uncertainty in the previous methodological steps (data fusion and long term projection) is being propagated to the estimates of cardinal dates and bloom magnitude. Although efforts have been dedicated to quantify these uncertainties, propagated uncertainty carries implications for the accuracy of the calculated cardinal dates.

Several existing methods are available to characterize phytoplankton blooms. [107] provides an exhaustive list of timing indices for quantifying phytoplankton phenology with advantages and disadvantages. These can be classified as biomass-based threshold methods, rate of change methods, and cumulative biomass-based threshold methods [40]. One might use the number of consecutive days that exceed a given threshold (elevated assessment level) defined by the literature. In the case of Dutch coastal waters this is around 12-15 mg/m^3 and 22-24 mg/m^3 for the Wadden Sea [160]. Alternatively, a low-pass method could be used for determining the start of the bloom [215], which is a temporal averaging algorithm acting as a low-pass filter, reducing the short-term fluctuations. [161] suggested using the date of the maximum and minimum values of daily change rates in the interpolated chlorophyll-a concentrations for the timing of the annual onset and breakdown of the phytoplankton bloom. The timing of the bloom can also be represented by another quantity, the center of gravity (COG) of the carbon content within the typical spring bloom period period [100]. Another possibility to characterise the spring bloom is to derive the cardinal dates of the mass development [172]. The cardinal dates are the beginning of the spring phytoplankton mass development, the maximum of the spring bloom (bloom peak), and the end of the spring mass development. Mathematical methods of describing cardinal dates were proposed by [172], such as finding the points of inflexion in the smoothed, log transformed, and differenced (1-week lag) data, deriving them from four linear segments (constant - increasing - decreasing - constant) fitted to the logarithmic values, or extracting the cardinal dates from the quantiles of a fitted parametric function (Weibull function). Similarly, [124] transformed phytoplankton biomass according to standard normal variation and took the first and third quartiles as cardinal dates, the beginning and the end of the spring bloom, respectively.

Several of the above mentioned methods (or listed by [107]) cannot properly deal with bi-modal data (require separation of the spring bloom) or large fluctuations in amplitude, some methods need parametric fitting (e.g. [209]), and most methods cannot deal with noisy data, hence require smoothing to pre-process the seasonal data before deriving the cardinal dates. As summarized by [107] if the seasonal time series is unimodal, from densely sampled and without noise, most methods will perform well. This is rarely the case, unless the data is interpolated and denoised. If that is not the case, more flexible approaches perform better which use less assumption on distribution patterns. For this reason to track long term changes in phytoplankton spring blooms we propose

to derive the cardinal dates using a non-parametric shape constrained method, namely log-concave regression [60, 87, 86]. Log-concave regression meets this flexibility requirement as it does not require any tuning parameters and can be directly applied on the annual bi-modal time series without any pre-processing. Consequently, our proposed method is less sensitive to bloom amplitude, missing data, and observational noise.

In summary, determining a mode of a unimodal (part of a) function, sometimes called 'bump hunting' is classically done using smoothing techniques, assuming some level of smoothness (which is reasonable) of the function. The advantage of using log-concave regression compared to techniques based on smoothing, is that it does not require tuning parameters (such as bandwidths) that heavily influence the outcome of the analysis. An alternative method one could use, would be unimodal regression, where no smoothness is used at all, resulting in discontinuous unimodal step functions as estimate of the regression function. The large class of log-concave functions contains unimodal functions that are continuous. Moreover estimation of these can be done in a stable manner.

In order to track long term changes in phytoplankton spring blooms we propose to derive the cardinal dates using a non-parametric shape constrained method, namely concave regression [60, 87, 86]. The concave or convex regression setup for a data set of size $\{n : (x_i, y_i) : i = 1, \ldots, n\}$ where $x_1 < x_2 < \ldots < x_n$ is the following:

$$Y_i = r_0(x_i) + \epsilon_i$$

for a concave function r_0 on $\mathbb{R}$, where $\{\epsilon_i : i = 1, \ldots, n\}$ are independent and identically distributed random variables and Y_i is the log chlorophyll-a concentration. Then, we apply concave regression on the log chlorophyll-a concentration data. We assume that the target of the estimation, $r_0 : \mathbb{R} \to \mathbb{R}$, is concave. Writing $\mathscr{K}$ for the set of concave functions on $\mathbb{R}$, the least squares estimate of r_0 is

$$\underset{r \in \mathscr{K}}{\operatorname{argmin}} \Phi(r), \quad \text{where} \quad \Phi(r) = \frac{1}{2} \sum_{i=1}^{n} \big(y_i - r(x_i)\big)^2$$

Utilizing this concave regression setup, the following two methodological steps are taken to identify the spring bloom cardinal dates (see Figure 5.7). The cardinal dates are the spring bloom beginning (B), -peak (P), and -end (E) dates expressed as the day of the year.

ISOLATING THE SPRING BLOOM

We take yearly time series of log chlorophyll-a concentrations (y_t), and assume that it is bi-modal separated by a boundary point t_b. In order to reduce computation time of the first step, we omit the first two months ($t_1 = 60$) and last two months ($t_2 = 300$) of the dataset since we know that the boundary that separates the spring and summer bloom will not be found there. It should be noted that omitting a portion of the yearly time series is only done in the first step during the identification of the boundary point. In the latter step, during the derivation of the spring bloom cardinal dates all dates on the "left side" of the boundary point are used $[0, t_b^{opt}]$. Omitting a portion of the yearly time series is optional. Then we fit $\Phi(t)$ on the data:

$$\Phi(t) = \begin{cases} \varphi_{t_b}(t) & t \le t_b \\ \tilde{\varphi}_{t_b}(t) & t > t_b \end{cases}$$

where $\varphi_{t_b}(t)$ is the concave regression of $(x_i, y_i) : x_i \le t_b$ on $[t_1, t_b]$, the "left side", and $\tilde{\varphi}_{t_b}(t)$ is the concave regression of $(x_i, y_i) : x_i > t_b$ on $[t_b{+}1, t_2]$, the "right side". Therefore both $\varphi_{t_b}(t)$ and $\tilde{\varphi_{t_b}}(t)$ are concave. The optimal boundary t_b^{opt} is found where the mean squared error of $\Phi(t)$ is minimal:

$$t_b^{opt} \rightarrow \underset{t_b}{\operatorname{argmin}}\, MSE_{t_b} + M\tilde{S}E_{t_b}$$

$$MSE_{t_b} = \frac{1}{t_b} \sum_{j=t_1}^{t_b} \left(y_j - \varphi_{t_b}(t_j)\right)^2$$

$$M\tilde{S}E_{t_b} = \frac{1}{t_2 - t_b} \sum_{j=t_b+1}^{t_2} \left(y_j - \tilde{\varphi_{t_b}}(t_j)\right)^2$$

This process of determining the boundary of spring and summer bloom is visually depicted in Figure 5.7A and Figure 5.7B.

DERIVE CARDINAL DATES OF THE SPRING BLOOM

After finding the boundary (t_b^{opt}) only the spring bloom ("left side") of the data is considered for further analysis where $t \in [0, t_b^{opt}]$. Then we take a continuous function $\Phi^*(t)$ which is defined as follows:

$$\Phi^*(t) = \begin{cases} c_l = \text{mean}\left(y_t : t \in [0, t_l]\right) & t \le t_l \\ \varphi(t) & t_l < t \le t_r \\ c_r = \text{mean}\left(y_t : t > t_r\right) & t > t_r \end{cases}$$

where c_l and c_r are constant and $\varphi(t)$ is the concave regression of $(x_i, y_i) : t_l < x_i \le t_r$. The points where the left constant function ends and the right constant function starts (t_l and t_r) will become the beginning and the end of the bloom (cardinal dates B and E). The third cardinal date, the peak of the bloom, is where $\varphi(t)$ takes its maximum. The points t_l and t_r are found where the mean squared error of $\Phi^*(t)$ is minimal:

$$(t_l, t_r) \rightarrow \underset{t_l, t_r}{\operatorname{argmin}} MSE_{c_l} + MSE_{c_r} + MSE_{\varphi}$$

$$MSE_{c_l} = \frac{1}{t_l} \sum_{j=0}^{t_l} \left(y_j - c_l(t_j)\right)^2$$

$$MSE_{c_r} = \frac{1}{t_J - t_r} \sum_{j=t_r}^{t_J} \left(y_j - c_r(t_j)\right)^2$$

$$MSE_{\varphi} = \frac{1}{t_r - t_l} \sum_{j=t_l}^{t_r} \left(y_j - \varphi(t_j)\right)^2$$

This final methodological step to identify t_l and t_r is shown in Figure 5.7C and Figure 5.7D. Finally, the cardinal dates together with the concave regression and the chlorophyll-a time series (transformed back to original values by taking their exponential function) are depicted in Figure 5.7E.

5

5.3. RESULTS

5.3.1. FUSED CHLOROPHYLL-A CONCENTRATION SIGNAL

The fused chlorophyll-a concentration signal, together with satellite observations, is depicted in Figure 5.8A and with in-situ observations in Figure 5.8B. One can observe that the fused signal almost perfectly follows the in-situ ("water") observations over the period in which only that type of measurements are available. From the moment that both in-situ and satellite date are available (1998), the fused signal lies between the two types but being closer to the in-situ observations according to the model formulation, since we have higher confidence in the field data. This is also reflected in the quantile-quantile plot and scatter plot of the fused signal compared to the in-situ data in Figure 5.8C-D, which lies almost perfectly on the diagonal, whereas the plot of the fused signal against the satellite observations deviates more from the diagonal. This enhancement of the historical chlorophyll-a signal has benefits for the projection step. Since the long-term projection is largely based on the observed correlations, if the input chlorophyll-a concentration time series is less accurate the statistical model will misrepresent the processes.

5.3.2. LONG TERM CHLOROPHYLL-A PROJECTION

The Bayesian structural time series model (introduced in Section 5.2.3) was trained (1976-2010) and tested (2010-2018) on the fused chlorophyll-a concentration signal and the historical measured solar radiation and air temperature data. Figure 5.9 visually depicts the validation of the in-sample forecast (1976-2010) and the forecast (2010-2018) against the fused data. The figure shows that most measurements (75%) lie within the predictive uncertainty band, indicating the model's reliability. The scatter plot of predictions is shown in Figure 5.10 whereas the performance metrics can be found in Table 5.1.

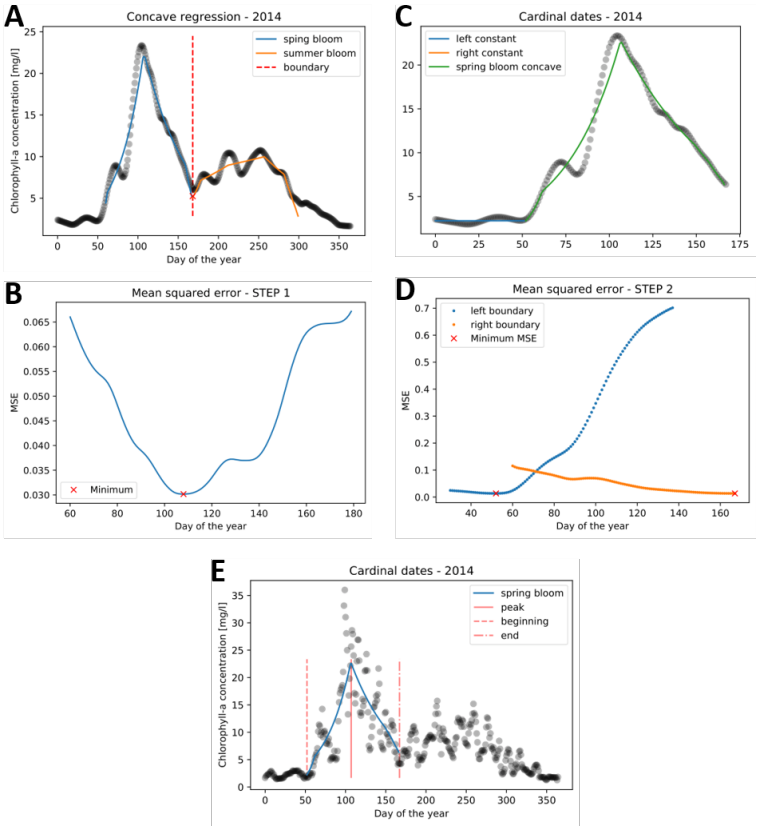


Figure 5.7: Steps to derive the cardinal dates of phytoplankton spring blooms: (1) Determining the boundary (t_b) for isolating the spring bloom **(A-B)**, and (2) concave regression to spring bloom **(C-D)**. The cardinal dates of the spring bloom are shown in **(E)**.

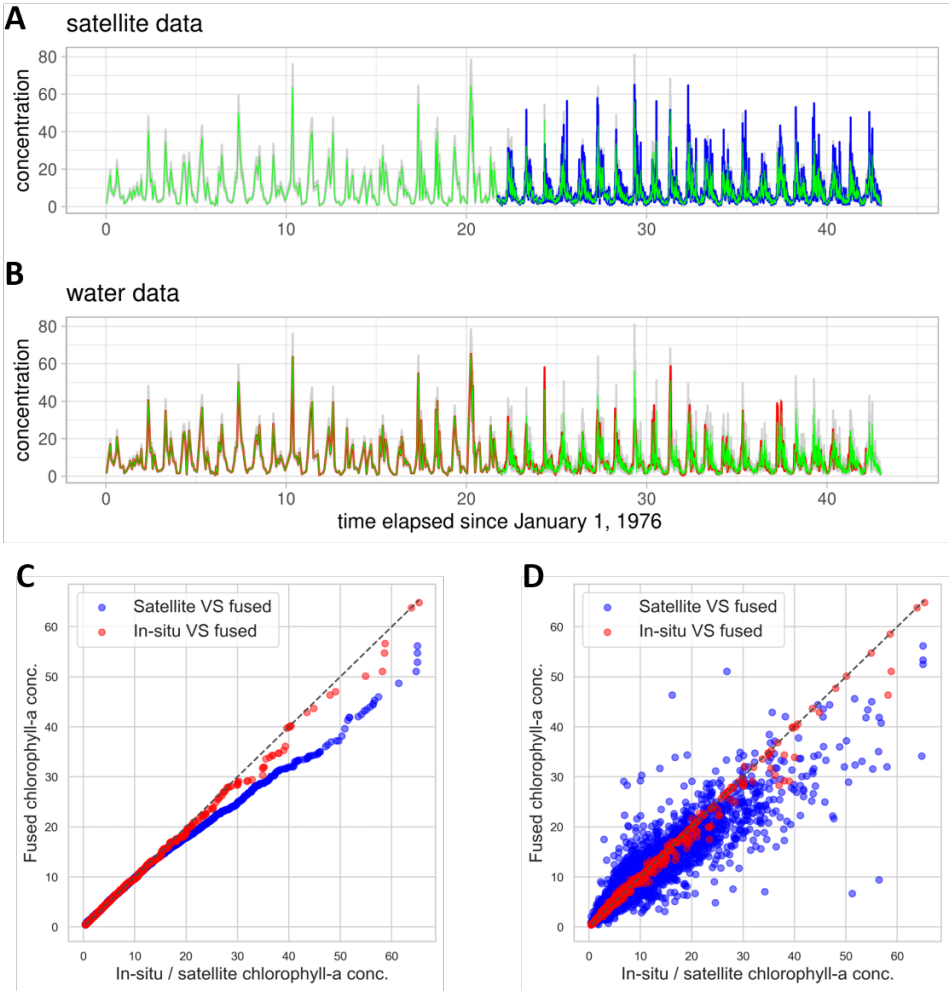


Figure 5.8: Data fusion results. The mean fused chlorophyll-a concentration signal (green) with uncertainty (grey) compared with satellite observations (blue) in **(A)**, and in-situ "water" observations (red) in **(B)**. Quantile-quantile plot of the fused signal compared to both in-situ and satellite observations in **(C)** and scatter plot in **(D)**.

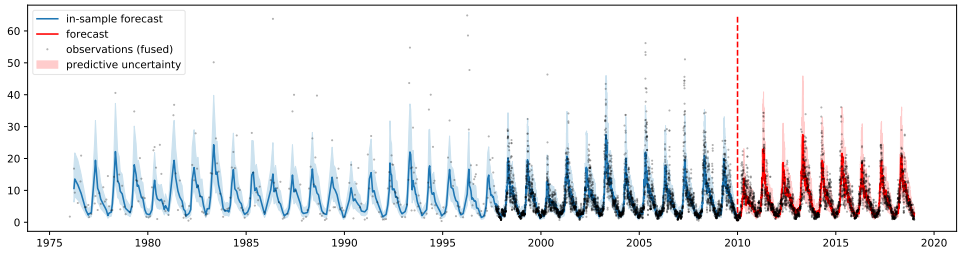


Figure 5.9: Time series forecasting validation against fused observations. Model fit between 1976-2010 (blue) and forecast between 2010-2018 (red). Predictive uncertainties in shaded area.

Table 5.1: Time series forecasting validation metrics against fused observations. Model forecast between 2010-2018 with N=3287 data points.

Performance metric	Value
N	3287.00
MAPE	0.38
RMSE	3.78
R2	0.51
% of obs in uncertainty band	75.63

While long-term data driven chlorophyll-a concentration prediction for climate impact assessment is not widespread, there have been few studies conducted on both inland water systems ([46, 114, 126, 137]) and marine systems ([105, 118, 31, 58]) that performed short term predictions. [31] predicted chlorophyll-a in the North Sea at different sites applying Generalized Additive Models (GAMs) with accuracies (R^2 values) ranging from 0.25 to 0.51 for hourly time scale, $0.15-0.22$ for daily time scale, and $0.27 to 0.63$ for bi-weekly time scale. Higher accuracy ($R^2 = 0.83$) was obtained in the North Atlantic, using a spatial GAM to predict month-to-month variation [105] or in a recent study by [58] where an R^2 value of more than 0.7 was achieved for a longer-term prediction (multi-year) with three different algorithms: Support Vector Machine Regressor (SVR), Random Forest, and Multi-layer Perceptron Regressor (MLP). SVR performed the best ($R^2 = 0.78$) with 17 predictor variables. Similar accuracies (R^2 values) were achieved in short-term prediction studies for lakes or reservoirs using Random Forest algorithm on monthly (0.2-0.6) and daily (0.6-0.8) data [126], as well as using Multiple-Layer Perceptron Neural Network (MLPNN) and Adaptive Network-based Fuzzy Inference System (ANFIS) 0.52-0.85 [137]. In comparison with these studies, we conclude that our model has acceptable accuracy, especially considering that we predict on a daily scale and eight years ahead, while most of the cited work focuses on much shorter prediction time frame. It should be noted that model comparability with other studies is hampered not only by the differences in ecosystem types (fresh water or open ocean instead of coastal waters) but also due to the fact that the predictor variables differ, and so as the experimental setup such as data splitting strategies, and prediction time frames.

After the calibration of hyperparameters and initial validation, the time series model

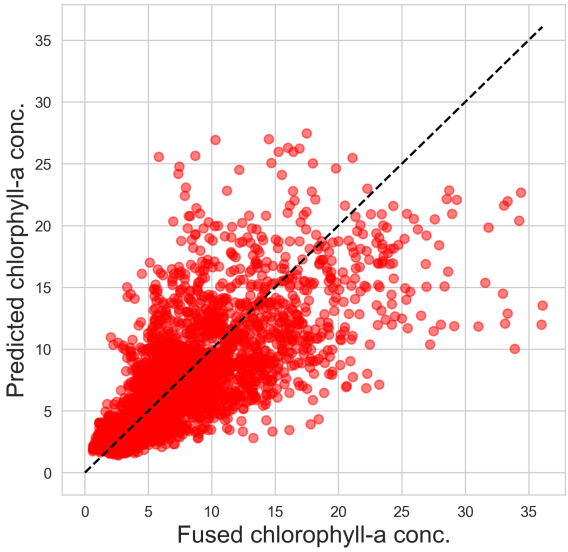


Figure 5.10: Scatter plot of predicted chlorophyll-a concentration against fused observations. Model forecast between 2010-2018 with N=3287 data points.

was retrained using the entire historical period (1976-2018), to better capture historical trends, and used for long-term chlorophyll-a concentration projection (2019-2089). Since the model contains log transformed solar radiation and air temperature as regressors, they need to be provided for the entire projection period. Consequently, after 2019 the bias corrected climate change projections are applied instead of the field observations. Given the numerous generated climate change projections (120 were used), the same number of future chlorophyll-a concentration trajectories were simulated, as shown in Figure 5.11. One can observe that the predictive uncertainty increases over time as we get farther from the projection start date. This predictive uncertainty originates from the trend component as explained in Section 5.2.3, and the modelling choices (e.g. changepoint prior scale) will influence it. We should emphasize that such long term projection is only a simplified approximation of the future chlorophyll-a signal, which follows a piecewise linear trend and continues to repeat its multi-seasonal behaviour, learnt from the past data, moreover includes linear effects of the two climate variables. These assumptions guarantee fast computation time, thus allowing numerous simulations for uncertainty quantification, which is the objective of this study. Nonetheless, it does not replace complex physically-based numerical models that are capable of simulating a wide range of ecological processes.

5.3.3. CHANGES IN PHYTOPLANKTON BLOOM DYNAMICS

The feature extraction step to derive the spring bloom cardinal dates (see Section 5.2.4) is first applied to the mean fused chlorophyll-a data to obtain the historical changes in spring bloom dynamics. Unfortunately, the cardinal dates could only be derived starting

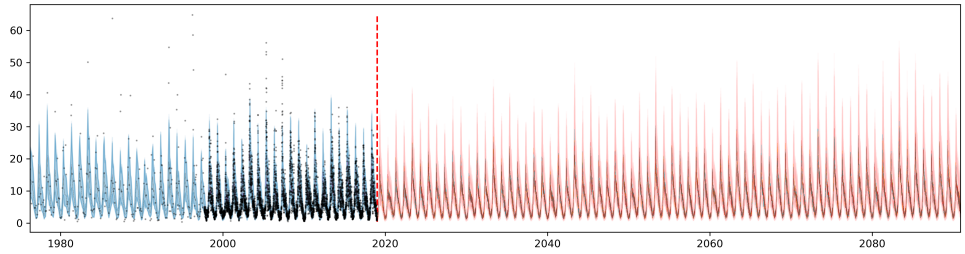


Figure 5.11: Long term chlorophyll-a concentration time series projection with radiation and temperature explanatory variables from generated climate projections (based on EURO-CORDEX). 120 solar radiation and air temperature projection scenarios were used to produce the 120 chlorophyll-a trajectories. Model fit between 1976-2018 (blue) and projection between 2019-289 (red). Predictive uncertainty in shaded area.

from 1998. This is due to the fact that between 1976 and 1998 only in-situ measurements were available which had a sparse temporal sampling frequency (10-20 per year). As previously argued, this number of yearly data points is insufficient to extract the cardinal dates. The historical phytoplankton bloom dynamics from 1998 to 2018 is depicted in Figure 5.12. The figure displays the three cardinal dates (beginning - green, peak - red, end - blue), the bloom duration (shaded blue area), and the bloom duration anomaly from the long-term mean bloom duration (bar chart). It can be observed that for certain years (2002, 2012, 2013) the bloom peak and bloom end cardinal dates lie very close to each other. These instances were visually confirmed. It was found that for 2002 and 2012 the feature extraction algorithm was accurate as a fast decay followed the bloom peak. On the other hand, in 2013 there was visibly no spring bloom observed, only a dominant summer bloom. This led the algorithm to falsely identify the spring bloom peak and end. This finding suggests that years where no spring bloom is observed should be removed from the dataset prior to applying the spring bloom cardinal detection algorithm. A possible extension of the method could be to report the type of seasonality (spring bloom, summer bloom, bi-modal, no bloom) [84] since changes in the type of seasonality are of interest, nevertheless, this is not part of the current implementation.

The feature extraction steps are then repeated on the projected future chlorophyll-a concentration between 2019-2089. The projected future spring bloom cardinal dates are depicted as boxplots in Figure 5.13A and as histograms in Figure 5.13B. The results indicate a relatively small variation, ~ 6 days, in the projected bloom peak timing (see Figure 5.14B), while a much higher level of uncertainty is observed for the bloom beginning, ~ 25 days, (see Figure 5.14A) and end timing, ~ 20 days (see Figure 5.14C). Bloom beginning and -peak resemble normal distributions, in the case of the bloom peak with a lower variance (higher peakedness). On the other hand, the bloom end resembles a right skewed log-normal distribution with relatively heavy tale due to the high number of outliers.

The bloom beginning is projected to slightly but consistently shift earlier, resulting in longer bloom duration towards the end of the century (see Figure 5.15A). The earlier spring bloom as an effect of climate change is in line with previous findings by [124] and [218] in laboratory trials (mesocosm experiments), by [59, 100, 161, 64] using histori-

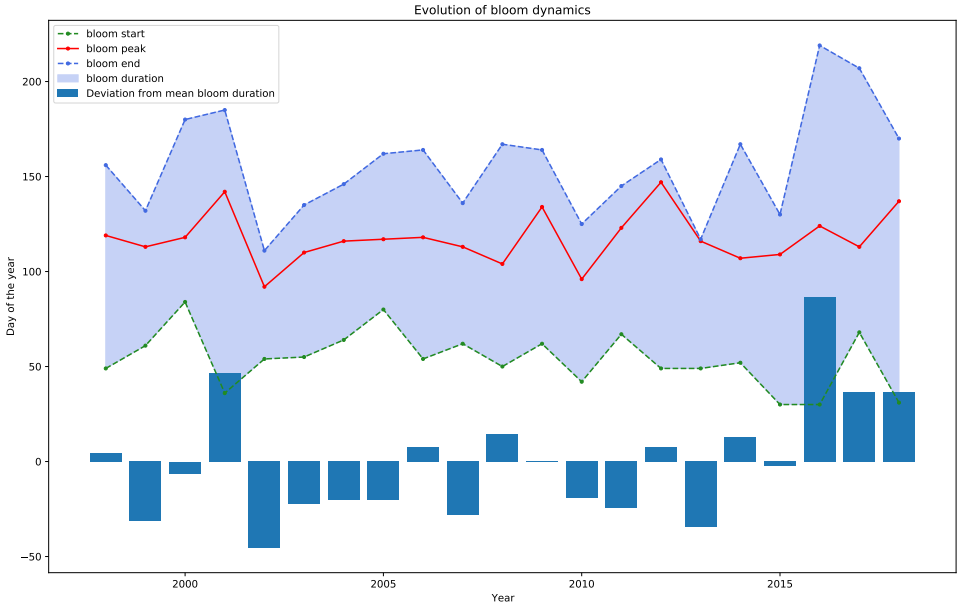


Figure 5.12: Historical spring bloom cardinal dates (beginning - green, peak - red, end - blue) and bloom duration (shaded blue area). The bar chart shows the yearly deviation (anomaly) from the long-term mean bloom duration.

cal data, or by [74] using numerical (hydrodynamic and ecological) prediction models forced by future climate change scenarios. Many of these studies found an even higher rate of spring bloom forward shift but in our case the accelerating effect of temperature rise might be moderated by the decreasing solar radiation trend. Despite the considerable uncertainty in the bloom end timing, no apparent trend can be observed. We emphasize that the actual day of the year of the derived cardinal dates may not be comparable to other findings in literature, since we used another method to obtain these cardinal dates. Thus, the projected trends and uncertainties carry the most value. We should also point out that the projected earlier spring blooms may not be a simple climatic response but could be the result of complex processes (physical and non-physical). Further investigation of these processes is necessary to fully understand the underlying mechanisms causing shifts in phytoplankton dynamics [100].

Apart from the cardinal dates, the chlorophyll-a concentration magnitude was also investigated. As Figure 5.15B shows, at the end of the 21^{st} century higher spring bloom peak magnitude can be expected. Considering the ensemble mean values, a $0.4\% year^{-1}$ trend is projected. This trend magnitude is comparable with the latest findings on chlorophyll-a historical trends in the North-West Shelf regions ($0.4-0.96\% year^{-1}$) [92], noting that this estimate was considering offshore marine waters, not coastal zones. It is also comparable to [221] who found nearly 20-30 % chlorophyll increase in the same study area between 1987-2012. Various numerical studies using climate models also project moderate increase in daily mean net primary production between 1980–1999

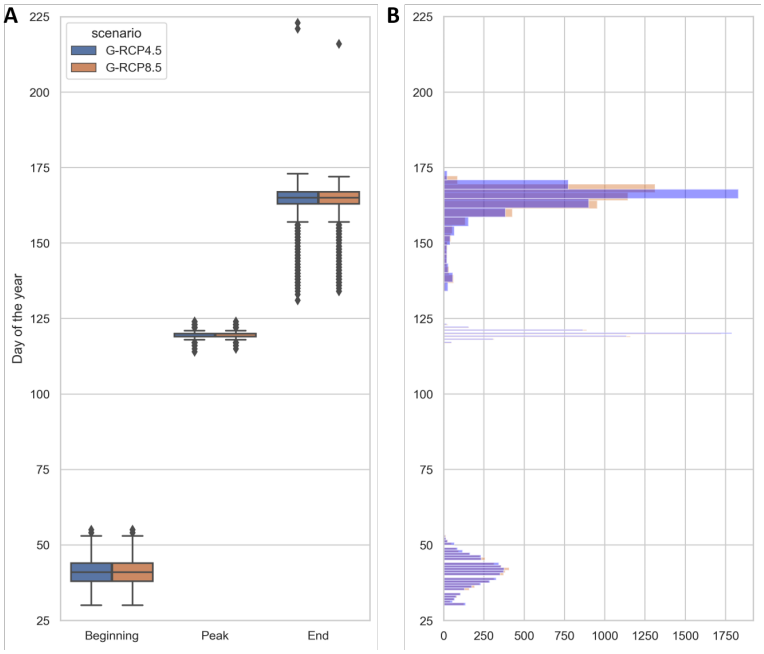


Figure 5.13: Range of projected future bloom cardinal dates **(A)** and their distributions **(B)** under 120 generated radiation and temperature projections (based EURO-CORDEX) (2019-2089). The statistics are grouped based on the generated projections corresponding to RCP scenarios (G-RCP4.5 and G-RCP8.5).

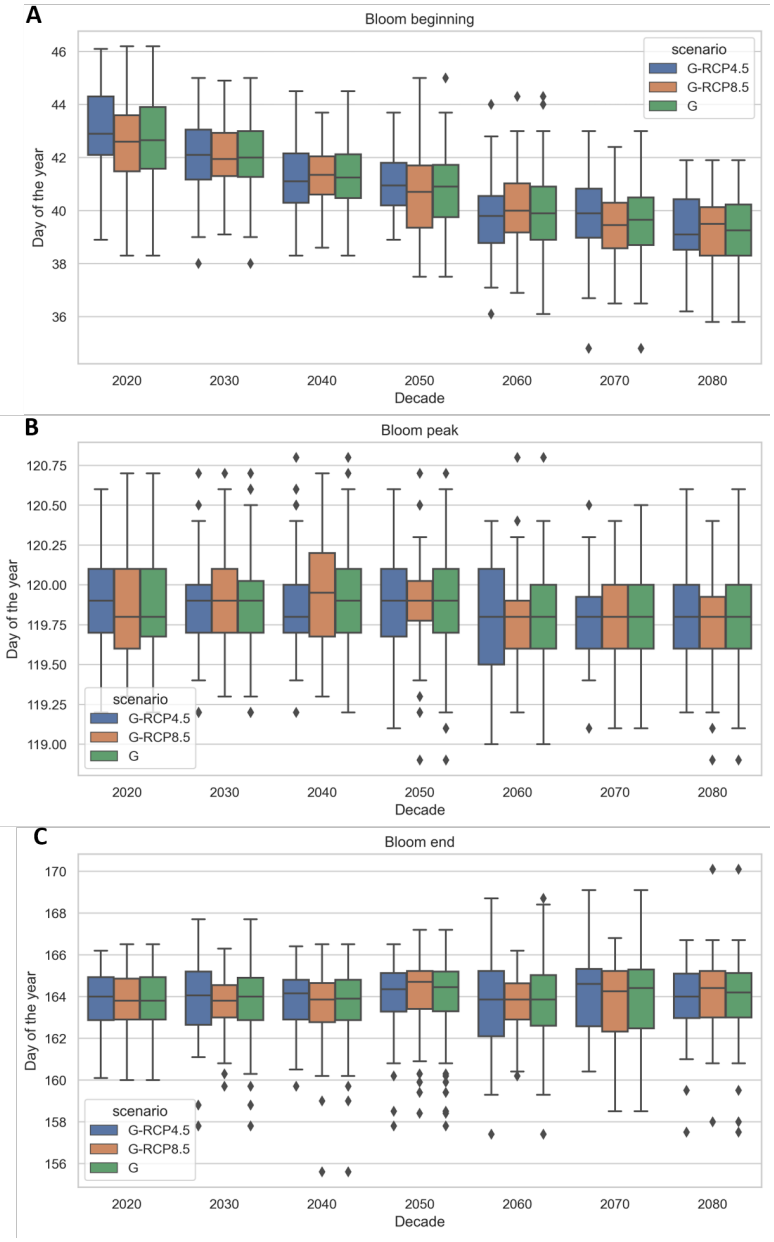


Figure 5.14: Projected future phytoplankton spring bloom beginning **(A)**, peak timing **(B)**, and end **(C)** under generated (G) radiation and temperature projections (based EURO-CORDEX) (2019-2089). The cardinal dates are grouped based on all generated projections (G), and generated projections corresponding to RCP scenarios (G-RCP4.5 and G-RCP8.5).

and 2080–2099 in the shallower southern North Sea [101, 102, 163]. We must emphasize that increasing chlorophyll concentration due to climate change is highly region specific (only occurring in some coastal areas) and very much debated [221]. In fact, some studies only report shift in spring bloom timing and species composition, but not in magnitude. In our study the projected positive trend is most probably driven by the linear trend component of the time series model and the rising air temperature as regressor, which have positive correlation to chlorophyll, based on the historical data. It should be noted, that in reality the correlation between air temperature and chlorophyll-a is non-linear and seasonally varying, moreover, it is different on a species or aggregate level. As the time series model could not incorporate non-linear correlations, it is assumed linear, hence, simulated interactions are only approximations of the real conditions. Nevertheless, in the season of interest (spring), when air temperature and solar radiation values did not reach their peak, this correlation is positive and the linearity assumption is a good approximation (see Figure 5.6). Furthermore, with chlorophyll-a concentration as a proxy we aim to describe aggregate level response, rather than species level response. We also emphasize that bloom magnitude is heavily influenced by nutrient concentration in the mixed layer depth [194, 21]. Although nutrient concentration was not used as an explanatory variable in this study we may expect that the correlation between air temperature and chlorophyll-a captured in historical data may include indirect effects such as thermal stratification, which influences nutrient availability in the mix layer depth.

The projected cardinal dates in Figures 5.13-5.15 are also grouped based on the generated projections corresponding to RCP scenarios. One observed difference is that in the last two decades bloom peak magnitudes are somewhat higher for RCP8.5. Perhaps counter intuitively, no other structural differences are visible between the RCP scenarios. The similarity between projected cardinal dates corresponding to RCP scenarios could be attributed to few reasons. Firstly, we must investigate the differences in solar radiation and air temperature projections between the RCP scenarios from Euro-CORDEX. As Figure 5.5 depicts, these differences for solar radiation are not apparent. For air temperature projections we see similar behaviour until the end of the century and differences in the last two decades become more articulate (RCP8.5 being higher), although few GCMS from both RCPs remain entangled and only one GCM from the RCP8.5 scenarios presents more extreme behaviour. This leads us to the second reason which might explain the lack of difference in cardinal dates between RCPs. The generated scenarios have been produced with a Bayesian stochastic generator introduced in [148]. This model assumes that Euro-CORDEX scenarios are exchangeable rather than independent, due to the fact that they originate from a common genealogy [190]. Consequently, the model formulation induces the phenomenom of “borrowing strength” where estimates for parameters over different scenarios are combined (“pooled”). This can correct outlier-like behaviour and makes the estimates statistically more robust [76, 78]. Thus, synthetic projections from this stochastic generator relax some of the distinct characteristics that input Euro-CORDEX RCP scenarios had. Although, new synthetic scenarios are generated per Euro-CORDEX scenario, due to the intentionally propagated uncertainty, the differences between synthetic scenarios of different RCP "families" may be less prominent. Additionally, the lack of clear response to the evident temperature difference increase in the past two decades may be attributed to a delayed feedback caused

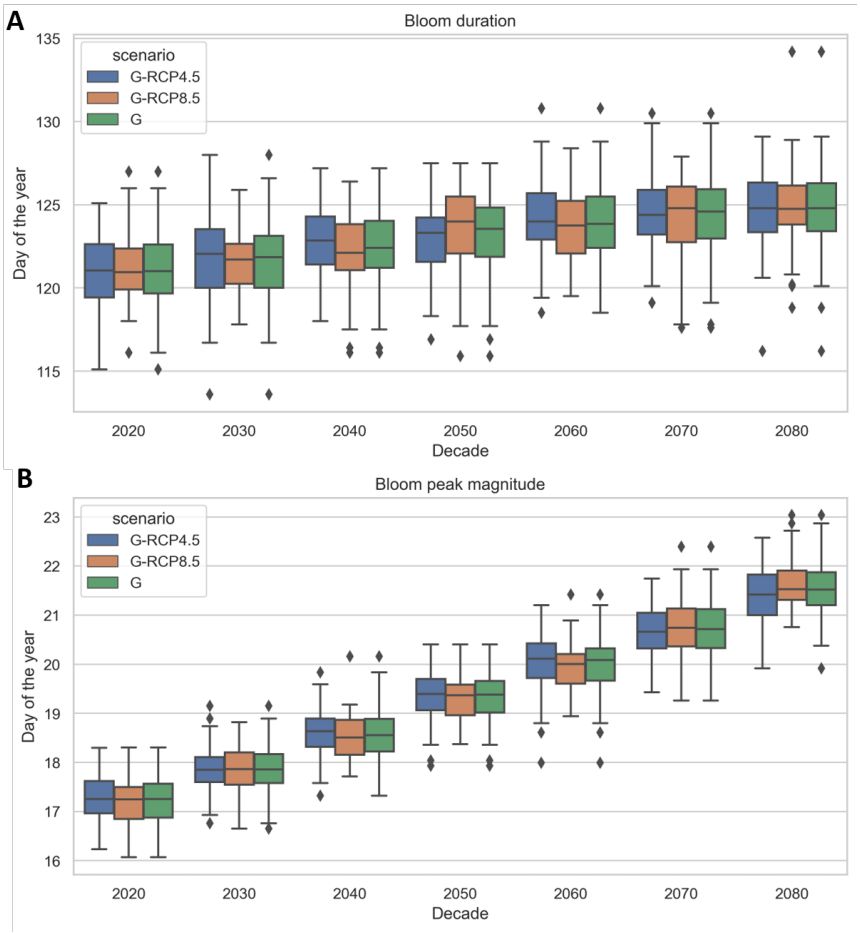


Figure 5.15: Projected future phytoplankton spring bloom duration **(A)** and peak magnitude **(B)** under generated radiation and temperature projections (based EURO-CORDEX) (2019-2089).

by ecosystem resilience [15]. Finally, and perhaps most importantly, it should be emphasized that generated scenarios serve as input into the structural time series model, which then feeds into log-concave regression step to derive the bloom metrics. As mentioned above, this adds further layers of uncertainties and the impacts of the various non-linear transformations may not be easily explained.

5.4. DISCUSSION

This paper presents an approach to study observed past and projected future marine phytoplankton phenology making use of statistical techniques, rather than physically-based models. The Bayesian setup in the data fusion and time series prediction models offer flexibility in model formulation and allow characterisation of predictive uncertain-

ties, which is crucial in climate change impact studies. In addition, for the extraction of phytoplankton cardinal dates we proposed a non-parametric regression model under shape constraints which has not been used before for such purposes, to our knowledge. Regarding the applied data, we aimed to make best use of the cross-disciplinary and multi-sourced measurements, covering marine biogeochemistry and atmospheric variables from field measurements, satellite imagery, numerical models, and synthetic generated scenarios.

We acknowledge the various sources of uncertainties in the data and models, which are considered and statistically quantified where possible. Firstly, uncertainty in the fusion of chlorophyll-a observations is quantified by the posterior distributions obtained through Bayesian parameter inference. Secondly, uncertainties in the climate projections are addressed using a large ensemble of generated stochastic scenarios, which cover numerous possible trajectories. Thirdly, in the Bayesian time series model we quantify uncertainties in two ways. On the one hand, uncertainty intervals of the future trend are computed individually for each projection, and on the other hand, this is repeated for a large number of projections, resulting in predictive uncertainty bands for each trajectory and for the entire ensemble. Lastly, uncertainty quantification in the feature extraction step is not possible explicitly, nevertheless, thanks to the ensemble approach a range of potential phytoplankton phenologies are simulated over the course of the 21^{st} century.

The main findings regarding phytoplankton phenology, the projected uncertainties in the beginning and the end of the spring bloom, as well as the prolonged bloom duration, increased peak magnitude and its forward shift (earlier bloom), may have repercussions on the marine food web. [73] found the same trends and attributed them to phenological mismatch between bloom timing and grazing pressure. When grazing pressure is shifted and predator-prey interactions are perturbed the phytoplankton loss by grazing is reduced resulting in higher bloom magnitude [26]. The forward shift in phytoplankton bloom phenology may also be explained by several other factors. These include increased early spring temperatures that accelerate phytoplankton cell division rates [19, 202, 104], change in stratification driven by temperature and/or wind trends, or change in the underwater light climate. Although, in our study slightly negative radiation trends are projected light availability can also be influenced by turbidity.

A consequence of these projected trends could be that energy transfer to higher trophic levels is disrupted as there is a tight coupling between the plankton trophic levels in marine pelagic ecosystems [169]. Such consequences are often described with the trophic match-mismatch hypothesis of [51]. Based on this hypothesis the reproductive success of higher trophic levels will be best when the phytoplankton phenology matches their requirements. Phenological shifts may therefore cause a temporal mismatch between zooplankton consumption (grazing) and phytoplankton production peak leading to higher mortality of the zooplankton, causing cascading effects towards the higher members of the food web [169, 202, 189, 31]. This has been documented in the North Sea [20], and other parts of the North Atlantic [162, 117]. The severity of these adverse effects in temperate productive systems is, however, debated [15]. Due to already high natural variability in the timing of predator consumption and its prey in temperate marine systems, compensating mechanisms may exist that could potentially reduce the impact of the

projected planktonic phenological shift [15, 59].

Our study aimed to quantify how uncertainty in environmental forcing, that influences the formation mechanism of spring blooms (through thermal stratification, mixed-layer temperatures, phytoplankton metabolic rates, and grazing) will impact the uncertainty in spring blooms dynamics. Since uncertainties in the spring bloom dynamics (especially timing [200]) are closely tied to uncertainties in secondary production, in the survival of larval populations, and ultimately in the recruitment to the adult stock [129], our results can inform further studies that attempt to propagate phytoplankton phenology related uncertainties to ecosystem response in higher trophic levels. An enhanced understanding of the variability of phytoplankton blooms is therefore a crucial step to estimate the impact on marine ecosystem functioning [216].

For future research the authors recommend to merge three components of the methodological framework into a single model. Integrating the Bayesian stochastic climate generator, the Bayesian data fusion model, and the Bayesian structural time series model would provide a consistent Bayesian hierarchical model that eliminates redundancies and offers a more elegant solution. It is worth noting that this integrated solution would be harder to re-use for researchers who are interested to take advantage of only a part of the model (stochastic generator, data fusion or projection) rather than the full chain. A further recommendation is to extend the approach to include spatial correlations, since currently only one location is considered. Extending the methodology in this way would allow us to make better use of the multi-dimensional data structure and include spatial gradients from coast to offshore locations.

As previously mentioned, chlorophyll-a concentration may not be an accurate proxy of phytoplankton biomass in the Southern North Sea [10]. In order to address this shortcoming, a potential avenue would be to apply novel satellite-derived products that consider phytoplankton functional types [220] or use phytoplankton carbon [23] instead of chorophyll-a. Although less frequently measured phytoplankton historical in-situ data is available in the North Sea, that could complement satellite derived indicators. In future research it should be evaluated if these indicators could better assess phytoplankton response to climate change. Another natural extension of the research is to further propagate uncertainties in spring bloom metrics to ecosystem behaviour. This could be achieved using statistical techniques or numerical models for predictive habitat distribution modelling (e.g. artificial neural networks, classification and regression trees).

An important limitation of the study is to only use air temperature and solar radiation as environmental covariates. Even though we confirmed that air temperature and solar radiation are the most dominant predictors for the study area and for the targeted temporal scale, inclusion of additional environmental factors impacting vertical mixing and bloom formation, such as nutrients, wind, salinity, dissolved oxygen, or mixed layer depth could improve the scenario analysis. Noting that the availability of long-term climate projections of any additional covariate is a prerequisite. Precipitation as a process related to ocean salinity has not been included for the following reason. According to [5] the salinity in the Wadden Sea is determined by fresh water input and its mixing with the North Sea and the influence of local climatic variations in precipitation can be ignored. Long-term variability of the salinity in our study area is in fact due to climatic variations in the precipitation over the river catchment areas (particularly the river Rhine) along

with other human induced changes and operation of waterways and sluices. Therefore, precipitation data at this site was not considered. Nevertheless, the air temperature and solar radiation variables indirectly impact ocean salinity through evaporation rates. Excluding vertical mixing processes such as wind can also be justified. While vertical mixing indeed affect nutrient conditions for phytoplankton blooms, the impacts depend on whether the area is already stratified [202]. According to [88], at the study location the water column is usually vertically well-mixed due to strong tidal mixing processes (strong flood and ebb currents) and persistent wind, which is a common feature here. This makes wind less relevant at this particular location. This was confirmed by both literature and our own data analysis.

Along with these points, we should also mention another important source of uncertainty in future climate studies focusing on the coastal zone, which is the role of anthropogenic interventions. Such interventions in the southern North Sea include coastal zone management efforts, aquaculture activities, sand mining, oil drilling, or fishing. Especially, large dredging and replenishment activities, like the major extension of the Port of Rotterdam (Maasvlakte 2) cause resuspension of buried inorganic nutrients into the water column and alter phytoplankton seasonality. In shallow coastal locations these uncertainties from anthropogenic impacts may outweigh the climate change induced ones, while moving towards transitional and offshore waters the effects are less prominent. Nonetheless, in this research human impacts are not addressed, only climatic ones.

Finally and most importantly, we recognize that our results related to climate change impacts on spring bloom dynamics will not resolve the ongoing debate on the complex and often contradictory findings. Especially, given the fact that the proposed data driven approach neglects the complicated and often non-linear ecological processes on species level. We reduced the marine biogeochemical response to climate change into a simple cause-effect relationship between two climate variables and chlorophyll-a concentration. As a consequence, our results are only an extrapolation of the observed correlations given projected changes in the climate using statistical models and giving appropriate attention to uncertainty quantification.

Despite the limitations, we believe that our proposed approach contributes to an integrated understanding of ecological responses to variable climate change through expressing future likelihoods of projected spring bloom dynamics and through the enhanced characterization of uncertainties associated to data and statistical methods.

6

CONCLUSIONS AND DISCUSSION

6.1. OVERARCHING CONCLUSIONS

This thesis presents a multi-disciplinary research where statistical techniques are applied to problems at the interface of marine and atmospheric processes. The keywords of the journals, in which the different chapters are published, demonstrate well the variety of topics where this thesis contributes to the scientific community:

- Hydroinformatics
- Coastal Environmental and Ecological Data Analysis
- Stochastic Environmental Research and Risk Assessment
- Marine Ecological Applications of Earth System Models and Regional Climate Models

The thesis contributes to these topics by providing statistical tools to understand the multi-dimensional climate and environmental datasets (and their relationships), as well as by offering ways for quantifying the uncertainties in the coastal ecological responses that are driven by the climatic variation. Due to the nature and scope of this research, the main conclusions are not related to the predicted trends and changes of coastal ecological response but instead they relate to the quantification of uncertainties, more specifically the necessary "way of thinking" as well as the statistical tools to derive those. The research conducted in the framework of this thesis will of course allow us to draw some conclusions on climate change induced trends and changes, such as the ones presented in Chapter 5, but the main value certainly lies elsewhere. **The main value of this thesis lies in the delicate balance of choosing our statistical and numerical techniques well to arrive at uncertainty quantification, sometimes making practical "engineering" assumptions and sometimes allowing room for more "elegant", transparent and integrated (but undoubtedly often mathematically more challenging) statistical solutions.** After all, that is the greatest benefit of statistics applied to environmental problems. While from the statisticians' point of view the research might be quite applied

and from the marine environmental scientists' point of view quite theoretical, it is the balancing of those two perspectives and making compromises (from both sides) that improves both scientific fields. If there must be one, this is the main conclusion of this thesis. Nevertheless, revisiting the main objectives presented in Chapter 1, the following practical conclusions were reached per research question:

Research Question 1: What is the value of probabilistic predictions for the coastal ecosystem state?

In other scientific fields, meteorology and hydrology for instance, using probabilistic predictions has been a common practice for a long while. For ecological modelling, however, deterministic predictions are still the standard, even more so in marine and coastal applications. While most efforts concentrate on improving the coastal water quality models for better deterministic predictions, in addition to these efforts (or instead of them) we should apply more probabilistic techniques to this field. Naturally, the most important value of probabilistic predictions for the coastal ecosystem state is that we attach predictive uncertainty to our mean prediction instead of trusting the best estimate, which is often rather erroneous due to the complexity and non-linearity of the modelled coastal ecological processes, the highly variable system dynamics and gradients, and the influence of land, to name a few. Obtaining prediction intervals (as shown in Figure 2.8, Figure 4.16, or Figure 5.11) gives the potential to express future likelihoods of predicted quantities (in our case chlorophyll-a concentrations) via pointwise-in-time predictive distributions. In a deterministic case we only have a single prediction (our best estimate) and with smaller ensembles we may derive ensemble mean and spread as a proxy of uncertainty, but it is the added value of fully probabilistic simulations to also derive the predictive distribution, which helps to achieve better characterization of uncertainties (e.g. distribution shape). This enhanced uncertainty quantification in turn supports better informed and rational decision making which often brings socio-economic and monetary benefits. In addition to this qualitatively defined value of using probabilistic simulations, the performance of deterministic and probabilistic predictions is also compared quantitatively in the thesis (in Chapter 2). A range of verification metrics that describe the goodness-of-fit, accuracy, reliability, and discrimination properties shows moderate improvement in the Dutch coastal waters when using a probabilistic model setup. The value of probabilistic predictions is therefore showcased in both statistical (derivation of pointwise-in-time predictive distribution) and "engineering" sense (improvement of verification metrics).

Research Question 2: How can we analyse, interpret and extract latent processes from complex multi-dimensional climate and environmental datasets to support coastal ecological impact studies?

When working with multi-sourced (model, satellite, in-situ) and multi dimensional (variable, time, space, scenarios) data we do have to dedicate sufficient attention to data exploration as applying or constructing statistical models based on such "multi-way" data is not trivial. The processes underlying the variations in these datasets are complex, the data may be noisy, and not all modelled variables are relevant to the studied processes.

In order to explore and reduce the data we propose a structured way of applying a set of dimension reduction techniques to incorporate spatial correlation via multi-way methods, temporal correlation through Dynamic Factor Analysis, and functional variability using Functional Data Analysis. This set of techniques was applied and evaluated in Chapter 3 and it can be concluded that they are indeed useful to guide the climate variables selection in coastal ecological impact studies. More specifically, with multi-way methods (particularly N-PLS) we were able to construct parsimonious models while including spatial correlation in the data structure. The dynamic factor model proved to be an appropriate tool to acquire information about underlying common trends in multivariate environmental time series, and to investigate the effects of atmospheric explanatory variables with the inclusion of temporal structure when constructing unobserved factors. Moreover, the functional PCA analysis found underlying functions that characterize the general shape of the environmental time series (mean function) and explain their functional variation, thereby reducing data complexity, and aid the interpretation of the underlying variability sources. While these dimension reduction methods have been separately already well documented in the literature (in fields such as chemometrics, econometrics and mathematics), structured and combined use of them for the multivariate analysis of atmospheric-ocean interactions to informing ecological impact studies is a novelty to the marine scientific community.

Research Question 3: How can we enrich existing climate projections to shift from scenario studies towards fully probabilistic climate impact studies?

Euro-CORDEX and other regional climate modelling experiments provide the users of their projections with a range of scenarios. These scenarios originate from a set of Representative Concentration Pathways and General Circulation Models. At present, most coastal climate impact studies only use these ensemble members to anticipate potential climate trajectories. While using around a dozen of scenarios is already helpful to show the variability in our future estimates, it does not produce proper uncertainty information. This thesis offers a paradigm shift from such scenario studies towards fully probabilistic studies by proposing a Bayesian stochastic climate generator that allows us to produce numerous climate trajectories around the existing ones. The main value of our proposed multi-layered (hierarchical) Bayesian climate generator is that it combines different climate scenarios into one model (rather than separately treating them), making model parameter estimates statistically more robust. This enhanced parameter uncertainty characterization permits us to represent the full range of plausible climate scenarios and subsequently the full range of impacts, once climate input is propagated through process-based models.

Our further conclusions and recommendations about enriching available climate projections include that: (1) simulated trajectories should directly incorporate long-term trend avoiding the common practice of simulating residuals which are then added to climatology (historical or climate change adjusted); (2) high-resolution Regional Climate Model ensembles (e.g. Euro-CORDEX) should be used as input to these generators as they describe local processes the best; and (3) a hierarchical Bayesian model can provide the necessary flexibility in model formulation. Even though some of these conceptual elements separately exist in the field of stochastic weather generators or more

broadly in the field of climate sciences and/or environmental sciences, the combination of these elements into one model is innovative. They provide the possibility to generate synthetic but representative Regional Climate Model scenarios, which saves the "cost" of producing further simulations with the climate models.

Research Question 4: How can we quantify climate change induced uncertainties in coastal phytoplankton spring bloom dynamics?

In order to answer this important last research question, which leads us closest to the title of this thesis and concludes the entire research, we build on the knowledge and results of previous chapters and make use of the statistical tools developed in those chapters. In fact, this question has several components.

- *What drives climate related changes and uncertainties in coastal phytoplankton dynamics?* (link to Chapter 3)

It is very hard to comprehensively answer the first sub-question. Nevertheless, imposing well-defined boundaries to this problem, such as focusing only on atmospheric drivers and limiting ourselves to the variables that have the largest influence on coastal chlorophyll-a concentration variability on daily to seasonal time scales, Chapter 3 concluded that we should use solar radiation and air temperature as proxies of climate related changes in this specific coastal area.

- *How to represent climate related uncertainties and statistically quantify those?* (link to Chapter 4)

Regarding the representation of climate related uncertainties, we should generate sufficiently many synthetic solar radiation and air temperature long-term trajectories (based on the methodology described in Chapter 4), that can be used as input to the fully probabilistic climate impact assessment. Since with this approach the predictive uncertainties of the synthetic solar radiation and air temperature time series are well characterized, the subsequently applied methods can use both the generated trajectories themselves and their statistical properties (e.g. mean, median, standard deviation, distribution shape), depending on the method.

- *How can we simulate long term evolution (an entire century) in phytoplankton dynamics?* (link to Chapter 5)

Due to the computational limitations of using physics-based numerical models to simulate long-term (century scale) chlorophyll-a trajectories for hundreds of scenarios, we have to employ fit-for-purpose model emulators. Chapter 5 has shown that a Bayesian structural time series is a good candidate for that. Its ability to produce a simplified approximation of the future long-term chlorophyll-a signal, which follows a piecewise linear trend and continues to repeat its multi-seasonal behavior learnt from the past data, and also includes linear effects of the two climate variables, makes it indeed fit-for-purpose for this objective. Training such a data-driven model with sufficient dataset is of paramount importance. In order to address this issue, we advise to make use of

multi-sensor data and fuse those complementary historical measurements into an enhanced signal. We have proposed a data fusion model for this purpose in Chapter 5 that combines the positive features of in-situ and satellite measurements, longer historical records and reliability on one side and higher frequency data on the other.

- *How do we relate phytoplankton spring bloom dynamics to quantities that can be simulated (chlorophyll-a)?* (link to Chapter 5)

Although deriving predictive uncertainties in long-term chlorophyll-a trajectories is useful in itself, the interest of the scientific community seems to concentrate on phytoplankton spring bloom dynamics. For this reason, Chapter 5 proposed a new feature extraction method to derive yearly spring bloom cardinal dates (beginning, peak, end) from chlorophyll-a concentrations as a proxy for spring bloom dynamics. The existing literature concludes that existing methods for this purpose perform well if the time series is uni-modal, densely sampled and without noise, but if this is not the case (as in most field- and remotely sensed data), more flexible approaches perform better which use less assumptions on distribution patterns. Our conclusions in this respect are that the previously mentioned difficulties can be overcome using a non-parametric shape constrained method, such as log-concave regression, as it meets this flexibility requirement, does not require any tuning parameters and can be directly applied on annual bi-modal time series without any pre-processing. Consequently, our proposed method is less sensitive to bloom amplitude variability, missing data, and observational noise.

6.2. MAIN LIMITATIONS

First of all, as already mentioned in Chapter 1, **not all uncertainty sources were addressed in the thesis**. We primarily focus on knowledge uncertainties related to input data and model parameters. We quantify (and attempt to reduce) uncertainty in the chlorophyll-a data by fusing in-situ and satellite observations, and in the climate data by generating synthetic climate projections. We also quantify uncertainty in the models used to generate synthetic climate projections, in the numerical or data-driven models to simulate chlorophyll-a projections, and in the feature extraction step to derive phytoplankton cardinal dates. All of these steps help us to express confidence in our simulations and provide likelihoods of the possible simulation outcomes. We do not consider how decisions will be adapted after having access to these likelihood estimates, however. Therefore, the decision uncertainty (how uncertainty estimates will be interpreted and incorporated into decisions) is neglected. We also neglect the impact of anthropogenic interventions. These interventions, such as coastal zone management efforts, aquaculture activities, dredging, oil drilling, or fishing, alter phytoplankton seasonality. In shallow coastal systems these uncertainties from anthropogenic impacts may outweigh the climate change induced ones. Nevertheless, in this research human impacts are not addressed, only climatic ones.

Secondly, throughout the thesis **the ecological response is simplified into a single indicator, chlorophyll-a concentration**, although in Chapter 5 features of the chlorophyll-a signal are extracted to derive phytoplankton spring bloom dynamics. The decision to focus only on chlorophyll-a concentration, which is a proxy for phytoplankton biomass,

was made to study changes in a primary indicator. Phytoplankton is the base of the marine food web and the higher trophic levels (on the predatory chain) depend on it. The main limitation with chlorophyll-a concentration as response variable is that we can only draw conclusions on phytoplankton dynamics at the aggregate level, not on species composition or population structure. Chlorophyll-a is therefore a very useful aggregate indicator to describe the ecosystem state, without having to consider the complex non-linear ecosystem processes within the trophic levels, but also limits the depth of conclusions we can make on phytoplankton populations. In the literature there are even debates that chlorophyll-a concentration may not be an accurate proxy of phytoplankton biomass in the Southern North Sea. Potential avenues would be to apply novel satellite-derived products that consider phytoplankton functional types or use phytoplankton carbon instead of chorophyll-a. While these new monitoring methods may better assess phytoplankton response to climate change in the future, at present chlorophyll-a concentration is used most prominently by the scientific community as it is historically measured and modelled by all monitoring types (in-situ, satellite, model). This relative abundance of data is crucial for the development and testing of the employed statistical techniques.

A further limitation is that **the study only uses atmospheric variables as climate signals**. Due to the complex interactions of climate forcing conditions with marine ecological processes, responses of phytoplankton to climate change are not trivial to estimate. In reality, apart from atmospheric climate variables there are other marine physical and non-physical factors impacting phytoplankton biomass, such as nutrients, salinity, dissolved oxygen or grazing. We have tried to limit the masking effect of trophic interactions, as far as this may be possible, by focusing on the spring phytoplankton bloom in Chapter 5. This is due to the fact that in temperate marine systems during spring the physical factors like temperature, light availability, and mixing are more prominent than the non-physical ones (e.g. grazing). We should note again that the availability of long-term historical data and future projections of any additional driver mentioned above is a prerequisite. Unfortunately most of those are not as extensively recorded and modelled as air temperature and solar radiation. Finally, the fact that the considered atmospheric drivers indirectly influence the formation mechanism of spring blooms in many ways (through thermal stratification, mixed-layer temperatures, phytoplankton metabolic rates, and grazing rates), reduces the importance of this limitation.

Another issue is the **relative fragmentation in the methods** developed in Chapter 4 and Chapter 5. Integrating the three main uncertainty quantification tools: the Bayesian stochastic climate generator, the Bayesian data fusion model, and the Bayesian structural time series model would provide a consistent Bayesian hierarchical model that eliminates redundancies and offers a more elegant solution. Although, apart from the additional effort that would be required to combine these methods, it is worth noting that this integrated solution would be harder to re-use for researchers who are interested to take advantage of only a part of the model (stochastic generator, data fusion or long-term projection) rather than the full chain.

Finally, a practical barrier to the uptake of the large number of generated synthetic climate scenarios offered by this thesis is the **computational time limitation** of the subsequent models that use those data as input. We have reduced this limitation by intro-

ducing a model emulator to replace the computationally expensive three dimensional physics-based model with a data-driven one. The construction, calibration and validation of such data-driven models takes considerable effort, however.

ACKNOWLEDGEMENTS

Naturally, the biggest debt of gratitude is owed to my supervisory trio: Frank, Geurt, and Ghada. Thank you for making my PhD possible and guiding me with enthusiasm, patience and professionalism!

Since practically all my higher educational steps and professional life is entwined with European grants, funds, and research projects, a huge acknowledgment goes the European Union for supporting science and innovation in the European Communities.

More specifically, this research has received funding from the European Union's Horizon 2020 research and innovation programme project ODYSSEA under grant agreement No 727277. The ODYSSEA project is about operating a network of integrated observatory systems in the Mediterranean Sea. Additional funding has been provided by the Directorate-General for European Civil Protection and Humanitarian Aid Operations project GREEN under grant agreement No ECHO/SUB/2016 / 740172/ PREV18. The GREEN project focused on Green infrastructures for disaster risk reduction protection: evidence, policy instruments and marketability. These two European projects made it financially possible for me to complete the PhD. Apart from the financial means, the diverse international project teams made the first steps of my professional life a very valuable learning experience. This spirit of international collaboration transformed the way I approach problems and find solutions to them, in my research and applied work alike.

I thank the people of The Netherlands and more specifically Delft for their hospitality. Delft became our true second home. Or maybe it is now the first ...?

On a personal note, I would like to thank both my Deltares and TU Delft families. This gratitude is extended to all my previous Deltares department colleagues over the four years: Marine and Coastal Management (MCM); Information, Resilience and Planning (IRP); and Data Science and Water Quality (DSW). Of course it includes all PhDs and staff members of the Statistics group of the Department of Applied Mathematics, who welcomed me very warmly and fully accepted me in their group, although I was a continuous but rare visitor.

Finally, I thank all my Hungarian friends and my family: my wife (Vanni), my mom (Anya), and my brother (Matyi) for supporting my endless educational cycle, which will finally come to an end.

Köszönöm szépen!

BIBLIOGRAPHY

[1] Nerilie Abram et al. *Framing and Context of the Report - Supplementary Material. In: IPCC Special Report on the Ocean and Cryosphere in a Changing Climate*. Tech. rep. IPCC, 2019. URL: https://www.ipcc.ch/site/assets/uploads/sites/3/2019/11/SROCC_Ch01-SM_FINAL.pdf.

[2] Nerilie Abram et al. *Framing and Context of the Report. In: IPCC Special Report on the Ocean and Cryosphere in a Changing Climate*. Tech. rep. IPCC, 2019. URL: https://www.ipcc.ch/site/assets/uploads/sites/3/2019/11/05_SROCC_Ch01_FINAL.pdf.

[3] E. Aguilar and M. Brunet. "Seasonal Patterns of Air Surface Temperature and Pressure Change in Different Regions of Antarctica". In: *Detecting and Modelling Regional Climate Change* (2001), pp. 215–228. DOI: 10.1007/978-3-662-04313-4_19. URL: https://link.springer.com/chapter/10.1007/978-3-662-04313-4_19.

[4] Jung Min Ahn, Sang Jin Lee, and Taeuk Kang. "Development of water quality forecasting system with ensemble stream prediction method in the Geum River Basin, Korea". In: *Desalination and Water Treatment* 57.2 (2016), pp. 670–683. ISSN: 19443986. DOI: 10.1080/19443994.2014.996010.

[5] Hendrik M. van Aken. "Variability of the water temperature in the western Wadden Sea on tidal to centennial time scales". In: *Journal of Sea Research* 60.4 (Nov. 2008), pp. 227–234. ISSN: 1385-1101. DOI: 10.1016/J.SEARES.2008.09.001.

[6] Mohamed Ali Ben Alaya et al. "Change point detection of flood events using a functional data framework". In: *Advances in Water Resources* 137 (Mar. 2020), p. 103522. ISSN: 0309-1708. DOI: 10.1016/J.ADVWATRES.2020.103522.

[7] Abdullah Alodah and Ousmane Seidou. "Influence of output size of stochastic weather generators on common climate and hydrological statistical indices". In: *Stochastic Environmental Research and Risk Assessment* 34.7 (July 2020), pp. 993–1021. ISSN: 14363259. DOI: 10.1007/s00477-020-01825-w.

[8] Abdullah Alodah and Ousmane Seidou. "The adequacy of stochastically generated climate time series for water resources systems risk and performance assessment". In: *Stochastic Environmental Research and Risk Assessment* 33.1 (Jan. 2019), pp. 253–269. ISSN: 14363259. DOI: 10.1007/s00477-018-1613-2.

[9] Santiago Alvarez-Fernandez, Han Lindeboom, and Erik Meesters. "Temporal changes in plankton of the North Sea: Community shifts and environmental drivers". In: *Marine Ecology Progress Series* 462 (Aug. 2012), pp. 21–38. ISSN: 01718630. DOI: 10.3354/meps09817. URL: www.int-res.com/articles/suppl/.

[10] Santiago Alvarez-Fernandez and Roel Riegman. “Chlorophyll in North Sea coastal and offshore waters does not reflect long term trends of phytoplankton biomass”. In: *Journal of Sea Research* 91 (Aug. 2014), pp. 35–44. ISSN: 13851101. DOI: 10.1016/j.seares.2014.04.005.

[11] A. Amengual et al. “A statistical adjustment of regional climate model outputs to local scales: Application to Platja de Palma, Spain”. In: *Journal of Climate* 25.3 (Feb. 2012), pp. 939–957. ISSN: 08948755. DOI: 10.1175/JCLI-D-10-05024.1.

[12] Claus A. Andersson and Rasmus Bro. “The N-way Toolbox for MATLAB”. In: *Chemometrics and Intelligent Laboratory Systems* 52.1 (2000), pp. 1–4. ISSN: 01697439. DOI: 10.1016/S0169-7439(00)00071-X.

[13] David Antoine and André Morel. “Oceanic primary production: 1. Adaptation of a spectral light-photosynthesis model in view of application to satellite chlorophyll observations”. In: *Global Biogeochemical Cycles* 10.1 (Mar. 1996), pp. 43–55. ISSN: 08866236. DOI: 10.1029/95GB02831. URL: http://doi.wiley.com/10.1029/95GB02831.

[14] L. Arentz et al. *Kalibratie slibtransport- en GEM model (Calibration of the sediment transport- and GEM model)*. Tech. rep. Delft, The Netherlands: Deltares, 2012. DOI: Report1205620-000-ZKS-0014.

[15] Angus Atkinson et al. “Questioning the role of phenology shifts and trophic mismatching in a planktonic food web”. In: *Progress in Oceanography* 137 (Sept. 2015), pp. 498–512. ISSN: 00796611. DOI: 10.1016/j.pocean.2015.04.023.

[16] Hanneke Baretta-bekker et al. “Report on the second application of the OSPAR Comprehensive Procedure to the Dutch marine waters”. In: 5 (2008), p. 79.

[17] Blanka Bartók et al. “Projected changes in surface solar radiation in CMIP5 global climate models and in EURO-CORDEX regional climate models for Europe”. In: *Climate Dynamics* 49.7-8 (Oct. 2017), pp. 2665–2683. ISSN: 14320894. DOI: 10.1007/s00382-016-3471-2. URL: https://link.springer.com/article/10.1007/s00382-016-3471-2.

[18] Gregory Beaugrand, Xavier Harlay, and Martin Edwards. “Detecting plankton shifts in the North Sea: A new abrupt ecosystem shift between 1996 and 2003”. In: *Marine Ecology Progress Series* 502 (Apr. 2014), pp. 85–104. ISSN: 01718630. DOI: 10.3354/meps10693. URL: www.int-res.com.

[19] Grégory Beaugrand and Philip C. Reid. “Long-term changes in phytoplankton, zooplankton and salmon related to climate”. In: *Global Change Biology* 9.6 (June 2003), pp. 801–817. ISSN: 13541013. DOI: 10.1046/j.1365-2486.2003.00632.x. URL: https://onlinelibrary.wiley.com/doi/full/10.1046/j.1365-2486.2003.00632.x%20https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1365-2486.2003.00632.x%20https://onlinelibrary.wiley.com/doi/10.1046/j.1365-2486.2003.00632.x.

[20] Grégory Beaugrand et al. “Plankton effect on cod recruitment in the North Sea”. In: *Nature* 426.6967 (Dec. 2003), pp. 661–664. ISSN: 00280836. DOI: 10.1038/nature02164. URL: https://www.nature.com/articles/nature02164.

[21] Michael J. Behrenfeld. "Abandoning sverdrup's critical depth hypothesis on phytoplankton blooms". In: *Ecology* 91.4 (Apr. 2010), pp. 977–989. ISSN: 00129658. DOI: 10.1890/09-1207.1. URL: https://esajournals.onlinelibrary.wiley.com/doi/full/10.1890/09-1207.1%20https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1890/09-1207.1%20https://esajournals.onlinelibrary.wiley.com/doi/10.1890/09-1207.1.

[22] Michael J. Behrenfeld and Emmanuel S. Boss. "Student's tutorial on bloom hypotheses in the context of phytoplankton annual cycles". In: *Global Change Biology* 24.1 (Jan. 2018), pp. 55–77. ISSN: 13541013. DOI: 10.1111/gcb.13858. URL: http://doi.wiley.com/10.1111/gcb.13858.

[23] Marco Bellacicco et al. "Improving the retrieval of carbon-based phytoplankton biomass from satellite ocean colour observations". In: *Remote Sensing* 12.21 (Nov. 2020), pp. 1–13. ISSN: 20724292. DOI: 10.3390/rs12213640. URL: www.mdpi.com/journal/remotesensing.

[24] Klemen Bergant and L. Kajfež-Bogataj. "N-PLS regression as empirical downscaling tool in climate change studies". In: *Theoretical and Applied Climatology* 81.1-2 (2005), pp. 11–23. ISSN: 0177798X. DOI: 10.1007/s00704-004-0083-2.

[25] Justus E. E. van Beusekom et al. "Wadden Sea Eutrophication: Long-Term Trends and Regional Differences". In: *Frontiers in Marine Science* 6.7 (July 2019), p. 370. ISSN: 2296-7745. DOI: 10.3389/FMARS.2019.00370. URL: www.frontiersin.org.

[26] Justus E.E. van Beusekom, Martina Loebl, and Peter Martens. "Distant riverine nutrient supply and local temperature drive the long-term phytoplankton development in a temperate coastal basin". In: *Journal of Sea Research* 61.1-2 (Jan. 2009), pp. 26–33. ISSN: 13851101. DOI: 10.1016/j.seares.2008.06.005.

[27] A.G. Birt et al. "A simple stochastic weather generator for ecological modeling". In: *Environmental Modelling & Software* 25.10 (Oct. 2010), pp. 1252–1255. ISSN: 1364-8152. DOI: 10.1016/J.ENVSOFT.2010.03.006.

[28] Meinte Blaas, Gerben De Boer, and Sandra Gaytan Aguilar. "Eutrophication assessment using remotely sensed and in situ Chlorophyll-a data". In: October (2013). DOI: 10.13140/RG.2.1.4259.9125.

[29] Blauw and Anouk. *UvA-DARE (Digital Academic Repository)*. Tech. rep. 2015. URL: http://dare.uva.nl.

[30] Anouk N. Blauw et al. "GEM: A generic ecological model for estuaries and coastal waters". In: *Hydrobiologia* 618.1 (2009), pp. 175–198. ISSN: 00188158. DOI: 10.1007/s10750-008-9575-x.

[31] Anouk N. Blauw et al. "Predictability and environmental drivers of chlorophyll fluctuations vary across different time scales and regions of the North Sea". In: *Progress in Oceanography* 161 (Feb. 2018), pp. 1–18. ISSN: 00796611. DOI: 10.1016/j.pocean.2018.01.005.

[32] Simon J. Bonner, Nathaniel K. Newlands, and Nancy E. Heckman. "Modeling regional impacts of climate teleconnections using functional data analysis". In: *Environmental and Ecological Statistics* 21.1 (Mar. 2014), pp. 1–26. ISSN: 13528505. DOI: 10.1007/S10651-013-0241-8/FIGURES/6. URL: https://link.springer.com/article/10.1007/s10651-013-0241-8.

[33] Glenn W. Brier. "Verification of Forecasts Expressed in Terms of Probability". In: *Monthly Weather Review* 78.1 (Jan. 1950), pp. 1–3. ISSN: 0027-0644. DOI: 10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2.

[34] R. Bro, A. K. Smilde, and S. De Jong. "On the difference between low-rank and subspace approximation: Improved model for multi-linear PLS regression". In: *Chemometrics and Intelligent Laboratory Systems* 58.1 (Sept. 2001), pp. 3–13. ISSN: 01697439. DOI: 10.1016/S0169-7439(01)00134-4. URL: http://www.sciencedirect.com/science/article/pii/S0169743901001344.

[35] Rasmus Bro. "Multi-way analysis in the food industry. Models, algorithms, and applications." PhD thesis. Universiteit van Amsterdam, 1998, p. 300.

[36] Rasmus Bro. "Multiway calibration. Multilinear PLS". In: *Journal of Chemometrics* 10.1 (1996), pp. 47–61. ISSN: 1099-128X. DOI: 10.1002/(SICI)1099-128X(199601)10:1<47::AID-CEM400>3.0.CO;2-C.

[37] Rasmus Bro. "PARAFAC. Tutorial and applications". In: *Chemometrics and Intelligent Laboratory Systems* 38.2 (1997), pp. 149–171. ISSN: 01697439. DOI: 10.1016/S0169-7439(97)00032-4. arXiv: arXiv:1408.1149.

[38] Rasmus Bro. "Review on multiway analysis in chemistry - 2000-2005". In: *Critical Reviews in Analytical Chemistry* 36.3-4 (2006), pp. 279–293. ISSN: 10408347. DOI: 10.1080/10408340600969965.

[39] Jochen Bröcker. "Evaluating raw ensembles with the continuous ranked probability score". In: *Quarterly Journal of the Royal Meteorological Society* 138.667 (July 2012), pp. 1611–1617. ISSN: 00359009. DOI: 10.1002/qj.1891. URL: http://doi.wiley.com/10.1002/qj.1891.

[40] Sarah R. Brody, M. Susan Lozier, and John P. Dunne. "A comparison of methods to determine phytoplankton bloom initiation". In: *Journal of Geophysical Research: Oceans* 118.5 (May 2013), pp. 2345–2357. ISSN: 21699275. DOI: 10.1002/jgrc.20167. URL: http://doi.wiley.com/10.1002/jgrc.20167.

[41] J. W. Campbell. "The lognormal distribution as a model for bio-optical variability in the sea". In: *Journal of Geophysical Research* 100.C7 (July 1995), pp. 13237–13254. ISSN: 01480227. DOI: 10.1029/95jc00458. URL: https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/95JC00458%20https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/95JC00458%20https://agupubs.onlinelibrary.wiley.com/doi/10.1029/95JC00458.

[42] Iñigo Capellán-Pérez et al. "Likelihood of climate change pathways under uncertainty on fossil fuel resource availability". In: *Energy and Environmental Science* 9.8 (Aug. 2016), pp. 2482–2496. ISSN: 17545706. DOI: 10.1039/c6ee01008c.

[43] C. K. Carter and R. Kohn. “On Gibbs Sampling for State Space Models”. In: *Biometrika* 81.3 (Aug. 1994), p. 541. ISSN: 00063444. DOI: 10.2307/2337125.

[44] Fateh Chebana, Sophie Dabo-Niang, and Taha B.M.J. Ouarda. “Exploratory functional flood frequency analysis and outlier detection”. In: *Water Resources Research* 48.4 (Apr. 2012), p. 4514. ISSN: 1944-7973. DOI: 10.1029/2011WR011040. URL: https://onlinelibrary.wiley.com/doi/full/10.1029/2011WR011040%20https://onlinelibrary.wiley.com/doi/abs/10.1029/2011WR011040%20https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2011WR011040.

[45] J. Chen, F.P. Brissette, and R. Leconte. “WeaGETS – a Matlab-based daily scale weather generator for generating precipitation and temperature”. In: *Procedia Environmental Sciences* 13 (2012), pp. 2222–2235. ISSN: 18780296. DOI: 10.1016/j.proenv.2012.01.211.

[46] Hyungmin Cho, U. Jin Choi, and Heekyung Park. “Deep learning application to time-series prediction of daily chlorophyll-a concentration”. In: *WIT Transactions on Ecology and the Environment* 215 (Oct. 2018), pp. 157–163. ISSN: 17433541. DOI: 10.2495/EID180141. URL: www.witpress.com,.

[47] Sy-Miin Chow et al. “Dynamic Factor Analysis Models With Time-Varying Parameters”. In: *Multivariate Behavioral Research* 46.2 (Apr. 2011), pp. 303–339. ISSN: 0027-3171. DOI: 10.1080/00273171.2011.563697.

[48] William S. Cleveland. “Robust locally weighted regression and smoothing scatterplots”. In: *Journal of the American Statistical Association* 74.368 (1979), pp. 829–836. ISSN: 1537274X. DOI: 10.1080/01621459.1979.10481038.

[49] Harriet Cole et al. “Mind the gap: The impact of missing data on the calculation of phytoplankton phenology metrics”. In: *Journal of Geophysical Research: Oceans* 117.C8 (Aug. 2012), n/a–n/a. ISSN: 01480227. DOI: 10.1029/2012JC008249. URL: http://doi.wiley.com/10.1029/2012JC008249.

[50] European Commission et al. *Algal bloom and its economic impact*. Publications Office, 2016. DOI: doi/10.2788/660478.

[51] D. H. Cushing. “Plankton production and year-class strength in fish populations: An update of the match/mismatch hypothesis”. In: *Advances in Marine Biology* 26.C (Jan. 1990), pp. 249–293. ISSN: 00652881. DOI: 10.1016/S0065-2881(08)60202-3.

[52] Domenico D’Alelio et al. “Machine learning identifies a strong association between warming and reduced primary productivity in an oligotrophic ocean gyre”. In: *Scientific Reports* 10.1 (2020). ISSN: 20452322. DOI: 10.1038/s41598-020-59989-y.

[53] Didier Dacunha-Castelle, T.T.H Hoang, and Sylvie Parey. “Modeling of air temperatures: preprocessing and trends, reduced stationary process, extremes, simulation”. In: *Journal de la Société Française de Statistique* 156.1 (2015), pp. 138–168. URL: http://journal-sfds.math.cnrs.fr/index.php/J-SFdS/article/view/421.

[54] T Daggers. “Validation of marine primary production model fo the North Sea using in-situ data”. PhD thesis. Utrecht University, Utrecht, the Netherlands, 2013. URL: https://dspace.library.uu.nl/bitstream/handle/1874/280657/Thesis%7B%5C_%7DT.D.Daggers.pdf?sequence=1.

[55] C. Dalelane et al. “A Pragmatic Approach to Build a Reduced Regional Climate Projection Ensemble for Germany Using the EURO-CORDEX 8.5 Ensemble”. In: *Journal of Applied Meteorology and Climatology* 57.3 (Mar. 2018), pp. 477–491. ISSN: 1558-8424. DOI: 10.1175/JAMC-D-17-0141.1. URL: https://journals.ametsoc.org/view/journals/apme/57/3/jamc-d-17-0141.1.xml.

[56] F Danuso. “Climak: a Stochastic Model For Weather Data Generation”. In: *Ital. J. Agron* 6 (2002), pp. 57–71.

[57] R. Daren Harmel and Patricia K. Smith. “Consideration of measurement uncertainty in the evaluation of goodness-of-fit in hydrologic and water quality modeling”. In: *Journal of Hydrology* 337.3-4 (Apr. 2007), pp. 326–336. ISSN: 00221694. DOI: 10.1016/j.jhydrol.2007.01.043.

[58] Felipe de L. L. de Amorim et al. “Evaluation of Machine Learning predictions of a highly resolved time series of Chlorophyll-a concentration”. In: *bioRxiv* (Jan. 2021), p. 2021.05.12.443749. DOI: 10.1101/2021.05.12.443749. URL: http://biorxiv.org/content/early/2021/05/12/2021.05.12.443749.abstract.

[59] Xavier Desmit et al. “Changes in chlorophyll concentration and phenology in the North Sea in relation to de-eutrophication and sea surface warming”. In: *Limnology and Oceanography* 65.4 (Apr. 2020), pp. 828–847. ISSN: 0024-3590. DOI: 10.1002/lno.11351. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/lno.11351.

[60] Charles R. Doss. “Concave regression: value-constrained estimation and likelihood ratio-based inference”. In: *Mathematical Programming* 174.1-2 (Mar. 2019), pp. 5–39. ISSN: 14364646. DOI: 10.1007/s10107-018-1338-5. URL: https://doi.org/10.1007/s10107-018-1338-5.

[61] Cintia B.S. Dotto et al. “Comparison of different uncertainty techniques in urban stormwater quantity and quality modelling”. In: *Water Research* 46.8 (May 2012), pp. 2545–2558. ISSN: 0043-1354. DOI: 10.1016/J.WATRES.2012.02.009. URL: https://www.sciencedirect.com/science/article/pii/S0043135412000978.

[62] J. L. Dufresne et al. “Climate change projections using the IPSL-CM5 Earth System Model: From CMIP3 to CMIP5”. In: *Climate Dynamics* 40.9-10 (2013). ISSN: 14320894. DOI: 10.1007/s00382-012-1636-1.

[63] Kyle F. Edwards et al. “Phytoplankton growth and the interaction of light and temperature: A synthesis at the species and community level”. In: *Limnology and Oceanography* 61.4 (July 2016), pp. 1232–1244. ISSN: 19395590. DOI: 10.1002/lno.10282.

[64] Martin Edwards and Anthony J. Richardson. “Impact of climate change on marine pelagic phenology and trophic mismatch”. In: *Nature* 430.7002 (2004). ISSN: 00280836. DOI: 10.1038/nature02808.

[65] Ghada Y El Serafy et al. “Improving the Description of the Suspended Particulate Matter Concentrations in the Southern North Sea through Assimilating Remotely Sensed Data”. In: *Ocean Sci. J* 46.3 (2011), pp. 179–204. ISSN: 1738-5261. DOI: 10.1007/s12601-011-0015-x. URL: http://dx.doi.org/10.1007/s12601-011-0015-x%7B%5C%%7D5Cnwww.springerlink.com.

[66] O. Essenwanger. “Correlation of wind direction observations and other surface elements”. In: *Geofisica pura e applicata 1962 51:1* 51.1 (Jan. 1962), pp. 251–290. ISSN: 1420-9136. DOI: 10.1007/BF01992668. URL: https://link.springer.com/article/10.1007/BF01992668.

[67] Stefania Favilla et al. “Assessing feature relevance in NPLS models by VIP”. In: *Chemometrics and Intelligent Laboratory Systems* 129 (Nov. 2013), pp. 76–86. ISSN: 01697439. DOI: 10.1016/J.CHEMOLAB.2013.05.013.

[68] A. Sofia Ferreira et al. “Accuracy and precision in the calculation of phenology metrics”. In: *Journal of Geophysical Research: Oceans* 119.12 (Dec. 2014), pp. 8438–8453. ISSN: 21699275. DOI: 10.1002/2014JC010323. URL: http://doi.wiley.com/10.1002/2014JC010323.

[69] Jerome Fiechter. “Assessing marine ecosystem model properties from ensemble calculations”. In: *Ecological Modelling* 242 (Sept. 2012), pp. 164–179. ISSN: 03043800. DOI: 10.1016/j.ecolmodel.2012.05.016.

[70] Christopher B. Field et al. “Primary production of the biosphere: Integrating terrestrial and oceanic components”. In: *Science* 281.5374 (July 1998), pp. 237–240. ISSN: 00368075. DOI: 10.1126/SCIENCE.281.5374.237/SUPPL_FILE/982246E_THUMB.GIF. URL: https://www.science.org/doi/abs/10.1126/science.281.5374.237.

[71] Eelke O Folmer et al. “Large-Scale Spatial Dynamics of Intertidal Mussel (Mytilus edulis L.) Bed Coverage in the German and Dutch Wadden Sea”. In: *Ecosystems* 17.3 (2014), pp. 550–566. ISSN: 1435-0629. DOI: 10.1007/s10021-013-9742-4. URL: https://doi.org/10.1007/s10021-013-9742-4.

[72] Gabriele Freni and Giorgio Mannina. “Bayesian approach for uncertainty quantification in water quality modelling: The influence of prior distribution”. In: *Journal of Hydrology* 392.1-2 (Oct. 2010), pp. 31–39. ISSN: 0022-1694. DOI: 10.1016/J.JHYDROL.2010.07.043. URL: https://www.sciencedirect.com/science/article/pii/S0022169410004828.

[73] Kevin D. Friedland et al. “Spring bloom dynamics and zooplankton biomass response on the US Northeast Continental Shelf”. In: *Continental Shelf Research* 102 (July 2015), pp. 47–61. ISSN: 18736955. DOI: 10.1016/j.csr.2015.04.005.

[74] Y. F. Friocourt et al. “Marine downscaling of a future climate scenario in the North Sea and possible effects on dinoflagellate harmful algal blooms”. In: *Food Additives and Contaminants - Part A Chemistry, Analysis, Control, Exposure and Risk Assessment* 29.10 (Oct. 2012), pp. 1630–1646. ISSN: 19440049. DOI: 10.1080/19440049.2012.714079.

[75] Masami Fujiwara and Michael S. Mohr. “Identifying environmental signals from population abundance data using multivariate time-series analysis”. In: *Oikos* 118.11 (Nov. 2009), pp. 1712–1720. ISSN: 00301299. DOI: 10.1111/j.1600-0706.2009.17570.x. URL: http://doi.wiley.com/10.1111/j.1600-0706.2009.17570.x.

[76] D Gamerman and H F Lopes. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference, Second Edition*. Vol. 1. Taylor & Francis, 2006, p. 245. ISBN: 9781584885870.

[77] P. Garnesson, A. Mangin, and M. Bretagnon. *OCEAN COLOUR PRODUCTION CENTRE Satellite Observation Copernicus-GlobColour Products*. Tech. rep. Copernicus Marine Environemnt Monitoring Service, 2020. URL: https://resources.marine.copernicus.eu/documents/QUID/CMEMS-OC-QUID-009-030-032-033-037-081-082-083-085-086-098.pdf.

[78] Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel / Hierarchical Models*. Analytical Methods for Social Research. Cambridge University Press, 2006. DOI: 10.1017/CBO9780511790942.

[79] Marco A. Giorgetta et al. “Climate and carbon cycle changes from 1850 to 2100 in MPI-ESM simulations for the Coupled Model Intercomparison Project phase 5”. In: *Journal of Advances in Modeling Earth Systems* 5.3 (2013). DOI: 10.1002/jame.20038.

[80] Eric Goberville et al. “Uncertainties in the projection of species distributions related to general circulation models”. In: *Ecology and Evolution* 5.5 (Mar. 2015), pp. 1100–1116. ISSN: 20457758. DOI: 10.1002/ece3.1411.

[81] F. Gohin, J. N. Druon, and L. Lampert. “A five channel chlorophyll concentration algorithm applied to Sea WiFS data processed by SeaDAS in coastal waters”. In: *International Journal of Remote Sensing* 23.8 (Apr. 2002), pp. 1639–1661. ISSN: 01431161. DOI: 10.1080/01431160110071879. URL: https://www.tandfonline.com/doi/abs/10.1080/01431160110071879.

[82] Igor Gómez, Sergio Molina, and Juan José Galiana-Merino. “Evaluating the influence of air pollution on solar radiation observations over the coastal region of Alicante (Southeastern Spain)”. In: *Journal of Environmental Sciences* 126 (Apr. 2023), pp. 633–643. ISSN: 1001-0742. DOI: 10.1016/J.JES.2022.05.004.

[83] Mengyi Gong et al. “State space functional principal component analysis to identify spatiotemporal patterns in remote sensing lake water quality”. In: *Stochastic Environmental Research and Risk Assessment* 35.12 (Dec. 2021), pp. 2521–2536. ISSN: 14363259. DOI: 10.1007/S00477-021-02017-W/FIGURES/7. URL: https://link.springer.com/article/10.1007/s00477-021-02017-w.

[84] Fernando González Taboada and Ricardo Anadón. "Seasonality of North Atlantic phytoplankton from space: impact of environmental forcing on a changing phenology (1998-2012)". In: *Global Change Biology* 20.3 (Mar. 2014), pp. 698–712. ISSN: 13541013. DOI: 10.1111/gcb.12352. URL: http://doi.wiley.com/10.1111/gcb.12352.

[85] Ulf Gräwe et al. "Seasonal variability in M2 and M4 tidal constituents and its implications for the coastal residual sediment transport". In: *Geophysical Research Letters* 41.15 (Aug. 2014), pp. 5563–5570. ISSN: 19448007. DOI: 10.1002/2014GL060517. URL: https://agupubs.onlinelibrary.wiley.com/doi/full/10.1002/2014GL060517%20https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2014GL060517%20https://agupubs.onlinelibrary.wiley.com/doi/10.1002/2014GL060517.

[86] Piet Groeneboom and Geurt Jongbloed. *Nonparametric estimation under shape constraints: Estimators, algorithms and asymptotics*. Cambridge University Press, Jan. 2014, pp. 1–416. ISBN: 9781139020893. DOI: 10.1017/CBO9781139020893. URL: /core/books/nonparametric-estimation-under-shape-constraints/881B662EEF5B5266E5E4D80E6153FCDA.

[87] Piet Groeneboom, Geurt Jongbloed, and Jon A. Wellner. "Estimation of a convex function: Characterizations and asymptotic theory". In: *Annals of Statistics* 29.6 (Dec. 2001), pp. 1653–1698. ISSN: 00905364. DOI: 10.1214/aos/1015345958. URL: https://projecteuclid.org/euclid.aos/1015345958.

[88] Sjoerd Groeskamp, Janine J. Nauw, and Leo R. M. Maas. "Observations of estuarine circulation and solitary internal waves in a highly energetic tidal channel". In: *Ocean Dynamics 2011 61:11* 61.11 (Aug. 2011), pp. 1767–1782. ISSN: 1616-7228. DOI: 10.1007/S10236-011-0455-Y. URL: https://link.springer.com/article/10.1007/s10236-011-0455-y.

[89] Tian Guo et al. "Impact of number of realizations on the suitability of simulated weather data for hydrologic and environmental applications". In: *Stochastic Environmental Research and Risk Assessment* 32.8 (Oct. 2018), pp. 2405–2421. ISSN: 14363259. DOI: 10.1007/s00477-017-1498-5.

[90] Hoshin Vijai Gupta, Soroosh Sorooshian, and Patrice Ogou Yapo. "Status of Automatic Calibration for Hydrologic Models: Comparison with Multilevel Expert Calibration". In: *Journal of Hydrologic Engineering* 4.2 (Apr. 1999), pp. 135–143. ISSN: 1084-0699. DOI: 10.1061/(ASCE)1084-0699(1999)4:2(135).

[91] Mohanned Abduljabbar Hael. "Modeling of rainfall variability using functional principal component method: a case study of Taiz region, Yemen". In: *Modeling Earth Systems and Environment* 7.1 (Mar. 2021), pp. 17–27. ISSN: 23636211. DOI: 10.1007/S40808-020-00876-W/FIGURES/6. URL: https://link.springer.com/article/10.1007/s40808-020-00876-w.

[92] Matthew L. Hammond et al. "Regional surface chlorophyll trends and uncertainties in the global ocean". In: *Scientific Reports* 10.1 (2020). ISSN: 20452322. DOI: 10.1038/s41598-020-72073-9.

[93] Andrew C. Harvey. *Forecasting, structural time series models and the Kalman filter*. Cambridge: Cambridge University Press, 1990, p. 554. ISBN: 9781107049994. DOI: 10.1017/CBO9781107049994. URL: http://ebooks.cambridge.org/ref/id/CBO9781107049994.

[94] M. Z. Hashmi, A. Y. Shamseldin, and B. W. Melville. "Statistical downscaling of precipitation: state-of-the-art and application of bayesian multi-model approach for uncertainty assessment". In: *Hydrology and Earth System Sciences Discussions* 6.5 (Oct. 2009), pp. 6535–6579. ISSN: 1812-2116. DOI: 10.5194/hessd-6-6535-2009. URL: https://hess.copernicus.org/preprints/6/6535/2009/.

[95] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Vol. 1. 2009, pp. 1–694. ISBN: 978-0-387-84857-0. DOI: 10.1007/b94608. arXiv: arXiv:1011.1669v3. URL: http://www.springerlink.com/index/10.1007/b94608.

[96] K. Hayhoe et al. "Ch. 4: Climate Models, Scenarios, and Projections. Climate Science Special Report: Fourth National Climate Assessment, Volume I". In: (2017). DOI: 10.7930/J0WH2N54. URL: https://science2017.globalchange.gov/chapter/4/.

[97] W. Hazeleger et al. "EC-Earth V2.2: Description and validation of a new seamless earth system prediction model". In: *Climate Dynamics* 39.11 (2012). ISSN: 14320894. DOI: 10.1007/s00382-011-1228-5.

[98] Brent Henderson. "Exploring between site differences in water quality trends: a functional data analysis approach". In: *Environmetrics* 17.1 (Feb. 2006), pp. 65–80. ISSN: 1099-095X. DOI: 10.1002/ENV.750. URL: https://onlinelibrary.wiley.com/doi/full/10.1002/env.750%20https://onlinelibrary.wiley.com/doi/abs/10.1002/env.750%20https://onlinelibrary.wiley.com/doi/10.1002/env.750.

[99] O.-C. Herrmann. "Validation of Ensemble Forecast Accuracy in Integrated Modelling in the North Sea. Identification of potential improvements of the ensemble forecast accuracy for hydrodynamics and water quality applications". PhD thesis. EUROAQUAE Hydroinformatics and Water Management, Cottbus, Germany, 2015.

[100] Olle Hjerne et al. "Climate Driven Changes in Timing, Composition and Magnitude of the Baltic Sea Phytoplankton Spring Bloom". In: *Frontiers in Marine Science* 6.7 (Aug. 2019), p. 482. ISSN: 2296-7745. DOI: 10.3389/fmars.2019.00482. URL: https://www.frontiersin.org/articles/10.3389/fmars.2019.00482/full.

[101] J. Holt et al. "Physical processes mediating climate change impacts on regional sea ecosystems". In: *Biogeosciences Discussions* 11.2 (Feb. 2014), pp. 1909–1975. ISSN: 1810-6285. DOI: 10.5194/bgd-11-1909-2014.

[102] Jason Holt et al. "Potential impacts of climate change on the primary production of regional seas: A comparative analysis of five European seas". In: *Progress in Oceanography* 140 (Jan. 2016), pp. 91–115. ISSN: 00796611. DOI: 10.1016/j.pocean.2015.11.004.

[103] Jiacong Huang and Junfeng Gao. “An ensemble simulation approach for artificial neural network: An example from chlorophyll a simulation in Lake Poyang, China”. In: *Ecological Informatics* 37 (Jan. 2017), pp. 52–58. ISSN: 15749541. DOI: 10.1016/j.ecoinf.2016.11.012.

[104] Kristen R. Hunter-Cevera et al. “Physiological and ecological drivers of early spring blooms of a coastal phytoplankter”. In: *Science* 354.6310 (Oct. 2016), pp. 326–329. ISSN: 10959203. DOI: 10.1126/science.aaf8536. URL: http://dx.doi.org/10.5061/dryad.jm8s7.

[105] Andrew J. Irwin and Zoe V. Finkel. “Mining a sea of data: Deducing the environmental controls of ocean chlorophyll”. In: *PLoS ONE* 3.11 (Nov. 2008), p. 3836. ISSN: 19326203. DOI: 10.1371/journal.pone.0003836. URL: www.plosone.org.

[106] Daniela Jacob et al. “EURO-CORDEX: new high-resolution climate change projections for European impact research”. In: *Regional Environmental Change* 14.2 (Apr. 2014), pp. 563–578. ISSN: 1436-378X. DOI: 10.1007/s10113-013-0499-2. URL: https://doi.org/10.1007/s10113-013-0499-2.

[107] Rubao Ji et al. “Marine plankton phenology and life history in a changing climate: Current research and future directions”. In: *Journal of Plankton Research* 32.10 (Oct. 2010), pp. 1355–1368. ISSN: 01427873. DOI: 10.1093/plankt/fbq062. URL: http://www.agu.org/meetings/chapman/2007/bcall/.

[108] C. Jiayuan. “Framework For Assessing Uncertainty in Ecological Risk Mapping. A Case Study of North Sea in ECOSTRESS Project”. PhD thesis. UNESCO-IHE Insitute for Water Education, Delft, The Netherlands, 2015.

[109] Sijmen de Jong. “Regression coefficients in multilinear PLS”. In: *Journal of Chemometrics* 12.1 (Jan. 1998), pp. 77–81. ISSN: 0886-9383. DOI: 10.1002/(SICI)1099-128X(199801/02)12:1<77::AID-CEM496>3.0.CO;2-7.

[110] A. M.Y. Kamel et al. “Using remote sensing to enhance modelling of fine sediment dynamics in the Dutch coastal zone”. In: *Journal of Hydroinformatics* 16.2 (Mar. 2014), pp. 458–476. ISSN: 14647141. DOI: 10.2166/hydro.2013.211. URL: http://iwaponline.com/jh/article-pdf/16/2/458/387313/458.pdf.

[111] Isidora Katara et al. “Atmospheric forcing on chlorophyll concentration in the Mediterranean”. In: *Essential Fish Habitat Mapping in the Mediterranean* (2008), pp. 33–48. DOI: 10.1007/978-1-4020-9141-4_4. URL: https://link.springer.com/chapter/10.1007/978-1-4020-9141-4_4.

[112] Matthias Katzfuss, Dorit Hammerling, and Richard L. Smith. “A Bayesian hierarchical model for climate change detection and attribution”. In: *Geophysical Research Letters* 44.11 (June 2017), pp. 5720–5728. ISSN: 00948276. DOI: 10.1002/2017GL073688. URL: http://doi.wiley.com/10.1002/2017GL073688.

[113] G. Keetels et al. *Winning suppletiezand Noordzee 2013-2017 (Sand mining North Sea 2013-2017)*. Tech. rep. Delft, The Netherlands: Deltares, 2012. URL: http://publications.deltares.nl/1204963%7B%5C_%7D000%7B%5C_%7D0040.pdf.

[114] Sina Keller et al. “Hyperspectral data and machine learning for estimating CDOM, chlorophyll a, diatoms, green algae and turbidity”. In: *International Journal of Environmental Research and Public Health* 15.9 (Sept. 2018), p. 1881. ISSN: 16604601. DOI: 10.3390/ijerph15091881. URL: www.mdpi.com/journal/ijerph.

[115] A. M. G. Klein Tank and G. Lenderink. *Climate change in the Netherlands : supplements to the KNMI'06 scenarios*. Tech. rep. De Bilt, Netherlands.: KNMI, 2009.

[116] Reto Knutti and Jan Sedláček. “Robustness and uncertainties in the new CMIP5 climate model projections”. In: *Nature Climate Change* 3.4 (Apr. 2013), pp. 369–373. ISSN: 1758678X. DOI: 10.1038/nclimate1716.

[117] P. Koeller et al. “Basin-scale coherence in phenology of shrimps and phytoplankton in the North Atlantic Ocean”. In: *Science* 324.5928 (May 2009), pp. 791–793. ISSN: 00368075. DOI: 10.1126/science.1170987. URL: https://science.sciencemag.org/content/324/5928/791%20https://science.sciencemag.org/content/324/5928/791.abstract.

[118] Vladimir Krasnopolsky et al. “Adjusting neural network to a particular problem: Neural network-based empirical biological model for chlorophyll concentration in the upper ocean”. In: *Applied Computational Intelligence and Soft Computing* 2018 (2018). ISSN: 16879732. DOI: 10.1155/2018/7057363.

[119] Ingrid Kröncke et al. “Comparison of biological and ecological long-term trends related to northern hemisphere climate in different marine ecosystems”. In: *Nature Conservation 34: 311-341* 34 (May 2019), pp. 311–341. ISSN: 1314-3301. DOI: 10.3897/NATURECONSERVATION.34.30209. URL: https://natureconservation.pensoft.net/article/30209/.

[120] Roman Krzysztofowicz. “The case for probabilistic forecasting in hydrology”. In: *Journal of Hydrology* 249.1-4 (2001), pp. 2–9. ISSN: 00221694. DOI: 10.1016/S0022-1694(01)00420-6.

[121] Yi-Ming Kuo, Hone-Jay Chu, and Tsung-Yi Pan. “Temporal precipitation estimation from nearby radar reflectivity using dynamic factor analysis in the mountainous watershed - a case during Typhoon Morakot”. In: *Hydrological Processes* 28.3 (Jan. 2014), pp. 999–1008. ISSN: 08856087. DOI: 10.1002/hyp.9639. URL: http://doi.wiley.com/10.1002/hyp.9639.

[122] Sonja van Leeuwen et al. “Stratified and nonstratified areas in the North Sea: Long-term variability and biological and policy implications”. In: *Journal of Geophysical Research: Oceans* 120.7 (2015), pp. 4670–4686. DOI: https://doi.org/10.1002/2014JC010485. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2014JC010485.

[123] Martin Leutbecher. “Ensemble size: How suboptimal is less than infinity?” In: *Quarterly Journal of the Royal Meteorological Society* 145.S1 (Sept. 2019), pp. 107–128. ISSN: 0035-9009. DOI: 10.1002/qj.3387. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/qj.3387.

[124] Aleksandra Lewandowska and Ulrich Sommer. "Climate change and the spring bloom: A mesocosm study on the influence of light and temperature on phytoplankton and mesozooplankton". In: *Marine Ecology Progress Series* 405 (Apr. 2010), pp. 101–111. ISSN: 01718630. DOI: 10.3354/meps08520.

[125] Zhijie Li et al. "Generalized likelihood uncertainty estimation method in uncertainty analysis of numerical Eutrophication models: Take BLOOOM as an example". In: *Mathematical Problems in Engineering* 2013 (2013). ISSN: 1024123X. DOI: 10.1155/2013/701923.

[126] Xia Liu, Jianfeng Feng, and Yuqiu Wang. "Chlorophyll a predictability and relative importance of factors governing lake phytoplankton at different timescales". In: *Science of the Total Environment* 648 (Jan. 2019), pp. 472–480. ISSN: 18791026. DOI: 10.1016/j.scitotenv.2018.08.146.

[127] M. Llope et al. "Effects of environmental conditions on the seasonal distribution of phytoplankton biomass in the North Sea". In: *Limnology and Oceanography* 54.2 (Mar. 2009), pp. 512–524. ISSN: 00243590. DOI: 10.4319/lo.2009.54.2.0512. URL: http://doi.wiley.com/10.4319/lo.2009.54.2.0512.

[128] C. E. M. Lloyd et al. "Discharge and nutrient uncertainty: implications for nutrient flux estimation in small streams". In: *Hydrological Processes* 30.1 (Jan. 2016), pp. 135–152. ISSN: 08856087. DOI: 10.1002/hyp.10574. URL: http://doi.wiley.com/10.1002/hyp.10574.

[129] Alan R. Longhurst. *Ecological Geography of the Sea*. Elsevier Inc., 2007. ISBN: 9780 1245 552 11. DOI: 10.1016/B978-0-12-455521-1.X5000-1.

[130] Eva Lopez-Fornieles et al. "Potential of Multiway PLS (N-PLS) Regression Method to Analyse Time-Series of Multispectral Images: A Case Study in Agriculture". In: *Remote Sensing 2022, Vol. 14, Page 216* 14.1 (Jan. 2022), p. 216. ISSN: 2072-4292. DOI: 10.3390/RS14010216. URL: https://www.mdpi.com/2072-4292/14/1/216/htm%20https://www.mdpi.com/2072-4292/14/1/216.

[131] F. J. Los and M. Blaas. "Complexity, accuracy and practical applicability of different biogeochemical model versions". In: *Journal of Marine Systems* 81.1-2 (2010), pp. 44–74. ISSN: 09247963. DOI: 10.1016/j.jmarsys.2009.12.011. URL: http://dx.doi.org/10.1016/j.jmarsys.2009.12.011.

[132] F. J. Los, T. A. Troost, and J. K L Van Beek. "Finding the optimal reduction to meet all targets-Applying Linear Programming with a nutrient tracer model of the North Sea". In: *Journal of Marine Systems* 131 (2014), pp. 91–101. ISSN: 09247963. DOI: 10.1016/j.jmarsys.2013.12.001. URL: http://dx.doi.org/10.1016/j.jmarsys.2013.12.001.

[133] F. J. Los, M. T. Villars, and M. W.M. Van der Tol. "A 3-dimensional primary production model (BLOOM/GEM) and its applications to the (southern) North Sea (coupled physical-chemical-ecological model)". In: *Journal of Marine Systems* 74.1-2 (2008), pp. 259–294. ISSN: 09247963. DOI: 10.1016/j.jmarsys.2008.01.002.

[134] Daniel P. Loucks and Eelco van Beek. *Water resource systems planning and management: An introduction to methods, models, and applications*. Springer International Publishing, Mar. 2017, pp. 1–624. ISBN: 9783319442341. DOI: 10.1007/978-3-319-44234-1.

[135] Haiping Lu, Konstantinos N. Plataniotis, and Anastasios Venetsanopoulos. *Multilinear Subspace Learning: Dimensionality Reduction of Multidimensional Data*. 2013, p. 296. ISBN: 9781439857243.

[136] Qunying Luo. "Necessity for post-processing dynamically downscaled climate projections for impact and adaptation studies". In: *Stochastic Environmental Research and Risk Assessment* 30.7 (Oct. 2016), pp. 1835–1850. ISSN: 14363259. DOI: 10.1007/s00477-016-1233-7. URL: www.longpaddock.qld.gov.au/silo,.

[137] Wenguang Luo et al. "Comparing artificial intelligence techniques for chlorophyll-a prediction in US lakes". In: *Environmental Science and Pollution Research* 26.29 (Oct. 2019), pp. 30524–30532. ISSN: 16147499. DOI: 10.1007/s11356-019-06360-y. URL: https://doi.org/10.1007/s11356-019-06360-y.

[138] Helmut Lütkepohl. *New introduction to multiple time series analysis*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 1–764. ISBN: 3540401725. DOI: 10.1007/978-3-540-27752-1. arXiv: arXiv:1011.1669v3. URL: http://link.springer.com/10.1007/978-3-540-27752-1.

[139] Saikat Maitra and Jun Yan. "Principle Component Analysis and Partial Least Squares: Two Dimension Reduction Techniques for Regression". In: *Casualty Actuarial Society, 2008 Discussion Paper Program* (2008), pp. 79–90.

[140] Douglas Maraun et al. "Towards process-informed bias correction of climate change simulations". In: 7.11 (Nov. 2017), pp. 764–773. ISSN: 17586798. DOI: 10.1038/nclimate3418. URL: https://www.nature.com/articles/nclimate3418.

[141] Sílvia Mas et al. "Application of chemometric methods to environmental analysis of organic pollutants: A review". In: *Talanta* 80.3 (Jan. 2010), pp. 1052–1067. ISSN: 0039-9140. DOI: 10.1016/J.TALANTA.2009.09.044.

[142] Amy McGovern et al. "Identifying predictive multi-dimensional time series motifs: An application to severe weather prediction". In: *Data Mining and Knowledge Discovery* 22.1-2 (Jan. 2011), pp. 232–258. ISSN: 13845810. DOI: 10.1007/s10618-010-0193-7. URL: https://link.springer.com/article/10.1007/s10618-010-0193-7.

[143] Abigail McQuatters-Gollop and Jan E. Vermaat. "Covariance among North Sea ecosystem state indicators during the past50 years — Contrasts between coastal and open waters". In: *Journal of Sea Research* 65.2 (Feb. 2011), pp. 284–292. ISSN: 1385-1101. DOI: 10.1016/J.SEARES.2010.12.004.

[144] Sushant Mehan et al. "Comparative Study of Different Stochastic Weather Generators for Long-Term Climate Data Simulation". In: *Climate* (2017). ISSN: 2225-1154. DOI: 10.3390/cli5020026.

[145] E. Meijgaard et al. "The KNMI regional atmospheric model RACMO version 2.1". In: *Tech. Rep. 302, KNMI* (Jan. 2008).

[146] Thomas Mendlik and Andreas Gobiet. “Selecting climate simulations for impact studies based on multivariate patterns of climate change”. In: *Climatic Change* 135.3-4 (Apr. 2016), pp. 381–393. ISSN: 15731480. DOI: 10.1007/S10584-015-1582-0/FIGURES/5. URL: https://link.springer.com/article/10.1007/s10584-015-1582-0.

[147] Lőrinc Mészáros and Ghada El Serafy. “Setting up a water quality ensemble forecast for coastal ecosystems: a case study of the southern North Sea”. In: *Journal of Hydroinformatics* (Mar. 2018), jh2018027. ISSN: 1464-7141. DOI: 10.2166/hydro.2018.027. URL: http://jh.iwaponline.com/lookup/doi/10.2166/hydro.2018.027.

[148] Lőrinc Mészáros et al. “A Bayesian stochastic generator to complement existing climate change scenarios: supporting uncertainty quantification in marine and coastal ecosystems”. In: *Stochastic Environmental Research and Risk Assessment* 35.3 (Mar. 2021), pp. 719–736. ISSN: 14363259. DOI: 10.1007/S00477-020-01935-5/FIGURES/17. URL: https://link.springer.com/article/10.1007/s00477-020-01935-5.

[149] Lőrinc Mészáros et al. “Climate change induced trends and uncertainties in phytoplankton spring bloom dynamics”. In: *Frontiers in Marine Science* 8 (2021), p. 1067. ISSN: 22967745. DOI: 10.3389/FMARS.2021.669951/BIBTEX.

[150] Jeff Harrison Mike West. *Bayesian Forecasting and Dynamic Models*. 1997, p. 680. ISBN: 0387947256. DOI: 10.1007/b98971. arXiv: arXiv:1011.1669v3.

[151] Joao Morim et al. “Robustness and uncertainties in global multivariate wind-wave climate projections”. In: *Nature Climate Change* 9.9 (Sept. 2019), pp. 711–718. ISSN: 17586798. DOI: 10.1038/s41558-019-0542-5. URL: https://www.nature.com/articles/s41558-019-0542-5.

[152] M. R. Najafi and H. Moradkhani. “A hierarchical Bayesian approach for the analysis of climate change impact on runoff extremes”. In: *Hydrological Processes* 28.26 (Dec. 2014), pp. 6292–6308. ISSN: 08856087. DOI: 10.1002/hyp.10113. URL: http://doi.wiley.com/10.1002/hyp.10113.

[153] Carlos Navarro-Racines et al. “High-resolution and bias-corrected CMIP5 projections for climate change impact assessments”. In: *Scientific Data* 7.1 (Dec. 2020), pp. 1–14. ISSN: 20524463. DOI: 10.1038/s41597-019-0343-8. URL: https://doi.org/10.1038/s41597-019-0343-8.

[154] Brian C. O’Neill et al. “IPCC reasons for concern regarding climate change risks”. In: *Nature Climate Change* 7.1 (Jan. 2017), pp. 28–37. ISSN: 17586798. DOI: 10.1038/nclimate3179. URL: https://www.nature.com/articles/nclimate3179.

[155] J.E. O’Reilly et al. “Ocean color chlorophyll a algorithms for SeaWiFS, OC2, and OC4: Version 4”. In: *SeaWiFS Postlaunch Calibration and Validation Analyses* 11 (2000), pp. 9–23.

[156] J.E. O’Reilly et al. “Ocean color chlorophyll algorithms for SeaWiFS”. In: *Journal of Geophysical Research Atmospheres* 103.1998 (1998), pp. 24937–24953.

[157] Natalie Packham and Wolfgang M Schmidt. “Latin hypercube sampling with dependence and applications in finance”. In: *Journal of Computational Finance* 13.3 (2008), pp. 81–111.

[158] Sylvie Parey. “Generating a Set of Temperature Time Series Representative of Recent Past and Near Future Climate”. In: *Frontiers in Environmental Science* 7.6 (June 2019), p. 99. ISSN: 2296-665X. DOI: 10.3389/fenvs.2019.00099. URL: https://www.frontiersin.org/article/10.3389/fenvs.2019.00099/full.

[159] N. C. Perez. “Validation of ecological models: DELFT3d-GEM”. PhD thesis. EUROAQUAE Hydroinformatics and Water Management, Cottbus, Germany, 2015.

[160] S.W.M. Peters et al. *Atlas of chlorophyll-a concentration for the North Sea (based on MERIS imagery of 2003)*. E-05/01. Instituut voor Milieuvraagstukken, 2005. ISBN: 9051920261.

[161] Catharina Johanna Maria Philippart et al. “Long-term field observations on seasonality in chlorophyll-a concentrations in a shallow coastal marine ecosystem, the Wadden Sea”. In: *Estuaries and Coasts* 33.2 (Dec. 2010), pp. 286–294. ISSN: 15592723. DOI: 10.1007/s12237-009-9236-y. URL: https://link.springer.com/article/10.1007/s12237-009-9236-y.

[162] Trevor Platt, César Fuentes-Yaco, and Kenneth T. Frank. “Marine ecology: Spring algal bloom and larval fish survival”. In: *Nature* 423.6938 (May 2003), pp. 398–399. ISSN: 00280836. DOI: 10.1038/423398b. URL: https://www.nature.com/articles/423398b.

[163] D. Pushpadas, C. Schrum, and U. Daewel. “Projected climate change impacts on North Sea and Baltic Sea: CMIP3 and CMIP5 model based scenarios”. In: *Biogeosciences Discussions* 12.15 (Aug. 2015), pp. 12229–12279. ISSN: 1726-4170. DOI: 10.5194/bgd-12-12229-2015.

[164] J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer Series in Statistics. New York, NY: Springer New York, 2005. ISBN: 978-0-387-40080-8. DOI: 10.1007/b98888. URL: http://link.springer.com/10.1007/b98888.

[165] James Ramsay, Giles Hooker, and Spencer Graves. *Functional Data Analysis with R and MATLAB*. Springer New York, 2009. DOI: 10.1007/978-0-387-98185-7.

[166] Jens Christian Refsgaard et al. “Uncertainty in the environmental modelling process - A framework and guidance”. In: *Environmental Modelling and Software* 22.11 (2007), pp. 1543–1556. ISSN: 13648152. DOI: 10.1016/j.envsoft.2007.02.004.

[167] Helen M. Regan, Mark Colyvan, and Mark A. Burgman. “A TAXONOMY AND TREATMENT OF UNCERTAINTY FOR ECOLOGY AND CONSERVATION BIOLOGY”. In: *Ecological Applications* 12.2 (Apr. 2002), pp. 618–628. ISSN: 1939-5582. DOI: 10.1890/1051-0761(2002)012[0618:ATATOU]2.0.CO;2.

[168] Brian J. Reich and Benjamin A. Shaby. "A hierarchical max-stable spatial model for extreme precipitation". In: *Annals of Applied Statistics* 6.4 (2012), pp. 1430–1451. ISSN: 19417330. DOI: 10.1214/12-AOAS591. URL: https://projecteuclid.org/euclid.aoas/1356629046.

[169] Anthony J. Richardson and David S. Schoeman. "Climate Impact on Plankton Ecosystems in the Northeast Atlantic". In: *Science* 305.5690 (Sept. 2004), pp. 1609–1612. ISSN: 0036-8075. DOI: 10.1126/SCIENCE.1100958. URL: https://science.sciencemag.org/content/305/5690/1609%20https://science.sciencemag.org/content/305/5690/1609.abstract.

[170] C. W. Richardson. "Stochastic simulation of daily precipitation, temperature, and solar radiation". In: *Water Resources Research* 17.1 (Feb. 1981), pp. 182–190. ISSN: 19447973. DOI: 10.1029/WR017i001p00182.

[171] Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer Texts in Statistics. New York, NY: Springer New York, 2004. ISBN: 978-1-4419-1939-7. DOI: 10.1007/978-1-4757-4145-2. URL: http://link.springer.com/10.1007/978-1-4757-4145-2.

[172] Susanne Rolinski et al. "Identifying Cardinal Dates in Phytoplankton Time Series to Enable the Analysis of Long-Term Trends". In: *Oecologia* 153.4 (2007), pp. 997–1008. URL: http://www.jstor.org/stable/40213046.

[173] Casey P. Ruff et al. "Salish Sea Chinook salmon exhibit weaker coherence in early marine survival trends than coastal populations". In: *Fisheries Oceanography* 26.6 (Nov. 2017), pp. 625–637. ISSN: 1365-2419. DOI: 10.1111/FOG.12222. URL: https://onlinelibrary.wiley.com/doi/full/10.1111/fog.12222%20https://onlinelibrary.wiley.com/doi/abs/10.1111/fog.12222%20https://onlinelibrary.wiley.com/doi/10.1111/fog.12222.

[174] K. Salacinska et al. "Sensitivity analysis of the two dimensional application of the Generic Ecological Model (GEM) to algal bloom prediction in the North Sea". In: *Ecological Modelling* 221.2 (2010), pp. 178–190. ISSN: 03043800. DOI: 10.1016/j.ecolmodel.2009.10.001.

[175] P Samuelsson et al. *The Surface Processes of the Rossby Centre Regional Atmospheric Climate Model (RCA4)*. Tech. rep. 1. SMHI, 2015. URL: https://www.smhi.se/en/publications/the-surfaceprocesses-of-the-rossby-centreregional-atmospheric-climate-modelrca4-1.89801.

[176] Benjamin M. Sanderson, Reto Knutti, and Peter Caldwell. "A Representative Democracy to Reduce Interdependency in a Multimodel Ensemble". In: *Journal of Climate* 28.13 (July 2015), pp. 5171–5194. ISSN: 0894-8755. DOI: 10.1175/JCLI-D-14-00362.1. URL: https://journals.ametsoc.org/view/journals/clim/28/13/jcli-d-14-00362.1.xml.

[177] Simo Särkkä. *Bayesian Filtering and Smoothing*. Institute of Mathematical Statistics Textbooks. Cambridge University Press, 2013. DOI: 10.1017/CBO9781139344203.

[178] Bertrand Saulquin, Francis Gohin, and Odile Fanton d'Andon. "Interpolated fields of satellite-derived multi-algorithm chlorophyll-a estimates at global and European scales in the frame of the European Copernicus-Marine Environment Monitoring Service". In: *Journal of Operational Oceanography* 12.1 (Jan. 2019), pp. 47–57. ISSN: 1755-876X. DOI: 10.1080/1755876X.2018.1552358. URL: https://www.tandfonline.com/doi/full/10.1080/1755876X.2018.1552358.

[179] Corinna Schrum et al. *Projected Change—North Sea*. Springer, Cham, 2016, pp. 175–217. DOI: 10.1007/978-3-319-39745-0_6. URL: https://link.springer.com/chapter/10.1007/978-3-319-39745-0%7B%5C_%7D6.

[180] M. M. Segovia-Gonzalez, F. M. Guerrero, and P. Herranz. "Explaining functional principal component analysis to actuarial science with an example on vehicle insurance". In: *Insurance: Mathematics and Economics* 45.2 (Oct. 2009), pp. 278–285. ISSN: 01676687. DOI: 10.1016/j.insmatheco.2009.07.003.

[181] Mikhail Semenov and Elaine Barrow. *LARS-WG A Stochastic Weather Generator for Use in Climate Impact Studies*. Tech. rep. User Manual, Hertfordshire, UK, Jan. 2002.

[182] Ghada Y El Serafy et al. "Data Assimilation of Satellite Data of Suspended Particulate Matter in Delft3D-WAQ for the North Sea". In: *Proceedings of the Joint EUMETSAT/AMS Conference,* (2007), pp. 1–8.

[183] Theodore J. Smayda. "What is a bloom? A commentary". In: *Limnology and Oceanography* 42.5part2 (July 1997), pp. 1132–1136. ISSN: 00243590. DOI: 10.4319/lo.1997.42.5_part_2.1132. URL: http://doi.wiley.com/10.4319/lo.1997.42.5_part_2.1132.

[184] Age K. Smilde. "Comments on multilinear PLS". In: *Journal of Chemometrics* 11.5 (1997), pp. 367–377. ISSN: 0886-9383. DOI: 10.1002/(SICI)1099-128X(199709/10)11:5<367::AID-CEM481>3.0.CO;2-I.

[185] AK Smilde, Rasmus. Bro, and Paul. Geladi. *Multi-way Analysis with Applications in the Chemical Sciences*. J. Wiley, 2004, p. 381. ISBN: 0-471-98691-7.

[186] Kimberly Smith, Courtenay Strong, and Firas Rassoul-Agha. "A new method for generating stochastic simulations of daily air temperature for use in weather generators". In: *Journal of Applied Meteorology and Climatology* 56.4 (Apr. 2017), pp. 953–963. ISSN: 15588432. DOI: 10.1175/JAMC-D-16-0122.1. URL: www.ametsoc.org/PUBSReuseLicenses.

[187] U. Sommer et al. "The PEG-model of seasonal succession of planktonic events in fresh waters". In: *Arch. Hydrobiol* 106 (1986), pp. 433–471. URL: https://www.researchgate.net/profile/Z-Gliwicz/publication/243710329_The_PEG-model_of_seasonal_succession_of_planktonic_events_in_fresh_waters/links/0c9605374cb07052cd000000/The-PEG-model-of-seasonal-succession-of-planktonic-events-in-fresh-waters.pdf.

[188] Ulrich Sommer and Kathrin Lengfellner. "Climate change and the timing, magnitude, and composition of the phytoplankton spring bloom". In: *Global Change Biology* 14.6 (June 2008), pp. 1199–1208. ISSN: 1365-2486. DOI: 10.1111/J.1365-2486.2008.01571.X. URL: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1365-2486.2008.01571.x%20https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2486.2008.01571.x%20https://onlinelibrary.wiley.com/doi/10.1111/j.1365-2486.2008.01571.x.

[189] Ulrich Sommer et al. "Beyond the Plankton Ecology Group (PEG) Model: Mechanisms Driving Plankton Succession". In: *Annual Review of Ecology, Evolution, and Systematics* 43 (Nov. 2012), pp. 429–448. DOI: 10.1146/ANNUREV-ECOLSYS-110411-160251. URL: https://www-annualreviews-org.tudelft.idm.oclc.org/doi/abs/10.1146/annurev-ecolsys-110411-160251.

[190] Scott Steinschneider et al. "The effects of climate model similarity on probabilistic climate projections and the implications for local, risk-based adaptation planning". In: *Geophysical Research Letters* 42.12 (June 2015), pp. 5014–5044. ISSN: 00948276. DOI: 10.1002/2015GL064529. URL: http://doi.wiley.com/10.1002/2015GL064529.

[191] Jamaludin Suhaila. "Functional data visualization and outlier detection on the anomaly of El Niño southern oscillation". In: *Climate* 9.7 (July 2021). ISSN: 22251154. DOI: 10.3390/CLI9070118.

[192] Jamaludin Suhaila and Zulkifli Yusop. "Spatial and temporal variabilities of rainfall data using functional data analysis". In: *Theoretical and Applied Climatology* 129.1-2 (July 2017), pp. 229–242. ISSN: 14344483. DOI: 10.1007/S00704-016-1778-X/TABLES/4. URL: https://link.springer.com/article/10.1007/s00704-016-1778-x.

[193] Jamaludin Suhaila et al. "Comparing rainfall patterns between regions in Peninsular Malaysia via a functional data analysis technique". In: *Journal of Hydrology* 411.3-4 (Dec. 2011), pp. 197–206. ISSN: 0022-1694. DOI: 10.1016/J.JHYDROL.2011.09.043.

[194] H. U. Sverdrup. "On Conditions for the Vernal Blooming of Phytoplankton". In: *ICES Journal of Marine Science* 18.3 (Jan. 1953), pp. 287–295. ISSN: 1054-3139. DOI: 10.1093/icesjms/18.3.287. URL: https://academic.oup.com/icesjms/article-lookup/doi/10.1093/icesjms/18.3.287.

[195] Karl E. Taylor, Ronald J. Stouffer, and Gerald A. Meehl. "An Overview of CMIP5 and the Experiment Design". In: *Bulletin of the American Meteorological Society* 93.4 (Apr. 2012), pp. 485–498. ISSN: 00030007. DOI: 10.1175/BAMS-D-11-00094.1. URL: https://journals.ametsoc.org/view/journals/bams/93/4/bams-d-11-00094.1.xml.

[196] Sean J Taylor and Benjamin Letham. "Forecasting at Scale". In: *PeerJ Inc.* (Sept. 2017). DOI: 10.7287/peerj.preprints.3190v2. URL: https://doi.org/10.7287/peerj.preprints.3190v2.

[197] Claudia Tebaldi and Bruno Sansó. "Joint projections of temperature and precipitation change from multiple climate models: a hierarchical Bayesian approach". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 172.1 (Jan. 2009), pp. 83–106. ISSN: 09641998. DOI: 10.1111/j.1467-985X.2008.00545.x. URL: http://doi.wiley.com/10.1111/j.1467-985X.2008.00545.x.

[198] Claudia Tebaldi et al. "Quantifying Uncertainty in Projections of Regional Climate Change: A Bayesian Approach to the Analysis of Multimodel Ensembles". In: *Journal of Climate* 18.10 (May 2005), pp. 1524–1540. ISSN: 0894-8755. DOI: 10.1175/JCLI3363.1. URL: http://journals.ametsoc.org/doi/10.1175/JCLI3363.1.

[199] E. Todini. "A model conditional processor to assess predictive uncertainty in flood forecasting". In: *International Journal of River Basin Management* 6.2 (2008), pp. 123–137. ISSN: 18142060. DOI: 10.1080/15715124.2008.9635342.

[200] David W. Townsend et al. "Causes and consequences of variability in the timing of spring phytoplankton blooms". In: *Deep-Sea Research Part I* 41.5-6 (May 1994), pp. 747–765. ISSN: 09670637. DOI: 10.1016/0967-0637(94)90075-2.

[201] Dennis Trolle et al. "Advancing projections of phytoplankton responses to climate change through ensemble modelling". In: *Environmental Modelling and Software* 61 (Nov. 2014), pp. 371–379. ISSN: 13648152. DOI: 10.1016/j.envsoft.2014.01.032.

[202] Ingrid Tulp, Ralf van Hal, and Adriaan Rijnsdorp. *Effects of climate change on North Sea fish and benthos*. Tech. rep. s. Wageningen: IMARES, 2006. URL: https://www.researchgate.net/publication/40112939_Effects_of_climate_change_on_North_Sea_fish_and_benthos.

[203] Shahid Ullah and Caroline F. Finch. "Applications of functional data analysis: A systematic review". In: *BMC Medical Research Methodology* 13.1 (Mar. 2013), pp. 1–12. ISSN: 14712288. DOI: 10.1186/1471-2288-13-43/TABLES/1. URL: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/1471-2288-13-43.

[204] Per Unden et al. *HIRLAM-5 Scientific Documentation*. 2002. URL: https://www.researchgate.net/publication/278962772_HIRLAM-5_scientific_documentation.

[205] Laura Uusitalo et al. "An overview of methods to evaluate uncertainty of deterministic models in decision support". In: *Environmental Modelling and Software* 63 (2015), pp. 24–31. ISSN: 13648152. DOI: 10.1016/j.envsoft.2014.09.017. URL: http://dx.doi.org/10.1016/j.envsoft.2014.09.017.

[206] Hendrik Jan Van Der Woerd and Reinold Pasterkamp. "HYDROPT: A fast and flexible method to retrieve chlorophyll-a from multispectral satellite observations of optically complex coastal waters". In: *Remote Sensing of Environment* 112.4 (Apr. 2008), pp. 1795–1807. ISSN: 00344257. DOI: 10.1016/j.rse.2007.09.001.

[207] Hans Van Haren, David K. Mills, and Lambertus P.M.J. Wetsteyn. "Detailed observations of the phytoplankton spring bloom in the stratifying central North Sea". In: *Journal of Marine Research* 56.3 (1998), pp. 655–680. DOI: 10.1357/002224098765213621.

[208] Tiffany C. Vance et al. "From the Oceans to the Cloud: Opportunities and challenges for data, models, computation and workflows". In: *Frontiers in Marine Science* 6.4 (Jan. 2019), p. 211. ISSN: 22967745. DOI: 10.3389/FMARS.2019.00211/BIBTEX.

[209] M. Vargas, C. W. Brown, and M. R.P. Sapiano. "Phenology of marine phytoplankton from satellite ocean color measurements". In: *Geophysical Research Letters* 36.1 (Jan. 2009). ISSN: 00948276. DOI: 10.1029/2008GL036006. URL: https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2008GL036006%20https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2008GL036006%20https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2008GL036006.

[210] Andrew Verdin et al. "BayGEN: A Bayesian Space-Time Stochastic Weather Generator". In: *Water Resources Research* 55.4 (Apr. 2019), pp. 2900–2915. ISSN: 0043-1397. DOI: 10.1029/2017WR022473. URL: https://onlinelibrary.wiley.com/doi/abs/10.1029/2017WR022473.

[211] Fosco M. Vesely et al. "Quantifying Uncertainty Due to Stochastic Weather Generators in Climate Change Impact Studies". In: *Scientific Reports* 9.1 (Nov. 2019). ISSN: 20452322. DOI: 10.1038/s41598-019-45745-4.

[212] A. Voldoire et al. "The CNRM-CM5.1 global climate model: Description and basic evaluation". In: *Climate Dynamics* 40.9-10 (May 2013), pp. 2091–2121. ISSN: 09307575. DOI: 10.1007/s00382-011-1259-y. URL: http://link.springer.com/10.1007/s00382-011-1259-y.

[213] Tue M. Vu et al. "Evaluation of multiple stochastic rainfall generators in diverse climatic regions". In: *Stochastic Environmental Research and Risk Assessment* 32.5 (May 2018), pp. 1337–1353. ISSN: 14363259. DOI: 10.1007/s00477-017-1458-0.

[214] Detlef P. van Vuuren et al. "The representative concentration pathways: an overview". In: *Climatic Change* 109.1 (Aug. 2011), p. 5. ISSN: 1573-1480. DOI: 10.1007/s10584-011-0148-z. URL: https://doi.org/10.1007/s10584-011-0148-z.

[215] Karen Helen Wiltshire et al. "Resilience of North Sea phytoplankton spring bloom dynamics: An analysis of long-term data at Helgoland Roads". In: *Limnology and Oceanography* 53.4 (July 2008), pp. 1294–1302. ISSN: 00243590. DOI: 10.4319/lo.2008.53.4.1294. URL: http://doi.wiley.com/10.4319/lo.2008.53.4.1294.

[216] Monika Winder and James E. Cloern. “The annual cycles of phytoplankton biomass”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 365.1555 (Oct. 2010), pp. 3215–3226. ISSN: 0962-8436. DOI: 10.1098/rstb.2010.0125. URL: https://royalsocietypublishing.org/doi/10.1098/rstb.2010.0125.

[217] Monika Winder and Ulrich Sommer. “Phytoplankton response to a changing climate”. In: *Hydrobiologia* 698.1 (2012), pp. 5–16. ISSN: 00188158. DOI: 10.1007/s10750-012-1149-2.

[218] Monika Winder et al. “Spring phenological responses of marine and freshwater plankton to changing temperature and light conditions”. In: *Marine Biology* 159.11 (Nov. 2012), pp. 2491–2501. ISSN: 00253162. DOI: 10.1007/s00227-012-1964-z.

[219] Svante Wold et al. “Multi-way principal components-and PLS-analysis”. In: *Journal of Chemometrics* 1.1 (1987), pp. 41–56. DOI: https://doi.org/10.1002/cem.1180010107.

[220] Hongyan Xi et al. “Global retrieval of phytoplankton functional types based on empirical orthogonal functions using CMEMS GlobColour merged products and further extension to OLCI data”. In: *Remote Sensing of Environment* 240 (Apr. 2020), p. 111704. ISSN: 00344257. DOI: 10.1016/j.rse.2020.111704.

[221] Xu Xu, Carsten Lemmen, and Kai W. Wirtz. “Less Nutrients but More Phytoplankton: Long-Term Ecosystem Dynamics of the Southern North Sea”. In: *Frontiers in Marine Science* 7 (Aug. 2020), p. 662. ISSN: 2296-7745. DOI: 10.3389/fmars.2020.00662. URL: https://www.frontiersin.org/article/10.3389/fmars.2020.00662/full.

[222] Liwei Yang et al. “Quantitative effects of air pollution on regional daily global and diffuse solar radiation under clear sky conditions”. In: *Energy Reports* 8 (Nov. 2022), pp. 1935–1948. ISSN: 2352-4847. DOI: 10.1016/J.EGYR.2021.12.081.

[223] Yi Yu et al. “The variability of chlorophyll-a and its relationship with dynamic factors in the basin of the South China Sea”. In: *Journal of Marine Systems* 200 (Dec. 2019), p. 103230. ISSN: 09247963. DOI: 10.1016/j.jmarsys.2019.103230.

[224] Xinshuo Zhang et al. “Estimation of Daily Ground-Received Global Solar Radiation Using Air Pollutant Data”. In: *Frontiers in Public Health* 10 (Apr. 2022), p. 617. ISSN: 22962565. DOI: 10.3389/FPUBH.2022.860107/XML/NLM.

[225] Qibin Zhao et al. “Higher order partial least squares (HOPLS): A generalized multilinear regression method”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.7 (2013), pp. 1660–1673. ISSN: 01628828. DOI: 10.1109/TPAMI.2012.254. arXiv: 1207.1230.

[226] A F Zuur, I D Tuck, and N Bailey. “Dynamic factor analysis to estimate common trends in fisheries time series”. In: *Canadian Journal of Fisheries and Aquatic Sciences* 60.5 (2003), pp. 542–552. ISSN: 0706-652X. DOI: 10.1139/f03-030.

[227] A. F. Zuur et al. “Estimating common trends in multivariate time series using dynamic factor analysis”. In: *Environmetrics* 14.7 (Nov. 2003), pp. 665–685. ISSN: 1180-4009. DOI: 10.1002/env.611. URL: http://doi.wiley.com/10.1002/env.611.

APPENDIX

CHAPTER 4

$X \sim N(\mu, \sigma^2)$ if the random variable X has density

$$f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2 / 2\sigma^2}$$

where μ and σ^2 are the mean and variance parameters respectively.

$X \sim G(a, b)$ if the random variable X has density

$$f(x; a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}$$

where a is the shape and b is the rate parameter.

$X \sim IG(\alpha, \beta)$ if the random variable X has density

$$f(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} e^{-\frac{\beta}{x}}$$

where α is the shape and β is the scale parameter.

CHAPTER 5

The Forward Filtering Backwards Sampling (FFBS)-algorithm steps [43, 177] are defined as follows, where the dynamic and measurement models are:

$$x_k = A_{k-1} x_{k-1} + a_{k-1} + N(0, Q_{k-1})$$
$$y_k = H_k x_k + N(0, R_k)$$

where $x_k \in R^n$ is the state, $y_k \in R^m$ is the measurement, $N(0, Q_{k-1})$ is the process noise, $N(0, R_k)$ is the measurement noise, A_{k-1} is the transition matrix of the dynamic model, H_k is the measurement model matrix, and the prior Gaussian $x_0 \sim N(m_0, P_0)$. The model can be written in probabilistic terms:

$$p(x_k | x_{k-1}) = N(x_k | A_{k-1} x_{k-1} + a_{k-1}, Q_{k-1})$$
$$p(y_k | x_k) = N(y_k | H_k x_k, R_k).$$

This implies that there exist vectors m_k^- and m_k, and matrices P_k^-, P_k, S_k^- such that

$$p(x_k|y_{1:k-1}) = N(x_k|m_k^-, P_k^-)$$
$$p(x_k|y_{1:k}) = N(x_k|m_k, P_k)$$
$$p(y_k|y_{1:k-1}) = N(y_k|H_k m_k^-, S_k^-)$$

Then the prediction and update steps are the following, where the the recursion is started from the prior mean m_0 and covariance P_0.

For $k \geq 1$

Prediction steps

$$m_k^- = A_{k-1} m_{k-1} + a_{k-1}$$
$$P_k^- = A_{k-1} P_{k-1} A_{k-1}^T + Q_{k-1}$$

Update steps

$$v_k = y_k - H_k m_k^-$$
$$S_k = H_k P_k^- H_k^T + R_k$$
$$K_k = P_k^- H_k^T S_k^{-1}$$
$$m_k = m_k^- + K_k v_k$$
$$P_k = P_k^- - K_k S_k K_k^T$$

Backward sampling:

$$G_k = P_k A_k^T [P_{k+1}^-]^{-1}$$
$$m_k^s = m_k + G_k [y_{k+1} - m_{k+1}^-]$$
$$P_k^s = P_k - G_k P_{k+1}^- G_k^T$$

CURRICULUM VITÆ

Lőrinc MÉSZÁROS

15-04-1991 Born in Nyíregyháza, Hungary.

EDUCATION

2017–2022 **PhD in Applied Statistics**
Delft University of Technology, Delft, The Netherlands

Thesis:	Climate change induced uncertainties in future coastal ecosystem state
Promotor:	Prof. dr. ir. G. Jongbloed
Promotor:	dr. ir. F.H. van der Meulen
Co-promotor:	dr. ir. G.Y.H. El Serafy

2014–2016 **Master of Science in Euro Hydroinformatics and Water Management**
(Euroaqae - Erasmus Mundus Joint Master Program)
- Brandenburg Technical University, Cottbus, Germany
- Newcastle University, Newcastle, United Kingdom
- Polytech Nice Sophia, Nice, France

2009–2014 **Bachelor of Science in Civil Engineering**
Budapest University of Technology and Economics, Budapest, Hungary

EXPERIENCE

2017–present **Researcher and Advisor at Deltares**
Department of Data Science and Water Quality
Unit of Marine and Coastal Systems

List of Publications

Journal Publications Related to Thesis

- **Mészáros, L.**, van der Meulen, F., Jongbloed, G., El Serafy, G. (2022) *Coastal environmental and atmospheric data reduction in the Southern North Sea supporting ecological impact studies*, Frontiers in Marine Science

- **Mészáros, L.**, van der Meulen, F., Jongbloed, G., El Serafy, G. (2021) *Climate Change Induced Trends and Uncertainties in Phytoplankton Spring Bloom Dynamics*, Frontiers in Marine Science **8**, 1067.

- **Mészáros, L.**, van der Meulen, F., Jongbloed, G., El Serafy, G. (2021). *A Bayesian stochastic generator to complement existing climate change scenarios: supporting uncertainty quantification in marine and coastal ecosystems*, Stochastic Environmental Research and Risk Assessment **35**, 719–736.

- **Mészáros, L.**, El Serafy, G. (2018). *Setting up a water quality ensemble forecast for coastal ecosystems: a case study of the southern North Sea*, Journal of Hydroinformatics **20** (4), 846–863.

Other Journal Publications

- Shettigar, N. A., Bhattacharya, B., **Mészáros, L.**, Spinosa, A., El Serafy, G. (2020). *3D ensemble simulation of seawater temperature – an application for aquaculture operations*, Frontiers in Marine Science **7**, 1020.

- Capet, A., Fernández, V., She, J., Dabrowski, T., Umgiesser, G., Staneva, J., **Mészáros, L.**, Campuzano, F., Ursella, L., Nolan, G., El Serafy, G. (2020). *Operational Modeling Capacity in European Seas—An EuroGOOS Perspective and Recommendations for Improvement*, Frontiers in Marine Science **7**, 129.

- Tintoré J, …, **Mészáros, L.**, et al. (2019). *Challenges for Sustained Observing and Forecasting Systems in the Mediterranean Sea* , Frontiers in Marine Science **6**, 568.