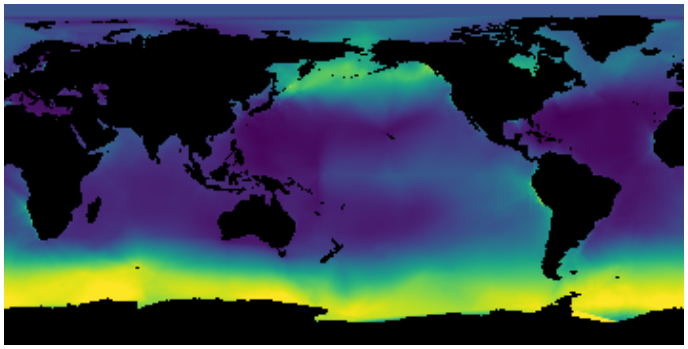
Optimization of Model Parameters, Uncertainty Quantification and Experimental Designs in Climate Research



M.Sc. Joscha Reimer

Dissertation
zur Erlangung des akademischen Grades
Doktor der Naturwissenschaften
(Dr. rer. nat.)

Technische Fakultät
Christian-Albrechts-Universität zu Kiel

Kiel
Deutschland
2019

Prüfungskommission:

1. Gutachter:	Prof. Dr. Thomas Slawig
2. Gutachter:	Prof. Dr. Andreas Oschlies
3. Prüfer:	Prof. Dr. Thomas Meurer
Vorsitz:	Prof. Dr. Wilhelm Hasselbring

Disputationstermin: 20. Februar 2020

Abstract

Several methods for optimization of model parameters, uncertainty quantification and uncertainty reduction by optimal experimental designs are studied and applied to models with different computational complexity from climate research.

The generalized least squares estimator and its special cases the weighted and the ordinary least squares estimator are described in detail together with their statistical properties. They are applied to several models using the SQP algorithm, a derivative based local optimization algorithm, in combination with the OQNLP algorithm, a globalization algorithm. This combination is proven to find model parameters well fitting to the measurement data with few function evaluations which is especially important for computationally expensive models.

The uncertainty in the estimated model parameters implied by the uncertainty in the measurement data as well as the resulting uncertainty in the model output is quantified in several ways using the first and second derivative of the model with respect to its parameters. The advantages and disadvantages of the different methods are highlighted.

The reduction of the uncertainty by additional measurements is predicted using optimal experimental design methods. It is determined how many measurements are advisable and how their conditions, like time, location and which process to be measured, should be chosen for an optimal uncertainty reduction. Robustification approaches, like sequential optimal experimental design and approximate worst case experimental designs are used to mitigate the dependency of predictions on the model parameters estimate.

A detailed statistical description of the measurements is important for the applied methods. Therefore, a statistical analysis of millions of marine measurement data is carried out. The climatological means, the variabilities, split into climatological and short scale variabilities, and correlations are estimated from the data. The associated probability distributions are examined for normality and log-normality using statistical testing and visual inspection.

To determine the correlations, an algorithm was developed that generates valid correlation matrices, i.e., positive semidefinite matrices with ones as diagonal values, from estimated correlation matrices. The algorithm tries to keep the changes as small as possible and to achieve a matrix with a low condition number. Its (worst case) execution time and memory consumption are asymptotically equal to those of the fastest algorithms to check positive semidefiniteness, making the algorithm applicable to large matrices. It is also suitable for sparse matrices because it preserves sparsity patterns. In addition to statistics, it can also be useful in numerical optimization.

In the context of this thesis, several software packages were developed or extended which are freely available as open source and extensively tested.

The results obtained from the models and data help to improve the understanding of the underlying processes. The applied methods are not limited to the application examples used here and can be applied to many data and models in climate research and beyond.

Zusammenfassung

Mehrere Methoden zur Optimierung von Modellparametern, Unsicherheitsquantifizierung und Unsicherheitsreduktion durch optimale Versuchsplanung werden untersucht und auf Modelle mit unterschiedlicher Komplexität aus der Klimaforschung angewandt.

Der verallgemeinerte Kleinste-Quadrate-Schätzer und seine Spezialfälle, der gewichtete und der gewöhnliche Kleinste-Quadrate-Schätzer, werden zusammen mit ihren statistischen Eigenschaften ausführlich beschrieben. Sie werden auf mehrere Modelle unter Verwendung des SQP-Algorithmus, einem ableitungsbasierten lokalen Optimierungsalgorithmus, in Kombination mit dem OQNLP-Algorithmus, einem Globalisierungsalgorithmus, angewendet. Diese Kombination hat sich bewährt, um gut zu den Messdaten passende Modellparameter mit wenigen Funktionsauswertungen zu finden, was besonders bei rechenintensiven Modellen wichtig ist.

Die Unsicherheit in den geschätzten Modellparametern, die sich aus der Unsicherheit in den Messdaten ergibt, sowie die daraus resultierende Unsicherheit in der Modellausgabe werden auf verschiedene Weise unter Verwendung der ersten und zweiten Ableitung des Modells bezüglich dessen Parameter quantifiziert. Die Vor- und Nachteile der verschiedenen Methoden werden aufgezeigt.

Die Reduzierung der Unsicherheit durch zusätzliche Messungen wird mit optimale Versuchsplanungsmethoden vorhergesagt. Es wird bestimmt, wie viele Messungen sinnvoll sind und wie deren Bedingungen, wie Zeit, Ort und zu messender Prozess, für eine optimale Unsicherheitsreduzierung gewählt werden sollten. Robustifizierungsansätze, wie sequentielle optimalen Versuchsplanung und approximative Worst-Case Versuchsplanung, werden verwendet, um die Abhängigkeit der Vorhersagen von der Schätzung der Modellparameter zu verringern.

Eine detaillierte statistische Beschreibung der Messungen ist für die angewandten Methoden wichtig. Daher wird eine statistische Analyse von Millionen von marinen Messdaten durchgeführt. Die klimatologischen Mittel, die Variabilitäten, unterteilt in klimatologische und kurzskalige Variabilitäten, und Korrelationen werden aus den Daten geschätzt. Die zugehörigen Wahrscheinlichkeitsverteilungen werden mittels statistischer Tests und visueller Inspektion auf Normalität und Log-Normalität untersucht.

Um die Korrelationen zu bestimmen, wurde ein Algorithmus entwickelt, der aus geschätzten Korrelationsmatrizen gültige Korrelationsmatrizen erzeugt, d.h. positive semidefinite Matrizen mit Einsen als Diagonalwerte. Der Algorithmus versucht, die Änderungen so klein wie möglich zu halten und dabei eine Matrix mit einer niedrigen Konditionszahl zu erzielen. Seine (ungünstigste) Ausführungszeit und Speicherverbrauch sind asymptotisch gleich zu denen des schnellsten Algorithmus zum Überprüfen von positiv Semidefinitheit, was den Algorithmus auf große Matrizen anwendbar macht. Er ist ebenfalls geeignet für dünnbesetzte Matrizen, da er Besetzungsstrukturen bewahrt. Neben der Statistik kann der Algorithmus auch in der numerischen Optimierung nützlich sein.

Im Rahmen dieser Arbeit wurden mehrere Softwarepakete entwickelt oder erweitert, welche als Open Source frei verfügbar sowie umfassend getestet sind.

Die Ergebnisse aus den Modellen und Daten helfen das Verständnis der zugrunde liegenden Prozesse zu verbessern. Die angewandten Methoden beschränken sich nicht auf die hier verwendeten Anwendungsbeispiele und können auf viele Daten und Modelle in der Klimaforschung und darüber hinaus angewendet werden.

Contents

1 Introduction **8**
- 1.1 Climate Change and Climate Research . . . 8
- 1.2 Marine Phosphate and its Impact on Climate Change . . . 8
- 1.3 Salt Marshes and their Role in Climate Change . . . 9
- 1.4 Computer Models in Climate Research . . . 10
- 1.5 Model Parameters and their Optimization . . . 10
- 1.6 Measurement Data and their Importance in Climate Research . . . 10
- 1.7 Uncertainty in Measurement Data . . . 11
- 1.8 Uncertainty in Climate Predictions . . . 11
- 1.9 Uncertainty Reduction by Optimal Experimental Designs . . . 12
- 1.10 Correlation Matrices Derived from Measurement Data . . . 12

2 Optimization of model parameters and experimental designs with the Optimal Experimental Design Toolbox (v1.0) exemplified by sedimentation in salt marshes **14**
- 2.1 Introduction . . . 14
- 2.2 Optimization of model parameters and experimental designs . . . 15
  - 2.2.1 The weighted least squares estimator . . . 15
  - 2.2.2 Asymptotic properties . . . 16
  - 2.2.3 Optimal experimental designs . . . 16
  - 2.2.4 Calculation of optimal experimental designs . . . 17
  - 2.2.5 Robust optimal experimental designs . . . 17
  - 2.2.6 Efficiency of experimental designs . . . 17
- 2.3 The Optimal Experimental Design Toolbox . . . 18
  - 2.3.1 Provision of the model function . . . 18
  - 2.3.2 Setup of the solver . . . 18
  - 2.3.3 Optimization of experimental designs and model parameters . . . 19
  - 2.3.4 Execution time and memory consumption . . . 19
  - 2.3.5 Changeable options . . . 19
  - 2.3.6 Help and documentation . . . 20
- 2.4 Application examples . . . 20
  - 2.4.1 The models . . . 20
  - 2.4.2 Numerical experiments . . . 21
  - 2.4.3 Accuracy of the parameter estimations . . . 22
  - 2.4.4 Efficiency for the experimental designs . . . 23
  - 2.4.5 Distribution of optimal measuring points . . . 24
- 2.5 Conclusions . . . 25

3 Approximation of Hermitian Matrices by Positive Semidefinite Matrices using Modified Cholesky Decompositions **29**
- 3.1 Introduction . . . 30
  - 3.1.1 Objectives of approximation algorithms . . . 30
  - 3.1.2 Existing approximation methods . . . 31
- 3.2 The approximation algorithm . . . 32
  - 3.2.1 The MATRIX and the DECOMPOSITION algorithm . . . 32
  - 3.2.2 Representation of the approximation matrix . . . 35
  - 3.2.3 Positive semidefinite approximation . . . 41
  - 3.2.4 Diagonal values . . . 42

3.2.5 Condition number . . . 43
3.2.6 Approximation error . . . 43
3.2.7 Choice of ω and d . . . 45
3.2.8 Permutation . . . 49
3.2.9 Complexity . . . 50
3.3 Implementation and numerical experiments . . . 51
3.3.1 Implementation . . . 52
3.3.2 Comparison with other approximation algorithms . . . 52
3.4 Conclusions . . . 55
3.5 Appendix . . . 55

4 Statistical Analysis of the Phosphate Data of the World Ocean Database 2013 61
4.1 Introduction . . . 61
4.2 Methods used in the Statistical Analysis . . . 62
4.2.1 Statistical Model . . . 62
4.2.2 Climatological Means . . . 64
4.2.3 Variabilities . . . 65
4.2.4 Statistical Dependencies . . . 66
4.2.5 Probability Distributions . . . 68
4.3 Results of the Statistical Analysis . . . 68
4.3.1 Spatial and Temporal Distribution . . . 69
4.3.2 Climatological Means . . . 69
4.3.3 Variabilities . . . 71
4.3.4 Statistical Dependencies . . . 75
4.3.5 Probability Distributions . . . 79
4.4 Conclusions . . . 80
4.5 Appendix . . . 81

5 Optimization of Model Parameters, Uncertainty Quantification and Experimental Designs for a Global Marine Biogeochemical Model 85
5.1 Introduction . . . 85
5.2 Methods for Parameter Estimation, Uncertainty Quantification and Experimental Design . . . 86
5.2.1 Model Parameter Estimation . . . 86
5.2.2 Uncertainty in Parameter Estimation . . . 88
5.2.3 Uncertainty in Model Output . . . 90
5.2.4 Uncertainty Reduction using Optimal Experimental Design Methods 91
5.2.5 Computational Details . . . 93
5.3 Marine Phosphorus Cycle as Application Example . . . 94
5.3.1 Circulation Model . . . 95
5.3.2 Biogeochemical Model . . . 95
5.3.3 Model Parameters . . . 97
5.3.4 Simulation and Spin-up . . . 98
5.3.5 Derivative . . . 99
5.3.6 Measurement Data . . . 100
5.4 Results for the Application Example . . . 101
5.4.1 Model Parameter Estimation . . . 101
5.4.2 Uncertainty in Parameter Estimation . . . 103
5.4.3 Uncertainty in Model Output . . . 105
5.4.4 Uncertainty Reduction by Additional Measurements . . . 107

5.5 Conclusions 109

6 Conclusions and Outlooks **116**
6.1 Marine Phosphate Data and their Statistical Analysis 116
6.2 Marine Phosphorus Model 117
6.3 Salt Marsh Models 118
6.4 Optimization of Model Parameters 119
6.5 Uncertainty Quantification 120
6.6 Uncertainty Reduction by Optimal Experimental Designs 121
6.7 Approximation Algorithm for Correlation Matrices 122
6.8 Developed Software Packages 123

References **125**

Acknowledgement **130**

1 Introduction

1.1 Climate Change and Climate Research

The earth's climate system is changing dramatically. The IPCC (Intergovernmental Panel on Climate Change) summarizes the current changes in the clarity needed: "Warming of the climate system is unequivocal, and since the 1950s, many of the observed changes are unprecedented over decades to millennia. The atmosphere and ocean have warmed, the amounts of snow and ice have diminished, and sea level has risen." [28, 1.1]

These changes have a drastic impact on nature and humankind [28, 1.3]. The risk of extinction increases for a large fraction of species [28, 2.3]. The frequency and intensity of extreme precipitation, storms and storm surges, inland and coastal flooding, sea level rise and land loss as well as drought and water scarcity will increase as climate change progresses [28, 2.3]. Food security will worsen over time and health risks will increase. [28, 2.3]. Abrupt and irreversible changes become more likely with progressing global warming [28, 2.4].

Since the world's well-being is at stake, there is an urgent need for action [28, 3.2]. 195 countries have agreed on the Paris Agreement [42] in 2015 to take action to limit the global warming to 1.5°C above pre-industrial levels. However, their efforts must be significantly increased to achieve this objective [29]. More than 11,000 scientists have pointed this out as well in 2019. They have warned that "planet Earth is facing a climate emergency" and that effective countermeasures are indispensable [58]. Another 11,000 have joined this warning within a few days [2].

For being able to take effective actions, it is essential to understand the causes and consequences of the climate change as well as the progress of transition and how it can be mitigated. Sound knowledge has been gained in this respect in recent decades [23, 24, 25, 26, 28, 51]. However, there are still many uncertainties and open questions [28, 51], proving that climate research is still a crucial and ongoing research.

This thesis aims to contribute to this research by demonstrating the applicability and usefulness of selected methods for data analysis, optimization of model parameters, uncertainty quantification and uncertainty reduction in climate research. For this purpose, salt marshes and the marine phosphorus cycle, as two parts of the earth system, were chosen as application examples. In addition to the presented methods and their assessment, the results obtained for these application examples represent another important contribution of this thesis.

1.2 Marine Phosphate and its Impact on Climate Change

The main reason for global warming is the anthropogenic increase in greenhouse gas emissions [28, 1.3.1]. Its contribution from 1951 to 2010 is estimated to be between 0.5°C to 1.3°C [28, 1.3.1]. Within the anthropogenic greenhouse gas emissions, CO_2 emissions are the most crucial. They account for 76% of total global warming due to anthropogenic greenhouse gas emissions in 2010. [28, 1.2.2] The cumulative CO_2 emissions will determine mainly the global warming in the coming decades and eventually even centuries [28, 2.1].

The CO_2 concentration in the atmosphere is strongly influenced by the ocean. The ocean has already absorbed about 30% of the emitted anthropogenic CO_2 [28, 1.2.2], by exchanging CO_2 between the surface ocean and the atmosphere due to CO_2 partial pressure differences [27, 6.1.1] [6, 3.2].

The carbon saturation of the surface water, therefore, is crucial for the CO_2 absorption

capacity of the ocean. The amount of dissolved carbon in the surface ocean in turn is influenced on the carbon transport within the ocean, which is determined by the solubility pump, the biological pump and the marine carbonate pump [27, 6.1.1].

In the following we focus on the biological pump which is defined as: "The process of transporting carbon from the ocean's surface layers to the deep ocean by the primary production of marine phytoplankton, which converts dissolved inorganic carbon (DIC) and nutrients into organic matter through photosynthesis. This natural cycle is limited primarily by the availability of light and nutrients such as phosphate, nitrate and silicic acid, and micronutrients, such as iron." [27, Glossary]

The phosphate concentrations, especially in the euphotic zone where enough light for photosynthesis is available, are therefore important for the CO_2 uptake of the ocean and thus also for the atmospheric CO_2 concentration and the intensity of global warming [6, 4].

A statistical analysis of millions of phosphate measurements in the ocean is provided in Section 4. The methods used for the statistical analysis are presented in Subsection 4.2. The results of the analysis of phosphate measurements are illustrated in Subsection 4.3. They are more comprehensive than results from previous statistical analyses and improve the understanding of phosphate concentrations in the global ocean.

A model describing the phosphate and dissolved organic phosphorus concentrations as well as the phytoplankton production in the global ocean is the subject of Section 5. A detailed description of the model is given in Subsection 5.3. More realistic model parameters were determined and associated uncertainties quantified. The applied methods are described in Subsection 5.2 and the results are illustrated in Subsection 5.4. The obtained results provide a better understanding of the model and the associated uncertainties, and enable better predictions using this model.

1.3 Salt Marshes and their Role in Climate Change

Another part of the earth system strongly influenced by climate change, but also influencing climate change itself, is salt marshes. Salt marshes are coastal areas which are flooded and drained by salt water due to tides and mostly covered with salt-tolerant grasslike plants.

Salt marshes sequester millions of tons of carbon annually. In this respect they are one of the most effective ecosystems in the world, considering the amount of sequestered carbon in relation to the area they require [38]. Therefore, they make an important contribution to mitigating climate change. In addition, they offer many other benefits to humans, like coastal protection and water purification, food and raw materials as well as a place of recreation and tourism. [4]

However, the accelerating sea level rises [27, 3.7] endangers salt marshes. Due to their many advantages for humans, it is important to study their chances of survival [34, 32]. This depends on whether or not their gain of elevation by sedimentary deposition can compensate the sea level rise.

Models describing the change of salt marsh elevations over time are subject to Section 2. The models are described in detail in Subsection 2.4. It was investigated how the parameters of the models can be adapted to the local salt marshes with minimal measuring effort. The results are included in Subsection 2.4 as well. They allow to increase the forecasting capability of the models with minimal measuring effort. A software package developed for this purpose is introduced in Subsection 2.3. The applied methods are presented in Subsection 2.2. The software package as well as the methods are not limited to these models.

1.4 Computer Models in Climate Research

Computer models are, in addition to measurements, the primary tools in climate research [19, 1.1], [20]. They range from simple energy balance models to complex earth system models [37, 39] [27, 9.1.2], [39, 5.4]. Their complexity depends on the number and accuracy of the modeled processes and their interactions as well as the resolution used to model the processes [37, 2.1].

A disadvantage of highly complex models is that they can be executed effectively only on high-performance computers [37, 1.2.1]. Even on these an execution can take days, weeks or even months.

Due to the limited availability of computing power, simpler models are still frequently used. They can be sufficient to answer certain problems and provide insights that might otherwise be hidden by the complexity of more sophisticated models [37, 1.2.1]. Furthermore, they are indispensable for testing and extending the concepts upon which more complex models are based [37, 1.1].

The two salt marsh models, which are studied in Section 2, are rather simple models while the model in Section 5, describing the marine phosphorus cycle, is more complex. The handling of the complexity and the associated long execution time is also addressed in the context of the methods presented in this Sections.

1.5 Model Parameters and their Optimization

Climate models, regardless of their complexity, often contain parameters whose values are usually not known exactly [37, 39]. These can be, e.g., constants or averages in biological or geological processes unable to be measured directly or only very imprecisely or only with much effort. On top the parameters are often the result of simplifications, also called parameterization, of processes of too small scale or too high complexity to be modeled directly [11, 18], [27, 9.1.3.1], [31], [39, 5.3], [37, 2.5].

Values for these parameters are sometimes just guessed by experts. Often a parameter optimization, also called model calibration, fitting, tuning or parameter estimation, is carried out [27, Box 9.11], [37, 2.5] where the parameter values are determined by numerical optimization methods, or sometimes even by hand, in such a way that the resulting model output matches measurement results [3, 60].

The resulting optimal model parameters not only depend on the measurement results but also on the selected numerical optimization algorithm and the metric which quantifies the difference between the model output and the measurement results. Hence the numerical optimization algorithm [40, 15], [63, 4], [60, 13,14] and the metric [63, 3], [60, 3], [16, 44, 10, 65] should be carefully selected.

An parameter optimization was also carried out for the models considered in this thesis. The applied methods are summarized in Subsection 2.2 and 5.2. The results for this models are presented in Subsection 2.4 and 5.4

1.6 Measurement Data and their Importance in Climate Research

Measurement data are the basis of natural science research [64, Preface], [50, Foreword]. They are essential for understanding the underlying processes and are necessary for optimizing model parameters and evaluating models.

For some of the climate related processes, millions of measurement data are available. An example of this is the World Ocean Database [8], established by the Intergovernmental Oceanographic Commission, where millions of oceanographic measurement data sets are

gathered. Due to the large amount of data, it is necessary to extract and summarize the main information regarding the measured process [35, 4].

This was done for the phosphate data of the World Ocean Database 2013 [7] by a statistical analysis presented in Section 4. The applied methods, which are not limited to this data, are described in Subsection 4.2. The results of this analysis are presented in Subsection 4.3. They were also used in the Section 5 in the parameter optimization of the marine phosphorus model addressed there.

1.7 Uncertainty in Measurement Data

Common scientific sense dictates that the result of a measurement can never be the true value of the measured quantity [50, 3]. There are always deviations which have several sources, like the measurement method, the measurement device, human interactions and environmental conditions [64, 1.2].

The total deviation is called measurement error and is not known either. The measurement uncertainty quantifies, typically in a statistical way, the expected measurement error. Measurement data without this quantification are useless [50, 3].

Besides the error in the measurement result, there is also an error in the conditions of the measurements, such as the time and location of the measurements. Usually this error can be neglected in climate research because climatic processes usually change very little on small scales.

This statistical analysis in Section 4 also includes quantification of errors in the measurement results. However, errors in the measurement conditions are neglected since they are very small compared to the resolution of the analysis.

1.8 Uncertainty in Climate Predictions

Climate predictions with climate models contain several uncertainties which originate from different sources [27, 1.4.2], [61, 1.1].

On the one hand these are model errors due to approximation or imprecise representation of the underlying real processes. This may be because the real processes are not understood completely or because processes have been simplified or omitted. If the real processes would be thoroughly understood, the model error could be quantified by the difference between the real processes and the model. Since, this is usually not the case in climate research, quantifying the model error is often very challenging [61, 12].

Other sources are numerical errors which result from the limited accuracy of the numerical algorithms used to solve the model equations and the computer arithmetic in general. Detailed error analyses, including numerical error bounds, are available for many numerical algorithms [21].

Model inputs which are not known exactly, like model parameters, initial or boundary conditions or exogenous forcing, are another source of uncertainty [63, 5],[61, 7]. All these uncertain model inputs shall be considered as model parameters for uncertainty quantification.

Uncertainty quantification is subject to Section 2 and 5. It is addressed how to quantify uncertainties in the model parameters resulting from uncertainties in the measurements as well as uncertainties in the model output resulting from uncertainties in the model parameters. In Subsection 2.2 and 5.2 methods to quantify these uncertainties with respect to the weighted least squares estimator and the generalized least squares estimator are discussed. Uncertainty quantification for the studied salt marsh models and the

studied marine phosphorus model are presented in Subsection 2.4 and 5.4, respectively. Quantification of model errors or numerical errors are not subject of this thesis.

1.9 Uncertainty Reduction by Optimal Experimental Designs

The uncertainty in the model parameters and the model outputs strongly depends on the measurement data used for the model parameter estimation which can be taken into account when planning measurements.

The entirety of all controllable conditions specifying the measurements, e.g., the time and location of the measurements, the applied measurement methods and the measured processes, are called measurement conditions. Optimal experimental design techniques [46, 47],[63, 6] allow to determine these measurement conditions in advance in such a way that the resulting uncertainty is minimized. Achieving a specified uncertainty with considerably fewer measurements can considerably reduce the effort and cost of measurements.

Depending on the used design criterion [47, 9], uncertainties in the model parameters or the model output are taken into account. Uncertainties in different model parameters or model outputs can also be weighted differently or be limited to a subset of the model parameters or the model outputs.

It can also be predicted how new measurements would reduce uncertainty, making it possible to determine in advance whether the associated effort and costs are worth it.

Uncertainty reduction by optimizing measurement conditions is one of the main topics in Section 2 and 5. Related methods are described in Subsection 2.2 and 5.2. Related results for the salt marsh models and the marine phosphorus model are presented in 2.4 and 5.4, respectively. A software toolbox for optimizing experimental designs is introduced in 2.3.

1.10 Correlation Matrices Derived from Measurement Data

Measurement data and their statistical properties are indispensable for the understanding the associated processes and their modeling. One of these statistical properties are covariances or correlations. E.g., they are taken into account by estimating model parameters with the generalized least squares estimator [60, 2.1.4], by corresponding uncertainty quantification and by uncertainty reduction using optimal experimental designs.

However, these correlations are usually not known and are thus estimated from the data resulting in a correlation matrix estimate. This could be not positive semidefinite, i.e., an invalid correlation matrix. In this case, the estimate must be replaced by a valid one which should of course be close to the original one. Special approximation algorithms are available for this purpose [59, 22, 49].

Further, it is important that the estimated correlation matrix is well conditioned for numerical reasons. E.g., in order to evaluate the generalized least squares estimator, a linear equation involving the correlation matrix must be solved. So even if the estimate is a valid correlation matrix, it might be useful to replace the estimate by a better conditioned correlation matrix which is still sufficiently close to the original one.

Section 3 focuses on algorithms for both of this problems. As pointed out in Subsection 3.1, existing approximation algorithms have different drawbacks. Hence, a new approximation algorithm, particularly suitable for this purpose, is presented in Subsection 3.2. A software package with an implementation of this algorithm is presented in Subsection 3.3 as well as results of numerical experiments that demonstrate the advantages of this algorithm.

The aim of this algorithm is to calculate a well-conditioned positive semidefinite approximation with minimal difference to the original matrix. It allows to predefine or bound the diagonal values of the approximation. Hence, it can be used for correlation matrices whose diagonal values have to equal one.

During the development of the algorithm, special attention was paid to a low execution time and a low memory consumption making it particularly suitable for large matrices. In addition, the sparsity pattern of the original matrix is preserved which makes it suitable for sparse matrices, too.

Besides statistics, numerical optimization might be another field of application for this algorithm because in some optimization algorithms nonpositive definite matrices have to be approximated by positive definite matrices [15, 40, 13].

Geosci. Model Dev., 8, 791–804, 2015
www.geosci-model-dev.net/8/791/2015/
doi:10.5194/gmd-8-791-2015

Geoscientific Model Development
Open Access

Optimization of model parameters and experimental designs with the Optimal Experimental Design Toolbox (v1.0) exemplified by sedimentation in salt marshes

J. Reimer[1], **M. Schuerch**[2], **and T. Slawig**[1]

[1]Institute of Computer Science, Future Ocean – Kiel Marine Sciences, Christian-Albrechts-University Kiel, 24098 Kiel, Germany
[2]Institute of Geography, Future Ocean – Kiel Marine Sciences, Christian-Albrechts-University Kiel, 24098 Kiel, Germany

Correspondence to: J. Reimer (jor@informatik.uni-kiel.de)

Received: 6 August 2014 – Published in Geosci. Model Dev. Discuss.: 26 September 2014
Revised: 23 January 2015 – Accepted: 9 March 2015 – Published: 25 March 2015

Abstract. The geosciences are a highly suitable field of application for optimizing model parameters and experimental designs especially because many data are collected.

In this paper, the weighted least squares estimator for optimizing model parameters is presented together with its asymptotic properties. A popular approach to optimize experimental designs called local optimal experimental designs is described together with a lesser known approach which takes into account the potential nonlinearity of the model parameters. These two approaches have been combined with two methods to solve their underlying discrete optimization problem.

All presented methods were implemented in an open-source MATLAB toolbox called the *Optimal Experimental Design Toolbox* whose structure and application is described.

In numerical experiments, the model parameters and experimental design were optimized using this toolbox. Two existing models for sediment concentration in seawater and sediment accretion on salt marshes of different complexity served as an application example. The advantages and disadvantages of these approaches were compared based on these models.

Thanks to optimized experimental designs, the parameters of these models could be determined very accurately with significantly fewer measurements compared to unoptimized experimental designs. The chosen optimization approach played a minor role for the accuracy; therefore, the approach with the least computational effort is recommended.

1 Introduction

Mathematical models often contain roughly known model parameters which are optimized based on measurements. The resulting accuracy of the model parameters depends on the conditions, also called experimental setups or experimental designs, under which these measurements are carried out. These experimental designs can be optimized so that the resulting accuracy is maximized. Thus, the effort and cost of measurements can be significantly reduced.

The optimization of experimental designs is therefore particularly interesting for geosciences, where much money is spent on data collection. However, few application examples exist in this field (see Guest and Curtis, 2009, for an overview). This article aims to promote this approach in geosciences and exemplarily apply it to an existing salt marsh accretion model (Schuerch et al., 2013).

In optimizing experimental design, the main problem is to quantify the information content of the measurements to be planned. In general, this can only be done approximatively. There are several approaches available. In Sect. 2, four different approaches to optimize experimental designs together with the weighted least squares estimator for model parameters are presented. Each of these four approaches makes a different trade-off between accuracy and computational effort.

One approach is based on the linearization of the model with respect to the parameters and is the most common used approach called local optimal experimental design. The second more robust approach takes into account the potential

nonlinearity of the model parameters. Both approaches are combined with two different approaches of solving the underlying discrete optimization problem.

To the author's knowledge, there is no open-source software available that can apply all four of these approaches. The only software using this robust approach is a closed-source software called VPLAN which was introduced in Körkel (2002). For the local optimal approach, several implementations are available, but there is no open-source software written in MATLAB. All four approaches, together with approaches to optimize model parameters, were implemented in a MATLAB toolbox called the *Optimal Experimental Design Toolbox*. Its structure and application is described in Sect. 3.

We have chosen two models as application examples, simulating the suspended sediment concentration on salt marshes during tidal inundation and resulting accretion rates on these marshes (Krone, 1987; Temmerman et al., 2003; Schuerch et al., 2013). Both models are zero-dimensional point models and differ in their complexity and number of parameters. These models can be used as a basis to predict the survival capability of salt marshes under the influence of expected global sea level rise.

To use these models for local salt marshes, their parameters have to be adapted to the local environmental conditions. The required measurements are very time-consuming and costly. Using the presented approaches, these measurements could be carried out much more efficiently. These two models are described together with the attendant numerical experiments and the associated results in Sect. 4.

2 Optimization of model parameters and experimental designs

The first step to the optimization of model parameters is the choice of the estimator. This maps the measurement results onto estimated model parameters. These estimated parameters are often defined so that they minimize a so-called misfit function. The misfit function quantifies the distance between the measurement results and the model output.

The estimator should be derived from the statistical properties of the measurement errors, for example, a maximum likelihood estimator. Often the measurement errors are assumed to be normally distributed; this leads to the least squares estimators. They are the most widely used class of estimators since their introduction by Gauss and Legendre (Stigler, 1981).

Their simplest form is the ordinary least squares estimator. Its misfit function is the sum of the squares of the differences between each measurement result and the corresponding model output. A generalization is the weighted least squares estimator which has advantages in the event of heteroscedastic measurement errors. This estimator and its asymptotic properties are presented in the following subsection. The generalized least squares estimator is a further generalization which takes into account the stochastic dependence of the measurement errors.

2.1 The weighted least squares estimator

In the following, the weighted least squares estimator is presented. For this purpose, some notations and assumptions are introduced.

The model function is denoted by

$$f : \Omega_x \times \Omega_p \to \mathbb{R}.$$

Here, $\Omega_x \subseteq \mathbb{R}^{n_x}$ is the set of feasible experimental designs, and $\Omega_p \subset \mathbb{R}^{n_p}$ is the set of feasible model parameters from which the unknown exact parameter vector $\hat{\boldsymbol{p}} \in \Omega_p$ is to be determined. Often these sets are defined by lower and upper bounds.

The measurement result for every design $\boldsymbol{x} \in \Omega_x$ is considered as a realization of a random variable $\eta_{\boldsymbol{x}}$. Each random variable $\eta_{\boldsymbol{x}}$ is assumed to be normally distributed with the expectation $f(\boldsymbol{x}, \hat{\boldsymbol{p}})$ and standard deviation $\sigma_{\boldsymbol{x}} > 0$.

A1a. $\eta_{\boldsymbol{x}} \sim \mathcal{N}(f(\boldsymbol{x}, \hat{\boldsymbol{p}}), \sigma_{\boldsymbol{x}}^2)$ for every $\boldsymbol{x} \in \Omega_x$.

Furthermore, these random variables are assumed to be pairwise stochastically independent.

A1b. $\eta_{\boldsymbol{x}}$ and $\eta_{\boldsymbol{x}'}$ are stochastically independent for every $\boldsymbol{x}, \boldsymbol{x}' \in \Omega_x$.

If we consider $n \geq n_p$ measurement results $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)^T \in \mathbb{R}^n$ with corresponding experimental designs $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \Omega_x$, the weighted least squares estimation $\boldsymbol{p}_n$ and the corresponding estimator P_n are defined as

$$\boldsymbol{p}_n := P_n(\boldsymbol{y}) := \underset{\boldsymbol{p} \in \Omega_p}{\arg\min}\ \psi_n(\boldsymbol{y}, \boldsymbol{p}), \tag{1}$$

where the misfit function ψ_n is defined as

$$\psi_n : \mathbb{R}^n \times \Omega_p \to \mathbb{R}, (\boldsymbol{y}, \boldsymbol{p}) \longmapsto \sum_{i=1}^{n} \left(\frac{\boldsymbol{y}_i - f(\boldsymbol{x}_i, \boldsymbol{p})}{\sigma_{\boldsymbol{x}_i}} \right)^2.$$

With the following assumptions, the existence of a minimum is ensured.

A2. $f(\boldsymbol{x}, \cdot)$ is continous for every $\boldsymbol{x} \in \Omega_x$.

A3. Ω_p is compact (closed and bounded).

If $\psi_n(\boldsymbol{y}, \cdot)$ is convex, the minimum is also unique.

The parameter estimation $\boldsymbol{p}_n$ in Eq. (1) can be calculated with an optimization method for continuous optimization problems. A possible method is the sequential quadratic programming (SQP) algorithm which is, for example, described in Nocedal and Wright (1999, chapter 18).

2.2 Asymptotic properties

Provided certain regularity conditions are met, the least squares estimators are consistent, asymptotically normally distributed and asymptotically efficient.

These asymptotic properties were first proved by Jennrich (1969) for the ordinary least squares estimator and also discussed in Malinvaud (1970) and Wu (1981). In White (1980), these properties were proved for the weighted least squares estimator and for the generalized least squares estimator in White and Domowitz (1984). A good summary for all three can be found in Amemiya (1983).

Consistency means that the estimated parameters converge in probability to the unknown exact parameters as the number of measurements goes to infinity; that is,

$$P_n \xrightarrow{p} \hat{\boldsymbol{p}} \text{ as } n \to \infty$$

for the weighted least squares estimator P_n with the unknown exact model parameters $\hat{\boldsymbol{p}}$.

For consistency, the following assumptions are sufficient in addition to the previous assumptions A1 to A3 (Seber and Wild, 2003, p. 565).

A4a. $n^{-1}B_n$ converges uniformly with $B_n : \Omega_p \times \Omega_p \to \mathbb{R}, (\boldsymbol{p}, \boldsymbol{p}') \longmapsto \sum_{i=1}^{n} f(\boldsymbol{x}_i, \boldsymbol{p}) f(\boldsymbol{x}_i, \boldsymbol{p}') \sigma_{\boldsymbol{x}_i}^{-2}$.

A4b. $\bar{D}(\boldsymbol{p}, \hat{\boldsymbol{p}}) = 0 \Rightarrow \boldsymbol{p} = \hat{\boldsymbol{p}}$ for all $\boldsymbol{p} \in \Omega_p$ with $\bar{D} := \lim_{n\to\infty} n^{-1} D_n$ and $D_n : \Omega_p \times \Omega_p \to \mathbb{R}, (\boldsymbol{p}, \boldsymbol{p}') \longmapsto \sum_{i=1}^{n} (f(\boldsymbol{x}_i, \boldsymbol{p}) - f(\boldsymbol{x}_i, \boldsymbol{p}'))^2 \sigma_{\boldsymbol{x}_i}^{-2}$ ($\bar{D}$ is well defined by assumption A4a).

An estimator is asymptotically efficient if its variance converges to the Cramér–Rao bound as the number of measurements goes to infinity. The Cramér–Rao bound (Cramér, 1946; Rao, 1945) is a lower bound for the variance of any unbiased estimator.

For asymptotic efficiency, the following assumptions are sufficient in addition to the previous assumptions A1 to A4 (Seber and Wild, 2003, p. 571).

A5. $\hat{\boldsymbol{p}}$ is an interior point of Ω_p. Let $\hat{\Omega}_p \subseteq \Omega_p$ be an open neighborhood of $\hat{\boldsymbol{p}}$.

A6. $f(\boldsymbol{x}_i, \cdot)$ is twice continuously differentiable in $\hat{\Omega}_p$.

A7. $n^{-1}M_n$ converges uniformly with $M_n : \hat{\Omega}_p \to \mathbb{R}^{n_p \times n_p}, \boldsymbol{p} \longmapsto \sum_{i=1}^{n} \nabla_p f(\boldsymbol{x}_i, \boldsymbol{p}) \nabla_p f(\boldsymbol{x}_i, \boldsymbol{p})^T \sigma_{\boldsymbol{x}_i}^{-2}$.

A8. $n^{-1}H_n$ converges uniformly with $H_n : \hat{\Omega}_p \to \mathbb{R}^{n_p \times n_p}, \boldsymbol{p} \longmapsto \left(\sum_{i=1}^{n} \left(\frac{\partial^2}{\partial p_i \partial p_j} f(\boldsymbol{x}_i, \boldsymbol{p})\right)^2 \sigma_{\boldsymbol{x}_i}^{-2} \right)_{i,j=1,\ldots,n_p}$.

A9. $\hat{M}(\hat{\boldsymbol{p}})$ is invertible with $\hat{M} := \lim_{n\to\infty} n^{-1} M_n$.

In this case, the Cramér–Rao bound of the weighted least squares estimator P_n is the inverse of the Fisher information matrix $M_n(\hat{\boldsymbol{p}})$.

Under these assumptions, the asymptotic behavior of the weighted least squares estimator can be summarized by its convergence in distribution as follows:

$$\sqrt{n}(P_n - \hat{\boldsymbol{p}}) \xrightarrow{d} \mathcal{N}(0, M_n(\hat{\boldsymbol{p}})^{-1} n) \text{ as } n \to \infty \tag{2}$$

(see, e.g., Seber and Wild, 2003, chapter 12 and Walter and Pronzato, 1997, chapter 3).

2.3 Optimal experimental designs

The accuracy of the weighted least square estimator P_n can be described by its covariance matrix. Due to the asymptotic distribution (Eq. 2), this can be approximated by the inverse of the information matrix $M_n(\boldsymbol{p}_n)$, provided the matrix $M_n(\boldsymbol{p}_n)$ is nonsingular, that is,

$$\operatorname{cov}(P_n) \approx M_n(\boldsymbol{p}_n)^{-1}. \tag{3}$$

Therefore, the unknown model parameters can be determined more accurately the smaller the (approximated) covariance matrix of the estimator is.

Criteria $\phi : \mathbb{R}^{n_p \times n_p} \to \mathbb{R}^+ \cup \{\infty\}$, such as the trace or determinant, are used in order to compare these matrices (see, e.g., El-Monsef et al., 2009, for an overview of various criteria). If the approximation (Eq. 3) is used and $M_n(\boldsymbol{p}_n)$ is singular, the value of ϕ is set to infinity.

In the context of optimizing experimental designs, we assume $n \geq 0$ measurements have been carried out and designs for additional measurements should be selected from m designs $\boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_m \in \Omega_x$. The choice for each design $\boldsymbol{x}'_i$ is expressed by a weight $w_i \in \{0, 1\}$, where 1 indicates the selection and 0 the contrary.

Hence, the resulting information matrix, depending on the choice $\boldsymbol{w} \in \{0, 1\}^m$ and the parameter vector $\boldsymbol{p}_n \in \Omega_p$, is defined as

$$M_n(\boldsymbol{w}, \boldsymbol{p}_n) := M_n(\boldsymbol{p}_n) + \sum_{i=1}^{m} \boldsymbol{w}_i \frac{\nabla_p f(\boldsymbol{x}'_i, \boldsymbol{p}_n) \nabla_p f(\boldsymbol{x}'_i, \boldsymbol{p}_n)^T}{\sigma_{\boldsymbol{x}'_i}^2}.$$

If the covariance matrix is approximated by the inverse of the information matrix, optimal (additional) designs, with respect to a criterion ϕ, are expressed by a solution of

$$\underset{\boldsymbol{w} \in \{0,1\}^m}{\arg\min}\ \phi(M_n(\boldsymbol{w}, \boldsymbol{p}_n)^{-1}). \tag{4}$$

These optimal designs are called local optimal designs because these designs are only optimal regarding the previous model parameter estimation $\boldsymbol{p}_n$ and not the unknown exact model parameters $\hat{\boldsymbol{p}}$.

Potential constraints on the choice of the designs can be realized by constraints on the weight $\boldsymbol{w}$. For example, the

number or the cost of the measurements can be limited by linear constraints on $\boldsymbol{w}$. These constraints have to be considered in the above optimization problem (Eq. 4).

The formulation (Eq. 4) is useful if additional experimental designs should be chosen from a finite number of experimental designs. Otherwise, the optimization problem can be reformulated so that the additional optimal design variables have to be optimized directly.

2.4 Calculation of optimal experimental designs

A straight-forward way to solve the optimization problem (Eq. 4) is to test all possible values of $\boldsymbol{w}$. This direct approach is only practical for small m.

For bigger m, the optimization problem (Eq. 4) is solved approximately. For this purpose, it is solved in the continuous rather than the discrete setting; that is, the constraint $\boldsymbol{w} \in \{0,1\}^m$ is relaxed to $\boldsymbol{w} \in [0,1]^m$. Accordingly, the problem

$$\underset{\boldsymbol{w}\in[0,1]^m}{\arg\min}\ \phi(M_n(\boldsymbol{w}, \boldsymbol{p}_n)^{-1}) \tag{5}$$

is solved.

A possible algorithm to solve this continuous optimization problem is the SQP algorithm which is, for example, described in Nocedal and Wright (1999, chapter 18).

After the continuous problem (Eq. 5) is solved, its solution is projected onto integers with heuristics. An easy way is to round the continuous solution. Another is to sum up all continuous weights and then to choose as many designs with the highest continuous weights. Potential constraints on w still have to be considered by solving the continuous problem and the following projection onto an integer solution. The second heuristic, for example, preserves constraints on the number of designs to choose.

Our numerical experiments with the application examples in Sect. 4 have shown that the solutions of the continuous problem (Eq. 5) are already close to integer values. This behavior was also observed, for example, in Körkel (2002) and Körkel et al. (2004).

2.5 Robust optimal experimental designs

The information matrix M_n depends on the estimated parameters $\boldsymbol{p}_n$ if the parameters are nonlinear. This may lead to suboptimal designs if $\nabla_p f(\,\cdot\,, \boldsymbol{p}_n)$ differs strongly from $\nabla_p f(\,\cdot\,, \hat{\boldsymbol{p}})$.

For this reason, we now consider a method which takes into account a possible nonlinearity of the parameters. This robust method was presented in Körkel (2002) and Körkel et al. (2004).

The main idea of the method is not to optimize the quality of the covariance matrix for a single parameter vector $\boldsymbol{p}_n$ as in Eq. (4), but to optimize the worst case quality within a whole domain which contains the unknown exact parameter vector $\hat{\boldsymbol{p}}$ with high probability.

For this purpose, a confidence region which contains $\hat{\boldsymbol{p}}$ with probability $\alpha \in (0,1)$ is approximated by

$$G_n(\alpha) := \left\{ \boldsymbol{p} \in \mathbb{R}^{n_p} \mid \|\boldsymbol{p} - \boldsymbol{p}_n\|^2_{M_n(\boldsymbol{p}_n)^{-1}} \le \gamma(\alpha) \right\}. \tag{6}$$

Here, $\gamma(\alpha)$ is the α quantile of the χ^2 distribution and $\|\boldsymbol{v}\|_{\boldsymbol{A}} := \sqrt{\boldsymbol{v}^T \boldsymbol{A} \boldsymbol{v}}$ denotes the energy norm of the vector $\boldsymbol{v} \in \mathbb{R}^{n_p}$ with respect to the positive definite matrix $\boldsymbol{A} \in \mathbb{R}^{n_p \times n_p}$. The approximation of the confidence region arises from linearization of the model function f in point $\boldsymbol{p}_n$ and the assumption $P_n \sim \mathcal{N}(\hat{\boldsymbol{p}}, M_n(\boldsymbol{p}_n)^{-1})$.

If the worst case quality in the entire region $G_n(\alpha)$ shall be optimized, the optimization problem (Eq. 4) becomes

$$\underset{\boldsymbol{w}\in\{0,1\}^m}{\arg\min}\ \max_{\boldsymbol{p}\in G_n(\alpha)} \phi(M_n(\boldsymbol{w}, \boldsymbol{p})^{-1}). \tag{7}$$

This minimum–maximum optimization problem can be solved only with considerably more computational effort compared to the optimization problem (Eq. 4). In order to reduce this effort, the function $\phi(M_n(\boldsymbol{w}, \,\cdot\,)^{-1})$ is linearized in point $\boldsymbol{p}_n$ in the following way:

$$\begin{aligned}&\phi(M_n(\boldsymbol{w}, \boldsymbol{p})^{-1}) \approx \\ &\quad \phi(M_n(\boldsymbol{w}, \boldsymbol{p}_n)^{-1}) + \nabla_p(\phi(M_n(\boldsymbol{w}, \boldsymbol{p})^{-1}))^T (\boldsymbol{p} - \boldsymbol{p}_n).\end{aligned}$$

The resulting inner maximization problem can be solved analytically. It is

$$\begin{aligned}&\max_{\boldsymbol{p}\in G_n(\alpha)} \phi(M_n(\boldsymbol{w}, \boldsymbol{p}_n)^{-1}) + \nabla_p(\phi(M_n(\boldsymbol{w}, \boldsymbol{p})^{-1}))^T (\boldsymbol{p} - \boldsymbol{p}_n) \\ &= \phi(M_n(\boldsymbol{w}, \boldsymbol{p}_n)^{-1}) + \gamma(\alpha)^{\frac{1}{2}} \|\nabla_p(\phi(M_n(\boldsymbol{w}, \boldsymbol{p}_n)^{-1}))\|_{M_n(\boldsymbol{p}_n)},\end{aligned}$$

as can be seen, for example, in Körkel (2002). With this approach the optimization problem (Eq. 7) is replaced by

$$\begin{aligned}&\underset{\boldsymbol{w}\in\{0,1\}^m}{\arg\min}\ \phi(M_n(\boldsymbol{w}, \boldsymbol{p}_n)^{-1}) \\ &\quad + \gamma(\alpha)^{\frac{1}{2}} \|\nabla_p(\phi(M_n(\boldsymbol{w}, \boldsymbol{p}_n)^{-1}))\|_{M_n(\boldsymbol{p}_n)}.\end{aligned} \tag{8}$$

This optimization problem again can be solved approximatively by solving the corresponding continuous problem and projecting this solution onto an integer solution as described in the previous subsection.

It should be noted that in this approach (Eq. 8), the first and second derivatives of the model are used. In contrast, only the first derivative is used for local optimal designs (Eq. 4).

2.6 Efficiency of experimental designs

A common way to describe the benefit of an experimental design is its efficiency. The efficiency of an experimental design $\boldsymbol{w} \in \{0,1\}^m$ regarding a criterion ϕ and with n previous measurements is defined as follows:

$$E_\phi(\boldsymbol{w}) := \min_{\hat{\boldsymbol{w}}\in\{0,1\}^m} \frac{\phi(M_n(\hat{\boldsymbol{w}}, \hat{\boldsymbol{p}})^{-1})}{\phi(M_n(\boldsymbol{w}, \hat{\boldsymbol{p}})^{-1})}. \tag{9}$$

```
model_object = model_explicit('p*t^2', 'p', 't')
% 1. input: the model function as symbolic formula
% 2. input: the parameter variable(s)
% 3. input: the experimental design variable(s)
% return: a model object which implements the model interface
```

Listing 1. Create a model with a symbolic model function.

It should be noted that the searched parameter vector $\hat{\boldsymbol{p}}$ is used here. If this is not known then the efficiency can not be calculated.

The efficiency is always between 0 and 1 and is larger the better the experimental design is.

3 The Optimal Experimental Design Toolbox

We implemented the methods presented in the previous section for optimization of model parameters and experimental designs as a MATLAB toolbox named the *Optimal Experimental Design Toolbox*.

MATLAB (MathWorks, 2014) was chosen because it supports vector and matrix operations and provides many numerical algorithms, especially for optimization. Moreover, MATLAB supports object oriented programming and therefore permits a simple structuring, modification and extension of the implementation. Another advantage of MATLAB is that it can easily interact with C and Fortran.

The toolbox is available at a Git repository (Reimer, 2015) under the GNU General Public License (Foundation, 2007). It includes extensive commented source code, a detailed help integrated in MATLAB and a user manual.

3.1 Provision of the model function

For the methods described in Sect. 2, the model function and its first and second derivative with respect to the model parameters are required.

Actually, the model function is required for the parameter optimization and, depending on the optimization method, also its first derivative. Its first derivative is also required for the experimental design optimization. If the robust method is used its second derivative is also required.

The *model* interface prescribes how to provide these functions. They need not be written in MATLAB itself, since MATLAB can call functions in C, C++ or Fortran.

The toolbox has several possibilities to provide the derivatives automatically. The *model_fd* class, for example, provides the derivatives by approximation with finite differences. If the model function is given as an explicit symbolic function, the *model_explicit* class can provide the derivatives by symbolic differentiation with the *Symbolic Math Toolbox*. Listing 1 shows, for example, how a *model_explicit* object is created.

In the event that the model function is given as a solution of an initial value problem, the *Optimal Experimental Design Toolbox* contains the *model_ivp* class. This class solves the parameter dependent initial value problem and calculates the necessary derivatives. Listing 2 shows how a *model_ivp* object is created.

```
model_object = model_ivp('-y+(t+1)*b', '[a,b]', 'y', 'a', 't', [1,10])
% 1. input: the right hand side of the differential equation
% 2. input: the model parameter variable(s)
% 3. input: the model function variable
% 4. input: the initial value of the model function
% 5. input: the dependent variable in the model function
% 6. input: the interval of integration
% return: a model object which implements the model interface
```

Listing 2. Create a model with a model function given as solution of an initial value problem.

```
solver_object.set_model(model_object)
% input: an object that implements the model interface
```

Listing 3. Set the model.

The class takes advantage of the fact that the integration and differentiation of the differential equation can be interchanged if the model function is sufficiently often continuously differentiable. Required derivatives of the differential equation and initial value are calculated again by symbolic differentiation with the *Symbolic Math Toolbox*. The resulting initial value problems are solved with MATLAB's *ode23s* function which can also solve stiff problems. Since the arising initial value problems for the derivatives are mutually independent, the solutions of the initial value problems can be calculated in parallel using the *Parallel Computing Toolbox*.

3.2 Setup of the solver

Methods for the optimization of model parameters and experimental designs are provided by the *solver* class. First, a *solver* object has to be created and the necessary information has to be passed.

On the one hand, this is the *model* which has to be set by the *set_model* method (see Listing 3).

On the other hand, an initial guess of the model parameters have to be set by the *set_initial_parameter_estimation* method (see Listing 4).

Potential accomplished measurements can be set via the *set_accomplished_measurements* method. These measurements consist of the corresponding experimental designs together with their variances of the measurement errors. Furthermore, the measurement results themselves have to be passed for a parameter estimation (see Listing 5).

Finally, if an optimization of experimental designs shall be performed, the selectable measurements have to be set by the *set_selectable_measurements* method (see Listing 6). These measurements consist of the experimental designs as well as the variances of the measurement errors.

```
solver_object.set_initial_parameter_estimation([1, 2])
% input: the initial estimation of the model parameters
```

Listing 4. Set the initial parameter estimation.

```
solver_object.set_accomplished_measurements((1:5)', 0.01*ones(5,1), -↩
    exp((1:5)'))
% 1. input: the experimental designs of accomplished measurements
% 2. input: the variances of the associated measurement errors
% 3. input: the associated measurement results
```

Listing 5. Set accomplished measurements.

3.3 Optimization of experimental designs and model parameters

Once the *solver* object is configured as described in the previous subsection, experimental designs or model parameters can be optimized via the *get_optimal_measurements* (see Listing 7) or the *get_optimal_parameters* (see Listing 8) method, respectively. Constraints on the experimental designs or model parameters can be passed to the corresponding method.

The *get_optimal_measurements* method can solve the optimization problem directly by trying all possible combinations or approximatively.

For the approximative solving, the continuous problem is solved with the SQP algorithm (see Nocedal and Wright, 1999, Chapter 18) provided by the *fmincon* function of the *Optimization Toolbox*. Its solution is projected onto an integer solution by the second heuristic described in Sect. 2.4.

The first derivative of the objective function is provided in analytical form. This saves much of the computing time compared to derivatives calculated by finite differences. The Hessian matrix is approximated by the Broyden–Fletcher–Goldfarb–Shanno (BFGS) update (Broyden, 1970; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970).

MATLAB's SQP algorithm can recover from infinity. If an infinite function value is reached during the optimization, the algorithm attempts to take a smaller step. Thus, if the optimization is started with a regular design, singular designs do not make any trouble.

The *get_optimal_parameters* method uses the trust-region-reflective (Coleman and Li, 1994, 1996) or the Levenberg–Marquard algorithm (Levenberg, 1944; Marquardt, 1963; Moré, 1977) provided by the *lsqnonlin* function of the *Optimization Toolbox* to solve the least squares problem resulting from the parameter estimation. The first derivative of the objective function is also provided analytically.

Furthermore, the expected quality of the resulting parameter estimation for any selection of experimental designs can be calculated using the *get_quality* method of the *solver* object. Thus, for example, the increase in quality by adding or removing experimental designs can be determined.

```
solver_object.set_selectable_measurements((6:10)', 0.01*ones(5, 1))
% 1. input: the selectable experimental designs
% 2. input: the variances of the associated measurement errors
```

Listing 6. Set selectable measurements.

```
optimal_measurements = solver_object.get_optimal_measurements(3)
% input: the maximum number of measurements allowed
% return: the optimal subset of the selectable measurements with a ↩
    number of measurements less or equal to the restriction
```

Listing 7. Optimize experimental designs.

3.4 Execution time and memory consumption

The total time required for the optimization of the model parameters or an experimental design depends crucially on the time required for evaluating the model function and its first and second derivative with respect to the model parameters.

When optimizing model parameters, the model function and its first derivative has to be evaluated several times with different model parameter vectors at the accomplished measuring points. When optimizing experimental designs, the model function and its first and second derivative has to be evaluated for one model parameter vector but at the accomplished and selectable measuring points.

Generally, the execution time increases with the number of parameters, the number of selectable measurements and the number of accomplished measurements.

The implementation of this toolbox favors a low execution time of a low memory consumption. For this reason, (intermediate) results within a method call and between successive method calls are saved and reused. An example is multiple occurring matrix multiplications within a method call. Another example is a re-optimization of designs with other constraints, such as another maximum number of allowed measurements. Here, the derivatives of the model function calculated in the previous optimization are reused.

Due to the described caching strategy, the total memory consumption depends linearly on the number of (accomplished and selectable) measurements and quadratically on the number of parameters. Nevertheless, it should be possible to solve problems with hundreds of parameters and thousands of measurements on a standard computer.

3.5 Changeable options

Many settings for the optimization of experimental designs or model parameters are changeable. These can be altered by the *set_option* method of the *solver* object (see Listing 9). The desired options can be set using property-value pairs, as already known from MATLAB.

Estimation method: The estimation method for the quality of experimental designs can be selected by the *estimation_method* option. The standard *point* estimation method and the robust *region* estimation method, both

```
optimal_parameters = solver_object.get_optimal_parameters([0,0],[9,9])
% 1. input: the lower bound of the model parameters
% 2. input: the upper bound of the model parameters
% return: a parameter estimation resulting from the accomplished ↩
    measurements which takes into account the passed constraints
```

Listing 8. Optimize model parameters.

```
solver_object.set_option('option_name', option_value)
% 1. input: the name of the option which should be changed
% 2. input: the new value of the option
```

Listing 9. Change an option.

presented in Sect. 2, are supported. The *region* estimation method is the default setting.

Confidence level: The level of confidence for the confidence region at the *region* estimation method, represented by α in Sect. 2.5, can be set by the *alpha* option. The default value is 0.95.

Prior parameter estimation: It can be chosen whether a parameter optimization should be performed before optimizing experimental designs. This can be set by the *parameter_estimation* option and the values *yes* or *no*. To save computational time no previous parameter optimization is performed by default.

Quality criterion: The quality criterion, which is applied to the covariance matrix and represented in Sect. 2.1 as ϕ, can also be chosen with the *criterion* option. The *criterion* interface prescribes the syntax of the criterion function and its necessary derivatives. The trace of the covariance is the default criterion and implemented by the *criterion_A* class.

Parameter scaling: It can be chosen whether model parameter should be scaled before optimizing experimental designs or the model parameters themselves. Scaling means a uniform impact of all model parameters and is enabled by default. The options are *edo_scale_parameters* and *po_scale_parameters* with the values *yes* and *no*.

Optimization algorithm for experimental design: The exact and the approximative approach for the optimization of an experimental design problem can be chosen with the *edo_algorithm* option and the values *direct* and *local_sqp*. For time reasons, by default the experimental design problem is solved by the approximative approach. Furthermore, the number of function evaluations and iterations by the SQP algorithm can be constrained by the options *edo_max_fun_evals* and *edo_max_iter*.

Optimization algorithm for parameter estimation: The optimization algorithm for the parameter estimation problem can be chosen with the *po_algorithm* option. The trust-region-reflective (Coleman and Li, 1994, 1996) and the Levenberg–Marquard algorithm (Levenberg, 1944; Marquardt, 1963; Moré, 1977) can be chosen with the values *trust-region-reflective* and Levenberg–Marquardt The trust-region-reflective algorithm is the default algorithm. Furthermore, the number of function evaluations and iterations can be limited through the options *po_max_fun_evals* and *po_max_iter*.

3.6 Help and documentation

The *Optimal Experimental Design Toolbox* also provides extensive integrated help. Besides system requirements and version information, a user's guide with step-by-step instructions on how to optimize experimental designs and model parameters is included. Demos show how to work with the toolbox in practice. In addition, a detailed description for every class and method is available.

The layout of the help for the *Optimal Experimental Design Toolbox* is based on the design of the help also used by MATLAB and other toolboxes. Thus, the user does not have to get reoriented with a new layout.

4 Application examples

In this section, numerical experiments together with their results regarding the optimization of model parameters and experimental designs are presented for two models of different complexity. Both models describe the sediment concentration in seawater during tidal inundation of coastal salt marshes.

Coastal salt marshes have an important ecological function with their diverse flora and as a nursery for migratory birds. Furthermore, they have the role of dissipating current and wave energy and therefore reducing erosional forces at dikes and coastal areas.

With these models, the vertical accretion of coastal salt marshes can be predicted. When considering expected global sea level rise (IPCC, 2013), the future ability of coastal salt marshes to adapt to rising sea levels and thus to survive can be estimated. Depending on these estimates, measures to protect these salt marshes can be taken.

Calibration of the model parameters requires measurements of suspended sediment concentration during tidal inundation, which are time-consuming and laborious. For this reason, it is advantageous to know under which conditions and how many of these measurements should be carried out.

4.1 The models

Both models are zero-dimensional point models, which describe the sediment concentration in seawater during tidal inundation of coastal salt marshes. The first model (C_2-model) has two model parameters, was described in Temmerman et al. (2003) and was adapted for a salt marsh in the Wadden

Sea (southeastern North Sea), located near Hoernum in the southern part of the island of Sylt (Germany), by Schuerch et al. (2013). The second model (C_3-model) has three model parameters, is an extension of the first model and subject of current research.

4.1.1 The C_2-model

The first model is called the C_2-model. Here, the sediment concentration in $\mathrm{kg\,m^{-3}}$ is modeled by the function $C : [t_S, t_E) \to \mathbb{R}^+$. Furthermore, t_S is the start time of the inundation of the salt marsh and t_E the end time. The concentration C is given implicitly as the solution of the initial value problem

$$C'(t) = \begin{cases} \frac{-w_S C(t) + (C_0 - C(t))h'(t)}{h(t)-E} & \text{if } h'(t) > 0 \\ \frac{-w_S C(t)}{h(t)-E} & \text{else} \end{cases}$$
$$\text{for all } t \in (t_S, t_E) \text{ and } C(t_S) = C_0. \quad (10)$$

Here, $C_0 \geq 0$ is the initial sediment concentration of the flooding seawater and $w_S \geq 0$ the settling velocity of the suspended sediment in $\mathrm{m\,s^{-1}}$. Moreover, the function

$$h : \mathbb{R} \to \mathbb{R}, t \longmapsto \frac{a}{1 + \left(\frac{t-x_0}{b}\right)^2} + h_{HW} - h_{MHW}$$

describes the time-dependent water surface elevation and E the elevation of the marsh both in meters and relative to a fixed datum. Here, a, b and x_0 are constants describing the change in the water level, h_{MHW} the mean high water level and h_{HW} the high water level of a certain tidal inundation in meters. The start t_S and end time t_E of the inundation are the points where the height h equals the elevation of the marsh E.

The sediment concentration C thus decreases continuously within a tidal cycle depending on the settling velocity w_S which is described by the term

$$-\frac{w_S C(t)}{h(t) - E}$$

in Eq. (10). During the flood phase, the reduced sediment concentration is partially compensated by new inflowing sea water. This is described by the term

$$\frac{(C_0 - C(t))h'(t)}{h(t) - E}$$

in the first case of Eq. (10).

The values used in the water surface elevation function h, for the local salt marsh, are shown in Table 1. These have been estimated by nonlinear regression analysis using local historic tide gauge data from 1999 to 2009 (at Hoernum Hafen, Germany). The continuous high-resolution (6 min) time series has, therefore, been split into the individual tidal cycles beforehand (Schuerch et al., 2013).

Table 1. Values used for the water surface elevation function h

	a	b	x_0	h_{MHW}	E
local value	3.7506	19447.1	−1301.0	3.75 m	1.3m

Table 2. Values for the C_2-model.

	C_0 [$\mathrm{kg\,m^{-3}}$]	w_S [$\mathrm{m\,s^{-1}}$]
reference value	0.1	10^{-5}
typical range	0.01–0.2	4×10^{-6}–4×10^{-4}
start value	5	2×10^{-7}
optimization bound	10^{-4}–10^4	10^{-8}–1

The high water level h_{HW} of the current tidal inundation is measured or taken from predictions.

The initial sediment concentration C_0 and the settling velocity w_S are only roughly known and therefore model parameters. Reference values derived from literature values and typical ranges can be found in Table 2 (see Bartholdy and Aagaard, 2001, for C_0 and Temmerman et al., 2003, for C_0 and w_S).

4.1.2 The C_3-model

The second model is an extension of the C_2-model and is called the C_3-model. Here the model parameters C_0 and w_S are substituted by

$$C_0 = k(h_{HW} - E),$$
$$w_S = r(C_0)^s = rk^s(h_{HW} - E)^s.$$

Where $k \geq 0$, $r \geq 0$ and $s \geq 0$ are unknown model parameters. Reference values derived from literature values and typical ranges (where available) can be found in Table 3 (see van Leussen, 1999, and Pejrup and Mikkelsen, 2010, for the settling index s and Temmerman et al., 2004, for k).

In this model, a linear relationship between the initial sediment concentration and the high water level is assumed, where during heavy flooding a higher sediment concentration is assumed (Temmerman et al., 2003; Schuerch et al., 2013). Additionally, a relationship between the initial sediment concentration and the settling velocity is assumed (Krone, 1987). This is an empirical approximation of the so-called flocculation process (Burt, 1986).

4.2 Numerical experiments

We performed several numerical experiments to compare the benefit of optimized with unoptimized measurement conditions. Also, the benefit of different approaches to optimization measurement conditions was compared. Using these results, an appropriate approach for the optimization of conditions for real measurements was selected.

Table 3. Values for the C_3-model.

	k	r	s
reference value	0.25	10^{-5}	0.5
typical range	0.04–0.2		0.5–3.5
start value	12.5	2×10^{-7}	3
optimization bound	10^{-4}–10^{4}	10^{-8}–1	10^{-1}–5

The approaches introduced in Sect. 2 and implemented by the *Optimal Experimental Design Toolbox* described in Sect. 3 were used for the numerical experiments. For that, we used the *model_ivp* class which allows for the calculation of the solution of an initial value problem and its first and second derivatives with respect to the model parameters. The C_2-model was implemented by the *model_C2* class and the C_3-model by the *model_C3* class which is a subclass of the *model_C2* class.

For our numerical experiments, we used the model output with the reference parameters in Tables 2 and 3 plus an additive normally distributed measurement error with zero expectation as artificial measurement results. As standard deviation of the measurement error, we chose 10^{-2} once and 10^{-1} once.

In our numerical experiments, we alternately selected a fixed number of experimental designs and estimated the model parameters with corresponding measurement results. We carried out each experiment 10 times and averaged the results to minimize the influence of randomness.

For the parameter estimation, the start values and bounds in Tables 2 and 3 were used. The bounds were chosen so that the typical range of values is covered, but also more extreme values are possible. The starting values were chosen slightly outside the typical ranges to represent a poor initial guess.

The experimental designs for these models consist of the time point of the measurement and the high water level of the tidal inundation. A set of thirty selectable experimental designs was specified. They were obtained by combining three different high water levels of the tidal inundation (1.5, 2.0 and 2.5 m) with 10 time points equidistantly spread over the inundation period.

For choosing the experimental designs, we compared the standard and the robust approach presented in Sect. 3 with the trace as quality criterion together with uniformly distributed experimental designs. In the robust approach, a confidence level of 95 % was used. The optimization problems for the experimental designs were once solved exactly and once approximatively (see Sect. 2.4). To evaluate all these methods, we compared the resulting parameter estimations with the reference model parameters.

We further investigated whether the number of measurements after which new experimental designs are optimized had an impact on the accuracy of the parameter estimation. For this purpose, different numerical experiments were performed where the parameters and experimental designs have been optimized after each one, three and five measurements. Altogether 50 measurements were simulated at each experiment with the C_2-model. For the C_3-model, 150 measurements were simulated at each experiment since the model is more complex and therefore a sufficiently accurate estimation of its parameters might be more difficult.

4.3 Accuracy of the parameter estimations

In this subsection, we compare the accuracy of the parameter estimations resulting from the previously described numerical experiments. Some results are illustrated in Figs. 1 and 2.

4.3.1 Results for the C_2-model

The accuracy of the parameter estimations for the C_2-model only improved marginally after four to twelve measurements independent of the choice of the experimental designs. The accuracy improved faster the more frequently the experimental designs and parameters were optimized. However, the best achieved accuracy was independent of the frequency.

With uniformly distributed experimental designs the best achieved accuracy was slightly worse than with optimized experimental designs. Four to six more measurements were needed compared to optimized experimental designs in order to achieve their accuracy.

Although the parameters occur nonlinearly in this model, it made close to no difference whether the standard or the robust approach for the optimization of the experimental designs was used.

The approximate solving of the discrete optimization problem has resulted in slightly worse accuracy at the first iterations compared to the exact solving. Thereafter, the difference was very small. The solutions of the relaxed continuous optimization problems were almost always nearly integer.

The different standard deviations of the measurement errors only influenced the best achieved accuracy which was of course worse at a higher standard deviation. This can be explained by the fact that different constant standard deviations only mean a different scaling of the objective of the experimental design optimization problem. Thus, different constant standard deviations do not affect its solution.

4.3.2 Results for the C_3-model

After 10–25 measurements, the accuracy of the parameter estimations for the C_3-model with optimized experimental designs only improved slightly. Again, the fewer measurements performed per iteration the faster the accuracy improved, and the best achieved accuracy was independent of the number of measurements per iteration.

With uniformly distributed experimental designs, the best accuracy was achieved after 24–60 measurements. Further-

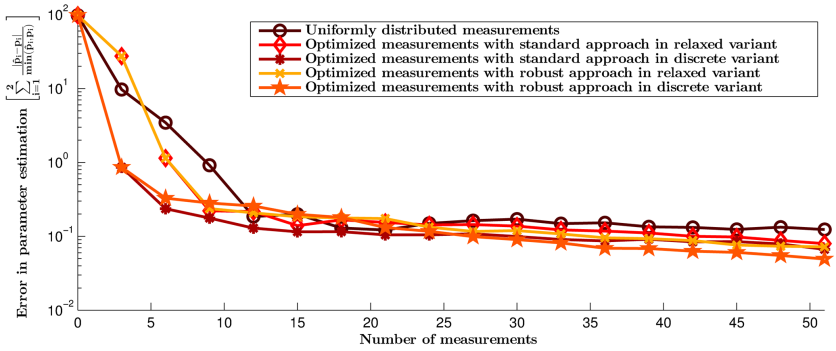


Figure 1. Averaged error in the parameter estimation from 10 optimization runs with the C_2-model and three measurement per iteration with standard deviation 10^{-2} of the measurement error.

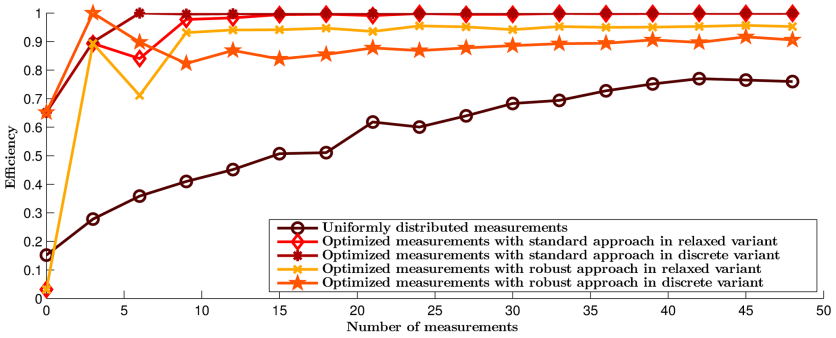


Figure 2. Averaged error in the parameter estimation from 10 optimization runs with the C_3-model and three measurement per iteration with standard deviation 10^{-2} of the measurement error.

more, the best achieved accuracy was worse by about a factor of 10 compared to the best accuracy achieved by (standard) optimized experimental designs.

The standard approach for optimizing experimental designs resulted in a slightly better accuracy compared to the robust approach.

For both approaches, the difference between the accuracy achieved with the exact solutions of the discrete optimization problem and the accuracy achieved with the approximate solutions was small but recognizable and almost constant over the iterations. Also in these experiments, the solutions of the relaxed continuous optimization problems were almost all nearly integer.

Again, the different standard deviations of the measurement errors only influenced the best achieved accuracy.

4.3.3 Conclusions regarding the approach for optimizing experimental designs

Optimized experimental designs provided a much more accurate parameter estimation than uniformly distributed experimental designs independent of the chosen optimization approach. Furthermore, only about half as many measurements were needed to archive the same accuracy with optimized experimental designs as with uniformly distributed experimental designs. In the more complex model, the difference was even bigger.

The robust approach did not achieved higher accuracy compared to the standard approach. In the complex model, the robust approach was even slightly less accurate. This may indicate that the gain in accuracy by taking into account the nonlinearity is offset by the additional approximations in the robust approach. Since a considerably higher computational effort is associated with the robust approach, the standard approach should be preferred, at least for these models.

The exact solutions of the discrete optimization problems yielded only slightly better accuracy gains compared to its approximate solutions. The fact that the approximate solutions were almost all nearly integer also argues for the approximate solving. This circumstance was also observed in Körkel (2002) and Körkel et al. (2004). For these reasons and because the exact solving requires much more computational effort, the approximate solving should be preferred, at least for these models.

4.4 Efficiency for the experimental designs

We also calculated the efficiencies of the used experimental designs. Some results are illustrated in Figs. 3 and 4.

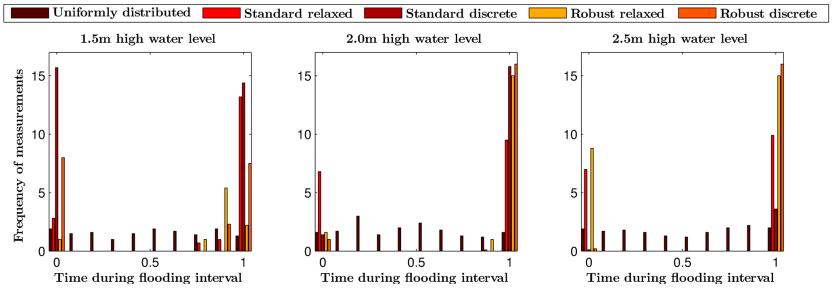


Figure 3. Averaged efficiency for the experimental designs from 10 optimization runs with the C_2-model and three measurement per iteration with standard deviation 10^{-2} of the measurement error.

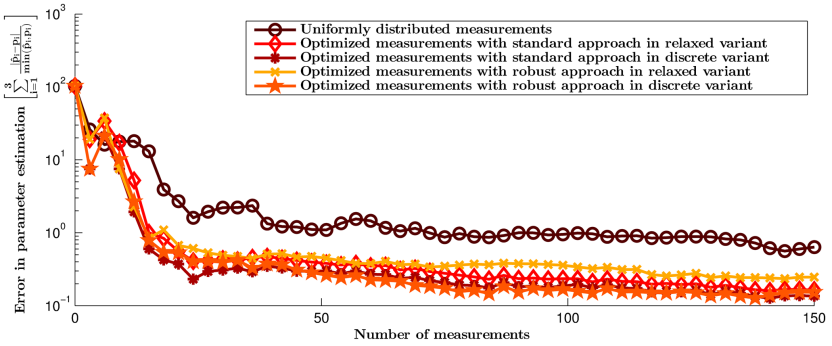


Figure 4. Averaged efficiency for the experimental designs from 10 optimization runs with the C_3-model and three measurement per iteration with standard deviation 10^{-2} of the measurement error.

The results emphasized the already seen importance of the optimization of the experimental designs. In particular, the advantage in the case of the few measurements carried out so far was highlighted. Again, the slight advantage of the standard approach over the robust approach was visible. With increasing number of accomplished measurements, the selection strategy of new measurements became less important as the amount and thus the influence of the new measurements compared to those of the accomplished measurements decreased.

4.5 Distribution of optimal measuring points

In this subsection, we compare the distribution of the measuring points optimized in the previously described numerical experiments. Graphical representation of the distribution of the measuring points from some numerical experiments are shown in Figs. 5 and 6.

4.5.1 Distribution for the C_2-model

The optimized measuring points were almost exclusively located at the start and end of the inundation periods. At the start of the inundation period, both approaches in the exact variant favored lower high water levels unlike both approaches in the approximate variant which favored higher high water levels. At the end of the inundation period, the standard approach in both variants favored lower high water levels unlike the robust approach in both variants which favored higher high water levels.

4.5.2 Distribution for the C_3-model

For the C_3-model the optimized measuring points accumulated at the end of the inundation periods. All approaches favored lower high water levels. With an increasing number of measurements per iteration, the robust approach in both variants also preferred measurements in the middle of the inundation periods with the highest high water level.

4.5.3 Conclusions regarding the distribution of optimal measuring points

The numerical experiments showed that measurements at the start and end of the inundation periods should be preferred for the C_2-model.

Measurements at the start of the inundations can be justified by the fact that one parameter of the model is the concentration at the start of the inundation. The fact that the settling velocity as second model parameter most affects the concentration at the end of the inundations justifies measurements here. This can be confirmed by an examination of the ordinary differential equation of the model derived with respect to the settling velocity. The derivative of the model with re-

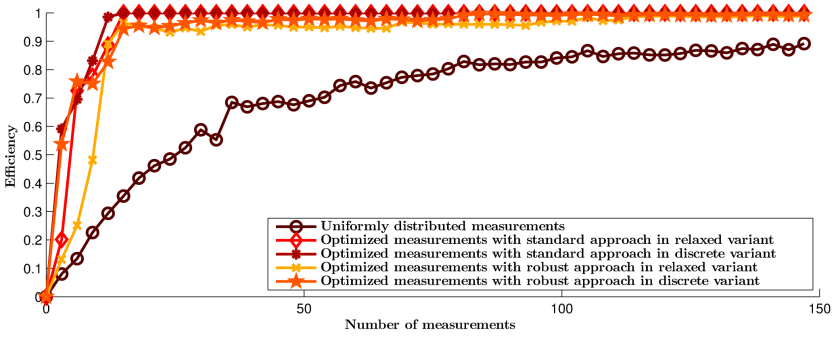


Figure 5. Averaged frequency of measurements from 10 optimization runs with the C_2-model and three measurement per iteration with standard deviation 10^{-2} of the measurement error.

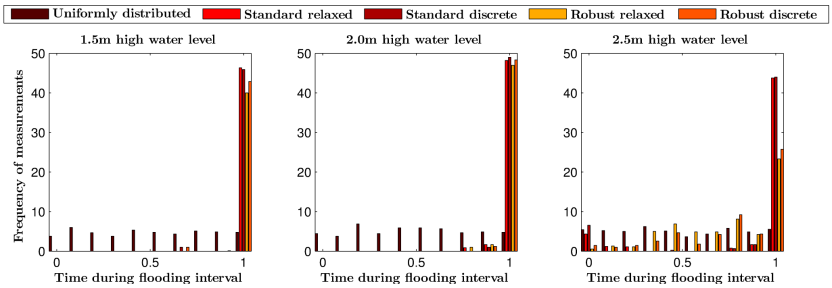


Figure 6. Averaged frequency of measurements from 10 optimization runs with the C_3-model and three measurement per iteration with standard deviation 10^{-2} of the measurement error.

spect to the settling velocity is zero at the start of the inundation and is getting smaller the further the inundation progresses. Its absolute greatest value it thus reached at the end of the inundation.

The experiments with the C_3-model showed that here measurements at end of the inundation periods should be preferred. In this model, the concentration at the start is no parameter but is affected by a parameter that also influences the settling velocity. For this reason, measurements are not suggested at the start.

For both models the high water level seemed to play a minor role for the choice of measuring points.

As a rule of thumb, one can say that measurements should be carried out at the end of an inundation period and also some at the start if the C_2-model is used.

5 Conclusions

In this paper we presented two different approaches for optimizing experimental design for parameter estimations. One method was based on the linearization of the model with respect to its parameters, the other takes into account a possible nonlinearity of the model parameters. Both methods were implemented in our presented *Optimal Experimental Design Toolbox* for MATLAB.

By employing the presented approach for two existing salt marsh models, we showed that model parameters can be determined much more accurately if the corresponding measurement conditions were optimized. In particular for time-consuming or costly measurements, it is useful to optimize the measurement conditions with the *Optimal Experimental Design Toolbox*.

This gain in accuracy is not limited to the application examples. In general, using the implemented methods, the accuracy of the parameters of any model can be maximized while minimizing the measurement cost, provided that the related assumptions are fulfilled. However, the required execution time for the optimization increases with the number of model parameters and (accomplished and selectable) measurements. Parallelization techniques in the optimization as well as in the model evaluation itself can counteract this effect.

In addition to the parallelization, the optimization in the toolbox could also be extended to techniques of globalization, so that the chance of getting into a local minimum is reduced.

The results concerning the application examples have not significantly differed despite the various approaches for optimizing experimental design. For this reason, the approach with the least computational effort is recommended. However, this recommendation can not be applied readily to other

(more complex) models. Here, the performance of the approaches should be compared again if possible.

Furthermore, it has been found that measurements at the beginning and end of the inundation period are particularly important for the application examples. The high water level of the inundation seemed to play a minor role. These results, however, can not be applied easily to other models. Usually, a separate optimization of experimental design makes sense here.

Code availability

The *Optimal Experimental Design Toolbox* is available under the GNU General Public License (Foundation, 2007) at a Git repository (Reimer, 2015). In addition to the toolbox, including commented source code and a user manual, an implementation of the application examples is also available.

Acknowledgements. We would like to thank the referees for their comments that helped us to clarify and improve this paper.

This project was funded by the Deutsche Forschungsgemeinschaft (DFG) as part of the Kiel cluster "The Future Ocean".

Edited by: J. Neal

References

Amemiya, T.: Nonlinear regression models, Handbook of econometrics, 1, 333–389, 1983.

Bartholdy, J. and Aagaard, T.: Storm surge effects on a back-barrier tidal flat of the Danish Wadden Sea, Geo-Mar. Lett., 20, 133–141, doi:10.1007/s003670000048, 2001.

Broyden, C. G.: The Convergence of a Class of Double-rank Minimization Algorithms: 2. The New Algorithm, IMA J. Appl. Math., 6, 222–231, doi:10.1093/imamat/6.3.222, 1970.

Burt, T.: Field Settling Velocities of Estuary Muds, in: Estuarine Cohesive Sediment Dynamics, edited by: Mehta, A., Vol. 14 of Lecture Notes on Coastal and Estuarine Studies, 126–150, Springer New York, doi:10.1007/978-1-4612-4936-8_7, 1986.

Coleman, T. F. and Li, Y.: On the convergence of reflective Newton methods for large-scale nonlinear minimization subject to bounds, Math. Program., 67, 189–224, 1994.

Coleman, T. F. and Li, Y.: An Interior, Trust Region Approach for Nonlinear Minimization Subject to Bounds, SIAM J. Optimiz., 6, 418–445, 1996.

Cramér, H.: Mathematical Methods of Statistics, Almqvist & Wiksells Akademiska Handböcker, Princeton University Press, 1946.

El-Monsef, M. M. E. A., Rady, E. A., and Seyam, M. M.: Relationships among Several Optimality Criteria, Interstat, 2009.

Fletcher, R.: A new approach to variable metric algorithms, Comput. J., 13, 317–322, doi:10.1093/comjnl/13.3.317, 1970.

Foundation, F. S.: GNU General Public License, available at: http://www.gnu.org/licenses/gpl-3.0-standalone.html (last access: 30 July 2014), 2007.

Goldfarb, D.: A family of variable-metric methods derived by variational means, Math. Comput., 24, 23–26, 1970.

Guest, T. and Curtis, A.: Iteratively constructive sequential design of experiments and surveys with nonlinear parameter-data relationships, J. Geophys. Res.-Sol., 114, B04307, doi:10.1029/2008JB005948, 2009.

IPCC: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, edited by: Stocker, T. F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S. K., Boschung, J., Nauels, A., Xia, Y., Bex, B., and Midgley, P. M., Cambridge University Press, Cambridge, United Kingdom, New York, NY, USA, doi:10.1017/CBO9781107415324, 2013.

Jennrich, R. I.: Asymptotic Properties of Non-Linear Least Squares Estimators, Ann. Math. Stat., 40, 633–643, doi:10.1214/aoms/1177697731, 1969.

Krone, R. B.: A method for simulating historic marsh elevations, in: Coastal Sediments (1987), pp. 316–323, ASCE, 1987.

Körkel, S.: Numerische Methoden für Optimale Versuchsplanungsprobleme bei nichtlinearen DAE-Modellen, PhD thesis, Ruprecht-Karls-Universität Heidelberg, 2002.

Körkel, S., Kostina, E., Bock, H. G., and Schlöder, J. P.: Numerical methods for optimal control problems in design of robust optimal experiments for nonlinear dynamic processes, Optim. Method. Softw., 19, 327–338, doi:10.1080/10556780410001683078, 2004.

Levenberg, K.: A method for the solution of certain problems in least squares, Q. Appl. Math., 2, 164–168, 1944.

Malinvaud, E.: The Consistency of Nonlinear Regressions, Ann. Math. Stat., 41, 956–969, 1970.

Marquardt, D. W.: An Algorithm for Least-Squares Estimation of Nonlinear Parameters, J. Soc. Ind. Appl. Math., 11, 431–441, 1963.

MathWorks: Matlab R2014a Primer, Natick, Massachusetts, release 2014a Edn., available at: http://www.mathworks.de/help/releases/R2014a/pdf_doc/matlab/getstart.pdf (last access: 12 March 2015), 2014.

Moré, J.: The Levenberg-Marquardt algorithm: Implementation and theory, in: Numerical Analysis, edited by: Watson, G., Vol. 630 of Lecture Notes in Mathematics, 105–116, Springer Berlin Heidelberg, doi:10.1007/BFb0067700, 1977.

Nocedal, J. and Wright, S.: Numerical Optimization, Springer series in operations research and financial engineering, Springer, New York, 1999.

Pejrup, M. and Mikkelsen, O. A.: Factors controlling the field settling velocity of cohesive sediment in estuaries, Estuar. Coast. Shelf Sci., 87, 177–185, doi:10.1016/j.ecss.2009.09.028, 2010.

Rao, R. C.: Information and the accuracy attainable in the estimation of statistical parameters, Bull. Calcutta Math. Soc., 37, 81–91, 1945.

Reimer, J.: oed: Optimal Experimental Design Toolbox: Version 1.0, doi:10.5281/zenodo.16004, 2015.

Schuerch, M., Vafeidis, A., Slawig, T., and Temmerman, S.: Modeling the influence of changing storm patterns on the ability of a salt marsh to keep pace with sea level rise, J. Geophys. Res.-Earth, 118, 84–96, doi:10.1029/2012JF002471, 2013.

Seber, G. A. F. and Wild, C. J.: Nonlinear regression, Wiley series in probability and statistics, Wiley-Interscience, 2003.

Shanno, D. F.: Conditioning of Quasi-Newton Methods for Function Minimization, Math. Comput., 24, 647–656, 1970.

Stigler, S. M.: Gauss and the Invention of Least Squares, Ann. Stat., 9, 465–474, 1981.

Temmerman, S., Govers, G., Meire, P., and Wartel, S.: Modelling long-term tidal marsh growth under changing tidal conditions and suspended sediment concentrations, Scheldt estuary, Belgium, Mar. Geol., 193, 151–169, doi:10.1016/S0025-3227(02)00642-4, 2003.

Temmerman, S., Govers, G., Wartel, S., and Meire, P.: Modelling estuarine variations in tidal marsh sedimentation: response to changing sea level and suspended sediment concentrations, Mar. Geol., 212, 1–19, doi:10.1016/j.margeo.2004.10.021, 2004.

van Leussen, W.: The variability of settling velocities of suspended fine-grained sediment in the Ems estuary, J. Sea Res., 41, 109–118, doi:10.1016/S1385-1101(98)00046-X, 1999.

Walter, É. and Pronzato, L.: Identification of parametric models from experimental data, Communications and control engineering, Springer, New York, 1997.

White, H.: Nonlinear Regression on Cross-Section Data, Econometrica, 48, 721–746, 1980.

White, H. and Domowitz, I.: Nonlinear Regression with Dependent Observations, Econometrica, 52, 143–162, 1984.

Wu, C.-F.: Asymptotic Theory of Nonlinear Least Squares Estimation, Ann. Stat., 9, 501–513, 1981.

Corrigendum to Geosci. Model Dev., 8, 791–804, 2015
https://doi.org/10.5194/gmd-8-791-2015-corrigendum

Geoscientific Model Development
Open Access
EGU

Corrigendum to "Optimization of model parameters and experimental designs with the Optimal Experimental Design Toolbox (v1.0) exemplified by sedimentation in salt marshes" published in Geosci. Model Dev., 8, 791–804, 2015

J. Reimer[1], **M. Schuerch**[2], **and T. Slawig**[1]

[1]Institute of Computer Science, Future Ocean – Kiel Marine Sciences, Christian-Albrechts-Universität zu Kiel, 24098 Kiel, Germany
[2]Institute of Geography, Future Ocean – Kiel Marine Sciences, Christian-Albrechts-Universität zu Kiel, 24098 Kiel, Germany

Correspondence: J. Reimer (jor@informatik.uni-kiel.de)

Published: 3 December 2019

In this corrigendum, we explain that the images in Figs. 2, 3, 4 and 5 were mixed up in the original publication. The captions are in the correct order. The original image of Fig. 2 should actually be Fig. 3; the image of Fig. 3 should actually be Fig. 5; the image of Fig. 4 should actually be Fig. 2; and the image of Fig. 5 should actually be Fig. 4.

Numerical Algorithms manuscript No.
(will be inserted by the editor)

Approximation of Hermitian Matrices by Positive Semidefinite Matrices using Modified Cholesky Decompositions

Joscha Reimer

May 4, 2019

Keywords linear algebra, matrix approximation algorithm, modified Cholesky decomposition, positive semidefinite matrix, positive definite matrix, Cholesky decomposition, Hermitian matrix, symmetric matrix

Abstract A new algorithm to approximate Hermitian matrices by positive semidefinite Hermitian matrices based on modified Cholesky decompositions is presented. In contrast to existing algorithms, this algorithm allows to specify bounds on the diagonal values of the approximation.

It has no significant runtime and memory overhead compared to the computation of a classical Cholesky decomposition. Hence it is suitable for large matrices as well as sparse matrices since it preserves the sparsity pattern of the original matrix.

The algorithm tries to minimize the approximation error in the Frobenius norm as well as the condition number of the approximation. Since these two objectives often contradict each other, it is possible to weight these two objectives by parameters of the algorithm. In numerical experiments, the algorithm outperforms existing algorithms regarding these two objectives.

A Cholesky decomposition of the approximation is calculated as a byproduct. This is useful, for example, if a corresponding linear equation should be solved.

A fully documented and extensively tested implementation is available. Numerical optimization and statistics are two fields of application in which the algorithm can be of particular interest.

Kiel University,
Department of Computer Science,
Algorithmic Optimal Control - CO_2 Uptake of the Ocean,
24098 Kiel, Germany
E-mail: joscha.reimer@email.uni-kiel.de

1 Introduction

Algorithms for approximating Hermitian matrices by positive semidefinite Hermitian matrices are useful in several areas. In stochastics they are needed to transform nonpositive semidefinite estimations of covariance and correlation matrices to valid estimations [50, 44, 27, 23]. In optimization they are needed to deal with nonpositive definite Hessian matrices in Newton type methods [20, 40, 8].

The existing algorithms have different disadvantages, which will be outlined below. A new algorithm without these disadvantages is presented in Section 2 where it is also examined in detail. An implementation is introduced in Section 3 together with numerical experiments and corresponding results. Conclusions are drawn in Section 4.

1.1 Objectives of approximation algorithms

In order to evaluate the existing algorithms, objectives of an ideal approximation algorithm are established. For this, let $A \in \mathbb{C}^{n\times n}$ be an Hermitian matrix and $B \in \mathbb{C}^{n\times n}$ its approximation. The first three objectives are the following:

(**O1**) B is positive semidefinite.

(**O2**) The approximation error $\|B - A\|$ is small.

(**O3**) The condition number $\kappa(B) = \|B\|\,\|B^{-1}\|$ is small.

In addition to the approximation error, the condition number of the approximation is usually important as well, since, for example, often linear equations including the approximation have to be solved.

The three objectives (**O1**), (**O2**) and (**O3**) are sometimes contradictory. Hence, an ideal algorithm would allow to prioritize between (**O2**) and (**O3**). The norm used in (**O2**) and (**O3**) may depend on the actual application. Typical choices are the spectral norm or the Frobenius norm.

Especially for large matrices, the execution time of the algorithm as well as the needed memory are important. The fastest way to test whether a matrix is positive definite is to try to calculate its Cholesky decomposition [24]. This needs $\frac{1}{3}n^3 + \mathcal{O}(n^2)$ basic operations in the dense real valued case. The approximation algorithm cannot be expected to be faster but at least asymptotically as fast. Thus, the next two objectives are:

(**O4**) The algorithm requires at most $\mathcal{O}(n^2)$ more basic operations than the calculation of a Cholesky decomposition of B.

(**O5**) The algorithm needs to store $\mathcal{O}(n)$ numbers besides A and B and allows to overwrite A with B.

If A is a sparse matrix, B should have the same sparsity pattern. This allows an effective overwriting and is essential if the corresponding dense matrix would be to big to store. Thus, the next objective is:

(**O6**) $A_{ij} = 0$ implies $B_{ij} = 0$.

For correlation matrices it is crucial that B has only ones as diagonal values. This is the reason for the last objective:

(**O7**) The diagonal of B can be predefined.

Similar objectives to (**O1**), (**O2**), (**O3**) and (**O4**) have been used in [53, 54,7,15]. Here, another objective has been established: If A is "sufficiently" positive definite, B should be equal to A. This objective is not explicitly listed here and should be covered by (**O2**).

1.2 Existing approximation methods

An overview of existing methods to approximate Hermitian matrices by positive semidefinite Hermitian matrices is provided next. They are evaluated using the objectives mentioned above.

The minimal approximation error can be achieved by computing an eigendecomposition and replacing negative eigenvalues [24,25]. This was done in statistics [32,50] as well as in optimization [40, Chapter 3.4], [20, Chapter 4.4.2.1]. However, this does not meet (**O4**), (**O6**) and (**O7**).

It is also possible to calculate approximations with minimal approximation error and the restriction that all diagonal values are one [27,3,4,28]. These methods could be extended so that the approximation has arbitrary predefined (nonnegative) diagonal values. Nevertheless, these methods do not meet (**O4**) and (**O6**).

Another method, especially common in optimization, is to add a predefined positive definite matrix multiplied by a sufficiently large scalar to the original matrix. The predefined matrix is usually the identity matrix or a diagonal matrix. The scalar is usually determined by increasing a value until the resulting approximation can be successfully Cholesky factorized. This method is also used in a modified Newton's method [21,8,40] and the Levenberg-Marquardt method [37,38,8]. However, (**O4**) and (**O7**) are not met.

A well-known method, in statistics, is a convex combination with a predefined positive definite matrix. In this context it is based on the concept of shrinkage estimator [55,14,50]. The positive definite matrix is again usually the identity matrix or a diagonal matrix. Only the convex combination factor has to be determined. This is usually done by examining the underlying statistical problem [6,16,30,35,36,52,58]. However, methods without using any statistical assumptions exist as well [23]. None of these meet (**O4**) and (**O7**).

Other methods used, especially in optimization, are modified Cholesky algorithms [19,20,53,54,39,7,15]. These compute a variant of a Cholesky decomposition like a LDL^T, a LBL^T or a LTL^T decomposition. Here L is a lower triangular matrix, D is a diagonal matrix, B is a block diagonal matrix with block size smaller of one or two and T is a tridiagonal matrix. During or after the calculation of these decompositions, their factors are modified so that they represent a positive definite matrix. The methods based on LBL^T

decomposition [39,7] do not meet (**O4**), (**O6**) and (**O7**), the ones based on LTL^T decomposition [15] do not meet (**O6**) and (**O7**) and the ones based on LDL^T decomposition [19,20,53,54,15] do not meet (**O7**).

Hence, none of the existing methods meet all objectives. However, methods that do not meet (**O7**) can be extended to meet this objective. For that the calculated approximation is multiplied by a suitable chosen diagonal matrix from both sides. This does not affect (**O1**), (**O4**), (**O5**) and (**O6**). So the modified Cholesky method based on LDL^T decomposition could meet all objectives if they are extended to meet (**O7**).

The new method presented in Section 2 is a modified Cholesky method based on LDL^T decomposition as well. Contrary to the already published methods of this kind, this methods modifies not only the matrix D but also the matrix L during their calculation. In this way, the algorithm meets all objectives. Furthermore it better meets (**O2**) and (**O3**) than the other extended methods based on LDL^T decomposition as shown in Section 3 by numerical experiments.

2 The approximation algorithm

The algorithm MATRIX which approximates Hermitian matrices by positive semidefinite matrices Hermitian is presented and analyzed in this section.

2.1 The MATRIX and the DECOMPOSITION algorithm

Previous modified Cholesky methods based on LDL^T decomposition [19,20, 53,54,15] applied to a symmetric matrix A try to calculate its LDL^T decomposition. While doing so, they increase some of the values in the diagonal matrix D. Hence, they result in a decomposition of a positive definite matrix $A + \Delta$, where Δ is a diagonal matrix with values greater or equal to zero. However, is this way, the approximation $A + \Delta$ cannot have predefined diagonal elements.

The key idea of the new algorithm is to modify the off-diagonal values of A instead or in addition to its diagonal values. In detail, the Hermitian positive definite approximation $B \in \mathbb{C}^{n\times n}$ of an Hermitian $A \in \mathbb{C}^{n\times n}$ is defined as

$$B_{ij} := \hat{\omega}_{ij} A_{ij} \text{ if } i \neq j \text{ and } B_{ii} := A_{ii} + \delta_i$$

where $\hat{\omega}_{ij} \in [0,1]$, $\hat{\omega}_{ij} = \hat{\omega}_{ji}$ and $\delta_i \in \mathbb{R}$ for all $i,j \in \{1,\ldots,n\}$.

If, for example, $\hat{\omega}_{ij} = 0$ and $\delta_i > |A_{ii}|$ for all $i,j \in \{1,\ldots,n\}$, then B is a diagonal matrix with only positive values and thus positive definite. If, on the other hand, $\hat{\omega}_{ij} = 1$ and $\delta_i = 0$ for all $i,j \in \{1,\ldots,n\}$, then $B = A$ and there is no approximation error.

The challenge is now to determine the values $\hat{\omega}_{ij}$ and δ_i such that the objectives established in Subsection 1.1 are met. This is where we use a (complex valued) modified Cholesky method based on LDL^H decomposition. During

the calculation of a LDL^H decomposition of A, we modify L and D if the matrix represented by the decomposition would become not positive definite, its condition number would become to high or the requirements on the diagonal values would be violated otherwise.

In detail, the off-diagonal values in the i-th row of L are multiplied by $\omega_i \in [0,1]$ and $\delta_i \in \mathbb{R}$ is added to the i-th diagonal value of D. This δ_i corresponds to the previously mentioned δ_i and ω to $\hat{\omega}$ such that $\hat{\omega}_{i,j} = \omega_{\max\{i,j\}}$ for all $i, j \in \{1, \ldots, n\}$. This relationship is discussed in Subsection 2.2. Furthermore symmetric permutation techniques are used to reduce the approximation error, the computational effort and the required memory.

The algorithm DECOMPOSITION, which computes the permuted modified LDL^H decomposition and the values ω and δ, is described in detail in Algorithm 1.

Algorithm 1 DECOMPOSITION

Input:
- $A \in \mathbb{C}^{n\times n}$ Hermitian, $x \in (\mathbb{R}\cup\{-\infty\})^n$, $y \in (\mathbb{R}\cup\{\infty\})^n$, $l \in \mathbb{R}\cup\{-\infty\}$, $u \in \mathbb{R}\cup\{\infty\}$, $\epsilon > 0$
- with $\max\{x_i, l\} \leq \min\{y_i, u\}$ for all $i \in \{1, \ldots, n\}$
- with $|x_i|, |l| \geq \epsilon$ or $|y_i|, |u| \geq \epsilon$ for all $i \in \{1, \ldots, n\}$

Output:
- $L \in \mathbb{C}^{n\times n}$, $d, \omega, \delta \in \mathbb{R}^n$, $p \in \{1, \ldots, n\}^n$

function DECOMPOSITION$(A, x, y, l, u, \epsilon)$
  $p_i \leftarrow i$ for all $i \in \{1, \ldots, n\}$
  $\alpha_i \leftarrow 0$ for all $i \in \{1, \ldots, n\}$
  for $i \leftarrow 1, \ldots, n$ **do**
    select $j \in \{i, \ldots, n\}$
    swap p_i with p_j and L_{ik} with L_{jk} for all $k \in \{1, \ldots, i-1\}$
    select $d_i \in [l, u]$, $\omega_{p_i} \in [0, 1]$ with $|d_i| \notin (0, \epsilon)$, $d_i + \alpha_{p_i}\omega_{p_i}^2 \in [x_{p_i}, y_{p_i}]$
    $L_{ij} \leftarrow \omega_{p_i} L_{ij}$ for all $j \in \{1, \ldots, i-1\}$
    $\delta_{p_i} \leftarrow d_i + \omega_{p_i}^2 \alpha_{p_i} - A_{p_i p_i}$
    for $j \leftarrow i+1, \ldots, n$ **do**
      if $d_i \neq 0$ **then**
        $L_{ji} \leftarrow \left(A_{p_j p_i} - \sum_{k=1}^{i-1} L_{jk}\overline{L}_{ik} d_k \right) (d_i)^{-1}$
        $\alpha_{p_j} \leftarrow \alpha_{p_j} + |L_{ji}|^2 d_i$
      else
        $L_{ji} \leftarrow 0$
      end if
    end for
  end for
  $L_{ii} \leftarrow 1$ and $L_{ij} \leftarrow 0$ for all $i, j \in \{1, \ldots, n\}$ with $j > i$
  return $(L, d, p, \omega, \delta)$
end function

The algorithm MATRIX, which computes the approximation B, is described in detail in Algorithm 2.

Algorithm 2 MATRIX

Input:
- $A \in \mathbb{C}^{n\times n}$ Hermitian, $x \in (\mathbb{R}\cup\{-\infty\})^n$, $y \in (\mathbb{R}\cup\{\infty\})^n$, $l \in \mathbb{R}\cup\{-\infty\}$, $u \in \mathbb{R}\cup\{\infty\}$, $\epsilon > 0$
- with $\max\{x_i, l\} \leq \min\{y_i, u\}$ for all $i \in \{1,\dots,n\}$
- with $|x_i|, |l| \geq \epsilon$ or $|y_i|, |u| \geq \epsilon$ for all $i \in \{1,\dots,n\}$

Output:
- $B \in \mathbb{C}^{n\times n}$

function MATRIX$(A, x, y, l, u, \epsilon)$
  $(L, d, p, \omega, \delta) \leftarrow$ DECOMPOSITION$(A, x, y, l, u, \epsilon)$
  $q_{p_i} \leftarrow i$ for all $i \in \{1,\dots,n\}$
  for $i \leftarrow 1,\dots,n$ **do**
    $B_{ii} \leftarrow A_{ii} + \delta_i$
    for $j \leftarrow i+1,\dots,n$ **do**
      if $q_i > q_j$ **then**
        $a \leftarrow j$, $b \leftarrow i$
      else
        $a \leftarrow i$, $b \leftarrow j$
      end if
      if $d_{q_a} \neq 0$ **or** $\omega_b = 0$ **then**
        $B_{ij} \leftarrow A_{ij}\omega_b$
      else
        $B_{ij} \leftarrow \sum_{k=1}^{q_a-1} L_{q_i k} d_k \overline{L}_{q_j k}$
      end if
    end for
  end for
  $B_{ji} \leftarrow \overline{B}_{ij}$ for all $i, j \in \{1,\dots,n\}$ with $j > i$
  return B
end function

The parameters l and u of the algorithms are lower and upper bounds on the diagonal values of D. The positive definiteness of B can be controlled by l as pointed out in Subsection 2.3. The parameters x and y are lower and upper bounds on the diagonal values of B as shown in Subsection 2.4. The condition number of B and the approximation error $\|B - A\|$ are influenced by x, y, l, u as demonstrated in Subsection 2.5 and 2.6, respectively. Moreover, they allow to prioritize a low approximation error or a low condition number. The numerical stability of the algorithms is controlled by ϵ.

The algorithms can be considered as a whole class of algorithms since there are many possibilities to choose the permutation and ω and δ as discussed in

Subsection 2.7 and 2.8. The algorithm is carefully designed, so that the overhead in computational effort and memory consumption compared to classical Cholesky decomposition algorithms is negligibly if ω and δ are chosen in a proper way, as shown in Subsection 2.9.

For the rest of this section, we use the following notation for the analysis of both algorithms.

Definition 1 Let

$$B := \text{MATRIX}(A, x, y, l, u, \epsilon)$$

where $(A, x, y, l, u, \epsilon)$ is some valid input for the algorithm with $A \in \mathbb{C}^{n\times n}$ and

$$(L, d, p, \omega, \delta) := \text{DECOMPOSITION}(A, x, y, l, u, \epsilon).$$

Define $D := \text{diag}(d)$ the diagonal matrix with d as the diagonal. Define $P \in \mathbb{R}^{n\times n}$ as the permutation matrix induced by p, which is

$$P_{ij} := \begin{cases} 1 \text{ if } j = p_i \\ 0 \text{ else} \end{cases} \quad \text{for all } i, j \in \{1, \ldots, n\}.$$

2.2 Representation of the approximation matrix

In this subsection it is shown that $B = P^T LDL^H P$. This means that MATRIX calculates the matrix represented by the decomposition calculated by DECOMPOSITION. This will be crucial for further investigation of MATRIX.

First, we prove that p is a permutation vector.

Lemma 1

$$\{p_i \mid i \in \{1, \ldots, n\}\} = \{1, \ldots, n\}.$$

Proof: In DECOMPOSITION, the variable p is initiated at line 2 of the algorithm so that $p_i = i$ for all $i \in \{1, \ldots, n\}$. After its initialization, the variable p is only changed in line 6. Here some of its components are swapped in each iteration. Thus $\{p_i \mid i \in \{1, \ldots, n\}\} = \{1, \ldots, n\}$ at the end of the algorithm.

□

Next it is shown how a corresponding inverse permutation vector can be defined.

Lemma 2 *Define*

$$q_{p_i} := i \text{ for all } i \in \{1, \ldots, n\}.$$

q is well defined and

$$p_{q_i} = i \text{ for all } i \in \{1, \ldots, n\}.$$

Proof: q is well defined due to Lemma 1. Let $i \in \{1, \dots, n\}$. Due to Lemma 1, a $j \in \{1, \dots, n\}$ exists with $p_j = i$. Furthermore $q_{p_j} = j$ due to the definition of q. Thus, $p_{q_i} = p_{q_{p_j}} = p_j = i$ follows.

□

A fast way to calculate LDL^H, using only A, ω, δ and p, is pointed out in the next lemma.

Lemma 3

$$(LDL^H)_{ii} = A_{p_i p_i} + \delta_{p_i}$$

and

$$(LDL^H)_{ij} = A_{p_i p_j}\omega_{p_{\max\{i,j\}}} \text{ if } d_{\min\{i,j\}} \neq 0 \text{ or } \omega_{p_{\max\{i,j\}}} = 0$$

for all $i, j \in \{1, \dots, n\}$ *with* $i \neq j$.

Proof: First some properties of the variable p during the execution of the algorithm are proved. Denote the for loop starting at line 4 of the algorithm the main for loop. Let $p^{(0)}$ be the value of the variable p directly before the main for loop and $p^{(i)}$ its value directly after its i-th iteration for each $i \in \{1, \dots, n\}$. Its final value is denoted by p.

Let $i \in \{1, \dots, n\}$. The variable p is initiated so that $p_i^{(0)} = i$. After its initialization, the variable p is only changed in line 6. Here the variables p_i and p_j are swapped for some $j \in \{i, \dots, n\}$ in the i-th iteration of the main for loop. Hence

$$\{p_i^{(j)} \mid i \in \{1, \dots, n\}\} = \{1, \dots, n\} \text{ for all } j \in \{1, \dots, n\}. \tag{1}$$

Furthermore the variable p_i is not changed anymore after the i-th iteration. Thus

$$p_i = p_i^{(j)} \text{ for all } i, j \in \{1, \dots, n\} \text{ with } i \leq j \tag{2}$$

and hence

$$p_i^{(i)} \neq p_j^{(j)} \text{ for all } i, j \in \{1, \dots, n\} \text{ with } i \neq j. \tag{3}$$

Next it is shown that all entries in the variables d, ω and δ are set once in the algorithm and are never changed after that. Hence, we do not need an index indicating the current iteration for this variables. Let d, ω and δ be the final value of the corresponding variables.

The value of d_i is set in the i-th iteration of the main for loop at line 7 and nowhere else. The values of ω_{p_i} and δ_{p_i} are set in the i-th iteration of the main for loop at line 7 and line 9 and due to (3) nowhere else. Furthermore ω_i and δ_i are set due to equation (2) and Lemma 1. Hence, all entries in the variables d, ω and δ are set once in the algorithm and are never changed after that.

Next properties of the variable L in the algorithm are proved which will lead to the result of this lemma. Denote with $L^{(i)}$ the value of the variable L directly after the i-th iteration of the main for loop for all $i \in \{1, \dots, n\}$. L denotes its final value.

Let $i, j \in \{1, \ldots, n\}$ with $j < i$. The variable L_{ij} is only changed in the j-th iteration at line 12 or line 15, in the i-th iteration at line 8 and maybe in the k-th iteration at line 6 for $k \in \{j+1, \ldots, i\}$. Thus, after the i-th iteration it is unchanged which means

$$L_{ij} = L_{ij}^{(k)} \text{ for all } i, j, k \in \{1, \ldots, n\} \text{ with } j < i \leq k. \tag{4}$$

In the i-th iteration, the variable L_{ij} might only be changed in line 6 and line 8. In line 6 the variable L_{ij} is only changed if it is swapped with the variable L_{kj} for some $k \in \{i+1, \ldots, n\}$. This is exactly the case if the variable p_i is swapped with the variable p_k. This together with line 8 and equation (1) implies

$$\begin{gathered} L_{ij}^{(i)} = \omega_{p_i^{(i)}} L_{kj}^{(i-1)} \text{ if } p_i^{(i)} = p_k^{(i-1)} \\ \text{for all } i, j, k \in \{1, \ldots, n\} \text{ with } j < i. \end{gathered}$$

This results with equation (2) and (4) in

$$\begin{gathered} L_{ij} = \omega_{p_i} L_{kj}^{(i-1)} \text{ if } p_i = p_k^{(i-1)} \\ \text{for all } i, j, k \in \{1, \ldots, n\} \text{ with } j < i. \end{gathered} \tag{5}$$

In the k-th iteration for all $k \in \{j+1, \ldots, i-1\}$, the variable L_{ij} might only be changed in line 6 due to a swap with the variable L_{kj}. This is exactly the case if the variable p_i is swapped with the variable p_k. This together with equation (1) implies

$$\begin{gathered} L_{ij}^{(l)} = L_{kj}^{(l-1)} \text{ if } p_i^{(l)} = p_k^{(l-1)} \\ \text{for all } i, j, k, l \in \{1, \ldots, n\} \text{ with } j < l < i. \end{gathered} \tag{6}$$

Equation (5) and (6) result in

$$\begin{gathered} L_{ij} = \omega_{p_i} L_{kj}^{(l)} \text{ if } p_i = p_k^{(l)} \\ \text{for all } i, j, k, l \in \{1, \ldots, n\} \text{ with } j \leq l < i. \end{gathered} \tag{7}$$

Now with this preparatory work, the main statement of this lemma can be proved. $L_{jj} = 1$ and $L_{jk} = 0$ for all $k \in \{j+1, \ldots, n\}$ due to line 19. This implies

$$(LDL^H)_{ij} = \sum_{k=1}^{n} L_{ik} \overline{L}_{jk} d_k = L_{ij} d_j + \sum_{k=1}^{j-1} L_{ik} \overline{L}_{jk} d_k.$$

Due to equation (1), a $l \in \{1, \ldots, n\}$ exists with $p_i = p_l^{(j)}$. Hence, equation (4) and (7) imply

$$L_{ij} d_j + \sum_{k=1}^{j-1} L_{ik} \overline{L}_{jk} d_k = L_{ij} d_j + \sum_{k=1}^{j-1} L_{ik} \overline{L}_{jk}^{(j)} d_k = \omega_{p_i} \left(L_{ij}^{(j)} d_j + \sum_{k=1}^{j-1} L_{ik}^{(j)} \overline{L}_{jk}^{(j)} d_k \right).$$

Thus

$$(LDL^H)_{ij} = \omega_{p_i} \left(L_{ij}^{(j)} d_j + \sum_{k=1}^{j-1} L_{ik}^{(j)} \overline{L}_{jk}^{(j)} d_k \right). \tag{8}$$

Due to line 12

$$A_{p_l^{(j)} p_j^{(j)}} = L_{lj}^{(j)} d_j + \sum_{k=1}^{i-1} L_{lk}^{(j)} \overline{L}_{jk}^{(j)} d_k \text{ if } d_j \neq 0.$$

Furthermore $p_i = p_l^{(j)}$ by definition of l and $p_j = p_j^{(j)}$ due to equation (2). This together with the previous two equations implies

$$(LDL^H)_{ij} = \omega_{p_i} A_{p_i p_j} \text{ if } d_j \neq 0.$$

Moreover with equation (8) it follows

$$(LDL^H)_{ij} = \omega_{p_i} A_{p_i p_j} \text{ if } \omega_{p_i} = 0.$$

D is a real-valued diagonal matrix and thus Hermitian. Hence, the matrix LDL^H is Hermitian as well. Since A is also Hermitian,

$$(LDL^H)_{ji} = (\overline{LDL^H})_{ij} = \overline{\omega_{p_i} A_{p_i p_j}} = \omega_{p_i} A_{p_j p_i} \text{ if } d_j \neq 0 \text{ or } \omega_{p_i} = 0. \tag{9}$$

The combination of the three previous equations results in

$$(LDL^H)_{ij} = A_{p_i p_j} \omega_{p_{\max\{i,j\}}} \text{ if } \omega_{p_{\max\{i,j\}}} \neq 0 \text{ or } d_{\min\{i,j\}} = 0$$
$$\text{for all } i, j \in \{1, \ldots, n\} \text{ with } i \neq j$$

which is one part of the statement of this lemma.

Since $L_{ii} = 1$ and $L_{ik} = 0$ for all $k \in \{i+1, \ldots, n\}$ due to line 19,

$$(LDL^H)_{ii} = \sum_{j=1}^{n} |L_{ij}|^2 d_j = d_i + \sum_{j=1}^{i-1} |L_{ij}|^2 d_j. \tag{10}$$

Define for every $k \in \{0, \ldots, i-1\}$ an $i_k \in \{1, \ldots, n\}$ with $p_i = p_{i_k}^{(k)}$ which exists uniquely due to equation (1). Then equation (7) implies

$$\sum_{j=1}^{i-1} |L_{ij}|^2 d_j = \omega_{p_i}^2 \sum_{k=1}^{i-1} |L_{i_k k}^{(k)}|^2 d_k. \tag{11}$$

Denote with $\alpha^{(0)}$ the value of the variable α directly before the main for loop and with $\alpha^{(i)}$ its value directly after its i-th iteration for each $i \in \{1, \ldots, n\}$.

Define for every $k \in \{0, \ldots, i-1\}$ an $i_k \in \{k+1, \ldots, n\}$ with $p_i = p_{i_k}^{(k)}$ which exists uniquely due to equation (1). Then

$$\alpha_{p_i}^{(i)} = \alpha_{p_{i_{i-1}}^{(i-1)}}^{(i-1)}$$

and

$$\alpha^{(k)}_{p^{(k)}_{i_k}} = \alpha^{(k-1)}_{p^{(k)}_{i_k}} + |L^{(k)}_{i_k k}|^2 d_k \text{ for all } k \in \{1, \ldots, i-1\}$$

due to line 13. Furthermore $\alpha^{(0)}_{p_{i_0}} = 0$ due to line 3. Hence

$$\alpha^{(i)}_{p_i} = \sum_{k=1}^{i-1} |L^{(k)}_{i_k k}|^2 d_k. \tag{12}$$

The combination of equation (10), (11) and (12) results in

$$(LDL^H)_{ii} = d_i + \omega^2_{p_i} \alpha^{(i)}_{p_i}.$$

Due to line 9 and equation (2), $d_i + \omega^2_{p_i} \alpha^{(i)}_{p_i} = \delta_{p_i} + A_{p_i p_i}$ and thus

$$(LDL^H)_{ii} = \delta_{p_i} + A_{p_i p_i}$$

which is the other part of the statement of this lemma.

□

The next lemma shows how B can be calculate using only A, δ, ω and p.

Lemma 4

$$B_{ii} = A_{ii} + \delta_i$$

and

$$B_{ij} = A_{ij}\omega_{b(i,j)} \text{ if } d_{q_{a(i,j)}} \neq 0 \text{ or } \omega_{b(i,j)} = 0$$

where

$$q_{p_i} := i,\ a(i,j) := \begin{cases} j \text{ if } q_i > q_j \\ i \text{ else} \end{cases},\ b(i,j) := \begin{cases} i \text{ if } q_i > q_j \\ j \text{ else} \end{cases}$$

for all $i, j \in \{1, \ldots, n\}$ *with* $i \neq j$.

Proof: First of all, q is well defined due to Lemma 2. Let $i \in \{1, \ldots, n\}$. In MATRIX, B_{ii} is set only at line 5 in the i-th iteration of the outer for loop at line 4. Due to this line $B_{ii} = A_{ii} + \delta_i$ and thus

$$B_{ii} = A_{ii} + \delta_i \text{ for all } i \in \{1, \ldots, n\}$$

Let $j \in \{i+1, \ldots, n\}$. In MATRIX, the variable B_{ij} is set only in line 13 or line 15 in the i-th iteration of the outer for loop at line 4 and the j-th iteration of the inner for loop at line 6. At this iteration the variables a and b have the the value $a(i,j)$ and $b(i,j)$, respectively, due to line 8 and line 10. Hence due to line 13,

$$B_{ij} = A_{ij}\omega_{b(i,j)} \text{ if } d_{q_{a(i,j)}} \neq 0 \text{ or } \omega_{b(i,j)} = 0$$
$$\text{for all } i, j \in \{1, \ldots, n\} \text{ with } i < j.$$

The variable B_{ji} is set only in line 19 so that $B_{ji} = \overline{B}_{ij}$. Hence, the previous equation implies

$$B_{ji} = \overline{B}_{ij} = \overline{A_{ij}\omega_{b(i,j)}} = \overline{A}_{ij}\omega_{b(i,j)} = A_{ji}\omega_{b(j,i)}$$
$$\text{if } d_{q_{a(j,i)}} \neq 0 \text{ or } \omega_{b(j,i)} = 0 \text{ for all } i, j \in \{1, \ldots, n\} \text{ with } i < j.$$

□

Next the main theorem of this subsection emphasizes the connection between MATRIX and DECOMPOSITION.

Theorem 1

$$B = P^T L D L^H P.$$

Proof: Define

$$q_{p_i} := i \text{ for all } i \in \{1, \ldots, n\}.$$

Due to Lemma 2, q is well defined and

$$p_{q_i} = i \text{ for all } i \in \{1, \ldots, n\}.$$

Let $i, j \in \{1, \ldots, n\}$ with $i < j$. Define a and b so that

$$q_a = \min\{q_i, q_j\} \text{ and } q_b = \max\{q_i, q_j\}.$$

This is well defined due to Lemma 1.

Due to line 15 of MATRIX and the definition of the variables a and b in the algorithm,

$$B_{ij} = \sum_{k=1}^{q_a - 1} L_{q_i k} d_k \overline{L}_{q_j k} \text{ if } d_{q_a} = 0 \text{ and } \omega_b \neq 0.$$

Since L is a lower triangular matrix and due to the definition of q_a,

$$L_{q_i k} = 0 \text{ or } L_{q_j k} = 0 \text{ for all } k \in \{q_a + 1, \ldots, n\}.$$

Thus,

$$B_{ij} = \sum_{k=1}^{n} L_{q_i k} d_k \overline{L}_{q_j k} = (L D L^H)_{q_i q_j} \text{ if } d_{q_a} = 0 \text{ and } \omega_b \neq 0.$$

Furthermore Lemma 3 and 4 and the definition of q imply

$$B_{ij} = A_{ij}\omega_b = A_{p_{q_i} p_{q_j}} \omega_{p_{q_b}} = (L D L^H)_{q_i q_j} \text{ if } d_{q_a} \neq 0 \text{ or } \omega_b = 0.$$

Due to line 19 of MATRIX,

$$B_{ji} = \overline{B}_{ij} = (\overline{L D L^H})_{q_i q_j} = (L D L^H)_{q_j q_i}$$

Lemma 3 and Lemma 4 imply

$$B_{ii} = A_{ii} + \delta_i = A_{p_{q_i} p_{q_i}} + \delta_{p_{q_i}} = (L D L^H)_{q_i q_i}.$$

Thus

$$B_{ij} = (L D L^H)_{q_i q_j} \text{ for all } i, j \in \{1, \ldots, n\}.$$

The definition of P implies

$$P_{q_i i} = 1 \text{ and } P_{q_j i} = 0 \text{ for all } i, j \in \{1, \ldots, n\} \text{ with } i \neq j.$$

Hence,

$$(L D L^H)_{q_i q_j} = \sum_{k=0}^{n} \sum_{j=0}^{n} P_{ki} (L D L^H)_{kl} P_{lj} = (P^T L D L^H P)_{ij}$$
$$\text{for all } i, j \in \{1, \ldots, n\}.$$

□

2.3 Positive semidefinite approximation

MATRIX can be forced to calculate positive definite or positive semidefinite matrices using $l > 0$ or $l \geq 0$, respectively as shown in Theorem 2. Thus, MATRIX meets objective (**O1**) if $l \geq 0$ is chosen. To prove this theorem, it is first shown that the values of d are bounded below by l. For subsequent proofs, it is also shown that the values of d are bounded above by u and y.

Lemma 5

$$d_i \in [l, u] \cap \mathbb{R} \text{ and } |d_i| \notin (0, \epsilon)$$

and if $l \geq 0$,

$$d_i \leq y_{p_i}$$

for all $i \in \{1, \ldots, n\}$.

Proof: Let $i \in \{1, \ldots, n\}$. In DECOMPOSITION the variable d is only changed in line 7. Here d_i is chosen at the i-th iteration of the surrounding for loop so that $d_i \in [l, u] \cap \mathbb{R}$ and $|d_i| \notin (0, \epsilon)$. Apart from that, the variable d_i is not set or changed anymore, so

$$d_i \in [l, u] \cap \mathbb{R} \text{ and } |d_i| \notin (0, \epsilon) \text{ for all } i \in \{1, \ldots, n\}.$$

The variable α in DECOMPOSITION is only changed in line 3 and line 13. Due to this lines and the previous equation,

$$\alpha_i \geq 0 \text{ if } l \geq 0.$$

In line 7, d_i is also chosen so that $d_i + \omega_{p_i}^2 \alpha_{p_i} \leq y_{p_i}$. This implies, together with the previous equation,

$$d_i \leq y_{p_i} \text{ if } l \geq 0 \text{ for all } i \in \{1, \ldots, n\}.$$

□

Theorem 2 *B is positive semidefinite if $l \geq 0$ and positive definite if $l > 0$.*

Proof: Theorem 1 implies

$$z^H B z = z^H P^T L D L^H P z = (L^H P z)^H D (L^H P z)$$

for all $z \in \mathbb{C}^n$. Moreover L and P are invertible. Hence, B is positive semidefinite if $D_{ii} = d_i \geq 0$ and positive definite if $D_{ii} = d_i > 0$ for all $i \in \{1, \ldots, n\}$. Thus, Lemma 5 implies that B is positive semidefinite if $l \geq 0$ and positive definite if $l > 0$.

2.4 Diagonal values

MATRIX allows to define lower and upper bounds for the diagonal values of B using x and y as proved in Theorem 3. This allows to predefined diagonal values of B by setting both bounds to the desired diagonal values. Thus, MATRIX meets objective (**O7**) by appropriately selecting the parameters x and y.

It should be taken into account that the algorithm requires $x_i \leq u$ and $l \leq y_i$ for all $i \in \{1, \ldots, n\}$. Hence, if positive semidefinite approximations are required, only nonnegative values can be used as predefined diagonal values. However, this is not an actual restriction, since positive semidefinite matrices always have nonnegative diagonal values.

Theorem 3

$$x_i \leq B_{ii} \leq y_i \text{ for all } i \in \{1, \ldots, n\}.$$

Proof: In the MATRIX, DECOMPOSITION is called first to calculate L, d, p, ω and δ. Let $i \in \{1, \ldots, n\}$. At the i-th iteration of the outer for loop in DECOMPOSITION,

$$d_i + \omega_{p_i}^2 \alpha_{p_i} \in [x_{p_i}, y_{p_i}]$$

due to line 7 and

$$\delta_{p_i} = d_i + \omega_{p_i}^2 \alpha_{p_i} - A_{p_i p_i}$$

due to line 9 and thus also

$$A_{p_i p_i} + \delta_{p_i} \in [x_{p_i}, y_{p_i}].$$

The variables p_i and δ_{p_i} are not changed anymore after that. Thus

$$A_{p_i p_i} + \delta_{p_i} \in [x_{p_i}, y_{p_i}] \text{ for all } i \in \{1, \ldots, n\}$$

at the end of the algorithm. Due to Lemma 1,

$$\{p_i \in \{1, \ldots, n\}\} = \{1, \ldots, n\}$$

and thus

$$A_{ii} + \delta_i \in [x_i, y_i].$$

Lemma 4 states that

$$B_{ii} = A_{ii} + \delta_i$$

and thus

$$x_i \leq B_{ii} \leq y_i.$$

□

2.5 Condition number

The condition number of B can be controlled by l, u and y as shown in Theorem 4. Hence, MATRIX meets objective (**O3**) with suitable chosen parameters.

Theorem 4 *Let $l > 0$. Then*

$$\kappa_2(L) \leq 2\left(\frac{a}{l}\right)^{\frac{n}{2}},\ \kappa_2(D) \leq \frac{b}{l} \text{ and } \kappa_2(B) \leq 4\frac{a^n b}{l^{n+1}}$$
$$\text{with } a := \frac{1}{n}\sum_{i=1}^{n} y_i \text{ and } b := \min\{u, \max_{i=1,\ldots,n} y_i\}.$$

Proof: P is a permutation matrix and thus $\text{trace}(PBP^T) = \text{trace}(B)$. Furthermore, PBP^T is positive definite because a permutation matrix is invertible and B is positive definite due to Theorem 2. Moreover, $\kappa_2(PBP^T) = \kappa_2(B)$ because a permutation matrix is also orthogonal. Thus, Theorem 3 implies

$$\frac{\text{trace}(PBP^T)}{n} = \frac{\text{trace}(B)}{n} \leq \frac{1}{n}\sum_{i=1}^{n} y_i = a.$$

Theorem 1 states that

$$PBP^T = LDL^H.$$

Lemma 5 implies

$$l \leq D_{ii} \leq \min\{u, y_{p_i}\} \leq b \text{ for all } i \in \{1, \ldots, n\}$$

since $l \geq 0$. Hence, Theorem 9 in the appendix implies

$$\kappa_2(L) \leq 2\left(\frac{a}{l}\right)^{\frac{n}{2}},\ \kappa_2(D) \leq \frac{b}{l} \text{ and } \kappa_2(B) \leq 4\frac{a^n b}{l^{n+1}}.$$

□

2.6 Approximation error

The approximation error $\|B - A\|$ can be expressed using A, δ, ω and p as shown in the next theorem where it is also proved that the approximation error is bounded. For that, it is first demonstrated that δ is bounded.

Lemma 6 *Let $l \geq 0$. Then*

$$|\delta_i| \leq a + b \text{ for all } i \in \{1, \ldots, n\}$$
$$\text{with } a := \max_{i=1,\ldots,n} y_i \text{ and } b := \max_{i=1,\ldots,n} |A_{ii}|.$$

Proof: Let $i \in \{1, \ldots, n\}$. B is positive semidefinite due to Theorem 2 since $l \geq 0$. Hence,

$$0 \leq B_{ii} \text{ and } B_{ii} \leq y_i \leq a$$

due to Theorem 3. Furthermore

$$B_{ii} = A_{ii} + \delta_i$$

due to Lemma 4. Thus

$$|\delta_i| = |B_{ii} - A_{ii}| \leq |B_{ii}| + |A_{ii}| \leq a + b.$$

□

Theorem 5 *Let $l > 0$ or otherwise $d_i = 0$ imply $\omega_j = 0$ for all $i, j \in \{1, \ldots, n\}$ with $j \geq i$. Define $E := B - A$. Then*

$$\begin{aligned}\|E\|_2 \leq \|E\|_1 &= \|E\|_\infty \\ &= \max_{i=1,\ldots,n} \left(|\delta_{p_i}| + (1-\omega_{p_i}) \sum_{j=1}^{i-1} |A_{p_i p_j}| + \sum_{j=i+1}^{n} (1-\omega_{p_j}) |A_{p_i p_j}| \right) \\ &\leq a + b + (n-1)c\end{aligned}$$

and

$$\begin{aligned}\|E\|_F^2 &= \sum_{i=1}^{n} \left(\delta_{p_i}^2 + 2(1-\omega_{p_i})^2 \sum_{j=1}^{i-1} |A_{p_i p_j}|^2 \right) \\ &\leq n((a+b)^2 + (n-1)c^2)\end{aligned}$$

with

$$a := \max_{i=1,\ldots,n} y_i, b := \max_{i=1,\ldots,n} |A_{ii}| \text{ and } c := \max_{i,j=1,\ldots,n; i \neq j} |A_{ij}|.$$

Proof: Let $i, j \in \{1, \ldots, n\}$. Lemma 3 and Theorem 1 imply

$$B_{p_i p_j} = \begin{cases} A_{p_i p_j} \omega_{p_{\max\{i,j\}}} & \text{if } i \neq j \\ A_{p_i p_i} + \delta_{p_i} & \text{else} \end{cases}.$$

Thus,

$$E_{p_i p_j} = \begin{cases} (\omega_{p_{\max\{i,j\}}} - 1) A_{p_i p_j} & \text{if } i \neq j \\ \delta_{p_i} & \text{else} \end{cases}.$$

Furthermore

$$\{p_i \mid i \in \{1, \ldots, n\}\} = \{1, \ldots, n\}$$

due to Lemma 1. Hence, E is Hermitian because A is Hermitian. Thus, the properties of the norms imply

$$\|E\|_2 \leq \|E\|_1 = \|E\|_\infty.$$

Moreover

$$\begin{aligned}\|E\|_\infty &= \max_{i=1,\dots,n} \sum_{j=1}^{n} |E_{p_i p_j}| = \max_{i=1,\dots,n} \left(|E_{p_i p_i}| + \sum_{j=1}^{i-1} |E_{p_i p_j}| + \sum_{j=i+1}^{n} |E_{p_i p_j}| \right) \\ &= \max_{i=1,\dots,n} \left(|\delta_{p_i}| + (1-\omega_{p_i}) \sum_{j=1}^{i-1} |A_{p_i p_j}| + \sum_{j=i+1}^{n} (1-\omega_{p_j}) |A_{p_i p_j}| \right) \\ &\leq a + b + (n-1)c\end{aligned}$$

because $|\delta_i| \leq a+b$ and $\omega_i \in [0,1]$ due to Lemma 6 and line 7 in DECOMPOSITION. Additionally

$$\begin{aligned}\|E\|_F^2 &= \sum_{i=1}^{n} \left(|E_{p_i p_i}|^2 + 2 \sum_{j=1}^{i-1} |E_{p_i p_j}|^2 \right) \\ &= \sum_{i=1}^{n} \left(\delta_{p_i}^2 + 2(1-\omega_{p_i})^2 \sum_{j=1}^{i-1} |A_{p_i p_j}|^2 \right) \\ &\leq n(a+b)^2 + n(n-1)c^2.\end{aligned}$$

□

2.7 Choice of ω and d

The choice of ω and d in line 7 in DECOMPOSITION is arbitrary apart from that they must be feasible. However, their choice is crucial for the approximation error due to Theorem 5 and line 9 of DECOMPOSITION.

Based on this theorem the algorithm MINIMAL_CHANGE, presented in Algorithm 3, is derived which chooses ω and d so that in each iteration the additional approximation error in the Frobenius norm is minimized. This does not guaranteed that the overall approximation error is minimized but still results in a small approximation error as numerical tests in Subsection 3.2 have shown. Hence, MATRIX meets objective (**O2**) when using MINIMAL_CHANGE. It can be incorporated by replacing line 7 in DECOMPOSITION with the code snippet CHOOSE_d_ω presented in Algorithm 4.

MINIMAL_CHANGE was designed so that its needed number of basic operations and memory is negligible compared to the number of basic operations and memory needed by MATRIX as discussed in Subsection 2.9. This makes it possible to meet objectives (**O4**) and (**O5**) while using MINIMAL_CHANGE.

It also ensures that $B = A$ if A already meets the requirements on B. In detail, these are $x_i \leq A_{ii} \leq y_i$ and $\max\{l, \epsilon\} \leq D_{ii} \leq u$ for all $i \in \{1, \dots, n\}$, where D is the diagonal matrix of the LDL^H decomposition of PAP^T.

If several pairs (d, ω) minimize the additional approximation error, the one with the biggest d is chosen in MINIMAL_CHANGE. This results in absolute smaller values in L which reduces the condition number of B, as shown in

the proof of Theorem 9. Moreover the numerical stability of the algorithms is increased because a division by d is part of the algorithms.

Algorithm 3 MINIMAL_CHANGE

Input:
- $x \in \mathbb{R} \cup \{-\infty\}$, $y, u \in \mathbb{R} \cup \{\infty\}$, $l, \epsilon, \alpha, \beta, \gamma \in \mathbb{R}$ with $l, \alpha, \beta \geq 0$, $\epsilon > 0$, $\max\{l, \epsilon, x\} \leq \min\{u, y\}$ and $\beta = 0 \Rightarrow \alpha = 0$

Output:
- $d, \omega \in \mathbb{R}$

function MINIMAL_CHANGE$(x, y, l, u, \epsilon, \alpha, \beta, \gamma)$
  if $\max\{l, \epsilon, x - \alpha\} \leq \gamma - \alpha \leq \min\{u, y - \alpha\}$ **then**
    return $(\gamma - \alpha, 1)$
  end if
  $C \leftarrow \emptyset$
  if $\max\{l, \epsilon, x - \alpha\} \leq \min\{u, y - \alpha\}$ **then**
    $C \leftarrow \{(\min\{\max\{l, \epsilon, x - \alpha, \gamma - \alpha\}, u, y - \alpha\}, 1)\}$
  end if
  if $\alpha \neq 0$ **then**
    for $d \in (\{\max\{l, \epsilon\}\} \cap [x - a, \infty)) \cup (\{u\} \cap (-\infty, y]))$ **do**
      for $\omega \in \mathbb{R}$ with $2\alpha^2\omega^3 + (2\alpha(d - \gamma) + \beta)\omega - \beta = 0$ **do**
        $\omega \leftarrow \min\{\max\{\omega, \sqrt{\frac{\max\{x-d,0\}}{\alpha}}\}, \sqrt{\frac{y-d}{\alpha}}, 1\}$
        $C \leftarrow C \cup \{(d, \omega)\}$
      end for
    end for
  end if
  if $l = 0$ and $x \leq 0$ and $2\gamma \leq \epsilon$ **then**
    $C \leftarrow C \cup \{(0, 0)\}$
  end if
  return $(d, \omega) \in C$ with smallest $((d + \omega^2\alpha - \gamma)^2 + (\omega - 1)^2\beta, -d, \omega)$ in lexicographical order
end function

Algorithm 4 CHOOSE_d_ω

for $k \leftarrow i, \ldots, n$ **do**
  if $i = 1$ **then**
    $\beta_{p_k} \leftarrow 0$
  else
    $\beta_{p_k} \leftarrow \beta_{p_k} + 2|A_{p_k p_{i-1}}|^2$
  end if
end for
$(d_{p_i}, \omega_{p_i}) \leftarrow$MINIMAL_CHANGE$(x_{p_i}, y_{p_i}, l, u, \epsilon, \alpha_{p_i}, \beta_{p_i}, A_{p_i p_i})$

The next Theorem stats that MINIMAL_CHANGE chooses feasible d and ω which minimize in each iteration the additional approximation error.

Theorem 6 *Let*

$$d, \omega := \text{MINIMAL\_CHANGE}(x, y, l, u, \epsilon, \alpha, \beta, \gamma)$$

where $(x, y, l, u, \epsilon, \alpha, \beta, \gamma)$ *is some valid input for the algorithm. Let*

$$\Phi_* := \{(d, \omega) \mid d \in [\max\{l, \epsilon\}, u], \omega \in [0, 1], d + \omega^2\alpha \in [x, y]\},$$

$$\Phi_0 := \begin{cases} \{(0,0)\} \text{ if } \max\{l, x\} \leq 0 \\ \emptyset \text{ else} \end{cases}, \Phi := \Phi_* \cup \Phi_0$$

and

$$\Psi := \{(d, \omega) \in \Phi \mid f(d, \omega) = \min_{(\hat{d}, \hat{\omega}) \in \Phi} f(\hat{d}, \hat{\omega})\}$$

with $f : \mathbb{R}^2 \to \mathbb{R}, (d, \omega) \mapsto (d + \omega^2\alpha - \gamma)^2 + (\omega - 1)^2\beta$. *Then* $(d, \omega) \in \Psi$.

Proof: Φ is compact and f is continuous. Thus, f has a minimum on Φ due to Weierstrass's theorem [51, Theorem 4.16]. Hence, $\Psi \neq \emptyset$ and thus,

$$\Psi \cap \Phi_*^\circ \neq \emptyset \text{ or } \Psi \cap \partial\Phi_* \neq \emptyset \text{ or } \Psi \cap \Phi_0 \neq \emptyset \tag{13}$$

where Φ_*° denotes the interior of Φ_* and $\partial\Phi_*$ its boundary. Next these three cases are considered.

First consider the case that $\Psi \cap \Phi_*^\circ \neq \emptyset$. Then

$$\nabla f(d, \omega) = 0 \text{ for all } (d, \omega) \in \Psi \cap \Phi_*^\circ$$

due to [40, Theorem 12.3]. Furthermore

$$\nabla f(d, \omega) = \begin{pmatrix} 2(d + \omega^2\alpha - \gamma) \\ 4\alpha\omega(d + \omega^2\alpha - \gamma) + 2\beta(\omega - 1) \end{pmatrix}$$

for all $(d, \omega) \in \Phi_*^\circ$. This implies

$$\omega = 1 \text{ and } d = \gamma - \alpha \text{ for all } (d, \omega) \in \Psi \cap \Phi_*^\circ \text{ if } \beta \neq 0.$$

If $\beta = 0$, the algorithm requires $\alpha = 0$, which implies

$$(\gamma - \alpha, 1) \in \Psi \text{ if } \Psi \cap \Phi_*^\circ \neq \emptyset \text{ and } \beta = 0.$$

Hence,

$$(\gamma - \alpha, 1) \in \Psi \text{ if } \Psi \cap \Phi_*^\circ \neq \emptyset.$$

Thus, $\Psi \cap \Phi_*^\circ \neq \emptyset$ implies $(\gamma - \alpha, 1) \in \Phi_*$. Hence, $(\gamma - \alpha, 1)$ is returned by the algorithm in line 3 if $\Psi \cap \Phi_*^\circ \neq \emptyset$.

If $\Psi \cap \Phi_*^\circ = \emptyset$, the algorithm constructs a candidate set C and returns a minimizer of f on C in line 20. Hence, it remains to prove that

$$C \cap \Psi \neq \emptyset \text{ if } \Psi \cap \partial\Phi_* \neq \emptyset \text{ or } \Psi \cap \Phi_0 \neq \emptyset.$$

Consider now the case $\Psi \cap \partial\Phi_* \neq \emptyset$. Let $(d,\omega) \in \Psi \cap \partial\Phi_*$ and define $a := \max\{l, \epsilon\}$. Then

$$d \in \{a, u\} \text{ or } d + \omega^2\alpha \in \{x, y\} \text{ or } \omega \in \{0, 1\}.$$

If $\omega = 1$, the definitions of f and Φ_* imply

$$\max\{a, x - \alpha\} \leq \min\{u, y - \alpha\}$$

and

$$(d,\omega) = (\min\{\max\{a, x - \alpha, \gamma - \alpha\}, u, y - \alpha\}, 1).$$

This value is included in C at line 7.

If $\alpha = 0$, $(d,\omega) \in \Psi$ implies $(d, 1) \in \Psi$ for all $(d,\omega) \in \Phi$. Hence, the case $\alpha = 0$ is covered by the previous case where $\omega = 1$. Thus, assume $\alpha \neq 0$.

If $d + \omega^2\alpha = c$ for $c \in \{x, y\}$, $d \leq c$ and $\omega = \sqrt{\frac{c-d}{\alpha}}$. Since

$$f\left(d, \sqrt{\tfrac{c-d}{\alpha}}\right) = (c - \gamma)^2 + \left(\sqrt{\tfrac{c-d}{\alpha}} - 1\right)^2 \beta$$

and (d,ω) is a minimizer of f on Ψ, it follows

$$d \in \{c - \alpha, a, u\} \text{ if } d + \omega^2\alpha = c.$$

$d = c - \alpha$ any $d + \omega^2\alpha = c$ imply $\omega = 1$. The case $\omega = 1$ was already considered.

The case $d \in \{a, u\}$ is considered now. Then $(d,\omega) \in \Phi$ is equivalent to $\omega \in [\check{\omega}_d, \hat{\omega}_d]$ with

$$\check{\omega}_d := \sqrt{\tfrac{\max\{x-d,0\}}{\alpha}}, \quad \hat{\omega}_d := \sqrt{\tfrac{\min\{y-d,\alpha\}}{\alpha}}.$$

Hence, $d = u$ implies $y \geq u$ and $d = a$ implies $x - \alpha \leq a$. Define

$$\Omega_d := \{\omega \in \mathbb{R} \mid \tfrac{\partial}{\partial\omega} f(d,\omega) = 0\}.$$

$\omega \in (\check{\omega}_d, \hat{\omega}_d)$ implies $\omega \in \Omega_d$. $\omega = \check{\omega}_d$ implies $\min \Omega_d \leq \check{\omega}_d$ and $\omega = \hat{\omega}_d$ implies $\max \Omega_d \geq \hat{\omega}_d$. Hence

$$\omega \in \{\min\{\max\{\omega, \check{\omega}_d\}, \hat{\omega}_d\}\} \mid \omega \in \Omega_d\}$$

These values are included in C in line 13.

The last case is $\omega = 0$. This implies $d = a$, because (d,ω) is a minimizer of f on Φ. The case $d = a$ was already considered.

Hence,

$$C \cap \Psi \neq \emptyset \text{ if } \Psi \cap \partial\Phi_* \neq \emptyset.$$

Thus, it remains to show that $C \cap \Psi \neq \emptyset$ if $\Psi \cap \Phi_0 \neq \emptyset$. Hence, consider now the case $\Psi \cap \Phi_0 \neq \emptyset$. The definition of Φ_0 implies then $(0, 0) \in \Psi$ and $\max\{l, x\} \leq 0$. This implies $\epsilon \leq u, y$ due to the requirements of the algorithm. Thus, $(\epsilon, 0) \in \Phi$. Hence, since $(0, 0) \in \Psi$,

$$\gamma^2 + \beta = f(0, 0) \leq f(\epsilon, 0) = (\epsilon - \gamma)^2 + \beta = \gamma^2 - 2\epsilon\gamma + \epsilon^2 + \beta.$$

Thus, $\Psi \cap \Phi_0 \neq \emptyset$ implies $2\gamma \leq \epsilon$. $(0,0)$ is included in C in line 18 in this case. Hence

$$C \cap \Psi \neq \emptyset$$

and the algorithm returns a value in Ψ in all cases.

□

2.8 Permutation

Another part of DECOMPOSITION with some flexibility in its design is the permutation step in line 5 where the row and column for the current iteration are chosen. This choice drastically affects the output of DECOMPOSITION and thus of MATRIX, too. Several strategies for the permutation are conceivable.

A strategy to reduce the approximation error is to choose the permutation that minimizes the additional approximation error. To achieve this, in each iteration the additional approximation error for all remaining indices is computed and the one with the lowest additional approximation error is chosen. If this is the same for several indices, a higher value in d is preferred. As already stated in the previous subsection, this reduces the condition number of the approximation and increases the numerical stability. If these values are the same as well, a lower ω and then a lower index is preferred.

Another strategy is to prioritize higher values in d instead of lower additional approximation errors. This improves the condition number and the numerical stability even further and does not necessarily increase the total approximation error as numerical experiments have shown.

To use this strategy, line 8 in CHOOSE_d_ω can be replaced by the following code snippet CHOOSE_p_d_ω presented in Algorithm 5. Furthermore in DECOMPOSITION , the swap in line 6 has to be moved after CHOOSE_p_d_ω and line 5 could be removed.

Algorithm 5 CHOOSE_p_d_ω

```
1:  d̂ ← −∞
2:  for k ← i, ..., n do
3:      (d̃, ω̃) ← MINIMAL_CHANGE(x_{p_k}, y_{p_k}, l, u, ε, α_{p_k}, β_{p_k}, A_{p_k p_k})
4:      f̃ ← (d̃ + ω̃² α_{p_k} − A_{p_k p_k})² + (ω̃ − 1)² β_{p_k}
5:      if (−d̃, f̃, ω̃, k) < (−d̂, f̂, ω̂, j) in lexicographical order then
6:          j ← k, d̂ ← d̃, ω̂ ← ω̃, f̂ ← f̃
7:      end if
8:  end for
9:  (d_{p_j}, ω_{p_j}) ← (d̂, ω̂)
```

For sparse matrices, the permutation also affects the sparsity pattern of the matrix L. Hence, it would be beneficial to choose a permutation which reduces the number of nonzero values in L and thus reduces also the computational effort and the memory consumption. However, minimizing the number of nonzero values is a NP-complete problem [60].

However, several heuristic methods exist, which can reduce the number of nonzero values significantly. These are band reducing permutation algorithms like the CuthillMcKee algorithm [9] and the reverse CuthillMcKee algorithm [17], symmetric approximate minimum degree permutation algorithms [18], like for example [1], or symmetric nested dissection algorithms. A good overview is provided by [12, chapter 7] and [13, Chapter 8]. It should be taken into account that only symmetric permutation methods are applicable in our context.

2.9 Complexity

In the context of large matrices and limited resources, the needed run time and memory of MATRIX and DECOMPOSITION are crucial.

The fastest way to check if $A \in \mathcal{C}^{n\times n}$ is positive definite is to try to calculate a (classical) Cholesky decomposition of A, that is a lower triangular matrix L with $A = LL^H$ [26, Chapter 10], [22, Chapter 4.2]. This needs at worst $\frac{1}{3}n^3 + \mathcal{O}(n^2)$ basic operations and stores $2n^2 + \mathcal{O}(n)$ numbers in the real valued case. The needed memory can be reduced if only the lower triangles of A and L are stored. This would result in $n^2 + \mathcal{O}(n)$ numbers. It can be reduced even more if A can be overwritten by L. This would result in $\frac{1}{2}n^2 + \mathcal{O}(n)$ numbers.

MATRIX and DECOMPOSITION using CHOOSE_p_d_ω need at worst $\frac{1}{3}n^3 + \mathcal{O}(n^2)$ basic operations and memory for $2n^2 + \mathcal{O}(n)$ numbers in the real valued case, too. For this only a few small modifications are necessary which are explained below. Hence, both algorithms have asymptotically the same worst case number of basic operations and memory as an algorithm which calculates a Cholesky decomposition. Thus, their overhead is negligible and vanishes asymptotically. With some small modifications, it is also possible to overwrite the input matrix A with the output matrices L and B. Thus, MATRIX meets objective (**O4**) and (**O5**).

For MINIMAL_CHANGE, the number of needed basic operations and numbers that have to be stored is $\mathcal{O}(1)$. Hence, CHOOSE_p_d_ω needs $\mathcal{O}(n)$ basic operations and stores $\mathcal{O}(1)$ numbers. If CHOOSE_p_d_ω is used in DECOMPOSITION to choose the permutation as well as d and ω, $\mathcal{O}(n^2)$ additional basic operations have to be performed and $\mathcal{O}(n)$ additional numbers have to be stored.

In DECOMPOSITION, a crucial part for the number of needed operations is the calculation of L in line 12. Here the effort can be reduced by calculating and storing $LD^{\frac{1}{2}}$ instead of L first. After that L can be calculated with an effort of $\mathcal{O}(n^2)$ basic operations. This approach results in an overall worst case

number of $\frac{1}{3}n^3 + \mathcal{O}(n^2)$ basic operations plus the basic operations needed for the permutation and the choice of d and ω. Furthermore $2n^2 + \mathcal{O}(n)$ numbers have to be stored in DECOMPOSITION despite the memory needed for the choice of the permutation, d and ω. Hence, if CHOOSE_p_d_ω is used, the overall worst case number of basic operations is $\frac{1}{3}n^3 + \mathcal{O}(n^2)$ and $2n^2 + \mathcal{O}(n)$ numbers have to be stored.

The needed storage can be reduced by storing only one triangle of A and the lower triangle of L and by overwriting A with L. This would result in $n^2 + \mathcal{O}(n)$ and $\frac{1}{2}n^2 + \mathcal{O}(n)$ numbers, respectively. However, the permutation in DECOMPOSITION must be taken into account here. For this, the indexing of A in line 12 must be suitably adapted or A must be permuted. However, these modifications would not influence the $\frac{1}{3}n^3$ as the dominant part in the number of basic operations.

For MATRIX, at most $\frac{1}{3}n^3 + \mathcal{O}(n^2)$ basic operations plus the basic operations for choosing the permutation, d and ω are needed as well. This is because for each execution of line 15 in MATRIX, the execution of line 12 in DECOMPOSITION is once omitted. Thus, the worst case number of needed basic operations of MATRIX increases only by $\mathcal{O}(n^2)$ compared to the worst case number of needed basic operations of DECOMPOSITION.

At most $2n^2 + \mathcal{O}(n)$ numbers have to be stored in MATRIX plus the numbers that need to be stored for choosing the permutation, d and ω. To achieve this, B must overwrite L. If the strict lower triangle of L is allowed to overwrite the strict lower triangle of A, at most $n^2 + \mathcal{O}(n)$ numbers have to be stored plus the numbers for the choice of the permutation, d and ω.

Hence, MATRIX, with the small modifications mentioned above, needs at most $\frac{1}{3}n^3 + \mathcal{O}(n^2)$ basic operations and stores at most $2n^2 + \mathcal{O}(n)$ numbers if CHOOSE_p_d_ω is used to choose the permutation, d and ω. It stores only $n^2 + \mathcal{O}(n)$ numbers if it is allowed to overwrite the input matrix.

It would also be possible to reduce the needed memory to $\frac{1}{2}n^2 + \mathcal{O}(n)$ numbers by passing only the lower triangle of the matrices A, L and B. Since in this case A is then no longer available after the calculation of L, B must be calculated in the more expensive way shown in Theorem 1. This would result in $\frac{1}{3}n^3 + \mathcal{O}(n^2)$ additional basic operations.

In the complex valued case, the main statement remains the same: The overhead of MATRIX and DECOMPOSITION using CHOOSE_p_d_ω is negligible and vanishes asymptotically compared to an algorithm which calculates a (classical) Cholesky decomposition. In a similar way, an analysis can be carried out for the case where A is a sparse matrix.

3 Implementation and numerical experiments

An implementation of the algorithms MATRIX and DECOMPOSITION is presented in this section together with the performed numerical experiments.

3.1 Implementation

The algorithms MATRIX and DECOMPOSITION presented in Section 2 are implemented in a software library written in Python [43] called matrix-decomposition library [48]. Their implementation uses the MINIMAL_CHANGE algorithm and provides both permutation algorithms described in Subsection 2.8 as well as several fill reducing permutation algorithms for sparse matrices. In addition, the library provides many more approximation and decomposition algorithms together with various other useful functions regarding matrices and its decompositions.

The library is available at github [46]. It is based on NumPy [41], SciPy [33, 59] and scikit-sparse [49]. It was extensively tested using pytest [34] and documented using Sphinx [5]. The matrix-decomposition library and all required packages are open-source.

They can be comfortably installed using the cross-platform package manager Conda [2] and the Anaconda Cloud [45]. Here all required packages are installed during the installation of the matrix-decomposition library. The library is also available on the Python Package Index [47] and is thus installable with the standard Python package manager pip [42] as well.

3.2 Comparison with other approximation algorithms

The MATRIX algorithm has been compared with other modified Cholesky algorithms based on LDL^T decomposition by the resulting approximation errors and the condition numbers of the approximations. For the results presented here, we have use the Frobenius norm. However, the results using the spectral norm look similar.

The other algorithms are GMW81 [20], which is a refined version of [19], GMW1 [15] and GMW2 [15] which are based on GMW81, SE90 [53] and its refined version SE99 [54] as well as SE1 [15] which in turn is based on SE99. All these algorithms are implemented in the matrix-decomposition library [48]. These algorithms have been extended so that the approximation can have predefined diagonal values. For this, the calculated approximation was scaled by multiplying with a suitable diagonal matrix on both sides.

MATRIX has been configured so that the permutation strategy which prefers high values in D is used and no upper bound on the values in D is applied.

Different test scenarios were used. The first three scenarios are random correlation matrices disturbed by some additive unbiased noise which should be approximated by valid correlation matrices. The random correlation matrices have been generated by the algorithm described in [10]. The off-diagonal values of the symmetric noise matrices have been drawn from a normal distribution with expectation value zero and 0.1, 0.2 or 0.3 as standard deviation depending on the scenario. The diagonal values of the noise matrices were zero in all scenarios.

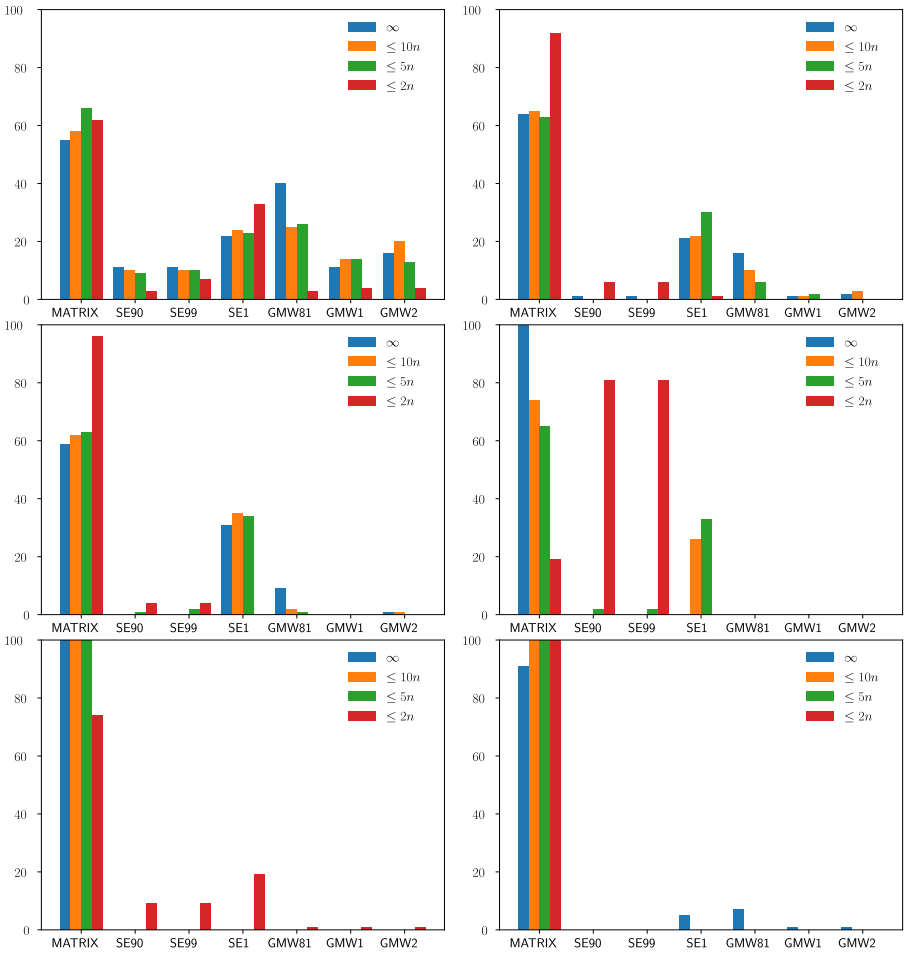


Fig. 1: Frequency, how often the algorithm achieved the smallest approximation error for the four different bounds on the condition number of the approximation (different colors) and for each of the six test scenarios (different plots).

The last three scenarios are randomly generated symmetric matrices with eigenvalues uniformly distributed in $[-10^4, 10^4]$, $[-10^4, 1]$ or $[-1, 10^4]$, depending on the scenario, which should be approximated by symmetric positive semidefinite matrices. Each of these random symmetric matrices has been generated by multiplying a random orthogonal matrix, generated with the algorithm described in [56], with a diagonal matrix with the chosen eigenvalues as diagonal values and then multiplying this with the transposed random orthogonal matrix. The eigenvalues have been drawn from uniform distributions and were altered so that each matrix has at least one negative and one positive eigenvalue.

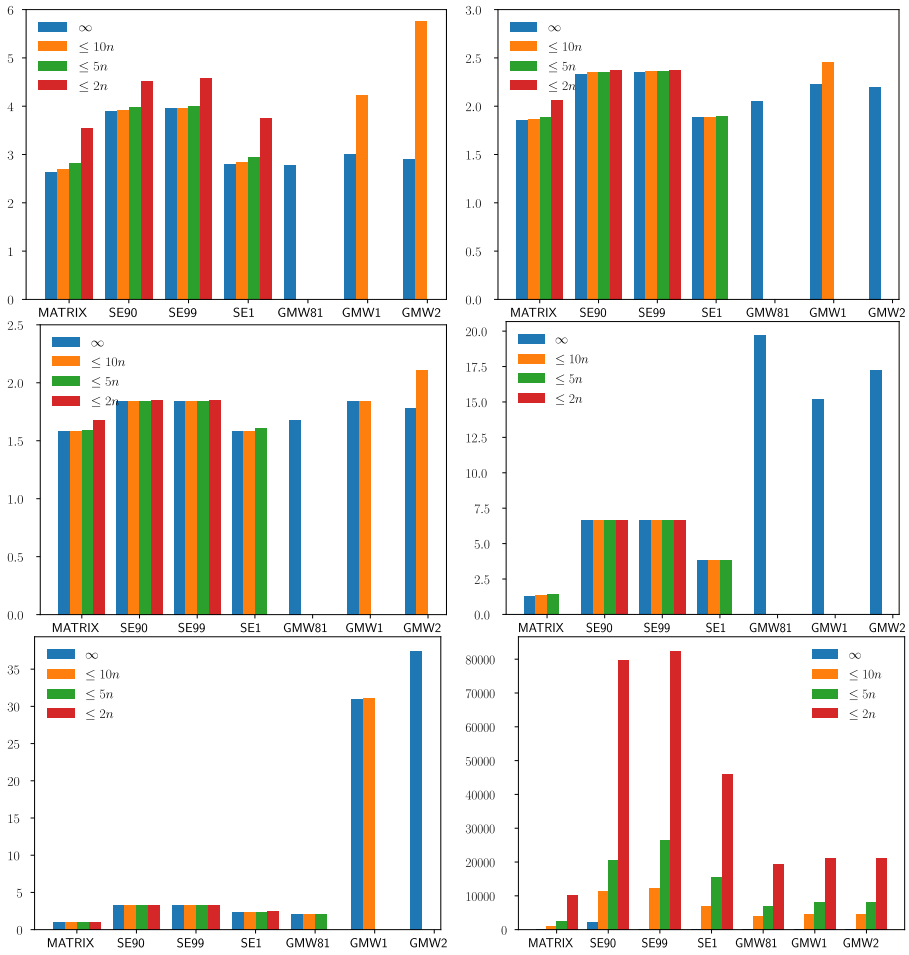


Fig. 2: Median of the approximation errors for the four different bounds on the condition number of the approximation (different colors) and for each of the six test scenarios (different plots).

For each of the six scenarios, 100 matrices have been generated with dimensions evenly distributed between 10, 20, 30, 40 and 50 and each of them was approximated.

The approximations are assessed according to the approximation error and their condition number using different objectives. The first one is to minimize the approximation error without caring about the condition number. The other three are to minimize the approximation error while getting a condition number lower or equal to $10n$, $5n$ and $2n$, respectively, where n is the dimension of the matrix. This corresponds to the requirement, that the condition number should be sufficiently small (but must not be minimal), which often occurs in application examples. Minimizing only the condition number without taking the approximation error into account is not useful.

Each of the algorithms has a parameter, representing a lower bound on the values of D, allowing to favor a low approximation error or a low condition number. Hence, each algorithm has been executed several times with different values for this parameter and for each of the four objectives only the approximation which best meets the objective was taken into account.

Figure 1 shows how many times each algorithm has computed the approximation with the smallest approximation error among all tested algorithms for the six scenarios and the four objectives. The MATRIX algorithm clearly outperforms all other tested algorithms in all scenarios.

Figure 2 shows the median of the approximation errors for each of the six scenarios and the four objectives. The approximation errors are relative to the minimal possible approximation errors which have been calculated using the methods described in [24] and [27]. No bar in Figure 2 indicates that the algorithm was not able to calculate an approximation with the restriction to the condition number for at least half of the test matrices.

The results show that MATRIX calculates approximations with approximation errors usually close to optimal and still sufficiently small condition numbers. In addition, it performs better, sometimes very considerably, than the other tested algorithms.

The numerical tests have also indicated that, for determining d_i, a varying lower bound $\hat{l}_i$, defined as

$$\hat{l}_i := \begin{cases} l & \text{if } \frac{1}{2}c_{p_i} < l \\ u & \text{if } \frac{1}{2}c_{p_i} > u \\ \frac{1}{2}c_{p_i} & \text{else} \end{cases} \quad \text{with } c_i := \begin{cases} x_i & \text{if } A_{ii} < x_i \\ y_i & \text{if } A_{ii} > y_i \\ A_{ii} & \text{else} \end{cases}$$

for each $i \in \{1 \ldots, n\}$, is useful in order to achieve a low approximation error and a low condition number. This varying lower bound is also choosable in the matrix-decomposition library.

4 Conclusions

A new algorithm to approximate Hermitian matrices by positive semidefinite Hermitian matrices was presented. In contrast to existing algorithms, it allows to specify bounds on the diagonal values of the approximation.

It tries to minimize the approximation error in the Frobenius norm and the condition number of the approximation. Parameters of the algorithm can be used to select which of these two objectives is preferred if not both objectives can be meet equally well. Numerical tests have shown that the algorithms outperforms existing algorithms regarding the approximation error as well as the condition number.

The algorithm is suitable for very large matrices, since it needs only $\frac{1}{3}n^3 + \mathcal{O}(n^2)$ basic operations and storage for $n^2 + \mathcal{O}(n)$ numbers in the real valued case. This is asymptotically the same number of basic operations as the computation of a Cholesky decomposition would need. Moreover the algorithm is

also suitable for sparse matrices since it preserves the sparsity pattern of the original matrix.

The LDL^H decomposition of the approximation is calculated as a byproduct. This allows to solve corresponding linear equations or to calculate the corresponding determinant very quickly. If such a decomposition should be calculated anyway, the algorithm has no significant overhead.

Two parts in the algorithm are realizable in many different ways. Various possibilities were presented, more are conceivable.

An open-source implementation of this algorithm is freely available. The implementation is fully documented and easy to install. Extensive numerical tests confirm the functionality of the algorithm and its implementation.

Numerical optimization and statistics are two fields of application in which the algorithm can be of particular interest.

A Appendix

Theorem 7 *Let $A \in \mathbb{C}^{n\times n}$ be a positive semidefinite matrix. A has a LDL^H decomposition. If A is positive definite this decomposition is unique.*

Proof: See [29, p. 13].

□

Theorem 8 *Let $L \in \mathbb{C}^{n\times n}$ be a lower triangular matrix with ones on the diagonal and $D \in \mathbb{R}^{n\times n}$ a diagonal matrix. LDL^H is*

a) invertible if and only if $D_{ii} \neq 0$ for all $i \in \{1,\dots,n\}$.
b) positive semidefinite if and only if $D_{ii} \geq 0$ for all $i \in \{1,\dots,n\}$
c) positive definite if and only if $D_{ii} > 0$ for all $i \in \{1,\dots,n\}$.

Proof: Sylvester's law of inertia [57] extended to Hermitian matrices [31] implies that LDL^H and D have the same number of negative, zero, and respectively positive eigenvalues. Since D is a diagonal matrix, the eigenvalues of D are its diagonal values. Hence LDL^H is invertible, positive semidefinite or positive definite if and only if the diagonal values of D are non-zero, non-negative or positive, respectively.

□

Theorem 9 *Let $A \in \mathbb{C}^{n\times n}$ be a positive definite matrix. Let L and D be the matrices of its LDL^H decomposition. Then*

$$\left(\frac{\operatorname{trace}(A)}{n\beta}\right)^{\frac{n}{2(n-1)}} \leq \kappa_2(L) \leq 2\left(\frac{\operatorname{trace}(A)}{n\alpha}\right)^{\frac{n}{2}},$$

$$\kappa_2(D) = \frac{\beta}{\alpha} \text{ and } \kappa_2(A) \leq 4\frac{\beta}{\alpha}\left(\frac{\operatorname{trace}(A)}{n\alpha}\right)^n$$

$$\text{with } \alpha := \min_{i=1,\dots,n} D_{ii} \text{ and } \beta := \max_{i=1,\dots,n} D_{ii}.$$

Proof: Define $B := LL^H$. The definition of B implies

$$\kappa_2(L) = \sqrt{\kappa_2(B)}$$

since $\kappa_2(B) = \kappa_2(LL^H) = \kappa_2(L)^2$.

L is a lower triangular matrix with ones on the diagonal. Hence, $\det(L) = 1$ and

$$\det(B) = \det(L)\det(L^H) = \det(L)^2 = 1.$$

Thus, [11] state that

$$c^{-\frac{1}{n-1}} \leq \kappa_2(B) \leq \frac{1+\sqrt{1-c}}{1-\sqrt{1-c}} \quad \text{with} \quad c := \left(\frac{n}{\operatorname{trace}(B)}\right)^n. \tag{14}$$

Besides,

$$\operatorname{trace}(B) = \operatorname{trace}(LL^H) = \sum_{i,j=1}^{n} L_{ij}\overline{L_{ij}} = \sum_{i,j=1}^{n} |L_{ij}|^2 = \|L\|_F^2. \tag{15}$$

and

$$\|L\|_F^2 = \sum_{i,j=1}^{n} |L_{ij}|^2 \geq \sum_{i=1}^{n} |L_{ii}|^2 = n.$$

Hence $0 \leq c \leq 1$, which implies

$$\frac{1+\sqrt{1-c}}{1-\sqrt{1-c}} = \frac{(1+\sqrt{1-c})^2}{(1-\sqrt{1-c})(1+\sqrt{1-c})} = \frac{(1+\sqrt{1-c})^2}{c} \leq \frac{2^2}{c}. \tag{16}$$

Equation (14), (15) and (16) result in

$$\left(\frac{\|L\|_F^2}{n}\right)^{\frac{n}{n-1}} \leq \kappa_2(B) \leq 4\left(\frac{\|L\|_F^2}{n}\right)^n$$

and thus

$$\left(\frac{\|L\|_F^2}{n}\right)^{\frac{n}{2(n-1)}} \leq \kappa_2(L) \leq 2\left(\frac{\|L\|_F^2}{n}\right)^{\frac{n}{2}}. \tag{17}$$

Theorem 8 implies $0 < \alpha$ because A is positive definite. Moreover, $\alpha \leq D_{ii} \leq \beta$ for all $i \in \{1, \ldots, n\}$ by definition of α and β. Thus

$$\begin{aligned}
\frac{\operatorname{trace}(A)}{\beta} &= \frac{1}{\beta}\sum_{i=1}^{n} A_{ii} = \frac{1}{\beta}\sum_{i=1}^{n}\sum_{j=1}^{n} L_{ij}D_{jj}\overline{L}_{ij} = \sum_{i=1}^{n}\sum_{j=1}^{n} |L_{ij}|^2 \frac{D_{jj}}{\beta} \\
&\leq \sum_{i=1}^{n}\sum_{j=1}^{n} |L_{ij}|^2 = \|L\|_F^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} |L_{ij}|^2 \\
&\leq \sum_{i=1}^{n}\sum_{j=1}^{n} |L_{ij}|^2 \frac{D_{jj}}{\alpha} = \frac{1}{\alpha}\sum_{i=1}^{n}\sum_{j=1}^{n} L_{ij}D_{jj}\overline{L}_{ij} = \frac{1}{\alpha}\sum_{i=1}^{n} A_{ii} = \frac{\operatorname{trace}(A)}{\alpha}.
\end{aligned}$$

Hence

$$\left(\frac{\operatorname{trace}(A)}{n\beta}\right)^{\frac{n}{2(n-1)}} \leq \kappa_2(L) \leq 2\left(\frac{\operatorname{trace}(A)}{n\alpha}\right)^{\frac{n}{2}}$$

with (17).

Furthermore

$$\kappa_2(D) = \frac{\max\limits_{i=1,\ldots,n} |D_{ii}|}{\min\limits_{i=1,\ldots,n} |D_{ii}|} = \frac{\beta}{\alpha}$$

since D is a diagonal matrix. Thus

$$\begin{aligned}
\kappa_2(A) = \kappa_2(LDL^H) &\leq \kappa_2(L)\kappa_2(D)\kappa_2(L^H) \\
&= \kappa_2(L)^2\kappa_2(D) \leq 4\frac{\beta}{\alpha}\left(\frac{\|L\|_F^2}{n}\right)^n,
\end{aligned}$$

because $\kappa_2(AB) \leq \kappa_2(A)\kappa_2(B)$ and $\kappa_2(A) = \kappa_2(A^H)$ for every invertible matrices $A, B \in \mathbb{C}^{n\times n}$.

□

References

1. Amestoy, P.R., Davis, T.A., Duff, I.S.: An Approximate Minimum Degree Ordering Algorithm. SIAM Journal on Matrix Analysis and Applications **17**(4), 886–905 (1996). DOI 10.1137/S0895479894278952
2. Anaconda, I.: Conda Package Manager. URL `https://conda.io/docs/index.html`
3. Borsdorf, R., Higham, N.J.: A Preconditioned Newton Algorithm for the Nearest Correlation Matrix. IMA J. Numer. Anal. **30**(1), 94–107 (2010). DOI 10.1093/imanum/drn085
4. Borsdorf, R., Higham, N.J., Raydan, M.: Computing a Nearest Correlation Matrix with Factor Structure. SIAM J. Matrix Anal. Appl. **31**(5), 2603–2622 (2010). DOI 10.1137/090776718
5. Brandl, G., et al.: Sphinx: Python documentation generator (2019). URL `www.sphinx-doc.org`. Version 2.0.1
6. Chen, Y., Wiesel, A., Eldar, Y.C., Hero, A.O.: Shrinkage Algorithms for MMSE Covariance Estimation. IEEE Transactions on Signal Processing **58**(10), 5016–5029 (2010). DOI 10.1109/TSP.2010.2053029
7. Cheng, S., Higham, N.: A Modified Cholesky Algorithm Based on a Symmetric Indefinite Factorization. SIAM Journal on Matrix Analysis and Applications **19**(4), 1097–1110 (1998). DOI 10.1137/S0895479896302898
8. Chong, E., Zak, S.: An Introduction to Optimization, 4th edn. Wiley Series in Discrete Mathematics and Optimization. Wiley (2013)
9. Cuthill, E., McKee, J.: Reducing the Bandwidth of Sparse Symmetric Matrices. In: Proceedings of the 1969 24th National Conference, ACM '69, pp. 157–172. ACM, New York, NY, USA (1969). DOI 10.1145/800195.805928
10. Davies, P.I., Higham, N.J.: Numerically Stable Generation of Correlation Matrices and Their Factors. BIT Numerical Mathematics **40**(4), 640–651 (2000). DOI 10.1023/A:1022384216930. URL `https://doi.org/10.1023/A:1022384216930`
11. Davis, P.J., Haynsworth, E.V., Marcus, M.: Bound for the P-condition number of matrices with positive roots. J. Res. Natl. Bur. Stand. B **65**, 13–14 (1961)
12. Davis, T.: Direct Methods for Sparse Linear Systems. Society for Industrial and Applied Mathematics (2006). DOI 10.1137/1.9780898718881
13. Davis, T.A., Rajamanickam, S., Sid-Lakhdar, W.M.: A survey of direct methods for sparse linear systems. Acta Numerica **25**, 383–566 (2016). DOI 10.1017/S0962492916000076
14. Devlin, S.J., Gnanadesikan, R., Kettenring, J.R.: Robust estimation and outlier detection with correlation coefficients. Biometrika **62**(3), 531–545 (1975)
15. Fang, H.r., O'Leary, D.P.: Modified Cholesky algorithms: a catalog with new approaches. Mathematical Programming **115**(2), 319–349 (2008). DOI 10.1007/s10107-007-0177-6
16. Fisher, T.J., Sun, X.: Improved Stein-type shrinkage estimators for the high-dimensional multivariate normal covariance matrix. Computational Statistics & Data Analysis **55**(5), 1909–1918 (2011). DOI 10.1016/j.csda.2010.12.006. URL `http://www.sciencedirect.com/science/article/pii/S0167947310004743`
17. George, A., Liu, J.W.: Computer Solution of Large Sparse Positive Definite. Prentice Hall Professional Technical Reference (1981)
18. George, A., Liu, J.W.: The Evolution of the Minimum Degree Ordering Algorithm. SIAM Review **31**(1), 1–19 (1989). DOI 10.1137/1031001
19. Gill, P.E., Murray, W.: Newton-type methods for unconstrained and linearly constrained optimization. Mathematical Programming **7**(1), 311–350 (1974). DOI 10.1007/BF01585529
20. Gill, P.E., Murray, W., Wright, M.H.: Practical optimization. Academic press (1981)
21. Goldfeld, S.M., Quandt, R.E., Trotter, H.F.: Maximization by Quadratic Hill-Climbing. Econometrica **34**(3), 541–551 (1966)
22. Golub, G.H., Van Loan, C.F.: Matrix Computations, third edn. The Johns Hopkins University Press, Baltimore, MD, USA (1996)
23. Higham, N., Strabi, N., ego, V.: Restoring Definiteness via Shrinking, with an Application to Correlation Matrices with a Fixed Block. SIAM Review **58**(2), 245–263 (2016). DOI 10.1137/140996112. URL `https://doi.org/10.1137/140996112`

24. Higham, N.J.: Computing a nearest symmetric positive semidefinite matrix. Linear Algebra and its Applications **103**, 103–118 (1988). DOI 10.1016/0024-3795(88)90223-6. URL http://www.sciencedirect.com/science/article/pii/0024379588902236
25. Higham, N.J.: Matrix Nearness Problems and Applications. In: M.J.C. Gover, S. Barnett (eds.) Applications of Matrix Theory, pp. 1–27. Oxford University Press (1989)
26. Higham, N.J.: Accuracy and Stability of Numerical Algorithms, 2nd edn. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA (2002). DOI 10.1137/1.9780898718027
27. Higham, N.J.: Computing the Nearest Correlation Matrix - A Problem from Finance. IMA J. Numer. Anal. **22**(3), 329–343 (2002). DOI 10.1093/imanum/22.3.329
28. Higham, N.J., Strabić, N.: Anderson Acceleration of the Alternating Projections Method for Computing the Nearest Correlation Matrix. Numerical Algorithms **72**(4), 1021–1042 (2016). DOI 10.1007/s11075-015-0078-3. URL https://doi.org/10.1007/s11075-015-0078-3
29. Householder, A.S.: The theory of matrices in numerical analysis. Blaisdell Publishing Company (1964)
30. Ikeda, Y., Kubokawa, T., Srivastava, M.S.: Comparison of linear shrinkage estimators of a large covariance matrix in normal and non-normal distributions. Computational Statistics & Data Analysis **95**, 95–108 (2016). DOI 10.1016/j.csda.2015.09.011. URL http://www.sciencedirect.com/science/article/pii/S0167947315002388
31. Ikramov, K.D.: On the inertia law for normal matrices. In: Doklady Mathematics C/C of Doklady Akademii Nauk, vol. 64, pp. 141–142 (2001)
32. Iman, R., Davenport, J.: An iterative algorithm to produce a positive definite correlation matrix from an approximate correlation matrix (with a program user's guide). Tech. rep., Sandia National Labs., Albuquerque, NM (USA) (1982). DOI 10.2172/5152227. URL https://doi.org/10.2172/5152227
33. Jones, E., Oliphant, T., Peterson, P., et al.: SciPy: library for scientific computing with Python (2019). URL http://www.scipy.org. Version 1.3
34. Krekel, H., et al.: pytest: helps you write better programs (2019). URL https://docs.pytest.or. Version 4.4.1
35. Ledoit, O., Wolf, M.: Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. Journal of Empirical Finance **10**(5), 603–621 (2003). DOI 10.1016/S0927-5398(03)00007-0. URL http://www.sciencedirect.com/science/article/pii/S0927539803000070
36. Ledoit, O., Wolf, M.: A well-conditioned estimator for large-dimensional covariance matrices. Journal of Multivariate Analysis **88**(2), 365–411 (2004). DOI 10.1016/S0047-259X(03)00096-4. URL http://www.sciencedirect.com/science/article/pii/S0047259X03000964
37. Levenberg, K.: A method for the solution of certain problems in least squares. Quarterly of applied mathematics **2**(2), 164–168 (1944)
38. Marquardt, D.W.: An Algorithm for Least-Squares Estimation of Nonlinear Parameters. Journal of the Society for Industrial and Applied Mathematics **11**(2), 431–441 (1963). URL http://www.jstor.org/stable/2098941
39. Moré, J.J., Sorensen, D.C.: On the use of directions of negative curvature in a modified newton method. Mathematical Programming **16**(1), 1–20 (1979). DOI 10.1007/BF01582091. URL https://doi.org/10.1007/BF01582091
40. Nocedal, J., Wright, S.: Numerical Optimization, second edn. Springer series in operations research and financial engineering. Springer, New York (2006)
41. Oliphant, T.E., et al.: NumPy: N-dimensional array package for Python (2019). URL http://www.numpy.org. Version 1.17
42. PyPA: The Python Package Installer. URL https://pip.pypa.io
43. Python Software Foundation: Python (2018). URL http://www.python.org. Version 3.7
44. Qi, H., Sun, D.: Correlation stress testing for value-at-risk: anunconstrained convex optimization approach. Computational Optimization and Applications **45**(2), 427–462 (2010). DOI 10.1007/s10589-008-9231-4. URL https://doi.org/10.1007/s10589-008-9231-4

45. Reimer, J.: Conda package for matrix-decomposition: a library for decompose (factorize) dense and sparse matrices in Python. URL https://anaconda.org/jore/matrix-decomposition
46. Reimer, J.: GitHub repository for matrix-decomposition: a library for decompose (factorize) dense and sparse matrices in Python. https://github.com/jor-/matrix-decomposition. URL https://github.com/jor-/matrix-decomposition
47. Reimer, J.: Python package for matrix-decomposition: a library for decompose (factorize) dense and sparse matrices in Python. URL https://pypi.org/project/matrix-decomposition/
48. Reimer, J.: matrix-decomposition: a library for decompose (factorize) dense and sparse matrices in Python (2019). DOI 10.5281/zenodo.3558540. URL https://doi.org/10.5281/zenodo.3558540. Version 1.2
49. Reimer, J., Grigorievskiy, A., Lee, A., Yuri, Barrett, L., Seljebotn, D.S., Smith, N., Cournapeau, D.: scikit-sparse: a library for sparse matrix manipulation in Python (2018). URL https://github.com/scikit-sparse/scikit-sparse. Version 0.4.4
50. Rousseeuw, P.J., Molenberghs, G.: Transformation of non positive semidefinite correlation matrices. Communications in Statistics–Theory and Methods **22**(4), 965–984 (1993)
51. Rudin, W.: Principles of Mathematical Analysis. International series in pure and applied mathematics. McGraw-Hill (1976)
52. Schäfer, J., Strimmer, K.: A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics. Statistical Applications in Genetics and Molecular Biology **4**(1), 1–32 (2005)
53. Schnabel, R., Eskow, E.: A New Modified Cholesky Factorization. SIAM Journal on Scientific and Statistical Computing **11**(6), 1136–1158 (1990). DOI 10.1137/0911064
54. Schnabel, R., Eskow, E.: A Revised Modified Cholesky Factorization Algorithm. SIAM Journal on Optimization **9**(4), 1135–1148 (1999). DOI 10.1137/S105262349833266X
55. Stein, C.: Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution. In: Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics, pp. 197–206. University of California Press, Berkeley, Calif. (1956)
56. Stewart, G.: The Efficient Generation of Random Orthogonal Matrices with an Application to Condition Estimators. SIAM Journal on Numerical Analysis **17**(3), 403–409 (1980). DOI 10.1137/0717034. URL https://doi.org/10.1137/0717034
57. Sylvester, J.J.: A demonstration of the theorem that every homogeneous quadratic polynomial is reducible by real orthogonal substitutions to the form of a sum of positive and negative squares. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science **4**(23), 138–142 (1852)
58. Touloumis, A.: Nonparametric Stein-type shrinkage covariance matrix estimators in high-dimensional settings. Computational Statistics & Data Analysis **83**, 251–261 (2015). DOI 10.1016/j.csda.2014.10.018. URL http://www.sciencedirect.com/science/article/pii/S0167947314003107
59. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, I., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P., SciPy: SciPy 1.0-Fundamental Algorithms for Scientific Computing in Python. CoRR **abs/1907.10121** (2019). URL http://arxiv.org/abs/1907.10121
60. Yannakakis, M.: Computing the Minimum Fill-In is NP-Complete. SIAM Journal on Algebraic Discrete Methods **2**(1), 77–79 (1981). DOI 10.1137/0602010

Statistical Analysis of the Phosphate Data of the World Ocean Database 2013

Joscha Reimer[1]

[1]Kiel University, Department of Computer Science, Algorithmic Optimal Control - CO_2 Uptake of the Ocean, 24098 Kiel, Germany

Correspondence: Joscha Reimer (joscha.reimer.edu@web.de)

Abstract. The phosphate data of the World Ocean Database 2013 are extensively statistically analyzed by splitting the measurement results into a long scale, i.e., climatological, and a short scale part. Means, medians, absolute and relative standard deviations, interquartile ranges, quartile coefficients of dispersion, correlations and covariances are estimated and analyzed. The underlying probability distributions are investigated using visual inspection as well as statistical tests. All presented methods are applicable to other data as long as they satisfy the postulated assumptions.

1 Introduction

Phosphate is a limiting nutrient for phytoplankton and therefore of central importance for understanding marine ecosystems (cf. (Bigg, 2003, 4)). In the context of marine biogeochemical and ecosystem modeling, phosphate data play a dominant role as they are available in a high number and by their good spatial and temporal coverage (cf. Kriest et al. (2010)).

A basic source for freely available and quality-controlled oceanic measurement data is the World Ocean Database, a project established by the Intergovernmental Oceanographic Commission of UNESCO. The most recent release is the World Ocean Database 2013 which is described in Boyer et al. (2013) and Johnson et al. (2013). Therein, millions of measurement data for phosphate concentrations in the ocean are available.

The World Ocean Atlas, a project by the National Oceanographic Data Center in the U.S., provides analyzed and aggregated data from the World Ocean Database. The most recent release of their analysis of phosphate data is introduced in Garcia et al. (2014) as part of the World Ocean Atlas 2013 version 2. Means, standard deviations and standard errors of the means are provided there on one and five degree spatial grids at selected depth levels, down to 500 m as monthly and seasonal averages and down to 5500 m as annual averages. The provided data are climatological data (cf. (Storch and Zwiers, 1999, 1.2.1)), i.e., data from different years have been used to calculate these properties of an average year. This has been done only for (space-time) grid boxes with enough data available. For all other grid boxes, the means (and only these) have been interpolated.

In this paper, we extended the analysis and aggregation provided in the World Ocean Atlas in several respects. The most important one is the split of the measurement results into a long scale part, i.e., climatological part, and a short scale part, i.e., noise part from a climatological perspective. This splitting allows to separately analyze and quantify the noise part in

measurements and a noise free climatological part. Moreover, the noise part is useful in assessing the spatial and temporal resolution used in the statistical analysis.

For each grid box, the climatological part is primarily described by its climatological mean, i.e., the average concentration in an average year, and the climatological variability, i.e., the usual deviation between the average concentration in an actual year and the climatological mean. The noise part is primarily described for each grid box by the usual deviation between the average concentration in an actual year and the result of its measurement, and thus includes the variation of the concentration inside a grid box as well as measurement errors.

Compared to the World Ocean Atlas, we used a slightly different spatial and temporal resolution. We took a monthly temporal resolution within each year in all depth layers down to the sea floor, allowing us to capture time-dependent changes in deeper layers, which can be especially important for fitting time-dependent models to these data. To counteract the sparseness of the measurement data, especially in deeper layers, the vertical resolution was decreased. Details are given in Appendix A.

Additionally to monthly means, we provide medians, absolute and relative standard deviations as well as interquartile ranges and quartile coefficients of dispersion, all of these for the entire ocean. Finally, we quantified the statistical dependencies using correlations and covariances and investigated the underlying probability distributions using visual inspection as well as statistical tests.

We feel confident that the results of this detailed statistical analysis may improve the understanding of the marine phosphate concentration. They may also be used to increase the accuracy of model fitting procedures (cf. (Walter and Pronzato, 1997, 4) or (Seber and Wild, 2003, 2.1.4)) of marine biogeochemical models, and the information gain through new measurements (cf. (Walter and Pronzato, 1997, 6) or (Pronzato and Pázman, 2013, 5)). Our approach is also applicable to other marine concentrations satisfying the assumptions made.

Our method for the statistical analysis is presented in Section 2, the results obtained for the phosphate concentration are presented in Section 3 and in Section 4, we draw our conclusion.

2 Methods used in the Statistical Analysis

To lay the foundation of the statistical analysis the statistical model is introduced and statistical assumptions postulated. Methods to estimate the associated expected values and to quantify the associated variabilities are presented afterwards. Thereafter, methods to quantify the statistical dependencies including covariances and correlations are introduced. The section closes with an investigation of the underlying probability distributions.

2.1 Statistical Model

We conducted our statistical analysis on a space-time grid described in detail in Appendix A and define:

- $\mathcal{X}_s$ as the set of all spatial grid boxes, identified e.g. by their center,
- $\mathcal{X}_t$ as the set of all time intervals within one year (which are the same for all years),

– $\mathcal{X}_a$ as the set of years,

for which a statistical analysis should be carried out. We then identify:

– $\mathcal{X} := \mathcal{X}_s \times \mathcal{X}_t \times \mathcal{X}_a \times \mathbb{N}$ as the set of all possible measurement point ,

where each measurement point $(s,t,a,n) \in \mathcal{X}$ in the grid box (s,t,a) has an unique index n and define:

– $X \subseteq \mathcal{X}$ as the set of all points where measurement data are actually available and

– $y(s,t,a,n)$ as the result of the measurement at $(s,t,a,n) \in X$.

For the analysis presented below we use the sets:

– $A(s,t) := \{a \mid (s,t,a,n) \in X\}$ with all years where measurements are available at $(s,t) \in \mathcal{X}_s \times \mathcal{X}_t$,

– $N(s,t,a) := \{n \mid (s,t,a,n) \in X\}$ with all indices of available measurements at $(s,t,a) \in \mathcal{X}_s \times \mathcal{X}_t \times \mathcal{X}_a$.

Random Fields and Statistical Assumptions

In the statistical analysis, we use the following main variables, which are considered as random fields:

– the measurement results η, defined on $\mathcal{X}$,

– the true concentration δ, without noise, averaged in each grid box, defined on $\mathcal{X}_s \times \mathcal{X}_t \times \mathcal{X}_a$,

– the noise ϵ, defined on $\mathcal{X}$, including the variability due to the discretization introduced by the space-time grid as well as by imperfect measuring instruments and methods.

Let $(s,t,a,n), (\hat{s},\hat{t},\hat{a},\hat{n}) \in \mathcal{X}$ represent arbitrary measurement points for the rest of this section. It is assumed that the noise is additive and unbiased:

$$\eta(s,t,a,n) = \delta(s,t,a) + \epsilon(s,t,a,n), \tag{1}$$

$$\mathbb{E}(\epsilon(s,t,a,n)) = 0. \tag{2}$$

Moreover, there is no interaction assumed between the true concentration and the noise as well as the noise at different points:

$$\delta(s,t,a) \text{ and } \epsilon(\hat{s},\hat{t},\hat{a},\hat{n}) \text{ are independent,} \tag{3}$$

$$\epsilon(s,t,a,n) \text{ and } \epsilon(\hat{s},\hat{t},\hat{a},\hat{n}) \text{ are independent if } (s,t,a,n) \neq (\hat{s},\hat{t},\hat{a},\hat{n}). \tag{4}$$

The following additional assumptions were made due to the sparseness of the available data: The noise is assumed to have equal distributions within the same grid box:

$$\epsilon(s,t,a,n) \stackrel{d}{=} \epsilon(s,t,a,\hat{n}). \tag{5}$$

The true concentration is assumed to have the same distribution at two points where only the year differs:

$$\delta(s,t,a) \stackrel{d}{=} \delta(s,t,\hat{a}). \tag{6}$$

The covariance of the true concentration is assumed to be invariant with respect to annual shifts:

$$\operatorname{cov}(\delta(s,t,a),\delta(\hat{s},\hat{t},\hat{a})) = \operatorname{cov}(\delta(s,t,a+z),\delta(\hat{s},\hat{t},\hat{a}+z)) \text{ with } a+z,\hat{a}+z \in \mathcal{X}_a. \tag{7}$$

The annual periodicity of the climatological process that is described is a justification for the last two assumptions.

2.2 Climatological Means

The measurement results η and the true concentration δ have the same expected values (cf. (Storch and Zwiers, 1999, 2.6.5)) due to the additivity (1) and unbiasedness (2) of the noise:

$$\mathbb{E}(\eta(s,t,a,n)) = \mathbb{E}(\delta(s,t,a)) + \mathbb{E}(\epsilon(s,t,a,n)) = \mathbb{E}(\delta(s,t,a)).$$

Due to the assumed annual periodicity (6) of the true concentration, its expected values do not depend on the year and thus can be defined as $\mu : \mathcal{X}_s \times \mathcal{X}_t \to \mathbb{R}$ with:

$$\mu(s,t) := \mathbb{E}(\delta(s,t,a)) = \mathbb{E}(\eta(s,t,a,n))$$

which corresponds to the climatological mean concentration in the grid box (s,t).

The climatological mean $\mu(s,t)$ could be estimated using the average of all measurement results available in the grid box (s,t). However, this could result in a very inaccurate estimate if the number of available measurements varies from year to year. An estimate would then tend to the average true concentration in years with the most measurements available which can be significantly differ from the climatological mean.

As a remedy, we first estimated the average true concentration in the grid box without noise for each year using the average of the measurement results within the same year:

$$c(s,t,a) := \frac{1}{|N(s,t,a)|} \sum_{n \in N(s,t,a)} y(s,t,a,n) \qquad \text{if } |N(s,t,a)| \geq 1.$$

The climatological mean in a grid box was then estimated by the average, i.e., the sample mean (cf. (Storch and Zwiers, 1999, 4.3.1)), of the estimated true concentrations in the grid box for different years:

$$\mu(s,t) \approx m(s,t) := \frac{1}{|A(s,t)|} \sum_{a \in A(s,t)} c(s,t,a) \qquad \text{if } |A(s,t)| \geq 1.$$

Alternatively, the median (cf. (Storch and Zwiers, 1999, 2.6.4)) instead of the (arithmetic) mean could be used in the previously described calculations for which the true concentration would be estimated as:

$$\hat{c}(s,t,a) := \operatorname*{median}_{n \in N(s,t,a)} y(s,t,a,n) \qquad \text{if } |N(s,t,a)| \geq 1$$

and the corresponding climatological mean as:

$$\mu(s,t) \approx \hat{m}(s,t) := \underset{a \in A(s,t)}{\text{median}}\ \hat{c}(s,t,a) \qquad \text{if } |A(s,t)| \geq 1.$$

In general, the median provides a more accurate estimate in case of outliers, otherwise the mean should be preferred (cf. (Linacre, 1992, 4)). If the number of measurements in several years is low, both estimates for μ might lack accuracy. Thus, it is reasonable to choose the required number of years with measurements $|A(s,t)|$ sufficiently high. We decided to require measurements for at least two years, preventing one extraordinary year to cause a poor estimate.

At the grid boxes lacking enough measurements, values for the climatological mean were interpolated according to Appendix B. Without sufficient data to achieve a meaningful interpolation, the average of the estimates could be used instead.

2.3 Variabilities

Due to the assumed independence of δ and ϵ, see Equation (3), the variances (var) (cf. (Storch and Zwiers, 1999, 2.6.7)) of the measurement results η are given by the sum of the variances of the true concentration δ and the measurement noise ϵ:

$$\mathrm{var}(\eta(s,t,a,n)) \quad = \mathrm{var}(\delta(s,t,a)) + \mathrm{var}(\epsilon(s,t,a,n)).$$

The variances of the true concentration, describing the climatological variabilities, do not depend on the year due to the assumed annual periodicity (6). Hence, the standard deviation (sd) (cf. (Storch and Zwiers, 1999, 2.6.7)) of the true concentration at a grid box was estimated by the sample standard deviation (cf. (Storch and Zwiers, 1999, 4.3.2)) of all estimated true concentrations in this grid box and different years:

$$\mathrm{sd}(\delta(s,t,a)) \approx \sqrt{\frac{1}{|A(s,t)|-1} \sum_{\hat{a} \in A(s,t)} (c(s,t,\hat{a}) - m(s,t))^2} \qquad \text{if } |A(s,t)| \geq 2.$$

Under the assumption of equal distributions of the noises in a grid box (5), the variance of the noises, describing the short scale variability, do not depend on the number of measurements. Hence, the standard deviation of the noise for a specific grid box and a specific year was estimated by the sample standard deviation of all measurement results in this grid box and year:

$$\mathrm{sd}(\epsilon(s,t,a,n)) \approx \sqrt{\frac{1}{|N(s,t,a)|-1} \sum_{\hat{n} \in N(s,t,a)} (y(s,t,a,\hat{n}) - c(s,t,a))^2} \qquad \text{if } |N(s,t,a)| \geq 2.$$

The sample interquartile range as an approximation of the interquartile range (iqr) can be used as well to quantify the variability:

$$\mathrm{iqr}(\delta(s,t,a)) \approx \underset{\hat{a} \in A(s,t)}{\mathrm{q}_{75\%}}\ c(s,t,\hat{a}) - \underset{\hat{a} \in A(s,t)}{\mathrm{q}_{25\%}}\ c(s,t,\hat{a}) \qquad \text{if } |A(s,t)| \geq 1,$$

and

$$\mathrm{iqr}(\epsilon(s,t,a,n)) \approx \underset{\hat{n} \in N(s,t,a)}{\mathrm{q}_{75\%}}\ y(s,t,a,\hat{n}) - \underset{\hat{n} \in N(s,t,a)}{\mathrm{q}_{25\%}}\ y(s,t,a,\hat{n}) \qquad \text{if } |N(s,t,a)| \geq 1.$$

Here, $q_{25\%}$ and $q_{75\%}$ denote the first and third quartile, respectively. The value of the quartile was linearly interpolated between two available values if necessary.

When interpreting the variability of a random variable, the variability relative to the expected value is usually more helpful than just the variability. Therefore, we calculate the relative standard deviations, i.e., the standard deviations divided by the means, and the quartile coefficients of dispersion, i.e., the interquartile ranges relative to the medians. Relative variabilities of the noise ϵ are meaningless since its expected values were assumed to be zero.

In the presence of outliers, interquartile ranges and quartile coefficients of dispersion perform better, otherwise standard deviations and relative standard deviations should be preferred.

Estimates of the variabilities in all variants are more accurate if a high number of measurements can be used. We computed these estimates only where at least three values were available and interpolated otherwise as described in Appendix B. The average of the estimates could be used in absence of sufficient data for interpolation.

2.4 Statistical Dependencies

From the additivity of the noise (1), the assumptions (3), (4) and the bilinearity of the covariance, we immediately deduce:

$$\left.\begin{aligned} &\operatorname{cov}(\delta(s,t,a),\epsilon(\hat{s},\hat{t},\hat{a},\hat{n})) = \operatorname{cov}(\epsilon(s,t,a,n),\epsilon(\hat{s},\hat{t},\hat{a},\hat{n})) = \operatorname{cov}(\eta(s,t,a,n),\epsilon(\hat{s},\hat{t},\hat{a},\hat{n})) = 0, \\ &\operatorname{cov}(\eta(s,t,a,n),\eta(\hat{s},\hat{t},\hat{a},\hat{n})) = \operatorname{cov}(\eta(s,t,a,n),\delta(\hat{s},\hat{t},\hat{a},\hat{n})) = \operatorname{cov}(\delta(s,t,a),\delta(\hat{s},\hat{t},\hat{a})) \end{aligned}\right\} \text{ if } (s,t,a,n) \neq (\hat{s},\hat{t},\hat{a},\hat{n}).$$

Thus, only the covariances of the true concentration δ had to be estimated. The corresponding correlations were calculated using the estimated standard deviations.

Pointwise Covariances and Correlations

The covariances of δ were assumed to be invariant with respect to annual shifts (7) and its covariance between two specific grid boxes and years was estimated by the sample covariance (cf. (Storch and Zwiers, 1999, 5.2.7)) of all pairs of estimated true concentrations in the same grid boxes and with the same difference in the years:

$$\operatorname{cov}(\delta(s,t,a),\delta(\hat{s},\hat{t},\hat{a})) \approx \frac{1}{|B|-1} \sum_{(b,\hat{b})\in B} (c(s,t,b)-m)(c(\hat{s},\hat{t},\hat{b})-\hat{m}), \qquad \text{if } |B| \geq 2$$

$$\text{with } B := \{(a+z,\hat{a}+z) \mid a+z,\hat{a}+z \in A(s,t)\},\ m = \frac{1}{|B|} \sum_{(b,\hat{b})\in B} c(s,t,b) \text{ and } \hat{m} = \frac{1}{|B|} \sum_{(b,\hat{b})\in B} c(\hat{s},\hat{t},\hat{b}).$$

A higher $|B|$ results, as usual, in more accurate estimate. The estimate of $\operatorname{cov}(\delta(s,t,a),\delta(s,t,a))$ is equal to the estimate of $\operatorname{var}(\delta(s,t,a))$ yielding a consistent estimate.

Covariance and Correlation Matrices

The pointwise covariance estimates were processed into a covariance matrix (cf. (Storch and Zwiers, 1999, 2.8.7)). For very large dimension, the covariance matrix can not be stored as a dense matrix due to limited memory. Hence, we used a sparse

matrix and assumed that the covariance is zero where no estimate was available, to that effect that, the number of stored entries corresponds to the number of estimated pointwise covariances, which can be controlled by $|B|$, the number of concentration estimates required for a covariance estimate. If estimated pointwise covariances close to zero are not stored and are thus implicitly assumed to be zero, the number of entries to be stored can be reduced even further.

In order to obtain a consistent estimate of a covariance matrix, it is not sufficient to just combine the individual estimates into a matrix. The resulting matrix has to be positive semidefinite and usually, even a well-conditioned positive definite estimate is preferred.

Often, so called shrinking methods are applied for this purpose (cf. Ledoit and Wolf (2004); Chen et al. (2010); Schäfer and Strimmer (2005)). These tend to pull the most extreme matrix entries towards more central values, achieved by a convex combination of the covariance matrix estimate and some suitable chosen target matrix. A disadvantage of these methods is the alternation of all matrix entries (i.e., pointwise covariances). In extreme cases, one poorly estimated covariance can lead to a large impact of all other well estimated covariances.

Due to this and other disadvantages of these methods, we used the approach described in Reimer (2019) and implemented in Reimer (2019a) where single off-diagonal entries are moved closer to zero generating a well-conditioned positive definite matrix. The $LDL^\top$ decomposition of the resulting covariance matrix were calculated by this approach as a byproduct and can be used to solve corresponding linear equations quickly. Furthermore permutation methods are applied to reduce the number of entries that must be stored allowing to efficiently process even larger matrices.

If the estimated variances, i.e. the diagonal values of the covariance matrix estimated, are considered sufficiently accurate, the diagonal values can be left unchanged by the algorithm, or the correlation matrix instead of the covariance matrix can be approximated by the algorithm. We decided to approximate the correlation matrix.

To get a well-conditioned matrix, we forced the approximation algorithm to ensure that each entry in the diagonal matrix D of the $LDL^\top$ decomposition is at least 0.01. This threshold directly affects the condition number and the approximation error and can be adjusted to prioritize one of these two.

Correlations Dependencies on Distances

In many applications in natural sciences, the correlations of random fields depend solely on the distance between the associated points and this we checked for the true concentration δ.

If the estimated correlations can be described by a function depending only on the distance of the associated points, the estimated correlations for points with the same distance must be (approximately) the same. Hence, we grouped all estimated correlations for points with the same distance. For each group, we calculated the interquartile range which must be close to zero. 0.1 could be a good threshold here, or 0.05 as a more restrictive value.

Moreover, if the correlations could be described even by a continuous function, the correlations must be close to each other if the distances of their associated points are close to each other. This can be also checked by grouping and calculating interquartile ranges.

Both can also be checked to some extend graphically by plotting the grouped correlations or the calculated interquartile ranges. To ensure significance of theses checks, the number of involved correlations has to be sufficiently large.

2.5 Probability Distributions

In addition to the estimation of statistical parameters like expectation value or standard deviation, the type of the underlying probability distribution (cf. (Storch and Zwiers, 1999, 2.6.3)) is of interest too. Visual inspection and statistical tests were used to analyze from what probability distribution the data may originate (cf. (Linacre, 1992, 4)). We focused on normal distribution (cf. (Storch and Zwiers, 1999, 2.7.3)) and log-normal distribution (cf. (Storch and Zwiers, 1999, 2.7.6)).

Histogram (cf. (Storch and Zwiers, 1999, 5.2.1)) and kernel density estimation (cf. Scott (2015)) were used to give an idea of the underlying probability distribution. Box plots are an alternative which indicate the expected value, the spread, the skewness and outliers. P-P (probability-probability) and Q-Q (quantile-quantile) plots are useful to study the data with respect to a particular probability distribution.

For a particular probability distribution, Usually several statistical tests are available to verify if data originate from a particular probability distribution. Tests applied in this thesis regarding normal distributions were the Shapiro-Wilk test introduced in Shapiro and Wilk (1965), the Anderson-Darling test introduced in Anderson and Darling (1952) and the D'Agostino-Pearson test introduced in D'Agostino (1971) and D'Agostino and Pearson (1973). To check if the data originate from log-normal distributions, first the natural logarithm and afterwards the tests for normal distributions were applied to the data.

To check the probability distribution at a particular point, by visual inspection or by statistical tests, all values originating from random variables with the same probability distribution were used:

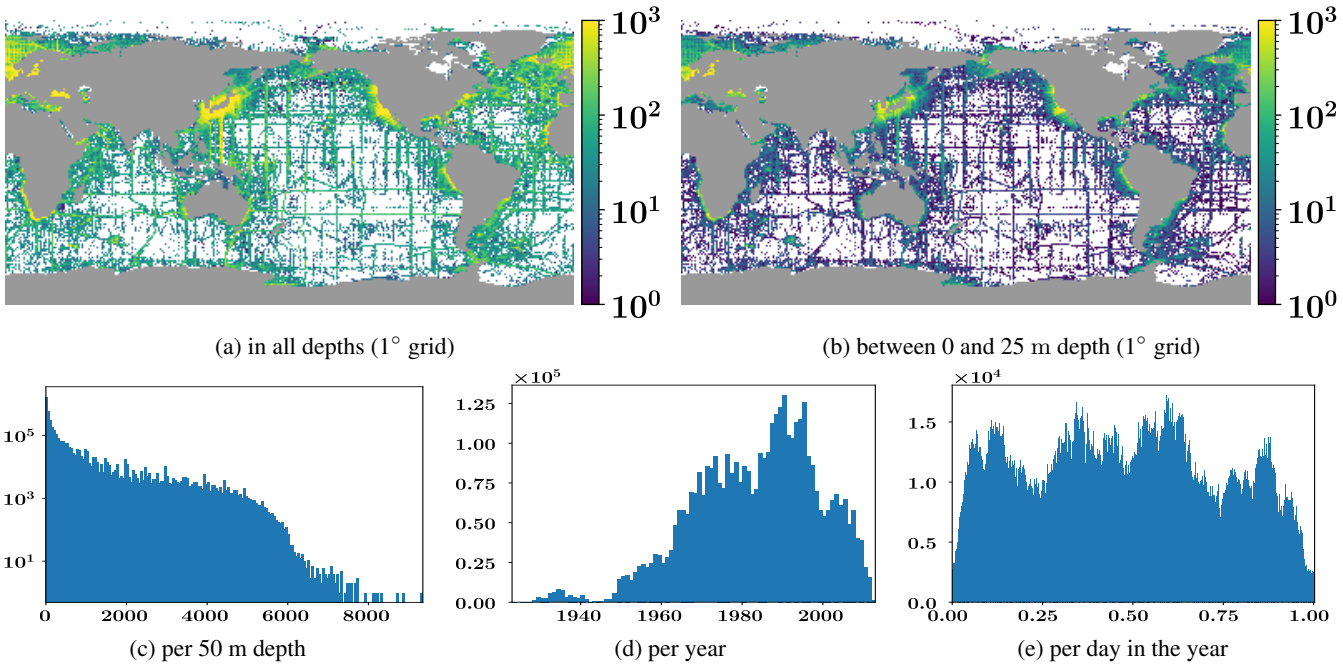
- For the measurement results η: all values $y(\hat{s},\hat{t},\hat{a},\hat{n})$ with $(\hat{s},\hat{t},\hat{a},\hat{n}) \in N(s,t,a)$.
- For the true concentration δ: all values $c(s,t,a)$ with $a \in A(s,t)$.
- For the noise ϵ: all values $\big(y(\hat{s},\hat{t},\hat{a},\hat{n}) - c(s,t,a)\big)$ with $(\hat{s},\hat{t},\hat{a},\hat{n}) \in N(s,t,a)$.

It should be noted that these statistical tests, as well as the graphical methods, cannot ensure whether data are really realizations from a normal distribution or not. They just can determine a reasonable certainty.

3 Results of the Statistical Analysis

We now present the results of our statistical analysis of the phosphate concentration data provided by the World Ocean Database 2013 presented in Boyer et al. (2013) and Johnson et al. (2013), obtained as described in Section 2, using the software mentioned in Appendix C and a one degree resolution with 33 depth layers and a monthly time resolution as described in Appendix A. They are not interpreted in a marine context.

The spatial and temporal distribution of the data is described in Subsection 3.1, the climatological means in Subsection 3.2, the long and short scale variabilities in Subsection 3.3, the covariances and correlations in Subsection 3.4, and the investigation of the probability distributions in Subsection 3.5.



(a) in all depths (1° grid)

(b) between 0 and 25 m depth (1° grid)

(c) per 50 m depth

(d) per year

(e) per day in the year

Figure 1. Number of phosphate measurements with respect to space and time.

3.1 Spatial and Temporal Distribution

All data in the World Ocean Database 2013 have been quality checked (cf. (Johnson et al., 2013, Section 3)). We used all phosphate data that passed quality control, which were more than 4.1 million measurements in total.

Regarding the spatial distribution, the number of measurements decreases with the distance to the coast (Figure 1a, 1b) and with growing depth (Figure 1c). In time, the measurements range from 1923 to 2012, with the majority between 1963 and 2009 (Figure 1d). Over most of the year, the measurements are uniformly distributed, significantly fewer measurements are only available in December and January (Figure 1e).

3.2 Climatological Means

The climatological mean concentrations, i.e. the concentrations in an average year, were estimated once using the arithmetic mean and once using the median, (compare Subsection 2.2). The results using the arithmetic mean are described next and are plotted in Figure 2.

The average estimated climatological mean is 2.17 $\mathrm{mmol\,m^{-3}}$. The time averaged climatological means near the surface are shown in Figure 2a. Here, the highest values are at the Southern Ocean ranging from 1.5 to 2.2 $\mathrm{mmol\,m^{-3}}$. Other high values are at the north of the Pacific Ocean, ranging from 1.0 to 1.5 $\mathrm{mmol\,m^{-3}}$, and at the southeast of the Pacific Ocean, around 1.0 $\mathrm{mmol\,m^{-3}}$. Elsewhere, they are usually below 0.5 $\mathrm{mmol\,m^{-3}}$.

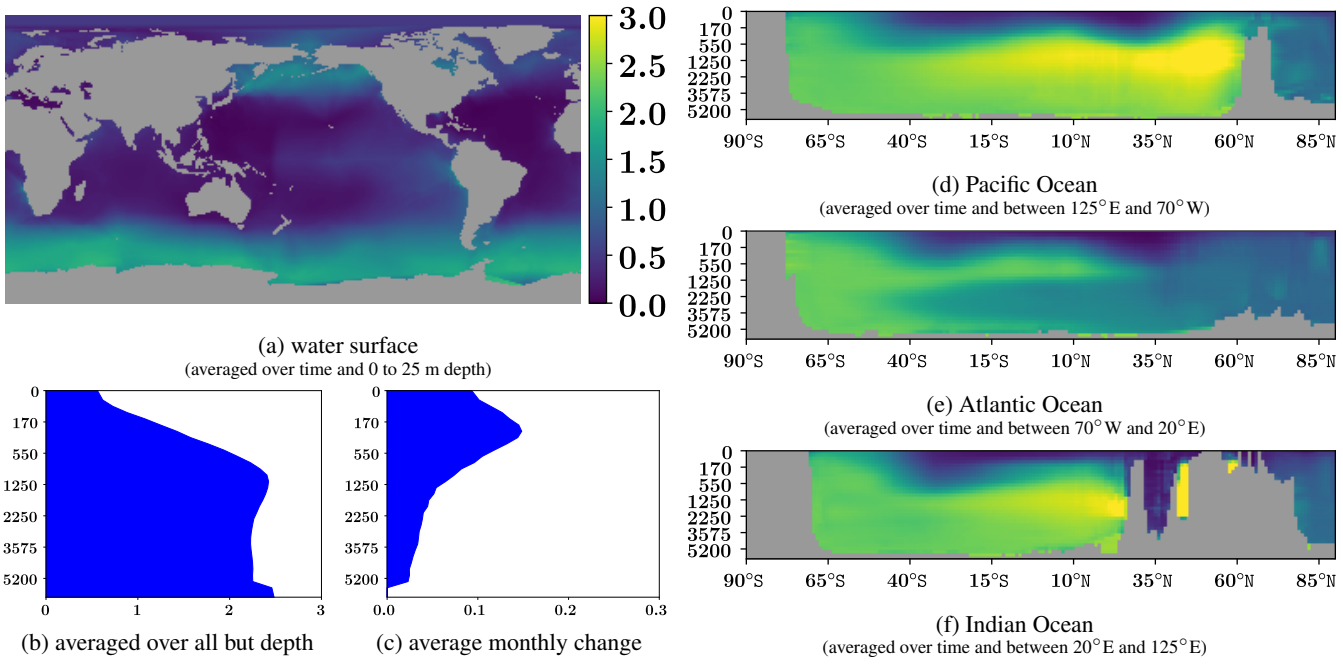


Figure 2. Climatological mean of phosphate in $\mathrm{mmol\,m^{-3}}$.

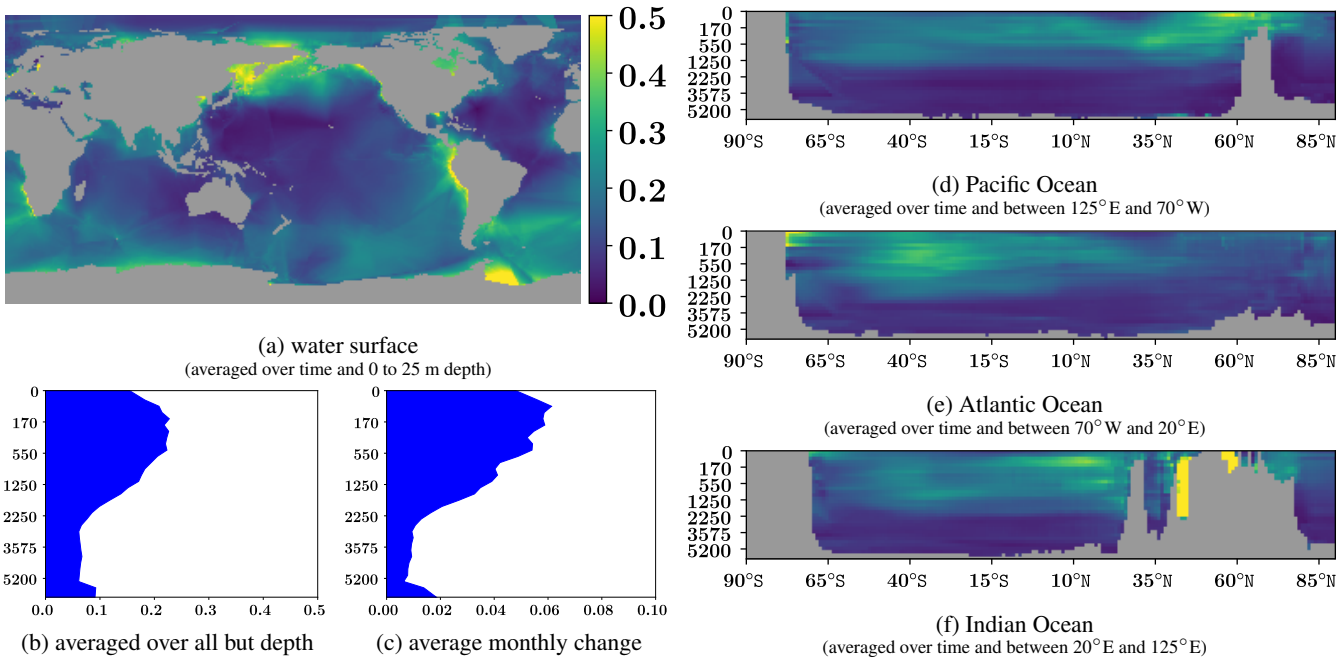
The climatological means averaged over all but the depth are presented in Figure 2b. They strictly increase from 0.6 to 2.5 $\mathrm{mmol\,m^{-3}}$ with growing depth up to 1200 m. After that they remain constant at approximately 2.25 $\mathrm{mmol\,m^{-3}}$.

The average absolute change in the climatological means after one month is about 0.05 $\mathrm{mmol\,m^{-3}}$. The average absolute monthly changes depending on the depth are plotted in Figure 2c. They strictly increase from 0.09 to 0.15 $\mathrm{mmol\,m^{-3}}$ with growing depth up to around 250 m. Afterwards they strictly decrease with growing depth. After a depth of 1500 m they are below 0.05 $\mathrm{mmol\,m^{-3}}$.

The climatological means in the Pacific Ocean depending on depth and latitude and averaged over longitude and time are shown in Figure 2d. Near the surface they are usually between 0.2 $\mathrm{mmol\,m^{-3}}$ and 1 $\mathrm{mmol\,m^{-3}}$. Only south of 50°S they are between 1.1 $\mathrm{mmol\,m^{-3}}$ and 1.7 $\mathrm{mmol\,m^{-3}}$. After a few hundred meter depth they increase rapidly. After a depth of 500 m, they are above 2 $\mathrm{mmol\,m^{-3}}$ and between 25°S and 60°N they are above 2.5 $\mathrm{mmol\,m^{-3}}$. Even deeper, they change only slightly.

The results for the Atlantic Ocean are shown in Figure 2e. They are similar to the one in the Pacific Ocean at the first few hundred meters depth as well as south of 50°S. However between 50°S and 30°N the average concentrations decrease from around 2.2 $\mathrm{mmol\,m^{-3}}$ to around 1.5 $\mathrm{mmol\,m^{-3}}$ at a depth of about 1250 m and then increase again slightly near the seafloor. North of 30°N, the values remain almost constant at 1.1 $\mathrm{mmol\,m^{-3}}$ at a depth greater than 1000 m.

Figure 2f illustrates the climatological means in the Indian Ocean. These are very similar to those in the Pacific Ocean at corresponding depths and longitudes.



(a) water surface
(averaged over time and 0 to 25 m depth)

(b) averaged over all but depth

(c) average monthly change

(d) Pacific Ocean
(averaged over time and between 125° E and 70° W)

(e) Atlantic Ocean
(averaged over time and between 70° W and 20° E)

(f) Indian Ocean
(averaged over time and between 20° E and 125° E)

Figure 3. Standard deviation of phosphate measurements (η) in $\mathrm{mmol\,m^{-3}}$.

The estimated climatological means using the median instead of the arithmetic mean look quite similar. Their average absolute difference is 0.004 $\mathrm{mmol\,m^{-3}}$. It is 0.011 $\mathrm{mmol\,m^{-3}}$ near the surface and decreases with growing depth.

3.3 Variabilities

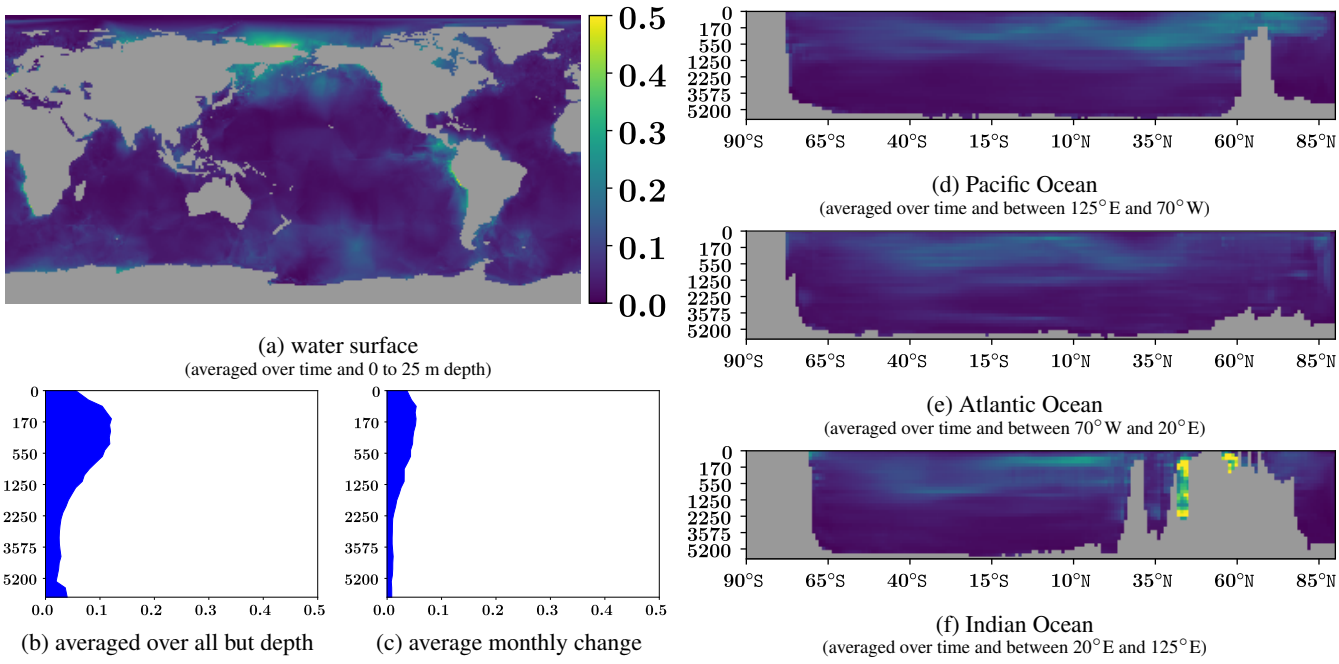
The estimated variabilities of the measurement results η as well as the short and the long scale variabilities, i.e., the variabilities of the noise ϵ and the true concentration δ, are described next using standard deviations and relative standard deviations.

Standard Deviations

The estimated standard deviations of the measurement results η are plotted in Figure 3. The average estimated standard deviation is 0.11 $\mathrm{mmol\,m^{-3}}$.

Its time averaged standard deviations near the surface are shown in Figure 3a. Here, the highest standard deviations are between 0.35 and 0.45 $\mathrm{mmol\,m^{-3}}$ around the eastern part of Russia, near the west and south of South America and the west coastal region of Southern Africa. Elsewhere near the coast or in the Southern Ocean, the standard deviations are between 0.15 and 0.25 $\mathrm{mmol\,m^{-3}}$. They are between 0.05 and 0.15 $\mathrm{mmol\,m^{-3}}$ in the remaining areas.

The standard deviations averaged over all but the depth is plotted in Figure 3b. It increases from 0.15 $\mathrm{mmol\,m^{-3}}$ at the surface to 0.23 $\mathrm{mmol\,m^{-3}}$ at around 120 m depth. Afterwards it barely changes up to a depth of 500 m. Then it strictly decreases to 0.06 $\mathrm{mmol\,m^{-3}}$ at a depth of 2500 m and hardly changes while going deeper.



(a) water surface
(averaged over time and 0 to 25 m depth)

(b) averaged over all but depth

(c) average monthly change

(d) Pacific Ocean
(averaged over time and between 125° E and 70° W)

(e) Atlantic Ocean
(averaged over time and between 70° W and 20° E)

(f) Indian Ocean
(averaged over time and between 20° E and 125° E)

Figure 4. Short scale, i.e. noise, standard deviation of phosphate concentration (ϵ) in $\mathrm{mmol\,m^{-3}}$.

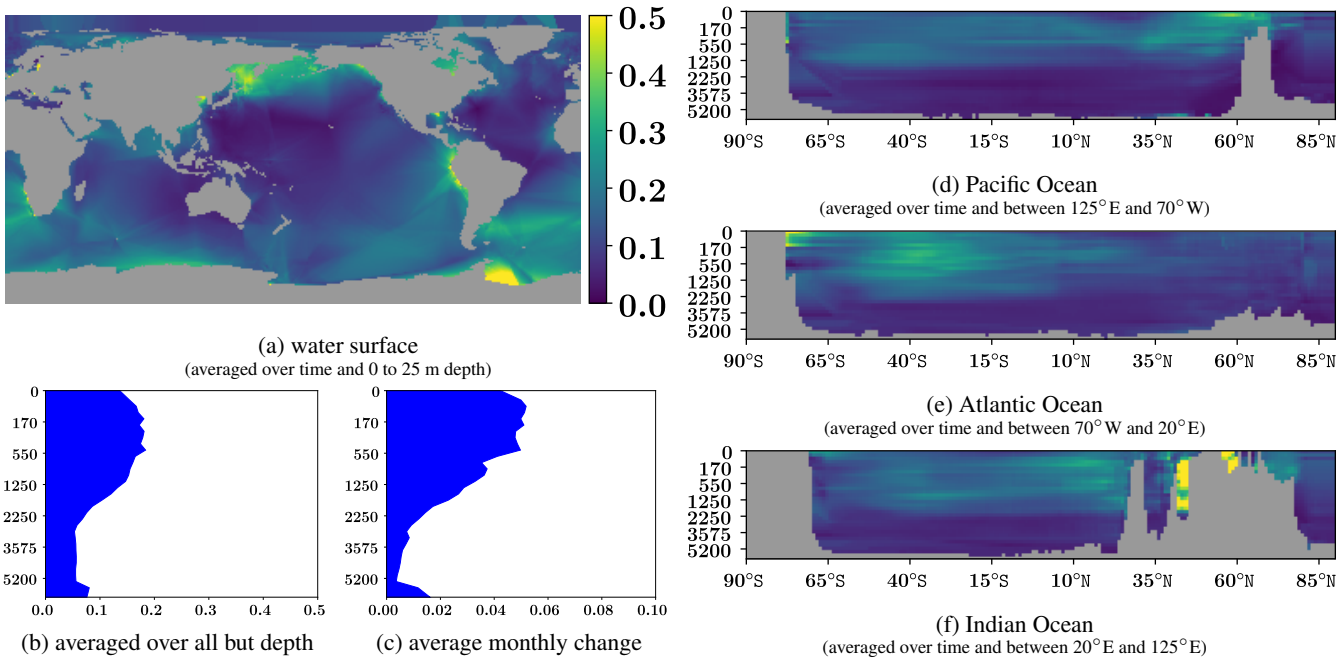
The average absolute change in the standard deviation after one month is shown in Figure 3c. Up to a depth of 500 m, the change is between 0.05 and 0.06 $\mathrm{mmol\,m^{-3}}$. Then it decreases to around 0.01 $\mathrm{mmol\,m^{-3}}$ at a depth of 2500 m whereupon it remains almost constant.

In Figure 3d, 3e and 3f, the standard deviations in the Pacific Ocean, the Atlantic Ocean and the Indian Ocean averaged over time and longitude are shown. They are between 0.25 and 0.35 $\mathrm{mmol\,m^{-3}}$ above 1000 m depth in the Pacific, the Indian Ocean and the South Atlantic Ocean. Elsewhere they are lower than 0.25 $\mathrm{mmol\,m^{-3}}$ and deeper than 1500 m even lower than 0.15 $\mathrm{mmol\,m^{-3}}$. Peaks between 0.40 and 0.45 $\mathrm{mmol\,m^{-3}}$ are located in the far north of the Pacific as well as in in the south of the Atlantic and the Southern Ocean.

As explained in Subsection 2.3, the standard deviations of the measurement results η are composed of the standard deviations of the true concentration δ and the noise ϵ and are plotted in Figure 4 and Figure 5. The standard deviation of the true concentration describes the climatological variability. In contrast, the standard deviation of the noise covers the short scale variabilities which include the variability within the grid boxes in specific years as well as measurement inaccuracies.

The average estimated standard deviation were 0.10 $\mathrm{mmol\,m^{-3}}$ for the true concentration and 0.05 $\mathrm{mmol\,m^{-3}}$ for the noise. Hence, the difference between the measurement results and the climatological mean arose to about two thirds from climatological variabilities and to about one third from short scale variabilities.

The standard deviations of the noise usually were below 0.10 $\mathrm{mmol\,m^{-3}}$ except around the eastern part of Russia, near the west and south of South America and the west coastal region of Southern Africa, where they were between 0.25 to 0.45 $\mathrm{mmol\,m^{-3}}$, 0.15 to 0.35 $\mathrm{mmol\,m^{-3}}$ and 0.10 to 0.30 $\mathrm{mmol\,m^{-3}}$, respectively, in the upper few hundred meters.



(a) water surface
(averaged over time and 0 to 25 m depth)

(b) averaged over all but depth

(c) average monthly change

(d) Pacific Ocean
(averaged over time and between 125° E and 70° W)

(e) Atlantic Ocean
(averaged over time and between 70° W and 20° E)

(f) Indian Ocean
(averaged over time and between 20° E and 125° E)

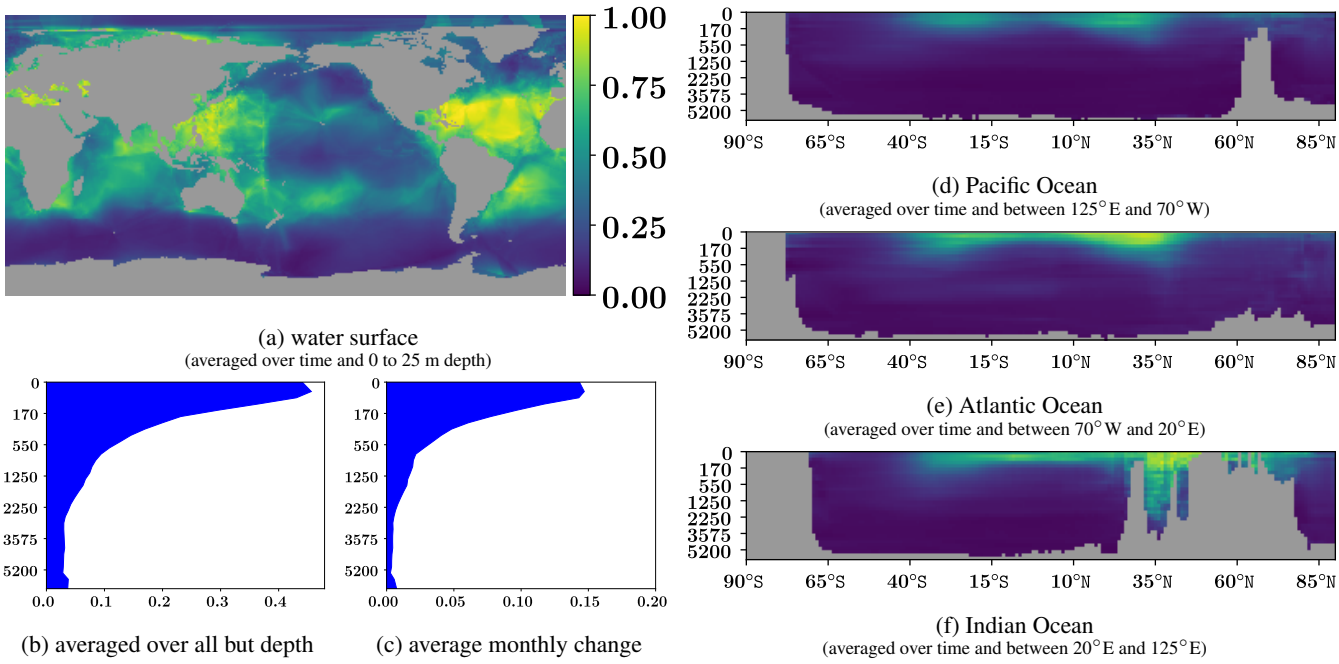
Figure 5. Climatological standard deviation of phosphate concentration (δ) in $\mathrm{mmol\,m^{-3}}$.

As the standard deviation of the noise includes the variability within a grid box in a specific year, a high standard deviation of the noise indicates that a finer resolution could provide more accurate climatological information and in contrast a low standard deviation indicates that the resolution could be reduced without loosing climatological information. A higher resolution thus seems to make sense for some coastal regions, whereas regions deep in the ocean and far away from coasts could also be resolved with a lower resolution. This indicates that a more non-uniform resolution may be advantageous and should be considered in any further analysis.

The standard deviations averaged over all but the depth are for the noises about half of that for the measurement results. The average monthly absolute change is usually 0.01 $\mathrm{mmol\,m^{-3}}$ lower for the noises than for the measurement results. Among the values for the noise averaged over time and longitude in the Pacific Ocean, the Atlantic Ocean and the Indian Ocean only the area around the eastern part of Russia stands out.

Usually the standard deviations of the true concentrations are approximately 0.05 $\mathrm{mmol\,m^{-3}}$ lower than the standard deviations of the measurement results except for areas where the standard deviations of the noise is high and thus the difference as well. This relationship arises because the variances of the measurement results are the sum of the variances of the true concentration and the noise.

We also quantified the variabilities of the true concentration δ and the noise ϵ by interquartile ranges. However, the plotted results look quite similar to the ones of the standard deviation and are therefore not included.



(a) water surface
(averaged over time and 0 to 25 m depth)

(b) averaged over all but depth

(c) average monthly change

(d) Pacific Ocean
(averaged over time and between 125° E and 70° W)

(e) Atlantic Ocean
(averaged over time and between 70° W and 20° E)

(f) Indian Ocean
(averaged over time and between 20° E and 125° E)

Figure 6. Relative standard deviation of phosphate measurements (η).

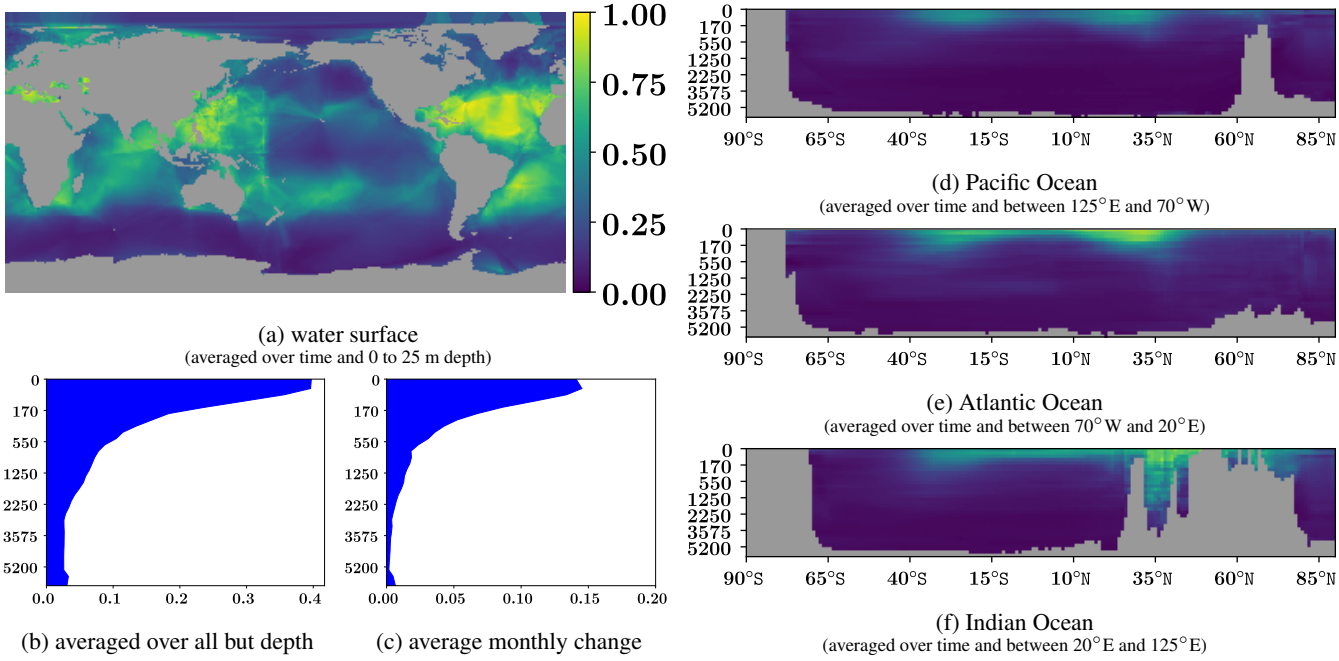
Relative Standard Deviations

The relative standard deviation is the standard deviation divided by the expected value and therefore allows to assess the variability relative to the typical size. Since the expected values of the noise ϵ are assumed to be zero, it only makes sense to look at the relative standard deviations of the true concentration δ and the measurement results η. The expected value of these two is the climatological mean.

The standard deviations for the true concentrations and the measurement results are quite similar as explained earlier. For this reason, the relative standard deviations of the true concentration and the measurement results are quite similar, too, as shown in Figure 6 and 7. The only differ notable near the eastern part of Russia where the standard deviations of the noise are very high.

The average relative standard deviation of the measurement results is 0.07 and that of the true concentration is 0.06. The highest values of the relative standard deviations of the measurement results are mostly between 0.4 and 1. This is mainly near the surface between 40°S and 45°N everywhere expect in the east of the Pacific Ocean. Elsewhere near the surface, the relative standard deviations are usually below 0.4 and in the Southern Ocean even below 0.2. Usually, high relative standard deviations result from low climatological means as well as low relative standard deviations from high climatological means.

The average relative standard deviation of the measurement results depending on the depth is 0.45 near the surface and decreases fast with growing depth. At 500 m depth it is 0.13, and below 2000 m depth it is below 0.05. The average absolute



(a) water surface
(averaged over time and 0 to 25 m depth)

(b) averaged over all but depth

(c) average monthly change

(d) Pacific Ocean
(averaged over time and between 125° E and 70° W)

(e) Atlantic Ocean
(averaged over time and between 70° W and 20° E)

(f) Indian Ocean
(averaged over time and between 20° E and 125° E)

Figure 7. Relative standard deviation of phosphate measurements (η).

difference after one month is 0.15 near the surface and decreases in a similar way. At 500 m depth it is 0.04, and below 2000 m depth it is below 0.01. Hence, the climatological variability is negligible after a few hundred meters of depth.

In the Pacific Ocean, the Atlantic Ocean and the Indian Ocean the previously mentioned areas between 40°S and 45°N and with a depth up to 200 m stand out with higher values.

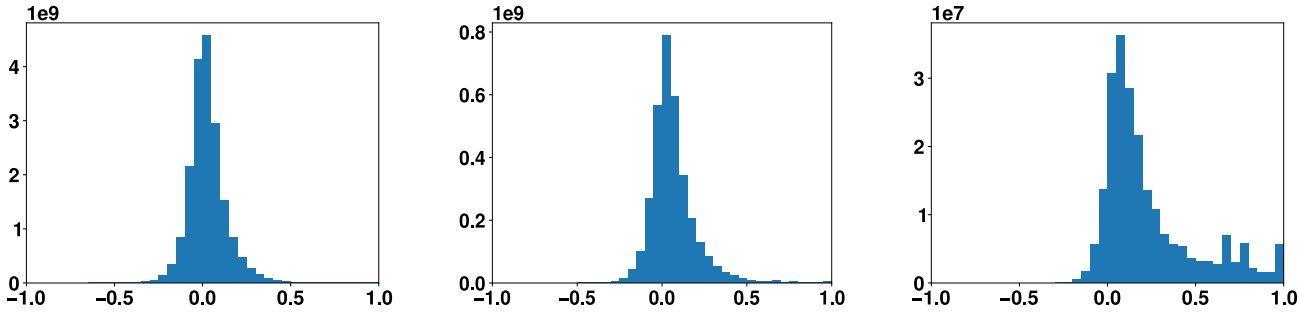
The quartile coefficients of dispersion of the true concentration δ and the measurement results η were also used to quantify its variabilities as an alternative to its relative standard deviations. However, the results look quite similar and are thus not illustrated here.

3.4 Statistical Dependencies

We quantified the statistical dependencies regarding δ, ϵ and η with covariances and correlations. It suffices to estimate the covariance or correlation of the true concentration δ. The other covariances and correlations result from these estimates (compare Subsection 2.4).

Pointwise Correlations

The calculated estimates depend on the required number of years where measurement results are available. We set up three estimates: The first covers measurements within at least 35 years, second within at least 40 years and for the final at least 45 years resulting in 1.9×10^{10}, 3.3×10^{9} and 2.2×10^{8} pointwise estimates, respectively. Estimating the correlation, we assumed



(a) 35 years required (b) 40 years required (c) 45 years required

Figure 8. Number of estimated correlations. (0.05 used as bin size in the histograms)

the standard deviation of the noise ϵ to be at least 0.1 $\mathrm{mmol\,m^{-3}}$, which corresponds to a coarse measurements accuracy. Otherwise some unrealistic low sample standard deviations would compromise our correlation estimation.

The number of estimated correlation values is shown in Figure 8 where estimates less than 0.01 in absolute value were not considered. The figure shows clearly that most of the estimated correlations are small in amount supporting our approach to assume that not estimated correlations are zero. High absolute values in the estimates are rare. Positive estimates are slightly more frequent than negative ones. Of course, these results depend on the spatial and temporal distribution of the measurement points and would possibly look different with a more uniform distribution of the measurement points.

Correlation Matrix

To generate a well-conditioned positive definite correlation matrix, we considered 4.1 million measurement points.

In average, every estimate was reduced in absolute value by 0.078 to generate this matrix. In total 67% of the estimates were modified. By saving the $LDL^\top$ decomposition instead of the generated correlation matrix itself, the amount of entries could be reduced by 36%. The sparsity pattern of the unmodified correlation matrix and the permuted well-conditioned correlation matrix are plotted in Figure 9.

(a) unpermuted (b) permuted

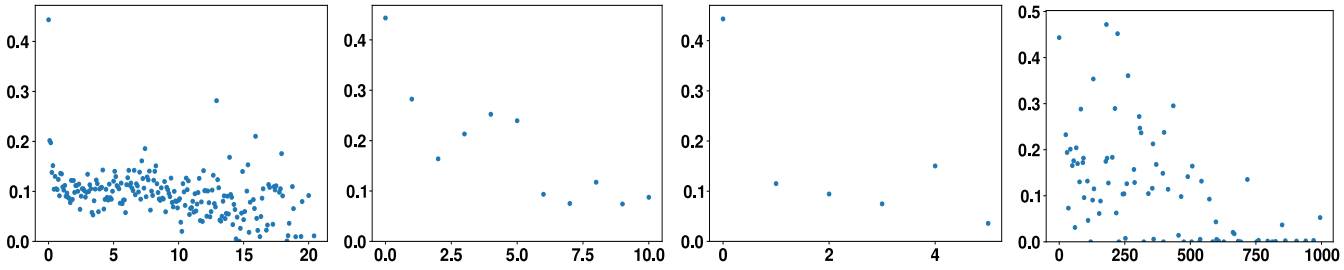
Figure 9. Sparsity pattern of correlation matrix.

Correlations Dependencies on Distances

We analyzed if the estimated correlation depends solely on the distance between the related measurement points (compare Subsection 2.4).

We sorted all estimated pointwise correlations in groups in which the associated measuring points differ by the same value. For each of these group with at least ten correlations, we calculated the interquartile range, resulting in more than 3×10^5 values in total. If the estimated correlations would depend only on the distance between the associated measuring points, most of these interquartile ranges must be close to zero. However, 50% are greater or equal than 0.05, 25% are greater than 0.10, and 5% are even greater than 0.20. Hence, it is unlikely that the estimated correlations can be described by a function depending only on the distance between the corresponding measurement points when using 0.05 as a threshold or even 0.10 as a less restrictive threshold.

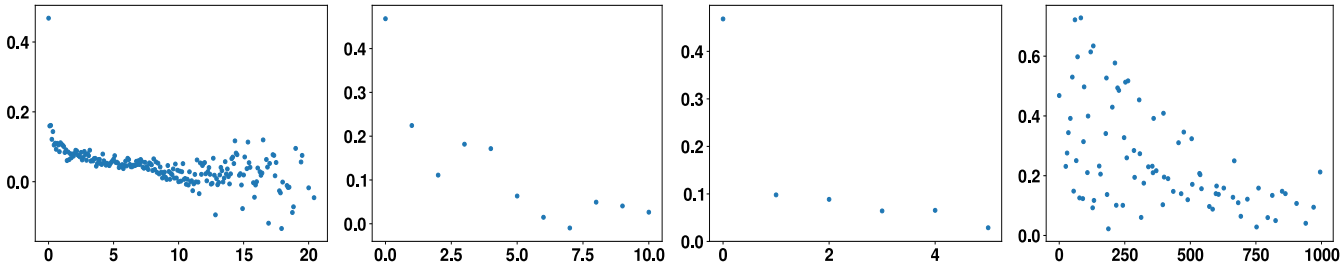
The calculated interquartile ranges, associated with measuring points which differ in a single direction, are plotted in Figure 10. They show, especially with regard to depth, a solely distance-related dependence is unlikely.



(a) time (b) longitude (c) latitude (d) depth

Figure 10. Interquartile ranges of estimated correlations associated with measuring points which differ in a single direction.

The means of the groups were plotted in Figure 11 for measuring points which differ in a single direction. They must be close to each other at points where the associated distances are close to each other, if the estimated correlations can be described by a continuous function that depends only on the distances. The graphs do not support this assumption, especially for points which differ by the depth or by long times.

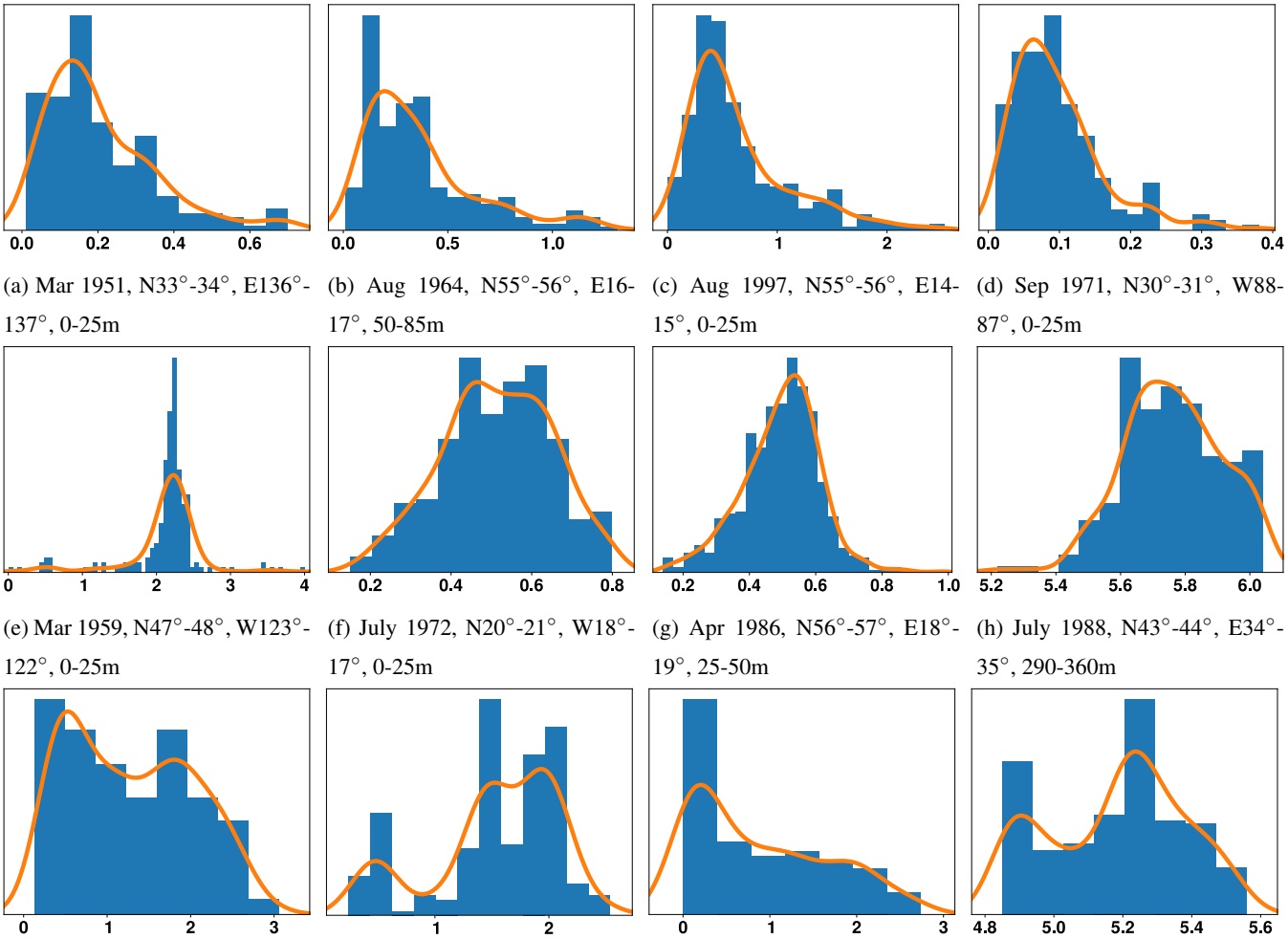


(a) time (b) longitude (c) latitude (d) depth

Figure 11. Means of estimated correlations associated with measuring points which differ in a single direction.

It should be noted that considerably less estimates for points that differ only in longitude or only in latitude are available than for points that differ only in time or only in depth. Hence, the results regarding the longitude and latitude lack significance.

The plots also show that the correlation tend to decrease in terms of absolute value with increasing distance between the measurement points.



(a) Mar 1951, N33°-34°, E136°-137°, 0-25m

(b) Aug 1964, N55°-56°, E16-17°, 50-85m

(c) Aug 1997, N55°-56°, E14-15°, 0-25m

(d) Sep 1971, N30°-31°, W88-87°, 0-25m

(e) Mar 1959, N47°-48°, W123°-122°, 0-25m

(f) July 1972, N20°-21°, W18°-17°, 0-25m

(g) Apr 1986, N56°-57°, E18°-19°, 25-50m

(h) July 1988, N43°-44°, E34°-35°, 290-360m

(i) June 1970, N44°-45°, W125-124°, 0-25m

(j) Apr 1975, N55°-56°, E15-E16°, 50-85m

(k) June 1982, N47°-48°, W125-124°, 0-25m

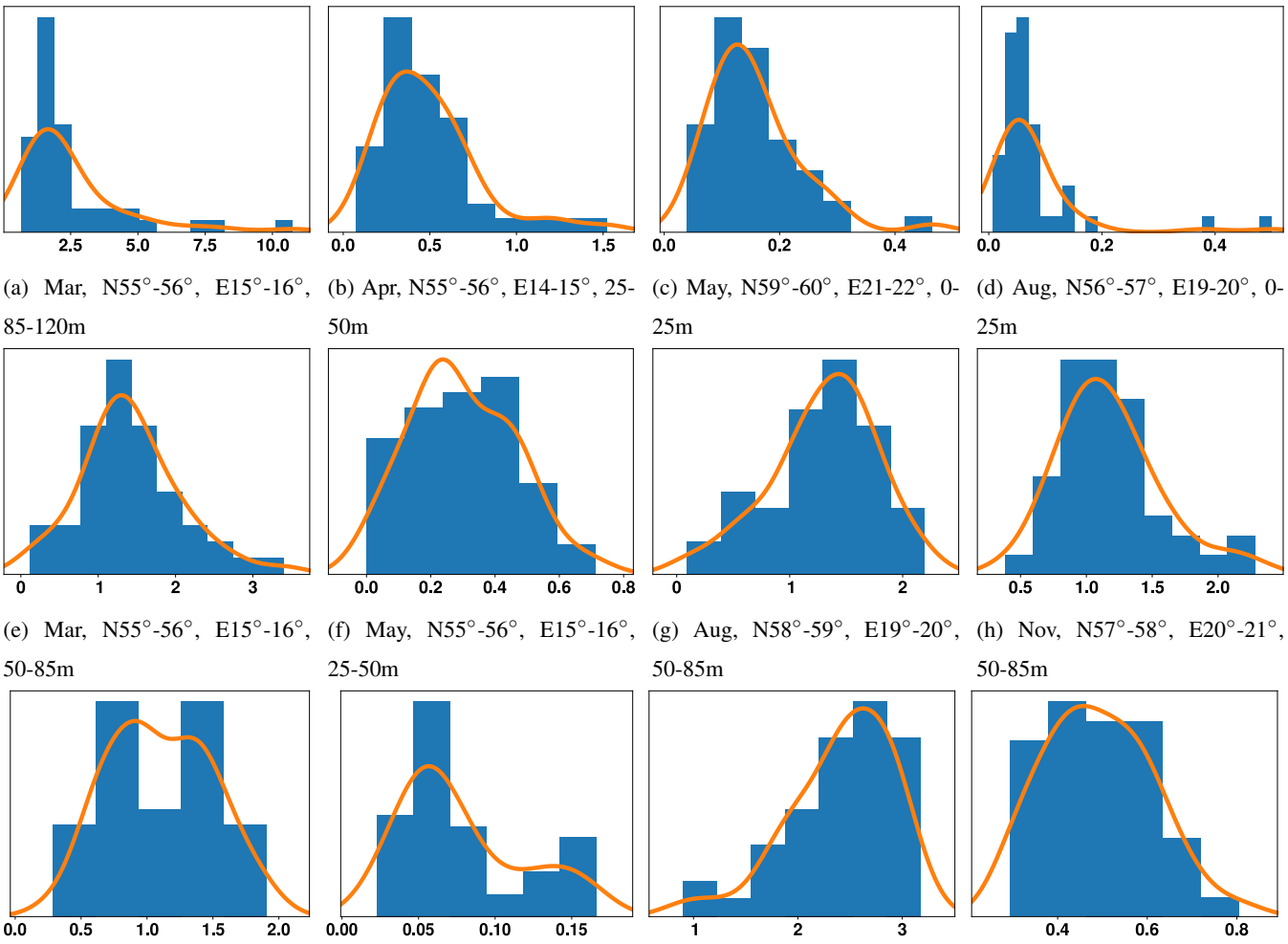
(l) June 1988, N42°-43°, E32-33°, 220-290m

Figure 12. Selected histograms and kernel density estimations for the measurement results (η). Most of them look like the ones in the first and second row, i.e., log-normal or normal distributed. A few look like the ones in the last row which do not look like either of the these two distributions.

3.5 Probability Distributions

We investigated the underlying probability distributions of the measurement results η, the true concentration δ and the noise ϵ, (compare Subsection 2.5).

To assess the probability distributions, we produced histograms and kernel density estimations (KDEs). For the histograms, the sizes of the bins were determined by the rule of Freedman and Diaconis (1981). For the KDEs, normal kernel were used with bandwidths determined by the rule of Scott (2015).



(a) Mar, N55°-56°, E15°-16°, 85-120m (b) Apr, N55°-56°, E14-15°, 25-50m (c) May, N59°-60°, E21-22°, 0-25m (d) Aug, N56°-57°, E19-20°, 0-25m

(e) Mar, N55°-56°, E15°-16°, 50-85m (f) May, N55°-56°, E15°-16°, 25-50m (g) Aug, N58°-59°, E19°-20°, 50-85m (h) Nov, N57°-58°, E20°-21°, 50-85m

(i) May, N57°-58°, E20-E21°, 50-85m (j) Aug, N52°-53°, E20-21°, 0-25m (k) Nov, N57°-58°, E20-21°, 85-120m (l) Dec, N57°-58°, E11-12°, 0-25m

Figure 13. Selected histograms and kernel density estimations for the true concentration (δ). Most of them look like the ones in the first and second row, i.e., log-normal or normal distributed. A few look like the ones in the last row which do not look like either of the these two distributions.

Histograms and KDEs with at least hundred values were generated for the measurement results η and with at least forty values for the true concentration δ, resulting in around eight hundred and four hundred plots, respectively. When most of the values are close to zero, the plots usually look like log-normal distributions and otherwise usually like normal distributions. However, there is no clear overall trend and some plots look like neither of these distributions.

A selection of these is presented in Figure 12 and 13. In the first rows typical plots which indicate log-normal distributions are presented, in the second rows typical plots which indicate normal distributions and in the third rows some of the rear obscure distributions.

As stated in Subsection 2.5, the values of the noise ϵ are just the values of η shifted by a constant. Hence, the histograms and KDEs for ϵ look similar to the ones in Figure 12, just with other values on the horizontal axis.

We also applied the statistical tests mentioned in Subsection 2.5 to test against normal and log-normal distributions. We used a significance level of 1% and tested only at points where at least 40 values are available.

Regarding the measurement results η, the tests rejected in average the normal distribution assumption for 73% of the cases. The log-normal distribution assumption was rejected in average for 65% of the cases. For the true concentration δ, the tests rejected in average the normal distribution assumption for 25% of the cases and the log-normal distribution assumption for 52% of the cases.

These high number of rejections gave a different impression than the visual inspection, indicating that at some grid boxes the probability distributions might neither be a normal nor a log-normal one. The high proportion of rejections may be explained by the fact that the measurement results have at most three significant digits. Thus they represent at best only heavily rounded realizations of a normal or log-normal distribution. However, these roundings are not taken into account in the tests.

4 Conclusions

Phosphate is a key component in understanding the marine ecosystem. Millions of measurement data for phosphate are available at the World Ocean Database but these are not statistically analyzed in such extent as it is done here.

The climatological mean and variability as well as the short scale variability were quantified. The results indicate that it makes sense to increase the resolution of the analysis in (some) coastal areas to obtain more accurate climatological information and to decrease the resolution in areas far away from coasts and deep in the ocean without loosing climatological information.

The correlation of climatological concentrations were estimated as well. They generally decrease with increasing spatial and temporal distance, but do not solely depend on the distance.

The climatological concentrations and the measurement results seem to be mostly normally or log-normally distributed. However, there is no clear trend and in some cases they do not seem to belong to either distribution. Hence, they do not seem to originate from a single type of distribution.

This extensive analysis is useful in understanding the marine phosphate concentration. It may also be valuable in the calibration of marine biogeochemical models where our estimated standard deviations and correlations could be incorporated to achieve a more accurate model calibration (cf. (Walter and Pronzato, 1997, 4) or (Seber and Wild, 2003, 2.1.4)).

The analysis is also helpful in planning new phosphate measurements. A lower short-scale standard deviation would mean that average concentration in this year could be be determined more accurate compared to a higher one. If the long-scale standard deviation is small as well the climatological concentration can also be determined more accurate.

If the measurements should be used to estimate parameters of a model or to determine the most realistic one among several models, optimal experimental design methods (cf. (Walter and Pronzato, 1997, 6) or (Pronzato and Pázman, 2013, 5)), which include statistical properties such as standard deviations and correlations, can be used to determine the places and times of measurements which provide the highest information gain.

The approaches in the statistical analysis are not limited to phosphate but can also be applied to other data which satisfies the assumptions.

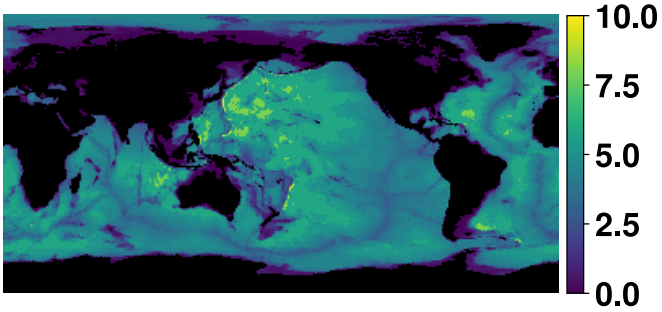
Appendix A: Spatial and Temporal Resolution

The spatial grid for our calculation was constructed using the one-degree spatial grid of the World Ocean Atlas 2013 (Garcia et al., 2014, Chapter 3.4), which is based on the global relief model ETOPO2v2 (National Geophysical Data Center (2006)).

The spatial grid of the World Ocean Atlas 2013 has a resolution of one degree and 137 vertical layers with increasing thickness. The annual data provided by the World Ocean Atlas 2013 are available up to a depth of 5500 meters which represents 102 layers. The seasonal and monthly data are available up to a depth of 500 meters which represents 37 layers.

We decided to reduce the vertical resolution to 33 layers in order to increase the number of data in each grid box. The new depth in each grid box was chosen as the highest depth in Table A1b which is less or equal to the depth in the World Ocean Atlas 2013. The resulting depths are plotted in Figure A1a.

For the temporal resolution one month was used at all layers, allowing to cover time-dependent changes in all layers.



(a) Depths in kilometer.

0	290	1080	2495	4510
25	360	1250	2740	4855
50	455	1420	3010	5200
85	550	1615	3280	6000
120	670	1810	3575	8000
170	790	2030	3870	10000
220	935	2250	4190	

(b) Depths possible for grid boxes in meters.

Figure A1. Depths in the spatial grid used for this statistical analysis.

Appendix B: Interpolation

The data in our calculations were linearly interpolated by triangulating the input data with the method of Qhull (Barber et al. (1996)) and performing linear barycentric interpolation on each triangle. Values for points outside the convex hull of the data points were interpolated using the value of the nearest data point. For this purpose, a kd-tree (Maneewongvatana and Mount (1999)) was used to rapidly look up the nearest neighbor of each point. We used a Python (Python Software Foundation (2018)) implementation of both algorithms, which is part of the SciPy library (Jones et al. (2019) and Virtanen et al. (2019)).

The annual periodicity and the periodicity with respect to the longitude were included in the interpolation by assuming the same values for data points plus/minus the period.

Instead of the depth, the number of the corresponding depths level described in Table A1b was used for the interpolation. Thus, a vertical distance deep down in the ocean is weighted less than near the surface. This takes into account that the changes of the values at great depth are smaller than those closer to the surface.

The points were scaled so that the distance between two consecutive depth levels corresponds to a distance of one degree, and the length of one year corresponds to the circumference of the earth.

Appendix C: Software

The results in Section 3 have been calculated and visualized using the measurements software package (Reimer (2019b)) which is based on Python (Python Software Foundation (2018)), NumPy (Oliphant et al. (2019)), SciPy (Jones et al. (2019) and Virtanen et al. (2019)), Matplotlib (Caswell et al. (2019) and Hunter (2007)), utillib (Reimer (2019c)) and the matrix-decomposition library (Reimer (2019a)).

The measurements software package contains functions for all methods described in Section 2 as well as for visualizing corresponding result. It is especially suited for data from the World Ocean Database because it provides special functions for processing these data. However, it is not limited to these data.

References

Anderson, T. W. and Darling, D. A.: Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes, The Annals of Mathematical Statistics, 23, 193–212, http://www.jstor.org/stable/2236446, 1952.

Barber, C. B., Dobkin, D. P., and H., H.: The Quickhull Algorithm for Convex Hulls, ACM Trans. Math. Softw., 22, 469–483, https://doi.org/10.1145/235815.235821, 1996.

Bigg, G. R.: The Oceans and Climate, Cambridge University Press, second edn., 2003.

Boyer, T., Antonov, J., Baranova, O., Coleman, C., Garcia, H., Grodsky, A., Johnson, D., Locarnini, R., Mishonov, A., O'Brien, T., Paver, C., Reagan, J., Seidov, D., Smolyar, I., and Zweng, M.: World Ocean Database 2013, Tech. rep., National Oceanic and Atmospheric Administration, Silver Spring, https://doi.org/10.7289/V5NZ85MT, s. Levitus, Ed.; A. Mishonov, Technical Ed., 2013.

Caswell, T. A., Droettboom, M., Hunter, J., Lee, A., Firing, E., Stansby, D., Klymak, J., de Andrade, E. S., Nielsen, J. H., Varoquaux, N., Hoffmann, T., Root, B., Elson, P., May, R., Dale, D., Lee, J.-J., Seppänen, J. K., McDougall, D., Straw, A., Hobson, P., Gohlke, C., Yu, T. S., Ma, E., Vincent, A. F., Silvester, S., Moad, C., Katins, J., Kniazev, N., Ariza, F., and Ernest, E.: Matplotlib: 3.1.1, https://doi.org/10.5281/zenodo.3264781, https://doi.org/10.5281/zenodo.3264781, 2019.

Chen, Y., Wiesel, A., Eldar, Y. C., and Hero, A. O.: Shrinkage Algorithms for MMSE Covariance Estimation, IEEE Transactions on Signal Processing, 58, 5016–5029, https://doi.org/10.1109/TSP.2010.2053029, 2010.

D'Agostino, R. B.: An omnibus test of normality for moderate and large size samples, Biometrika, 58, 341–348, https://doi.org/10.1093/biomet/58.2.341, http://biomet.oxfordjournals.org/content/58/2/341.abstract, 1971.

D'Agostino, R. B. and Pearson, E. S.: Tests for departure from normality. Empirical results for the distributions of b_2 and $\sqrt{b_1}$, Biometrika, 60, 613–622, https://doi.org/10.1093/biomet/60.3.613, http://biomet.oxfordjournals.org/content/60/3/613.abstract, 1973.

Freedman, D. and Diaconis, P.: On the Histogram as a Density Estimator: L_2 Theory, Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete, 57, 453–476, https://doi.org/10.1007/BF01025868, 1981.

Garcia, H. E., Locarnini, R. A., Boyer, T. P., Antonov, J. I., Baranova, O. K., Zweng, M. M., Reagan, J. R., and Johnson, D. R.: World Ocean Atlas 2013, Volume 4: Dissolved Inorganic Nutrients (phosphate, nitrate, silicate), Tech. rep., National Oceanic and Atmospheric Administration, s. Levitus, Ed., A. Mishonov Technical Ed.; NOAA Atlas NESDIS 76, 25 pp., 2014.

Hunter, J. D.: Matplotlib: A 2D graphics environment, Computing in Science & Engineering, 9, 90–95, https://doi.org/10.1109/MCSE.2007.55, 2007.

Johnson, D. R., Boyer, T. P., Garcia, H. E., Locarnini, R. A., Baranova, O. K., and Zweng, M. M.: World Ocean Database 2013 User's Manual, Tech. Rep. NODC Internal Report 22, National Oceanographic Data Center, https://doi.org/10.7289/V5DF6P53, sydney Levitus, Ed.; Alexey Mishonov, Technical Ed.; NODC Internal Report 22, NOAA Printing Office, Silver Spring, MD, 172 pp, 2013.

Jones, E., Oliphant, T., Peterson, P., et al.: SciPy: Open source scientific tools for Python, http://www.scipy.org, version 1.3, 2019.

Kriest, I., Khatiwala, S., and Oschlies, A.: Towards an assessment of simple global marine biogeochemical models of different complexity, Progress in Oceanography, 86, 337–360, https://doi.org/10.1016/j.pocean.2010.05.002, http://eprints.ifm-geomar.de/9204/, 2010.

Ledoit, O. and Wolf, M.: A well-conditioned estimator for large-dimensional covariance matrices, Journal of Multivariate Analysis, 88, 365–411, https://doi.org/10.1016/S0047-259X(03)00096-4, http://www.sciencedirect.com/science/article/pii/S0047259X03000964, 2004.

Linacre, E.: Climate Data and Resources, Routledge, 1992.

Maneewongvatana, S. and Mount, D. M.: It's okay to be skinny, if your friends are fat, in: Center for Geometric Computing 4th Annual Workshop on Computational Geometry, vol. 2, pp. 1–8, 1999.

National Geophysical Data Center: ETOPO2v2 2-minute Global Relief Model, https://doi.org/10.7289/v5j1012q, 2006.

Oliphant, T. E. et al.: NumPy: N-dimensional array package for Python, http://www.numpy.org, version 1.17.3, 2019.

Pronzato, L. and Pázman, A.: Design of Experiments in Nonlinear Models: Asymptotic Normality, Optimality Criteria and Small-Sample Properties, Lecture Notes in Statistics, Springer, New York, NY, 2013.

Python Software Foundation: Python, http://www.python.org, version 3.7, 2018.

Reimer, J.: Approximation of Hermitian Matrices by Positive Semidefinite Matrices using Modified Cholesky Decompositions, ArXiv e-prints, https://arxiv.org/abs/1806.03196v2, 2019.

Reimer, J.: matrix-decomposition: a library for decompose (factorize) dense and sparse matrices in Python, https://doi.org/10.5281/zenodo.3558540, version 1.2, 2019a.

Reimer, J.: measurements library: Python functions to handle, statistically analyze and plot measurement data., https://doi.org/10.5281/zenodo.3558700, version 0.3, 2019b.

Reimer, J.: utillib library: Python functions used in several other projects., https://doi.org/10.5281/zenodo.3558698, version 0.3, 2019c.

Schäfer, J. and Strimmer, K.: A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics, Statistical Applications in Genetics and Molecular Biology, 4, 1–32, 2005.

Scott, D.: Multivariate Density Estimation: Theory, Practice, and Visualization, Wiley Series in Probability and Statistics, Wiley, 2015.

Seber, G. A. F. and Wild, C. J.: Nonlinear Regression, Wiley series in probability and statistics, Wiley-Interscience, 2003.

Shapiro, S. S. and Wilk, M. B.: An Analysis of Variance Test for Normality (Complete Samples), Biometrika, 52, 591–611, http://www.jstor.org/stable/2333709, 1965.

Storch, H. v. and Zwiers, F. W.: Statistical Analysis in Climate Research, Cambridge University Press, https://doi.org/10.1017/CBO9780511612336, 1999.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, I., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy: SciPy 1.0-Fundamental Algorithms for Scientific Computing in Python, CoRR, abs/1907.10121, http://arxiv.org/abs/1907.10121, 2019.

Walter, É. and Pronzato, L.: Identification of Parametric Models from Experimental Data, Communications and control engineering, Springer, New York, 1997.

Optimization of Model Parameters, Uncertainty Quantification and Experimental Designs for a Global Marine Biogeochemical Model

Joscha Reimer[1]

[1]Kiel University, Department of Computer Science, Algorithmic Optimal Control - CO_2 Uptake of the Ocean, 24098 Kiel, Germany

Correspondence: Joscha Reimer (joscha.reimer.edu@web.de)

Abstract. Methods for model parameter estimation, uncertainty quantification and experimental design are summarized in this paper. They are based on the generalized least squares estimator and different approximations of its covariance matrix using the first and second derivative of the model regarding its parameters.

The methods have been applied to a model for phosphate and dissolved organic phosphorus concentrations in the global ocean. As a result, model parameters have been determined which considerably improved the consistency of the model with measurement results.

The uncertainties regarding the estimated model parameters caused by uncertainties in the measurement results have been quantified as well as the uncertainties associated with the corresponding model output implied by the uncertainty in the model parameters. This allows to better assess the model parameters as well as the model output.

Furthermore, it has been determined to what extent new measurements can reduce these uncertainties. For this, the information content of new measurements has been predicted depending on the measured process as well as the time and the location of the measurement. This is very useful for planning new measurements.

1 Introduction

Computer models are a primary tools in natural sciences and contain parameters which are usually insufficiently known (cf. McGuffie and Henderson-Sellers (2005); Neelin (2010)). These parameters are usually estimated using noisy measurement data (cf. Aster et al. (2013); Seber and Wild (2003)). This noise implies uncertainty in the estimated parameters as well as in the corresponding model output. This uncertainty is often not quantified, which, on the contrary, is essential to correctly assess the model parameters and the model output.

In order to counter this, we are going to summarize some techniques to estimate unknown model parameters and to quantify and reduce associated uncertainties. The presented methods are suited for computational complex models. We are going to demonstrate this using a model describing the phosphate and dissolved organic phosphorus concentrations in the global ocean. Phosphate is a limiting nutrient for marine phytoplankton and therefore influences the growth of phytoplankton and the absorption of atmospheric CO_2 by the ocean (cf. (Bigg, 2003, Chapter 4)).

Only uncertainties resulting from measurements are subject of this article. Model errors, i.e., the discrepancies between models and their modeled processes, are not captured as well as numerical errors, i.e., discrepancies between mathematical models and their (discretized) implementations.

The methods, including methods for parameter estimation, uncertainty quantification and experimental design, are presented in Section 2. The marine model as application example is introduced in Section 3. The results obtained for this model are presented in Section 4. Finally, a conclusion is drawn in Section 5.

2 Methods for Parameter Estimation, Uncertainty Quantification and Experimental Design

The generalized least squares estimator, as a model parameter estimation method, is summarized in this section together with its statistical assumptions and properties. Based on this, methods to quantify the uncertainty in the model parameters estimate and its corresponding model output are presented. They are built on approximations of the covariance matrix of the estimator and resulting approximate confidence intervals. Finally, optimal experimental design methods, which allow to reduce the uncertainty by optimally planned additional measurements, are briefly introduced.

2.1 Model Parameter Estimation

An estimate $\hat{\theta}_n$ of the model parameters is usually obtained as the minimizer of an objective function ϕ_n:

$$\hat{\theta}_n := \underset{\theta \in \Omega}{\operatorname{argmin}} \, \phi_n(\theta), \tag{1}$$

where Ω is some set of feasible model parameters and n is the number of measurements.

By far, the most commonly used estimate is the (ordinary) least squares estimate (cf. (Seber and Wild, 2003, Section 2.1), (Pronzato and Pázman, 2013, Section 3.1), (Smith, 2013, Section 4.3) and (Walter and Pronzato, 1997, Section 3.1)) where the objective function is:

$$\phi_n^{OLS}(\theta) := \|y_n - f_n(\theta)\|_2^2.$$

Here, y_n denotes the vector of the measurement results and $f_n(\theta)$ the vector of model outputs corresponding to the measurement points and depending on the model parameters θ.

However, we will use the generalized least squares estimate (cf. (Seber and Wild, 2003, Subsection 2.1.4)) with the objective function:

$$\phi_n(\theta) := (y_n - f_n(\theta))^\mathsf{T} {\mathcal{C}_n}^{-1} (y_n - f_n(\theta)), \tag{2}$$

where $\mathcal{C}_n$ is some positive definite matrix. If $\mathcal{C}_n$ is the identity matrix, this equals the ordinary least squares estimate and thus can be interpreted as a generalization. If $\mathcal{C}_n$ is a diagonal matrix, this corresponds to weighted least squares estimates.

The estimator corresponding to the generalized least squares estimate is the random vector:

$$\Theta_n := \underset{\theta \in \Omega}{\operatorname{argmin}} (Y_n - f_n(\theta))^\mathsf{T} {\mathcal{C}_n}^{-1} (Y_n - f_n(\theta)),$$

where Y_n is a random vector of which the measurement results y_n are a realization.

The estimator has some appealing properties under certain regularity conditions (cf. Jennrich (1969), Amemiya (1983), (Seber and Wild, 2003, Section 12.2), (Walter and Pronzato, 1997, Subsection 3.3.3) and (Pronzato and Pázman, 2013, Section 3.1)): It is consistent, asymptotically unbiased, asymptotically normal and asymptotically efficient. Hence, the estimator Θ_n converges almost surely to the desired model parameters if the number of measurements n goes to infinity, making it the most accurate estimator among all asymptotically unbiased estimators.

One of the regularity conditions is that the statistical model, which includes the model function f and the measurement noise, are correctly specified. For the generalized least squares estimator, this means that some true model parameters $\theta^* \in \Omega$ exist with:

$$Y_n \sim \mathcal{N}(f_n(\theta^*), \sigma^2 \mathcal{C}_n), \tag{3}$$

where σ is some positive scalar. This implies that the model can describe the modeled process with appropriate parameters and that the measurement noise is unbiased and normally distributed with covariance matrix $\sigma^2 \mathcal{C}_n$.

If the assumed statistical model in (3) is correct, the generalized least squares estimator is the maximum likelihood estimator. The desired model parameters, that should be estimated, are then θ^* and the consistency means then almost surely convergence to θ^*.

If the assumed statistical model is incorrect, the generalized least squares estimator Θ_n is the quasi maximum likelihood estimator, also known as pseudo maximum likelihood estimator, regarding the set of probability distributions:

$$\mathcal{P}_n := \{\mathcal{N}(f_n(\theta), \sigma^2 \mathcal{C}_n) \mid \theta \in \Omega\}.$$

This estimator is still consistent, asymptotically unbiased and asymptotically normal under certain regularity conditions (cf. White (1981) and White (1982)). However, the estimator no longer has to be efficient. Consistency means in this case the almost sure convergence to some $\theta^* \in \Omega$ so that $\mathcal{N}(f_n(\theta^*), \sigma^2 \mathcal{C}_n) \in \mathcal{P}_n$ has minimal difference to the probability distribution of Y_n among all probability distributions in $\mathcal{P}_n$.

The other regularity conditions vary slightly depending on the reference. They usually includes, that the model function f is twice continuously differentiable and that Ω is closed and bounded. Furthermore, θ^* must be uniquely identifiable, implying the measurement points must be chosen such that the model output at this points differs sufficiently for the model parameters θ^* compared to other model parameters $\theta \in \Omega$.

It should be noted that at some references only the ordinary least squares estimator is considered. However, their statements can be extended to the generalized least squares estimator by considering:

$$\phi_n(\theta) = \|\tilde{y}_n - \tilde{f}_n(\theta)\|_2^2,$$

as an ordinary least squares estimation with $\tilde{y}_n := {\mathcal{C}_n}^{-0.5} y_n$ and $\tilde{f}_n(\theta) := {\mathcal{C}_n}^{-0.5} f_n(\theta)$ (cf. (Seber and Wild, 2003, Subsection 2.1.4)).

3

2.2 Uncertainty in Parameter Estimation

The uncertainty in the estimated model parameters θ_n implied by the noise in the measurement results can be described by the probability distribution of the estimator Θ_n. Hence, we derive different approximations of this probability distribution in this subsection and calculate from these approximate confidence intervals for the unknown model parameters θ^*.

In order to approximate the probability distribution of Θ_n, we assume that:

$$\Theta_n \sim \mathcal{N}(\theta^*, \mathcal{V}_n), \tag{4}$$

where $\mathcal{V}_n$ is the covariance matrix of Θ_n. This is reasonable due to the asymptotically normal distribution and the asymptotically unbiasedness of the estimator Θ_n. They are ensured if the previous mentioned regularity conditions are fulfilled regardless of whether the assumed statistical model (3) is correct or not.

The error made due to assumption (4) is small, if n is sufficiently large. If the model function f_n is linear regarding the model parameters and the statistical model (3) is correct, (4) is a consequence (cf. (Smith, 2013, Section 7.2) and (Tenorio, 2017, Section 2.6)). Thus, even if n is low, the error made due to assumption (4) is small if the second and higher derivatives of f_n are close to zero and the statistical model assumed in (3) is sufficiently close to reality.

To approximate the covariance matrix $\mathcal{V}_n$, we use three different approximations. In order to introduce these, we first define $J_n(\hat{\theta}_n)$ as the Jacobian matrix of the model function f_n and $H_n(\hat{\theta}_n)$ as the Hessian matrix of the objective function $\frac{1}{2}\phi_n$ both at the estimate $\hat{\theta}_n$ as well as:

$$F_n(\hat{\theta}_n) := J_n(\hat{\theta}_n)^\mathsf{T} \mathcal{C}_n^{-1} J_n(\hat{\theta}_n). \tag{5}$$

$F_n(\hat{\theta}_n)$, sometimes instead $H_n(\hat{\theta}_n)$, is called the Fisher information matrix for the model parameter $\hat{\theta}_n$ (cf. (Pukelsheim, 2006, Section 3.10)).

The three approximations of $\mathcal{V}_n$ are:

$$\mathcal{V}_n^{(F)}(\hat{\theta}_n) := \sigma^2 F_n(\hat{\theta}_n)^{-1}, \tag{6}$$

$$\mathcal{V}_n^{(H)}(\hat{\theta}_n) := \sigma^2 H_n(\hat{\theta}_n)^{-1}, \tag{7}$$

$$\mathcal{V}_n^{(F,H)}(\hat{\theta}_n) := \sigma^2 H_n(\hat{\theta}_n)^{-1} F_n(\hat{\theta}_n) H_n(\hat{\theta}_n)^{-1}. \tag{8}$$

$\mathcal{V}_n^{(F)}(\hat{\theta}_n)$ is the most common of these approximations (cf. (Seber and Wild, 2003, Subsection 2.1.2), (Smith, 2013, Section 7.3), (Tenorio, 2017, Section 5.2), (Walter and Pronzato, 1997, Subsection 5.3.1) and Donaldson and Schnabel (1987)). It is derived by assuming the correctness of the statistical model (3) and applying the linear least squares theory (cf. (Smith, 2013, Section 7.2) and (Tenorio, 2017, Section 2.6)) to the linearized model $l_n(\theta) := f_n(\hat{\theta}_n) + J_n(\hat{\theta}_n)(\theta - \hat{\theta}_n)$.

$\mathcal{V}_n^{(H)}(\hat{\theta}_n)$ is justified by the asymptotic theory for nonlinear least squares estimation (cf. (Seber and Wild, 2003, Subsection 12.2.3)), where under the assumed regularity conditions, $\frac{1}{n}F_n(\hat{\theta}_n)$ equals $\frac{1}{n}H_n(\hat{\theta}_n)$ asymptotically if the statistical model (3) is correct.

$\mathcal{V}_n^{(F,H)}(\hat{\theta}_n)$ is derived by the asymptotic theory of quasi maximum likelihood estimators (cf. White (1982)) where it is not necessary that the assumed statistical model (3) is correct.

If the statistical model (3) is correct, all three approximations are asymptotically equal and approach asymptotically the true asymptotic covariance matrix of Θ_n (cf. White (1982)). Nevertheless, they are usually not equal for a finite number of measurements, because:

$$H_n(\hat{\theta}_n) = F_n(\hat{\theta}_n) + \sum_{k=1}^{n} H_k^f(\hat{\theta}_n) \left(\mathcal{C}_n^{-1}(y_n - f_n(\hat{\theta}_n)) \right)_k ,$$

where $H_k^f(\hat{\theta}_n)$ is the Hessian matrix of the model at the k-th measurement point with respect to its parameters evaluated at $\hat{\theta}_n$. However, if f is a linear function, all three approximations are equal, since $H_k^f(\hat{\theta}_n) = 0$.

It is not obvious which of these three approximations entails the smallest error if the statistical model (3) is correct and f is nonlinear. Different recommendations can be found in the literature (cf. Donaldson and Schnabel (1987), Cao and Spall (2012), Cao and Spall (2009) and Efron and Hinkley (1978)).

However, if the statistical model (3) is not correct, which is the common case, only $\mathcal{V}_n^{(F,H)}(\hat{\theta}_n)$ approaches asymptotically the true asymptotic covariance matrix of Θ_n (cf. White (1982)) and, hence, should be preferred.

If σ is unknown, it can be estimated by:

$$\hat{\sigma}_n^2 := \tfrac{1}{n}\phi(\hat{\theta}_n). \tag{9}$$

This is an estimation of a consistent and asymptotically efficient estimator for σ^2, if the assumed statistical model in (3) is correct (cf. (Seber and Wild, 2003, Subsection 2.2.1)). Otherwise $\hat{\sigma}_n^2$ converges almost surely to $\sigma^2 + e$, where $e \geq 0$ is the prediction mean square error, (cf. (Seber and Wild, 2003, Subsection 12.2.4) and (Pazman and Pronzato, 2006, Theorem 1)). Hence, $\hat{\sigma}_n^2$ usually overestimates σ^2 in this case.

After the covariance matrix $\mathcal{V}_n$ is approximated by $\hat{\mathcal{V}}_n$, approximate confidence intervals for the unknown model parameters θ^* can be constructed. For this, we first note that (4) considered component by component implies:

$$(\Theta_n)_i \sim \mathcal{N}\left((\theta^*)_i, (\mathcal{V}_n)_{ii}\right), \qquad \text{for all} \quad i \in \{1, \ldots, m\},$$

where m is the number of model parameters. Thus a confidence interval $(\mathcal{I}_n)_i$ for the i-th unknown true model parameter $(\theta^*)_i$ with approximate confidence level γ can be constructed as:

$$(\mathcal{I}_n)_i := [(\hat{\theta}_n)_i - (\alpha_n)_i, (\hat{\theta}_n)_i + (\alpha_n)_i], \qquad \text{with} \quad (\alpha_n)_i := q\left(\tfrac{1+\gamma}{2}, n-m\right) \sqrt{(\hat{\mathcal{V}}_n)_{ii}}, \tag{10}$$

(cf. (Seber and Wild, 2003, Section 5.1) and (Smith, 2013, Section 7.3)) where $q(\beta, k)$ denotes the β percentile of the t-distribution with k degrees of freedom. Typical values are listed in Table 1.

$\gamma \setminus k$	10	10^2	10^3	10^4	10^5
90%	1.812	1.660	1.646	1.645	1.645
95%	2.228	1.984	1.962	1.960	1.960
98%	2.764	2.364	2.330	2.327	2.326
99%	3.169	2.626	2.581	2.576	2.576

Table 1. Typical values for $q\left(\frac{1+\gamma}{2}, k\right)$ rounded to three decimal places.

The justification of (10) is that:

$$\frac{(\Theta_n)_i - (\theta^*)_i}{\sqrt{(\mathcal{V}_n)_{ii}}} \sim t_{n-m},$$

(cf. (Seber and Wild, 2003, Section 5.1)) where t_{n-m} is the t-distribution with $n-m$ degrees of freedom.

The advantage of the previously described approach to quantify the uncertainty is that it is calculable without much computational effort, provided that the associated derivatives can be evaluated, at least approximately, without too much effort.

Another option to quantify the uncertainty regarding the model parameters are Monte Carlo simulations (cf. (Walter and Pronzato, 1997, Section 5.2)). Here fictitious measurement data vectors are generated several times and each time the resulting model parameters are estimated. From these estimates, confidence intervals for the unknown true model parameters p could be calculated as well as statistical properties, like the expected value or the covariance matrix, of the estimator Θ_n.

The fictitious measurement data vectors can be generated using sampling methods like random sampling or Latin hypercube sampling as well as resampling methods like jackknife of bootstrap methods.

This Monte Carlo approach provides more accurate results than the previously described approximations if the number of fictitious measurement data vectors is large. However, the computational effort using this approach is enormous in comparison to the described above because a parameter estimation has to be performed several times. Hence, it is not applicable to our computational expensive model.

2.3 Uncertainty in Model Output

The uncertainty in the model parameters implies an uncertainty in the model output. This can be quantified in the same ways as the uncertainty in the model parameters. First a probability distribution of the corresponding random vector and then confidence intervals are approximated.

The uncertainty can be considered on the whole model output or only at some points of interest. Let $\tilde{f}$ denote the function that maps the model parameters to the model output whose uncertainty should be quantified. This can be the hole model output or only a subset.

The probability distribution of $\tilde{f}(\Theta_n)$ can then be used to describe the uncertainty in the model output due to the uncertainty in the model parameters. It can be approximated by:

$$\tilde{f}(\Theta_n) \sim \mathcal{N}(\tilde{f}(\theta^*), \mathcal{W}_n(\hat{\theta}_n)), \tag{11}$$

with

$$\mathcal{W}_n(\hat{\theta}_n) := J_{\tilde{f}}(\hat{\theta}_n)\mathcal{V}_n(\hat{\theta}_n)J_{\tilde{f}}(\hat{\theta}_n)^{\mathsf{T}}. \tag{12}$$

where $J_{\tilde{f}}(\hat{\theta}_n)$ is the Jacobian matrix of $\tilde{f}$ evaluated at $\hat{\theta}_n$ and $\mathcal{V}_n(\hat{\theta}_n)$ is an approximation of the covariance matrix of Θ_n.

The approximations (11) and (12) can be derived by assuming $\Theta_n \sim \mathcal{N}\left(\theta^*, \mathcal{V}_n(\hat{\theta}_n)\right)$ and calculating the probability distribution of $\tilde{l}(\Theta_n)$ where $\tilde{l}$ is a linearization of $\tilde{f}$. Another justification is the delta method (cf. (Tenorio, 2017, Theorem 2.27))

which allows to calculate the asymptotic probability distribution of $\tilde{f}(\Theta_n)$ if the asymptotic probability distribution of Θ_n is known.

Several approximations of the covariance matrix of Θ_n, namely $\mathcal{V}_n^{(F,H)}(\hat{\theta}_n)$, $\mathcal{V}_n^{(F)}(\hat{\theta}_n)$ and $\mathcal{V}_n^{(H)}(\hat{\theta}_n)$, were introduced in the previous subsection. Define $\mathcal{W}_n^{(F,H)}(\hat{\theta}_n)$, $\mathcal{W}_n^{(F)}(\hat{\theta}_n)$ and $\mathcal{W}_n^{(H)}(\hat{\theta}_n)$ as described in (12) using $\mathcal{V}_n^{(F,H)}(\hat{\theta}_n)$, $\mathcal{V}_n^{(F)}(\hat{\theta}_n)$ and $\mathcal{V}_n^{(H)}(\hat{\theta}_n)$, respectively. $\mathcal{W}_n^{(F,H)}(\hat{\theta}_n)$ is a good choice if the assumed statistical model might be incorrect. $\mathcal{W}_n^{(F)}(\hat{\theta}_n)$ and $\mathcal{W}_n^{(H)}(\hat{\theta}_n)$ are also reasonable if the assumed statistical model (3) is correct.

Looking at a single point of interest, (11) implies:

$$(\tilde{f}(\Theta_n))_i \sim \mathcal{N}\left((\tilde{f}(\theta^*))_i, (\mathcal{W}_n(\hat{\theta}_n))_{ii}\right). \tag{13}$$

Thus a confidence interval $\tilde{\mathcal{I}}_i$ for $(\tilde{f}(\Theta_n))_i$ with approximate confidence level γ can be constructed, in the same way as in the previous subsection, as:

$$(\tilde{\mathcal{I}}_n)_i := [(\tilde{f}(\hat{\theta}_n))_i - (\tilde{\alpha}_n)_i, (\tilde{f}(\hat{\theta}_n))_i + (\tilde{\alpha}_n)_i] \qquad \text{with} \quad (\tilde{\alpha}_n)_i := q\left(\tfrac{1+\gamma}{2}, n-m\right)\sqrt{(\mathcal{W}_n)_{ii}}. \tag{14}$$

Again, the advantage of these approximation is that they are calculable without much computational effort.

Instead of this approximation, Monte Carlo simulations could again be used to quantify the uncertainty. This time several model parameter vectors have to be generated from the assumed probability distribution of Θ_n. For each of this model parameter vectors, the model output at the points of interest have to be evaluated. From these model evaluations, confidence intervals could be calculated as well as statistical properties of $\tilde{f}(\Theta_n)$, like its expected value or covariance matrix.

Again, this approach provides more accurate results than the approximation (11) and (12) if the number of generated model parameter vectors is large but the computational effort is extensive compared to these approximations. Hence, it is not applicable here.

2.4 Uncertainty Reduction using Optimal Experimental Design Methods

The uncertainty in the model parameters as well as the model output can be reduced by additional measurements. However, not all measurements reduce the uncertainty equally. The idea of optimal experimental design methods (cf. Pukelsheim (2006), (Walter and Pronzato, 1997, Chapter 6) and (Seber and Wild, 2003, Subsection 5.13)) is to design the measurements such that the resulting uncertainty is minimized and, hence, the information gain is maximized.

The design of a measurement includes everything that characterizes the measurement, involving the place and time of the measurement. If several different processes are modeled, it also includes which process is measured. Furthermore, multiple measuring techniques might be choosable which might result in different measurement accuracies.

One of the key observations for optimal experimental design methods is that for a given $\hat{\theta}_n$, the actual measurement results are not needed for the calculation of $\mathcal{V}_n^{(F)}(\hat{\theta}_n)$. Hence, it can also be calculated including planned measurements that have not yet been carried out. The same applies to $\mathcal{W}_n^{(F)}(\hat{\theta}_n)$. Thus the new uncertainty resulting from additional measurements can be predicted with these values.

Using $\mathcal{V}_n^{(F)}(\hat{\theta}_n)$ and $\mathcal{W}_n^{(F)}(\hat{\theta}_n)$ is justified if the assumed statistical model (3) is correct. Otherwise they may not be consistent estimations of the corresponding covariance matrices. Nevertheless, they can be used under certain regularity condi-

tions (cf. Pazman and Pronzato (2006)) to assess measurement designs. $\mathcal{V}_n^{(H)}(\hat{\theta}_n)$ and $\mathcal{V}_n^{(F,H)}(\hat{\theta}_n)$ as well as $\mathcal{W}_n^{(F)}(\hat{\theta}_n)$ and $\mathcal{W}_n^{(F,H)}(\hat{\theta}_n)$ can not be used to predict the uncertainty reduction because they depend on the measurement results.

To compare the uncertainty or the information gain resulting from different measurement designs criteria (cf. (Pukelsheim, 2006, Chapter 5), (Walter and Pronzato, 1997, Section 6.1) and (Pronzato and Pázman, 2013, Chapter 5)) are established. These criteria quantify the uncertainty with a single value by mapping covariance matrices to scalar values. Typical design criteria are the sum of the diagonal values, the determinant and the maximal eigenvalue (cf. (Pukelsheim, 2006, Chapter 6), (Walter and Pronzato, 1997, Section 6.1) and (Pronzato and Pázman, 2013, Subsection 5.1.2)). The lower the values of these criteria are, the stronger the measurements would reduce the uncertainty.

The choice of an appropriate design criterion depends on the purpose of the additional measurements. In particular, whether the uncertainty in the model parameters or in the model outputs should be reduced and how much emphasis is placed on the reduction of individual model parameters or model outputs.

We have used two different design criteria. They are easy to calculate and to interpret. The first one aims at reducing uncertainty in the model parameters itself and is defined as:

$$\psi(\mathcal{V}_n^{(F)}(\hat{\theta}_n), \hat{\theta}_n) := \frac{1}{m} \sum_{k=1}^{m} \frac{\sqrt{(\mathcal{V}_n^{(F)}(\hat{\theta}_n))_{ii}}}{(\hat{\theta}_n)_i}. \tag{15}$$

This is the average of the relative uncertainty in each model parameter, quantified by the standard deviation of the corresponding estimator divided by the parameter value. Designs are therefore preferred which evenly reduce the uncertainty in each of the model parameters. The average of the absolute uncertainties, i.e., the average of the uncertainties not divided by the parameter values, is less useful, if typical model parameters are of different orders of magnitude.

The second design criteria is:

$$\psi_{\tilde{f}}(\mathcal{W}_n^{(F)}(\hat{\theta}_n), \hat{\theta}_n) := \frac{1}{l} \sum_{k=1}^{l} \left(\sum_{i \in I_k} \sqrt{(\mathcal{W}_n^{(F)}(\hat{\theta}_n))_{ii}} \right) \left(\sum_{i \in I_k} \tilde{f}_i(\hat{\theta}_n) \right)^{-1} \tag{16}$$

where I_k is the set of indices corresponding to the output of the k-th modeled process of the numbered l modeled processes. This criterion prefers designs which reduce the uncertainty in the model output at the points of interest evenly over all modeled processes.

Again, the absolute uncertainty might be less useful if the typical total model output for each process and thus its typical total absolute uncertainty differs by several orders of magnitude. The uncertainty relative to each individual model output is not useful either if some model outputs are zero or close to zero.

It should be straight forward to modify the criteria to the specific purpose of the additional measurements or to formulate new ones specially suited. Designs that minimize the criterion among all feasible designs are called (local) optimal designs. Local refers to the dependency on the parameter estimate $\hat{\theta}_n$.

The information gain by additional measurements can be quantified by subtracting the value of the criteria using only the previous designs with the value of the criteria using the previous and the additional designs.

8

Sometimes, it might be useful to assign a cost value to each design that quantifies the financial cost or the time effort associated with this measurement, so the predicted information gains relative to their costs can be considered. This allows to define optimal designs in relation to their costs or to choose designs up to a certain cost limit.

After carrying out the chosen additional measurements, their measurement results should be used together with the previous measurement results to update the estimate of the model parameters. Using this new estimate new measurements could be designed. This allows to include the information in the previous measurements in the planning of the next measurements. This iterative process is called sequential optimal experimental design (cf. (Walter and Pronzato, 1997, Subsection 6.4.2) and (Seber and Wild, 2003, Subsection 5.13.3)) and is particularly suitable if new measurements have to be planned repeatedly.

2.5 Computational Details

Several computational details regarding the estimation of the model parameters, as described in Subsection 2.1, are summarized in the following subsection.

Optimization Algorithm

A number of optimization algorithms exist which can be used to calculate the model parameter estimate $\hat{\theta}_n$ by minimizing the objective function ϕ_n. They can basically be divided into two categories: derivative based (cf. Gill et al. (1981) and Nocedal and Wright (2006)) and derivative free algorithms (cf. Conn et al. (2009) and Rios and Sahinidis (2013)).

Usually derivative based optimization algorithms need fewer function evaluations to find a local minimum compared to derivative free optimization algorithms. However, they usually have more difficulties finding a global minimum. We try to take advantage of the rapid convergence of the derivative based optimization algorithm SQP discussed in (Nocedal and Wright, 2006, Chapter 18) and try to avoid its difficulty with finding global minimum by combining it with the globalization algorithm OQNLP introduced in Ugray et al. (2007).

This OQNLP algorithm finds the minimizer by starting multiple local minimizations from promising start points. To generate start points, a scatter-search algorithm similar to that described in Glover (1998) is used. Thereafter, iteratively, local minima are searched by a local optimization algorithm from one of the start points. After each search, unpromising start points are removed from the set of start points by considering their value of the objective function and their distance to already found local minima. The algorithm terminates if all start points are used or removed. The local minimum with the lowest objective value is then identified as global minimum. This OQNLP algorithm is implemented in MATLAB (cf. MathWorks (2015a)) as GlobalSearch algorithm in the Global Optimization Toolbox (cf. (MathWorks, 2015b, Chapter 3)).

This SQP algorithm iteratively solves the Karush–Kuhn–Tucker (KKT) equations, introduced in Karush (1939) and Kuhn and Tucker (1951), of the constrained optimization problem. For this purpose, a constrained quadratic subproblem is solved in each iteration using an active set strategy like described in Gill et al. (1981, 1991). The solution of the quadratic subproblem is used as search direction for a line search procedure similar to that described in Han (1977), Powell (1978b) and Powell (1978a). The quadratic subproblem is formulated using the value of the objective function and its first derivative as well as an approximation of its second derivative. The BFGS method, developed by Broyden (1970), Fletcher (1970), Goldfarb (1970)

and Shanno (1970), is used as a quasi-Newton update for this approximation together with an correction technique, described in Powell (1978b), which keeps the approximated Hessian positive define. The SQP algorithm is implemented in MATLAB as fmincon algorithm in the Optimization Toolbox (cf. (MathWorks, 2015c, Chapter 6)).

Scaling of the Objective Function

Many optimization algorithms, like the one we have used, are not invariant to scaling therefore it is essential to scale the objective function (cf. (Gill et al., 1981, Section 7.5 and 8.7), (Smith, 2013, Section 7.3), (Nocedal and Wright, 2006, Section 2.2) and (Dennis and Schnabel, 1996, Section 7.1)) for a fast and accurate determination of a minimum.

Hence, for the estimation of the model parameters, the model parameters in the objective function and the objective function values are scaled, as described in (Gill et al., 1981, Section 7.5 and 8.7). The scaled parameters typically range from -1 to 1 and the objective function values typically be around 1.

Evaluating of the Objective Function

The objective function of the generalized least squares estimator (2) can be evaluated in many different ways. In the following, we describe a fast and numerically accurate way.

The objective function value that should be evaluated is:

$$\phi(p) = (y - f(p))^\mathsf{T} \mathcal{C}^{-1} (y - f(p)).$$

We have omitted the index n for the sake of simplicity. Define $S := \mathrm{diag}(\mathcal{C})^{0.5}$ to be the diagonal matrix containing the square root of the diagonal values of $\mathcal{C}$ and $B := S^{-1} A S^{-1}$.

Decompose B by LDL^T representation, meaning a lower triangle matrix L with ones on the diagonal and diagonal matrices D with positive values. It is important to notice that this has to be done only once and not for every evaluation of the objective function.

The objective function ϕ is then evaluated by first evaluating:

$$\psi(p) := D^{-0.5} L^{-1} S^{-1} (y - f(p)),$$

from right to left, where instead of the inverse of L the corresponding linear equation is solved using forward substitution. Then objective function value is evaluated by:

$$\phi(p) = \psi(p)^\mathsf{T} \psi(p).$$

3 Marine Phosphorus Cycle as Application Example

We use a model for the phosphate and dissolved organic phosphorus concentrations in the global ocean as application example for the parameter estimation, uncertainty quantification and experimental design methods described in Section 2.

First, the used circulation model and the biogeochemical model are introduced in this section. Then, the model parameters are described together with different guesses of their values. Next, the calculation of an annual periodic state is explained as well as a fast way to calculate the derivative of the model output regarding the model parameters. Finally, the measurement data used for parameter estimation are described.

3.1 Circulation Model

We have used the Transport Matrix Method (TMM) introduced in Khatiwala et al. (2005) to simulate the advection and diffusion of passive tracers in the ocean (cf. Khatiwala (2007)). This method has already been used in various studies (cf. Weber and Deutsch (2010); Kriest et al. (2010, 2012); Prieß et al. (2013); Graven et al. (2012)).

The TMM utilizes that the continuous advection-diffusion equation:

$$\frac{\partial Y_i}{\partial t} = \underbrace{\nabla \cdot (K \nabla \cdot Y_i))}_{\text{diffusion}} - \underbrace{\nabla \cdot (V Y_i)}_{\text{advection}} + \underbrace{S_i(Y_1, \ldots, Y_m, \theta)}_{\text{sources and sinks}} \text{ for } i \in 1, \ldots, m, \tag{17}$$

where $Y_1, \ldots, Y_m$ denote the concentrations of the m tracers, K the diffusion coefficient, V the velocity and $S_1, \ldots, S_m$ the source-and-sink-terms depending on the model parameters θ, can be written in the discretized form as a matrix equation:

$$y^{(n+1)} = A_i^{(n)}(A_e^{(n)} y^{(n)} + s^{(n)}(\theta)\Delta t). \tag{18}$$

$y^{(n)}$ denotes the vector of all tracer concentrations at all grid points of the circulation model in the discretized form at time step n, $s^{(n)}(\theta)$ the discretized version of the source-and-sink-terms depending on the model parameters θ and Δt the time step in the discretization. The matrices $A_i^{(n)}$ and $A_e^{(n)}$, called transport matrices, result from the discretization of the advection and diffusion terms where $A_i^{(n)}$ belongs to the implicit part and $A_e^{(n)}$ to the explicit part of the discretization.

The approach of the TMM is to determine the elements of these matrices by utilizing a general circulation model. For this, the general circulation model is executed several times with different suitable chosen tracer concentrations.

We use monthly averaged transport matrices (cf. Khatiwala et al. (2005); Khatiwala (2007)), calculated with the MIT general circulation model (cf. Marshall et al. (1997a, b, 1998)). At the middle of each month the corresponding transport matrix has been used. Elsewhere a linear interpolation of the two transport matrix closest to the point in time were used.

A spatial resolution of 2.8125 degree and 15 vertical layers with increasing depths was used at the construction of the transport matrices. Hence, this is also the resolution of our circulation model. The resolution corresponds to 64 boxes in north-south direction and 128 boxes in west-east direction.

For the temporal resolution, $\Delta t = 2880^{-1}$ y has been chosen which corresponds to a time step of roughly three hours. Hence, daytime dependent processes can be resolved.

3.2 Biogeochemical Model

The biogeochemical model contains phosphate (PO_4) and dissolved organic phosphorus (DOP) and is part of the ocean carbon model, described in Dutkiewicz et al. (2005), of the MIT Integrated Global System Model Version 2 (IGSM2), described in

Sokolov et al. (2005). This model and some variants are used frequently(cf. Parekh et al. (2005, 2006); Najjar and Orr (1998); Najjar et al. (2007); Kwon et al. (2009); Kriest et al. (2010, 2012); Prieß et al. (2013)). It is briefly described in the following where we stick to the notation in Dutkiewicz et al. (2005).

The concentration of PO_4 and DOP at layer i are described by the following source-minus-sink terms:

$$S_{PO_4}(i) = -J_{prod}(i) + \kappa_{re} DOP(i) + \Delta F(i), \tag{19}$$

$$S_{DOP}(i) = f_{DOP} J_{prod}(i) - \kappa_{re} DOP(i). \tag{20}$$

Here, J_{prod} denote the biological production (net community productivity). A fraction f_{DOP} of this biological production remains suspended as DOP. The remainder $(1 - f_{DOP})$ becomes particulate organic phosphorus (POP) which sinks to depths and instantly remineralizes to PO_4 which is modeled by $\Delta F(i)$. The DOP remineralizes back to PO_4 with rate κ_{re}. f_{DOP} and κ_{re} are model parameters.

The biological production:

$$J_{prod}(i) := \alpha \frac{PO_4(i)}{PO_4(i) + \kappa_{PO_4}} \frac{I(i)}{I(i) + \kappa_I}, \tag{21}$$

is modeled by Michaelis-Menten kinetics depending on the available light I and the nutrient PO_4 similar to McKinley et al. (2004). The corresponding half saturation constants κ_{PO_4} and κ_I are model parameters as well as the maximum community production rate α.

The available light:

$$I(i) := f_{PAR} Q_{SW} e^{-k z_c(i)}, \tag{22}$$

is modeled, as that portion of the short wave radiation Q_{SW} that is photo-synthetically available and has not been attenuated by water. The short wave radiation Q_{SW} is calculated by the atmosphere component of the IGSM2 as a function of time, latitude and ice cover (cf. Paltridge and Platt (1976) and Brock (1981)). The light attenuation coefficient of water k is treated as model parameter.

The fraction of photo-synthetically available radiation is described by f_{PAR}. It only enters into the biological production J_{prod} where only the ratio $\frac{\kappa_I}{f_{PAR}}$ is relevant. Due to this linear dependence, f_{PAR} and κ_I would not be uniquely identifiable if f_{PAR} would be a model parameter as well. For this reason, f_{PAR} is set constant to $f_{PAR} := 0.4$. This values is also used in the IGSM2.

Let n be the numbers of layers. For each layer i, let $z_t(i)$, $z_c(i)$ and $z_b(i)$ be its top, centered and bottom depth, respectively and $\Delta z(i) := z_b(i) - z_t(i)$ its thickness.

The portion of the biological production which is exported as POP from layer i to deeper layers is denoted by:

$$E(i) := (1 - f_{DOP}) J_{prod}(i) \Delta z(i). \tag{23}$$

It is assumed that the sinking speed increases with depth following a power law relationship (cf. Najjar and Orr (1998)) and that the exported POP instantly remineralizes to PO_4. The flux $F(i)$ into layer $i \geq 2$ is then modeled as follows, where i_e is the

last layer in the euphotic zone:

$$F(i) := \sum_{j=1}^{\min(i_e, i-1)} E(i) \left(\frac{z_b(i-1)}{z_b(j)} \right)^{-a_{re}}. \tag{24}$$

The change of the PO_4 concentration in layer $1 < i < n$ due to the flux is then:

$$\Delta F(i) := \sum_{j=1}^{\min(i_e, i-1)} E(i) \left(\left(\frac{z_b(i-1)}{z_b(j)} \right)^{-a_{re}} - \left(\frac{z_b(i)}{z_b(j)} \right)^{-a_{re}} \right) (\Delta z(i))^{-1}. \tag{25}$$

It is also assumed that no POP is lost to the sediment. This means, all POP that enters the deepest box is instant remineralized:

$$\Delta F(n) := \sum_{j=1}^{\min(i_e, i-1)} E(i) \left(\frac{z_b(i-1)}{z_b(j)} \right)^{-a_{re}} (\Delta z(i))^{-1}. \tag{26}$$

In the topmost layer no PO_4 arise from sunk and remineralized POP:

$$\Delta F(1) := 0. \tag{27}$$

3.3 Model Parameters

The seven parameters of the biogeochemical model, described in Subsection 3.2, are considered as unknown model parameters. Furthermore, the global average phosphorus concentration, which is used to spin-up the model into annual periodic concentrations as described in Section 3.4, is considered as an unknown model parameter as well. All these model parameters are listed in Table 2 and their values shall be estimated.

Parameter	Description	Unit
κ_{re}	remineralization rate of DOP	y^{-1}
α	maximum community production rate	$\mathrm{mmol\,m^{-3}\,y^{-1}}$
f_{DOP}	fraction new production going to DOP	-
κ_{PO_4}	half saturation constant of PO_4	$\mathrm{mmol\,m^{-3}}$
κ_I	half saturation constant of light	$\mathrm{W\,m^{-2}}$
k	light attenuation coefficient of water	m^{-1}
a_{re}	power law remineralization coefficient	-
p	average phosphorus concentration	$\mathrm{mmol\,m^{-3}}$

Table 2. Parameters of the marine phosphorus cycle model.

Our initial guesses and bounds for the unknown model parameters are summarized in Table 3. They are based on values used in other publications which are outlined next.

	κ_{re}	α	f_{DOP}	κ_{PO_4}	κ_I	k	a_{re}	p
initial guess	0.5	2	0.67	0.5	30	0.02	0.86	2.17
lower bound	0.05	0.2	0.05	0.01	5	0.001	0.5	0.4
upper bound	10	20	0.95	10	200	0.2	2	10

Table 3. Bounds and initial guesses for model parameters.

For κ_{re}, 2 y^{-1} was used in Najjar and Orr (1998) and in Dutkiewicz et al. (2005) based on Najjar and Orr (1998). 0.5 y^{-1} was used in Parekh et al. (2005) and in Kriest et al. (2010) based on Parekh et al. (2005). In Najjar and Orr (1998) different studies are summarized which had suggested that $\kappa_{re} \in [\frac{10}{7}, 5]\ \mathrm{y}^{-1}$.

3 $\mathrm{mmol\,m^{-3}\,y^{-1}}$ was used in Dutkiewicz et al. (2005) for α, 2 $\mathrm{mmol\,m^{-3}\,y^{-1}}$ in Kriest et al. (2010) and 6 $\mathrm{mmol\,m^{-3}\,y^{-1}}$ in Parekh et al. (2005). In McKinley et al. (2004), different values were used for different ocean regions with 2.5 $\mathrm{mmol\,m^{-3}\,y^{-1}}$ as average value.

$f_{DOP} \in [0, 1]$ by definition of f_{DOP}. 0.67 was used in Dutkiewicz et al. (2005), Kriest et al. (2010), Najjar and Orr (1998) and Parekh et al. (2005) all based on Yamanaka and Tajika (1997). 0.7 was suggested in Platt et al. (1989). Different studies are cited in Najjar and Orr (1998) which had estimated f_{DOP} in $[0.58, 0.77]$, $[0.65, 0.95]$, $[0.6, 0.7]$ or $[0.4, 0.8]$.

For κ_{PO_4}, 0.5 $\mathrm{mmol\,m^{-3}}$ was used in Dutkiewicz et al. (2005) and Kriest et al. (2010) and 0.01 $\mathrm{mmol\,m^{-3}}$ in McKinley et al. (2004).

25 $\mathrm{W\,m^{-2}}$ was used in Dutkiewicz et al. (2005) for κ_I and 30 $\mathrm{W\,m^{-2}}$ in Dutkiewicz et al. (2001), Kriest et al. (2010), McKinley et al. (2004) and Parekh et al. (2005). In Dutkiewicz et al. (2001), it was stated that κ_I varies from 5 $\mathrm{W\,m^{-2}}$ to 100 $\mathrm{W\,m^{-2}}$ for different species of phytoplankton based on several cited studies.

For k, 0.02 m^{-1} was used in Dutkiewicz et al. (2005) and Kriest et al. (2010).

0.9 was used in Dutkiewicz et al. (2005) for a_{re} based on Yamanaka and Tajika (1997) and Sarmiento et al. (1990). 0.858 was used in Martin et al. (1987) and in Kriest et al. (2010) based on Martin et al. (1987). Since the choice of a_{re} is closely related to $z_b(i_e)$, the depth of the euphotic zone, its common values are presented as well. 100 m was chosen in Yamanaka and Tajika (1997), Martin et al. (1987) and Maier-Reimer (1993) and 75 m in Najjar and Orr (1998). 120 m was used in Kriest et al. (2010) and 130 m in McKinley et al. (2004). We selected 120 m as well.

2.1701 $\mathrm{mmol\,m^{-3}}$ was used in Kriest et al. (2010) for p. 2.17 $\mathrm{mmol\,m^{-3}}$ is also the average phosphorus concentration of the climatological data provided by the World Ocean Atlas 2013 Garcia et al. (2014) and Reimer (2019b) which are both based on the data of the World Ocean Database 2013 introduced in Boyer et al. (2013).

3.4 Simulation and Spin-up

The previously described model has been simulated using the simulation package (Reimer (2019c)) which is based on Python (Python Software Foundation (2018)), NumPy (Oliphant et al. (2019)), SciPy (Jones et al. (2019) and Virtanen et al. (2019)), Matplotlib (Caswell et al. (2019) and Hunter (2007)), utillib (Reimer (2019d)), the measurements software package (Reimer

(2019b)) and the matrix-decomposition library (Reimer (2019a) and Reimer (2019a)). The simulation package also includes many pre- and post-processing functions. For the actual parallelized evaluation of the model, it uses the simulation framework METOS3D (Piwonski and Slawig (2016), Piwonski and Slawig (2013)) which is based on PETSc (Portable, Extensible Toolkit for Scientific Computation) (Balay et al. (2019a), Balay et al. (2019b)).

For each model simulation, the model has been spun up from constant concentrations to annual periodic concentrations. These constant concentrations were chosen so that the average phosphorus concentration p was achieved. To check if an annual periodicity is reached, the concentrations at the beginning of two consecutive model years were compared. If these are equal, a periodic state is reached.

Usually, it took 5000 to 7500 model years until roughly annual periodic concentrations were achieved. Sometimes even more model years were needed. We used at most 10.000 model years. Thereafter, the average difference between concentrations at two consecutive model years was around 10^{-7}.

A model simulation with a spin-up of 10.000 model years has taken about four hours on four connected computer nodes with sixteen cores each and a clock rate of 2.1 GHz, respectively.

3.5 Derivative

Besides the model output itself, the derivatives of the model output regarding the model parameters are needed for the estimation of the model parameters as well as the uncertainty quantification and the design of additional measurements. We have approximated them using finite difference quotients (cf. (Abramowitz and Stegun, 1972, Section 25.3), (Dennis and Schnabel, 1996, Section 4.2), (Nocedal and Wright, 2006, Section 8.1) and (Gill et al., 1981, Section 8.6)). For this, appropriate finite difference quotients and step sizes must be selected.

Central finite difference quotients ((Abramowitz and Stegun, 1972, Equation 25.3.21) and (Gill et al., 1981, Subsubsection 8.6.1.2)) have been used for the first order partial derivatives. They have a second order approximation error and are, thus, very accurate with an appropriate step size. For the second order partial derivatives, finite difference quotients (Abramowitz and Stegun, 1972, Equation 25.3.23 and Equation 25.3.27), with a second order approximation error as well, have been used.

Two additional function evaluations are needed for each approximation of the first order partial derivative. If the same step size is used for approximating the second order partial derivatives, two more function evaluations are needed for the second order partial derivative regarding two different variables and no additional function evaluations are needed for the second order partial derivative regarding one variable.

To reduce the number of additional function evaluations, finite difference quotients with first order approximation error (cf. (Dennis and Schnabel, 1996, Section 4.2) and (Gill et al., 1981, Section 8.6)), like forward or backward finite difference quotients for the first order partial derivatives, could be used. However, in our application example, the additional function evaluations correspond only to a small part of the total computational effort, as explained below. For that reason, we use the more accurate finite difference quotients described in the previous paragraphs.

The choice of the step size in the finite difference quotients is always a compromise between a small error in replacing the derivative by the finite difference quotient and a small error in the floating point arithmetic. Recommended step sizes are

usually a constant, depending on the used finite difference quotient, multiplied by the typical magnitude of the model parameter (cf. (Dennis and Schnabel, 1996, Section 5.6) and (Nocedal and Wright, 2006, Section 8.1)). These constants are the third and the fourth root of the machine precision, for the first order and second order finite difference quotient, respectively (cf. (Gill et al., 1981, Subsection 8.6.1)). These are roughly 10^{-5} and 10^{-4} for 64 bit floating point numbers.

Larger and smaller step sizes were also tested. However, too strong deviation from the recommended step size, usually by more than two orders of magnitude, result in unrealistic values.

To evaluate the finite difference quotients, we have first spun up the model with the unchanged model parameters. Usually this spin-up is needed anyway. The spin-ups for the slightly changed model parameters in the finite difference quotients were then started with the annual periodic concentrations, obtained from the spin-up with the unchanged model parameters, instead of the usually used constant concentrations. For the derivative regarding the average phosphorus concentration, the annual periodic concentration were slightly modified to match the average phosphorus concentration.

Using the annual periodic concentrations from the spin-up with the unchanged model parameters accelerates the evaluation of the derivative significantly because much fewer model years are needed to achieve annual periodic concentrations for the slightly changed model parameters. Tests with different model parameters have shown that usually only a few hundred model years are needed. Thus, we have used at most 500 years for the spin-ups for the slightly changed model parameters. Hence, the complete evaluation of the first derivative with central finite differences needs at most 80% more computational effort than the evaluation of the model itself.

3.6 Measurement Data

We used the measurement data for phosphate provided by the World Ocean Database 2013, presented in Boyer et al. (2013) and Johnson et al. (2013), for the model parameter estimation. We limited ourselves to the data that have passed all quality checks (Johnson et al., 2013, Section 3) and where the measurement points are inside the computational domain. These were about 2.2 million measurements.

For dissolved organic phosphorus, generally far less measurement data were available. We used almost 400 measurements obtained from Landolfi (2005), Landolfi et al. (2008) and Yoshimura et al. (2007). These data were quality checked as well and implausible data were removed together with data outside of the computational domain.

The corresponding standard deviations and the correlation matrix were estimated as described in Reimer (2019b) using the spatial resolution described in Subsection 3.1 and a monthly temporal resolution.

Here, the standard deviation in each space-time grid box was estimated using the sample standard deviation in each grid box where at least four values are available. Otherwise the standard deviation was interpolated for phosphate. For dissolved organic phosphorus, the average of its estimated standard deviations was used, since too few data are available for a meaningful interpolation. Furthermore, we used 0.1 as a lower bound for the standard deviations. This corresponds to the usual accuracy of the measurement data and prevents a disproportional weighting of measurement results with a very small sample standard deviation.

The correlation between different space-time grid boxes was estimated using the sample correlation where at least thirty-five value pairs were available. Otherwise the correlation is assumed to be zero. From these individual estimates, a valid correlation matrix was calculated using the algorithm described in Reimer (2019a).

The objective of the algorithm is to find a valid correlation matrix which is close to the original matrix and has a low condition number. A low condition number is important because otherwise small inaccuracies by numerical methods or measurements, are amplified and could dominate the evaluation of the objective function. The algorithm has a parameter which controls the weighting between a small difference to the original matrix and a small condition number. We have chosen 0.1 as value for this parameter which makes both objectives quite well achieved. Furthermore, the algorithm calculates the LDL^T decomposition of the correlation matrix as byproduct which was used for a fast and accurate evaluation of the objective function as described in Subsection 2.5.

4 Results for the Application Example

We applied the methods for parameter estimation, uncertainty quantification and experimental design introduced in Section 2 to the model for phosphate and dissolved organic phosphorus concentrations introduced in Section 3. The results are presented in the following.

4.1 Model Parameter Estimation

We used the generalized least squares estimator, as described in Subsection 2.1, to estimate the model parameters based on the measurement data described in Subsection 3.6. For this, the objective function was evaluated over 30.000 times with different model parameters.

Different model parameters and their objective function values are presented in Table 4. The first row contains the initial guess of the model parameters presented in Subsection 3.3. The last three rows contain the model parameters which minimize the objective function of the generalized least squares estimator (GLS), the weighted least squares estimator (WLS) and the ordinary least squares estimator (OLS), respectively. The objective function values in the table have been divided by the number of measurements to obtain values easier to interpret.

GLS	WLS	OLS	κ_{re}	α	f_{DOP}	κ_{PO_4}	κ_I	k	a_{re}	p
1.74	4.30	0.22	0.5	2.0	0.67	0.50	30	0.020	0.86	2.17
1.22	2.70	0.20	3.6	11.4	0.83	0.19	154	0.010	1.53	2.17
1.23	**2.69**	0.20	4.7	10.2	0.88	0.14	100	0.011	1.48	2.19
1.29	2.88	**0.19**	5.9	18.2	0.89	0.14	200	0.011	1.26	2.20

Table 4. objective function values for different model parameters. (GLS: generalized least squares estimator , WLS: weighted least squares estimator, OLS: ordinary least squares estimator)

Depending on the estimator, the optimal model parameters vary. However, all are better than the initial guess regardless of which estimator is considered.

We focus, as before, on the generalized least squares estimator and the corresponding optimal model parameters. Some of them differ significantly compared to their initial guess. Most conspicuous is the high value of κ_I and thus the low value of k. However, the objective function is rather insensitive to changes in these two parameters. Therefore more plausible values can be chosen for these parameters without major losses. The value of α seems a bit high as well which and is the consequence of the high value of κ_I. The values of the other model parameters seem plausible. κ_{re}, f_{DOP} and a_{re} are slightly higher than the initial guess and κ_{PO_4} is slightly lower. p is equal to the initial guess which is reasonable since this parameter could already be estimated very well directly from the measurement data.

The optimization process has also shown that the objective function has many local minima that are not global minima and that its value in some cases changes very little for changes in the model parameters. This shows that it is very challenging to find a global minimum and that this may not be unique. Even if the model parameters found represent the measurement results better than the initial model parameters, they may nevertheless not be a global minimum.

If the statistical assumption (3), on which the generalized least squares estimator is based is correct, the estimator is χ^2 distributed with n degrees of freedom, where n is the number of measurements. This implies that the estimator divided by n has an expected value of one and a variance of $\frac{2}{n}$. A confidence interval for this normalized estimator with a confidence level of 99% is approximately $[0.998, 1.002]$. However, the obtained value is 1.22 indicating that the statistical assumption (3) is not precisely fulfilled or no global minimum was found. Nevertheless with a value of 1.22, the model parameters might be quite close to a global minimum.

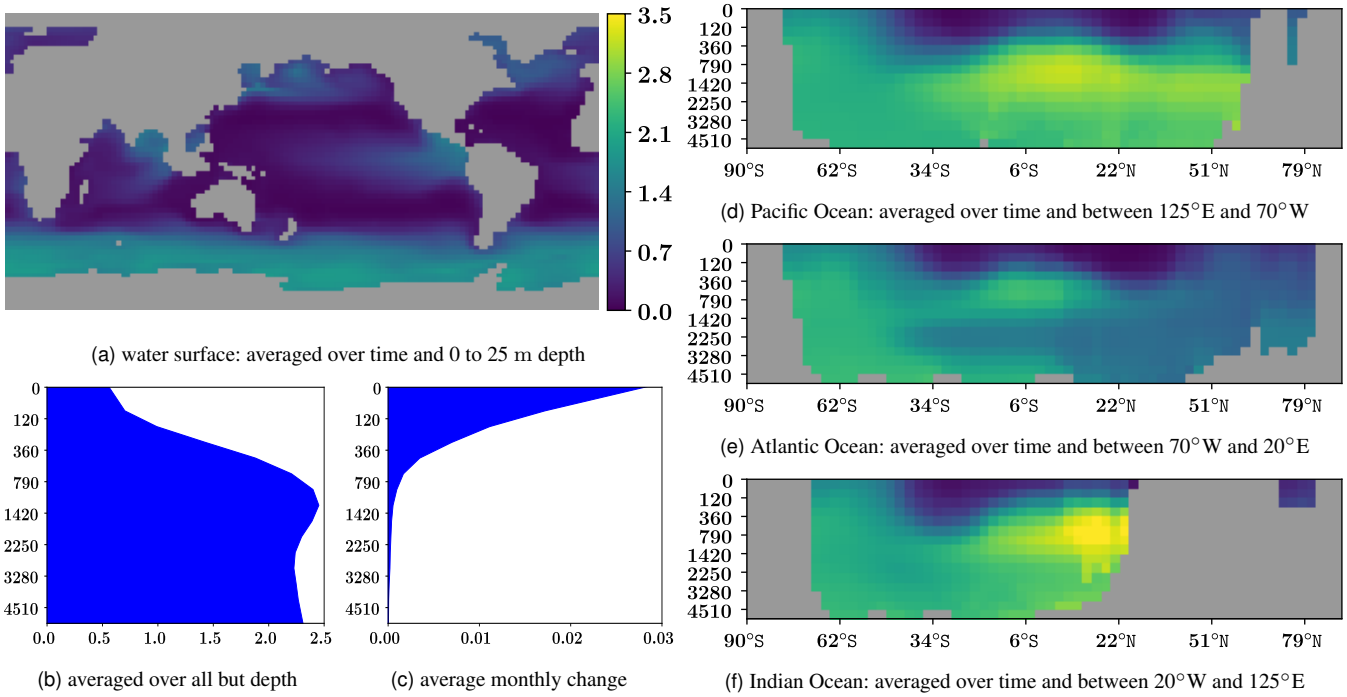
The model output with the optimal model parameters regarding the generalized least squares estimator (second row in Table 4) is summarized in Figure 1 and 2. The time averaged output at the water surface is plotted in Figure 1a and 2a. The average model output depending on the depth is shown in Figure 1b and 2b. The average absolute change after one month is plotted in Figure 1c and 2c. Figure 1d and 2d, 1e and 2e as well as 1f and 2f show the model output in the Pacific Ocean, the Atlantic Ocean and the Indian Ocean, respectively, depending on depth and latitude and averaged over time and between the corresponding longitudes.

The average phosphate concentration at the surface is roughly 0.6 $\mathrm{mmol\,m^{-3}}$. It increases with growing depth. Deeper than 700 meters the average is approximate constant 2.3 $\mathrm{mmol\,m^{-3}}$.

The temporal variability decreases with growing depth. The average monthly change of the concentrations is around 0.03 $\mathrm{mmol\,m^{-3}}$ at the surface. There are almost no changes over time deeper than 700 m.

At the water surface, the highest concentrations are at the Southern Ocean with around 2.2 $\mathrm{mmol\,m^{-3}}$ and at the north-eastern part of the Indian Ocean, the northern and middle-east part of the Pacific Ocean as well as the northern part of the Atlantic Ocean ranging from 1 to 2 $\mathrm{mmol\,m^{-3}}$.

The phosphate concentration is highest in each of the Pacific Ocean, the Atlantic Ocean and the Indian Ocean around the equator at a depth between 500 and 1500 meters. The lowest concentrations in each of these oceans is around the equator near the water surface.



(a) water surface: averaged over time and 0 to 25 m depth

(b) averaged over all but depth

(c) average monthly change

(d) Pacific Ocean: averaged over time and between 125°E and 70°W

(e) Atlantic Ocean: averaged over time and between 70°W and 20°E

(f) Indian Ocean: averaged over time and between 20°W and 125°E

Figure 1. Model output for phosphate (in $\mathrm{mmol\,m^{-3}}$) with model parameters estimated by GLS.

The average dissolved organic phosphorus concentration is almost 0.3 $\mathrm{mmol\,m^{-3}}$ at the water surface and decreases quickly with growing depth. It is close to zero below 500 m.

The temporal variability decreases rapidly as well with growing depth. The average monthly change of the concentrations is around 0.02 $\mathrm{mmol\,m^{-3}}$ at the surface and there are almost no changes over time below 500 m.

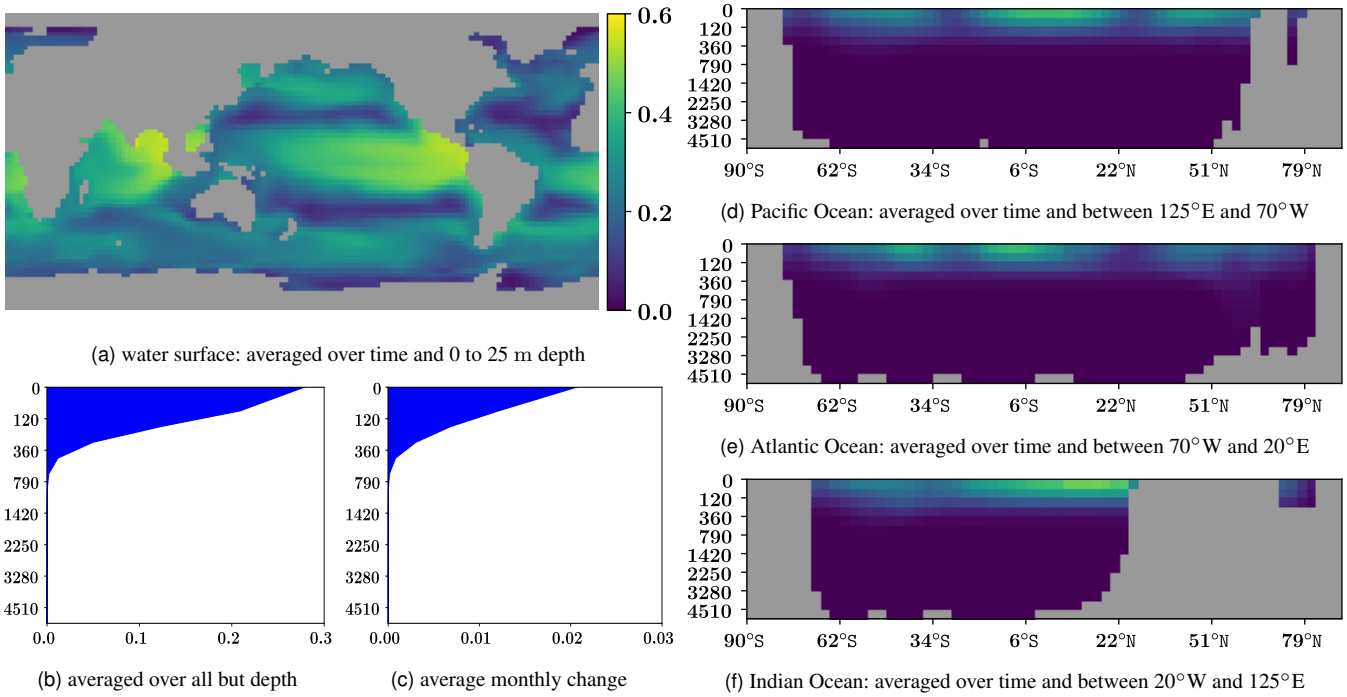
The highest dissolved organic phosphorus concentrations of almost 0.6 $\mathrm{mmol\,m^{-3}}$ are at the surface around the equator. Other high values with around 0.4 $\mathrm{mmol\,m^{-3}}$ are in areas around 45°S and 45°N.

The previously described behavior applies to the Pacific Ocean, the Atlantic Ocean as well as the Indian Ocean.

4.2 Uncertainty in Parameter Estimation

The uncertainty in the parameter estimation has been quantified as described in Subsection 2.2. For this, we have approximated the covariance matrix of the parameter estimator and confidence intervals with a confidence level of 99% using Equation (8) and (10).

For each model parameter, the length of its confidence interval relative to its estimated value is plotted in Figure 3. The estimates with the greatest uncertainty are those for κ_{re} and κ_{PO_4} with six to seven percent. These are followed by α and κ_I with around three percent. A lower uncertainty of about one percent is associated with f_{DOP}, k and a_{re}. The slightest uncertainty of one per mill is associated with p.



(a) water surface: averaged over time and 0 to 25 m depth

(b) averaged over all but depth

(c) average monthly change

(d) Pacific Ocean: averaged over time and between 125°E and 70°W

(e) Atlantic Ocean: averaged over time and between 70°W and 20°E

(f) Indian Ocean: averaged over time and between 20°W and 125°E

Figure 2. Model output for dissolved organic phosphorus (in $\mathrm{mmol\,m^{-3}}$) with model parameters estimated by GLS.

These values are consistent with our experience with the model. Its output is sensitive to changes in the parameters f_{DOP}, k and a_{re} and very sensitive to changes in p. Hence, it is reasonable that these parameter could be estimated quite accurately.

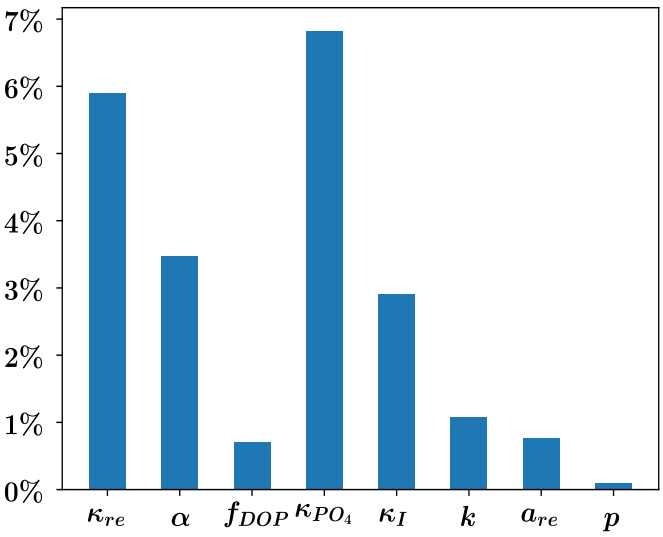


Figure 3. Confidence intervals length relative to estimated model parameters with 99 % confidence level.

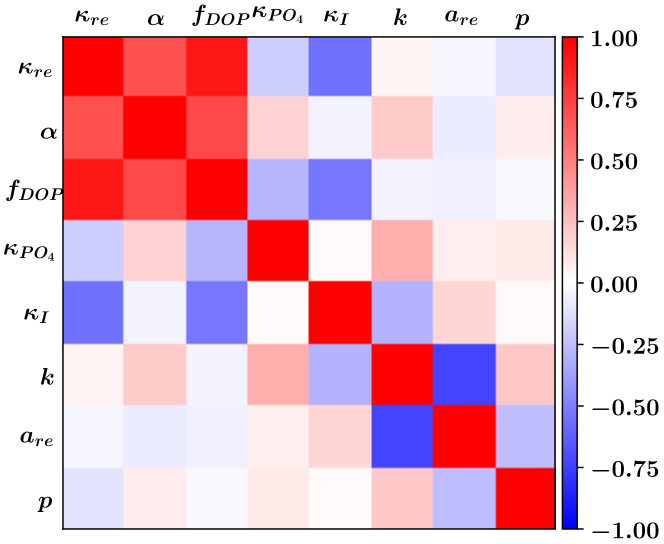
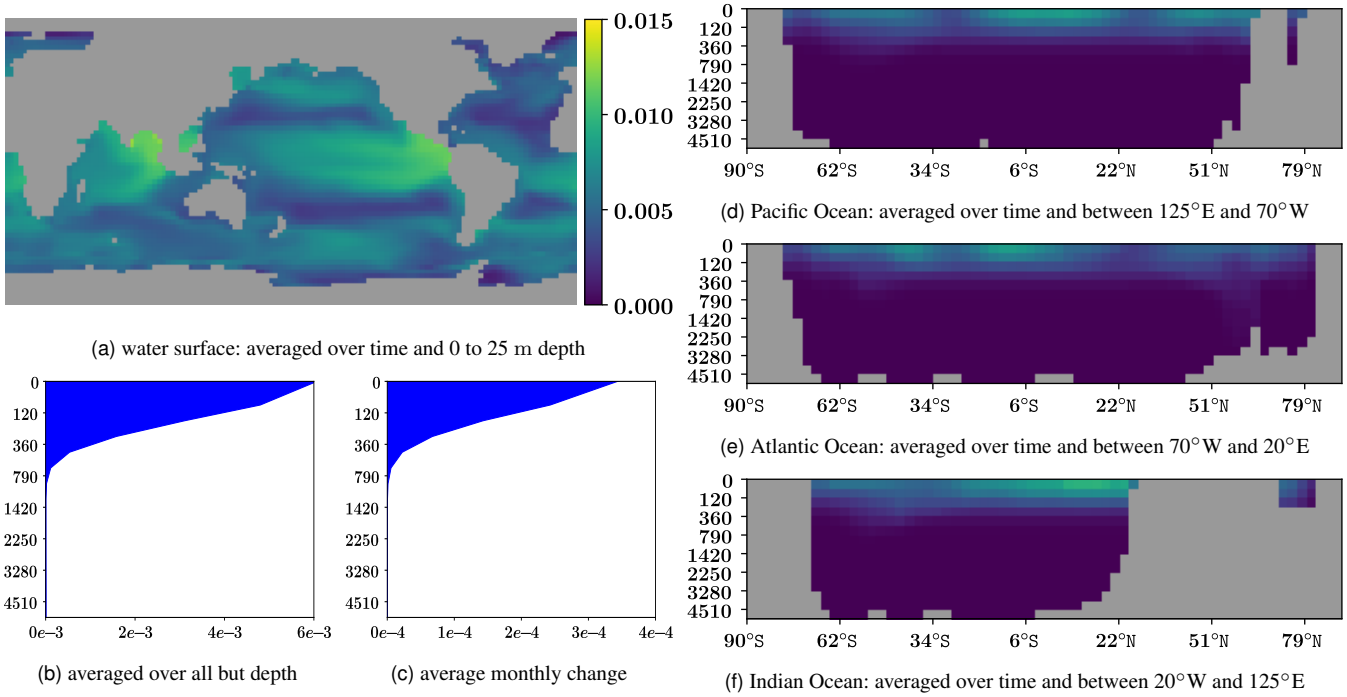


Figure 4. Correlation matrix of the model parameter estimator (generalized least squares estimator).



(a) water surface: averaged over time and 0 to 25 m depth

(b) averaged over all but depth

(c) average monthly change

(d) Pacific Ocean: averaged over time and between 125°E and 70°W

(e) Atlantic Ocean: averaged over time and between 70°W and 20°E

(f) Indian Ocean: averaged over time and between 20°W and 125°E

Figure 5. Confidence intervals length for phosphate model output (in $\mathrm{mmol\,m^{-3}}$) with 99 % confidence level.

The correlation matrix of the parameters estimator is plotted in Figure 4. Here strong positive correlations between κ_{re}, α and f_{DOP} are conspicuous. They imply that if the true value of one of these model parameters is higher or lower than its estimate, it is very likely that the same applies to the other two parameters. Especially κ_{re} and f_{DOP} have a correlation close to one.

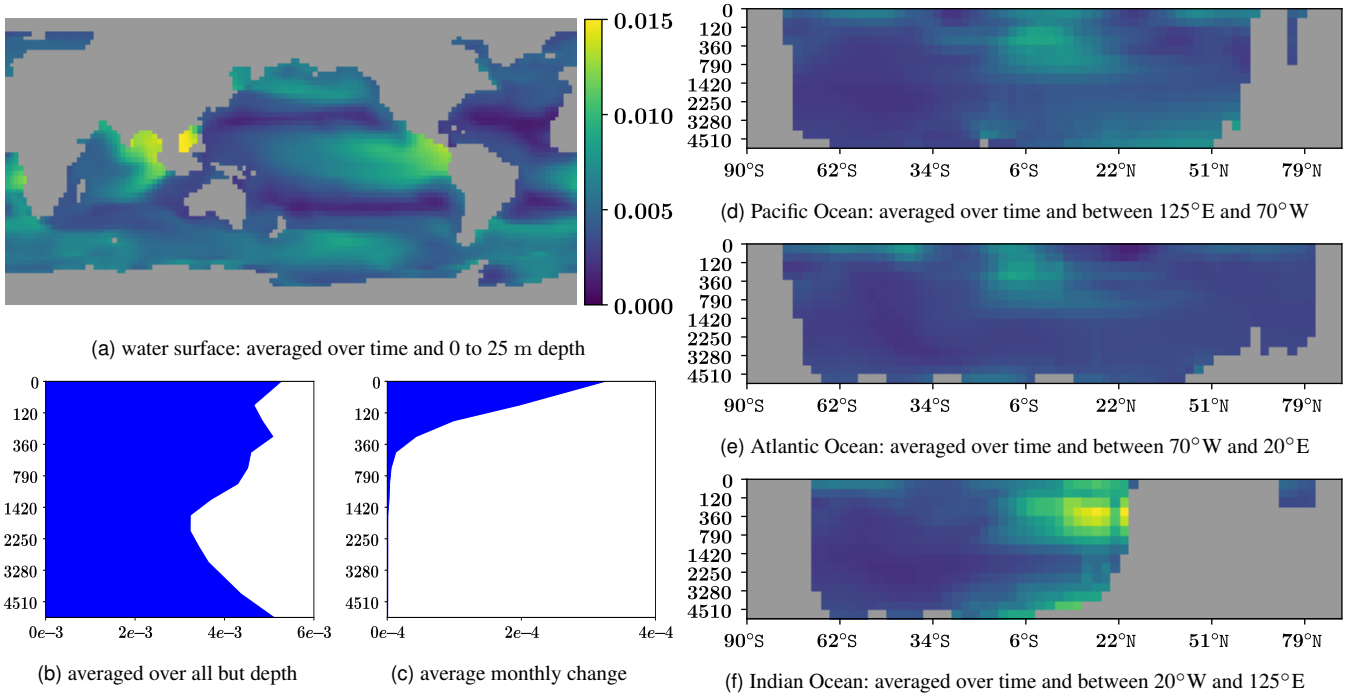
A strong negative correlation close to minus one is between k and a_{re}. This means that if the true value of one of these parameters is greater than its estimate, then it is very likely that the other is smaller and vice versa.

We also compared the different approaches to approximated the covariance matrix of the parameter estimator described in Equation (6), (7) and (8). All three approximations provide similar results. They differ in each component usually at most by a factor between one half and two. The similarity of the three approximations indicates that the statistical assumption (3) might be not far from reality.

4.3 Uncertainty in Model Output

The uncertainty in the model parameters implies uncertainty in the model output. This has been quantified as described in Subsection 2.3. For each model output the uncertainty is quantified by the length of corresponding confidence intervals with confidence level of approximately 99 %. Their lengths are plotted in Figure 5 and 6.

The average uncertainty at the water surface is 6×10^{-3} $\mathrm{mmol\,m^{-3}}$ for phosphate and 5×10^{-3} $\mathrm{mmol\,m^{-3}}$ for dissolved organic phosphorus. This corresponds to an uncertainty relative to the average model output of around 1 % for phosphate and



(a) water surface: averaged over time and 0 to 25 m depth

(b) averaged over all but depth

(c) average monthly change

(d) Pacific Ocean: averaged over time and between 125°E and 70°W

(e) Atlantic Ocean: averaged over time and between 70°W and 20°E

(f) Indian Ocean: averaged over time and between 20°W and 125°E

Figure 6. Confidence intervals length for dissolved organic phosphorus model output (in $\mathrm{mmol\,m^{-3}}$) with 99 % confidence level.

around 2 % for dissolved organic phosphorus. The uncertainty at the surface is high for both tracers right there where the dissolved organic phosphorus concentration itself is high.

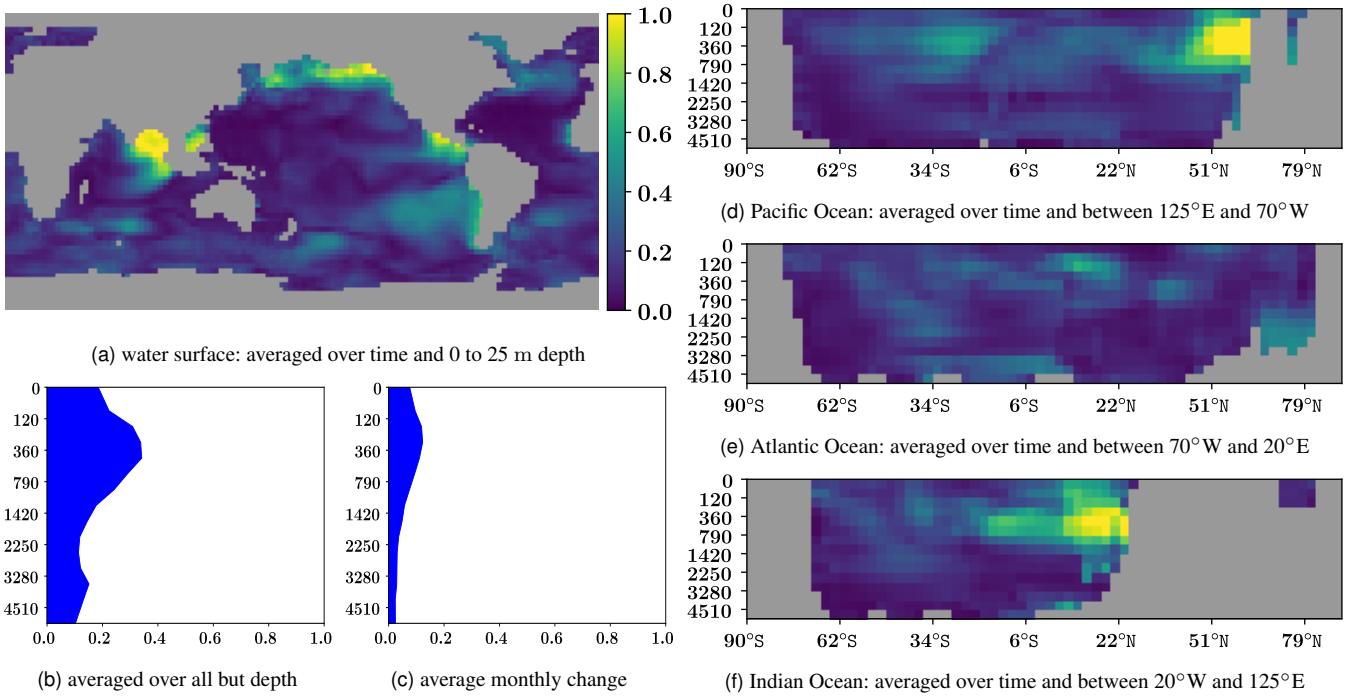
With growing depth, the average uncertainty for phosphate decreases strictly monotonically. It is close to zero after roughly 700 m. In contrast, the uncertainty for dissolved organic phosphorus is almost constant over all depths.

The uncertainties near the surface change on average by 7% per month for both tracers. The temporal variations regarding the uncertainties decrease with growing depth. There is almost no change over time deeper than 700 m for phosphate and deeper than 450 m for dissolved organic phosphorus. This corresponds to the model output itself which is almost constant over time from these depths on.

The absolute difference in climatological mean concentration of phosphate as described by the model and the one estimated directly from measurement data as described in Reimer (2019b) are shown in Figure 7. The differences are very high at the northeast of the Indian Ocean and at the north and east coast of the Pacific Ocean with values close to one.

These differences are significantly higher than the uncertainty in the model output resulting from the uncertainty in model parameters. This indicates that a significant model error occurs at these regions or that the estimated model parameters are not optimal, at least for these regions. Here the model should be improved or maybe model parameters specially suited for these regions should be estimated. The model error can originate in the biogeochemical model or the circulation model or both.

From a depth of around 1000 meters on, the differences relative to the concentrations are small, but still high compared to the uncertainties resulting from the uncertainties in the model parameters. Since the parameters of the biogeochemical



(a) water surface: averaged over time and 0 to 25 m depth

(b) averaged over all but depth

(c) average monthly change

(d) Pacific Ocean: averaged over time and between 125°E and 70°W

(e) Atlantic Ocean: averaged over time and between 70°W and 20°E

(f) Indian Ocean: averaged over time and between 20°W and 125°E

Figure 7. Absolute difference in phosphate model output and climatological mean estimated from measurement data (in $\mathrm{mmol\,m^{-3}}$).

model influence the concentration only very slightly at these depths, this indicates that the transport model is also erroneous. Nevertheless, the model reflects the climatological concentrations quite well in many regions.

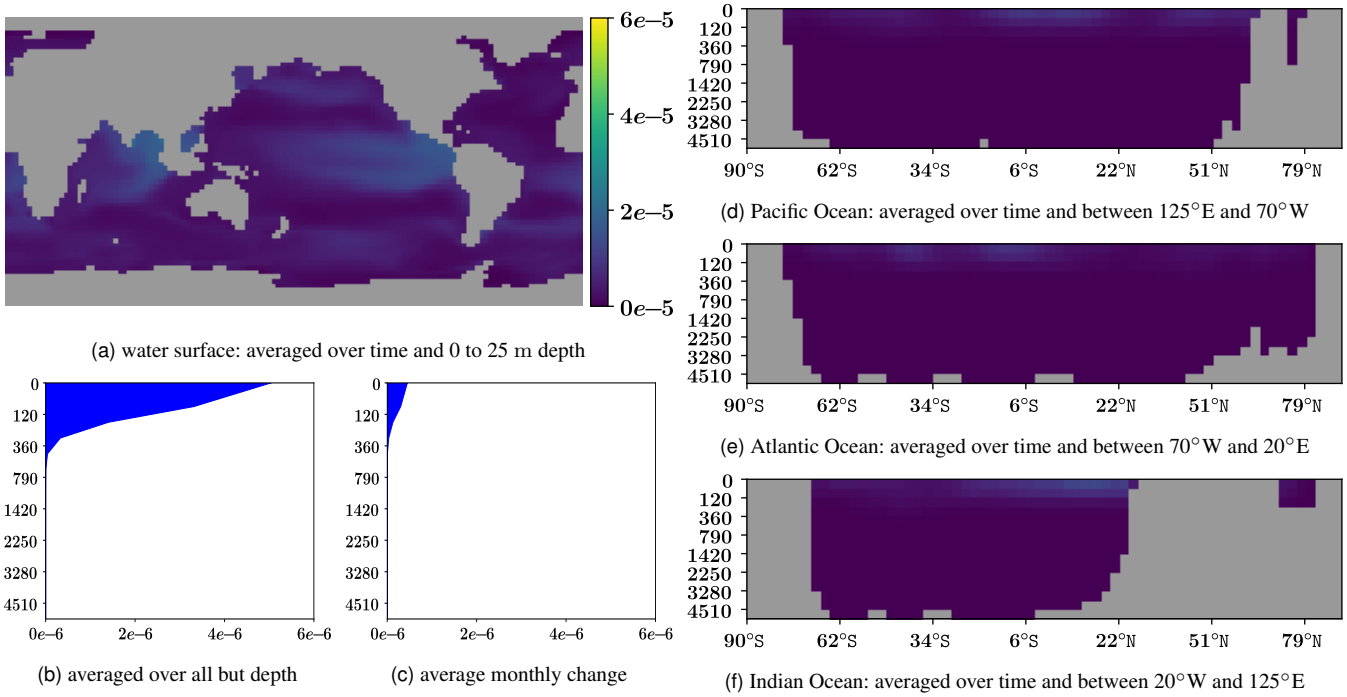
Due to the small number of dissolved organic phosphorus measurements, a corresponding comparison is not possible.

4.4 Uncertainty Reduction by Additional Measurements

The uncertainty regarding the model parameters as well as the model outputs can be reduced by additional measurements as described in Subsection 2.4.

In order to find out which measurement design significantly reduce the uncertainties and which result only in a slight information gain, we have analyzed the average model uncertainty equally weighted for both tracers as described in Equation (16) resulting for one additional measurement. Figure 8 and 9 show by what proportion the average model uncertainty is reduced by one additional measurement at this point.

The most informative measurements are located at the water surface. The information content decreases rapidly with growing depth. Compared to the the information content at the surface, it is below one third for phosphate measurements deeper than 150 m and for dissolved organic phosphorus measurements deeper than 80 m. Deeper than 400 m it is close to zero for phosphate measurements and deeper than 200 m it is approximately constant one sixth for dissolved organic phosphorus measurements. The time of the measurement seems to have little effect on their information content.



(a) water surface: averaged over time and 0 to 25 m depth

(b) averaged over all but depth

(c) average monthly change

(d) Pacific Ocean: averaged over time and between 125°E and 70°W

(e) Atlantic Ocean: averaged over time and between 70°W and 20°E

(f) Indian Ocean: averaged over time and between 20°W and 125°E

Figure 8. Uncertainty reduction by one phosphate measurement at this location (in $\mathrm{mmol\,m^{-3}}$).

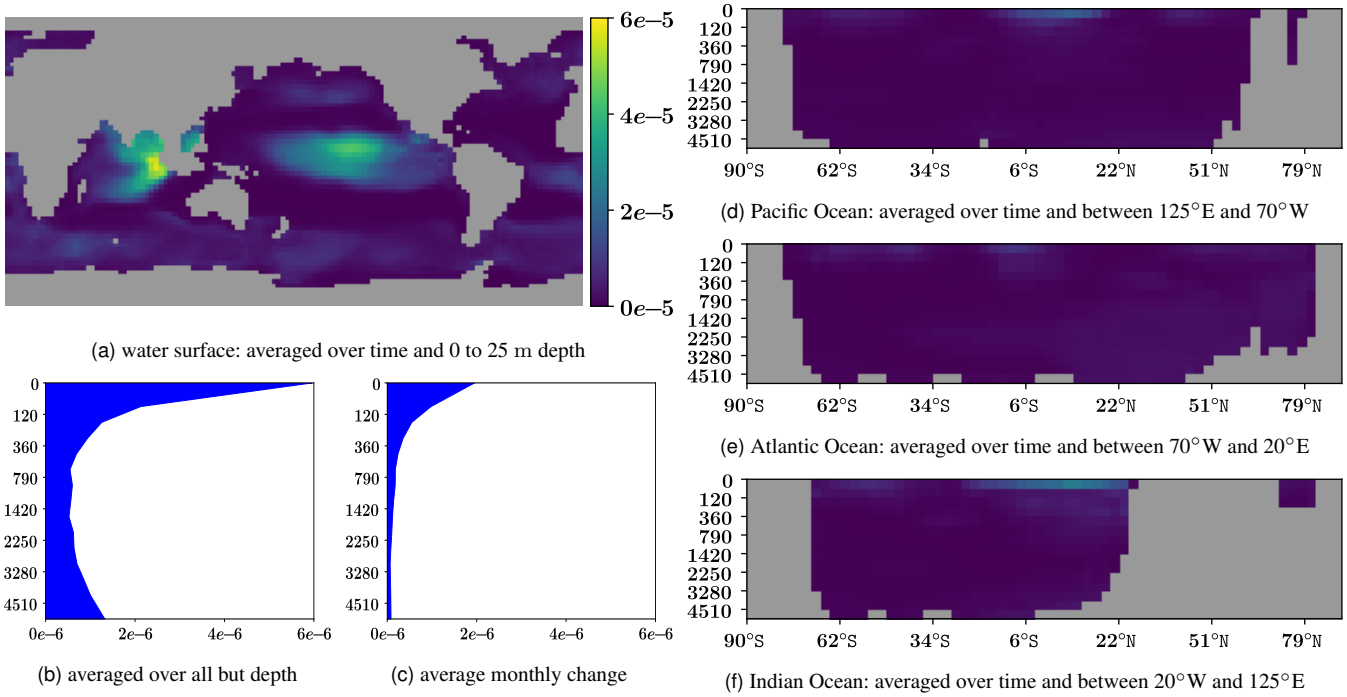
Phosphate measurements have the highest information content at the north-eastern part of the Indian Ocean and at the middle-east part of the Pacific Ocean. This indicates that measurements in areas where the concentration of dissolved organic phosphorus is high are especially worthwhile for phosphate measurements.

The highest information content of dissolved organic phosphorus measurements is at the surface of the north-eastern part of the Indian Ocean and the middle of the Pacific Ocean both around the equator. This indicates that measurements in areas where the concentration of dissolved organic phosphorus is high but that of phosphorus is low are especially worthwhile for dissolved organic phosphorus measurements.

A dissolved organic phosphorus measurement contains usually twice as much information as a phosphate measurements. However, carrying out a dissolved organic phosphorus measurement is many times more complex and expensive than carrying out a phosphate measurement. This means that carrying out dissolved organic phosphorus measurements is not worthwhile for reducing the model uncertainty.

A single additional measurement can reduce the average model uncertainty at most by roughly a twenty thousandth part. Hence, for a significant reduction, many additional measurements are required. This is plausible, since more than four million measurements have already been carried out and result in the current uncertainty.

As described in the previous subsection, the model appears to be erroneous in the northeast of the Indian Ocean and the north and east coast of the Pacific Ocean which speaks in favor of carrying out additional measurements there as well. These could then help to improve the model.



(a) water surface: averaged over time and 0 to 25 m depth

(b) averaged over all but depth

(c) average monthly change

(d) Pacific Ocean: averaged over time and between 125°E and 70°W

(e) Atlantic Ocean: averaged over time and between 70°W and 20°E

(f) Indian Ocean: averaged over time and between 20°W and 125°E

Figure 9. Uncertainty reduction by one dissolved organic phosphorus measurement at this location (in $\mathrm{mmol\,m^{-3}}$).

5 Conclusions

In this article we have presented several methods for model parameter estimation and uncertainty quantification. They are based on the generalized least squares estimator which has been described together with its statistical properties.

Several approximations of the covariance matrix of the estimator of the model parameters as well as the corresponding model output have been introduced. They are based on the first and second derivative of the model regarding its parameters. Their advantages and disadvantages have been emphasized. Approximate confidence intervals were provided as another way to quantify uncertainties.

Optimal experimental design methods have been briefly introduced which allow to predict the uncertainty reduction by additional measurements and to design new measurements in such a way that the information gain is maximized.

We have applied all these methods to a model for phosphate and dissolved organic phosphorus concentrations in the global ocean. For this, we have introduced the model briefly as well as its evaluation and corresponding measurement data.

We were able to find model parameters which are significantly more consistent with the measurement data compared to our initial guess. The individual model parameters of the model are subject to very diverse uncertainties. The uncertainties vary from 0.1 % to 7 % of the parameter values.

The uncertainties in the associated model output vary greatly as well, depending on location, time and tracer. The largest uncertainties are at the water surface, where, they are in average around 1 % of the phosphate concentrations and around 2 % of

the dissolved organic phosphorus concentrations. Usually, they are high where the dissolved organic phosphorus concentration is high. With increasing depth the uncertainty for phosphate decreases rapidly while remaining more or less constant for dissolved organic phosphorus.

In the northeast of the Indian Ocean and near the north and east coast of the Pacific Ocean, the difference between the climatological phosphate concentration described by the model and calculated from measurement data are much higher than the uncertainty implied by the uncertainty in the model parameters. This indicates that the model is erroneous here.

New measurements are most informative if they are close to the water surface. Phosphate measurements are especially worthwhile where the concentration of dissolved organic phosphorus is high. Taking into account the additional effort and costs associated with dissolved organic phosphorus measurements they are not worthwhile. If dissolved organic phosphorus measurements should be carried out nevertheless, they should be carried out where the dissolved organic phosphorus concentration is high and the phosphorus concentration is low.

The results obtained for this model help to better assess its parameters and output as well as to plan new measurements. The applicability and usefulness of the presented methods has been shown with this application example and are applicable to a wide range of models.

26

References

Abramowitz, M. and Stegun, I. A.: Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. National Bureau of Standards Applied Mathematics Series 55. Tenth Printing., ERIC, 1972.

Amemiya, T.: Nonlinear regression models, Handbook of econometrics, 1, 333–389, 1983.

Aster, R. C., Borchers, B., and Thurber, C. H.: Parameter Estimation and Inverse Problems, Elsevier, second edn., 2013.

Balay, S., Abhyankar, S., Adams, M. F., Brown, J., Brune, P., Buschelman, K., Dalcin, L., Dener, A., Eijkhout, V., Gropp, W. D., Karpeyev, D., Kaushik, D., Knepley, M. G., May, D. A., McInnes, L. C., Mills, R. T., Munson, T., Rupp, K., Sanan, P., Smith, B. F., Zampini, S., Zhang, H., and Zhang, H.: PETSc Users Manual, Tech. Rep. ANL-95/11 - Revision 3.11, Argonne National Laboratory, https://www.mcs.anl.gov/petsc, 2019a.

Balay, S., Abhyankar, S., Adams, M. F., Brown, J., Brune, P., Buschelman, K., Dalcin, L., Dener, A., Eijkhout, V., Gropp, W. D., Karpeyev, D., Kaushik, D., Knepley, M. G., May, D. A., McInnes, L. C., Mills, R. T., Munson, T., Rupp, K., Sanan, P., Smith, B. F., Zampini, S., Zhang, H., and Zhang, H.: PETSc Web page, https://www.mcs.anl.gov/petsc, 2019b.

Bigg, G. R.: The Oceans and Climate, Cambridge University Press, second edn., 2003.

Boyer, T., Antonov, J., Baranova, O., Coleman, C., Garcia, H., Grodsky, A., Johnson, D., Locarnini, R., Mishonov, A., O'Brien, T., Paver, C., Reagan, J., Seidov, D., Smolyar, I., and Zweng, M.: World Ocean Database 2013, Tech. rep., National Oceanic and Atmospheric Administration, Silver Spring, https://doi.org/10.7289/V5NZ85MT, S. Levitus, Ed.; A. Mishonov, Technical Ed., 2013.

Brock, T. D.: Calculating solar radiation for ecological studies, Ecological Modelling, 14, 1–19, https://doi.org/10.1016/0304-3800(81)90011-9, http://www.sciencedirect.com/science/article/pii/0304380081900119, 1981.

Broyden, C. G.: The Convergence of a Class of Double-rank Minimization Algorithms: 2. The New Algorithm, IMA Journal of Applied Mathematics, 6, 222–231, https://doi.org/10.1093/imamat/6.3.222, http://imamat.oxfordjournals.org/content/6/3/222.abstract, 1970.

Cao, X. and Spall, J. C.: Preliminary results on relative performance of expected and observed fisher information, in: Proceedings of the 48h IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference, pp. 1538–1543, https://doi.org/10.1109/CDC.2009.5400435, 2009.

Cao, X. and Spall, J. C.: Relative performance of expected and observed fisher information in covariance estimation for maximum likelihood estimates, in: 2012 American Control Conference (ACC), pp. 1871–1876, https://doi.org/10.1109/ACC.2012.6315584, 2012.

Caswell, T. A., Droettboom, M., Hunter, J., Lee, A., Firing, E., Stansby, D., Klymak, J., de Andrade, E. S., Nielsen, J. H., Varoquaux, N., Hoffmann, T., Root, B., Elson, P., May, R., Dale, D., Lee, J.-J., Seppänen, J. K., McDougall, D., Straw, A., Hobson, P., Gohlke, C., Yu, T. S., Ma, E., Vincent, A. F., Silvester, S., Moad, C., Katins, J., Kniazev, N., Ariza, F., and Ernest, E.: Matplotlib: 3.1.1, https://doi.org/10.5281/zenodo.3264781, 2019.

Conn, A. R., Scheinberg, K., and Vicente, L. N.: Introduction to derivative-free optimization, vol. 8, Siam, 2009.

Dennis, J. and Schnabel, R.: Numerical Methods for Unconstrained Optimization and Nonlinear Equations, Society for Industrial and Applied Mathematics, Philadelphia, https://doi.org/10.1137/1.9781611971200, 1996.

Donaldson, J. R. and Schnabel, R. B.: Computational Experience with Confidence Regions and Confidence Intervals for Nonlinear Least Squares, Technometrics, 29, 67–82, http://www.jstor.org/stable/1269884, 1987.

Dutkiewicz, S., Follows, M., Marshall, J., and Gregg, W. W.: Interannual variability of phytoplankton abundances in the North Atlantic, Deep Sea Research Part II: Topical Studies in Oceanography, 48, 2323–2344, https://doi.org/10.1016/S0967-0645(00)00178-8, http://

www.sciencedirect.com/science/article/pii/S0967064500001788, {JGOFS} Research in the North Atlantic Ocean: A Decade of Research, Synthesis and modelling, 2001.

Dutkiewicz, S., Sokolov, A. P., Scott, J., and Stone, P. H.: A three-dimensional ocean-seaice-carbon cycle model and its coupling to a two-dimensional atmospheric model: uses in climate change studies, Tech. Rep. 122, MIT - Massachusetts Institute of Technology, Cambridge, 2005.

Efron, B. and Hinkley, D. V.: Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information, Biometrika, 65, 457–483, https://doi.org/10.1093/biomet/65.3.457, 1978.

Fletcher, R.: A new approach to variable metric algorithms, The Computer Journal, 13, 317–322, https://doi.org/10.1093/comjnl/13.3.317, http://comjnl.oxfordjournals.org/content/13/3/317.abstract, 1970.

Garcia, H. E., Locarnini, R. A., Boyer, T. P., Antonov, J. I., Baranova, O. K., Zweng, M. M., Reagan, J. R., and Johnson, D. R.: World Ocean Atlas 2013, Volume 4: Dissolved Inorganic Nutrients (phosphate, nitrate, silicate), Tech. rep., National Oceanic and Atmospheric Administration, S. Levitus, Ed., A. Mishonov Technical Ed.; NOAA Atlas NESDIS 76, 25 pp., 2014.

Gill, P. E., Murray, W., and Wright, M. H.: Practical Optimization, Academic Press, London, 1981.

Gill, P. E., Murray, W., Wright, M. H., et al.: Numerical linear algebra and optimization, vol. 5, Addison-Wesley Redwood City, 1991.

Glover, F.: A template for scatter search and path relinking, in: Artificial Evolution, edited by Hao, J.-K., Lutton, E., Ronald, E., Schoenauer, M., and Snyers, D., vol. 1363 of *Lecture Notes in Computer Science*, pp. 1–51, Springer Berlin Heidelberg, https://doi.org/10.1007/BFb0026589, 1998.

Goldfarb, D.: A family of variable-metric methods derived by variational means, Mathematics of computation, 24, 23–26, 1970.

Graven, H. D., Gruber, N., Key, R., Khatiwala, S., and Giraud, X.: Changing controls on oceanic radiocarbon: New insights on shallow-to-deep ocean exchange and anthropogenic CO_2 uptake, Journal of Geophysical Research: Oceans, 117, https://doi.org/10.1029/2012JC008074, c10005, 2012.

Han, S.: A globally convergent method for nonlinear programming, Journal of Optimization Theory and Applications, 22, 297–309, https://doi.org/10.1007/BF00932858, 1977.

Hunter, J. D.: Matplotlib: A 2D graphics environment, Computing in Science & Engineering, 9, 90–95, https://doi.org/10.1109/MCSE.2007.55, 2007.

Jennrich, R. I.: Asymptotic Properties of Non-Linear Least Squares Estimators, The Annals of Mathematical Statistics, 40, 633–643, https://doi.org/10.1214/aoms/1177697731, 1969.

Johnson, D. R., Boyer, T. P., Garcia, H. E., Locarnini, R. A., Baranova, O. K., and Zweng, M. M.: World Ocean Database 2013 User's Manual, Tech. Rep. NODC Internal Report 22, National Oceanographic Data Center, https://doi.org/10.7289/V5DF6P53, Sydney Levitus, Ed.; Alexey Mishonov, Technical Ed.; NODC Internal Report 22, NOAA Printing Office, Silver Spring, MD, 172 pp, 2013.

Jones, E., Oliphant, T., Peterson, P., et al.: SciPy: Open source scientific tools for Python, http://www.scipy.org, version 1.3, 2019.

Karush, W.: Minima of functions of several variables with inequalities as side constraints, Ph.D. thesis, Master's thesis, Dept. of Mathematics, Univ. of Chicago, 1939.

Khatiwala, S.: A computational framework for simulation of biogeochemical tracers in the ocean, Global Biogeochemical Cycles, 21, https://doi.org/10.1029/2007GB002923, 2007.

Khatiwala, S., Visbeck, M., and Cane, M. A.: Accelerated simulation of passive tracers in ocean circulation models, Ocean Modelling, 9, 51–69, 2005.

Kriest, I., Khatiwala, S., and Oschlies, A.: Towards an assessment of simple global marine biogeochemical models of different complexity, Progress in Oceanography, 86, 337–360, https://doi.org/10.1016/j.pocean.2010.05.002, http://eprints.ifm-geomar.de/9204/, 2010.

Kriest, I., Oschlies, A., and Khatiwala, S.: Sensitivity analysis of simple global marine biogeochemical models, Global Biogeochemical Cycles, 26, https://doi.org/10.1029/2011GB004072, gB2029, 2012.

Kuhn, H. W. and Tucker, A. W.: Nonlinear Programming, in: Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, pp. 481–492, University of California Press, Berkeley, Calif., http://projecteuclid.org/euclid.bsmsp/1200500249, 1951.

Kwon, E. Y., Primeau, F., and Sarmiento, J. L.: The impact of remineralization depth on the air–sea carbon balance, Nature Geoscience, 2, 630–635, 2009.

Landolfi, A.: The importance of dissolved organic nutrients in the biogeochemistry of oligotrophic gyres, Ph.D. thesis, University of Southampton, http://eprints.soton.ac.uk/25117/, 2005.

Landolfi, A., Oschlies, A., and Sanders, R.: Organic nutrients and excess nitrogen in the North Atlantic subtropical gyre, Biogeosciences, 5, 1199–1213, https://doi.org/10.5194/bg-5-1199-2008, http://www.biogeosciences.net/5/1199/2008/, 2008.

Maier-Reimer, E.: Geochemical cycles in an ocean general circulation model. Preindustrial tracer distributions, Global Biogeochemical Cycles, 7, 645–677, https://doi.org/10.1029/93GB01355, 1993.

Marshall, J., Adcroft, A., Hill, C., Perelman, L., and Heisey, C.: A finite-volume, incompressible Navier Stokes model for studies of the ocean on parallel computers, Journal of Geophysical Research, 102, 5753–5766, https://doi.org/10.1029/96JC02775, 1997a.

Marshall, J., Hill, C., Perelman, L., and Adcroft, A.: Hydrostatic, quasi-hydrostatic, and nonhydrostatic ocean modeling, Journal of Geophysical Research: Oceans, 102, 5733–5752, https://doi.org/10.1029/96JC02776, 1997b.

Marshall, J., Jones, H., and Hill, C.: Efficient ocean modeling using non-hydrostatic algorithms, Journal of Marine Systems, 18, 115–134, https://doi.org/10.1016/S0924-7963(98)00008-6, http://www.sciencedirect.com/science/article/pii/S0924796398000086, 1998.

Martin, J. H., Knauer, G. A., Karl, D. M., and Broenkow, W. W.: VERTEX: carbon cycling in the northeast Pacific, Deep Sea Research Part A. Oceanographic Research Papers, 34, 267–285, https://doi.org/10.1016/0198-0149(87)90086-0, http://www.sciencedirect.com/science/article/pii/0198014987900860, 1987.

MathWorks: Matlab R2015a Primer, Natick, Massachusetts, R2015a edn., http://www.mathworks.de/help/releases/R2015a/pdf_doc/matlab/getstart.pdf, visited on 2015-10-30, 2015a.

MathWorks: Matlab Global Optimization Toolbox R2015a User's Guide, Natick, Massachusetts, R2015a edn., http://www.mathworks.de/help/releases/R2015a/pdf_doc/gads/gads_tb.pdf, visited on 2015-10-30, 2015b.

MathWorks: Matlab Optimization Toolbox R2015a User's Guide, Natick, Massachusetts, R2015a edn., http://www.mathworks.de/help/releases/R2015a/pdf_doc/optim/optim_tb.pdf, visited on 2015-10-30, 2015c.

McGuffie, K. and Henderson-Sellers, A.: A Climate Modelling Primer, Wiley, Chichester, 3 edn., 2005.

McKinley, G. A., Follows, M. J., and Marshall, J.: Mechanisms of air-sea CO2 flux variability in the equatorial Pacific and the North Atlantic, Global Biogeochemical Cycles, 18, 2004.

Najjar, R. and Orr, J.: Design of OCMIP-2 simulations of chlorofluorocarbons, the solubility pump and common biogeochemistry, http://www.cgd.ucar.edu/oce/klindsay/OCMIP/design.pdf, 1998.

Najjar, R. G., Jin, X., Louanchi, F., Aumont, O., Caldeira, K., Doney, S. C., Dutay, J.-C., Follows, M., Gruber, N., Joos, F., Lindsay, K., Maier-Reimer, E., Matear, R. J., Matsumoto, K., Monfray, P., Mouchet, A., Orr, J. C., Plattner, G.-K., Sarmiento, J. L., Schlitzer, R., Slater, R. D., Weirig, M.-F., Yamanaka, Y., and Yool, A.: Impact of circulation on export production, dissolved organic matter, and dissolved oxygen in

the ocean: Results from Phase II of the Ocean Carbon-cycle Model Intercomparison Project (OCMIP-2), Global Biogeochemical Cycles, 21, https://doi.org/10.1029/2006GB002857, gB3007, 2007.

Neelin, J. D.: Climate Change and Climate Modeling, Cambridge University Press, 2010.

Nocedal, J. and Wright, S.: Numerical Optimization, Springer series in operations research and financial engineering, Springer, New York, second edn., 2006.

Oliphant, T. E. et al.: NumPy: N-dimensional array package for Python, http://www.numpy.org, version 1.17.3, 2019.

Paltridge, G. and Platt, C.: Radiative processes in meteorology and climatology, Developments in atmospheric science, Elsevier Scientific Pub. Co., 1976.

Parekh, P., Follows, M. J., and Boyle, E. A.: Decoupling of iron and phosphate in the global ocean, Global Biogeochemical Cycles, 19, https://doi.org/10.1029/2004GB002280, 2005.

Parekh, P., Follows, M. J., Dutkiewicz, S., and Ito, T.: Physical and biological regulation of the soft tissue carbon pump, Paleoceanography, 21, https://doi.org/10.1029/2005PA001258, pA3001, 2006.

Pazman, A. and Pronzato, L.: Asymptotic criteria for designs in nonlinear regression with model errors, Mathematica Slovaca, 56, 543–553, 2006.

Piwonski, J. and Slawig, T.: Metos3D: the Marine Ecosystem Toolkit for Optimization and Simulation in 3-D, http://metos3d.github.io, 2013.

Piwonski, J. and Slawig, T.: Metos3D: the Marine Ecosystem Toolkit for Optimization and Simulation in 3-D – Part 1: Simulation Package v0.3.2, Geoscientific Model Development, 9, 3729–3750, https://doi.org/10.5194/gmd-9-3729-2016, http://www.geosci-model-dev.net/9/3729/2016/, 2016.

Platt, T., Harrison, W., Lewis, M., Li, W., Sathyendranath, S., Smith, R., and Vezina, A.: Biological Production of the Oceans - The Case for a Consensus, Marine Ecology Progress Series, 52, 77–88, 1989.

Powell, M.: A fast algorithm for nonlinearly constrained optimization calculations, in: Numerical Analysis, edited by Watson, G., vol. 630 of *Lecture Notes in Mathematics*, pp. 144–157, Springer Berlin Heidelberg, https://doi.org/10.1007/BFb0067703, 1978a.

Powell, M.: The convergence of variable metric methods for non-linearly constrained optimization calculations, in: Nonlinear Programming 3, edited by Mangasarian, O. L., Meyer, R. R., and Robinson, S. M., pp. 27–63, Academic Press, https://doi.org/10.1016/B978-0-12-468660-1.50007-4, http://www.sciencedirect.com/science/article/pii/B9780124686601500074, 1978b.

Prieß, M., Piwonski, J., Koziel, S., Oschlies, A., and Slawig, T.: Accelerated parameter identification in a 3D marine biogeochemical model using surrogate-based optimization, Ocean Modelling, 68, 22–36, https://doi.org/10.1016/j.ocemod.2013.04.003, http://www.sciencedirect.com/science/article/pii/S1463500313000693, 2013.

Pronzato, L. and Pázman, A.: Design of Experiments in Nonlinear Models: Asymptotic Normality, Optimality Criteria and Small-Sample Properties, Lecture Notes in Statistics, Springer, New York, NY, 2013.

Pukelsheim, F.: Optimal Design of Experiments, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2006.

Python Software Foundation: Python, http://www.python.org, version 3.7, 2018.

Reimer, J.: Approximation of Hermitian Matrices by Positive Semidefinite Matrices using Modified Cholesky Decompositions, ArXiv e-prints, https://arxiv.org/abs/1806.03196v2, 2019a.

Reimer, J.: Statistical Analysis of the Phosphate Data of the World Ocean Database 2013, ArXiv e-prints, https://arxiv.org/abs/1912.07384, 2019b.

Reimer, J.: matrix-decomposition: a library for decompose (factorize) dense and sparse matrices in Python, https://doi.org/10.5281/zenodo.3558540, version 1.2, 2019a.

30

Reimer, J.: measurements library: Python functions to handle, statistically analyze and plot measurement data., https://doi.org/10.5281/zenodo.3558700, version 0.3, 2019b.

Reimer, J.: simulation library: Python functions for simulating mathematical models, estimating model parameters, quantifying uncertainties and visualizing results.., https://doi.org/10.5281/zenodo.3558702, version 0.4, 2019c.

Reimer, J.: utillib library: Python functions used in several other projects., https://doi.org/10.5281/zenodo.3558698, version 0.3, 2019d.

Rios, L. M. and Sahinidis, N. V.: Derivative-free optimization: A review of algorithms and comparison of software implementations, Journal of Global Optimization, 56, 1247–1293, 2013.

Sarmiento, J. L., Thiele, G., Key, R. M., and Moore, W. S.: Oxygen and nitrate new production and remineralization in the North Atlantic subtropical gyre, J. Geophys. Res, 95, 303–18, 1990.

Seber, G. A. F. and Wild, C. J.: Nonlinear Regression, Wiley series in probability and statistics, Wiley-Interscience, 2003.

Shanno, D. F.: Conditioning of Quasi-Newton Methods for Function Minimization, Mathematics of Computation, 24, 647–656, http://www.jstor.org/stable/2004840, 1970.

Smith, R. C.: Uncertainty Quantification: Theory, Implementation, and Applications, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2013.

Sokolov, A. P., Schlosser, C. A., Dutkiewicz, S., Paltsev, S., Kicklighter, D. W., Jacoby, H. D., Prinn, R. G., Forest, C. E., Reilly, J. M., Wang, C., et al.: MIT integrated global system model (IGSM) version 2: model description and baseline evaluation, Tech. Rep. 124, MIT - Massachusetts Institute of Technology, 2005.

Tenorio, L.: An Introduction to Data Analysis and Uncertainty Quantification for Inverse Problems, Society for Industrial and Applied Mathematics, Philadelphia, PA, https://doi.org/10.1137/1.9781611974928, https://epubs.siam.org/doi/abs/10.1137/1.9781611974928, 2017.

Ugray, Z., Lasdon, L., Plummer, J., Glover, F., Kelly, J., and Martí, R.: Scatter search and local NLP solvers: A multistart framework for global optimization, INFORMS Journal on Computing, 19, 328–340, 2007.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, I., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy: SciPy 1.0-Fundamental Algorithms for Scientific Computing in Python, CoRR, abs/1907.10121, http://arxiv.org/abs/1907.10121, 2019.

Walter, É. and Pronzato, L.: Identification of Parametric Models from Experimental Data, Communications and control engineering, Springer, New York, 1997.

Weber, T. S. and Deutsch, C.: Ocean nutrient ratios governed by plankton biogeography, Nature, 467, 550–554, https://doi.org/10.1038/nature09403, 2010.

White, H.: Consequences and Detection of Misspecified Nonlinear Regression Models, Journal of the American Statistical Association, 76, 419–433, http://www.jstor.org/stable/2287845, 1981.

White, H.: Maximum Likelihood Estimation of Misspecified Models, Econometrica, 50, 1–25, http://www.jstor.org/stable/1912526, 1982.

Yamanaka, Y. and Tajika, E.: Role of dissolved organic matter in the marine biogeochemical cycle: Studies using an ocean biogeochemical general circulation model, Global Biogeochemical Cycles, 11, 599–612, https://doi.org/10.1029/97GB02301, 1997.

Yoshimura, T., Nishioka, J., Saito, H., Takeda, S., Tsuda, A., and Wells, M. L.: Distributions of particulate and dissolved organic and inorganic phosphorus in North Pacific surface waters, Marine Chemistry, 103, 112–121, https://doi.org/10.1016/j.marchem.2006.06.011, http://www.sciencedirect.com/science/article/pii/S0304420306001162, 2007.

31

6 Conclusions and Outlooks

Methods for optimization of model parameters, uncertainty quantification and uncertainty reduction by optimal experimental designs were successfully applied in this thesis to application examples in climate research of varying complexity. The methods were presented in detail and have proven be well applicable even for more complex models.

The optimized model parameters allow more realistic forecasts. The performed uncertainty quantifications are extremely useful in assessing the model outputs and the measurement data. Optimal experimental design methods were used to predict the uncertainty reduction by additional measurements allowing to predict whether and how many additional measurements are useful and how these measurements should be carried out. The results of the uncertainty quantification and optimal experimental design methods provide a good insight how the measurement data affects the estimated model parameters and the corresponding model output.

We presented a statistical analysis method for climate data which splits the effects into a climatological and a short scale part. In addition to the climatological mean, short and long scale variations and correlations were quantified. A new algorithm was developed to determine valid correlation matrices. The analysis method and the algorithm could be successfully applied to large sets of marine phosphate data resulting in a more detailed insight into the climatological phosphate concentration in the ocean.

For each topic covered, more detailed conclusions and an outlook for further research is presented in the following.

6.1 Marine Phosphate Data and their Statistical Analysis

Conclusions

A detailed statistical analysis of the phosphate measurement data provided by the World Ocean Database 2013 has been carried out in Section 4 providing further insights in phosphate concentrations in the global ocean, in particular from a climatological point of view.

In this analysis, the phosphate concentrations in the global ocean were estimated for a climatological, i.e., average, year. The variability around this climatological year has been quantified and was split into the climatological variability, i.e., the usual deviation between the average concentration in a specific year and the concentration in a climatological year as well as the short scale variability, i.e., the usual deviation between the measured concentration and the average concentration in this year.

The climatological variability is usually the one of interest in climate research. Nevertheless, the short scale variability contains useful information in allowing to assess the resolution of the statistical analysis. A high short scale variability indicates that the resolution should be increased to incorporate further details, where a low short scale variability indicates that the resolution could be decreased without losing much information. Further, it shows the amount of measurements necessary to determine the average concentration and thus could prove useful when planning new measurements.

Correlations, which are often neglected in climate research, were analyzed. Especially between points that are close to each other, significant correlations were identified.

The probability distributions were examined with regard to whether they are normal or log-normal. Unfortunately no clear trend was found and some distributions seemed to be neither.

The methods applied in this statistical analysis are not limited to phosphate data and can be used for other marine climate data as well. A software package was provided for

this purpose.

Outlooks

The information obtained by this analysis about appropriate resolutions can be used for new analyses. The best way to do this is to use an adaptive resolution where strongly variable regions are resolved more precisely and less variable regions less precisely.

The World Ocean Database 2018 is going to be published soon which is going to contain further phosphate measurements which could be included in a new analysis and are expected to provide more accurate results.

To discover climatological changes, the measurement data could be grouped over certain periods of time, e.g., decades, and each of these groups could be analyzed separately. Significant differences between the groups would indicate climatological changes.

Other interpolation methods, like splines [43], could be used to estimate values where not enough data is available. This could result in a smoother values.

The analysis regarding the probability distribution can be extended to other kinds. It seems, however, that the values do not originate from a single type of probability distributions.

The basis of the analysis is the assumption that the measurement results are the sum of a long scale and a shot scale component. A less common alternative is the assumption that the relationship is multiplicative instead of additive. It could be investigated if an analysis based on this assumption would provide significantly different results.

6.2 Marine Phosphorus Model

Conclusions

A model describing the phosphate, the dissolved organic phosphorus concentrations and the biological production in the global ocean is subject of Section 5. The model parameters have been estimated using the generalized least squares estimator, the phosphate data from the World Ocean Database 2013 and the results presented in Section 4.

The parameters could be determined in such a way that the resulting model output fits the measurement results significantly better. The model reproduce the climatological phosphate concentrations calculated from measurement data, as described in Section 4, quite well in most regions. However, the model seams to be erroneous at the northeast of the Indian Ocean and some regions near the north and east coast of the Pacific Ocean.

The uncertainty in the parameters implied by the uncertainty in the measurement results have been quantified. The average uncertainty is around three percent of the parameters value and range from seven percent to one permille. The uncertainties in the corresponding model output range from 0.012 to 0 mmol m^{-3}. Averaged over all but the depth, they are 0.006 mmol m^{-3} near the water surface and decrease with growing depth for phosphate and are almost constant 0.005 mmol m^{-3} at all depths for dissolved organic phosphorus. This results in an uncertainty relative to the model output of 1 % and 2% for phosphate and dissolved organic phosphorus, respectively, near the surface. The relative uncertainty decreases with growing depth for phosphate and increases for dissolved organic phosphorus.

Overall we consider this uncertainty as realistic since solely uncertainties in the data were taken into account and uncertainties from model structure and numerical errors were not quantified. The total uncertainty, including these as well, is higher than the

the calculated uncertainty which can thus be considered as lower bound for the total uncertainty.

The predicted reduction of the uncertainty by additional measurements revealed that additional measurements further reduce the uncertainty the closer they are to the water surface, where data can be collected with less effort than at great depths. Moreover an additional dissolved organic phosphorus measurement usually reduces the uncertainty twice as much as an additional phosphate measurement. It should be noted that dissolved organic phosphorus measurements are considerably more expensive and therefore additional measurements should be limited to phosphate measurements. The reduction varies strongly at different measuring locations as opposed to different measurement times. One additional measurement reduces the average uncertainty at most by a ten thousandth part of its value and hence, a significant improvement requires many additional measurements. However, the effort to achieve reductions can be considerably reduced by well chosen additional measurements.

Outlooks

Several other models for the marine phosphorus cycle exist [33] to which the methods presented in this thesis could be applied as well. After the model parameter optimization, it could be examined how realistic those models are and whether the higher complexity in some models is justified. Optimal experimental design methods for model discrimination [63, 6.6.3] can be used for this purpose as outlined in Subsection 6.6.

If a climatological change in the marine phosphorus cycle would exist, it would be helpful to determine the model parameters once before and once after the change. This would allow more accurate forecasts for each of these two periods. Further, it could help to understand this change and contribute to the model development. It may also be useful to determine separate model parameters for those regions where the model is erroneous with the current model parameters.

Due to many numerical experiments, model outputs for more than thirty thousand different sets of model parameters have been accumulated. These could be used directly to quantify the uncertainty in the model output. This could be based on ranges, where the uncertainty in the model output is quantified by the range of the existing model outputs with model parameters in a reasonable predefined range. Or it could be done based on probability distributions, where the uncertainty is quantified by a probability distribution based on the existing model outputs for a reasonable predefined probability distribution of the model parameters.

6.3 Salt Marsh Models

Conclusions

For the two presented models for salt marshes, measurement conditions have been optimized in such a way that a minimum number of measurements are required to determine the model parameters. Thus, it is possible to adapt the models to local salt marshes with minimal measuring effort.

It turned out that for the model with two parameters, about ten measurements would suffice and three of them should be carried out at the beginning, and the remaining at the end of the flooding of the salt marsh. For the model with three parameters, about twenty to twenty five measurements should be carried out at the end of the flooding.

Outlooks

Possible correlations of the measurement noise were neglected in the parameter estimation and optimization of measurement conditions. It might be justified in this case, however, it is recommended to confirm this assumption by measuring experiments. If correlations should be included after all, the generalized least squares estimator can be used as it was realized for the marine phosphorus model.

Since we have two competing models, it would make sense to compare their closeness to reality. For this purpose, measurement conditions can be optimized according to Subsection 6.6.

6.4 Optimization of Model Parameters

Conclusions

The parameters of the salt marsh models were estimated using the weighted least squares estimator and the SQP algorithm, a derivative based local optimization algorithm (see Subsection 2.2). The generalized least squares estimator and the SQP algorithm combined with OQNLP (see Subsection 5.2), a globalization algorithm, were used to estimate the parameters of the marine phosphorus model.

The weighted least squares estimator is reasonable if the measurement noises are uncorrelated, which is assumed for the salt marsh measurements. For marine phosphate measurements, the statistical analysis (see Section 4) revealed correlations for which the generalized least squares estimator should be preferred.

The SQP algorithm required only few function evaluations to determine (local) minima. However, depending on the application example, local minima might not be global minima. The SQP algorithm had reliably found global minima for the salt marsh models but not for the marine phosphorus model which resulted in the use of the globalization algorithm for the marine phosphorus model.

The applied methods, especially with the globalization algorithm, have been proven to be well applicable even for more complex models, although, the first derivative of the model with respect to its parameters is needed, which, however, is needed anyway for the uncertainty quantification. If necessary, the derivative can be calculated by algorithmic differentiation [17] or finite differences [1, 25.3], whereby algorithmic differentiation should be preferred over finite differences which are associated with longer execution time and lower accuracy.

The parameter estimate of the marine phosphorus model obtained by the generalized least squares estimator were compared with the estimates obtained by the weighted and ordinary least squares estimators. As expected, the estimates differ considerably. This shows the importance of including standard deviations and correlations in the parameter estimation as done by the generalized least squares estimator.

Outlooks

There exists a huge number of other estimators based on different assumptions, providing other advantages and disadvantages. A comparison of the discussed parameter estimates with the results from other methods, like total least squares methods [14, 8], Bayesian estimation methods [3, 11], [60, 2.7], maximum likelihood estimation methods [60, 2.2] based on, e.g., log-normal probability distributions, and finally regularization methods [14, 6-9], [12, 5], [60, 3.4], would be helpful.

6.5 Uncertainty Quantification

Conclusions

Several methods to quantify uncertainties in the model parameters as well as uncertainties in the model output were used in this thesis (see Section 5). They are based on the first and second derivative of the model with respect to its parameters.

The method based only on the first derivative, is easier to apply but is less accurate if the model massively simplifies the modeled process. In this case, the method which uses the first as well as the second derivative should be preferred.

An advantage of the presented methods is that they can be applied with relative small computational effort and are, thus, applicable to more complex models like the marine phosphorus model. The resulting uncertainties differ slightly depending on the used method.

A drawback is that the first and, depending on the method, also the second derivative is required. If they are not available, they can be approximated by algorithmic differentiation [17] or finite differences [1, 25.3]. Finite differences with some model specific accelerations were used for the marine phosphorus model.

Only uncertainties implied by the measurement data and the model parameters were quantified. Hence, the quantified uncertainty is lower than the total uncertainty and can thus be be considered as a lower bound.

Outlooks

Derivatives calculated by algorithmic differentiation are more accurate than derivatives calculated by finite differences. Hence, it would be of interest to check if the results for the marine phosphorus model change significantly if algorithmic differentiation would be used instead of finite differences. However, applying algorithmic differentiation, especially on complex models, is challenging and it is unclear whether it is worth the effort.

There are several robustification approaches reducing the dependency of the uncertainty quantification on the estimated model parameters, e.g., averaging uncertainties ([63, 6.4.3], [46, 8.1]) and worst case uncertainties ([63, 6.4.4], [46, 8.2]). These approaches could be applied to the application examples discussed here. However, robustification approaches considerably increase the computational effort and, therefore, may not be worth the additional computational effort if the uncertainty is quite insensitive with respect to the parameter estimate.

The approximation of the worst uncertainty, as it was done in Subsection 2.2 in the context of experimental design, is computational less expensive. However, this approach needs the derivatives in the next higher order. Numerical experiments have revealed that this robustification approach does not significantly influence the resulting uncertainty for the salt marsh models. For the marine phosphorus model, numerical experiments have shown that its uncertainties are quite insensitive to small changes in the model parameter estimate. This makes robustification needless for our application examples.

It would be interesting to apply other methods for uncertainty quantification, e.g., Monte Carlo based methods, to this application example and compare the result with the results already obtained. However, Monte Carlo based methods imply considerably more computational effort making a less complex application example like the salt marsh models more appropriate for a comparison.

A further step in quantifying the total uncertainty in our application examples would be to quantify the model error [61, 12], i.e., the uncertainty in the model function itself,

and the numerical error [21], i.e., the uncertainty introduced by numerical imprecisions. The model parameter estimation, as intermediate step for quantifying the model error, has already been completed.

6.6 Uncertainty Reduction by Optimal Experimental Designs

Conclusions

Conditions of additional measurements, such as time and location, were determined for the salt marsh models and the marine phosphorus model in such a way that the uncertainty reduction would be maximal when their results are incorporated. Optimal experimental design techniques were used for this purpose which allow to predict the uncertainty reduction by additional measurements without carrying them out. The details are described in Subsection 2.2 and 5.2.

By choosing an appropriate design criterion, it can be decided which uncertainty should be minimized. The average uncertainty in the model parameters and in the model output, in relative and absolute terms, were used in this thesis.

A robustification approach, which reduces the dependency on the previously estimated model parameters, was applied for the salt marsh models. However, the results did not significantly deviate from the results without robustification. For the marine phosphorus model, numerical experiments showed only small changes in the uncertainty for small changes in the model parameters estimate. Hence, the robustification approach, which is computationally more expensive and needs a derivative of additional order, was not used for the marine phosphorus model.

Moreover, sequential optimal experimental design ([46, 8.5], [63, 6.4.2]) where used for the salt marsh models. This is an iterative process where in each iteration only few optimal measurements are determined and carried out and afterwards the model parameters are estimated with the measurements made so far. The advantage of this is that the information from previous measurements is already included in the planning of later measurements. In general, this seems to be a very effective approach. However, for the salt marsh models, it turned out not to be crucial, since the optimal experimental designs are mostly independent of the model parameter estimates. For the marine phosphorus models it was unnecessary as well due to the vast amount of already available measurements.

Outlooks

The applied methods for optimizing measurement conditions all aimed to reduce the uncertainty in the model parameters and the model output. However, there are optimal experimental design methods for model discrimination [63, 6.6.3]. They optimize measurement conditions such that the most realistic one in a selection of models can be identified, allowing to reduce the model error, i.e., the uncertainty regarding the model itself. These methods could be very useful for the application examples discussed here since for the salt marshes as well as the marine phosphorus several competing models exist.

6.7 Approximation Algorithm for Correlation Matrices

Conclusions

An algorithm has been developed which allows to approximate Hermitian matrices by positive semidefinite Hermitian matrices. Two objectives of the algorithm are to minimize the approximation error as well as the condition number of the approximation. As it is usually not possible to achieve both objectives in an optimal way, one of them can be prioritized. Numerical tests have shown that the algorithm outperforms existing algorithms with regard to these two objectives.

The algorithm has asymptotically the same (worst case) execution time and memory consumption as the fastest algorithm to verify positive definiteness. Hence, it is usable for very large matrices. Moreover, it preserves the sparsity pattern of sparse matrices and is thus suitable for sparse matrices as well.

The unique feature of the algorithm is that diagonal values of the approximation can be bounded in advance. This makes it possible to obtain correlation matrices with this algorithm and makes it relevant for statistics. The algorithm may also be of interest in numerical optimization where often Hessian matrices must be approximated by positive definite matrices.

A decomposition of the approximation is calculated as a by-product and can be used, e.g., to solve associated linear equations in a fast and numerically stable way. An extensively tested, well documented and easy to install implementation of the algorithm is freely available.

The algorithm has already been used successfully in the statistical analysis of marine phosphate data where correlation matrices with billions of entries have been generated. It can certainly be useful in determining correlation matrices of other (climate) data and therefore help to understand the measured processes.

Outlooks

The permutation step in the algorithm offers many more ways of realization. It is yet not obvious which performs best and others could be developed and compared with those presented. In addition, one could try to merge different permutation strategies, e.g., a permutation method that reduces the approximation error and one that reduces the memory consumption for sparse matrices.

The algorithm allows to determine the modification of the original matrix in several ways. An iterative approach, which in each iteration modifies a part of the original matrix so that this modification is minimal in the Frobenius norm, was presented. This approach results usually in a small approximation error, however, not in the smallest possible. Other strategies may also be conceivable, especially when another norm, such as the spectral norm, is considered.

For lower restrictions on execution time and memory consumption, it would be possible to develop further strategies for permutation and modification. They could provide lower approximation errors and lower condition numbers, however, at the expense of execution time and memory consumption.

Besides, it certainly seems worth the effort to incorporate the algorithm into numerical optimization algorithms [15, 4.4] and benchmark the resulting optimization algorithm with a suitable selection of optimization problems.

6.8 Developed Software Packages

Conclusions

Several software packages were developed or extended in conjunction with this thesis. They are all available under open source licenses, were extensively tested and most of them are fully documented. Their source code and documentation meet the style conventions of the particular programming language. This allows to use the language specific help and documentation features and makes it easier to get familiar with the program and its source code.

The Optimal Experimental Design Toolbox [56] was developed to easily optimize model parameters and experimental designs in MATLAB [36]. Its documentation includes a step by step introduction based on application examples.

The matrix-decomposition library [52] provides several approximation and decomposition algorithms for dense and sparse matrices in Python [48] including the presented algorithms to approximate Hermitian matrices by positive semidefinite Hermitian matrices.

The measurements package [53] allows to process, analyze and visualize marine measurement data in Python. It is especially suited, but not limited, to data from the World Ocean Database.

The simulation package [54] allows to simulate marine models on high performance clusters. It provides an easy to use interface to run simulations and handles all necessary interactions with the cluster and the underlying simulation software METOS3D [45]. It allows to spin up models into an annual periodic state, approximate derivatives, estimate model parameters, quantify uncertainties, optimize experimental designs and visualize results.

The utillib package [55] is used in the measurements and the simulation package. It offers many utility functions that simplify, e.g., plotting, working with the file system and caching values. Using this package, computationally intensive results in the measurements and the simulation package are stored automatically in a database and are thus accessible any time without additional computational effort.

Significant contributions have also been made to the SciPy [62, 30] and scikit-sparse [57] packages. SciPy is an extensive and by far the most popular library for scientific computing in Python. The contributions were mainly related to sparse matrix handling. Existing functions have been considerably accelerated and new functions, e.g., for writing and reading sparse matrices to and from files, have been added. The software package scikit-sparse allows to decompose sparse matrices. Its functionality has been considerably extended by adding several new features.

In total, the self-written source code includes thirty-five thousand lines by now. This corresponds to about one thousand pages. Since parts of the source code were rewritten during the development of these packages, the total amount of written source code is considerably higher.

Outlooks

The Optimal Experimental Design Toolbox currently supports only the ordinary and the weighted least squares estimator so it might be useful to extend the toolbox by the generalized least squares estimator.

Although the implementation of the approximation algorithm described in Subsection 3.2, which is part of the matrix-decomposition library, was carefully tweaked for

computational efficiency, it could be accelerated even further by converting time consuming parts to a faster programming language like C or Fortran and embedding this code in the matrix-decomposition library using Cython [5, 9] or F2PY [41].

The globalization method used to find global minima in the model parameter estimation has turned out to be very beneficial. However it is only available in a closed source MATLAB toolbox. Since the algorithm itself is freely available, it would certainly be valuable to implement it in Python and make it freely available to the scientific community within the numerical optimization subpackage of SciPy.

References

[1] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. National Bureau of Standards Applied Mathematics Series 55. Tenth Printing.* ERIC, 1972.

[2] Alliance of World Scientists. World Scientists' Warning of a Climate Emergency. https://scientistswarning.forestry.oregonstate.edu/, 2019.

[3] R. C. Aster, B. Borchers, and C. H. Thurber. *Parameter Estimation and Inverse Problems*. Elsevier, second edition, 2013.

[4] E. B. Barbier, S. D. Hacker, C. Kennedy, E. W. Koch, A. C. Stier, and B. R. Silliman. The value of estuarine and coastal ecosystem services. *Ecological Monographs*, 81(2):169–193, 2011.

[5] S. Behnel, R. Bradshaw, C. Citro, L. Dalcin, D. Seljebotn, and K. Smith. Cython: The Best of Both Worlds. *Computing in Science Engineering*, 13(2):31 –39, Mar. 2011.

[6] G. R. Bigg. *The Oceans and Climate*. Cambridge University Press, second edition, 2003.

[7] T. Boyer, J. Antonov, O. Baranova, C. Coleman, H. Garcia, A. Grodsky, D. Johnson, R. Locarnini, A. Mishonov, T. O'Brien, C. Paver, J. Reagan, D. Seidov, I. Smolyar, and M. Zweng. *World Ocean Database 2013*. Silver Spring, 2013. S. Levitus, Ed.; A. Mishonov, Technical Ed.

[8] T. Boyer, O. Baranova, C. Coleman, H. Garcia, A. Grodsky, R. Locarnini, A. Mishonov, C. Paver, J. Reagan, D. Seidov, I. Smolyar, K. Weathers, and M. Zweng. *World Ocean Database 2018*. NOAA Atlas NESDIS 87, 2018.

[9] R. Bradshaw, S. Behnel, D. S. Seljebotn, G. Ewing, and et al. The Cython compiler. https://cython.org.

[10] P. Cadule, P. Friedlingstein, L. Bopp, S. Sitch, C. D. Jones, P. Ciais, S. L. Piao, and P. Peylin. Benchmarking coupled climate-carbon models against long-term atmospheric CO_2 measurements. *Global Biogeochemical Cycles*, 24(2), 2010.

[11] E. P. Chassignet and J. Verron. *Ocean modeling and parameterization*, volume 516. Springer Science & Business Media, 2012.

[12] G. Chavent. *Nonlinear Least Squares for Inverse Problems*. Springer Netherlands, 2010.

[13] E. Chong and S. Zak. *An Introduction to Optimization*. Wiley Series in Discrete Mathematics and Optimization. Wiley, 4th edition, 2013.

[14] A. Doicu, T. Trautmann, and F. Schreier. *Numerical regularization for atmospheric inverse problems*. Springer-Verlag Berlin Heidelberg, 2010.

[15] P. E. Gill, W. Murray, and M. H. Wright. *Practical Optimization*. Academic Press, London, 1981.

[16] P. J. Gleckler, K. E. Taylor, and C. Doutriaux. Performance metrics for climate models. *Journal of Geophysical Research: Atmospheres*, 113(D6), 2008.

[17] A. Griewank and A. Walther. *Evaluating Derivatives*. Society for Industrial and Applied Mathematics, second edition, 2008.

[18] S. Griffies, C. Böning, F. Bryan, E. Chassignet, R. Gerdes, H. Hasumi, A. Hirst, A.-M. Treguier, and D. Webb. Developments in Ocean Climate Modelling. *Ocean Modelling*, 2:123–192, 12 2000.

[19] S. M. Griffies. *Fundamentals of Ocean Climate Models*. Princeton University Press, Princeton, New Jersey, 2004.

[20] S. M. Griffies. Science of ocean climate models. *Elsevier*, 2008.

[21] N. J. Higham. *Accuracy and Stability of Numerical Algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2nd edition, 2002.

[22] N. J. Higham. Computing the Nearest Correlation Matrix - A Problem from Finance. *IMA J. Numer. Anal.*, 22(3):329–343, 2002.

[23] IPCC (Intergovernmental Panel on Climate Change). *Climate change: the 1990 and 1992 IPCC assessments; IPCC first assessment report overview and policymaker summaries and 1992 IPCC supplement*. Intergovernmental Panel on Climate Change, 1992.

[24] IPCC (Intergovernmental Panel on Climate Change). *IPCC second assessment; Climate Change 1995*. Intergovernmental Panel on Climate Change, 1995.

[25] IPCC (Intergovernmental Panel on Climate Change). *Climate Change 2001: Synthesis Report. A Contribution of Working Groups I, II, and III to the Third Assessment Report of the Integovernmental Panel on Climate Change*. Intergovernmental Panel on Climate Change, 2001.

[26] IPCC (Intergovernmental Panel on Climate Change). *Climate Change 2007: Synthesis Report. Contribution of Working Groups I, II and III to the Fourth Assessment Report of the Integovernmental Panel on Climate Change*. Intergovernmental Panel on Climate Change, 2008.

[27] IPCC (Intergovernmental Panel on Climate Change). *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013.

[28] IPCC (Intergovernmental Panel on Climate Change). *Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Integovernmental Panel on Climate Change*. Intergovernmental Panel on Climate Change, 2015.

[29] IPCC (Intergovernmental Panel on Climate Change). *IPCC, 2018: Global Warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty*. Intergovernmental Panel on Climate Change, 2018.

[30] E. Jones, T. Oliphant, P. Peterson, et al. SciPy: Open source scientific tools for Python. https://www.scipy.org, 2019. version 1.3.

[31] H. Kaper and H. Engler. *Mathematics and Climate*. Society for Industrial and Applied Mathematics, 2013.

[32] M. L. Kirwan and S. M. Mudd. Response of salt-marsh carbon accumulation to climate change. *Nature*, 489(7417):550–553, Sept. 2012.

[33] I. Kriest, S. Khatiwala, and A. Oschlies. Towards an assessment of simple global marine biogeochemical models of different complexity. *Progress in Oceanography*, 86(3-4):337–360, 2010.

[34] J. A. Langley, K. L. McKee, D. R. Cahoon, J. A. Cherry, and J. P. Megonigal. Elevated CO_2 stimulates marsh elevation gain, counterbalancing sea-level rise. *Proceedings of the National Academy of Sciences*, 106(15):6182–6186, 2009.

[35] E. Linacre. *Climate Data and Resources*. Routledge, 1992.

[36] MathWorks. *Matlab R2018a Primer*. Natick, Massachusetts, R2018a edition, Mar. 2018. visited on 2019-12-06.

[37] K. McGuffie and A. Henderson-Sellers. *A Climate Modelling Primer*. Wiley, Chichester, third edition, 2005.

[38] W. J. Mitsch and J. G. Gosselink. *Wetlands*. Wiley, Hoboken, NJ, fourth edition, 2007. XI, 582 S., zahlr. Ill., graph. Darst., Kt., 24 cm.

[39] J. D. Neelin. *Climate Change and Climate Modeling*. Cambridge University Press, 2010.

[40] J. Nocedal and S. Wright. *Numerical Optimization*. Springer series in operations research and financial engineering. Springer, New York, second edition, 2006.

[41] T. E. Oliphant et al. NumPy: N-dimensional array package for Python. https://www.numpy.org, 2019. Version 1.17.3.

[42] U. N. F. C. on Climate Change (UNFCCC). Paris agreement, 2015.

[43] G. M. Phillips. *Interpolation and approximation by polynomials*, volume 14. Springer Science & Business Media, 2003.

[44] R. Pincus, C. P. Batstone, R. J. P. Hofmann, K. E. Taylor, and P. J. Glecker. Evaluating the present-day simulation of clouds, precipitation, and radiation in climate models. *Journal of Geophysical Research: Atmospheres*, 113(D14), 2008.

[45] J. Piwonski and T. Slawig. Metos3D: the Marine Ecosystem Toolkit for Optimization and Simulation in 3-D – Part 1: Simulation Package v0.3.2. *Geoscientific Model Development*, 9:3729–3750, 2016.

[46] L. Pronzato and A. Pázman. *Design of Experiments in Nonlinear Models: Asymptotic Normality, Optimality Criteria and Small-Sample Properties*. Lecture Notes in Statistics. Springer, New York, NY, 2013.

[47] F. Pukelsheim. *Optimal Design of Experiments*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2006.

[48] Python Software Foundation. Python. http://www.python.org, 2018. version 3.7.

[49] H. Qi and D. Sun. Correlation stress testing for value-at-risk: an unconstrained convex optimization approach. *Computational Optimization and Applications*, 45(2):427–462, Mar 2010.

[50] P. Regtien, F. Van Der Heijden, M. J. Korsten, and W. Otthius. *Measurement Science for Engineers*. Elsevier, 2004.

[51] T. Reichler and J. Kim. How well do coupled models simulate today's climate? *Bulletin of the American Meteorological Society*, 89(3):303–312, 2008.

[52] J. Reimer. matrix-decomposition: a library for decompose (factorize) dense and sparse matrices in Python. https://doi.org/10.5281/zenodo.3558540.

[53] J. Reimer. measurements library: Python functions to handle, statistically analyze and plot measurement data. https://doi.org/10.5281/zenodo.3558700.

[54] J. Reimer. simulation library: Python functions for simulating mathematical models, estimating model parameters, quantifying uncertainties and visualizing results.. https://doi.org/10.5281/zenodo.3558702.

[55] J. Reimer. utillib library: a collection of helpful Python functions. https://doi.org/10.5281/zenodo.3558698.

[56] J. Reimer. The Optimal Experimental Design Toolbox for MATLAB. http://dx.doi.org/10.5281/zenodo.591918, Mar. 2015.

[57] J. Reimer, A. Grigorievskiy, A. Lee, Yuri, L. Barrett, D. S. Seljebotn, N. Smith, and D. Cournapeau. scikit-sparse: a library for sparse matrix manipulation in Python. https://github.com/scikit-sparse/scikit-sparse, Nov. 2018. version 0.4.4.

[58] W. J. Ripple, C. Wolf, T. M. Newsome, P. Barnard, and W. R. Moomaw. World Scientists' Warning of a Climate Emergency. *BioScience*, 11 2019. biz088.

[59] P. J. Rousseeuw and G. Molenberghs. Transformation of non positive semidefinite correlation matrices. *Communications in Statistics–Theory and Methods*, 22(4):965–984, 1993.

[60] G. A. F. Seber and C. J. Wild. *Nonlinear Regression*. Wiley series in probability and statistics. Wiley-Interscience, 2003.

[61] R. C. Smith. *Uncertainty Quantification: Theory, Implementation, and Applications*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2013.

[62] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, I. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy. SciPy 1.0-Fundamental Algorithms for Scientific Computing in Python. *arXiv*, abs/1907.10121, 2019.

[63] É. Walter and L. Pronzato. *Identification of Parametric Models from Experimental Data*. Communications and control engineering. Springer, New York, Jan. 1997.

[64] Z. Wang, D. Yi, X. Duan, J. Yao, and D. Gu. *Measurement data modeling and parameter estimation*. CRC Press, 2012.

[65] D. W. Waugh and V. Eyring. Quantitative performance metrics for stratospheric-resolving chemistry-climate models. *Atmospheric Chemistry and Physics*, 8(18):5699–5713, 2008.

Acknowledgement

My deepest gratitude goes to my supervisor Prof. Dr. Thomas Slawig for his support, our friendly and inspiring discussions as well as his valuable advice. I appreciate that I have been part of his research group for several years. He is also the one who raised my interest in numerical optimization and the related subjects for which I thank him as well.

I am grateful to Prof. Dr. Andreas Oschlies for his support as my co-supervisor and his helpful suggestions and explanations. I thank both of them for the opportunity to work on such an exciting and interesting interdisciplinary subject.

Many gratitude also goes to my colleagues who made it such a pleasure to come to work every day. I am grateful for the many interesting conversations with them, not only related to our research. In particular, I appreciate the proofreading of this thesis.

Last but not least, I thank my family and friends for their support and their patience and apologize for the time that has gone into this thesis instead of spending it with them.