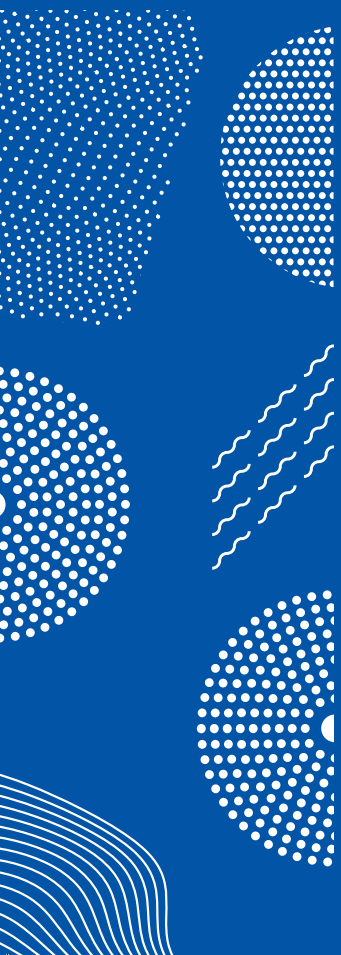
ILMATIETEEN LAITOS
METEOROLOGISKA INSTITUTET
FINNISH METEOROLOGICAL INSTITUTE

154
CONTRIBUTIONS

APPLICATIONS OF BAYESIAN COMPUTATIONAL STATISTICS AND MODELING TO LARGE-SCALE GEOSCIENTIFIC PROBLEMS

JOUNI SUSILUOTO



Finnish Meteorological Institute
Contributions 154

Applications of Bayesian computational statistics and modeling to large-scale geoscientific problems

Jouni Susiluoto

Academic dissertation

To be presented, with the permission of the Faculty of Science of the University of Helsinki, for public criticism in the hall "Athena" (#107), Siltavuorenpenger 3A, Helsinki, on October 16th 2019 at 12 o'clock noon.

Department of Mathematics and Statistics
University of Helsinki
Helsinki, Finland

Supervisor
Prof. Samuli Siltanen, University of Helsinki, Finland
Prof. Marko Laine, Finnish Meteorological Institute

Pre-examiners
Prof. Matti Vihola, University of Jyväskylä, Finland
Prof. Ralph Smith, North Carolina State University, USA

Opponent
Prof. Kody Law, University of Manchester, UK

Custos
Prof. Samuli Siltanen, University of Helsinki, Finland

Contact information
Department of Mathematics and Statistics
P.O. Box 64 (Gustav Hällströmin katu 2)
FI-00014 University of Helsinki
Finland
URL: http://mathstat.helsinki.fi/
Telephone: +358 29 419 11

Finnish Meteorological Institute
P.O. Box 503 (Erik Palménin aukio 1)
00101 Helsinki
Finland
URL: https://ilmatieteenlaitos.fi/
Telephone: +358 29 539 1000

ISSN 0782-6117
ISBN 978-952-336-080-8 (paperback)
ISBN 978-952-336-081-5 (PDF)
Helsinki 2019
Edita Prima Oy

ILMATIETEEN LAITOS
METEOROLOGISKA INSTITUTET
FINNISH METEOROLOGICAL INSTITUTE

Published by Finnish Meteorological Institute
(Erik Palménin aukio 1), P.O. Box 503
FIN-00101 Helsinki, Finland

Series title, number, and report code of publication
Contributions, 154, FMI-CONT-154
Date 30.9.2019

Author
Jouni Susiluoto

Title
Applications of Bayesian computational statistics and modeling to large-scale geoscientific problems

Abstract

Climate change is one of the most important, pressing, and furthest reaching global challenges that humanity faces in the 21st century. Already affecting daily lives of many directly and everyone indirectly, changes in climate are projected to have many catastrophic consequences. For this reason, researching climate and climate change is needed.

Studying complex geoscientific phenomena such as climate change consists of a patchwork of challenging mathematical, statistical, and computational problems. To solve these problems, local and global process models and statistical models are combined with both small *in situ* observation data sets with only few observations, and equally well with enormous global remote sensing data products containing hundreds of millions of data points. This integration of models and data can be done in a Bayesian inverse modeling setting if the algorithms and computational methods used are chosen and implemented carefully. The methods used in the four publications on which this thesis is based range from high-dimensional Bayesian spatial statistical models and Markov chain Monte Carlo methods to time series modeling and point estimation via optimization.

The particular geoscientific problems considered are: finding the spatio-temporal distribution of atmospheric carbon dioxide based on sparse remote sensing data, quantifying uncertainties in modeling methane emissions from boreal wetlands, analyzing and quantifying the effect of climate change on growing season in the boreal region, and using statistical methods to calibrate a terrestrial ecosystem model.

In addition to analyzing these problems, the research and the results help to understand model performance and how modeling uncertainties in very large computational problems can be approached, also providing algorithm implementations on top of which future efforts may be built.

Publishing unit
Climate System Research

Classification (UDC)
519.676, 551.588.7

Keywords
climate change, computational statistics, Markov chain Monte Carlo, Gaussian processes, Bayesian hierarchical models, carbon cycle, remote sensing, wetlands, methane emissions

ISSN and series title
0782-6117
Finnish Meteorological Institute Contributions

ISBN
978-952-336-080-8 (paperback)
978-952-336-081-5 (pdf)

DOI
https://doi.org/10.35614/isbn.9789523360815

Language
English

Pages
165

Selected publications

I J. Susiluoto, A. Spantini, H. Haario, Y. Marzouk. *Efficient multi-scale Gaussian process regression for massive remote sensing data with satGP v0.1*. Geosci. Model Dev. Discuss. 2019. URL https://doi.org/10.5194/gmd-2019-156

J.S. wrote the satGP program, designed most and implemented all of the novel computational ideas, designed and carried out the experiments, analyzed the results, and prepared the manuscript and the figures for publication.

II J. Susiluoto, M. Raivonen, L. Backman, M. Laine, J. Mäkelä, O. Peltola, T. Vesala, and T. Aalto. *Calibrating the sqHIMMELI v1.0 wetland methane emission model with hierarchical modeling and adaptive MCMC*. Geosci. Model Dev., 11, 1199-1228, 2018. doi: 10.5194/gmd-11-1199-2018. URL www.geosci-model-dev.net/11/1199/2018/

J.S. designed the research, implemented the algorithms, carried out the simulations, analyzed the results, and prepared the manuscript and the figures for publication.

III J. Pulliainen, M. Aurela, T. Laurila, T. Aalto, M. Takala, M. Salminen, M. Kulmala, A. Barr, M. Heimann, A. Lindroth, A. Laaksonen, C. Derksen, A. Mäkelä, T. Markkanen, J. Lemmetyinen, J. Susiluoto, S. Dengel, I. Mammarella, J-P. Tuovinen, and T. Vesala. *Early snowmelt significantly enhances boreal springtime carbon uptake.* Proceedings of the National Academy of Sciences, 114(42):11081–11086, 2017. ISSN 0027-8424. doi: 10.1073/pnas.1707889114. URL http://www.pnas.org/content/114/42/11081.

J.S. performed the climate model simulations, including the various auxiliary runs for creating the initial model states, post-processed the model output, and participated in analyzing the data.

IV J. Mäkelä, J. Susiluoto, T. Markkanen, M. Aurela, H. Järvinen, I. Mammarella, S. Hagemann, and T. Aalto. *Constraining ecosystem model with adaptive metropolis algorithm using boreal forest site eddy covariance measurements.* Nonlinear Processes in Geophysics, 23(6):447–465, 2016. doi: 10.5194/npg-23-447-2016. URL http://www.nonlin-processes-geophys.net/23/447/2016/.

J.S. developed and tested the sampling framework used for performing the simulations, provided initial modifications for the JSBACH model to efficiently work with that framework, and contributed to preparing the manuscript.

Partial list of symbols

Symbol	Meaning
$\mathbb{N}$	$\{0, 1, 2, 3, \ldots\}$
$\mathbb{N}^+$	$\{1, 2, 3, \ldots\}$
$[a, b)$	$\{x \in \mathbb{R} : a \leq x < b\}$
$\mathbb{R}^+$	$[0, \infty)$
$\triangleq$	Is defined to be
$\approx$	Is approximately equal to
$\propto$	Is proportional to
$\sim$	Is distributed according to
$a \leftarrow b$	a is assigned the value b
$\mathcal{N}(\mu, \Sigma)$	Multivariate normal distribution with mean vector μ and covariance Σ
Λ	Lebesgue measure
$f \circ g$	Composition of functions f and g
$\|x\|_\Gamma$	$\sqrt{x^T \Gamma^{-1} x}$
$\|K\|$	Determinant of matrix K
$\Gamma(s)$	Gamma function with argument s
$\xrightarrow{\mathrm{d}}$	Converges in distribution to
$\delta_{x,x'}$	Dirac delta function
$a \wedge b$	$\min(a, b)$
$\mu \ll \nu$	μ is absolutely continuous with respect to ν
$X \perp\!\!\!\perp Y$	X and Y are independent random variables
$1_{[a,b]}$	$1_{[a,b]}(x) = 1$ if $a \leq x \leq b$, otherwise 0.

Partial list of abbreviations

Abbreviation	Stands for
ARMA	Autoregressive moving average model (2.56)
DAG	Directed acyclic graph
ECMWF	European Center for Medium-Range Weather Forecasts
GPP	Gross primary production
GSSD	Growing season starting date
IPCC	Intergovernmental Panel on Climate Change
MAP	Maximum *a posteriori*
MCMC	Markov chain Monte Carlo
MLE	Maximum likelihood estimate
MRF	Markov random field
NLL	Negative log-likelihood
OCO-2	Orbiting Carbon Observatory 2 (NASA satellite)
PDE	Partial differential equation
i.i.d.	Independent and identically distributed
pdf	Probability density function
pmf	Probability mass function
s.t.	Such that

Acknowledgments

This work would not have been possible without the input of a large number of people: colleagues, family, friends, and others. For that simple reason, the following list is by no means exhaustive.

I would like to thank Profs. Samuli Siltanen from University of Helsinki (UHEL) and Marko Laine from the Finnish Meteorological Institute (FMI) for more and less formally advising me through the long doctoral process. I was lucky to get to work with you.

Thanks are also in order to Prof. Ari Laaksonen and head of the greenhouse gases group Tuomas Laurila, who initially employed me at the FMI greenhouse gases group solely based on a recommendation by Profs. Timo Vesala and Eero Nikinmaa (UHEL), both of who also deserve my gratitude. I sincerely thank Doc. Tuula Aalto, the group leader of the carbon cycle modeling group at the FMI, and Prof. Heikki Järvinen (UHEL), who initially pushed me closer to the applied statistics community at and around FMI.

I would like to thank all my co-authors, and in particular (in alphabetic order) colleague and friend Jarmo Mäkelä, Prof. Jouni Pulliainen, and Dr. Maarit Raivonen. Discussions with Prof. Johanna Tamminen and Dr. Janne Hakkarainen lead me to start looking into remote sensing-related problems, and that has been both a positive challenge and a pleasure.

From the years as a researcher at FMI, I want to thank the companionship and help of Drs. Leif Backman, Yao Gao, Sauli Lindberg (UHEL), Tiina Markkanen, Pirkka Ollinaho, Joni-Pekka Pietikäinen, Tea Thum, Aki Tsuruta, Simo Tukiainen, and others, for all kinds of (non-)peer support. Especially in the very first years, the company of Dr. Antti-Ilari Partanen (FMI) and Dr. Antti Solonen (LUT) was important.

For help with technical problems, which were not few nor far between, I would like to thank my friends Ilmari Karonen, Seppo Varho, and Jussi Virolainen. Veronika Gayler, Christian Reich and Reiner Schnur from Max Planck Institute for Meteorology in Hamburg, and Declan O'Donnell (FMI) helped me with ECHAM/JSBACH modeling issues. Regarding the text, I would like to thank Prof. Lassi Roininen for reading an early version and providing a long list of helpful comments.

The thesis work was improved immensely by the patience and bottomless knowledge of professors Heikki Haario (LUT) and Youssef Marzouk from Massachusetts

Institute of Technology (MIT), who made it possible for me to undertake and complete major parts of this work at MIT. Neither of you were my formal supervisors, but if senior colleagues were to be ranked by supervision hours, both of you would be very high on the list. Equally important was that working with you was a lot of fun.

I tip my imaginary hat to Alessio Spantini (MIT) for all the nice dinners and for all the things I learned from our collaboration, and also to Sebastian Springer (LUT) for the irreplaceable companionship in Boston. The uncertainty quantification group at the AeroAstro department of MIT, and in particular many of its graduate students and post-docs, taught me astonishingly much and it was a privilege to get to experience the extraordinarily supportive atmosphere that you all created. Some of these people are: Ricardo Baptista, Daniele Bigoni, Remy Lam, Alexander Marquez, Andrea Scarinci, Chaitanya Talnikar, Anirban Chaudhuri, Zheng Wang, and Benjamin Zhang.

A few very special mentions of people whose names did not come up yet: thank you to my long time friends Theo Kurten, Juuso Laatikainen, Pirkka-Pekka Laurila, Ari Keränen, Daniel Marszalec, and Esko Oksanen for all the good times.

Markku Koskinen, Matti Lindholm, and Tuukka Susiluoto took me many times to admire and appreciate the nature behind all the modeling, and those experiences always helped to remember why wasting hours in the lab and at the computer matters. Pekka Heinonen helped with various disasters in Finland while I was concentrating on research in Boston and that gave me peace of mind to work – you probably have no idea about how much it meant to me.

My parents Ilkka and Liisa always provided support and practical advise, and without them I would have stumbled many more times in the process. Thank you for always placing us children first in your priorities. My siblings Anne, Elina, Kaisa and Tuukka along with their families taught me irreplaceable life lessons, which emerged useful when navigating the frustrations and the unavoidable setbacks with research.

Last, I want to thank Emilija Rožukaitė for many good things, but most importantly for just being yourself. Without you, much of this work and many other even more exciting adventures in life would never have happened.

Cambridge, MA, May 2019
Jouni Susiluoto

Contents

1 Introduction

Climate change is one of the most critical challenges that humankind faces in the twenty-first century. It will potentially cause huge economic, societal and environmental disruptions, which the general public is in many parts of the world slowly starting to realize. In recent years politicians, scientists and news outlets among others have attributed events such as devastating hurricanes, forest fires, giant icebergs splintering away from glaciers, floods, catastrophic losses in biodiversity in pristine rainforests, extreme droughts, crop failures, and so on, to climate change. The research presented in this thesis strives to explain parts of the underlying mechanisms better.

Climate change is caused by heat-trapping gases, most notably carbon dioxide (CO2) and methane (CH4), that are released to the atmosphere both naturally and by humans. The radiative forcing potential of atmospheric carbon dioxide compared to the pre-industrial level is currently at 1.68 W/m^2 whereas that of methane is at 0.97 W/m^2, according to the latest report by the Intergovernmental Panel on Climate Change (IPCC) (IPCC, 2013). Other gases effect the radiative balance as well, but CO2 and CH4 are the most important ones, and by a wide margin.

The major source of CO2 is the burning of fossil fuels to power factories, cars, power plants, etc. The natural CO2 sources are dwarfed in comparison to these anthropogenic sources, without which the atmospheric CO2 concentrations would be stable. For methane, the anthropogenic sources include leaks from oil and natural gas fields, farming, landfills, and coal mining, but wetlands, where peat is anaerobically decomposed by *archaea* (prokaryotic organisms), are also an important component. The inner workings of wetlands differ widely from one to another, depending for instance on temperature, local plant species, soil chemistry, and availability of nutrients.

How carbon circulates in air, land, and water, is complicated and consists of a large number of processes. Much of the carbon dioxide emitted to the atmosphere is dissolved in water, little by little lowering the pH of the oceans. Another part is photosynthesized by plants, adding to the terrestrial carbon pool. The leftover CO2 stays to increase the atmospheric concentration, which during the last 30 years has risen by almost 20%. The second order mechanisms are complex – for instance changes in terrestrial and marine ecosystems affect their responses to the changing levels of atmospheric carbon dioxide.

An important part of most modern analyses of the carbon cycle is uncertainty

quantification. Uncertainty quantification tries to formally analyze and attribute sources of uncertainty in any estimates to different parts of the estimation process, such as the uncertainty arising from measurement and modeling errors. Evaluating the uncertainties sometimes critically changes outcomes of research, as was shown for instance by an Oxford study from summer 2018: after accounting for uncertainties, it was found to be plausible that there is no alien life in the Milky Way, contrary to the usual opposite conclusion from a non-probabilistic application of the Drake Equation.

For producing actionable climate-related scientific results, sources, sinks, and stocks of carbon need to be estimated, typically with complicated climate models and/or sophisticated statistical techniques. This task is not made easier by the intimate coupling of Earth's water and carbon cycles. In the research presented in this thesis, several such complications that arise from intertwining and interacting processes are looked at. Photosynthesis takes place by the action of plants opening their stomata, which inevitably lets out water vapor. In times of drought, wetland carbon decomposition changes from anaerobic to aerobic, but this behavior is difficult to model due to the nonlinear changes in the microbial populations affecting decomposition of organic matter. Finally, climate change affects the snow clearing date across the Boreal ecosystems, which is reflected in the total gross primary production during the growing season.

This thesis looks into both modeling different aspects of the carbon cycle and evaluating the associated uncertainties. The work utilizes site-level carbon dioxide and methane flux measurement data with time series analysis and Markov chain Monte Carlo (MCMC) (Gamerman, 1997) techniques, global climate modeling with large amounts of flux measurement data from all over the world, and remote sensing CO2 data from the NASA Orbiting Carbon Observatory 2 (OCO-2) satellite (Crisp et al., 2012; O'Dell et al., 2012) with stochastic processes and spatial statistics techniques.

The simplest way to estimate a quantity by modeling is to obtain a prediction by initializing the model according to best expert knowledge and data available and performing a single model simulation. This method is often used when the computer model in question is extremely expensive, as is the case with computational fluid dynamics models, which are used for e.g. climate and weather models and rocket engine or aircraft component performance simulations. An example of such direct simulation is also part of **Paper III**, where the gross primary production response to changing snow clearance date is evaluated and compared against flux measurement data with the ECHAM5/JSBACH/CBALANCE family of models (Roeckner et al., 2003a) from the Max Planck Institute for Meteorology in Hamburg. Due to a single simulation taking several weeks, no uncertainty quantification was possible.

If the model is computationally less demanding, statistical methods utilized for uncertainty quantification can be more sophisticated. **Paper IV** employs an MCMC algorithm to evaluate parameter posteriors of several parameters affecting the carbon and water cycles. Similarly to Paper III, Paper IV uses the JSBACH model, but restricts the spatial domain to measurement sites instead of performing the simu-

lations for a larger region. Pre-computed weather data is used for model forcing, saving remarkably in computation time by refraining from solving the complicated and expensive atmospheric component at each time step.

A wetland methane emission model is utilized in **Paper II** to analyze what parts of the methane production process are constrained by flux measurement data. Since the model is less complex, a more sophisticated modeling approach can be used for modeling uncertainties. An adaptive MCMC algorithm is employed in a Metropolis within Gibbs setting with hierarchical modeling of annually changing environmental parameters with an autoregressive moving average time series model for defining the error model. The results indicate that without further measurement data, it is very challenging to state the importance of the different processes. This is an important result, since there are enormous quantities of peat in the boreal wetlands, which might be eventually released by the thawing of the Siberian permafrost. The full parameter posterior uncertainty of such models has not been extensively evaluated earlier in literature, nor has a hierarchical approach been used.

In contrast to the flux measurement observations used in Papers II-IV, **Paper I** utilizes remote sensing CO2 measurements from the OCO-2 satellite. Remote sensing of greenhouse gas concentrations has obvious benefits compared to *in situ* measurements in that remote sensing provides almost global coverage and measures similarly everywhere. However, the approach also brings problems: gaps in data due to clouds, unknown biases and errors, and long distances between satellite trajectories. Utilizing remote sensing data for constructing time-dependent high resolution CO2 distributions is therefore still work in progress, and so far estimates published in the literature show overly smoothed CO2 fields, not being able to utilize the data to its full potential. The results we present should hence be an important opening: an open source multi-scale Gaussian process software, able to compute the demanding spatial statistics problem with enormous amounts of data (at least hundreds of millions of observations), and retaining the local fine structure. The computation enables calculating the posterior mean and marginal variances, but also drawing samples of random functions and calibrating covariance kernel parameters based on data. We additionally describe several novel ideas for covariance modeling, some of which have not been used either in this or any other context before, such as wind-informed kernels, multi-scale kernels, and periodic kernels. We validate the multi-scale approach with synthetic studies, and show initial applications of the methodology to the OCO-2 v9 data product.

The Papers described above underline the multidisciplinary nature of climate science. This thesis has to deal with all of those disciplines and therefore it contains some more fundamental aspects of mathematics (measure theory, probability, random functions), more applied topics (statistics), computational paradigms (programming, graphical models, inference algorithms), physical modeling (process/forward models), and climate science (analyzing the results). To communicate that full scientific process, most of these technical aspects are presented in Chapters 2 and 3. The

presentation of the mathematical theory in chapter 2 is not always comprehensive, since a full treatment would take up too much space. The topics are well known in literature, however, and the reader is referred to the literature cited below for further details.

For a general introduction to inverse problems, see e.g. Mueller and Siltanen (2012); Tarantola (2005). For general Bayesian and non-Bayesian statistics, see Gelman et al. (2013), Casella and Berger (2002), or Bickel and Doksum (2015, 2016). For the measure-theoretic foundations of probability, (Williams, 1991) is a good starting point. Kalman filter and dynamic linear models are treated in Särkkä (2013) and Durbin and Koopman (2012). For Gaussian processes, good references, and the ones mainly used for this thesis are Rasmussen and Williams (2006) and Santner et al. (2003). For an interesting measure-theoretic exposition of random functions but technically beyond the level required in this work, see e.g. Karatzas and Shreve (1998); Stroock (2018). A solid general treatment of Bayesian inverse problems, emphasizing infinite-dimensional settings, is also given by Stuart (2010), while Gamerman (1997) describes more comprehensively the fundamentals of MCMC.

This thesis is structured in the following way: Chapter 2 will explore theory, starting from a very short review of basic probability, introducing Bayes' theorem and inverse problems. It will then cover uncertainty, linear models, Gaussian random functions, and graphical models, through which the algorithms used in the Papers are explained. This is followed by a presentation of the Monte Carlo techniques used in the Papers, such as MCMC including Gibbs sampling, and importance sampling. The chapter ends with a discussion of hierarchical Bayesian models and autoregressive moving average (ARMA) time series modeling. Chapter 3 discusses the research in the Papers against the theoretical background, concentrating on computational issues and climate science. Chapter 4 contains a short discussion of how the analyses could be further improved, where the current limitations of the presented approaches are, and how some of the most straightforward lines of future work look like.

2 Theory

Quantifying uncertainty is based on the notion of probability. The uncertainty of predicted CO2 concentration in June 2050 in Helsinki can be given as a *credible interval*: with a given probability the concentration is between $x - \delta$ and $x + \delta$ ppm. Another way of describing the uncertainty is describing the *distribution* of possible values, for instance by stating that the predicted concentration is a random variable distributed according to, say, normal distribution with mean x and variance $\frac{\delta^2}{4}$.

Uncertainty quantification in geosciences is important, since it affects how to best evaluate risk. This includes for instance how to minimize expected (arbitrary) loss due to climate change, deforestation, particulate emissions, wildfires and natural disasters, among others.

Uncertainties are in this work predominantly quantified using the Bayesian paradigm and the essential theory for doing so is presented in this chapter. The readers who are intimately familiar with Bayesian models, time series, random functions, and associated algorithms, may choose to skip this review and only use it as reference when necessary. Likewise, the reader with very little mathematical background may just want to skip the chapter, since it is rather condensed and not suitable for self-study – for that purpose the references at the end of chapter 1 can serve as starting points. For the reader who is familiar with the problems described in especially Papers I-II, this chapter may provide valuable information about how to in practice go about solving the associated modeling and computational problems.

2.1 Probabilistic background and notation

2.1.1 Probability and random variables

Let Ω be the set of possible outcomes of an experiment. A σ-algebra of Ω, $\mathscr{F}$, is a set of subsets of Ω such that complements and countable unions of any $F \in \mathscr{F}$ are also members of $\mathscr{F}$, as is the full space Ω. A *probability space* is a triplet $(\Omega, \mathscr{F}, \mu)$, where μ is a probability measure, $\mu : \mathscr{F} \to [0, 1]$ s.t. $\mu(\emptyset) = 0$ and $\mu(\Omega) = 1$ (Bickel and Doksum, 2015). The sets $F \in \mathscr{F}$ are then called μ-*measurable*. Given spaces $(\Omega, \mathscr{F})$ and $(\Omega', \mathscr{F}')$, a mapping $h : \Omega \to \Omega'$ is a *measurable function* if $\forall F' \in \mathscr{F}'$ it is true that $h^{-1}(F') \in \mathscr{F}$ (Williams, 1991).

Measure μ is *absolutely continuous* with respect to ν, written $\mu \ll \nu$, if $\forall F \subset \mathscr{F}$,

it holds that $\nu(F) = 0 \Rightarrow \mu(F) = 0$. The measure ν is *σ-finite* if $\Omega = \cup_{i=1}^{\infty} F_i$ with $F_i \in \mathscr{F}$ and $\forall i$, $\nu(F_i) < \infty$. The *Radon-Nikodym Theorem* (Williams, 1991) states that given $\mu \ll \nu$ with ν σ-finite, there exists a function $g : \Omega \to [0, \infty]$ such that

$$\mu(F) = \int_F g(x) \mathrm{d}\nu(x). \tag{2.1}$$

The function g is called the *Radon-Nikodym derivative of μ with respect to ν*. Given in the above setting a second measurable space $(\Omega', \mathscr{F}')$ and a measurable function $h : \Omega \to \Omega'$, the *pushforward measure of* μ is denoted by $h_*(\mu) : \Omega' \to [0, \infty)$ with

$$h_*(\mu)(F') \triangleq \mu(h^{-1}(F')), \tag{2.2}$$

where $F' \in \mathscr{F}'$ (Peyré and Cuturi, 2018).

A *random variable* X is a measurable function from a probability space to a measurable space, $X : (\Omega, \mathscr{F}, \mu) \to (S, \mathscr{S})$, where $\mathscr{S}$ is a σ-algebra on the nonempty set S (e.g. Williams (1991)). In this work the random variables are generally real-valued, $S = \mathbb{R}$. The set Ω is called the *sample space* (Casella and Berger, 2002).

Given a real-valued random variable X as above, the *law of* X, $\mathcal{L}_X$, is defined as $\mathcal{L}_X \triangleq X^{-1} \circ \mu$. This can be thought of as the pushforward $X_*(\mu)(U)$ for sets $U \in \mathcal{B}(\mathbb{R})$ via (2.2), where $\mathcal{B}(\mathbb{R})$ denotes the standard Borel σ-algebra over $\mathbb{R}$. The (cumulative) *distribution function of* X (cdf) is then defined (Williams, 1991) by

$$F_X(a) \triangleq \mathcal{L}_X((-\infty, a]) = \mu(\{\omega \in \Omega : X(\omega) < a\}). \tag{2.3}$$

For all practical purposes, finite-dimensional random variables are often associated with probability density functions (pdf), and discrete with probability mass functions (pmf). The pdf of a real-valued random variable X, $f_X(x)$, if it exists, is defined via $\int_a^b f_X(x) \mathrm{d}\Lambda(x) = \Pr(a \leq X \leq b)$, where Λ denotes the standard Lebesgue measure. If $\mathcal{L}_X \ll \Lambda$, the pdf can also be described (Williams, 1991) as the Radon-Nikodym derivative of the law of X with respect to the Lebesgue measure,

$$f_X = \frac{\mathrm{d}\mathcal{L}_X}{\mathrm{d}\Lambda}. \tag{2.4}$$

A real-valued *random vector* is a random variable, which maps the sample space onto $\mathbb{R}^q$ for some $q \in \mathbb{N}$ (Casella and Berger, 2002). The definitions of pmf, pdf, and cdf generalize trivially. Functions of random variables are random variables.

For random variables X and Y the *joint density* is the pdf/pmf of the random vector (X, Y), and is denoted by $f_{X,Y}(x, y)$. The joint density can be *marginalized* over either of the arguments by integration, for instance by $f_X(x) = \int_{-\infty}^{\infty} f(x, y) \mathrm{d}y$ (Williams, 1991; Casella and Berger, 2002). Conditioning the random variable X on Y, denoted $X|Y$ defines a new random variable whose density function is called the *conditional density*, *of X given Y* and it is denoted and defined by $f_{X|Y}(x) = f_{X,Y}(x, y)/f_Y(y)$. The elementary *chain rule* is the definition of conditional probability written in the form $f_{X,Y}(x, y) = f_{X|Y}(x) f_Y(y)$.

The *expectation* of a function g of a random variable or vector $X : (\Omega, \mathscr{F}, \mu) \to \mathbb{R}^q$ is given by $\mathbb{E}[g(X)] = \int_\Omega g(X(\omega))\mathrm{d}\mu$ (Casella and Berger, 2002), which, given a density function $f_X(x)$ for X, boils down to $\mathbb{E}[g(X)] = \int_\Omega g(x)f_X(x)\mathrm{d}\Lambda(x)$. With that the *covariance* of a random variables X and Y is given by $\mathrm{Cov}(X, Y) = \mathbb{E}[X - \mathbb{E}[X]]\mathbb{E}[Y - \mathbb{E}[Y]]$, with *variance* of X defined as $\mathrm{Cov}(X, X)$ and written as $\mathbb{V}[X]$. The covariance matrix C of a random vector X has elements $C_{ij} = \mathrm{Cov}(X_i, X_j)$ (Casella and Berger, 2002). The *correlation* between random variables X and Y is defined as $\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathbb{V}[X]\mathbb{V}[Y]}}$.

For finite samples $x^1 \dots x^N$ from any distribution, the unbiased estimates of the mean and covariance are given by $\overline{x} = \frac{1}{N}\sum_{i=1}^N x^i$ and $S = \frac{1}{N-1}\sum_{i=1}^N (x^i - \overline{x})(x^i - \overline{x})^T$ respectively (Casella and Berger, 2002). In this thesis sample sizes are generally very large, and therefore sample covariances are also often denoted by letter C. The elements C_{kl} describe the covariance between the k^{th} and l^{th} dimension of vectors x. With pairs of vector data $(x^1 \dots x^n, y^1 \dots y^n)$, correlation refers to a matrix with the *Pearson correlation coefficients* as elements, as in

$$\mathrm{Corr}(x^1 \dots x^n, y^1 \dots y^n)_{kl} = \frac{1}{N-1}\sum_{i=1}^N \frac{(x_k^i - \overline{x}_k)(y_l^i - \overline{y}_l)^T}{\sqrt{\mathbb{V}[X_k]\mathbb{V}[X_l]}}. \tag{2.5}$$

Two sub-σ-algebras $\mathscr{F}_1, \mathscr{F}_2$ of $\mathscr{F}$ – that is, they are σ-algebras and $\mathscr{F}_1, \mathscr{F}_2 \subseteq \mathscr{F}$ – are independent if $\Pr(x \in F_1 \cap F_2) = \Pr(x \in F_1)\Pr(x \in F_2)$ for all $F_1 \in \mathscr{F}_1$, $F_2 \in \mathscr{F}_2$. Two random variables *independent*, if their σ-algebras are independent, written $X \perp\!\!\!\perp Y$ (Williams, 1991). In practice for distributions with well-behaving density functions this translates to that two random variables X and Y are independent if $f_{X,Y}(x, y) = f_X(x)f_Y(y)$ (Casella and Berger, 2002). In addition, they are *conditionally independent* given a third random variable Z, written $X \perp\!\!\!\perp Y|Z$, if $(X|Z) \perp\!\!\!\perp (Y|Z)$ (Bickel and Doksum, 2015).

In the Bayesian formulation of probability (e.g. Gelman et al. (2013)), which is followed in this work, it is conventional to write $p(x)$ instead of $f_X(x)$ for a random vector X and from here on that notation is adopted, except for where reference to the particular random vector is explicitly desired. While traditionally in the *frequentist* view, any model parameters are treated as unknown constants, in the *Bayesian* setting they are modeled as random variables and the object of interest then is how those parameters are distributed.

The famous *Bayes' theorem*, on which Bayesian probability theory and statistics are based (Gamerman, 1997), states that for arbitrary random variables X and Y,

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}, \tag{2.6}$$

and is directly proven by noting that $p(y|x)p(x) = p(y, x) = p(x|y)p(y)$. The term on the left hand side is the *posterior distribution* of X, the term $p(y|x)$ is called the *likelihood*, and despite the notation is considered to be a function of x, $p(x)$ is the

prior distribution that codifies all our *a priori* knowledge of the parameters X, and $p(y)$ is the *marginal likelihood* of observations y, or sometimes *evidence* (Gelman et al., 2013; Tarantola, 2005), that usually cannot be computed in closed form.

Bayes' formula presents a way to update our knowledge regarding a random variable by updating the prior distribution with new data. Let μ^y denote the *posterior measure*, the measure corresponding to $p(x|y)$ in (2.6) as its pdf, and let μ^0 similarly denote the *prior measure* with pdf $p(x)$. With fixed data y, using (2.1) and (2.6), and given some μ^y-measurable set F,

$$\begin{aligned} \mu^y(F) = \int_F p(x|y)\mathrm{d}x \propto \int_F p(y|x)p(x)\mathrm{d}x = \int_F p(y|x)\mathrm{d}\mu^0 \\ \Rightarrow p(y|x) \propto \frac{\mathrm{d}\mu^y}{\mathrm{d}\mu^0} \end{aligned} \tag{2.7}$$

meaning that the likelihood is proportional to the Radon-Nikodym derivative of the posterior with respect to the prior. The benefits of this approach is discussed in detail by e.g. Stuart (2010).

2.1.2 Model calibration via Bayes' theorem

Bayesian calibration of a nontrivial model M, where the evidence term cannot be evaluated in closed form (e.g. a weather model or some other partial differential equation (PDE) model) is carried out by evaluating the nominator of the right hand side of the Bayes' theorem (denominator can be dealt with algorithmically, see section 2.6.1).

Models used in geophysics and in this work are often discretized in time but the dynamics evolve continuously in time, meaning that the time parameter is in some continuous space, $t \in \mathbb{R}$. Let

$$x \triangleq M(\theta, x_0) \tag{2.8}$$

be the output of a discretization of such a dynamical model for the time-evolution of some initial state vector x_0 governed by parameters $\theta \in \Theta$, where Θ is some set, typically $\mathbb{R}^q$ with some $q \in \mathbb{N}^+$. The observations are denoted by $y \in \mathcal{Y}$, and a function, $\phi \in \mathcal{Y}^{\mathcal{X}}$, called the *observation operator*, is used for mapping the space of model paths $\mathcal{X}$ to the space of observables, $\mathcal{Y}$ (Stuart, 2010). These are in principle some Banach spaces (normed complete linear spaces, see e.g. Rudin (1987)) – for example L^p-spaces – but for discussing a finite number of states and observations, finite-dimensional vector spaces suffice. In practice, in this work the mapping ϕ is the identity map, since the models are (unrealistically) thought to represent real physical quantities.

For Bayesian estimation of parameters in the context of such a system, the *observation equation* or *model equation* (Stuart, 2010) can in case of additive error – perhaps the most common situation – be written as

$$y = \phi(x) + \epsilon, \tag{2.9}$$

where $\epsilon \sim \nu$, where ν is the density function of some probability measure. This density ν, according to which the model-observation mismatch is modeled, is in this work sometimes referred to as an *error model*.

Equation (2.9) defines the likelihood term in (2.6),

$$p(y|\theta) = \nu\left(y - \phi(x)\right), \tag{2.10}$$

where explicit dependence on the initial state has been suppressed. Time-discretized versions of (2.8) and (2.9) can then be written as

$$x_i \triangleq M(t_i; \theta, x_0) \quad \text{and} \tag{2.11}$$

$$y_i \triangleq \phi(x)_i + \epsilon_i, \tag{2.12}$$

with states x_i and observations $y_i \triangleq y(t_i)$ taken at times $t_{i=1\ldots N}$ and with the discretization of the continuous states mapped by the observation operator defined as $\phi(x)_i \triangleq \phi(x)_{|t=t_i}$. Another common model for the observations y substitutes a multiplicative error for the additive one in (2.9).

The choice of ν dictates how the model-observation mismatch is expected to be distributed and the particular form of ν is a modeling choice, which can be justified by making sure that the residuals obtained by sampling the posterior $p(\theta|y)$ are distributed according to ν. This is a difficult step: first, the residuals may change unexpectedly with θ, especially with models with chaotic dynamics, and second, changing the values of any auxiliary model parameters or how the autocorrelation structures are modeled also affect how ν should be picked. Since in reality the model-observation mismatch may be a result of several different processes with different statistical characteristics (e.g. Tarantola (2005), Ex. 1.22), the final choice of ν is often a well-justified compromise.

2.1.3 Forward models and inverse problems

The computer implementation of a mathematical model $M(\theta, x_0)$ as described in (2.8) is called a *forward model*, and it is used to solve a *forward problem* (Mueller and Siltanen, 2012; Tarantola, 2005), yielding a discretization of the continuous model trajectory x given initial conditions and any required parameters. If M is computationally extremely demanding, solving the forward problem only once may be the best available approach. Executing the forward model alone does not normally, however, provide information about the model parameter uncertainties, nor does it produce new information about the values of the model parameters.

The *inverse problem* (Mueller and Siltanen, 2012; Tarantola, 2005) associated with the forward problem can be solved to provide estimates of θ and x_0, either with uncertainties or without. For obtaining point estimates, the problem can take any of

the forms

$$\hat{\theta} = \arg\min_{\theta} \mathscr{L}(\theta) \tag{2.13}$$

$$\hat{x}_0 = \arg\min_{x_0} \mathscr{L}(x_0) \tag{2.14}$$

$$\hat{\theta}, \hat{x}_0 = \arg\min_{\theta, x_0} \mathscr{L}(\theta; x_0), \tag{2.15}$$

where $\mathscr{L}$ is a suitable loss function, for instance a negative logarithm of a likelihood (NLL) given some observations and an observation model. Variations of this particular form, $\mathscr{L}(\theta) = -\log p(y|\theta)$, are used widely in this work. They are treated in more detail in section 2.1.4.

The famous *Hadamard conditions* (Mueller and Siltanen, 2012) state, that a problem is *ill-posed*, if it does not have a solution, if the solution is not unique, or if the solution is not a continuous function of the initial conditions. If none of these conditions (i.e. remove the word not) hold, the problem is *well-posed*. In geophysics all three conditions are often true, meaning that problems are strongly ill-posed, and the practical implications of this often is that the inversion presented in (2.13 - 2.15) gets stuck in local minima since the optimization problems are very rarely convex.

For linear problems with Gaussian errors, (2.13) has a closed-form solution. However, adding noise often quickly shatters the usability of the naive inversion – inverting the forward model – in many systems. This can be avoided by perturbing the setup and adding a regularization term, the most commonly used one of which is the *ridge regression* or *Tikhonov regularization* (Mueller and Siltanen, 2012), which in the Bayesian setting with log-likelihood as the loss function amounts to incorporating a Gaussian prior to the optimization problem in (2.13 - 2.15),

$$\hat{\theta} = \arg\min_{\theta} \mathscr{L}(\theta) + \alpha\|\theta\|_2^2, \tag{2.16}$$

where α is a regularization parameter controlling the prior weight. By using different forms of the regularization term such as $\|\theta\|_\Gamma^2$ for some positive definite matrix Γ, different types of prior formulations can be prescribed.

In the context of geophysical models, closed-form solutions for the inversion are not available and optimization algorithms need to be used. The present work utilizes L-BFGS (Nocedal, 1980), BOBYQUA (Powell, 2009), and Nelder-Mead (Nelder and Mead, 1965) algorithms for solving for point estimates in various inverse problems. For state estimation, the 4DVAR algorithm (Dimet and Talagrand, 1986) is commonly used in numerical weather prediction, and the Kalman filter family of algorithms can be utilized for both state and parameter estimation.

While prescribing a prior alone may work, several other approaches are available to work around the problem of local minima in the response surface of the loss function. For obtaining point estimates, stochastic optimization algorithms such as stochastic gradient descent have become popular recently, especially in the machine learning

community. While this is a viable approach, it is not feasible when the gradients are not available. For moderate parameter dimension, scaling down $\mathscr{L}$ to ensure mixing and using a Markov chain Monte Carlo algorithm to find $\mathbb{E}[\theta]$ can work reliably and sometimes be faster than using global optimization algorithms such as ISRES (Runarsson and Xin Yao, 2005).

2.1.4 STANDARD POINT ESTIMATION AND CROSS VALIDATION METHODS

If $\mathscr{L}(\theta) = -\log p(y|\theta)$ in (2.13), the corresponding $\hat{\theta}$ is called a *maximum likelihood estimate* of θ. By setting $\mathscr{L}(\theta) = -\log p(\theta|y)$, the *maximum a posterior estimate* (MAP) is obtained instead (Casella and Berger, 2002; Stuart, 2010). Since log is a monotonous function, its presence above is not necessary, but often convenient. The MAP estimate corresponds to $\hat{\theta}$ in (2.16) when $\mathscr{L}$ in that equation is the NLL. While maximum likelihood estimation is useful and often used, it may overestimate the confidence in the predictive performance of the model. If observation/model error covariance is not fully known, it is relatively common practice in geosciences to use a diagonal covariance model with a Gaussian likelihood as a best guess (for an example, see e.g. Mäkelä et al. (2016)).

From a Bayesian perspective, overconfidence with predictive performance is a general issue for point estimation, since any predictive quantities obtained by using a point estimate for model parameters do not reflect the uncertainty that should be carried over by the propagation of uncertainty in those model parameters. This may be overcome by using *cross-validation* (Gelman et al., 2013), where the cross-validation prediction error for a set of observations A_i is estimated by excluding that set from the training set, obtaining an estimate for the model parameters θ denoted by $\hat{\theta}^i_{\mathrm{XV}}$ using that training set, and then predicting the observations in A_i using the parameters $\hat{\theta}^i_{\mathrm{XV}}$ and finally comparing to the true observed quantities. Usually, $\cup_{i=1}^{M} A_i = A$ and $A_i \cap A_j = \emptyset$ when $i \neq j$. A much-used special case is when, for all i, $A_i = \{y_i\}$. This is called *leave one out cross validation* (LOOCV) (Gelman et al., 2013). Cross validation is used in a regression modeling setting in Paper II to evaluate what independent variables best explain annual model parameter variations produced by the hierarchical model used.

2.2 UNCERTAINTY

2.2.1 SOURCES OF UNCERTAINTY

The term ϵ in (2.9) describes the total uncertainty in the model-observation mismatch $y - \phi(x)$. In reality it needs to describe errors from various sources. If characteristics of separate sources of model-observation mismatch are known, ϵ can and should be split into several different components (Stuart, 2010). In Paper II, where the model-data mismatch is known to change in time, ARMA modeling is used to describe

the correlation structure in the time series while additional modeling accounts for heteroscedasticity.

The most straightforward error source to describe is often the *measurement error*, which describes the error contribution from sensor noise of the instrument making the measurement. This error component is typically assumed to be independent and identically distributed (i.i.d.) Gaussian. However, for instance in the case of CH4 flux measurements in Paper II, it is better described by the Laplace distribution (Richardson et al., 2006).

For discretized dynamical models, *representation error* describes how averaging over a finite domain (e.g. time-space hypercube of a grid point from one model time step to the next) to compare with localized observations induces error (e.g. Ganesan et al. (2014)). This source is controlled by the exact form of ϕ.

Other sources are *random model error* and *model bias* due to for instance rare or unmodeled events or incomplete information about the initial conditions, *autocorrelation* of the observation errors, and *numerical error* from finite machine precision.

While problems arising from machine precision can be a nuisance, e.g. when calculating a Cholesky factorization, $C = L^T L$ (Trefethen and Bau, 1997; Boyd and Vandenberghe, 2004) of a covariance matrix with a large condition number using single precision, a greater difficulty with geophysical models (and many other models as well) is caused by model bias and unmodeled events. An example of such an event is extreme drought in Finland, where photosynthesis is normally not limited by the availability of water, and models typically have not needed to take that to account.

2.2.2 DISTRIBUTIONS FOR UNCERTAINTY MODELING

In the context of any specific problem, the form of ϵ in the observation equation (2.9) needs to be prescribed. The choices utilized in the various problems tackled in this thesis are presented in this section.

A random vector X following the *normal* or *Gaussian distribution* (Casella and Berger, 2002) with mean μ and covariance matrix C has the probability density function

$$\mathcal{N}(\mu, C) \triangleq f_X(x) = (2\pi)^{-\frac{n}{2}} |C|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\|x-\mu\|_C^2\right), \tag{2.17}$$

where $\|x-\mu\|_C$ stands for $\sqrt{(x-\mu)^T C^{-1}(x-\mu)}$. If the random vector X is split into two parts of sizes p and q, i.e. $X = (X_1, X_2)^T$, then the joint distribution can be written as

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} C(X_1, X_1) & C(X_1, X_2) \\ C(X_2, X_1) & C(X_2, X_2) \end{bmatrix}\right) \tag{2.18}$$

and the conditional distribution $f_{X_1|X_2}$ is Gaussian with its moments given by

$$\mathbb{E}[X_1|X_2] = \mu_1 + C(X_1, X_2)C(X_2, X_2)^{-1}(X_2 - \mu_2) \tag{2.19}$$

$$\mathrm{Cov}(X_1|X_2) = C(X_1, X_1) - C(X_1, X_2)C(X_2, X_2)^{-1}C(X_2, X_1). \tag{2.20}$$

The right hand side in (2.20) is known as the *Schur complement* of $C(X_2, X_2)$ (Boyd and Vandenberghe, 2004), and in the setting of (2.18), the marginal distribution $\int_{\mathbb{R}^q} f_{X_1,X_2}(x_1, x_2)\mathrm{d}x_2$ is also Gaussian.

For any random variable Z, with mean μ_z and finite variance σ_z^2, the *central limit theorem* (CLT) states, that at the limit when $N \to \infty$, $\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{Z_i - \mu_z}{\sigma_z} \xrightarrow{\mathrm{d}} \mathcal{N}(0, 1)$ (e.g. Williams (1991); Vershynin (2018)). In practice this means that large-sample averages are well-behaved in that their tails are controlled by the squared exponent in (2.17). The CLT does not, however, state how fast the tail probabilities vanish – this depends on what kind of a random variable Z is. For instance for sub-Gaussian random variables (tails probabilities decaying at least at squared exponential speed) Hoeffding's inequality gives the exact tail bounds (Vershynin, 2018).

The χ^2*-distribution* with $k \in \mathbb{N}^+$ degrees of freedom describes the distribution of the sum of squares of k standard normal $\mathcal{N}(0, 1)$ random variables and hence is supported on $x > 0$. The weighted sum of squares from the quadratic form in the log-likelihood of a normal observation model, (2.17), is χ^2-distributed, given that in that equation the Cholesky factor of C whitens the residuals $x_i - \mu_i$.

The *scaled inverse* χ^2*-distribution* (Gelman et al., 2013) adds a scale parameter $s > 0$, and has the pdf

$$f_X(x) = \left(\frac{k}{2}\right)^{\frac{k}{2}} e^{\left(\frac{2x}{ks^2}\right)} \frac{s^k x^{-(\frac{k}{2}+1)}}{\Gamma(k/2)} \tag{2.21}$$

with $\mathbb{E}[X] = \frac{s^2 k}{k-2}$ and $\mathbb{V}[X] = \frac{2k^2 s^4}{(k-2)^2(k-4)}$. It can be used for e.g. prescribing priors for variance parameters.

Often in finite sample sizes the tails of the normal distribution are too thin. A heavier-tailed version to be used in finite-sample settings would be the *Student's t-distribution*, but we utilize the *two-sided exponential* or *Laplace distribution* (Casella and Berger, 2002) instead, with pdf

$$f_X(x) = \frac{1}{2\sigma} \exp\left(-\frac{|x - \mu|}{\sigma}\right), \tag{2.22}$$

where μ and $\sigma > 0$ are the location and scale parameters, respectively. Additionally, $\mathbb{E}[X] = \mu$ and $\mathbb{V}[X] = 2\sigma^2$. Flux measurements done with instruments designed to be used for measuring trace gas fluxes at the biosphere-atmosphere boundary have been reported to follow the Laplace distribution instead of the normal distribution.

The *uniform distribution* is denoted by $\mathcal{U}_{[0,1]}$ and has the probability density function $f_X(x) = 1_{[0,1]}$. If X follows the discrete *Bernoulli distribution* with parameter p, denoted $X \sim \mathrm{Ber}(p)$, it has the probability mass function $f_X(x) \in \{0, 1\}$ s.t. $\Pr(X = 1) = p$ (Casella and Berger, 2002).

All of the aforementioned continuous distributions belong to or are closely related to the *Gamma family* of distributions, an exponential family (Bickel and Doksum, 2015). This is convenient and by design, since using distributions from the same

family results in conjugacy that can be exploited when used in e.g. hierarchical models, as described in section 2.7.

2.3 LINEAR REGRESSION

One of the most commonly used tools in any context to statistically analyze data is *linear regression* (Casella and Berger, 2002), which amounts to fitting parameters controlling the orientation of a hyperplane to minimize squared error between that plane and some data. Let $A \in \mathbb{R}^{p\times n}$ be a data matrix containing n vector-valued measurements, called *independent variables*, of length p in the columns, let $y \in \mathbb{R}^n$ be a vector of *dependent variables*, and let $\beta \in \mathbb{R}^p$ be a vector of *regression coefficients* with prior covariance Σ. The regression problem is written as

$$A^T\beta = y + \epsilon, \tag{2.23}$$

where $\epsilon \sim \mathcal{N}(0, \Gamma)$ is the measurement error associated with y. For an exactly determined or overdetermined system, $\text{rank}(A) \geq p$, the (Tikhonov-)*regularized least squares solution* of (2.23) and its covariance are given by (Tarantola, 2005)

$$\begin{aligned}\mathbb{E}[\beta|y] = \hat{\beta} &= \arg\min_{\beta}\left\{\|A^T\beta - y\|_\Gamma^2 + \|\beta\|_\Sigma^2\right\} \qquad (2.24)\\ &= (A\Gamma^{-1}A^T + \Sigma^{-1})^{-1}A\Gamma^{-1}y, \text{ and}\\ \text{Cov}(\beta|y) &= (A\Gamma^{-1}A^T + \Sigma), \qquad (2.25)\end{aligned}$$

where the notation $\|\cdot\|_\Sigma$ was defined in the context of (2.17). If in this equation $\Sigma^{-1} = 0$, $\hat{\beta}$ is the *ordinary least squares solution*. As an alternative, the use of sparsity-inducing ℓ^1-norm for regularization is customary in big data applications, but this comes at the cost of needing to use an algorithm such as the *least absolute shrinkage and selection operator* (LASSO) for obtaining $\hat{\beta}$ (Tibshirani, 1996). In this work only ridge regression-type regularization and Gaussian priors are used, however. For further details, see e.g. (Lassas and Siltanen, 2004).

2.4 GAUSSIAN PROCESSES

Given an index set $\mathscr{D}$, a *stochastic process* or *random function* is an indexed set of random variables $X_d : (\Omega, \mathscr{F}, \mu) \to (S, \mathscr{S})$ for all $d \in \mathscr{D}$ (Williams, 1991), and the space S is usually taken to be $\mathbb{R}^d$ with the Borel σ-algebra $\mathcal{B}(\mathbb{R}^d)$. A classical example of a stochastic process is the Wiener process (random walk), for which $\mathscr{D} = \mathbb{R}^+$, and it holds that

$$d_1 < d_2 < d \Rightarrow (X_{d_1} \perp\!\!\!\perp X_d)|X_2, \text{ and} \tag{2.26}$$

$$X_d|X_{d_2} \sim \mathcal{N}(x_{d_2}, d - d_2). \tag{2.27}$$

A *Gaussian process*, or *Gaussian random function* is a stochastic process indexed over an index set $\mathscr{D}$ defined by a *mean function* $m(d)$ and a *covariance function* $k(d, d')$ in that for any finite collection of size N of indices $D \subset \mathscr{D}$, the joint distribution of random variables $\{X_d : d \in D\}$ is multivariate Gaussian (Rasmussen and Williams, 2006) with

$$X_D \sim \mathcal{N}(m, K), \tag{2.28}$$

where $X_D = (X_{d_1}, \ldots, X_{d_N})^T$, $m = (m(d_1), \ldots, m(d_N))^T$ and K is a matrix with elements $K_{ij} = k(d_i, d_j)$. The requirement that K is a covariance matrix implies that k is symmetric in its arguments. While the measure-theoretic treatment of stochastic processes leads to many interesting results, this level of mathematical detail is not needed here; see e.g. (Karatzas and Shreve, 1998; Stroock, 2018; Rozanov, 1998; Øksendal, 2010) for further details.

If the index set $\mathscr{D}$ is two-dimensional, the term *Gaussian field* is often used in the literature. For a time-dependent process, the index set is usually taken to be $\mathbb{R}^+$ and, non-surprisingly, the letter t is used. For the Gaussian processes in Paper I, a spatio-temporal index set is used and the index set elements are denoted by x. A Gaussian process is *stationary* if its unconditional mean and covariance functions do not change under translation.

That a quantity of interest Ψ is modeled as a Gaussian process with mean and covariance functions $m(d)$ and $k(d, d'; \theta)$, is written $\Psi \sim \mathrm{GP}(m(d), k(d, d'; \theta))$, following Rasmussen and Williams (2006) and Gelman et al. (2013). Given a set of observations $\psi_D \in \mathbb{R}^n$ of the quantity of interest Ψ indexed by some index set $D \subset \mathscr{D}$, the log marginal likelihood for a given set of covariance kernel parameters θ can be directly evaluated (Rasmussen and Williams, 2006) via (2.17) as

$$\log p(\psi_D|\theta) = -\frac{1}{2}\|\psi_D - m\|_K^2 - \frac{1}{2}\log|K| - \frac{n}{2}\log(2\pi). \tag{2.29}$$

The mean function selection affects the maximum likelihood estimate of the covariance parameters given observed data, and the decision of what to include in the mean function and what to leave for the covariance function is a modeling choice.

For calculating the marginal distribution of Ψ at some *test input* $d^* \notin D$, $d^* \in \mathscr{D}$ conditioned on observations ψ_D, equations (2.18 - 2.20) can be directly employed, with $X_1 = \Psi^*$ and $X_2 = \psi_D$. Here Ψ^* denotes the marginal distribution of random field Ψ at test input d^*. The covariance $K(\psi_D, \psi_D)$ then has the elements $k(d, d')$ with $d, d' \in D$.

An alternative way to model the evolution of randomness in a dynamical system is a *dynamic linear model* (DLM), in which a state space model is used in conjunction with the Kalman filter or Kalman smoother algorithms for parameter estimation (Durbin and Koopman, 2012; Gamerman, 1997). Given a linear model M and a Gaussian observation model in (2.9), the Kalman filter first predicts $x_t|x_{t-1}$, the state at time t given the state at time $t-1$ and its covariance, and then updates those estimates using any available observations at time t. In Paper I the DLM approach could

have been utilized for state estimation, much like was done in Laine et al. (2014), even though due to the size of the problem and the nature of the data this would have been challenging.

2.4.1 A PARAMETRIC FORM FOR THE GAUSSIAN PROCESS MEAN FUNCTION

In case the mean of a Gaussian process prior is not zero, a convenient way of prescribing it is via the parametrization (Santner et al., 2003)

$$m(d) = \sum_{i=1}^{p} \zeta_i(d)\beta_p, \tag{2.30}$$

where ζ_i are some functions of index d that are expected to capture the dynamics of the variation, and β_i are coefficients that can be determined for best fit.

Let F be a matrix with elements $F_{ij} = \zeta_i(d_j)$ and θ be the covariance function parameters. The least-squares solution for the β-parameters are once again given by (2.23 - 2.25), with $A \leftarrow F$, $\Gamma \leftarrow K$, and $y \leftarrow \psi$, yielding

$$\beta|\psi, \theta \sim \mathcal{N}\left((FK^{-1}F^T)^{-1}FK^{-1}\psi, (FK^{-1}F^T)^{-1}\right). \tag{2.31}$$

While this form is useful in that knowing the covariance model it allows one to get closed-form point estimates of the β-factors and their uncertainties, it does not cover non-linear cases such as parameters appearing in the arguments of a non-linear ζ_i.

2.4.2 GAUSSIAN PROCESS COVARIANCE KERNELS

There are several standard parameterized forms for describing covariance kernels, and of those the *exponential*, *Matérn*, and *periodic kernels* are utilized in Paper I. The notation presented here is from that paper, and it is reused in chapter 3.

Let θ be the set of parameters controlling a covariance kernel, often containing at least a *scale parameter* ℓ, and a *maximum covariance parameter* τ^2. As a shorthand, let

$$\xi^{\gamma}_{\ell_I}(d, d') = \sum_{c \in I} \frac{|d_c - d'_c|^{\gamma}}{\ell_c^{\gamma}}, \tag{2.32}$$

where $I \ni c$ is the set of dimensions c of the members of the index set $\mathscr{D}$. For instance for a Gaussian process indexed with both a time and space dimension s.t. $d = (d_x, d_t)^T \in \mathbb{R}^2$, it is natural that the time and space axes can have different covariance scale parameters ℓ_x and ℓ_t.

The $\gamma-$*exponential family* of covariance kernels (Rasmussen and Williams, 2006) with $\theta = (\gamma, \ell, \tau^2)$, is defined by the covariance function

$$k_{\exp}(d, d'; \theta, I) = \tau^2 \exp\left(-\xi^{\gamma}_{\ell_I}(d, d')\right), \tag{2.33}$$

which, with $\gamma = 2$, yields infinitely differentiable realizations of the random process that look very smooth.

The *Matérn family* of covariance kernels (Rasmussen and Williams, 2006), with $\theta = (\nu, \ell, \tau^2)$, is given by

$$k_{\mathrm{M}}(d, d'; \theta) = \frac{\tau^2 s^\nu}{\Gamma(\nu) 2^{\nu-1}} K_\nu(s), \tag{2.34}$$

where $s = 2\sqrt{\nu}\xi^1_{\ell_I}(d, d')$, $\nu = \alpha - \frac{q}{2}$, where α is a smoothness parameter and q is the dimensionality of s, and K_ν is the modified Bessel function of the second kind of order ν. The value $\alpha = \infty$ corresponds to the squared exponential kernel and $\alpha = 1$ corresponds to the exponential kernel with $\gamma = 1$. Despite this similarity between the Matérn and exponential kernels, the realizations of the random function from the processes with values $1 < \alpha < \infty$ do not correspond to those from the kernel k_{exp} with any values of γ. The smoothness parameter ν is not estimated in this work, but that can also be done, see e.g. (Roininen et al., 2018).

The Matérn kernel is expensive to evaluate for any ν that is not a half-integer, since K_ν is an infinite series that truncates only for the half-integer values. Figure 2.1 gives a visual example of how realizations from exponential and Matérn can look like.

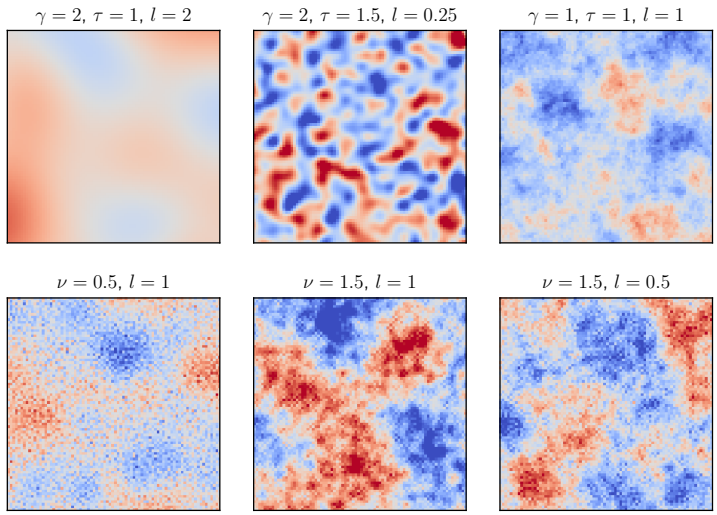


Figure 2.1: Example draws from Gaussian processes with exponential (first row) and Matérn (second row) covariance kernels show how the smoothness and scale change when the covariance kernel parameters θ are varied. These draws were generated by explicitly calculating the covariance matrix K by evaluating the covariance function in question between all pixel pairs, and drawing from that covariance by multiplying an i.i.d. standard normal vector by the Cholesky factor of K.

A *periodic kernel* with $\theta = (\tau^2, \ell_{\mathrm{per}}, \theta_{\mathrm{exp}})$ is defined in Paper I based on (Gelman

et al., 2013) by

$$k_{\text{per}}(d, d'; \theta, I) = \tau^2 \exp\left(-2\ell_{\text{per}}^{-2} \sin^2\left(\frac{\pi(t - t')}{\Delta_t}\right) - \xi^{\gamma}_{\ell_{I\setminus\{t\}}}(d, d')\right), \qquad (2.35)$$

in which the term θ_{exp} defines the scale parameters for the exponential functions ξ controlling the spatial component, and ℓ_{per} gives the periodic (e.g. inter-annual) covariance width for the temporal dimension. Normally the exponential spatial dependence, the last term in the exponent (2.35) is not present, but in the context of this work the periodicity is wanted to be restricted to the time dimension only.

The periodic kernel can be used to describe situations, where the dynamics of the data is expected to be periodic. For instance carbon dioxide fluxes do have an annual cycle due to the seasons repeating every year. If the mean function has periodic bias, this also can be caught by the periodic kernel. The periodic kernel is particular in that covariance over large temporal distances is possible, and, as in the context of Paper I, it can therefore be thought of having predictive capabilities even outside the temporal domain of the available observations.

A symmetric matrix $C \in \mathbb{R}^{n\times n}$ is *positive semi-definite* (PSD) if $\|x\|_C \geq 0$ for all $x \in \mathbb{R}^n$ (e.g. (Gruber, 2013)). In this work PSD matrices are always symmetric, even though sometimes the notion is taken to more generally refer to the situation where $\frac{1}{2}(C^T + C)$ is PSD. Since sums of PSD matrices are PSD, linear combinations of covariance functions are also valid covariance functions. This allows for lots of flexibility in describing combined effects of covariances of different scale, roughness, and amplitude.

A *multi-scale covariance kernel*, as defined in Paper I, captures covariances at various length scales. Given observation error variances of σ_x^2 for each observation at x, the multi-scale covariance function may then have for instance the form

$$k(x, x'; \theta) = \delta_{x,x'}\sigma_x^2 + k_{\text{per}}(x, x'; \theta, I_S) + k_{\text{M}}(x, x'; \theta) + k_{\text{exp}}(x, x'; \theta, I_{ST}). \qquad (2.36)$$

These kernel components are called *subkernels* in Paper I. What complexity and how many scale levels or kernel components are needed depend on the data. The identifiability of the parameters given data sampled using a multi-scale kernel is looked at in section 3.3.3.

The Gaussian process prediction problem can be solved locally using covariance tapering, as done in Paper I. Such *Vecchia approximations* (Vecchia, 1988) have been recently studied also by others, e.g. with a satellite remote sensing application to chlorophyll fluorescence data presented by Katzfuss et al. (2018). There are also various additional types of approximations to Gaussian processes to make them tractable with large data sets. A recent comparison of these methods is presented in Heaton et al. (2017).

2.5 Graphical models

A *graphical model* (Lauritzen, 1996; Wainwright and Jordan, 2008) is a model on a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ is the set of *vertices* or *nodes*, and $\mathcal{E}$ is the set of *edges*. The vertices correspond to random variables and the edges describe how those random variables depend on each other. Graphical models facilitate describing the conditional dependence structure, such as Markov structure, in Bayesian models. They are often used in situations where these dependency structures are complex and approximate inference algorithms are used to make the inference task tractable. The objective of the inference task is typically find out marginal distributions of the nodes, or the joint MAP estimate.

The graphical model framework can be also seen as an approach to looking at (typically high-dimensional) statistical inference problems. For different classes of graphical models there exist standard algorithms for performing inference (Wainwright and Jordan, 2008). This usually amounts to calculating expectations or point estimates or sampling from posterior distributions, conditionals, and marginals.

2.5.1 Directed graphical models

A *directed graphical model* (Wainwright and Jordan, 2008) or a *Bayesian network* describes the conditional dependence structure of a *directed acyclic graph* (DAG), where edges $e_{i \to j} \in \mathcal{E}$ have a direction and where there are no loops. In that case, the meaning of the DAG is, that the joint distribution of all the vertices factorizes according to

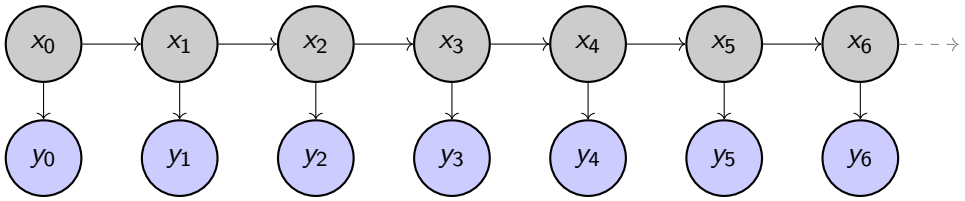
$$p(\{\nu \in \mathcal{V}\}) = \prod_{\nu \in \mathcal{V}} p(\nu | \nu_{\text{parents}}). \tag{2.37}$$

The *Kalman filter* (KF) (Kalman (1960), for a modern exposition see e.g. Särkkä (2013) or Law et al. (2015)) can be used as example of such a model, as is shown in figure 2.2, where the conditional dependence structure described by the arrows implies that the joint distribution $p(x_0 \dots x_N, y_0 \dots y_N)$ may be computed as in (2.37) as $p(y_0|x_0)p(x_0)\prod_{i=1}^{N} p(y_i|x_i)p(x_i|x_{i-1})$. This decomposition is a modeling choice, without which use of the KF algorithm would not be justified. The KF can also be described as an algorithm for solving the *hidden Markov model* (HMM) represented by the graph and its decomposition – hidden in that the states x_i are not directly observed, and Markov since $Y_i|X_0 \dots X_i = Y_i|X_i$. Alternatively, a *state space model* could be used, by specifying e.g. $x_{i+1} = f(x_i) + \xi_i$ and $y_i = x_i + \epsilon_i$ with $\xi_i \sim \mathcal{N}(0, \Gamma_i)$ and $\epsilon_i \sim \mathcal{N}(0, \Sigma_i)$ for some covariance matrices Σ_i and Γ_i. In the case of the standard KF, the function f would be linear. The state space approach is developed thoroughly for time series data by e.g. Durbin and Koopman (2012). Due to not using the KF in the Papers, the KF update formulas are not presented here.

In addition to the Kalman filter, for instance Markov chain Monte Carlo algorithms and hierarchical Bayesian models, both of which are described later, can be described as directed graphical models. Such models can be thought of as being *generative* in

that given parents ν_{parents}, realizations of ν can be generated directly. This paradigm is widely used in machine learning with for example generative adversarial networks and other models, see e.g. Goodfellow et al. (2014).

Figure 2.2: In the Kalman filter algorithm, the mean and the covariance of state x are updated at each time step i whenever observations $y_i \sim Y_i$ become available, as is represented by this DAG.



2.5.2 UNDIRECTED GRAPHICAL MODELS

An *undirected graphical model* or *Markov random field* (MRF) is a graphical model whose edges are not directed. These undirected edges determine if the global, local, or pairwise *Markov properties* hold (Lauritzen, 1996). These properties are equal if $p(\nu_{\mathcal{V}})$ is always strictly positive. For the algorithms in this work, the global Markov property is assumed, stating that any two different vertices ν_i and ν_j of a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, which are separated by a set of nodes $\mathcal{A}$ (in other words $e_{\nu_i\nu_j} \notin \mathcal{E}$) are conditionally independent given $\mathcal{A}$.

With the condition $p(\nu_{\mathcal{V}}) > 0$, the joint distribution of the graph can be written as a *maximal clique factorization*,

$$p(\{\nu \in \mathcal{V}\}) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \psi(\nu_c), \tag{2.38}$$

where Z is the normalizing *partition function*, and the set of maximal cliques $\mathcal{C}$, with $\cup\mathcal{C} = \mathcal{V}$, contains maximal sets of nodes c such that if ν_i, $\nu_j \in c$, then $e_{\nu_i\nu_j} \in \mathcal{E}$. The functions ψ are called *potential* or *compatibility functions*. For a lattice graph the maximal cliques are the adjacent pairs of random variables.

According to Hammersley and Clifford (1971), the maximal clique factorization and the conditional independence structure given by the graph are essentially identical. This suggests that efficient algorithms can be derived by working with maximal cliques of a graph.

The mean function modeling in Paper I is an example of an undirected graphical model, where the spatial dependence of the mean function parameters of a Gaussian process is modeled according to (3.3) and these β-parameters are allowed to change from one location to another based on spatially local observations.The important difference between the undirected and (acyclic) directed graphical models is that the

interdependence of the nodes is bi-directional ruling out straightforward sequential inference by just following the arrows of a DAG. Connections between Gaussian MRFs and Matérn class Gaussian processes is discussed e.g. by Lindgren et al. (2011).

An example of approximate inference in an undirected graph is given in figure 2.3, where marginalization in a lattice graph can be carried out effectively by diagonally calculating marginals corner-to-corner and back and carefully accounting for propagating beliefs (calculated marginals). In the variable elimination algorithm, the reconstituted graphs after elimination would have diagonal edges. In Paper I those are not considered for performance reasons, since the nodes diagonal to each other can then be computed in parallel due to absence of diagonal edges in the graph. Since solving the β coefficients in (2.31) involves inverting the covariance matrix K, and since this inversion is an $\mathcal{O}(n^3)$ process in the number of observations, computing the marginals in parallel separately is for large grids around 100 times faster even with a 12-threaded standard desktop workstation.

Not using the exact variable elimination algorithm does introduce an approximation error, but in the application of Paper I it is for several reasons typically either small or very small. First, with remote sensing data there are generally a large number of observations available for computing the β for each vertex ν. This means that when the spatial resolution of the grid is not excessively fine, the covariance with the observations selected for the other vertices of the reconstituted graph, referring to matrix K of the joint system in (2.31), would be much smaller. Second, when there are not that many observations available for fitting the parameters at each vertex the different vertices will share observations leading to similar β coefficients due to shared data. Third, at present only the modes are actually used, and therefore the joint and marginal variances of the β factors are not of paramount importance. Implementation of the exact variable elimination algorithm is planned to be added to the software tool presented in Paper I in the future.

2.6 MONTE CARLO ALGORITHMS

In this section X denotes a real-valued random variable or vector, i.e. $X : (\Omega, \mathscr{F}, \mu) \to \left(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \Lambda\right)$, and y denotes the observed data. Formulations with more general state spaces are common, but in this work $\mathbb{R}^d$ with Lebesgue measure Λ suffices.

Given any distribution $f_X(x)$ that samples need to be generated from, if its cumulative distribution function $F_X(x)$ is available, then independent samples can be trivially generated with the *inverse cumulative distribution function sampling* or *inversion sampling* (Tarantola, 2005): If $u \sim \mathcal{U}_{[0,1]}$, then obviously $F_X^{-1}(u) \sim f_X(x)$. A closed form of the cumulative distribution function is, however, rarely available.

In the Bayesian inverse problem setting Bayes' theorem (equation (2.6)) is used for finding the posterior distribution $p(x|y)$ of x by evaluating the likelihood function, $p(y|x)$. When $p(y|x)$ is computationally demanding to evaluate, usually also the evidence term $p(y) = \int_{\mathbb{R}^d} p(y|x)p(x)\mathrm{d}x$ is intractable, and for finding the pos-

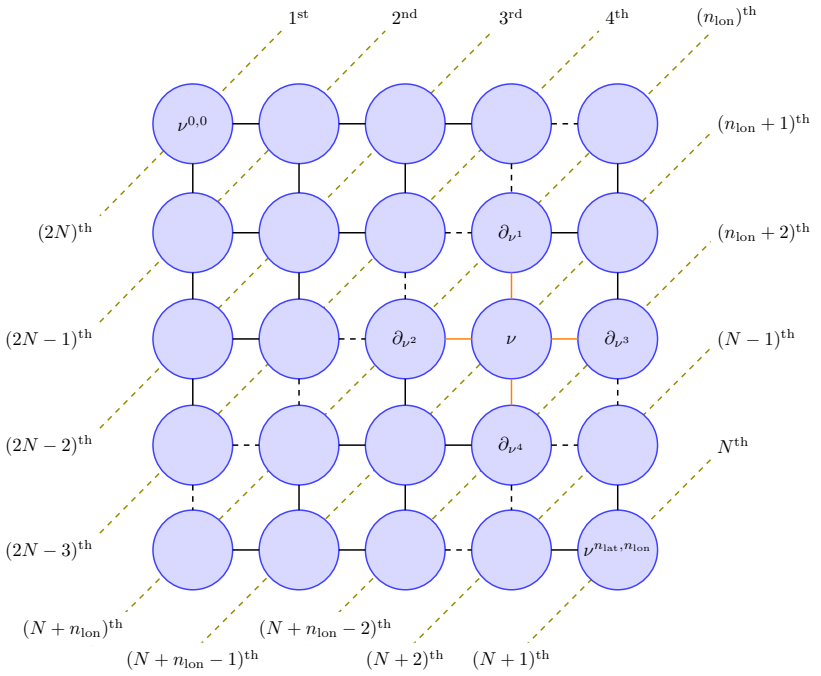


Figure 2.3: Figure from Paper I with ρ_i denoted by ∂_{ν^i}. The marginal distribution of vertex ν, $p(\nu)$, is conditional only on the neighbors $\partial_{\nu^1} \dots \partial_{\nu^4}$ (red edges) due to the Markov structure in the pictured lattice graph. This graph is used for solving for mean function coefficients in Paper I. Each connected pair is a maximal clique in this particular case. For effective solving, the vertices on the diagonal dashed lines are computed simultaneously. The order numbers labeling the diagonal lines represent an ordering in which the diagonals can be computed in parallel to get all the marginals in $\mathcal{O}(N)$ wall time, where $N = n_{\text{lat}} + n_{\text{lon}} - 1$. The $(N+1)^{\text{th}}$ computation in the corner is not conditioned on already-computed neighbors to avoid double counting data.

terior distribution clever algorithms are needed. *Monte Carlo algorithms*, nowadays discussed in a multitude of standard references such as Gamerman (1997); Gelman et al. (2013); Bickel and Doksum (2016); Tarantola (2005), are algorithms that utilize randomness for calculating expectations or drawing samples from a distribution. They are useful and necessary when a closed-form expression of the likelihood function is not available, which is always the case with complex geophysical process models.

One of the simplest such methods is the *rejection sampling* (Bickel and Doksum, 2016) algorithm: given an unknown unnormalized distribution $f(x)$, a constant M, and a known distribution $g_X(x)$, such that i.i.d. samples can be generated from $g_X(x)$, and given that for all $x \in \text{support}(f)$ it holds that $f(x) \leq Mg(x)$, samples x_i drawn from g are accepted as samples from f if $Y_i = 1$, where $Y_i \sim \text{Ber}\left(\frac{f(x_i)}{Mg(x_i)}\right)$. While this method works well for very low-dimensional targets if a good guess at M and g are available, the curse of dimensionality quickly destroys its performance. For this reason more sophisticated algorithms are utilized, which despite often producing correlated samples offer far superior performance. The Monte Carlo methods used in this work and described in this section are *Gibbs sampling*, *Adaptive Metropolis* MCMC, and *sampling-importance resampling* (SIR).

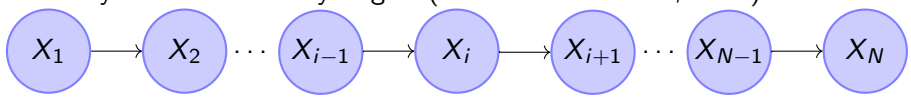
2.6.1 Markov chain Monte Carlo

A *Markov chain* is a sequence of random variables $X_1, X_2 \dots X_N$ such that for all i, the Markov property

$$p(x_i|x_1 \dots x_{i-1}) = p(x_i|x_{i-1}) \tag{2.39}$$

is satisfied (Gamerman, 1997). This extremely simple generative model is described in figure 2.4. As a conceptual bridge to section 2.4, Markov chains can be characterized as random functions with a discrete index set $\mathscr{D} = \mathbb{N}^+$ (Williams, 1991) and as with stochastic processes in section 2.4, it is useful to think about X_i as states of a dynamical system and of the indexes i as time. An obvious difference to the rejection sampling and inverse cdf sampling methods is that the samples generated by MCMC are not independent.

Figure 2.4: The random variables in a Markov chain depend only on the value of the preceding member. In MCMC algorithms this is exploited for efficiently generating correlated samples from a desired target distribution $p(x)$, since in theory $X_i \sim p(x)$ approximately for all sufficiently large i (Bickel and Doksum, 2016).



A Markov chain is *homogeneous* (Bickel and Doksum, 2015) if (2.39) is satisfied and $f_{X_i|X_{i-1}}(x_i|x_{i-1}) = f_{X_2|X_1}(x_2|x_1)\ \forall i \in \mathbb{N}$, $i \geq 2$, in other words the conditionals are not dependent on the index i. The evolution of a homogeneous chain is determined

by the *Markov transition kernel*,

$$q(x_1, x_2) = p(x_2|x_1), \tag{2.40}$$

which is a function giving the probability of transitioning from x_1 to x_2. If the state space is finite, then q is a matrix, whose elements q_{ij} give the transition probabilities from x_i to x_j. Given an MCMC chain with state space $\mathcal{X}$ and with transition kernel $q(\cdot, \cdot)$, the *stationary distribution* of the chain, π, if it exists and is unique, is given by

$$\pi(y) = \int_{\mathbb{R}^d} \pi(x)q(x, y)\mathrm{d}x \tag{2.41}$$

with some states $x, y \in \mathbb{R}^d$. If a chain can be constructed in such a way that it can be thought of as integrating over the state space as in the formula above, then the MCMC chain, after discarding some *burn-in* (or *warm-up* according to Gelman et al. (2013)) period to forget the starting point of the chain, can be seen as representing correlated draws from π. For if $X_i \sim \pi$, then $X_j \sim \pi$ for all $j > i$. Not any q will do, however, and the conditions for allowing this interpretation are clarified below. The following definitions are presented e.g. in Gamerman (1997) and Bickel and Doksum (2015).

A Markov chain is *stationary* if $\forall k \in \mathbb{N}$ and $\forall m \in \mathbb{N}^+$,

$$f_{X_1,\ldots,X_m}(x_1, \ldots, x_m) = f_{X_{k+1},\ldots,X_{k+m}}(x_{k+1}, \ldots, x_{k+m}). \tag{2.42}$$

A finite state Markov chain is *aperiodic* if it does not revisit the same state at fixed intervals, and for continuous state spaces such as $\mathbb{R}^n$, aperiodic chains do not visit any sets $F \subset \mathbb{R}^n$ s.t. $\Lambda(F) > 0$ at fixed intervals.

A finite state Markov chain is *positive recurrent*, if the expected visit time to any state is finite. For continuous states, the concept of *Harris recurrence* is used instead: a chain is Harris recurrent if the probability of revisiting any set $F \subseteq \mathbb{R}^q$ s.t. $\Lambda(F) > 0$ in a finite number of steps is one.

A finite state Markov chain is *irreducible* if any state x in the state space is reachable from any other state x' in a finite number of steps. For continuous state spaces, let ν be a measure on some state space $(S, \mathscr{S})$, typically $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. The chain is *$\nu$-irreducible* if for any $x \in S$ and any $F \subset S$ with $\nu(F) > 0$ there is an i_0 such that $\Pr(X_{i+i_0} \in F|X_i = x) > 0$. The chain is irreducible if any such ν exists.

An irreducible and aperiodic finite-state Markov chain is called *ergodic*. In the continuous case a Markov chain is ergodic if it is irreducible and Harris-recurrent. Furthermore, if the chain satisfies *detailed balance*,

$$\pi(x)q(x, y) = \pi(y)q(y, x), \tag{2.43}$$

then the chain is said to be *reversible*. This can be stated in that the *probability flow* from x to y is the same as from y to x. For an ergodic MCMC chain, (2.43) can be trivially manipulated with

$$\int_{\mathbb{R}^d} \pi(x)q(x, y)\mathrm{d}x = \int_{\mathbb{R}^d} \pi(y)q(y, x)\mathrm{d}x = \pi(y). \tag{2.44}$$

Equation (2.44) is actually just (2.41), meaning that transitioning from x, which is a random draw from the stationary distribution π, with the kernel q, yields another draw from π.

This development leads to the conclusion that devising transition kernels which generate ergodic reversible chains is desirable, since such chains ultimately automatically produce samples from the target distribution. The effectiveness, however, depends on the *mixing time* – how fast the random state variables of a Markov chain initialized at random end up being distributed according to π – and how correlated the samples are.

The *Metropolis-Hastings (MH) algorithm* is by far the most famous MCMC algorithm, and it is satisfies the detailed balance condition (Gamerman, 1997). Let $t(y;x)$ be the *proposal density*, which is a probability density on $\mathbb{R}^d$ evaluated at y.

The MH transition kernel $q(x,y)$ is defined with the help of the *acceptance probability*

$$\alpha(x,y) \triangleq \left(1 \wedge \frac{\pi(y)t(x;y)}{\pi(x)t(y;x)}\right), \tag{2.45}$$

using which it can be written in the form

$$q(x,y) = t(y;x)\alpha(x,y) \qquad\qquad x \neq y. \tag{2.46}$$

From this it follows that the probability the chain stays put is

$$\Pr(X_{i+1} = X_i) = 1 - \int\limits_{z\in\mathbb{R}^d\setminus\{x_i\}} q(x_i,z)\mathrm{d}z, \tag{2.47}$$

as demonstrated by Gamerman (1997) in the discrete state space case. The function t is most often parametrized by the location x of the chain at the previous iteration. A notable exception to this rule are independent proposals, which do not depend on the current chain location x.

While ergodicity depends on the proposal distribution, detailed balance for the MH algorithm follows from a direct calculation: for $\pi(y)t(x;y) > \pi(x)t(y;x)$ and $\pi(x)t(y;x) > \pi(y)t(x;y)$, respectively,

$$\begin{aligned}
\pi(x)q(x,y) &= \pi(x)t(y;x)\alpha(x,y) = \pi(x)t(y;x) = \alpha(y,x)t(x;y)\pi(y) = \pi(y)q(y,x),\\
\pi(y)q(y,x) &= \pi(y)t(x;y)\alpha(y,x) = \pi(y)t(x;y) = \alpha(x,y)t(y;x)\pi(x) = \pi(x)q(x,y),
\end{aligned} \tag{2.48}$$

and the case $\pi(y)t(x;y) = \pi(x)t(y;x)$ is trivial. When the proposal is symmetric, $\frac{t(x;y)}{t(y;x)} = 1$ and the MH-algorithm reduces to the *Metropolis algorithm* (Tarantola, 2005).

The power of the MH algorithm (and MCMC algorithms in general) is in how the density π in (2.46) is evaluated. When sampling a posterior distribution as given by Bayes' theorem (2.6), the observed data is fixed and whether a proposed point is

accepted or not depends only on the outcome of a Bernoulli trial whose parameter is the ratio of the posterior density evaluated at the proposed and current points. This makes evaluating the evidence term unnecessary.

Since the samples generated with MCMC are correlated, calculating the *effective sample size* (ESS) is useful. The (1-d) ESS is defined by (Gamerman, 1997)

$$\eta = N \left(1 + 2 \sum_{i=1}^{\infty} \rho_i(x_n)_{n=1}^{\infty} \right)^{-1}, \tag{2.49}$$

where N is the length of the Markov chain $(x_n)_{n=1}^{N}$, $\rho_i \triangleq \mathrm{Corr}(x_1 \dots x_{N-i}, x_{i+1} \dots x_N)$ is its lag-i *autocorrelation coefficient*, and the series is in practice truncated due to finite chain length and due to that after the initial decay in autocorrelation the terms tend to only contribute noise. There are many options for computing essential sample sizes for multivariate chains, but a canonical version of the ESS does not exist. One common way is to compute the ESS for each coordinate projection separately.

Draws generated with MCMC algorithms should be seen as draws from the posterior only after the chain has mixed well (Gelman et al., 2013), since only after some i_0 it is true that $X_{i_0+i} \sim \pi$ for all $i \in \mathbb{N}$. A practical way to find such an i_0 is to run multiple chains initialized at random points and to include as posterior samples from each chain the tails s.t. the inter-chain statistics agree with the within-chain statistics. If a single chain is used for e.g. computational reasons, whether the chain finds the target distribution or not can be usually also seen by looking at the chain for each state variable separately. In the rare case when π is multi-modal, comparing 2-d pairwise marginals with varying degrees of burn-in may be more revealing. While MCMC is used in Papers I, II, and IV, only the experiments with real-world data in Paper I exhibited multi-modal features (not shown).

As for the other Monte Carlo estimates, central limit theorems apply, implying that the variance of the estimator for the mean of a scalar target density π behaves according to $|\mathbb{E}[\pi] - \frac{1}{N-i_0} \sum_{i_0}^{N} x_i| \sim \mathcal{N}(0, \frac{\sigma^2}{\eta})$, where η is given by (2.49) and iterations before i_0 have been discarded as burn-in.

The applications presented in Papers I, II and IV use a variation of the MH algorithm, the *Adaptive Metropolis* (AM) algorithm (Haario et al., 2001), which produces non-homogeneous chains and therefore is not reversible. The chains are, however, ergodic and converge to the target distribution when the chain length tends to infinity. The sampling procedure with AM is identical to that of standard MH, except for that the covariance of the Gaussian proposal density is recalculated every once in a while[1] to match the sample covariance, scaled by the factor $\frac{2.4^2}{d}$, where d is the parameter dimension. This choice yields an optimal acceptance ratio for Gaussian targets (Roberts et al., 1997).

[1] How often the adaptation is done is implementation-dependent. It is known, however, that adapting at every iteration may lead to the algorithm misbehaving. As an example, in Paper II the adaptation was done whenever the iteration number was the square of an integer.

If the target is non-Gaussian, AM will still work, but not quite as effectively. In order to decrease the correlatedness of the samples, the *Delayed Rejection* (DR) algorithm (Tierney and Mira, 1999) may be implemented on top of AM, resulting in the *Delayed rejection adaptive Metropolis* (DRAM) algorithm (Haario et al., 2006). When a proposed state is rejected for the first time, the DR algorithm, instead of immediately repeating the previous value in the chain, proposes other points from scaled proposal densities. These later proposals are accepted with a modified acceptance probability that takes care of that the chain remains reversible. Practical experience, for instance from preliminary simulations for Paper II, showed that while DRAM works it will in many cases not improve nor deteriorate the performance of the sample generation. With multi-modal targets, the performance can, however, be dramatically improved with DR (see comment SC1 by Laine, Susiluoto, Tamminen, and Haario in the discussion of Lu et al. (2017)).

There are many alternative MCMC algorithms and new ones are, such as (Titsias and Papaspiliopoulos, 2016; Bouchard-Côté et al., 2015), are continuously being developed. Many of these more modern algorithms as well as the older Metropolis-adjusted Langevin (Grenander and Miller, 1994) and Hamiltonian Monte Carlo (Duane et al., 1987), utilize gradient information of the posterior density to improve the quality of the samples. Without gradients or good guesses at the covariances in the posterior, the AM method in practice performs well, as is shown in the Papers.

2.6.2 Gibbs sampling and Metropolis within Gibbs

Gibbs sampling is a Markov chain Monte Carlo method, where a multivariate target π of state $X = (X^1, \ldots, X^d)^T$ with random variable elements X^i is sequentially sampled component by component with the Markov transition kernel

$$q(x, y) = \prod_{i=1}^{d} p(y^i | z^{-i}), \tag{2.50}$$

where $z^{-i} = (y_1, \ldots, y_{i-1}, x_{i+1}, \ldots, x_d)$ (Gamerman, 1997). The resulting chain is homogeneous, and Gibbs sampling has been shown to have the full joint posterior distribution as the stationary distribution (Tierney, 1994). Gelman et al. (2013) presents Gibbs sampling as a special case of the MH algorithm.

Given the form of the proposal, it is natural to use Gibbs sampling in situations where analytic forms of the conditionals are available. This situation arises when a hierarchical statistical model is constructed utilizing conjugate priors, as outlined in section 2.7. When some of the conditionals are not available in closed form, other forms of sampling may be employed, such as rejection sampling or the Metropolis algorithm. In the latter case, the algorithm is then called Metropolis within Gibbs (Gamerman, 1997). It is used in Paper II.

Generally the number of samples needed in Monte Carlo sampling scales very poorly with parameter dimension even with best methods. Limiting the dimension

of the Metropolis-sampled part may help somewhat due to that in Gibbs sampling proposed parameters are always accepted (Gamerman, 1997). In the presence of parameter correlations, Monte Carlo algorithms proposing the correlated parameters together are generally superior. If an approximation of the joint posterior density is available, this can be used to rotate the parameter axes for achieving better mixing. Often-cited methods improving sampling efficiency based on this idea, applicable in generic Monte Carlo sampling settings, include for instance Active subspaces (Constantine et al., 2015), Likelihood-informed subspaces (Cui et al., 2014), and truncating the standard singular value decomposition (Mueller and Siltanen, 2012; Gruber, 2013).

2.6.3 IMPORTANCE SAMPLING AND RESAMPLING

Let random variable $X \sim \pi$ take values $x \in \mathcal{X}$, and let $\pi_b(x)$ be a distribution from where samples can be generated, called the *biasing distribution*, with the corresponding measure denoted by μ_b. *Importance sampling* is a method for estimating expectations of a function $g(x)$ by evaluating it at samples drawn from π_b and re-weighting those samples according to the ratio of g and π_b. More formally,

$$\mathbb{E}[g(X)] = \int_{\mathcal{X}} \frac{g(x)\pi(x)}{\pi_b(x)} \mathrm{d}\mu_b(x), \tag{2.51}$$

which with a finite sample of size N becomes

$$\mathbb{E}[g(X)] \approx \hat{g} \triangleq \sum_{x \sim \pi_b} \frac{g(x)}{\pi_b(x)/\pi(x)}, \tag{2.52}$$

where $\hat{g}$ is called the *importance sampling estimate of* $\mathbb{E}[g(X)]$. Gelman et al. (2013) treat $\pi(x)$ as an unscaled posterior density, but while this may be a useful depiction, conditioning on data is not generally necessary for describing importance sampling.

Importance sampling is particularly useful for rare event simulation with computationally demanding models; in case only tails of a parameter distribution trigger a rare event such as a catastrophic drought, flood, nuclear reactor meltdown or a flu pandemic, compared to naive Monte Carlo sampling the accuracy of the calculated expectation can be increased dramatically by using a biasing distribution with most of the mass in this rare event triggering region. The condition $\mathrm{support}(g) \subseteq \mathrm{support}(\pi_b)$ needs to be satisfied for importance sampling to work – otherwise it could happen that no samples are used from an area in $\mathcal{X}$ where g is large, and this would introduce bias to $\hat{g}$. The optimal choice for the biasing distribution that minimizes the variance of $\hat{g}$ is $\pi_b(x) \propto |g(x)|\pi(x)$ (Casella and Berger, 2002).

If the samples for computing the sum over the biasing distribution in (2.52) are taken from a previous Monte Carlo sample, the procedure of computing $\hat{g}$ can also be used to generate samples from g by re-weighting those Monte Carlo samples with $\frac{\pi(x)g(x)}{\pi_b(x)}$ and drawing independently according to the obtained weights. With $g(x) \equiv 1$

this is called *Sampling-Importance-Resampling* (SIR) (Gelman et al., 2013). The method can be useful e.g. if after conducting a Monte Carlo experiment there is need for an adjustment of the likelihood function, or if adequate data exists for repurposing output of model evaluations for calculation of additional statistics. The SIR method is utilized in Paper II.

2.7 Hierarchical Bayesian models

A *hierarchical Bayesian model* (Gamerman, 1997; Gelman et al., 2013) is a modeling approach for situations where parameters of a distribution need to be modeled as random variables. A typical example and the one utilized in this work involves creating a model for an ensemble of related experiments indexed with $i \in \{1, \dots n\}$. These experiments are conducted in such a way that the dependency structure of the data on any associated random variables is shared but observations y_i differ for each ensemble member.

In the hierarchical model described in figure 2.5 the parameters θ_i, possibly associated with each ensemble member, share a common prior distribution $p(\theta_i|\nu)$ parameterized with parameters ν, but the parameters θ_i can have different posterior marginal distributions $p(\theta_i|y)$ depending on the data y. The ν are called *hyperparameters* and their priors are called *hyperpriors* (Gelman et al., 2013). It is possible to have the data also depend on auxiliary parameters τ that are prescribed a fixed prior. In such a setting the full joint posterior distribution can be written as

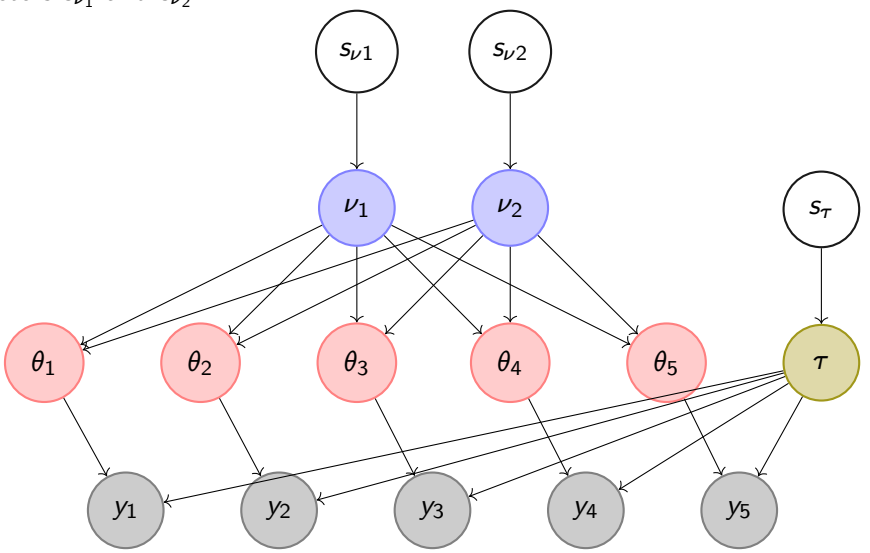
$$p(\theta, \tau, \nu|y) \propto p(\nu)p(\tau) \prod p(y_i|\theta_i, \tau)p(\theta_i|\nu). \tag{2.53}$$

Figure 2.5 expresses the model described above as a directed graphical model, as is customary for hierarchical models (Wainwright and Jordan, 2008). The graphical description intuitively reveals the conditional independence structure in (2.53).

More complex hierarchical models describing multiple levels of shared dependence structure may also be constructed. For instance the (hyper)parameters ν could again depend on other random variables ζ with assigned hyperprior distributions $p(\zeta)$, instead of just depending on the fixed parameters s.

Sampling from posterior distributions of these models often is facilitated by using *conjugate priors*, meaning that the families of the prior distributions are chosen to be such that the conditional distributions have closed forms and are easy to sample from (Gelman et al., 2013; Gamerman, 1997). This is for instance the case in the example of (2.53) if $\tau \sim \mathcal{N}(\mu_\tau, \Sigma_\tau)$ with some vector τ and covariance Σ_τ, $\nu = (\mu_\theta, \sigma_\theta^2)$, $\theta_i \sim \mathcal{N}(\mu_\theta, \sigma_\theta^2)$, $\mu_\theta \sim \mathcal{N}(\mu_0, \sigma_0^2)$, and $\frac{1}{\sigma_\mu^2} \sim \text{Scale-inv-}\chi^2(s)$ with some hyperprior parameters μ_0, σ_0^2, and s. If the θ_i parameters are not mutually correlated, and especially if furthermore the τ parameters are not correlated with the θ_i, then the θ_i parameters may be effectively Gibbs-sampled and the dimension of Metropolis-sampling the τ parameters remains smaller as mentioned earlier in section 2.6.2.

Figure 2.5: An example of the simple hierarchical model in (2.53) described as a DAG. Observations y_i are generated by parameters θ_i, all of whose priors depend on hyperparameters ν_1 and ν_2, and parameters τ which have a fixed prior with parameters s_τ. The priors of the hyperparameters ν are the hyperprior distributions with fixed parameters s_{ν_1} and s_{ν_2}.



This is called *Metropolis within Gibbs* or *Gibbs within Metropolis*, depending on the source.

2.8 Bayesian modeling with time series data

In geosciences, trace gas flux measurements are often done at flux measurement sites with fixed instruments producing time series data. Typical time series measurement data y_t with time index $t \in \{1 \dots T\}$ is evenly distributed in time, even though more often than not there are gaps in the data due to various reasons such as instrument malfunction, power outages, or weather. In Paper II time series flux measurement data is used for Bayesian model calibration. As with other Bayesian modeling, also in that setting the posterior shape depends strongly on how the residual autocorrelations in (2.10) are modeled.

The model M, designed to produce states x related to observations y as in (2.8) and (2.9) has unknown biases and random errors which may be time-dependent, as is the case in e.g. Paper II. The probability model for the model-observation mismatch needs to account for any such structure generated by any error source, see also section 2.1.2.

2.8.1 AR, MA, and ARMA models

Let the model-data residuals be denoted by r_t with $t \in \{1 \dots T\}$. A possible model for the residual autocorrelations is the *autoregressive model* (Harvey, 1990; Durbin and Koopman, 2012; Chatfield, 1989) of order n, AR(n), which models the time series r with

$$r_i = \sum_{j=1}^{n} \phi_j r_{i-j} + \epsilon_i, \tag{2.54}$$

where ϕ_j is the *lag-j autocorrelation coefficient* and ϵ_i are some random variables, which are assumed to be independent. However, if after fitting the ϕ_j parameters with a reasonable choice of order n the ϵ_i still *de facto* end up being autocorrelated, other models may be tried. A second much used model for time series data is the *moving average model* of order m, MA(m) (*ibid.*), given by

$$r_i = \sum_{k=1}^{m} \xi_k \epsilon_{i-k} + \epsilon_i, \tag{2.55}$$

where the difference to the AR models is that while AR models add random error on top of a weighted sum of the previous data values, the MA model adds the random error on top of a weighted sum of previous random errors. These models can be combined to form an *autoregressive moving average model of order* (m, n), denoted ARMA(m, n) (*ibid.*), with

$$r_i = \sum_{j=1}^{n} \phi_j r_{i-j} + \sum_{k=1}^{m} \xi_k \epsilon_{i-k} + \epsilon_i. \tag{2.56}$$

There are several variations to these models such as introducing nonlinearities or exogenous inputs (Durbin and Koopman, 2012). However, since model complexity is a liability when interpreting the results, these were not used in Paper II.

2.8.2 Practical parameter estimation in the ARMA setting

This section follows the presentation in Paper II, where the r_i in (2.56) are generated by a non-trivial dynamical model via (2.9). To perform Bayesian inference with Monte Carlo methods, the parameters ξ and ϕ add another layer of difficulty since the likelihood given any model parameters θ, $p(y|\theta)$, depends on both the ARMA parameters ϕ and ξ in addition to θ. While the fully Bayesian way of doing this would be finding the full joint posterior distribution $p(\theta, \phi, \xi|y)$ via evaluating $p(y|\theta, \phi, \xi)p(\theta, \phi, \xi)$, where the prior parameters would usually be independent, this optimization problem is not generally convex and both minimization algorithms and MCMC algorithm may get stuck in local minima and/or drift to nonphysical parameter regions. It may also happen that the model parameters are not constrained by the data under the statistical model used.

An alternative to finding the full joint parameter posterior distribution is to find a point estimate $\hat{\theta}$ of the model parameters by minimizing some statistic of the data (residuals), e.g. sum of absolute values, running mean, or sum of squares of the residuals or their subset, and then find point estimates for the error model parameters ξ and ϕ. Monte Carlo simulations to find the posterior distribution of the model parameters can then be performed given these estimates.

Given $\hat{\theta}$, finding the order (m, n) of the model and point estimates of the parameters ϕ and ξ can be done by minimizing the *Bayesian information criterion* (BIC) (Bickel and Doksum, 2016) – a standard method for model selection – giving

$$(\hat{\phi}, \hat{\xi}, m, n) = \underset{(\phi,\xi,m,n)}{\arg\min}\, \mathrm{BIC} = \underset{(\phi,\xi,m,n)}{\arg\min} \left\{ n_{\mathrm{par}} \log(n_{\mathrm{obs}}) - 2 \log\left(p(r|\phi, \xi)\right) \right\}. \quad (2.57)$$

In the above expression, n_{obs} is the number of observations, which in the absence of gaps in data equals T, and $n_{\mathrm{par}} = m + n$ is the number of parameters. Other popular criteria for model selection, such as the Akaike information criterion (AIC) (Bickel and Doksum, 2016), which uses the penalty $2n_{\mathrm{par}}$ instead of $n_{\mathrm{par}} \log(n_{\mathrm{obs}})$, often produce similar (but not identical) results. For finding the ARMA(2,1) parameters used in Paper II, residuals r were simulated by random sampling fifty parameter vectors θ from an approximate posterior, and the vast majority of those residual time series resulted in optimal ϕ and ξ parameters close to each other.

The resulting time series of error terms ϵ in (2.56) can be checked to not to be autocorrelated by calculating the *Durbin-Watson statistic* (Durbin and Watson, 1950, 1951),

$$T(\epsilon) = \frac{\sum_{i=2}^{T} (\epsilon_i - \epsilon_{i-1})^2}{\epsilon^T \epsilon}, \quad (2.58)$$

where a gapless observation series has been assumed, but gaps can be taken care of if needed by discarding any indexes with no observations. If $T(\epsilon)$ is close to 2, the time series has no substantial lag-1 autocorrelation (see p. 26).

Model residuals as in (2.10) are usually expected to be zero-mean, since any constant term could be simply added to the definition of model M in (2.8). To make sure the error model is correct and the obtained posterior shape is accurate, the appropriate scale parameters for the distribution of the ϵ_i in (2.56) need to be found. If the magnitude of the error terms varies in time as is often the case in geophysical applications, the error is called *heteroscedastic* (Harvey, 1990). To utilize such time series for the likelihood formulation, an easy way to proceed is to preprocess the series using a parametric model γ with parameters α, resulting in a new homoscedastic time series $\epsilon^* = \gamma(\epsilon; \alpha)$.

These α can for uncorrelated zero-mean residuals be found by minimizing some distributional distance such as the *Kullback-Leibler divergence* (KL-divergence) (discussed e.g. in Peyré and Cuturi (2018)) between an empirical distribution (histogram) $\eta_{\gamma(\epsilon;\alpha)}(x)$ of the ϵ^*-terms, and the actual error model $\nu(x)$. The KL-divergence for

two continuous real-valued distributions p and q is given by

$$D_{\mathrm{KL}}\left(p(x) \| q(x)\right) = \int_{\mathbb{R}^d} p(x) \log\left(\frac{p(x)}{q(x)}\right) \mathrm{d}x \tag{2.59}$$

and the appropriate parameters for the error model can be expressed with it as

$$\hat{\alpha} = \arg\min_{\alpha} D_{\mathrm{KL}}\left(\eta_{\gamma(\epsilon;\alpha)}(x) \| \nu(x)\right). \tag{2.60}$$

The process described in this section outlines one simple approach to do *covariance estimation* in a time series context, which is an important and often also a difficult part of Bayesian parameter estimation studies. This difficulty reflects the complexities of the error structures arising from combining real-world data and complicated computer models. The outline of the parameter estimation procedure presented here can be compared with the parameter estimation procedure of Gaussian process covariance kernels in (2.29) and mean functions parameter fields in section 2.5.2, which however omit the model selection and heteroscedasticity considerations. For the GP work in Paper I, these considerations still remain to be fully addressed.

3 Applications to geosciences

3.1 Overview of scientific contributions

The *large-scale geoscientific problems* in the title of this thesis refers to the topics presented in Papers I-IV and this section provides motivational context for that work while the details are discussed later. More emphasis is given to Papers I and II than to Papers III and IV.

3.1.1 Spatio-temporal high resolution CO2 distributions

The research carried out in **Paper I** tries to answer the following question: *Where is the carbon dioxide in the atmosphere?* This question is important in its own right since the general public has shown much interest in it, but the answer could be applied to atmospheric flux inversion, or statistical emission models could be developed based on the results. Since Gaussian processes provide uncertainty information via the theory in section 2.4, the results can also be applied to validation schemes and hypothesis testing.

Current scientific literature of Gaussian processes or kriging applied to atmospheric remote sensing of global CO2 does not contain any high resolution studies that the authors of Paper I would be aware of. The high number of observations leads to computational compromises which often result in overly smoothed posterior fields. However, the multi-scale approach presented in Paper I is able to produce *arbitrarily high resolution CO2 maps* with both fine and coarse scale features.

The spatial statistics software presented is not constrained to CO2 or the OCO-2, but *can be used with any remote sensing data* in principle. The software is able to *learn kernel and mean function parameters*, and is able to *sample from extremely high-dimensionally discretized posterior or prior distributions* as defined by the multi-scale kernel description.

3.1.2 Uncertainties in Boreal wetland CH4 emission processes

Out of all methane emissions, those from natural wetlands have the highest uncertainty. While this in itself is more than enough reason to study uncertainties in process-based wetland emission models, there is another important reason as well:

changing climate increases uncertainty regarding future emissions. **Paper II** studies a Finnish boreal wetland site with a model that was developed in tandem with writing Paper II (Raivonen et al., 2017). The central questions that we try to answer are *how much uncertainty is there in the model parameters controlling the physical processes*, and *do the model parameters and hence the wetland's behavior react to environmental changes*.

Many wetland methane emission models have been written, but their systematic calibration has in general not been a research priority in the community. More specifically, at the time of writing we were unaware of any Bayesian calibration studies of wetland emission models. For this reason it is valuable that the work answers questions such as, *given flux measurements, model, and input data, how are the model parameters correlated in the posterior distribution*, and *how much interchangeability is there between the methane production and transportation processes*.

One difficulty that this study does not yet tackle arises from that the many different types of wetlands all over the boreal region all behave differently. Understanding the functioning of these different environments would require a calibration process for all these types and to accomplish that a spatial statistics or regression/classification study of boreal wetland distributions would be needed. Despite this opportunity for future research, Paper II can be seen as *groundwork for future larger-scale studies* of uncertainties related to boreal wetland CH4 emissions.

3.1.3 Effects of climate change on growing season and gross primary production

Carbon emissions to the atmosphere are the main driving force of climate change and while the overall mechanisms have been known for a long time, how climate changes is actually a complex process. The emissions are balanced partly by the uptake of carbon from the atmosphere by plants, and the magnitude of this uptake is controlled by many factors. In the boreal region, the date of snow clearance regulates when the growing season starting date (GSSD). **Paper III** answers the questions *how many days earlier does the growing season start than in the 1970s*, and *how much additional carbon is getting photosynthesized in this process?*

To answer these questions, Paper III utilizes a wide latitude of flux measurements from all over the boreal region, and compares that with global climate model simulations forced with data from the European Center for Medium-Range Weather Forecasts (ECMWF).

While boreal ecosystems have been studied widely, the connection between snow clearance date and gross primary production (GPP) has not been studied previously in this fashion. This study is a pure simulation study in the sense that, unlike in the other Papers, uncertainties are not quantified (except for providing the p-values of regression estimates).

3.1.4 Monte Carlo estimates of land surface scheme hydrology and gas exchange parameters

In land surface schemes (LSS) of climate models the model hydrology description often poses difficulties since changes in hydrological conditions may produce nonlinear effects on other modeled variables such as GPP. In models the hydrology-related subroutines are intimately connected to the carbon exchange via stomata, since stomata control both CO2 and water transport in the gaseous phase. Since models utilize discrete plant functional types[1] for land cover description, the parameter values controlling model behavior for each type are generic, average best guesses. Therefore, in case of a rare event such as a major drought, the model may perform worse than it could with calibrated parameter values. Furthermore, the generic parameter values are generally not the best ones available for a particular measurement site.

Against this background, **Paper IV** looks at *how the hydrology-related parameters of the land surface scheme correlate in the posterior distribution*, and asks *whether the model is able to capture a rare event (drought) with calibrated parameters*. In addition, the MCMC calibration is done with different temporal poolings of the data, allowing to look at *how the uncertainty estimates and model performance change depending on the data averaging performed*.

3.1.5 Other related work

Two additional publications co-authored by J.S., Raivonen et al. (2017) and Mäkelä et al. (2019), are intimately connected to Papers II and IV, respectively. Even though they are not a part of this thesis work, they are briefly mentioned here to give context to Papers II and IV.

In Raivonen et al. (2017), the HIMMELI wetland methane emission model and the physical processes are described in detail, and this article provides motivation behind the modeling choices and more clarity regarding the underlying biology than Paper II does. In its approach, it is purely a model development manuscript and does not explicitly employ the techniques described in Chapter 2.

As a continuation of Paper IV, Mäkelä et al. (2019) evaluates how different stomatal conductance formulations in the JSBACH LSS are or are not able to explain measured fluxes under different environmental conditions. It looks at a wider variety of flux measurement sites (10 as opposed to two in Paper IV), uses an adaptive population importance sampler (APIS) for carrying out the Monte Carlo sampling, and has, due to the different conductance models, a model selection flavor. The more comprehensive and methodical approach than the one taken in Paper IV brings the findings closer to upstream integration to improve the performance of JSBACH in the boreal region.

[1] Plant functional types are collections of parameters with which the model distinguishes ecosystem types from each other. These parameters contain values for e.g. maximum leaf area index, biomass, nitrogen deposition rate, etc.

3.2 Models, observations, and algorithms control computational cost

3.2.1 Parallel models and parallel algorithms

Geoscience is a versatile field of applied science that often serves as a testing ground for novel computational methods. Despite this versatility, a large variety of these problems, especially when it comes to uncertainty quantification, are computationally constrained, as is easily seen from the descriptions of the Monte Carlo algorithms in section 2.6. While parallel resources for computation are nowadays readily available, creating efficiently scalable code is a challenge, often due to data and memory bandwidth limitations, but also due to the sequential nature of many sampling techniques.

Figure 3.1 describes the computational cost of the problems tackled in this thesis and helps to explain why the inference algorithms and modeling paradigms were chosen in the very way they were. In **Paper I**, (light blue arrow, bottom right in figure 3.1) the Gaussian process software is able to compute marginals globally in a half-degree grid for every day for four and a half years with OCO-2 data (with reasonable settings) in ten months' time on one CPU core. The inbuilt OpenMP parallelization brings this down to a few days on a modern supercomputer node, but since utilizing several nodes would require architectural changes, the maximum size of the problems is currently limited by available single-node resources. The current implementation requires keeping all observations in memory, and therefore problems with the largest numbers of observations can not be computed on a modern laptop. On a supercomputer node, computing with billions of observations is possible. This also applies to generating gridded draws from the GP.

In **Paper II**, (red and orange arrows in the middle), the forward model – a wetland methane emission model – runs parallelized (downward component of arrows) to yield a speed-up in computation. The experiment was designed so that the forward model simulations of the different years for any given parameter in the MCMC chain were run on different cores simultaneously, meaning that the temporal domain was split into several parts. This guided the inference algorithm choice towards the Metropolis within Gibbs MCMC paradigm, which is well suited for hierarchical modeling, see sections 2.7 and 2.6.1. In the first preliminary experiments (orange arrow), available in the discussion paper of Paper II, several experiments were performed with different model discretizations. While this aspect was dropped from the final version, the parallelization scheme employed decreased the amount of simultaneous model evaluations by the number of different discretizations (leftward component). Together these choices took the core hour requirement down from several years to less than a month. In the final simulations a single MCMC chain was computed and therefore no algorithmic parallelization was possible. However, the speed-up from the time domain decomposition remained.

Paper IV employed several parallel MCMC chains to generate posterior estimates

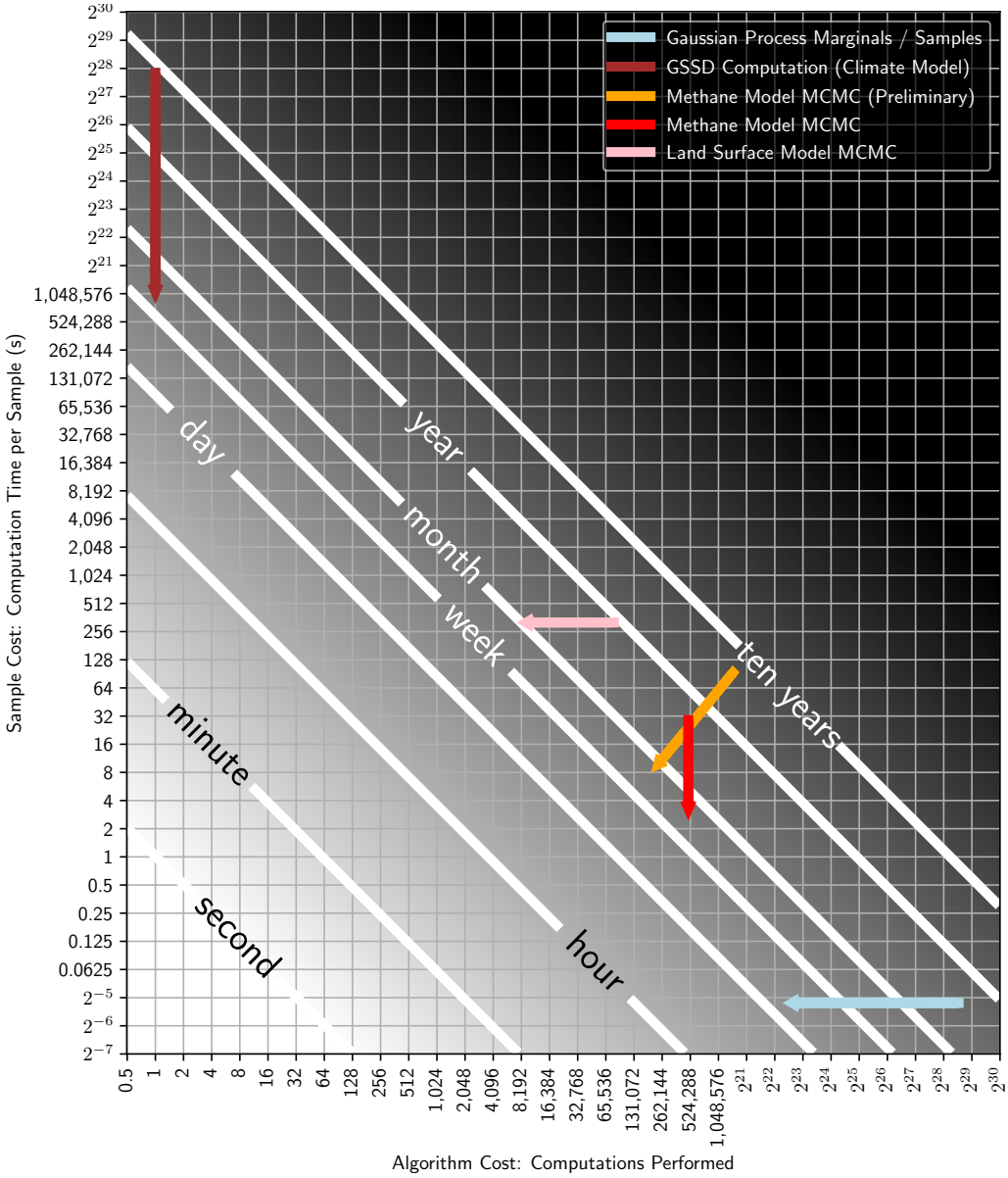


Figure 3.1: This figure shows how the computational cost is divided between sampling schemes and sample generation. The x-axis shows the number of samples, and the y-axis shows the cost per sample. The diagonal white lines show the contours for constant 1-core computation time. Arrows start at the total computational cost of a problem, and end at the computational cost that takes to account both model and algorithm parallelization. All the experiments conducted are constrained by available CPU time and the logarithmic scales along both axes provide perspective to how expensive the most demanding simulations in Papers I-IV were. The black area in the very upper right is unfeasible without massive parallelization or with dedicated accelerators. Increasing Gaussian process model or climate model resolution would easily extend the light blue and the brown arrows to that area. The initialism GSSD in the legend stands for growing season starting date.

of a set of land surface model parameters (pink arrow in the middle). The uncoupled simulations were run on a fast laptop with four hyperthreaded cores. More effective parallelization possibilities could have been utilized on a supercomputer, but that was avoided here to facilitate code development and avoid code porting.

The very opposite to Paper I in terms of parallelization is **Paper III** (brown arrow in the top left corner), where no algorithm was used – just a single climate model simulation with reanalyzed ECMWF forcing data and with the objective of producing data for a regression analysis to back up and quantify other scientific reasoning based on multiple sources of *in situ* measurement data. The simulation was carried out on ten supercomputer nodes (parallelized using the Message Passing Interface library) and in several segments due to model instability. Performing the final simulation required generating initial carbon and water pools, which doubled the amount of computation.

The common denominator of these computational challenges is that only algorithms which *in practice* yield results in a month's walltime are feasible, and the experiments were designed accordingly. While the arrows in figure 3.1 are not normalized in that they represent true computation times in different computing environments, they still reveal where the practical computational constraints are in the research reported in this thesis.

Even though a month is a reasonable amount of time to be spent on computer simulations, for all the above experiments that time is only the tip of an iceberg. Different model configurations needed to be tested and tried before the final product-yielding simulations could be performed, and many of the computational problems in the Papers contained smaller but still important computational sub-problems, such as calculating the MLE of model parameters, creating initial conditions, etc. Those are not pictured in figure 3.1.

3.2.2 The role of the observation data

Observations are used in the Papers for four primary purposes; (1) forward model forcing, (2) forward model calibration, (3) error model calibration, and (4) forward model validation. The term forward model is reserved here for dynamical models – the statistical models describing the model-observation mismatch are called error models as was discussed in section 2.1.2.

The number of observations and the way they are utilized in the Papers varies wildly and therefore a summary of observation usage is given in Table 3.1. In the table an observation vector may contain several variables, which are detailed in the Papers themselves.

Paper I uses OCO-2 satellite observations, of which there are 249 million for the time period considered, and even after selecting data according to quality-flagging, 116 million observations remain. There is no dynamical forward model, only a statistical model describing the data. Since the model parameters are fitted to all the data, separate validation is not needed. The calculated set of marginals can be effectively

Table 3.1: Number or observation vectors used for each Paper, and the ways in which observations are utilized. Here *forcing* refers to model forcing (setting model states based on data), *learn M* refers to learning parameters θ controlling the model M, as in (2.8), *learn* ϵ refers to learning error model parameters as in (2.9), and *validate* refers to whether observations are used in a direct validation scheme, such as cross validation.

	#obs vectors	Forcing	Learn M	Learn ϵ	Validate	Comments
I	~120,000,000	N/A	N/A	yes	no	Details in section 3.3.1
II	~2,500	yes	yes	yes	yes	
III	~100,000	yes	N/A	N/A	no	Full reanalysis fields
IV	~100,000	yes	yes	no	yes	Data is aggregated

regarded as a statistic $T(y)$ that adequately (to the modeler) summarizes the huge number of observations. The approximate GP algorithm, presented in section 3.3.1, describes how the large number of inputs is handled.

Papers II and IV utilize time series flux observations for constraining the models, and while Paper II also does cross validation for the hierarchical model, in Paper IV a straightforward simple validation is performed on an alternate site. The difference between number of observations is explained by that Paper IV uses half-hourly data, whereas Paper II uses daily means, since the model used in II does not realistically describe the diurnal cycle and therefore using the half-hourly observations would amount to fitting noise. Both of these Papers utilize measurement data to force the forward model, but the error model calibration in Paper IV is not rigorous, while Paper II actually uses draws from an approximate posterior predictive distribution to calibrate the error model parameters before the final Bayesian model calibration is performed.

Paper III does not contain a calibration step, and therefore parameter finding is not applicable. The 100,000 forcing fields are full T63-resolution reanalysis fields from ECMWF – either ERA Interim or ERA-40, depending on the year. Paper III utilizes also flux measurement data from 10 sites in Finland, Sweden, Russia, and Canada, but these are not directly tied to the modeling – only via aggregate statistics in Table 1 of Paper III – and therefore they are not reported here.

With a large number of data, a common complication with Bayesian model calibration in geosciences is that the posterior density may in practice contract towards a point estimate that is sometimes not realistic. This behavior is aggravated by any (often unavoidable) model misspecification, but it also takes place without it in the small observation error limit of overdetermined Bayesian inverse problems, as described e.g. by Stuart (2010). With MCMC the practical implication is that the observation error variance in the observation model may need to be inflated to allow posterior exploration. As a result, the size of the posterior is in the end not necessarily reliable. Paper IV solves this problem by building the statistical model

for data averages instead of individual points, and Paper II utilizes an exploratory MCMC based on which an error model used by the SIR algorithm is calibrated. Despite the calibration and due to model misspecification, the choices the modeler has to make are apparent in how the posterior looks like. Parameter correlations in the posterior are more resistant to log-posterior scaling than e.g. marginal variances, since $\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathbb{V}(X)\mathbb{V}(Y)}}$. For this reason physics-based interpretation and analysis of results of the Bayesian model calibration is justified to be based on analyzing the posterior correlation structures. Among the Papers, this is most emphasized in Paper II.

3.3 Efficient multi-scale Gaussian processes for massive remote sensing data

This and the following sections present the research of each individual Paper in more detail than was done in sections 3.1 and 3.2. For even further details, please consult the Papers themselves.

The first Paper of the thesis, Susiluoto et al. (2019), deals with a computational spatial statistics approach to regularize a sparse set of satellite observations into a spatio-temporal grid with arbitrary resolution. The method used is Gaussian process regression, and both marginals and samples from both the prior and the posterior are obtained. The space-dependent mean function of the Gaussian process is learned utilizing an approximate elimination algorithm on a regular lattice graph to learn the modes of the marginal distributions over a Markov Random Field.

The Gaussian process theory is described in section 2.4, the MRF and the elimination algorithm are described in section 2.5.2, the objectives and highlights of this part were briefly stated in section 3.1.1, and an outline of computational cost and size of the problem was given in section 3.2. While these will be slightly expanded here, the main focus is on additional key details, computation, and discussion.

Several kriging/GP studies such as Zeng et al. (2013, 2017); Nguyen et al. (2014); Hammerling et al. (2012b,a); Tadić et al. (2017); Zammit-Mangion et al. (2015), and Zammit-Mangion et al. (2018) have been conducted with remote sensing CO2 data over the years. The majority of those have used data from the GOSAT satellite, while a handful of exploratory publications related to the OCO-2 satellite have been published. For details, see the introduction section in Paper I.

Compared to other CO2 measuring instruments the sun-synchronous OCO-2 satellite is particularly interesting, since it provides high resolution column-integrated dry air CO2 mole fraction (XCO2) measurements. It does so by applying an algorithm to retrieved absorption spectra of reflected sunlight. The footprint of a single measurement is only 1.29 by 2.25 kilometers in size, with eight measurements abreast. Clouds and aerosols often result in quality-flagged and missing measurements. The approximate revisiting time to any particular location is 16 days, but obviously not

all area between two trajectories is covered during one 16-day period, and the closer to the equator the satellite is, the larger the uncharted area.

Despite the high spatial resolution of the satellite measurements, there are at the moment, as far as we know, no published CO2 maps based on only data and showing any of that finer structure. The central problem is computation: in order to calculate the Gaussian process posterior, the covariance matrix of the observations needs to be inverted. This is lots of work with hundreds of millions of observations. How the calculations are performed algorithmically is described next.

3.3.1 Gaussian process model algorithm description

The random field Ψ, in Paper I the spatio-temporal XCO2-field, was defined in section 2.4 to be a Gaussian process, denoted $\Psi \sim \mathrm{GP}(m(x), k(x, x'))$, if the joint distribution of the process at any finite set of points was multivariate normal. The function m had a parametric form given below in (3.3) and exponential, Matérn, and periodic covariance kernels were supported by the software. An additional non-stationary kernel, the wind-informed kernel, is proposed and discussed below in section 3.3.6.

The Gaussian process model computation in practice comes down to computing conditional expectations and variances of the multivariate Gaussian distribution given in (2.19) and (2.20). These distributions are enormous - in the largest simulation in Paper I the dimension n equals 116489343 and storing or solving this size of a linear system, which is an $\mathcal{O}(n^3)$ operation, is not directly possible. For this reason an efficient algorithm and its implementation are needed. The satGP program, consisting of roughly 4000 lines of highly optimized C code and presented in Paper I, is able to approximately compute (level of approximation is controllable with input parameters) the desired spatio-temporal grid of marginals in

$$\text{cost} = \mathcal{O}\left(\frac{A n_{\text{times}}}{\omega^2}\left[(n_{\text{ker}}\kappa)^3 + \sum_{l=1}^{n_{\text{ker}}}(r_l \log(r_l) + \kappa \log \kappa)\right]\right) \tag{3.1}$$

time. In this equation, A is the grid area, n_{times} is the number of time steps, ω the grid resolution, n_{ker} the number of subkernels as in (2.36), κ the maximum subkernel size, and $r_l \propto \prod_{i=1}^{q} \ell_i^l$ is a factor determining the size of the hyper-ellipse outside which covariance with the test input is less than the prescribed covariance threshold σ^2_{min}. The values of r_l also depend on the maximum covariance parameters τ^2. This scaling is linear in number of marginals, and the parts in the brackets — first term for inverting the constructed full multi-scale covariance and second for finding observations that are informative for each test input where the marginal is computed — is highly optimized. For additional details regarding observation selection and multiple other computational aspects, please see Paper I and the satGP source code.

The downside of obtaining the linear scaling with the number of test inputs of course is that the full posterior covariance will not be retrieved, only the marginal variances. The posterior covariances can still in principle be calculated from posterior

sample covariances. Another natural possibility is constructing a multi-grid or multi-fidelity Gaussian process (Peherstorfer et al., 2018; Kennedy and O'Hagan, 2000), and this extension would not be impossible to implement in satGP.

The satGP program can draw from the random process by conditioning on previous predictions. Computing the Gaussian field roughly amounts to interpolation by solving a linear system of equations locally at the test input using the observations that are within a desired radius. If the ordering for generating the field is chosen so that instead of interpolation, extrapolation is performed (for instance if in the 1-d case the sampled points would reside at $x_1 = 0$, $x_2 = 0.1$, $x_3 = 0.2\ldots$), such ordering may in practice lead to oscillations in the generated data. For this reason a sparse ordering is used, both in space and time: if the number of inputs where the field is generated is $n_{\text{tot}} = n_{\text{lat}} n_{\text{lon}} n_{\text{times}}$, then the m^{th} computed point is number $(mp \bmod (n_{\text{tot}}))$ in the linear ordering along axes (time, latitude, longitude), with the last of these changing fastest. In the above, p is taken to be the largest prime number under $0.9 n_{\text{tot}}$.

The satGP software also contains routines to learn the maximum marginal likelihood estimates (marginalized over the Gaussian process realizations) of the covariance function parameters θ using an approximate random-sampling based method

$$\hat{\theta}_{\text{MLE}} = \arg\min_{\theta} \sum_{x_i \in E_{\text{ref}}} \left\{ \|\psi_i^{\text{obs}}\|_{\tilde{K}_i} + \log|\tilde{K}_i| \right\}, \tag{3.2}$$

where E_{ref} is a set of randomly sampled points from the specified spatio-temporal domain. The vector $\psi_i^{\text{obs}} \in \mathbb{R}^{d_i}$ contains at most $n_{\text{ker}}\kappa$ observations closest in covariance to x_i from which the mean function value at x_i has been subtracted, and $\tilde{K}_i$ is the corresponding covariance matrix determined by the covariance function with parameters θ and the observations ψ_i^{obs}. Due to randomly selecting E_{ref}, this procedure results in the log-likelihood including an unknown multiplicative coefficient and hence an unknown multiplier of covariance in the exponent. Therefore, while posterior mean estimates (for unimodal symmetric), posterior medians, and MAP estimates remain valid, the true size of e.g. credible regions is not known.

The most important input parameters needed by satGP together with the algorithm description illustrate how the software works, and they are shown in table 1 and figure 4 in Paper I.

3.3.2 OBTAINING THE GP MEAN FUNCTION FROM A GAUSSIAN MRF

For describing the XCO2 field observed by the OCO-2, the mean function (2.30) is assigned the explicit form

$$m(x, t; \beta, \delta) = f(t, \delta)^T \beta = \beta_0 \sin\left(\frac{2\pi t}{\Delta_t} + \delta\right) + \beta_1 \cos\left(\frac{4\pi t}{\Delta_t} + \delta\right) + \beta_2 + Ct, \tag{3.3}$$

where Δ_t is the length of the period, that is, one year, and where the spatial dependence denoted by the argument x comes from the selection of observations for fitting

the coefficients of the mean function. The particular form of (3.3) was chosen for its ability to represent the increase in the CO2 concentration as a global trend, and also because with this form it is possible to describe the seasons both in the tropics and closer to the poles. The resulting mean function coefficients are shown in figure 3.2.

For spatial smoothness, a Gaussian MRF utilizing the setting presented in section 2.5.2 is used. Since in addition to the β-parameters also the δ-parameter varies from place to place, (2.31) cannot be used directly due to the δ-parameter not conforming to its form. Instead, a first pass calibration is performed utilizing the BFGS gradient-based optimization algorithm to find the mode of all parameters for each vertex, by minimizing

$$l_\nu(\beta, \delta) = \frac{1}{n_{\text{obs}}} \sum_{i=1}^{n_{\text{obs}}} (m(x, t; \beta, \delta) - y_i)^2 + \sum_{\nu' \in \partial\nu} \psi(\nu, \nu'), \tag{3.4}$$

where the latter sum is over the edge potentials corresponding to Gaussian priors defined by the modes of the neighbors. The scaling is arbitrary since the objective is to merely fit the δ-parameter to produce fields that look smooth to enable computing $p(\beta|\delta, y)$.

In propagating the posterior marginals (beliefs) when computing the β-factors, the precision of the neighboring points is scaled according to the distance to those points on the latitude-longitude grid, since close to the poles the grid points are closer to each other than on the equator. For fitting the parameters with (2.31) at each grid point, observations that are nearest in spatial covariance (disregarding the time component) are chosen, and the marginals are computed conditioning on the optimized δ. The uncertainties of the β-factors are given by (2.31), but they are also approximated by the BFGS-algorithm, which therefore in principle could also be used. However, in Paper I the exact computation via (2.31) was utilized.

While only a minor part of Paper I, Fig. 3.2 shows several intriguing features. The constant term β_2 has high values where emission hotspots are known to be. The parameter controlling slow oscillations, β_0, shows the reversed seasons between the northern and southern hemispheres, and β_1 shows a semiannual signal of higher amplitude in the Congo area. The phase shift parameter δ appears noisy in areas where the β_0 and β_1 parameters are close to zero, which is exactly when the δ-parameter plays very little role. To conclude it is worth remembering that since the parameters act together to describe the CO2 field, drawing far-reaching conclusions from individual maps should be avoided.

3.3.3 Identifiability of multi-scale parameters

The justification for using the multi-scale covariance kernel formulation, (2.36), is not obvious — it could be that the parameters of the multi-scale model would not be identifiable in practice. In Paper I, multi-scale kernel parameters are recovered from synthetic data generated by drawing a sample from the GP prior. While this

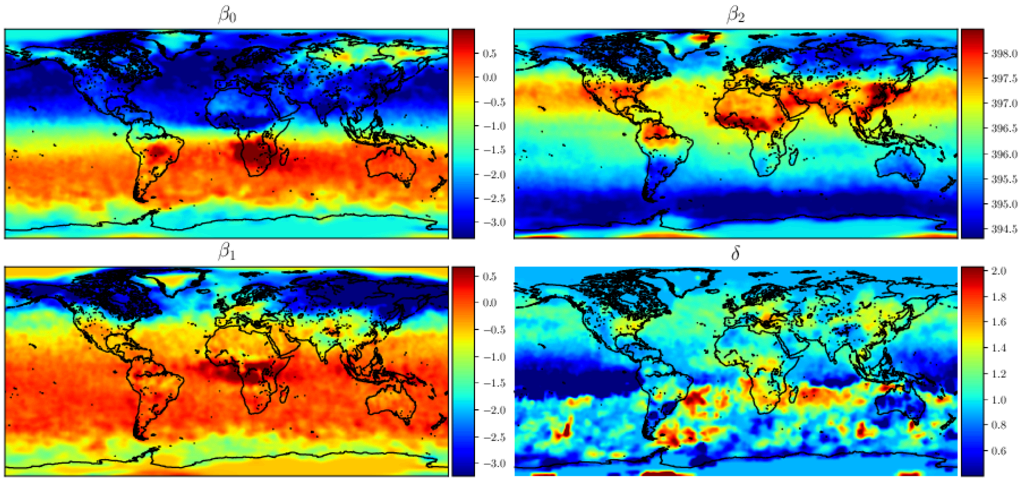


Figure 3.2: Local β factors, and the δ phase-shift of the mean function as in (3.3). The trend component, C, has been fitted to global values, and does not vary spatially.

parameter inversion fails with optimization algorithms, MCMC can be successfully used as a stochastic optimizer. The posterior mean value is a good estimate for parameters as shown in figure 3.3 in a synthetic study with two subkernels. Notice that while the true values are not in the very centers, the scales of the axes reveal that the true values are within a small distance from the center in the parameter space.

This synthetic study validates the multi-scale approach in that since the parameters of the different subkernels are recoverable, the different kernels may indeed be needed for describing the field. In Paper I, a three-component kernel is shown, and while there the length-scale parameters of the smaller-size kernels are slightly overestimated, the ability to approximately find the true parameter values remains.

Table 3.2: Covariance function parameter values learned from OCO-2 data. First column shows the Matérn subkernel parameters, and the second column the parameters of the exponential subkernel.

	$(\cdot) = \text{mat}$	$(\cdot) = \text{exp}$
$\tau^{(\cdot)}$	0.899	2.72
$\ell_{lat}^{(\cdot)}$	0.00513	0.0418
$\ell_{lon}^{(\cdot)}$	0.0363	0.397
$\ell_{t}^{(\cdot)}$	20h 22min	16d 20h 12min

After validating the parameter estimation process for the multi-scale kernel, the

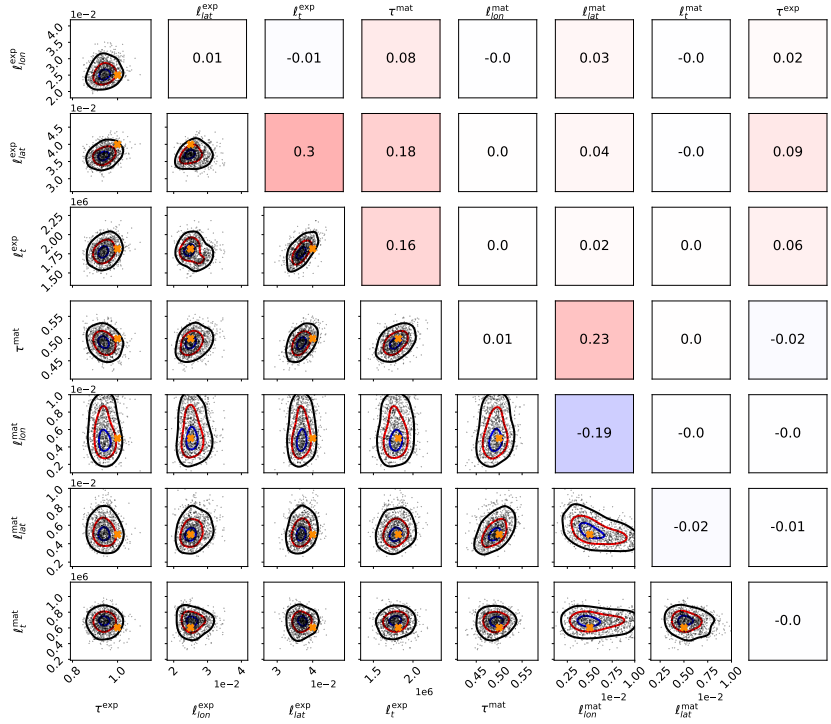


Figure 3.3: Approximate posterior with unknown scaling of the log-posterior from a synthetic study with two subkernels. The first subkernel is of Matérn type and the second an exponential one with smaller and larger length scale parameters respectively. The data was sampled using a random spatial pattern from the prior and 1% noise was added, after which the parameters were learned.

parameters corresponding on the OCO-2 data were learned. No data thinning was applied, and the number of reference points in E_{ref} was set to be 12 with $\kappa = 256$. The resulting parameter values are shown in Table 3.2. The notable aspect of the parameter values is the elongation of covariance ellipses of both kernels in the more informative zonal direction.

3.3.4 Learning multi-scale kernel parameters from OCO-2 data

The multi-scale kernel allows larger scale features to be combined with local enhancements. In figure 3.4 a covariance kernel consisting of a single subkernel alone with large length scale parameters was compared with that same subkernel combined with a subkernel with shorter length-scale parameters. Observations from the OCO-2 v9 data product were used.

The parameters of the kernels are given by Table 3.2. The total kernel size was kept at 1024 ($\kappa = 512$ for (a-b) and $\kappa = 1024$ for (c-d)) in both experiments. Random data thinning with $\zeta_{\text{train}} = 5$ was applied: the parameter determines how

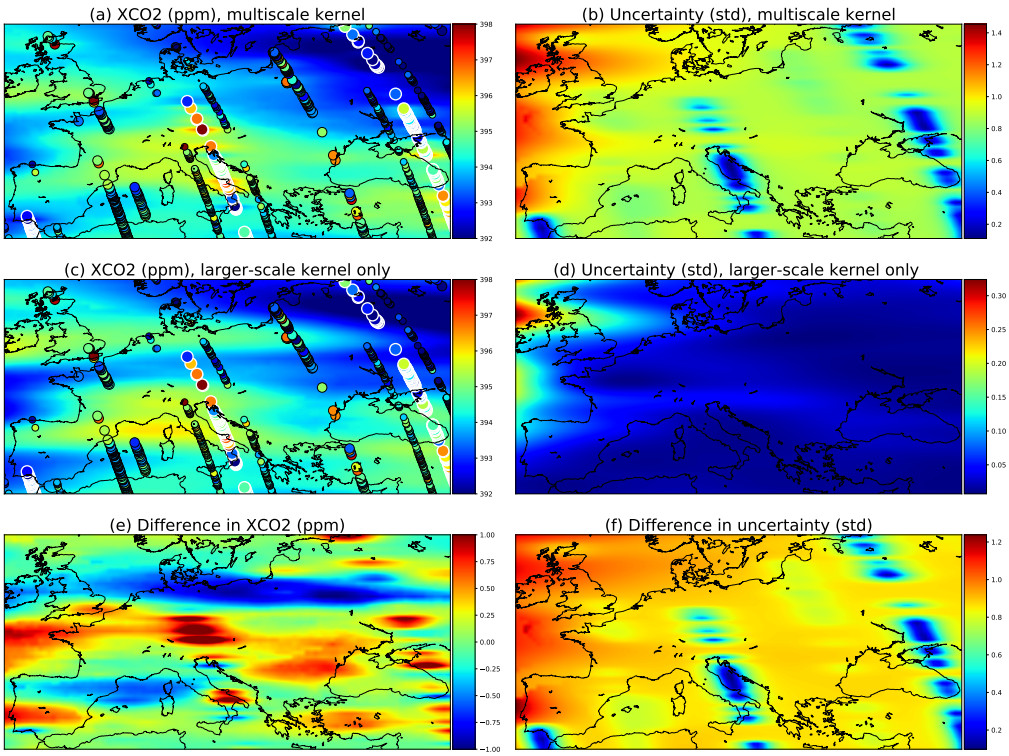


Figure 3.4: Comparison of a multi-scale kernel with two components with the parameters shown in Table 3.2, and a kernel containing only the exponential subkernel in Table 3.2. The observations used are shown in panels (a) and (c) as circles. The large ones with white borders are observations from the present day, September 15$^{\text{th}}$ 2014, medium circles are observations from 14$^{\text{th}}$ and 16$^{\text{th}}$, and small circles from 13$^{\text{th}}$ and 17$^{\text{th}}$.

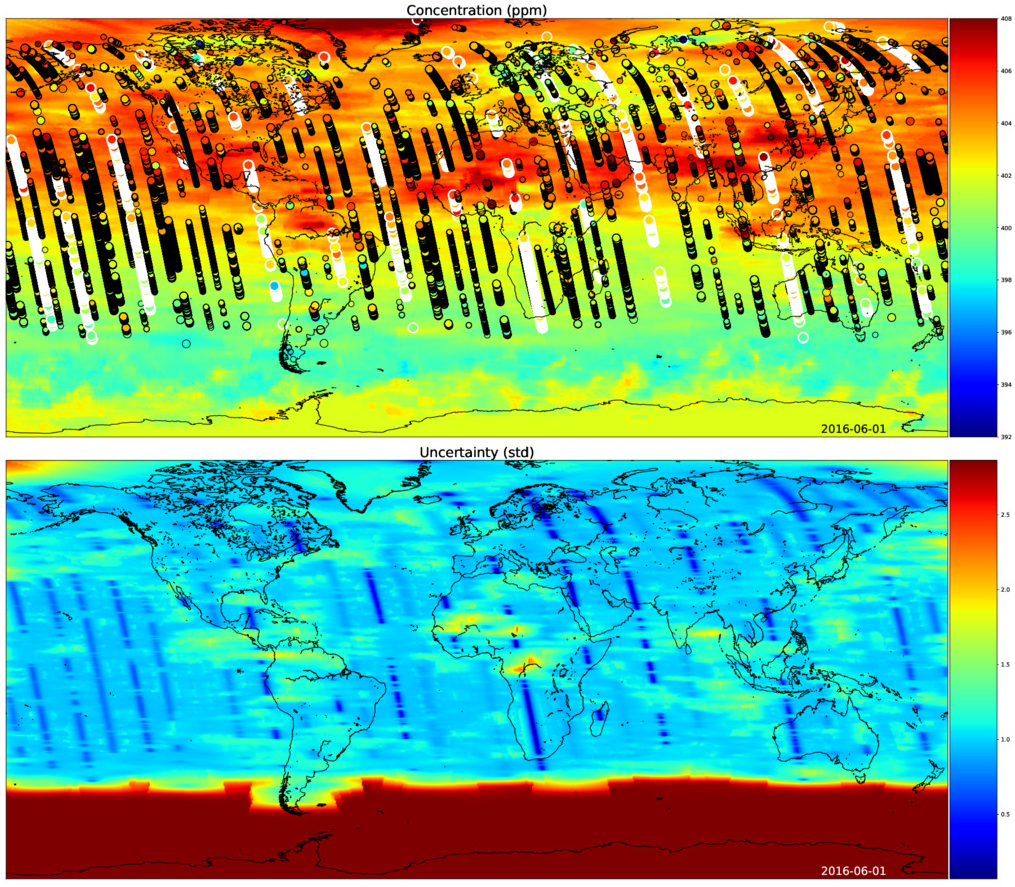
likely including the next observation is, and this probability depends on the distance to the previously added observation with $\Pr(\text{add } y|y_{\text{prev}}) = \frac{|x-x_{\text{prev}}||_2}{\omega\zeta_{\text{train}}}$, where ω is the grid resolution and x and x_{prev} denote, as earlier, the locations where observations y and y_{prev} were made. Such thinning discourages observations very close to each other from being included; for further details, see Paper I. Earlier, Tadić et al. (2017) has also used a distance-based probabilistic approach for observation selection, even though the inclusion probability is different. In both of the experiments $\omega = \frac{1}{2}$ was used and the exact same set of observations was utilized for calculating the marginals.

The figure clearly shows where present day observations are found as local enhancements. With the single subkernel with the larger length-scale parameters, the uncertainties are unreasonably low.

3.3.5 Posterior XCO2 fields

A central reason for creating the Gaussian process software for remote sensing data is to be able to get better estimates of the spatio-temporal distributions of the quantity of interest with uncertainties. Figure 3.5 shows the means and uncertainties of the Gaussian process posterior calculated via (2.19) and (2.20) in a grid. The slight edginess far from observations, especially visible where the uncertain portion starts on the bottom of the lower part, is due to capping the search radius at 1100 km (10 equatorial degrees) in order to facilitate computation. In total 351 million marginals were computed with $\kappa = 256$ and using no data thinning, with parameter values from Table 3.2. The total number of observations used was 116 million.

Figure 3.5: Global GP posterior marginals with uncertainties on first of June 2016. In the summer months, the coverage of the satellite does not reach the South Pole due to lack of sunlight. The circles with the white edges are the current-day observations, the medium circles are observations from one day away, and the smallest circles are observations from two days away. Notice how the uncertainty increases from day to day due to the smaller kernel reducing local uncertainty less and less.



3.3.6 Wind-informed kernel

One of the novel ideas in Paper I is the *wind-informed covariance kernel*, which rotates the covariance ellipse according to the wind axes. Given zonal and meridional wind vectors u and v, it is defined by parameters $\theta = (\tau, \ell, \rho, w^*)$. The kernel itself is an exponential kernel whose length-scale component to the direction of the wind, $\ell^{\parallel}$, is scaled by $\sqrt{1 + |w^*|\rho}$, where w^* is the wind velocity vector at the test input x^* (listed above as a parameter since it does not depend on individual inputs x and x'). The additional parameter ρ determines how large a role the wind speed should play. The length scale parameter perpendicular to wind, $\ell^{\perp}$, is not scaled, i.e. $\ell^{\perp} \leftarrow \ell$. Figure 3.6 shows equicovariance contours for various combinations of ρ and w^*.

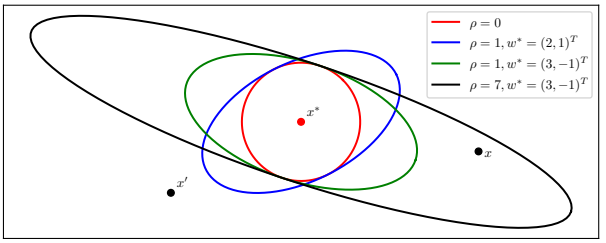


Figure 3.6: Equicovariance ellipses from the wind-informed kernel with various wind vectors w^* and values of ρ. The wind velocities are taken at the test input x^* but the covariance function k is of course evaluated also for each pair of observations x and x'.

The rationale behind the formulation of the wind-informed kernel is that e.g. trace gases are spread by winds and therefore the covariance direction should change according to wind direction. This subkernel type may also be combined with others in a multi-scale kernel.

The wind kernel parameters were calibrated by finding the medians of the approximate posterior calculated with the approximate marginal maximum likelihood method given by (3.2). The parameters found were $\tau = 2.07$, $\ell = 0.038$, and $\rho = 56.7$, and the values of $\kappa = 1024$ and grid resolution of 0.7° were used with the thinning parameter $\zeta_{\text{train}} = 1$ introducing some thinning. An example of the results is shown in figure 3.7, and as expected the uncertainty is clearly reduced where wind is blowing directly towards or away from the observations. The predicted mean of the concentration field is also spread due to the winds. The posterior marginal mean field looks less monotonous than the fields from fixed-direction kernels.

The wind-informed covariance kernel could be formulated in various ways and which formulation works best with what data still needs to be studied further. The winds used in figure 3.7 were processed from the local winds that come with the OCO-2 data. Obviously, winds derived from an actual wind data product would provide better accuracy, especially when the test inputs x^* are far from any observations.

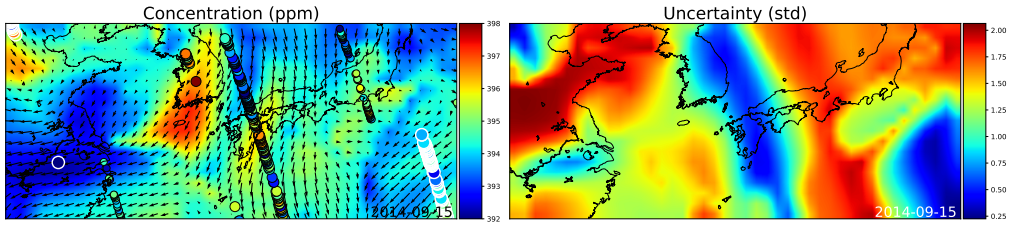


Figure 3.7: Japan, Koreas, China: GP posterior marginal mean field of XCO2 and the corresponding uncertainties produced with the wind-informed kernel. As before, circles with the white edges are present-day observations, medium ones are from adjacent days, and the smallest ones are from two days away. Wind direction and magnitude are given by the arrows.

3.4 Bayesian inference of physics of a Boreal wetland with hierarchical MCMC

Boreal wetlands and peatlands are a major source of CH4 emissions to the atmosphere, and it is likely that the magnitude of these emissions will grow as climate change progresses. In addition to CH4, wetlands – in particular drained and managed wetlands – release and/or have the potential to release substantial amounts of CO2. How substantial these emissions are and will be is not fully known, since peatland carbon emission estimates currently have high uncertainties (or uncertainties are not reported) and Bayesian analysis in the field of wetland emission modeling remains rare.

The research in Paper II and its objectives and results were briefly introduced in section 1 and 3.1.2, related work was mentioned in section 3.1.5, and the computational cost was discussed in section 3.2. The published literature pertaining to Bayesian modeling or model calibration in the context of wetland CH4 emission models is covered in the introduction section of Paper II. This section describes the computational problem referencing section 2 and discusses some of the main results.

3.4.1 The HIMMELI forward model

The wetland methane emission model HIMMELI[2], developed in collaboration between University of Helsinki and Finnish Meteorological Institute (Raivonen et al., 2017), is a 1-d partial differential equation model discretized by soil layers of variable thickness. In addition to CH4, explicit formulations of CO2 and O2 are also included. The model contains processes for CH4 production from root exudate decomposition and anaerobic peat decay. The transportation of the gases to the surface takes place in three ways: diffusion, transport via stems of aerenchymatous plants, and transportation

[2]HIMMELI stands for *HelsinkI Model of MEthane buiLd-up and emIssion for peatlands*.

due to bubble formation, called *ebullition*.

Methane is produced predominantly when oxygen is not available and this is in HIMMELI controlled by the water table depth (WTD). The exudate input is provided as pre-calculated net primary production (NPP), fraction of which is passed on to the roots. The root depth distribution determines at which depth the exudates are deposited. If the water table level is above that deposition depth, methane may be produced.

The model version in Paper II, called sqHIMMELI, contains also the processes dealing with root exudates and peat decay, whereas in Raivonen et al. (2017) those processes are described as external functions for generating input. The 21 equations defining much of the sqHIMMELI model and the role of the model parameters are described in sections 3.4 and 3.5 of Paper II.

In addition to NPP and WTD, the model takes in soil temperature profiles and leaf area index (LAI) data, which broadly speaking tells how many layers of leaves in the canopy intercept solar radiation. The simulations and the study were performed utilizing measurement time series of the inputs and CO2 and CH4 fluxes from a research station in Hyytiälä, Southern Finland. Data from years 2005-2014 was used. For some input variables, filling gaps or other additional modeling were needed, see section 2 in Paper II.

3.4.2 Bayesian Inference

The posterior distribution of the parameters controlling most parts of the model physics was computed with Monte Carlo methods. The posterior is a joint distribution of 14 parameters, which are presented in Table 3.3. The parameters partly affect the same processes, and all of the processes are coupled in the model code. For this reason, using samples from the posterior some correlations are to be expected in both the parameters and also between predicted quantities.

The Bayesian calibration was conducted via a hierarchical model described in section 2.7 and shown in figure 2.5. The parameters were divided into two sets: one where the parameters have a changing hyperprior, whose parameters have a fixed prior, and another where the parameters only have a fixed prior. The former are called here (and in Paper II) "hierarchical" and the latter "non-hierarchical" parameters, even though this terminology is not universal. The hierarchical parameters ζ_{exu} and Q_{10} varied from year to year, and their normal priors shared common hyperparameters, with $\frac{1}{\sigma^2} \sim$ Scale-inv-$\chi^2(k, s)$ and $\mu \sim (\mu_0, \sigma_0^2)$, with fixed k, s, μ_0 and σ_0^2. These parameters were sampled with Gibbs sampling (section 2.6.2), and the non-hierarchical parameters (see third column in table 3.3) were sampled with an Adaptive Metropolis step (section 2.6.1).

The sqHIMMELI model calculates CH4 fluxes from the wetland given the model initial state, input data, and parameters. The observation operator is not well known, since even the footprint area of the measurements depends on time-varying external factors such as wind at the surface. Partly for this reason, a heavier tailed Laplace-

Table 3.3: Parameters examined in Paper II. The first column contains parameter symbols, second lists the primary process to which the parameter contributes, and the third lists whether the parameter was modeled in a hierarchical fashion or not. A short functional description of the parameters is given in the last column. The symbol "$\rightarrow$" reads "decomposition into" and T stands for temperature. See also Table 3 in Paper II, which gives the prior limits, units, and references.

	Relevant to	Hier.	Parameter controls...
τ_{exu}	CH4 prod.	no	decay rate of exudates
ζ_{exu}	CH4 prod.	yes	fraction of NPP converted to exudates
τ_{cato}	CH4 prod.	no	rate of peat$\rightarrow$CH4
Q_{10}	CH4 prod.	yes	dependence on T of peat$\rightarrow$CH4
$f_{\text{exu}}^{\text{CH}_4}$	CH4 prod.	no	fraction of anaerobic peat$\rightarrow$ CH4
V_{R0}	Resp.	no	heterotrophic respiration rate
ΔE_R	Resp.	no	dependence of heterotrophic respiration on T
V_{O0}	CH4 oxid.	no	base rate of CH4 oxidation
ΔE_{oxid}	CH4 oxid.	no	dependence of CH4 oxidation on T
λ_{root}	Gas transport	no	root depth
ρ	Gas transport	no	root ending area per biomass
τ	Gas transport	no	root tortuosity parameter
$f_{D,a}$	Gas transport	no	diffusion rate in air-filled peat
$f_{D,d}$	Gas transport	no	diffusion rate in water-filled peat

distributed error model was used with the scaling of the error depending on the day of year, and for this heteroscedasticity model two additional parameters were fitted (see Appendix A of Paper II). The residuals were assumed to be correlated and their covariance structure was described with an ARMA(2,1) model, see section 2.8.1. The ARMA(2,1) parameters were learned as described in section 2.8.2 by minimizing the KL-divergence between the formal error model and the empirical distribution of the residuals. This was done after an initial, exploratory MCMC experiment was conducted to find an approximate posterior mean. The final posterior distribution was estimated using importance resampling, see section 2.6.3, with the exploratory posterior used as a biasing distribution.

3.4.3 Results and discussion

The setting presented in the previous section allows for lots of analysis. Figure 3.8 shows the output fluxes from the posterior mean parameter values, including credible intervals as shaded areas generated by random sampling the error model. The figure visually verifies that the calibrated model is able to produce fluxes that look realistic. The exudate pool sizes and the CH4 emissions closely follow the NPP input and the predictive credible intervals look reasonable.

The parameter posterior distribution shown in figure 3.9 contains various correlations reflecting interchangeability between the processes given the likelihood function

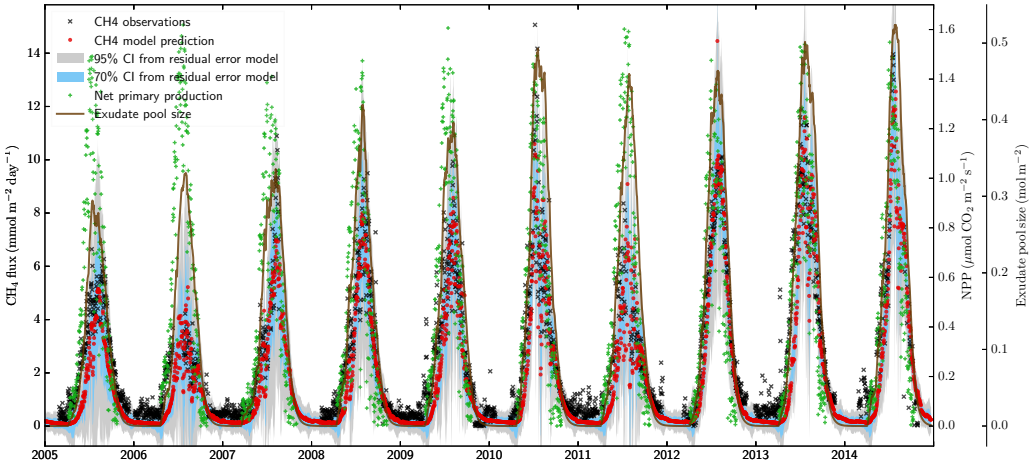


Figure 3.8: Output from the model with posterior mean parameter values. While the fit is good, the calibration is performed with the same observation dataset and therefore the residuals are only relevant for training error.

and the observed flux data. When the model is run with random samples drawn from the posterior distribution, correlations between the processes can be evaluated, as shown in figure 5 in Paper II for year 2012. That figure reveals that plant transport of CH4 (via hollow stems) is driven by exudate decomposition, and that ebullition is in practice perfectly correlated with diffusion, raising the question of whether modeling ebullition is actually an unnecessary complication. With additional data, such as soil gas profiles, the processes might become better separated. Some of the correlations shown in figure 3.9 are strong, and they are rooted in the model equations, but often indirectly. These correlations are thoroughly discussed in sections 5.3 and 5.4 of Paper II.

For prediction, the hierarchical parameter calibration is of course not possible, and therefore other methods needed to be used for obtaining the Q_{10} and ζ_{exu} parameters for predictive purposes. Two schemes were used in Paper II: simply using the mean of the hierarchical parameters, and constructing a regression model for the ζ_{exu} and the Q_{10} parameters. The latter was performed by taking the posterior mean estimates for all the annually changing Q_{10} and ζ_{exu} parameters and then regressing those values against the mean soil temperature at 35 cm depth of the first 10 weeks of each year for Q_{10}, and against the NPP of 130 first days of each year for ζ_{exu}. The annual errors are shown in figure 3.10. In the figure plant transport is missing since it is the complement of diffusion. The term "all ebullition" refers to any ebullition that is released from the underwater part of the peat layer to air, and since water table is most of the time at least slightly under the surface, this is not a real flux, since the gas will be emitted to the atmosphere ultimately via diffusion in the air. On the right, the regression-based predictions are shown to not produce better annual predictions than

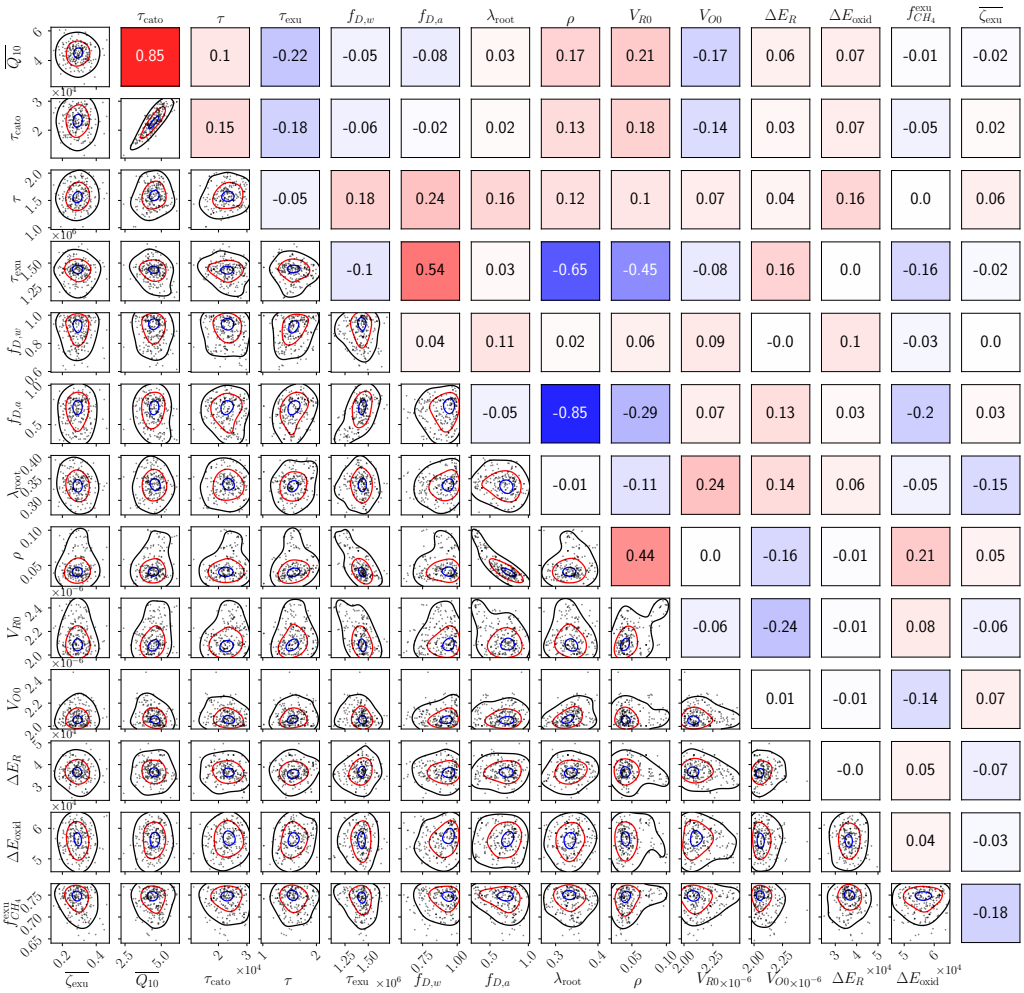


Figure 3.9: Lower triangle on the left: pairwise posterior marginal distributions between parameters, with labels on the left and bottom. The 10%, 50%, and 90% contours, calculated from a kernel density estimate, are shown. The upper right triangle shows pairwise correlations of the parameters with labels on the top and left. For the hierarchically modeled Q_{10} and ζ_{exu} parameters the distributions of the prior means are shown.

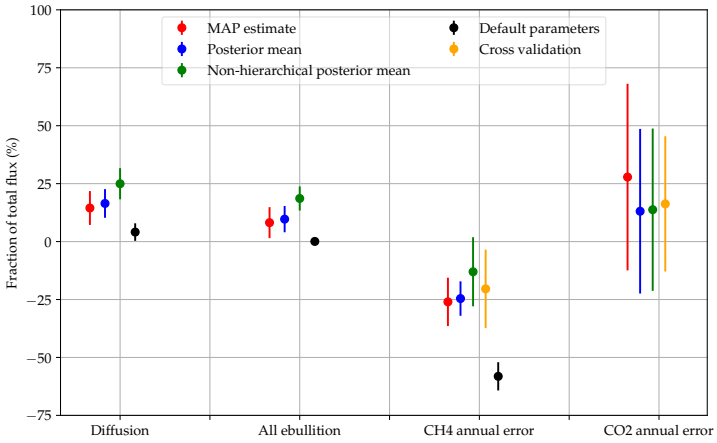


Figure 3.10: Transport component breakdown and annual errors. Plant transport is the complement of diffusion, and ebullition is effectively zero in all cases. For "All ebullition", see text. Predictive results from cross validation are shown in orange in the error descriptions.

the non-hierarchically modeled parameters, implying that either assessing auxiliary performance metrics – such as using time intervals shorter than a year – is needed, or more complicated parametric models need to be constructed for modeling the time dependence of the parameters.

3.5 Climate and land surface modeling

Papers III and IV are different from Papers I and II in many respects, but most importantly they both utilize a significantly more complicated forward model, the JSBACH land surface scheme from the Max Planck Institute for Meteorology (MPI-M) in Hamburg. That model is a part of the ECHAM[3] climate model and it describes processes interfacing the biosphere and the atmosphere. JSBACH makes independent predictions at each grid point based on external forcing. In Paper III that forcing comes from the atmospheric component of the ECHAM6 climate model, and in Paper IV from measured meteorological conditions at flux measurement sites.

This section describes briefly the research findings of Papers III and IV. The objectives and background for them were discussed in section 1, 3.1.3, and 3.1.4. The computational cost and observations used were described briefly in section 3.2.

[3]The JSBACH name creatively stands for Jena Scheme for Biosphere-Atmosphere Coupling in Hamburg. Inspiration for such naming came from a previous model called MOZART. Furthermore, the EC in ECHAM stands for the European Center, from where initial code was adapted, while the HAM part of that name once again refers to Hamburg. For code availability, as of 2019, see mpimet.mpg.de/en/science/models/mpi-esm/jsbach.html.

3.5.1 The ECHAM/JSBACH forward model

The JSBACH model is a complicated PDE model described partly in Roeckner et al. (2003a), more fully in the official model documentation available with the source code, and for relevant parts also in Appendix A of Mäkelä et al. (2016). The ECHAM model, which is used to produce the forcing data for JSBACH in Paper III, solves the atmospheric part including transport of species such as water vapor and trace gases, among everything else. It is a very complicated and heavily parametrized model, and its version 5 is described in the technical reports (Roeckner et al., 2003a,b). The performance of version 6 is further described in Stevens et al. (2013).

The JSBACH model describes different terrain types with *plant functional types*, which summarize the average physical functions of the different terrain types from glaciers to tropical rain forests. In Paper III, particularly changes in areas with the plant functional type *extratropical coniferous forest* were evaluated, and while that same type was used in Paper IV, there the associated parameters were adapted to the local conditions. The most important output variables for the purposes of the Papers were gross primary production, net primary production, evapotranspiration (ET), and snow coverage. All of these variables have to do with the carbon, water, or energy balance of the biosphere-atmosphere boundary.

3.5.2 Paper III – climate change has shifted the growing season

Paper III by Pulliainen et al. (2017) utilizes flux measurement data from the Boreal region, passive microwave retrievals of snow clearance date (SCD), modeling, and meteorological reanalysis data to evaluate how much earlier the starting date of spring recovery (SR) has shifted due to climate change, and how much that shift has affected the carbon balance in the first 180 days of the year. The result is that the onset of spring has become 0.23 days earlier each year, translating into an increase in the uptake of carbon of 52 megatons per decade.

The inference process to produce these estimates was the following: the passive microwave remote sensing data was used to retrieve snow clearance dates, and those data were used with in-situ flux measurements of CO2 to learn the parameters of a regression model for predicting the timing of SR based on SCD. The ECHAM6-JSBACH model was used to calculate the GPP, and earlier SR was found to be weakly correlated with higher springtime GPP.

To produce reliable quantities with modeling, carbon pools in the model were spun up with a 2000-year initial simulation with a lightweight model, CBALANCE, after which a hydrology spin-up was performed using ECHAM with no outside forcing from year 1870 up until 1958. From 1959 onwards the ERA-interim reanalysis dataset was used to nudge the model to keep the meteorology close to the observed, and starting 1979 the ERA-40 dataset was used for that same purpose. The GPP/SR trends were calculated for each grid point from the 36-year period of 1979-2014.

In addition to the global results quoted above, the combination of modeling with

flux measurements allowed looking at the changes regionally. It appears that in Eurasia the change in springtime GPP per decade was proportionally higher (6.8%) than in North America (5.5%). Similarly, the shift of the starting date of spring recovery is also larger in Eurasia, where this figure is a remarkable 3.0 days per decade, while in North America the shift is smaller but still sizable at 1.3 days per decade.

3.5.3 Paper IV – constraining LSS parameters with flux data with adaptive MCMC

In the last included work, Paper IV, parameters of the JSBACH land surface model were calibrated using the Adaptive Metropolis MCMC algorithm. This work has been introduced in section 1, 3.1.4, and 3.2. Markov chain Monte Carlo was described in section 2.6.1.

The work in Mäkelä et al. (2016) utilizes flux data from two measurement sites. The first of these is in Hyytiälä (61°51′N, 24°17′E), and the second one is in Sodankylä (67°22′N, 26°38′E). These sites are long-running measurement sites where the predominant tree species is the Scots pine (*Pinus Sylvesteris*). For Hyytiälä, half-hourly measurements of CO2 and H2O fluxes were used from 1999-2008, while for Sodankylä, the time period was 2000-2008. The JSBACH model calibration used the Hyytiälä data from 2000-2004, whereas for generating the initial conditions for the model the year 1999 was used. For Sodankylä, this spin-up was done with data from all the years and no calibration was performed, instead reusing the data for validation. The aim of the spin-up process was to stabilize the fast carbon pools and the water pools so that local conditions would be represented in initial states of the model.

Since the objective of the study was to improve and better understand how the gas exchange processes in the model are able to describe conditions at these particular sites, the parameters chosen for the calibration were related to gas exchange. These 15 parameters are described in Table 1 of Paper IV. The parameters were calibrated using three different loss functions: one with seasonally averaged data, another one with daily averaged data, and the third one with the original half-hourly data. Three of the parameters were only calibrated with the first one of these.

Even though MCMC usually gives a statistically meaningful posterior distribution, in this work rigorous uncertainty quantification was not attempted as the distributions of the model-observation residuals were not carefully analyzed. The cost functions used were of the standard quadratic form corresponding to a Gaussian observation model

$$\mathscr{L}(\theta) = \sum_i (x - y)^T \Gamma^{-1} (x - y), \tag{3.5}$$

where θ is the model parameter vector, the model output x depends on θ, and the sum is over the (potentially averaged) observations. For the calibration with seasonally averaged data, the vectors x and y contained residuals of mean GPP, mean ET, and maximum LAI, and the diagonal Γ matrix contained, for each period, means

of the observed GPP and ET squared and maximum of the observed LAI squared. For daily and half-hourly calibration LAI was not used, and the elements of Γ were further multiplied with the square root of the number of corresponding observations, inflating the size of the posterior.

A principal component analysis of the MCMC chains revealed that estimates of two parameters controlling bare-soil evaporation – soil dryness-based relative humidity and skin reservoir field capacity (how much water can be held at the very top of the vegetation in a layer of some millimeters) – are in this calibration the least reliable ones. Using the posterior mean values from the MCMC run of the calibration period for Hyytiälä, model performance as measured by (3.5) improved for all the validation runs with the exception that the seasonal calibration in Hyytiälä lead to degraded performance as measured by the daily and half-hourly cost function values. For the Sodankylä site, performance improved with all calibration methods and all metrics when compared to the default parameter values, implying that parameter calibration is generalizable from one site to a similar site at a different location.

The calibrated model was not able to describe a rare drought event in 2006 in Hyytiälä (GPP drop in August 2006 in figure 2 of Paper IV). However, since there were no extended dry periods in the calibration data, the failure of the calibrated model to accommodate for this anomaly was not unexpected.

4 Conclusions and future work

The methods presented in section 2 represent a small and relatively simple subset of the very large number of techniques nowadays used for uncertainty quantification and data science. Similarly, the context provided by climate change, and more generally geosciences, is huge, and therefore this work scratches only a corner or two of an immense problem space. In this sense it is fortunate that the mathematical theory is agnostic to the applications and the methods and algorithms can easily be reused.

Each of the Papers presented contained three building blocks: models, data, and algorithms. These building blocks were together used to answer specific climate change-related research questions: statistical models marry process models and observational data, and carefully analyzing the different aspects of the model-observation mismatch enabled the utilization of Bayes' theorem for solving inverse problems, either with Monte Carlo methods or via point estimation.

While models and data were used in all the Papers, only the first two utilized non-trivial statistical estimation techniques to try to understand the statistical properties of the data. Still, even in those two publications, much room was left for further analysis, and in Papers III and IV the price for omitting Bayesian uncertainty quantification was that the posterior and posterior predictive uncertainties remained unknown. On one hand this lack of uncertainty quantification adversely affects how actionable the results are, but on the other hand when expensive computational models are used, conducting Bayesian analysis is often impossible. This was in particular the case with Paper III.

Certain themes recur when evaluating how the research could have been improved. When model-generated input data – for instance wind data in Paper I or leaf area index, net primary production, and water table depth data in Paper II – were used, the propagation of uncertainties pertaining to those quantities were overlooked. While disregarding uncertainties in input data is often necessary, the implications of that are that uncertainty estimates from settings involving modeled input data and complex models need to be approached with caution. The flip side of the coin is that even when all modeling is perfect, the results of any inference are only as good as the data that is used. This was most evident in Paper I, where the quality of the uncertainty information provided with the XCO2 observations was not always reliable.

The work in the Papers may be critiqued in more specific ways to guide future research endeavors. In Paper I the covariance between measurement errors of the

individual measurements are not known, and neither are the various biases that are known to exist in the data. Regarding satGP, there is room for development in how observations for each subkernel are selected, and the effects of this still need to be analyzed and minimized. The β coefficient fields with their uncertainties may provide further useful information that can be used to devise better formulations of the mean function. Other possible next steps include applying the satGP software to other problems, combining multiple data products, performing model selection to select the best combinations of subkernels for the multi-scale kernels, and general code development and usability enhancements.

The most pressing issue in Paper II is the lack of uncertainty quantification for input data generation. Following that, the error modeling can be further enhanced by treating the instrument error and other error sources separately in the observation equation, potentially yielding improved models for describing the data. Cross validation at other measurement sites and computing regional fluxes with uncertainties would be valuable, both in terms of the actual results and in terms of learning how well the modeled processes actually describe what they are intended to describe.

The regression plots in Paper III show large deviations, which tend to disproportionally affect the trends when Gaussian errors are assumed (e.g. figure 4 in Paper III). Furthermore, while Paper III includes uncertainties in the presentation of the springtime GPP increase due to changes in the spring recovery date (Table 1 in Paper III), those trends were calculated using data from only two measurement stations in both Eurasia and North America, and this may lead to increased representation error.

The *ad hoc* nature of the cost function formulations in Paper IV rules out proper uncertainty quantification, and with it e.g. the possibility to compute Monte Carlo estimates of future carbon balance based on parameter posteriors. The differences in the optimal parameter values between the different loss function formulations shows how important data selection and averaging are, and points out that the design of any model calibration exercise must be based on future modeling needs. The incapability of the model to describe the dry event in the summer of 2006 suggests that process modifications need to be carried out. That this work was undertaken in Mäkelä et al. (2019) (see section 3.1.5) serves as an example of how process models may and should be improved based on statistical analyses.

When research is constrained by the availability and quality of observations, collecting more data and refining the analyses little by little provides more and more confidence in the conclusions. This is what the IPCC reports describe, with each new version having more weight and urgency in both the details and the overall message.

The research presented in this thesis consists of technical results related to climate change, carbon cycle, models, and data. These technicalities, however, hide an important aspect of the work, which is to underline that climate change has already advanced very far (Papers I,III), and this results in unpredictable and difficult-to-model phenomena (Papers II-IV). For these reasons, action needs to be taken to address the problems reported by the scientific community in addition to performing

and funding more research. While a scientist can use Bayesian analysis to improve the models, that same analysis can also be used by policy makers and voters as a small ingredient in cooking up a way to save the world from the most catastrophic climate change scenarios.

References

P. Bickel and K. Doksum. *Mathematical Statistics 2e*, volume 1. CRC Press, 1st edition, 2015. ISBN 9781498723800.

P. Bickel and K. Doksum. *Mathematical Statistics 2e*, volume 1. CRC Press, 2nd edition, 2016. ISBN 9781498722681.

A. Bouchard-Côté, S. J. Vollmer, and A. Doucet. The Bouncy Particle Sampler: A Non-Reversible Rejection-Free Markov Chain Monte Carlo Method. *arXiv e-prints*, art. arXiv:1510.02451, Oct 2015.

S. P. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004. ISBN 0521833787.

G. Casella and R. Berger. *Statistical Inference*. Duxbury advanced series in statistics and decision sciences. Thomson Learning, 2002. ISBN 9780534243128.

C. Chatfield. *The analysis of time series: an introduction*. Chapman and Hall, 4th edition, 1989.

P. G. Constantine, C. Kent, and T. Bui-Thanh. Accelerating MCMC with active subspaces. *arXiv e-prints*, art. arXiv:1510.00024, Sep 2015.

D. Crisp, B. M. Fisher, C. O'Dell, C. Frankenberg, R. Basilio, H. Bösch, L. R. Brown, R. Castano, B. Connor, N. M. Deutscher, A. Eldering, D. Griffith, M. Gunson, A. Kuze, L. Mandrake, J. McDuffie, J. Messerschmidt, C. E. Miller, I. Morino, V. Natraj, J. Notholt, D. M. O'Brien, F. Oyafuso, I. Polonsky, J. Robinson, R. Salawitch, V. Sherlock, M. Smyth, H. Suto, T. E. Taylor, D. R. Thompson, P. O. Wennberg, D. Wunch, and Y. L. Yung. The ACOS CO_2 retrieval algorithm – ndash; Part II: Global XCO_2 data characterization. *Atmospheric Measurement Techniques*, 5(4):687–707, 2012. doi: 10.5194/amt-5-687-2012. URL https://www.atmos-meas-tech.net/5/687/2012/.

T. Cui, J. Martin, Y. M. Marzouk, A. Solonen, and A. Spantini. Likelihood-informed dimension reduction for nonlinear inverse problems. *Inverse Problems*, 30(11): 114015, oct 2014. doi: 10.1088/0266-5611/30/11/114015. URL https://doi.org/10.1088%2F0266-5611%2F30%2F11%2F114015.

F.-X. L. Dimet and O. Talagrand. Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects. *Tellus A*, 38A(2):97–110, 1986. doi: 10.1111/j.1600-0870.1986.tb00459.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1600-0870.1986.tb00459.x.

S. Duane, A. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid monte carlo. *Physics Letters B*, 195(2):216 – 222, 1987. ISSN 0370-2693. doi: https://doi.org/10.1016/0370-2693(87)91197-X. URL http://www.sciencedirect.com/science/article/pii/037026938791197X.

J. Durbin and S. Koopman. *Time Series Analysis by State Space Methods: second Edition*. Oxford Statistical Science Series. OUP Oxford, 2012. ISBN 9780199641178.

J. Durbin and G. Watson. Testing for serial correlation in least-squares regression, I. *Biometrika*, 37:409–428, 1950.

J. Durbin and G. Watson. Testing for serial correlation in least-squares regression, Ii. *Biometrika*, 38:159–178, 1951.

D. Gamerman. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 1997. ISBN 9780412818202.

A. L. Ganesan, M. Rigby, A. Zammit-Mangion, A. J. Manning, R. G. Prinn, P. J. Fraser, C. M. Harth, K.-R. Kim, P. B. Krummel, S. Li, J. Mühle, S. J. O'Doherty, S. Park, P. K. Salameh, L. P. Steele, and R. F. Weiss. Characterization of uncertainties in atmospheric trace gas inversions using hierarchical bayesian methods. *Atmospheric Chemistry and Physics*, 14(8):3855–3864, 2014. doi: 10.5194/acp-14-3855-2014. URL https://www.atmos-chem-phys.net/14/3855/2014/.

A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 3rd edition, 2013.

I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Networks. *arXiv e-prints*, art. arXiv:1406.2661, Jun 2014.

U. Grenander and M. I. Miller. Representations of knowledge in complex systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(4):549–603, 1994. ISSN 00359246. URL http://www.jstor.org/stable/2346184.

M. Gruber. *Matrix Algebra for Linear Models*. Wiley, 2013. ISBN 9781118608814.

H. Haario, E. Saksman, and J. Tamminen. An adaptive metropolis algorithm. *Bernoulli*, 7(2):223–242, 2001.

H. Haario, M. Laine, A. Mira, and E. Saksman. Dram: Efficient adaptive mcmc. *Statistics and Computing*, 16(4):339–354, 2006. ISSN 1573-1375. doi: 10.1007/s11222-006-9438-0. URL http://dx.doi.org/10.1007/s11222-006-9438-0.

D. M. Hammerling, A. M. Michalak, and S. R. Kawa. Mapping of CO2 at high spatiotemporal resolution using satellite observations: Global distributions from OCO-2. *Journal of Geophysical Research: Atmospheres*, 117(D6), 2012a. doi: 10.1029/2011JD017015. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2011JD017015.

D. M. Hammerling, A. M. Michalak, C. O'Dell, and S. R. Kawa. Global CO2 distributions over land from the Greenhouse Gases Observing Satellite (GOSAT). *Geophysical Research Letters*, 39(8), 2012b. doi: 10.1029/2012GL051203. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2012GL051203.

J. Hammersley and P. Clifford. Markov random fields on finite graphs and lattices. 1971.

A. C. Harvey. *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, 1990. doi: 10.1017/CBO9781107049994.

M. J. Heaton, A. Datta, A. Finley, R. Furrer, R. Guhaniyogi, F. Gerber, R. B. Gramacy, D. Hammerling, M. Katzfuss, F. Lindgren, D. W. Nychka, F. Sun, and A. Zammit-Mangion. A Case Study Competition Among Methods for Analyzing Large Spatial Data. *arXiv e-prints*, art. arXiv:1710.05013, Oct 2017.

IPCC. *Summary for Policymakers*, book section SPM, pages 1–30. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013. ISBN ISBN 978-1-107-66182-0. doi: 10.1017/CBO9781107415324.004. URL www.climatechange2013.org.

R. Kalman. A new approach to linear filtering and prediction problems. *Transactions of ASME – Journal of Basic Engineering*, 82:35–45, 1960.

I. Karatzas and S. E. Shreve. *Brownian Motion and Stochastic Calculus*. Springer Graduate Texts in Mathematics. Springer-Verlag, "2nd" edition, 1998. doi: 10.1007/978-1-4684-0302-2.

M. Katzfuss, J. Guinness, and W. Gong. Vecchia approximations of Gaussian-process predictions. *arXiv e-prints*, art. arXiv:1805.03309, May 2018.

M. C. Kennedy and A. O'Hagan. Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, 87(1):1–13, 2000. ISSN 00063444. URL http://www.jstor.org/stable/2673557.

M. Laine, N. Latva-Pukkila, and E. Kyrölä. Analysing time-varying trends in stratospheric ozone time series using the state space approach. *Atmospheric Chemistry and Physics*, 14(18):9707–9725, 2014. doi: 10.5194/acp-14-9707-2014. URL https://www.atmos-chem-phys.net/14/9707/2014/.

M. Lassas and S. Siltanen. Can one use total variation prior for edge-preserving Bayesian inversion? *Inverse Problems*, 20:1537, 08 2004. doi: 10.1088/0266-5611/20/5/013.

S. Lauritzen. *Graphical Models*. Oxford Statistical Science Series. Clarendon Press, 1996. ISBN 9780191591228.

K. Law, A. Stuart, and K. Zygalakis. *Data Assimilation*. Springer International Publishing, 2015. ISBN 9783319203249.

F. Lindgren, H. Rue, and J. Lindström. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498, 2011. doi: 10.1111/j.1467-9868.2011.00777.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2011.00777.x.

D. Lu, D. Ricciuto, A. Walker, C. Safta, and W. Munger. Bayesian calibration of terrestrial ecosystem models: a study of advanced markov chain monte carlo methods. *Biogeosciences*, 14(18):4295–4314, 2017. doi: 10.5194/bg-14-4295-2017. URL https://www.biogeosciences.net/14/4295/2017/.

J. Mäkelä, J. Susiluoto, T. Markkanen, M. Aurela, H. Järvinen, I. Mammarella, S. Hagemann, and T. Aalto. Constraining ecosystem model with adaptive Metropolis algorithm using boreal forest site eddy covariance measurements. *Nonlinear Processes in Geophysics*, 23(6):447–465, 2016. doi: 10.5194/npg-23-447-2016. URL http://www.nonlin-processes-geophys.net/23/447/2016/.

J. Mäkelä, J. Knauer, M. Aurela, A. Black, M. Heimann, H. Kobayashi, A. Lohila, I. Mammarella, H. Margolis, T. Markkanen, J. Susiluoto, T. Thum, T. Viskari, S. Zaehle, and T. Aalto. Parameter calibration and stomatal conductance formulation comparison for boreal forests with adaptive population importance sampler in the land surface model jsbach. *Geoscientific Model Development*, 12(9):4075–4098, 2019. doi: 10.5194/gmd-12-4075-2019. URL https://www.geosci-model-dev.net/12/4075/2019/.

J. Mueller and S. Siltanen. *Linear and Nonlinear Inverse Problems with Practical Applications*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2012. doi: 10.1137/1.9781611972344. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611972344.

J. A. Nelder and R. Mead. A Simplex Method for Function Minimization. *The Computer Journal*, 7(4):308–313, 01 1965. ISSN 0010-4620. doi: 10.1093/comjnl/7.4.308. URL https://doi.org/10.1093/comjnl/7.4.308.

H. Nguyen, M. Katzfuss, N. Cressie, and A. Braverman. Spatio-temporal data fusion for very large remote sensing datasets. *Technometrics*, 56(2):174–185, 2014. doi: 10.1080/00401706.2013.831774. URL https://doi.org/10.1080/00401706.2013.831774.

J. Nocedal. Updating Quasi-Newton matrices with limited storage. *Math. Comput.*, 35(151):773–782, 07 1980.

C. W. O'Dell, B. Connor, H. Bösch, D. O'Brien, C. Frankenberg, R. Castano, M. Christi, D. Crisp, A. Eldering, B. Fisher, M. Gunson, J. McDuffie, C. E. Miller, V. Natraj, F. Oyafuso, I. Polonsky, M. Smyth, T. Taylor, G. C. Toon, P. O. Wennberg, and D. Wunch. Corrigendum to "The ACOS CO_2 retrieval algorithm – Part 1: Description and validation against synthetic observations" published in atmos. meas. tech., 5, 99–121, 2012. *Atmospheric Measurement Techniques*, 5(1):193–193, 2012. doi: 10.5194/amt-5-193-2012. URL https://www.atmos-meas-tech.net/5/193/2012/.

B. Øksendal. *Stochastic Differential Equations: An Introduction with Applications*. Universitext. Springer Berlin Heidelberg, 2010. ISBN 9783642143946.

B. Peherstorfer, K. Willcox, and M. Gunzburger. Survey of multifidelity methods in uncertainty propagation, inference, and optimization. *SIAM Review*, 60(3): 550–591, 2018. doi: 10.1137/16M1082469. URL https://doi.org/10.1137/16M1082469.

G. Peyré and M. Cuturi. Computational Optimal Transport. *arXiv e-prints*, art. arXiv:1803.00567, Mar 2018.

M. Powell. The BOBYQA algorithm for bound constrained optimization without derivatives. Report, Department of Applied Mathematics and Theoretical Physics, Cambridge, UK, 08 2009.

J. Pulliainen, M. Aurela, T. Laurila, T. Aalto, M. Takala, M. Salminen, M. Kulmala, A. Barr, M. Heimann, A. Lindroth, A. Laaksonen, C. Derksen, A. Mäkelä, T. Markkanen, J. Lemmetyinen, J. Susiluoto, S. Dengel, I. Mammarella, J.-P. Tuovinen, and T. Vesala. Early snowmelt significantly enhances boreal springtime carbon uptake. *Proceedings of the National Academy of Sciences*, 114(42): 11081–11086, 2017. ISSN 0027-8424. doi: 10.1073/pnas.1707889114. URL http://www.pnas.org/content/114/42/11081.

M. Raivonen, S. Smolander, L. Backman, J. Susiluoto, T. Aalto, T. Markkanen, J. Mäkelä, J. Rinne, O. Peltola, M. Aurela, A. Lohila, M. Tomasic, X. Li, T. Larmola, S. Juutinen, E.-S. Tuittila, M. Heimann, S. Sevanto, T. Kleinen, V. Brovkin,

and T. Vesala. HIMMELI v1.0: HelsinkI Model of MEthane buiLd-up and emission for peatlands. *Geoscientific Model Development*, 10(12):4665–4691, 2017. doi: 10.5194/gmd-10-4665-2017. URL https://www.geosci-model-dev.net/10/4665/2017/.

C. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. Adaptative computation and machine learning series. University Press Group Limited, 2006. ISBN 9780262182539.

A. D. Richardson, D. Y. Hollinger, G. G. Burba, K. J. Davis, L. B. Flanagan, G. G. Katul, J. W. Munger, D. M. Ricciuto, P. C. Stoy, A. E. Suyker, S. B. Verma, and S. C. Wofsy. A multi-site analysis of random error in tower-based measurements of carbon and energy fluxes. *Agricultural and Forest Meteorology*, 136(1–2): 1 – 18, 2006. ISSN 0168-1923. doi: http://dx.doi.org/10.1016/j.agrformet.2006.01.007. URL http://www.sciencedirect.com/science/article/pii/S0168192306000281.

G. O. Roberts, A. Gelman, and W. R. Gilks. Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.*, 7(1):110–120, 02 1997. doi: 10.1214/aoap/1034625254. URL http://dx.doi.org/10.1214/aoap/1034625254.

E. Roeckner, G. Bäuml, L. Bonaventura, R. Brokopf, M. Esch, M. Giorgetta, S. Hagemann, I. Kirchner, L. Kornblueh, E. Manzini, A. Rhodin, U. Schlese, U. Schulzweida, and A. Tompkins. The atmospheric general circulation model ECHAM5. Part i: Model description. Report 349, Max-Planck-Institut für Meteorologie, Hamburg, 2003a.

E. Roeckner, G. Bäuml, L. Bonaventura, R. Brokopf, M. Esch, M. Giorgetta, S. Hagemann, I. Kirchner, L. Kornblueh, E. Manzini, A. Rhodin, U. Schlese, U. Schulzweida, and A. Tompkins. The atmospheric general circulation model ECHAM5. Part ii: Simulated climatology and comparison with observations. Report 354, Max-Planck-Institut für Meteorologie, Hamburg, 2003b.

L. Roininen, S. Lasanen, M. Orispää, and S. Särkkä. Sparse approximations of fractional Matérn fields. *Scandinavian Journal of Statistics*, 45(1):194–216, 2018. doi: 10.1111/sjos.12297. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/sjos.12297.

Y. Rozanov. *Random Fields and Stochastic Partial Differential Equations*. Kluwer Academic Publishers, 1998. ISBN 0792349849.

W. Rudin. *Real and Complex Analysis*. McGraw-Hill, 1987. ISBN 0070542341.

T. P. Runarsson and Xin Yao. Search biases in constrained evolutionary optimization. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 35(2):233–243, May 2005. doi: 10.1109/TSMCC.2004.841906.

T. Santner, B. Williams, and W. Notz. *The Design and Analysis of Computer Experiments*. Springer Series in Statistics. Springer Verlag New York, 2003. ISBN 0387954201.

S. Särkkä. *Bayesian Filtering and Smoothing*. Bayesian Filtering and Smoothing. Cambridge University Press, 2013. ISBN 9781107030657.

B. Stevens, M. Giorgetta, M. Esch, T. Mauritsen, T. Crueger, S. Rast, M. Salzmann, H. Schmidt, J. Bader, K. Block, R. Brokopf, I. Fast, S. Kinne, L. Kornblueh, U. Lohmann, R. Pincus, T. Reichler, and E. Roeckner. Atmospheric component of the MPI-M Earth system model: ECHAM6. *Journal of Advances in Modeling Earth Systems*, 5(2):146–172, 2013. doi: 10.1002/jame.20015. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/jame.20015.

D. Stroock. *Elements of Stochastic Calculus and Analysis*. CRM Short Courses. Springer International Publishing, 2018. ISBN 9783319770383.

A. M. Stuart. Inverse problems: A Bayesian perspective. *Acta Numerica*, 19:451–559, 2010. doi: 10.1017/S0962492910000061.

J. Susiluoto, A. Spantini, H. Haario, and Y. Marzouk. Efficient multi-scale gaussian process regression for massive remote sensing data with satgp v0.1. *Geoscientific Model Development Discussions*, 2019:1–30, 2019. doi: 10.5194/gmd-2019-156. URL https://www.geosci-model-dev-discuss.net/gmd-2019-156/.

J. M. Tadić, X. Qiu, S. Miller, and A. M. Michalak. Spatio-temporal approach to moving window block kriging of satellite data v1.0. *Geoscientific Model Development*, 10(2):709–720, 2017. doi: 10.5194/gmd-10-709-2017. URL https://www.geosci-model-dev.net/10/709/2017/.

A. Tarantola. *Inverse Problem Theory and Methods for Model Parameter Estimation*. Society for Industrial and Applied Mathematics, 2005. ISBN 9780898715729.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. ISSN 00359246. URL http://www.jstor.org/stable/2346178.

L. Tierney. Markov chains for exploring posterior distributions. *Ann. Statist.*, 22(4): 1701–1728, 12 1994. doi: 10.1214/aos/1176325750. URL https://doi.org/10.1214/aos/1176325750.

L. Tierney and A. Mira. Some adaptive monte carlo methods for bayesian inference. *Statistics in Medicine*, 18(17-18):2507–2515, 1999. doi: 10.1002/(SICI)1097-0258(19990915/30)18:17/18⟨2507::AID-SIM272⟩3.0.CO;2-J. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/.

M. K. Titsias and O. Papaspiliopoulos. Auxiliary gradient-based sampling algorithms. *arXiv e-prints*, art. arXiv:1610.09641, Oct 2016.

L. Trefethen and D. Bau. *Numerical Linear Algebra*. Society for Industrial and Applied Mathematics (SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104), 1997. ISBN 9780898719574.

A. V. Vecchia. Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, 50(2):297–312, 1988. ISSN 00359246. URL http://www.jstor.org/stable/2345768.

R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008. doi: 10.1561/2200000001. URL http://www.eecs.berkeley.edu/~wainwrig/Papers/WaiJor08_FTML.pdf.

D. Williams. *Probability with Martingales*. Cambridge University Press, 1991. ISBN 9780521406055.

A. Zammit-Mangion, N. Cressie, A. L. Ganesan, S. O' Doherty, and A. J. Manning. Spatio-temporal bivariate statistical models for atmospheric trace-gas inversion. *ArXiv e-prints*, Sept. 2015.

A. Zammit-Mangion, N. Cressie, and C. Shumack. On statistical approaches to generate level 3 products from satellite remote sensing retrievals. *Remote Sensing*, 10(1), 2018. ISSN 2072-4292. doi: 10.3390/rs10010155. URL http://www.mdpi.com/2072-4292/10/1/155.

Z. Zeng, L. Lei, L. Guo, L. Zhang, and B. Zhang. Incorporating temporal variability to improve geostatistical analysis of satellite-observed CO2 in China. *Chinese Science Bulletin*, 58(16):1948–1954, Jun 2013. ISSN 1861-9541. doi: 10.1007/s11434-012-5652-7. URL https://doi.org/10.1007/s11434-012-5652-7.

Z.-C. Zeng, L. Lei, K. Strong, D. B. A. Jones, L. Guo, M. Liu, F. Deng, N. M. Deutscher, M. K. Dubey, D. W. T. Griffith, F. Hase, B. Henderson, R. Kivi, R. Lindenmaier, I. Morino, J. Notholt, H. Ohyama, C. Petri, R. Sussmann, V. A. Velazco, P. O. Wennberg, and H. Lin. Global land mapping of satellite-observed CO2 total columns using spatio-temporal geostatistics. *International Journal of Digital Earth*, 10(4):426–456, 2017. doi: 10.1080/17538947.2016.1156777. URL https://doi.org/10.1080/17538947.2016.1156777.

ILMATIETEEN LAITOS
METEOROLOGISKA INSTITUTET
FINNISH METEOROLOGICAL INSTITUTE

FINNISH METEOROLOGICAL INSTITUTE
Erik Palménin aukio 1
P.O. Box 503
FI-00560 HELSINKI
tel. +358 29 539 1000
WWW.FMI.FI

FINNISH METEOROLOGICAL INSTITUTE
CONTRIBUTIONS No. 154
ISBN 978-952-336-080-8 (paperback)
ISSN 0782-6117
Edita Prima Oy 2019

ISBN 978-952-336-081-5 (pdf)
https://doi.org/10.35614/isbn.9789523360815
Helsinki 2019