Satistical Modelling Of CO2 Exchange Between Land And Atmosphere

Using Stochastic Optimisation And Gaussian Markov Random Fields

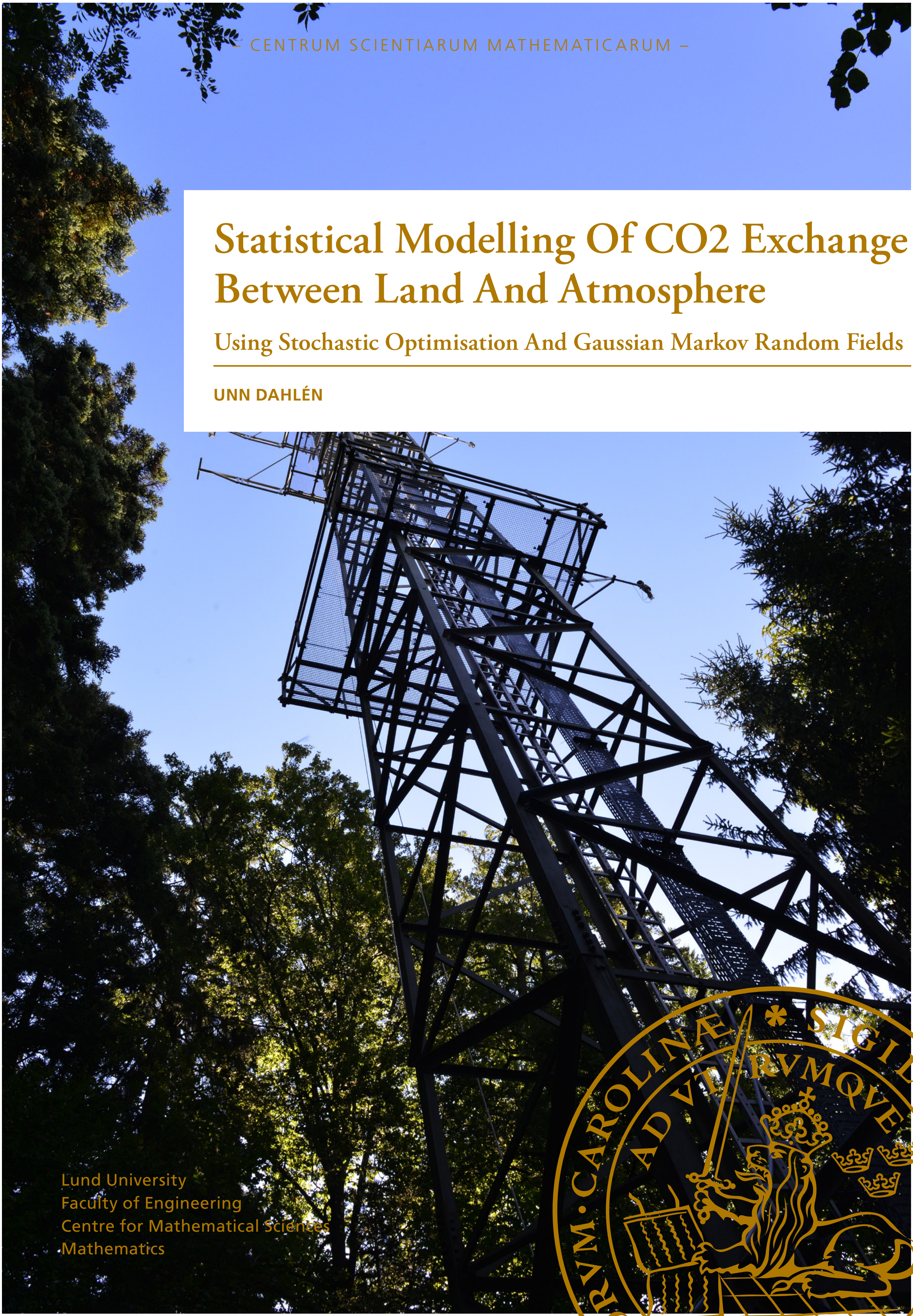
Dahlén, Unn

– CENTRUM SCIENTIARUM MATHEMATICARUM –

Statistical Modelling Of CO2 Exchange Between Land And Atmosphere

Using Stochastic Optimisation And Gaussian Markov Random Fields

UNN DAHLÉN



Lund University
Faculty of Engineering
Centre for Mathematical Sciences
Mathematics

Printed by Media-Tryck, Lund 2019 NORDIC SWAN ECOLABEL 3041 0903

LUND
UNIVERSITY

Doctoral Theses in Mathematical Sciences 2019:03

ISBN 978-91-7895-175-8
ISSN 1404-0034

Statistical Modelling of CO2 Exchange between Land and Atmosphere

Using Stochastic Optimisation and Gaussian Markov Random Fields

Unn Dahlén

LUND
UNIVERSITY

DOCTORAL DISSERTATION

by due permission of the Faculty of Engineering, Lund University, Sweden. To be defended at the Center of Mathematical Sciences, room MH:R. Date 2019-09-20 and time 09.15.

Faculty opponent
Martin Sköld

Organization
LUND UNIVERSITY

Centre of Mathematical Sciences
Mathematical Statistics
Box 118
SE-221 00 Lund, Sweden

Document name
DOCTORAL DISSERTATION

Date of issue
2019-09-20

Author(s)
Unn Dahlén

Sponsoring organization

Title and subtitle
Statistical modelling of CO2 exchange between land and atmosphere:
Using stochastic optimisation and Gaussian Markov random fields

Abstract

This thesis focuses on the development and application of efficient mathematical tools for estimating and modelling the exchange of carbon dioxide (CO2) between the Earth and its atmosphere; here referred to as the global CO2 surface flux. There are two main approaches for estimating the CO2 flux: Processed based (bottom up) modelling and atmospheric inversion (top down) modelling. The first part of the thesis focuses on applying and improve methods for estimating unknown or uncertain parameters in ecosystem models. This can partly be seen as an optimization problem, since the task is to find the parameter set which gives a modelled flux output closest to the flux observations with respect to certain model assumptions. Standard gradient based optimization methods are seldom applicable since the derivatives are commonly unknown and, due to the complex interactions between flux output and model parameters, the system is highly nonlinear and often multimodal. We show that a popular model-based search method, Gradient Adaptive Stochastic Search (GASS), which combines importance sampling with some second order gradient information, can be used for efficient parameter inference. Furthermore, the importance sampling for this method is improved by forming probabilistic distributions based on good samples from previous iterations in the algorithm.

Secondly, the thesis deals with atmospheric inversions, where time series of CO2 concentrations taken from a global network of measurement stations are used together with an atmospheric transport model, to obtain a reconstruction of the CO2 surface flux. For this application we introduce a new concept of modelling the surface flux, by using Gaussian Markov Random Fields (GMRF) defined on a continuous spatial domain. In contrast to previous inversion methods, the modelled concentrations are obtained from a highly resolved spatial integration, while keeping a discrete temporal resolution. The smooth representation of the flux reduces aggregation errors present in traditional flux representations restricted to a grid, and allows the flux covariance to be estimated on a continuous spatial domain. Modelling the CO2 flux using GMRFs open up for the use of numerical methods for sparse matrices. The last part of the thesis presents methods for improving the inference on our GMRF model, by using Markov Chain Monte Carlo methods. We show that using Crank Nicholson based proposals, significantly reduces the computational time needed for estimating CO2 flux in atmospheric inverse modelling.

Key words
Gaussian Markov random fields, Stochastic optimization, Markov chain Monte Carlo, Atnospheric inverse mondelling, CO2 flux.

Classification system and/or index terms (if any)

Supplementary bibliographical information

Language

ISSN and key title
1404-0034

ISBN
978-91-7895-175-8

Recipient's notes

Number of pages
245

Price

Security classification

I, the undersigned, being the copyright owner of the abstract of the above-mentioned dissertation, hereby grant to all reference sources permission to publish and disseminate the abstract of the above-mentioned dissertation.

Signature

Date 2019-08-12

Statistical Modelling of CO2 Exchange Between Land and Atmosphere

using stochastic optimisation and Gaussian Markov Random fields

Unn Dahlén

Lund University

LTH
Faculty of Engineering

Faculty of Engineering
Centre for Mathematical Sciences
Mathematical Statistics

Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Box 118
SE-221 00 Lund
Sweden

http://www.maths.lth.se/

Doctoral Theses in Mathematical Sciences 2019:03
ISSN 1404-0034

ISBN 978-91-7895-175-8 (print)
ISBN 978-91-7895-176-5 (pdf)
LUTFMS-1047-2019
Cover photo by Peter Rüegg, ETH Zurich

Printed in Sweden by Media Tryck, Lund 2019

Contents

List of papers

This thesis consists of the following papers:

Paper A Unn Dahlén, Marko Scholze, Stefan Ohlin, Andrew McRobert, and Johan Lindström, ”Parameter Optimisation of Terrestrial Ecosystem Models using Gradient Adaptive Stochastic Search”. To be submitted to *Mathematics of Climate and Weather Forecasting*.

Paper B Unn Dahlén, Johan Lindström and Marko Scholze: ”Using memory-based importance sampling to improve stochastic gradient optimisation of vegetation models”. To be submitted to *Computational Optimization and Applications*.

Paper C Unn Dahlén, Johan Lindström and Marko Scholze: ”Inverse modelling of spatio-temporal CO_2 flux fields using Gaussian Markov Random Fields” Submitted to *Environmetrics*.

Paper D Johan Lindström, Unn Dahlén: ”An efficient MCMC method for parameter inference in atmospheric inverse modelling of CO_2 using Gaussian Markov Random Fields”. Manuscript in preparation.

Paper E Unn Dahlén, Nils Rydén and Andreas Jakobsson: ”Damage Identification in Concrete using Impact Non-Linear Reverberation Spectroscopy”. *NDT & E International*, vol. 53, April 2015.

Additional papers not included in the thesis:

1. Nils Ryden, Unn Dahlén and Andreas Jakobsson, ”Characterization of Progressive Damage in Concrete Using Impact Nonlinear Reverberation Spectroscopy”. *International Symposium Non-Destructive Testing in Civil Engineering* (NDT-CE 2015).

2. Nils Ryden, Unn Dahlén, Per Lindh, Andreas Jakobsson, ”Impact non-linear reverberation spectroscopy applied to non-destructive testing of building materials”.*The Journal of the Acoustical Society of America*, vol. 140, no. 4, pp. 3327-3327, 2016.

Acknowledgements

Five years of hard work is finally conveyed to a safe place; my thesis. I would like to thank everyone who has joined me on this journey: colleges, friends, and family. Especially thanks to my supervisor, Johan Lindström, for all support and guidance through this time. Many thanks also to my second supervisor, Marko Scholze, for always being encouraging, and for providing me with a lot of data. I would like to thank all people in the administration at the mathematical statistics department, present and former, for providing help of all kinds. James, for being one of the most great-hearted people I know, always caring and full of energy. Maria, for being a super-sweet person that I care for a lot. A special thanks to my favourite PhD-colleges, Rachele, Mareile and Carl; spending time with you has made this time so much funnier. Rachele, thanks for always being so supportive, Carl, particularly thanks for all the computer support.

On a private level, I want to thank my dear friends, Lina, Rebecca, Maria, and Therese, for being present in my life and for always being encouraging. I would like to thank my family, who have always supported me, whatever choice I made. Finally, I would like to thank my partner Ruben. I am so grateful to share everyday life with you, and I look so much forward to have a family of our own. That means more than anything else. I love you.

Unn Dahlén
Lund, August 2019

Abstract

This thesis focuses on the development and application of efficient mathematical tools for estimating and modelling the exchange of carbon dioxide (CO_2) between the Earth and its atmosphere; here referred to as the global CO_2 surface flux.

There are two main approaches for estimating the CO_2 flux: Processed based (bottom-up) modelling and atmospheric inversion (top-down) modelling. The first part of the thesis focuses on applying and improve methods for estimating unknown or uncertain parameters in ecosystem models. This can partly be seen as an optimization problem since the task is to find the parameter set which gives a modelled flux output closest to the flux observations with respect to certain model assumptions. Standard gradient-based optimization methods are seldom applicable since the derivatives are commonly unknown and, due to the complex interactions between flux output and model parameters, the system is highly non-linear and often multimodal.

We show that a popular model-based search method, Gradient Adaptive Stochastic Search (GASS), which combines importance sampling with some second-order gradient information, can be used for efficient parameter inference. Furthermore, the importance sampling for this method is improved by forming probabilistic distributions based on good samples from previous iterations in the algorithm.

Secondly, the thesis deals with atmospheric inversions, where time series of CO_2 concentrations taken from a global network of measurement stations are used together with an atmospheric transport model, to obtain a reconstruction of the CO_2 surface flux.

For this application, we introduce a new concept of modelling the surface flux, by using Gaussian Markov Random Fields (GMRF) defined on a continuous spatial domain. In contrast to previous inversion methods, the modelled concentrations are obtained from a highly resolved spatial integration, while keeping a discrete temporal resolution. The smooth representation of the flux reduces aggregation errors present in traditional flux representations restricted to a grid and allows the flux covariance to be estimated on a continuous spatial domain.

Modelling the CO_2 flux using GMRFs open up for the use of numerical methods for sparse matrices. The last part of the thesis presents methods for improving the inference on our GMRF model, by using Markov Chain Monte Carlo methods. We show that using Crank Nicholson based proposals significantly reduces the computational time needed for estimating CO_2 flux in atmospheric inverse modelling.

Populärvetenskaplig sammanfattning

Den främsta orsaken till en ökad växthuseffekt anses idag bero på en högre koncentration av koldioxid i atmosfären. Sedan den industriella revolutionen har människan påverkat mängden koldioxid i atmosfären genom aktiviteter så som förbränning av fossila bränslen, avskogning, förändringar i markanvändning, samt cementproduktion. För att kunna motverka framtida klimatförändringar krävs en fördjupad kunskap om de processer som styr koldioxidens kretslopp, och en ökad förståelse om hur dessa processer eventuellt påverkas av klimatförändringar. Denna avhandling försöker uppnå detta genom att studera utbytet av koldioxid mellan vår jord och dess atmosfär med hjälp av matematiska modeller.

Så kallade *ekosystemmodeller* utnyttjar kunskapen om olika fysikaliska processer för att beskriva ekosystemet, inklusive hur vegetation och markanvändning påverkar koldioxidflödet. Genom att mäta koldioxidflödet på en eller flera platser kan vi, genom olika matematiska metoder, förbättra våra modeller så att koldioxidflödet från modellerna matchar det observerade flödet. Första delen av avhandlingen fokuserar på att hitta de parametrar, som ger störst likhet mellan det modellerade och obsvererade flödet. Istället för att fokusera på enskilda parametrar, använder vi robusta statistiska metoder som justerar en sannolikhetsmodell över alla möjliga parametrar. Genom att slumpvis dra olika parameter-kombinationer och betygsätta deras kvalitet, kan vår sannolikhetsmodell hitta de parametrar som ger bäst matchning mellan det modellerade och observerade flödet.

Information om koldioxidytflödet kan också fås genom att titta på variationer av atmosfäriska koncentrationer av koldioxid i tid och rum, genom så kallad atmosfärisk invers modellering. En ökning eller minskning av den lokala koldioxidkoncentrationen i atmosfären, har sitt ursprung i ytflödet, d.v.s. från upptag och utsläpp av koldioxid genom olika fysikaliska processer vid jordytan. Genom att använda transportmodeller som approximerar hur luftmassor har färdats från jordytan, genom atmosfären, till de enskilda mätstationerna, kan vi försöka hitta ett ytflöde som ger en bra överensstämmelse med de observerade koncentrationerna. De okända parametrarna består här av ett stort antal komponenter som beskriver det okända ytflödet, ofta med fler komponenter än observerade koncentrationer. Detta innebär att flera olika representationer av ytflödet kan ge likvärdiga approximationer av data. Därför krävs ytterligare restriktioner på ytflödets struktur, i form av t.ex. en lämplig modell. Denna avhandling fokuserar på nya sätt att definiera dessa modeller, vilka ger fördelaktiga beräkningsegenskaper och möjliggör bättre approximationer till det sanna ytflödet.

Introduction

1 Estimating CO_2 surface fluxes

The increase in atmospheric greenhouse gases (GHG), such as carbon dioxide (CO_2), and their link to increasing temperature and other climate impacts, has motivated numerous research studies on the different components of the CO_2 cycle. The terrestrial biosphere is of major interest since the terrestrial productivity has a strong connection with CO_2 concentration in the atmosphere. To limit the future increase in CO_2 concentrations, it is important to understand the underlying physical processes behind sources and sinks of CO_2, and how these processes interact with a varying climate.

There are mainly two types of modelling approaches when analysing the CO_2 surface flux, i.e. the exchange of carbon dioxide between land and atmosphere. Ecosystem or ”bottom-up” modelling is based on process understanding, for which knowledge of physical sub-processes and interactions between climate and ecosystem is inferred to predict future CO_2 surface flux. These models simulate CO_2 exchange for either undisturbed ecosystems (natural CO_2 exchange), or for human-influenced ecosystems by including effects from agriculture (crops), forestry, and deforestation. Forcing the ecosystem models with different climate scenarios, bottom-up modelling can be used for predicting future source and sinks of CO_2.

The other important tool for analysis of surface flux is atmospheric inversion modelling, also known as ”top-down” modelling. Here, the idea is to reconstruct the historical surface fluxes[1] by using observations of atmospheric concentrations together with a transport model, which quantifies the sensitivity of observations to the surface fluxes. The (linear) transport models are based on numerical meteorological models and give information on how winds have transported the source fluxes to the different observational sites. Due to the under-determined nature

[1] In atmospheric inverse modelling, the flux is often referred in plural form as CO_2 surface fluxes. The flux has traditionally been restricted to a grid; thereof the plural form. Even when a continuous representation of the flux exists the transport models still have a finite resolution, in which case surface fluxes typically refer to the flux resolved on the transport grid.

of atmospheric inverse modelling, fluxes are often informed by assigning structure to the fluxes using a Bayesian formalism. The top-down method can help to identify locations of emissions not accounted for in the ecosystem process models, but can also locate (unreported) anthropogenic emission sources. Moreover, ecosystem modelling and atmospheric modelling can easily be combined by using the output from ecosystem models as input, or prior flux belief, in the atmospheric inversion. Due to the possibility of quantifying carbon sources and sinks, inverse modelling has the potential for being used in decision making, as a tool for understanding and reducing emissions of CO_2 to the atmosphere.

1.1 Bottom-up modelling

Process-based dynamical ecosystem models (Cox, 2001, Knorr, 2000, Sitch et al., 2003, e.g.) are based on mathematical descriptions of ecological systems and are often highly complex due to the many interacting components involved. These models include detailed descriptions of processes that govern the carbon cycle such as: the creation of carbon through photosynthesis and solar radiation, the allocation of carbon to plant growth, release of carbon to the atmosphere via plant respiration and decomposition, etc. Each process is described by its own set of model parameters, which are fixed in time. Dynamical state variables of the system, such as carbon- and water balance, and vegetation structure and composition, are used to describe the current state of the system.

Today's ecosystem models have large sources of errors (Heimann and Körner, 1996). One source of errors is due to uncertainties in the model parameters. The calibration of parameters are usually performed on small-scale field experiments, and parameters may need to be tuned when used in *Dynamical Global Vegetation Models* (DGVMs) (Cramer et al., 1999). This can be done by comparing flux output from DGVMs against flux tower observations. The problem of tuning the parameters is in literature commonly called *parameter estimation*. Due to the highly non-linear relation between parameters and model output, the problem of finding the (statistically) optimal parameters is complex. Moreover, the lack of gradient information restricts the number of optimisation tools that can be used for the problem. The first part of this thesis, Part I (Paper A and B), deals with parameter estimation in ecosystem models. The main focus is on parameter estimation for the LPJ-GUESS vegetation model, introduced in the following section, using CO_2 flux tower observations.

1.1.1 LPJ-GUESS

The parameter estimation in this thesis is applied to the LPJ-GUESS (Lund-Potsdam-Jena General EcoSystem Simulator, Smith et al., 2014), a DGVM originally developed at Lund University in collaboration with the Postdam Institute for Climate Impact Research, and the Max-Planck Institute. It is an advanced ecosystem model driven by regional climate conditions and atmospheric CO_2 concentrations, with vegetation dynamics resulting from competition among plants for light, space, and nutrients. A simplified schedule of LPJ-GUESS is shown in Figure 1.

Each grid cell in the LPJ-GUESS model has its own set of climate input. The grid cells include a number of replicate patches, that aim at describing the distribution of vegetation stands in that grid cell. The patches include plants in different stages of development, that are exposed to different disturbances. To limit the possible number of components, plants of similar type are grouped into plant functional types (PFT:s), which are treated as a unit within the same cohort (age group) and patch. In daily time steps, the LPJ-GUESS model includes processes such as soil hydrology, photosynthesis, plant respiration, phenology, and microbial decomposition. At annual time scale, that year's net production of carbon is allocated to leaves, fine roots and stem wood, etc; with the exact allocation and the resulting growth in biomass depending on PFT class (Smith et al., 2014).

The net release/uptake of CO_2 flux from land to the atmosphere is an output from DVGMs. By matching the output of CO_2 flux from the model with observations of CO_2 flux, the aim is to fine-tune the parameters in LPJ-GUESS that are sensitive to the CO_2 flux.

1.1.2 Data – Fluxes of CO_2

In the first part of the thesis, Part I, the main focus is on applying new methods for parameter estimation, as well as developing current methods. The ability to recover the global solution, i.e. the true parameters, is tested by introducing a simulation (or twin) experiment. The model parameters are optimised using pseudo-observations of CO_2 flux simulated from the LPJ-GUESS model, assuming a perfectly-known process-model, and no observational noise. Thus, the experiment tests the capability of reconstructing model parameters under optimal conditions. In Paper A, we also investigate the ability to constrain the LPJ-GUESS model parameters using daily CO_2 flux observations from a single boreal site in

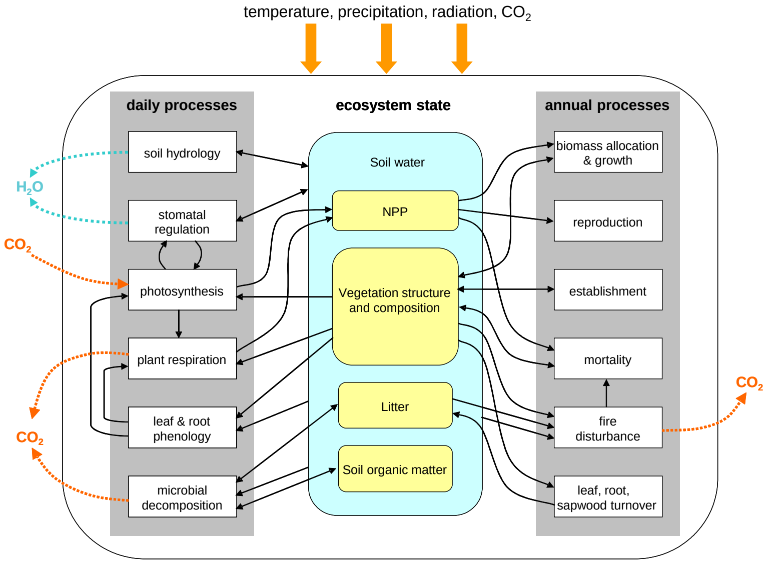


Figure 1: Schedule of process-dynamics in the vegetation model LPJ-GUESS.

Northern Sweden, during the time period 1997-2003. Currently, several flux towers in Europe are under construction. These will be part of the Integrated Carbon Observation System (ICOS); a European research infrastructure. In the future, the aim is to estimate model parameters using all available flux towers in ICOS.

1.2 Top-down modelling

Atmospheric inversion modelling aims to use spatio-temporal observations of CO_2 concentration in combination with atmospheric transport, to reconstruct historical sinks and sources of CO_2 (Ciais et al., 1995, Gurney et al., 2002, Rayner et al., 1999).

New and larger sampling networks for measuring concentrations of atmospheric CO_2 enables more detailed reconstructions from atmospheric inverse modelling, and introduces a need for more efficient computational methods. In the second part of this thesis, Part II (Paper C and D), we introduce flux models based on Gaussian Markov Random Fields (GMRF). The precision matrices of

GMRFs, i.e. the inverse covariance matrices, have only a few non-zero elements. This sparsity promotes computational tools, which enable efficient and fast numerical operations. Another important difference to previous inversion methods is that our CO_2 fluxes are defined on a continuous spatial domain, whereas previous inversion methods have fluxes confined to a longitude-latitude grid. Thus, we limit aggregation errors arising from the assumption of a piecewise flat fluxes in each grid cell, and moreover, the resulting spatial covariance model is derived on a continuous spatial domain.

A general problem for inverse methods is the dense observational matrix that arises from the flux integration, yielding expensive computations when using standard inference tools. In the last paper, Paper D, the maximum-likelihood estimation in Paper C, is replaced by a Markov Chain Monte Carlo (MCMC) method, which avoids the most expensive calculations, improving computational efficiency.

1.2.1 Data – Concentrations of CO_2

The raw data in the atmospheric inversion consists of CO_2 surface flask measurements at a global surface network, with samples collected and analysed by several institutions (e.g. NOAA). Weekly concentrations are averaged to monthly mean CO_2 concentration when processed according to the procedure described by Rödenbeck (2005) for the Jena Carboscope. The spatial distribution of measurements is shown in Figure 2, with several marine sites in the Pacific Ocean, and a high number of measurements in North America, and Western Europe. Few stations are located in sparsely populated or underdeveloped regions such as Siberia, tropical South America, tropical Africa, and the Southern Ocean, where the natural fluxes are assumed to be high.

1.2.2 Atmospheric transport

The link between fluxes and observations are obtained through an atmospheric transport model. Because CO_2 is an inert gas, i.e., a gas without chemical interaction to the surrounding atmosphere, CO_2 concentrations are tractable based solely on knowledge regarding transport, sinks, and sources. The atmospheric transport of CO_2 is governed by the past meteorological state. Mathematically, atmospheric transport models are obtained by numerically solving the mass continuity equations, based on the state of the atmosphere, found from three-

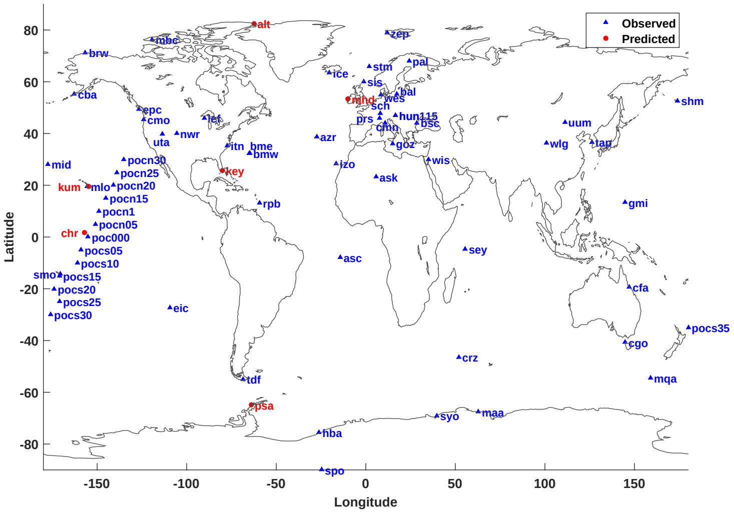


Figure 2: Network of global measurement stations. The observations locations marked with blue triangles are used to estimate fluxes, whereas observation locations indicated with red circles is used for validating the different model performances (Fig. 2 in Paper C).

dimensional weather forecast models - or atmospheric general circulation models running in climate mode (Heimann and Körner, 2003). Large studies on the effect of transport errors on CO_2 flux inversions were performed in the Transcom 3 intercomparison project (Baker et al., 2006, Gurney et al., 2003, 2004). In our inversion system, we use the global atmospheric tracer model TM3 (Heimann and Körner, 2003). In similarity with other grid-scale inversions (high resolution inversions), the sensitivity is defined on a regular longitude-latitude grid. Here with 48 boxes in latitude and 72 boxes in longitude.

1.3 Model framework

1.3.1 Data assimilation

In Geoscience, parameter estimation and atmospheric inverse modelling is often treated under a common framework; referred to as data assimilation (Rayner et al., 2018, Wang et al., 2009). The goal of data assimilation is to integrate (statistical) methods with observations to obtain optimal estimates of unknown states, and/or

Table 1: The different components in data assimilation when applied on parameter estimation v.v. atmospheric inversion.

	Parameter estimation	**Atmosperic inversion**
External forcing (u)	Climate and atmospheric conditions	-
Model (m)	Non-linear DGVM	Linear transport model
Target (x)	DGVM model parameters	CO_2 surface fluxes
Observations (y)	CO_2 surface fluxes	CO_2 atmospheric concentrations

parameters; here referred to as the *target* variables. Data assimilation essentially has up to four components: (1) External forcings, (2) A model that relates the state, model parameter and external input to observations, (3) Observations, and (4) an optimisation technique (Wang et al., 2009). For the parameter estimation in part I, the model is a non-linear DGVM forced by climate and atmospheric conditions, the target is the DGVM model parameters, and observations consist of CO_2 surface fluxes. Whereas in atmospheric inversion the model is the linear atmospheric transport, the target is the CO_2 surface fluxes and the observations consists of monthly CO_2 concentrations (see Table 1 for a comparison).

A common reference for relating target variables, $\mathbf{x}$ ($n_x \times 1$), and external forcing, $\mathbf{u}$, to observations, $\mathbf{y}$ ($n_y \times 1$), is here referred to as the model, $\mathbf{m}(\mathbf{x}, \mathbf{u})$. Data assimilation aims at finding target variables such that the model output gives a good match to the observations, i.e., $\mathbf{m}(\mathbf{x}, \mathbf{u}) \approx \mathbf{y}$. This is in general handled by introducing a cost function, J, to be minimized. Typically, the cost function is based on an assumption of additive errors using an observation model

$$\mathbf{y}|\mathbf{x}, \mathbf{u} = \mathbf{m}(\mathbf{x}, \mathbf{u}) + \boldsymbol{\varepsilon}, \tag{1}$$

where the observation error, $\boldsymbol{\varepsilon}$, includes both measurement noise and model noise arising from an imperfect model. Since the external forcing, $\mathbf{u}$, is usually assumed known (apart from in prediction studies), the subscript $\mathbf{u}$ is from now on suppressed. The cost function is typically formed by minimizing the sum of the squared errors:

$$\mathrm{J}(\mathbf{x}) = ||\mathbf{y} - \mathbf{m}(\mathbf{x})||_2^2 \tag{2}$$

where $||\cdot||$ is the Euclidean norm, corresponding to an assumption of $\boldsymbol{\varepsilon}$ being iid Gaussian errors. If different weights are assigned to different observation, the cost function extends to

$$\mathrm{J}(\mathbf{x}) = \sum_{i=1}^{n_y} \frac{\left(\mathbf{y}_i - \mathbf{m}_i(\mathbf{x})\right)^2}{\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon},ii}} \tag{3}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$ is a diagonal matrix, in general assumed to be known. If we allow for dependent errors, the matrix $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$ represents the error covariance matrix, $\mathrm{Cov}(\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon})$, and $\mathrm{J}(\mathbf{x}) = (\mathbf{y} - \mathbf{m})^{\mathsf{T}}\boldsymbol{\Sigma}_{\varepsilon}(\mathbf{y} - \mathbf{m})$.

In Part I of the thesis, the model $\mathbf{m}(\mathbf{x})$ is highly non-linear, and therefore minimizing J becomes a difficult problem. Since no explicit solution exists, the optimisation of the cost function is based on iterative methods.

In Part 2, the model $\mathbf{m}(\mathbf{x})$ is linear, but the number of unknowns are typically much larger than the number of knowns, i.e., $n_x \gg n_y$. This results in an ill-conditioned problem, which requires additional information on the target variables.

Historically, different kinds of regularizations have been attempted to make the inversion system in Part 2, solvable. Shrinkage estimators (Krakauer et al., 2004) penalize the overall size of the target variables, whereas smoothing operators penalizes rapid variations in the target variables (Enting, 1987, McIntosh and Veronis, 1993). The most common form of shrinkage regularization is Tikhonov regularization (Tikhonov and Arsenin, 1977), which adds a penalization term $||\mathbf{Tx}||_2^2$ to the cost function (3), where $\mathbf{T} = \sigma^2 \cdot \mathbf{I}_{\mathbf{n_x}}$ is a scaled identity matrix of size n_x. A popular smoothing regularization is the smoothing spline (Wahba, 1990), which assumes target variables to be dependent. The smoothing spline can be directly linked to Gaussian random fields, through its definition based on Stochastic Partial Differential Equations (SPDEs), as described in Lindgren et al. (2011), Nychka (2000). Thus, it can be shown that the Wahba splines correspond to a Gaussian target distribution with Matérn covariance of infinite range (Kimeldorf and Wahba, 1970, Nychka, 2000, Wahba, 1990).

Current inversion methods replace the standard regularization tools by assigning a (prior) distribution for the (unknown) target variables. This is commonly referred to as a Bayesian formalism since we assign a density to the unobserved or ”latent” variables/parameters. Introducing a prior on the target variables can be seen as a way of combining shrinkage and/or smoothing regularization. A Gaussian prior yields a penalizing term that shrinks the solution to a prior mean, $\boldsymbol{\mu}_{\mathbf{x}}$, with the smoothness of the solution related to the prior covariance matrix, $\boldsymbol{\Sigma}_{\mathbf{x}} = \mathrm{Cov}(\mathbf{x}, \mathbf{x})$.

By adding a prior distribution to the target we obtain a modified cost function, related to the posterior distribution of the target variables, $p(\mathbf{x}|\mathbf{y})$, introduced in the following section. Assuming Gaussian observation errors and a Gaussian prior, the cost function takes the form:

$$\mathrm{J}(\mathbf{x}) = (\mathbf{y} - \mathbf{m}(\mathbf{x}))^\intercal \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1}(\mathbf{y} - \mathbf{m}(\mathbf{x})) + (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}}). \tag{4}$$

based on the relation, $\mathrm{J}(\mathbf{x}) = -2\log p(\mathbf{x}|\mathbf{y})$. Thus, minimizing the cost function (4) yields the maximum of the posterior density.

1.3.2 Hierarchical Modelling (HM)

From a statistical view point data assimilation, with a prior distribution on the target, can be translated into hierarchical modelling (HM) (Carlin et al., 2014, Wikle et al., 1998) consisting of two or three layers:

$$\begin{aligned} \text{Data model: } & p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) \\ \text{Process/Prior model: } & p(\mathbf{x}|\boldsymbol{\theta}) \\ \text{Parameter/Prior model: } & p(\boldsymbol{\theta}) \end{aligned}$$

where $\mathbf{y}$ are the observed data, $\mathbf{x}$ is the latent field (or target variables), and $\boldsymbol{\theta}$ are the model parameters. The posterior density of the target can be derived based on Bayes theorem (Bayes, 1763). In data assimilation, the model parameters, $\boldsymbol{\theta}$, are often assumed to be fixed and known. In this case, the posterior density of the latent field, $p(\mathbf{x}|\mathbf{y})$, is given by :

$$p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}) = \frac{p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})p(\mathbf{x}|\boldsymbol{\theta})}{p(\mathbf{y})} \propto p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})p(\mathbf{x}|\boldsymbol{\theta}). \tag{5}$$

where $p(\mathbf{y})$ is the probability distribution of the observations.

If model parameters are assumed to be fixed, but unknown, this is called Empirical Hierarchical Modelling; or empirical-Bayesian modelling (Cressie and Wikle, 2015). Then, $p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$ is replaced by $p(\mathbf{x}|\mathbf{y}, \hat{\boldsymbol{\theta}})$ in Equation (5), where $\hat{\boldsymbol{\theta}}$ is an estimator of $\boldsymbol{\theta}$, typically obtained from maximum likelihood (ML) or maximum aposteriori probability (MAP) estimation.

In the third case, a probability distribution is assigned to the model parameters, $\boldsymbol{\theta}$, resulting in a Bayesian Hierarchical Model (BHM) (Berliner, 1996). The joint posterior density of the latent field and model parameters then becomes

$$p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}) \propto \frac{p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})}. \tag{6}$$

The posterior of the parameters can be obtained by marginalizing over the latent field:

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\boldsymbol{\theta}) \int p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta}) d\mathbf{x}, \tag{7}$$

and for the latent field, the posterior is found from:

$$p(\mathbf{x}|\mathbf{y}) \propto \int p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}. \tag{8}$$

Hence, the cost function (4) is related to the posterior (5), given the assumption of fixed (possibly unknown) model parameters, $\boldsymbol{\theta}$.

2 Part I: Parameter estimation

The goal of parameter estimation for DGVMs is to find the most likely model parameters, $\mathbf{x}^* \in \mathcal{X} \subseteq \mathbb{R}^d$, based on statistical assumptions and observations. If prior information on model parameters is included, the data assimilation adopts a Bayesian framework and the cost function (4) will be proportional to the negative log-posterior (5). The statistically optimal parameters are found by minimizing the cost function (maximizing the posterior):

$$\mathbf{x}^* = \underset{\mathbf{x} \in \mathcal{X}}{\arg\min}\, \mathrm{J}(\mathbf{x}). \tag{9}$$

Even though the data assimilation problem (1) is typically not ill-conditioned for parameter estimation, prior information on the parameters based on process knowledge can still be helpful to constrain the parameters to reasonable values. Due to the non-linear model, $\mathbf{m}(\mathbf{x})$ (see Section 1.3.1), the posterior density is very complex, and therefore iterative methods are required to optimise the cost function.

The posterior uncertainty, $\mathsf{V}(\mathbf{x}|\mathbf{y})$, is often approximated using the Hessian matrix, i.e. by the second-order partial derivative of the posterior calculated, possibly using finite differences, at the mode:

$$\mathsf{V}(\mathbf{x}|\mathbf{y}) \approx \left[\frac{\partial^2}{\partial \mathbf{x}^2} \log p(\mathbf{x}|\mathbf{y}) \bigg|_{\mathbf{x}=\mathbf{x}^*} \right]^{-1}. \tag{10}$$

In some studies on parameter estimation (e.g. Ziehn et al. 2012), the goal is to estimate the entire posterior distribution, $p(\mathbf{x}|\mathbf{y})$. This can be obtained by e.g. Markov Chain Monte Carlo methods (introduced in Part II, Section 3.4.2). However, the parameter estimation in this thesis focuses on using and developing stochastic optimisation methods for finding the optimal parameter values. Therefore, the discussion on parameter estimation is focused on optimisation methods, and on stochastic theory for improving a certain class of random search methods.

2.1 optimisation

Solving the optimisation problem (9), is a complex problem due to the very unstructured and multi-modal dynamic vegetation model used as $\mathbf{m}(\mathbf{x})$. In this section, some standard optimisation methods are discussed, as well as some modern stochastic optimisation tools.

2.1.1 Gradient-based

Gradient-based methods are iterative methods that use the gradient of the function at the current point, to calculate a search direction. The common update form for gradient-based algorithms is

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \alpha_t \mathbf{h}_t, \tag{11}$$

where t is the iteration index, $\alpha_t > 0$ is a scaler at iteration t, helpful for controlling step length and convergence, and $\mathbf{h}_t$ is the step direction at iteration t. In order for $\mathbf{h}$ to be a descent direction, i.e., a direction that guarantees that J can be reduced, the step direction must satisfy $\mathbf{h}^\mathsf{T}\mathrm{J}'(\mathbf{x}_t) < 0$. By defining the step direction on the form:

$$\mathbf{h} = -B_t^{-1}\mathrm{J}'(\mathbf{x}_t), \tag{12}$$

a descent direction is obtained whenever B_t is a positive definite matrix. Letting $B_t^{-1} = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix, one obtains the steepest descent method. The standard Newton update is obtained with $B_t = \mathrm{J}''(\mathbf{x}_t)$. If J is non-linear, as is typically the case in applications to DGVMs, the Hessian $\mathrm{J}''(\mathbf{x}_t)$ is not necessarily positive definite and therefore $\mathbf{h}$ does not have to be a descent direction. However, if the optimisation is started close enough to the global minima of a ”nice” function, then $\mathrm{J}''(\mathbf{x}_t)$ will be positive definite, and the algorithm converges quadratically to the solution (Fletcher, 1987). In quasi-Newton methods, B_t is a positive definite approximation to the Hessian which is updated at each iteration. Some examples of quasi-Newton methods are *Broyden-Fletcher-Goldfarb-Shanno* (BFGS) (Byrd et al., 1995) method and *symmetric-rank-1* (SR1) (Byrd et al., 1996). Quasi-Newton methods are often implemented to avoid the computationally expensive evaluation and storage of large Hessian matrices, but also to guarantee a decent direction.

2.1.2 Gradient-free

Although there are many gradient-based algorithms which provide good convergence properties under mild condition, gradient-based methods are in general informed locally, which is inappropriate when working with highly multi-modal objective functions. Moreover, in the application on parameter estimation for the LPJ-GUESS model, derivative information is hard; or even impossible, to obtain. Thus, traditional gradient-based methods would require numerical approximations of the gradient. Moreover, some parameters of the DVGM model can abruptly change the dynamics of certain sub-processes, leading to discontinuities in the objective function.

Gradient-free methods work on global domains, with several cost functions evaluated at each step in the algorithm, and are therefore less sensitive to local minima. In the following, an overview of a few gradient-free methods, and families of stochastic gradient-free methods is provided:

Nelder-Mead One popular gradient-free method is the Nelder-Mead algorithm (Nelder and Mead, 1965), which is based on an iterative update of the vertices in a simplex; a polytype of $n + 1$ vertices that define a search domain in dimension n. This approach is effective for smaller dimensions, but for larger dimensions and with a non-structured objective functions, the risk of converging to a local solution is high (Lagarias et al., 1998).

Random Search methods Random search methods are a collection of optimisation methods which are based on a stochastic mechanism to generate candidate solutions, i.e., populations of possible solutions to the optimisation problem (Zlochin et al., 2004). These methods are attractive for unstructured objective functions, since their stochastic nature, provides opportunities to escape local minima. Depending on the mechanism of generating new candidate solutions, random search methods are divided into instance-based or model-based algorithms.

In **Instance-based** algorithms, new candidate solutions depend explicitly on previous candidate solutions. Genetic algorithms (Sastry et al., 2005) are a class of instance-based methods that have been used for parameter estimation in vegetation modelling (Barrett, 2002, Roxburgh et al., 2006). Another example of instance-based methods is simulated annealing (Kirkpatrick et al., 1983).

Model-based methods are a newer generation of random search methods (Hu et al., 2012b, Pelikan et al., 2002). In these algorithms, the candidate solutions are generated via an intermediate probabilistic model, and the algorithm describes how to propagate this probabilistic model using candidate solutions from previous steps. In general, many model-based (and instant-based) methods are of a heuristic nature and lacks support for convergence. However, quite recently, Zhou and Hu (2014) were able to connect the framework for some model-based methods to stochastic approximation (SA), which provided better insight into the asymptotic behaviour and convergence properties of these algorithms.

For the parametrization problem of DVGMs, we will apply and extend a method called Gradient Adaptive Stochastic Search (GASS), which belongs to the class of model-based methods.

2.2 GASS

The idea of GASS (Hu et al., 2012a, Zhou and Hu, 2014) and other model-based methods is to form a sequence of distributions, $\{f(\mathbf{x}, \boldsymbol{\theta}_k)\}$, where the limiting distribution $f^* = \lim_{k\to\infty} f(\mathbf{x}, \boldsymbol{\theta}_k)$, assigns most of its probability mass to the set of optimal solutions. The procedure is usually divided into two main steps:

1. Sample candidate solutions, $\mathbf{x}_i$, $i = 1, \dots N$, from the current distribution, $f(\mathbf{x}, \boldsymbol{\theta}_k)$.

2. Use the candidate solutions to update the model parameters $\boldsymbol{\theta}_k$ to new model parameters $\boldsymbol{\theta}_{k+1}$, such that $f(\mathbf{x}, \boldsymbol{\theta}_{k+1})$ is biased towards the candidate solutions of high quality.

Model-based methods differ mainly in the specification of the distributions, $\{f(\mathbf{x}, \boldsymbol{\theta}_k)\}$, and in the update rule. Moreover, the construction of the distribution $f(\mathbf{x}, \boldsymbol{\theta}_k)$, and/or the update rule, depends on the objective function which is not always available in explicit form. In GASS, the original optimisation problem over the solution space $\mathcal{X} \in \mathbb{R}^n$:

$$\mathbf{x}^* = \arg\max_{\mathbf{x} \in \mathcal{X}} H(\mathbf{x}), \tag{13}$$

is exchanged with optimisation over the parameter space $\boldsymbol{\Theta}$:

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \int S_{\boldsymbol{\theta}'}(H(\mathbf{x})) f(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{x}, \tag{14}$$

where $\boldsymbol{\theta}' \in \boldsymbol{\Theta}$ is a fixed but arbitrary parameter, and S is a monotone shape function for which $S(H(\mathbf{x}))$ is non-negative. The maximum of (14) is obtained when the distribution, $f(\mathbf{x}, \boldsymbol{\theta})$, has all its probability mass concentrated at the global solution, $\mathbf{x}^*$. By restricting the distribution to the exponential family, given on the form:

$$f(\mathbf{x}; \boldsymbol{\theta}) = h(\mathbf{x}) \exp(\boldsymbol{\theta}^\intercal T(\mathbf{x}) - \phi(\boldsymbol{\theta})), \tag{15}$$

where $T(\mathbf{x})$ is the sufficient statistics and $\boldsymbol{\theta}$ is the vector of natural parameters, the model parameters are updated efficiently based on a standard Newton step, e.g. (11). At step k, the gradient and Hessian of (14) given $\boldsymbol{\theta}' = \boldsymbol{\theta}_k$, are approximated based on current samples. The gradient, Φ, is given by

$$\Phi = \mathsf{E}_{p(\cdot, \boldsymbol{\theta}_k)}[T(\mathbf{X})] - \mathsf{E}_{f(\cdot, \boldsymbol{\theta}_k)}[T(\mathbf{X})], \tag{16}$$

where the distribution, $p(\cdot, \boldsymbol{\theta})$, is defined as:

$$p(\mathbf{x}; \boldsymbol{\theta}) \triangleq \frac{S_{\boldsymbol{\theta}}(H(\mathbf{x})) f(\mathbf{x}; \boldsymbol{\theta})}{\int S_{\boldsymbol{\theta}}(H(\mathbf{x})) f(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{x}}. \tag{17}$$

The second term in (16) can be calculated explicitly, whereas the first term is estimated by Monte Carlo integration using the candidate solutions, $\{\mathbf{x}^i\}_{i=1}^N$, to perform importance sampling. The Hessian is approximated by:

$$-\left(\mathsf{V}_{f(\cdot, \boldsymbol{\theta}_k)}[T(\mathbf{x})] + \varepsilon \mathbf{I}\right), \tag{18}$$

where the variance term is estimated from current samples and the diagonal matrix $\varepsilon\mathbf{I}$ is added to ensure that the Hessian approximation can be inverted in the Newton update step.

2.3 Importance sampling

The unknown form of the objective function, $H(\mathbf{x})$, prevents the analytical calculation of the gradient in (16). If we would be able to draw independent candidate solutions directly from the distribution $p(\mathbf{x}, \boldsymbol{\theta})$, the standard Monte Carlo estimate of the gradient (16) is

$$\widehat{\Phi} = \frac{1}{N}\sum_{i=1}^{N} T(\mathbf{x}_i), \qquad \mathbf{x}_i \sim p(\mathbf{x}; \boldsymbol{\theta}_k), \tag{19}$$

where $\widehat{\Phi} \to \Phi$ when $N \to \infty$, by the law of large numbers. Moreover, the Monte Carlo estimation of Φ provides $\mathcal{O}(\sqrt{N})$ converges, independent of the dimension of $\mathbf{x}$. However, in the application to DGVMs, $H(\mathbf{x}) = -\mathrm{J}(\mathbf{x})$ (see Eqn. 4) is a complex function, and even though the exact form of $H(\mathbf{x})$ was known, the distribution $p(\mathbf{x}; \boldsymbol{\theta})$ would probably not allow for direct sampling.

An alternative is to estimate the gradient (16) based on Importance Sampling (IS). This method was originally introduced as a way for reducing the variance of Monte Carlo estimates. Introducing an arbitrary instrumental distribution, $g(\mathbf{x})$, we have that:

$$\Phi = \int T(\mathbf{x}) p(\mathbf{x}; \boldsymbol{\theta}_k) d\mathbf{x} = \int T(\mathbf{x}) \frac{p(\mathbf{x}; \boldsymbol{\theta}_k)}{g(\mathbf{x})} g(\mathbf{x}) d\mathbf{x}, \tag{20}$$

if $g(\mathbf{x}) > 0$ whenever $T(\mathbf{x})p(\mathbf{x}; \boldsymbol{\theta}_k) \neq 0$. Thus, one can sample from $g(\mathbf{x})$, and introduce weights, $\omega(\mathbf{x}) = \frac{p(\mathbf{x};\boldsymbol{\theta})}{g(\mathbf{x})}$, resulting in:

$$\widehat{\Phi} = \frac{1}{N}\sum_{i=1}^{N} \omega(\mathbf{x}_i) T(\mathbf{x}_i) \qquad \mathbf{x}_i \sim g(\mathbf{x}). \tag{21}$$

In the GASS method, the gradient is estimated using importance sampling with the instrumental density being the current probabilistic model, i.e., $g(\mathbf{x}) = f(\mathbf{x}; \boldsymbol{\theta}_k)$, where f belongs to the exponential family. The corresponding weights satisfy the proportionality:

$$\omega_i \propto \frac{p(\mathbf{x}_i^k; \boldsymbol{\theta}_k)}{f(\mathbf{x}_i^k, \boldsymbol{\theta}_k)} \propto S_{\boldsymbol{\theta}_k}(H(\mathbf{x})). \tag{22}$$

For the unstructured objective functions occuring when estimating parameters in DGVMs, a more efficient instrumental distribution for estimating the gradient

(16), is one that better captures the multi-modal shape of the objective function. In Paper B of this thesis, our aim is to improve the efficiency of the importance sampling, by introducing an instrumental distribution, given on a multi-modal form:

$$g(\mathbf{x};\boldsymbol{\theta},\{\pi_j,\psi_j\}_{j=1}^{N_\nu}) = (1-\delta)f(\mathbf{x},\boldsymbol{\theta}) + \delta\sum_{j=1}^{N_\nu}\pi_j\nu_j(\mathbf{x};\psi_j). \tag{23}$$

Here the second term is a Gaussian mixture model fitted to samples from previous iterations, with components $\nu_j(\mathbf{x};\psi_j)$, and weights $\{\pi_j\}$ satisfying $\sum_j^{N_\nu}\pi_j = 1$.

2.4 Expectation Maximization method

The unknown parameters of the mixture components in (23) can be estimated using the well-known expectation maximization (EM) algorithm (Dempster et al., 1977, McLachlan and Krishnan, 2007). The EM algorithm is often used when the likelihood or posterior is easier to obtain by including some kind of ”missing” or ”hidden” data. It does not have to be missing data in the conventional sense, such as missing observations in a time series, but could also be more hypothetical, like un-observable variables, e.g. which mixture component each observation belongs to.

The EM algorithm is an iterated two-step procedure, where the **E-step** requires the calculation of the complete/full likelihood, conditional on the observations and current iterate, or guess, of the unknown parameters to be estimated. The **M-step** maximizes the Q-function derived in the E-step based on some optimisation procedure. These two steps are repeated until the (incomplete) likelihood converges.

Example For the above mixture model, (23), assume that we have observations $\vec{\mathbf{x}} = \{\mathbf{x}_i\}_{i=1}^n$, and ”missing” data being the class/mixture-belongings, z, taking some value in the integer set $\{1,\ldots,N_\nu\}$. Let the aim be to estimate the unknown parameters of the mixture model: $\boldsymbol{\Psi} = \{\pi_j,\psi_j\}_{j=1}^{N_\nu}$. The logarithm of the full likelihood takes the form

$$\log p(\vec{\mathbf{x}},z|\boldsymbol{\Psi}) = \sum_{j=1}^{N_\nu}\Big[\log(\pi_j) + \log\nu_j(\mathbf{x}_i;\psi_j)\Big]\,\mathbb{I}(z_i = j),$$

where $\mathbb{I}(z_i = j)$ is an indicator function, taking value one if observation i belongs to class j, and zero otherwise. The Q-function at iteration t of the EM algorithm is given by

$$\begin{aligned} \mathsf{Q}(\boldsymbol{\Psi}, \boldsymbol{\Psi}^t) &= \mathsf{E}_{\boldsymbol{\Psi}^t}\left[\log p(\vec{\mathbf{x}}, z|\boldsymbol{\Psi})\right] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{N_\nu}\left[\log(\pi_j) + \log \nu_j(\mathbf{x}_i; \phi_j)\right] \mathsf{P}(z_i = j|\mathbf{x}_i, \boldsymbol{\Psi}^t), \end{aligned} \tag{24}$$

where $\mathsf{P}(z_i = j|\mathbf{x}_i, \boldsymbol{\Psi}^t)$ is the probability of z_i taking value j, given $\mathbf{x}_i$ and $\boldsymbol{\Psi}^t$, i.e. the probability of observation i belonging to class j. The M-step of the EM algorithm optimises the Q-function with respect to the unknown parameters:

$$\boldsymbol{\Psi}^{t+1} = \arg\max_{\boldsymbol{\Psi}} \mathsf{Q}(\boldsymbol{\Psi}, \boldsymbol{\Psi}^t). \tag{25}$$

The EM steps, (24) and (25), are repeated until the likelihood $L(\boldsymbol{\Psi}) = p(\mathbf{x}|\boldsymbol{\Psi})$ has converged, e.g. until $L(\boldsymbol{\Psi}^{t+1}) - L(\boldsymbol{\Psi}^t)$ becomes sufficiently small.

Assigning a prior distribution to any component in $\boldsymbol{\Psi}$, one would instead maximize the log posterior

$$\log p(\boldsymbol{\Psi}|\mathbf{x}) = \log p(\mathbf{x}|\boldsymbol{\Psi}) + \log p(\boldsymbol{\Psi}). \tag{26}$$

Using the same Q-function as in (24), the new M-step for the posterior becomes

$$\boldsymbol{\Psi}^{t+1} = \arg\max_{\boldsymbol{\Psi}} \left[\mathsf{Q}(\boldsymbol{\Psi}, \boldsymbol{\Psi}^t) + \log p(\boldsymbol{\Psi})\right].$$

The introduction of a "nice" Bayesian prior on the unknown parameters, $\boldsymbol{\Psi}$, will almost always result in a more concave objective function, aiding the optimisation.

3 Part II: Atmospheric inverse modelling

This part of the introduction describes the spatial-temporal statistics, used in the atmospheric inversion.

Due to the ill-conditioned nature of the atmospheric inverse problem, the main focus in the research on atmospheric inverse modelling has been on either restricting the solution to an identifiable form, e.g. by the design of the target;

or by using some kind of ancillary information, such as regularization. Since the introduction of Bayesian methods to atmospheric inverse modelling (Enting et al., 1993, 1995), most focus has been on including appropriate prior knowledge, and on developing methods for estimating the resulting targets.

In atmospheric inverse modelling, the targets are usually discretized fluxes on a grid (Baker et al., 2006, Gurney et al., 2003, 2004, Law et al., 2003, e.g.), or weights in some basis function expansion (Dahlén et al., 2019, Rödenbeck et al., 2003). Target variables might also arise from some regression type models of the mean fluxes, as in (Michalak et al., 2004). In all studies, the target variables or parts of the target variables can be modelled using spatio-temporal random fields.

In what follows; the spatial processes are first introduced (Section 3.1), whereafter the model is expanded to spatio-temporal processes (Section 3.2). In Section 3.3 the computational benefits arriving from sparse matrices and matrices on Kronecker form are considered. The standard model set-up and inference is discussed in Section 3.4, followed by some more detailed information on inference methods.

3.1 Spatial statistics

A spatial process has no temporal structure, and might be an instantaneous state of a spatio-temporal process, or arise from some aggregation in time. Often, nearby process values tend to be more similar than those further apart, due to physical processes that interact in a spatio-temporal continuous domain. When aggregated over time, this typically results in higher spatial correlations between nearby values (Cressie and Wikle, 2015). Therefore, the introduction of spatial dependence into the process model can help predict process-values at unobserved locations.

Let us denote a continuous spatial random process by $x(\mathbf{s})$, $\mathbf{s} \in \mathcal{D}$, where $\mathcal{D} \subset \mathbb{R}^d$. Typically, the spatial dimension is either $d = 2$ (a surface) or $d = 3$ (a volume). A spatial process is defined based on it's multivariate distribution over a finite set of spatial locations, $\{\mathbf{s}_1, \ldots \mathbf{s}_n\} \in \mathcal{D}$, given by

$$F(x_1, \ldots, x_n; \mathbf{s}_1, \ldots, \mathbf{s}_n) = \mathsf{P}(x(\mathbf{s}_1) \leq x_1, \ldots x(\mathbf{s}_n) \leq x_n). \tag{27}$$

If the above distribution is invariant to spatial shifts, $\mathbf{h}$, i.e., if

$$F(x_1, \ldots, x_n; \mathbf{s}_1, \ldots, \mathbf{s}_n) = F(x_1, \ldots, x_n; \mathbf{s}_1 + \mathbf{h}, \ldots, \mathbf{s}_n + \mathbf{h}). \tag{28}$$

the process is called *strictly stationary*. A process for which the two first moments exist and are invariant to spatial shifts is called a *weakly* or *second-order stationary* process. Then, the process mean is constant, and the covariance depends only on a vector between two points, i.e.

$$\mathsf{E}[x(\mathbf{s})] = \mu \quad \forall \mathbf{s} \in \mathcal{D}$$

and

$$\mathrm{Cov}[x(\mathbf{s}), x(\mathbf{s}+\mathbf{h})] = \mathsf{C}(\mathbf{h}) \quad \forall \mathbf{s}, \mathbf{s}+\mathbf{h} \in \mathcal{D}$$

where $\mathsf{C}(\mathbf{h})$ is the covariance function. If the covariance only depends on the distance and not the direction of the vector separating the points, the covariance is *isotropic*, and the covariance is given by: $\mathrm{Cov}[x(\mathbf{s}), x(\mathbf{s}+\mathbf{h})] = \mathsf{C}(||\mathbf{h}||)$, where $||\cdot||$ denotes the Euclidean length (see Gelfand et al. (2010) for many more details).

The Matérn covariance function (Matérn, 1960) is a popular isotropic covariance function defined as

$$\mathsf{C}(||\mathbf{h}||) = \sigma^2 \frac{(\kappa||\mathbf{h}||)^\nu K_\nu(\kappa||\mathbf{h}||)}{2^{\nu-1}\Gamma(\nu)}, \tag{29}$$

where $||\cdot||$ is a distance on a metric space, K_ν is the modified Bessel function of the second kind, σ^2 is the marginal variance ($\mathsf{C}(0) = \sigma^2$), and $\nu > 0$ and $\kappa > 0$ are the smoothing and scaling parameters, respectively. The Matérn model is commonly used in geostatistics due to its flexible form (Stein, 2012), and the family incorporates common covariance families such as the exponential ($\nu = \frac{1}{2}$) and Gaussian ($\nu \to \infty$) (Guttorp and Gneiting, 2006).

3.1.1 Gaussian Fields

In this thesis, the stochastic processes are typically modelled by Gaussian fields, where the joint density of a random vector, $\mathbf{x} = \left[x(\mathbf{s}_1), \dots x(\mathbf{s}_n)\right]$, has a Multivariate Gaussian distribution:

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right), \tag{30}$$

with mean, $\boldsymbol{\mu} \in \mathbb{R}^n$, and covariance matrix, $\boldsymbol{\Sigma}_{ij} = \mathrm{Cov}(x(\mathbf{s}_i), x(\mathbf{s}_j)) \in \mathbb{R}^{n\times n}$. This is typically denoted as $\mathbf{x} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Note that the Gaussian process is defined by its two first (central) moments, $\mathsf{E}[\mathbf{x}]$, and $\mathsf{V}[\mathbf{x}]$. Therefore, a weakly stationary Gaussian process, is also strictly stationary.

3.1.2 Gaussian Markov Random Fields

Gaussian Markov Random Fields (GMRFs) (Lauritzen, 1996, Rue and Held, 2005) are Gaussian fields characterized by their sparse precision matrix; i.e. with many elements being zero in the inverse of the covariance matrix, here denoted $\mathbf{Q} = \boldsymbol{\Sigma}^{-1}$. This facilitates efficient matrix algebra using mathematical tools for sparse matrices (see e.g. Rue and Held 2005), and reduces the complexity of computing determinants and solving equation systems involving the precision matrix (see Section 3.3.1). This is particularly useful for ML inference, as demonstrated in Section 3.4.1, and in MCMC methods (see Section 3.4.2), where one needs to simulate from the random field.

A random vector $\mathbf{x} \sim N(\boldsymbol{\mu}, \mathbf{Q}^{-1})$, with elements, $\mathbf{x}_i = x(\mathbf{s}_i)$, is called a GMRF if the joint distribution of $\mathbf{x}$ satisfies $p(\mathbf{x}_i|\mathbf{x}_{-i}) = p(\mathbf{x}_i|\mathbf{x}_{\mathcal{N}_i}), \forall i$. Here, $\mathbf{x}_{-i}$ denotes the full vector $\mathbf{x}$ without element $\mathbf{x}_i$, and $\mathcal{N}_i$ is a subset of neighbourhood variables $\{\mathbf{x}_j\}_{j \in \mathcal{N}_i}$, which in some sense are close or connected to $\mathbf{x}_i$. It further holds that

$$\mathbf{x}_i \perp \mathbf{x}_j | \mathbf{x}_{-\{i,j\}} \Longleftrightarrow \mathbf{Q}_{ij} = 0 \Longleftrightarrow j \notin \mathcal{N}_i.$$

Thus, $\mathbf{Q}_{i,j}$ being zero implies that $\mathbf{x}_i$ and $\mathbf{x}_j$ are conditionally independent, and moreover, j is not a neighbour of i (Rue and Held, 2005).

3.1.3 Link between Gaussian Fields and Gaussian Markov Random Fields

The application of GMRFs has historically been restricted to models on gridded discrete domains (Besag, 1974, Besag et al., 1991, Rue and Tjelmeland, 2002). However, the link between Gaussian fields and GMRFs introduced in Lindgren et al. (2011), extends the application of GMRFs to continuous spatial domains. The foundation of the method builds on the research by Whittle (1954, 1963) showing that Gaussian fields with Matérn covariance are solutions to the stochastic differential equation:

$$(\kappa^2 - \Delta)^{\alpha/2} \tau x(\mathbf{s}) = \mathcal{W}(\mathbf{s}), \tag{31}$$

where $\Delta = \sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2}$ is the *Laplacian* operator, $\mathcal{W}(\mathbf{s})$ is Gaussian white noise, and τ is a scaling parameter. The parametric relation between the SPDE and Matérn covariance function in (29) is given by, $\alpha = \nu + d/2$, and τ relates to the marginal field variance, as

$$\mathsf{C}(\mathbf{0}) = \sigma^2 = \frac{\Gamma(\nu)}{\Gamma(\nu + d/2)(4\pi)^{d/2}\kappa^{2\nu}\tau^2}.$$

The idea in Lindgren et al. (2011) involved expressing the Gaussian field using a finite element representation:

$$x(\mathbf{s}) = \sum_{i=1}^{n} \phi_i(\mathbf{s})\omega_i, \tag{32}$$

with Gaussian distributed weights $\{\omega_i\}_{i=1}^{n}$ and piecewise linear basis functions $\{\phi_i(\mathbf{s})\}_{i=1}^{n}$, defined on a Delaunay triangulation (see Lindgren et al. 2011 for details). The link to GMRF is obtained by solving the distribution of weights such that the approximation (32) follows the stochastic weak formulation of the SPDE:

$$\begin{bmatrix} \langle(\kappa^2 - \Delta)^{\alpha/2}\tau x, \psi_1\rangle \\ \vdots \\ (\kappa^2 - \Delta)^{\alpha/2}\tau x, \psi_n\rangle \end{bmatrix} \overset{d}{=} \begin{bmatrix} \langle\mathcal{W}, \psi_1\rangle \\ \vdots \\ \langle\mathcal{W}, \psi_n\rangle \end{bmatrix} \tag{33}$$

where $\{\psi_j\}_{j=1}^{n}$ is a specific set of test functions, $\overset{d}{=}$ denotes equality in distribution, and $\langle\cdot,\cdot\rangle$ denotes the inner product defined as $\langle f, g\rangle = \int f(\mathbf{s})g(\mathbf{s})d\mathbf{s}$.

The common Galerkin solution is obtained by letting $\psi_i = \phi_i$, and $\alpha = 2$, for which element, j, on the left-hand side of (33), can be written:

$$\begin{aligned} \langle(\kappa^2 - \Delta)\tau x, \phi_j\rangle &= \sum_{i=1}^{n} \omega_i \, \langle(\kappa^2 - \Delta)\phi_i, \phi_j\rangle \, \tau \\ &= \sum_{i=1}^{n} \omega_i \, \left(\kappa^2\langle\phi_i, \phi_j\rangle - \langle\Delta\phi_i, \phi_j\rangle\right) \tau. \end{aligned}$$

By specifying matrices $\mathbf{C}$ and $\mathbf{G}$ with elements $\mathbf{C}_{ij} = \langle\phi_i, \phi_j\rangle$, and $\mathbf{G}_{ij} = \langle\Delta\phi_i, \phi_j\rangle$, and forming a matrix $\mathbf{K}_{\kappa^2} = (\kappa^2\mathbf{C} + \mathbf{G})$, and a vector $\boldsymbol{\omega} = \begin{bmatrix}\omega_1 & \dots & \omega_n\end{bmatrix}$, the left-hand side of (33) is given by $\tau\mathbf{K}_{\kappa^2}\boldsymbol{\omega}$.

The right hand side (33) is multivariate Gaussian, with mean zero and pairwise covariance, $\mathrm{Cov}(\langle\mathcal{W}, \phi_i\rangle, \langle\mathcal{W}, \phi_j\rangle) = \mathrm{Cov}(\langle\phi_i, \phi_j\rangle) = \mathbf{C}_{ij}$. Thus, (33) can be reformulated as:

$$\tau\mathbf{K}_{\kappa^2}\boldsymbol{\omega} \overset{d}{=} \mathsf{N}(\mathbf{0}, \mathbf{C}),$$

yielding weights, $\boldsymbol{\omega} \in \mathsf{N}(\mathbf{0}, \tau^{-1}\mathbf{K}_{\kappa^2}^{-1}\mathbf{C}\mathbf{K}_{\kappa^2}^{-T}\tau^{-1})$. Thus, the corresponding precision matrix is:

$$\mathbf{Q} = \tau\mathbf{K}_{\kappa^2}^{\mathsf{T}}\mathbf{C}^{-1}\mathbf{K}_{\kappa^2}\tau = \tau(\kappa^2\mathbf{C} + \mathbf{G})^{\mathsf{T}}\mathbf{C}^{-1}(\kappa^2\mathbf{C} + \mathbf{G})\tau. \tag{34}$$

Due to the specification of the basis functions, $\{\psi_j\}_{j=1}^{n}$, both $\mathbf{C}$ and $\mathbf{G}$ are sparse matrices, with non-zero values, obtained only where the basis functions overlap. However, $\mathbf{C}^{-1}$, is dense, but[2] can, with minor effects, be replaced by a diagonal matrix $\widetilde{\mathbf{C}}$ with elements $\widetilde{\mathbf{C}}_{ii} = \langle \phi_i, 1 \rangle$. For a stationary covariance both τ and κ are constants, but a non-stationary GMRF can easily be formed by introducing τ and κ as diagonal matrices. The SPDE approach introduced above also generalizes to manifolds, such as the sphere.

For other values of α, GMRFs are only obtained if $\alpha \in \mathbb{Z}^+$ (Rozanov, 1977), and have precision matrix, $\mathbf{Q}_\alpha = \tau \mathbf{Q}_{\alpha,\kappa^2} \tau$, with $\mathbf{Q}_{\alpha,\kappa^2}$ given by the recursion

$$\begin{cases} \mathbf{Q}_{1,\kappa^2} = \mathbf{K}_{\kappa^2} = (\kappa^2 \widetilde{\mathbf{C}} + \mathbf{G}) \\ \mathbf{Q}_{2,\kappa^2} = \mathbf{K}_{\kappa^2}^{\mathsf{T}} \widetilde{\mathbf{C}}^{-1} \mathbf{K}_{\kappa^2} \\ \mathbf{Q}_{\alpha,\kappa^2} = \mathbf{K}_{\kappa^2} \widetilde{\mathbf{C}}^{-1} \mathbf{Q}_{\alpha-2,\kappa} \widetilde{\mathbf{C}}^{-1} \mathbf{K}_{\kappa^2}, \quad \text{if } \alpha = 3, 4, \ldots \end{cases} \tag{35}$$

3.2 Spatio-temporal statistics

The extension from spatial to spatio-temporal random fields is obtained by adding a temporal component, t, yielding processes on the form

$$\{\mathbf{x}(\mathbf{s}, t) : (\mathbf{s}, t) \subseteq \mathbb{R}^d \times \mathbb{R}\}. \tag{36}$$

The difference from viewing the process (36) as a spatial process on $\mathbb{R}^{d+1}$, mainly arrives from physical interpretation of time moving in only one (forward) direction. Hence, a good model should account for this difference. Still, several technical results for spatial processes, related to covariance functions and inference, applies to spatio-temporal processes when viewed as spatial processes on $\mathbb{R}^{d+1}$.

For example, a spatio-temporal process on $\mathbb{R}^d \times \mathbb{R}$ is *second-order stationary* if it is *second-order stationary* on the Euclidean domain $\mathbb{R}^{d+1}$. The corresponding covariance is expressed through a *space-time covariance function*:

$$\mathrm{Cov}(\mathbf{x}(\mathbf{s}, t), \mathbf{x}(\mathbf{s} + \mathbf{h}, t + v)) = \mathsf{C}(\mathbf{h}, v) \tag{37}$$

where $\mathbf{h} \in \mathbb{R}^d$ and $v \in \mathbb{R}$. The margins, $\mathsf{C}(\cdot, 0)$ and $\mathsf{C}(0, \cdot)$ are purely spatial and temporal covariance functions, respectively.

In this thesis, we use *separable* space-time covariance functions expressed on the form

$$\mathsf{C}(\mathbf{h}, v) = \mathsf{C}_S(\mathbf{h}) \cdot \mathsf{C}_T(v). \tag{38}$$

[2] all $\mathbf{C}$-matrices

Thus, the spatial and temporal covariances are fully separated. The covariance matrix, $\mathbf{\Sigma}$, of a spatio-temporal vector with separable covariance can be written as a Kronecker product (Blangiardo and Cameletti, 2015) between the temporal covariance matrix, $\mathbf{\Sigma}_T$, and the spatial covariance matrix, $\mathbf{\Sigma}_S$, i.e.:

$$\mathbf{\Sigma} = \mathbf{\Sigma}_T \otimes \mathbf{\Sigma}_S.$$

Some computational benefits arising from the Kronecker structure are discussed in the next section.

3.3 Reductions in computational costs and memory requirements

Often, high dimensional targets arise when the unknown of interest lies in an infinite-dimensional space and is approximated using a finite-dimensional representation. This is the case for e.g. the latent field, $\mathbf{x}$, in the atmospheric inversion problem. For large-dimensional targets the computational time and storage often limit the applicability of standard inference methods.

Modelling a stochastic field (vector) using a GMRF typically leads to computational advantages and reduced memory requirements. Commonly, the computations involving the sparse precision matrix $\mathbf{Q}$ ($n \times n$) are implemented by calculating and using the Cholesky factorization, $\mathbf{Q} = \mathbf{L}\mathbf{L}^\intercal$, where $\mathbf{L}$ is a lower triangular matrix[3] that inherits much of the sparse structure in $\mathbf{Q}$. The cost of computing the Cholesky factorization for a sparse and a full matrix $\mathbf{Q}$, respectively; and for different problem dimensions (1D, 2D and 3D), are given in Table 2. Here, we also show how the number of non-zero elements in the matrices $\mathbf{Q}$ and $\mathbf{L}$: $n_\mathbf{Q}$ and $n_\mathbf{L}$, scales with the size of the stochastic field, n (Rue and Held 2005, chapter 2.4.3; George and Liu 1981, chapter 8).

The Kronecker structure of the spatio-temporal covariance matrix, emerging from the separable space-time covariance function (38), also reduces the computational costs and memory requirements. Both the precision matrix and the Cholesky factorization inherits the Kronecker structure of the covariance,

$$\mathbf{Q} = (\mathbf{\Sigma}_T \otimes \mathbf{\Sigma}_S)^{-1} = \mathbf{\Sigma}_S^{-1} \otimes \mathbf{\Sigma}_T^{-1} = \mathbf{Q}_T \otimes \mathbf{Q}_S \tag{39}$$

$$\mathbf{L} = \mathbf{L}_T \otimes \mathbf{L}_S \tag{40}$$

which can be used efficiently in matrix operations involving $\mathbf{Q}$.

[3] The Choleskey factorization is sometimes given as $\mathbf{Q} = \mathbf{R}\mathbf{R}^\intercal$ where $\mathbf{R} = \mathbf{L}^\intercal$ is an upper, or right, triangular matrix.

Table 2: The number of non-zero elements in matrices **Q** and **L**: $n_{\mathbf{Q}}$ and $n_{\mathbf{L}}$, and the cost for computing the Cholesky decomposition, $\mathbf{Q} = \mathbf{L}\mathbf{L}^{\mathsf{T}}$, for sparse versus full precision matrices. For fields with n different elements in 1D, 2D and 3D defined on a regular mesh (Rue and Held 2005, chapter 2.4.3; George and Liu 1981, chapter 8).

	Sparse **Q**			Full **Q**
	1D	2D	3D	1D - 3D
Field dimension	$n \times 1$	$\sqrt{n} \times \sqrt{n}$	$n^{1/3} \times n^{1/3} \times n^{1/3}$	-
$n_{\mathbf{Q}}$	$\mathcal{O}(n)$	$\mathcal{O}(n)$	$\mathcal{O}(n)$	n^2
$n_{\mathbf{L}}$	$\mathcal{O}(n)$	$\mathcal{O}(n \log n)$	$\mathcal{O}(n^{4/3})$	n^2
Cholesky	$\mathcal{O}(n)$	$\mathcal{O}(n^{3/2})$	$\mathcal{O}(n^2)$	$\mathcal{O}(n^3)$

For Gaussian fields, computations of log-likelihoods and conditional expectations needed in the inference typically involve three key matrix expressions: 1) log-determinants $\log|\mathbf{Q}|$, 2) matrix-vector multiplication $\mathbf{Qv}$, and 3) solving matrix-vector equation systems $\mathbf{Q}^{-1}\mathbf{v}$.

In the following, we discuss the computational advantages of sparse matrices, and matrices on Kronecker form, when applied to solve these three matrix expressions.

3.3.1 Computations with sparse matrices

The logarithm of the determinant can be computed via the Cholesky decomposition as:

$$\log|\mathbf{Q}| = \log|\mathbf{L}\mathbf{L}^{\mathsf{T}}| = 2\log|\mathbf{L}| = 2\sum_{i=1}^{n} \log \mathbf{L}_{ii}$$

where the last equality follows from the fact that **L** is a lower triangular matrix. The above expression also holds for Cholesky factors of full matrices. However, the cost of computing the Cholesky factorization of a sparse matrix is much lower (see Table 2).

The matrix-vector product is often not the limiting calculation in the inference. For sparse **Q** matrices, the i:th element in $\mathbf{Qv}$ is efficiently computed as

$$[\mathbf{Qv}]_i = \mathbf{Q}_{ii}\mathbf{v}_i + \sum_{j \in \mathcal{N}_i} \mathbf{Q}_{ij}\mathbf{v}_j$$

where $\mathcal{N}_i$ is the neighbourhood of i (see Section 3.1.2). Thus, the number of operations for computing all elements in the matrix-vector product scales linearly with $n_{\mathbf{Q}}$; the number of non-zero elements in $\mathbf{Q}$.

The matrix-vector equation system can be computed via a two-step procedure using the Cholesky matrix, $\mathbf{L}$. We note that

$$\mathbf{Q}^{-1}\mathbf{v} = (\mathbf{L}^{\mathsf{T}}\mathbf{L})^{-1}\mathbf{v} = \mathbf{L}^{-1}(\mathbf{L}^{-\mathsf{T}}\mathbf{v})$$

where $\mathbf{b} = \mathbf{L}^{-\mathsf{T}}\mathbf{v}$, and $\mathbf{L}^{-1}\mathbf{b}$ can be solved efficiently using forward and backward substitution. The cost of these operations scales with the number of non-zero elements in $\mathbf{L}$, i.e. $n_{\mathbf{L}}$. Since the amount of non-zero element is typically smaller then the cost of obtaining the Cholesky factorization (see Table 2), the main computational cost arrives from calculating $\mathbf{L}$.

3.3.2 Kronecker structure

With $\mathbf{Q}$ represented on a separable form (39), and using mathematical rules for the Kronecker product (further useful quantities can be found in Harville, 1997, Petersen and Pedersen, 2012), we have that

$$\log|\mathbf{Q}| = n_S \log|\mathbf{Q}_T| + n_T \log|\mathbf{Q}_S| \tag{41}$$

where n_S and n_T are the sizes of $\mathbf{Q}_S$ and $\mathbf{Q}_T$, respectively. Note that the cost of computing the Cholesky factorization for $\mathbf{Q}_T$ is of order n_T, since the temporal domain is one dimensional (see Table 2). Thus, for a process with separable space-time covariance function, the cost of computing the log determinant mainly scales with the computational cost of computing the Cholesky factorization for $\mathbf{Q}_S$.

Moreover, the matrix-vector multiplication, $\mathbf{Qv}$, can be expressed as

$$\mathbf{Qv} = (\mathbf{Q}_T \otimes \mathbf{Q}_S)\mathbf{v} = \mathrm{vec}\left(\mathbf{Q}_S \,\mathrm{ivec}(\mathbf{v})\, \mathbf{Q}_T^{\mathsf{T}}\right) \tag{42}$$

by using the vec-operator (Petersen and Pedersen 2012, p.60), illustrated in Figure 3. Hence, the high dimensional matrix-vector multiplication is replaced by the calculation of two smaller matrix-matrix multiplications. The corresponding computational cost is reduced from $\mathcal{O}(n_T^2 n_S^2)$ to $\mathcal{O}(n_T n_S^2)$ (assuming $n_T < n_S$).

Using the same calculations with vec-operators as introduced above, the matrix expression, $\mathbf{Q}^{-1}\mathbf{v}$, is given by

$$\mathbf{Q}^{-1}\mathbf{v} = \mathrm{vec}\left(\mathbf{Q}_S^{-1}\mathrm{ivec}(\mathbf{v})\mathbf{Q}_T^{-\mathsf{T}}\right) = \mathrm{vec}\left(\left[\mathbf{Q}_T^{-1}\left(\mathbf{Q}_S^{-1}\mathrm{ivec}(\mathbf{v})\right)^{\mathsf{T}}\right]^{\mathsf{T}}\right).$$

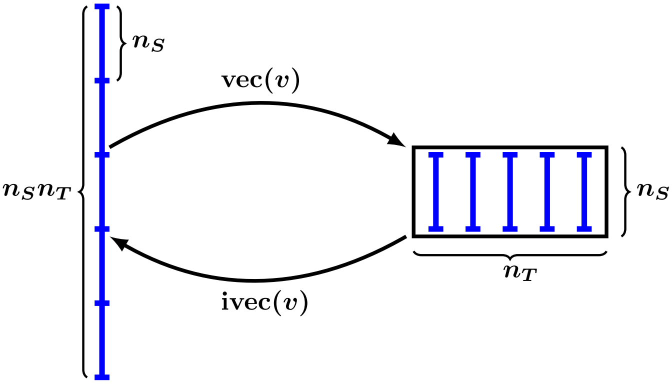


Figure 3: Illustration of the vec and ivec (inverse vec) operations that transform a $n_S n_T \times 1$ vector to a $n_S \times n_T$ matrix.

Note that the two-equation systems inside the above brackets can be solved efficiently using Cholesky factorization and back/forward substitution as explained in the previous section (3.3.1). Thus, the problem of solving one large equation system of size $n_S \cdot n_T$, is reduced to solving several smaller equation systems of size n_S and n_T. Due to the low cost of solving the forward and backward substitution, the cost of solving $\mathbf{Q}^{-1}\mathbf{v}$ scales with the computational cost of computing the Cholesky factorization for $\mathbf{Q}_S$.

Finally, we note that by using the Kronecker form (39), computation and storage of the whole ($n \times n$) precision matrix, $\mathbf{Q}$, can be avoided.

3.4 Inference

In atmospheric inverse modelling, the model introduced in Section 1.3 is linear, and can typically be expressed in the form, $\mathbf{m}(\mathbf{x}) = \mathbf{y_0} + \mathbf{A}\mathbf{x}$, where $\mathbf{A}$ is a dense matrix that approximates an integral over the latent field. The model arrives from the originally continuous representation of a single CO_2 concentration as

$$y(\mathbf{s}, t) = y_0 + \int_{t_0}^{t} \int_{S^2} J(\mathbf{s}, t, \mathbf{u}, \tau) x(\mathbf{u}, \tau) d\mathbf{u} d\tau, \tag{43}$$

where y_0 is the background concentration at time t_0, and J is the sensitivity of concentration $y(\mathbf{s}, t)$ to the flux at time τ and spatial location $\mathbf{u}$. The discretization of J is often called the transport matrix (see section 1.2.2).

The observational error is typically assumed to be independent and Gaussian with variance, σ_ε^2, resulting in a precision matrix, $\mathbf{Q}_{\boldsymbol{\varepsilon}} = \mathbf{I}\sigma_\varepsilon^{-2}$, and a data model:

$$\mathbf{y}|\mathbf{x} \sim \mathsf{N}(\mathbf{y}_0 + \mathbf{A}\mathbf{x}, \mathbf{Q}_{\boldsymbol{\varepsilon}}^{-1}) \tag{44}$$

Historically, the latent field (target), $\mathbf{x}$, has been modelled as a Gaussian field. The introduction of spatial and temporal correlation structures on fluxes corresponds to a smoothing regularization, which enables fluxes to be resolved on higher resolution (Rödenbeck et al., 2003). In both Paper C and D, the latent field is modelled as a zero-mean GMRF,

$$\mathbf{x}|\boldsymbol{\theta} \sim \mathsf{N}(\mathbf{0}, \mathbf{Q}(\boldsymbol{\theta})^{-1}) \tag{45}$$

where the precision matrix, $\mathbf{Q} = \mathbf{Q}_T \otimes \mathbf{Q}_S$, is constructed from sparse temporal and spatial precision matrices: $\mathbf{Q}_T$ and $\mathbf{Q}_S$. Conditional on parameters, $\boldsymbol{\theta}$, a Gaussian latent field with Gaussian observations yields a Gaussian posterior (5). Explicit expressions for the posterior expectation and variance are given by (see e.g. Rue and Held 2005)

$$\mathsf{E}[\mathbf{x}|\mathbf{y}] = \boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}} = \mathbf{Q}_{\mathbf{x}|\mathbf{y}}^{-1}\mathbf{A}^\mathsf{T}\mathbf{Q}_\varepsilon\left(\mathbf{y} - \mathbf{y}_0\right) \tag{46}$$

$$\mathsf{V}^{-1}[\mathbf{x}|\mathbf{y}] = \mathbf{Q}_{\mathbf{x}|\mathbf{y}} = (\mathbf{Q} + \mathbf{A}^\mathsf{T}\mathbf{Q}_{\boldsymbol{\varepsilon}}\mathbf{A}) \tag{47}$$

where the precision matrices, $\mathbf{Q}$ and $\mathbf{Q}_\varepsilon$, depend on the model parameters $\boldsymbol{\theta}$.

Models on the typical form explained above commonly have a sparse observation matrix, $\mathbf{A}$, arising from observing a single or a few components of the latent field. However, in atmospheric inversions, the observation matrix approximates an integration of the latent field and is therefore dense. This leads to a dense posterior precision, $\mathbf{Q}_{\mathbf{x}|\mathbf{y}}$ in (47), which obstructs the use of efficient inference tools, such as INLA (Rue et al., 2009).

In Paper C, the model parameters, $\boldsymbol{\theta}$ are assumed to be fixed but unknown leading to an Empirical Hierarchical Model (see Section 1.3.2). The inference is based on a maximum likelihood estimation of the unknown parameters (see Section 3.4.1), where an effort has been made to reduce the cost of computing the log-likelihood.

In Paper D, the aim is to speed up the inference by modelling the inverse problem using a BHM; assigning suitable priors to the unknown model parameters. Here, we sample the joint posterior distribution (6) using a MCMC method (see Section 3.4.2). The high dimensional latent field is sampled based on an efficient and computational beneficial Crank Nicholson proposals (Cotter et al., 2013) (see Section 3.4.3).

3.4.1 Maximum Likelihood

The maximum likelihood estimate of $\boldsymbol{\theta}$ is found by maximizing the likelihood, $L(\boldsymbol{\theta}) = p(\mathbf{y}|\boldsymbol{\theta})$, with respect to parameters, $\boldsymbol{\theta}$. The likelihood has an expression similar to (7), and is given by

$$p(\mathbf{y}|\boldsymbol{\theta}) = \int p(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta}) d\mathbf{x} = \int p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) p(\mathbf{x}|\boldsymbol{\theta}) d\mathbf{x}.$$

The above integral is hard to compute, and instead we use a ”trick” similar to the one introduced in Rue et al. (2009). Noting that

$$p(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta}) = p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}) p(\mathbf{y}|\boldsymbol{\theta}) = p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) p(\mathbf{x}|\boldsymbol{\theta}),$$

we solve for $p(\mathbf{y}|\boldsymbol{\theta})$ and obtain the likelihood,

$$p(\mathbf{y}|\boldsymbol{\theta}) = \frac{p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) p(\mathbf{x}|\boldsymbol{\theta})}{p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})}, \quad \forall \mathbf{x}. \tag{48}$$

The above expression is valid for any $\mathbf{x}$, and completely tractable since our Gaussian observations give a Gaussian posterior $p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$. By choosing $\mathbf{x} = \boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}}$, we obtain a numerical beneficial expression for the posterior distribution, for which $p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}) \propto |\mathbf{Q}_{\mathbf{x}|\mathbf{y}}|^{1/2}$. Computing (48) still requires the calculation of both the inverse (for computing $\boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}}$) and the determinant of the dense posterior precision matrix (47), both computed at a cost of n^3. In Paper C, we combine the Woodbury matrix identity (Woodbury, 1950), the structure of $\mathbf{A}$, and computations for sparse matrices (section 3.3.1), to minimize the computational cost of the likelihood (48).

3.4.2 Markov Chain Monte Carlo

The inference for BHMs is often obtained by sampling from the posterior distribution(s) using Markov Chain Monte Carlo (MCMC) (see e.g. Gilks et al. 1995, Wikle et al. 1998). MCMC is an iterative method, where the idea is to construct an irreducible and aperiodic Markov chain with the posterior, or target density, as stationary distribution. In contrast with most other sampling methods, the samples from a Markov Chain are dependent.

Suppose we want to compute $\mu = \mathsf{E}_f[\Phi(\mathbf{X})]$, i.e. the expectation of some function $\Phi(\mathbf{X})$ with respect to a target density, $f(\mathbf{x})$. In terms of BHM this can be seen as computing a posterior expectation. The Law of Large Numbers for Markov Chains implies that

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} \Phi(\mathbf{X}_i) \xrightarrow{\mathsf{P}} \int \Phi(\mathbf{x})f(\mathbf{x})d\mathbf{x} \quad as\ n \to \infty, \tag{49}$$

where $\{\mathbf{X}_i\}_{i=1}^{n}$ is a Markov Chain with stationary distribution $f(\mathbf{x})$. In other words, $\hat{\mu}_n$ converges to μ in probability. Moreover, the uncertainty in the estimate of $\hat{\mu}$ is given by the Markov Chain Central Limit Theorem (here without providing detailed conditions for the Central Limit Theorem) as:

$$\hat{\mu}_n \sim \mathsf{N}(\mu, \frac{1}{n}\left(\mathsf{V}(\Phi(\mathbf{X}_i)) + 2\sum_{i=1}^{\infty} \mathrm{Cov}\left(\Phi(\mathbf{X}_i), \Phi(\mathbf{X}_{i+k})\right)\right). \tag{50}$$

Thus, the accuracy of the MCMC estimator is connected to the pairwise covariances between samples $\Phi(\mathbf{X}_i)$ and $\Phi(\mathbf{X}_j)$. In practise the uncertainty of (50) can never be calculated, but needs to be estimated from a Markov Chain of finite length.

A popular way of assessing the quality of the Markov Chain is the mixing property; i.e., how efficiently the state space is explored by the Markov Chain. A good mixing means that $\mathsf{E}[\mathbf{X}_k - \mathbf{X}_{k-1}]$ is large. Optimal mixing properties have been derived for certain algorithms under specific conditions (the optimal acceptance rate for a Gaussian proposal is given in the next section).

3.4.3 The Metropolis-Hastings algorithm

One of the most common method for constructing a Markov Chain with a stationary distribution, f, is the Metropolis-Hastings (MH) algorithm (Hastings, 1970, Metropolis et al., 1953). Given the value of the current step of a Markov Chain, $\mathbf{X}_k = \mathbf{x}_k$, a new sample $\mathbf{x}_{k+1}$ is obtained according to the following procedure:

1. Sample a candidate value $\mathbf{x}^*$ from a proposal density, $q(\mathbf{x}|\mathbf{x}_k)$.

2. Calculate the acceptance ratio

$$a(\mathbf{x}^*, \mathbf{x}_k) = \min\left(1, \frac{f(\mathbf{x}^*)q(\mathbf{x}_k|\mathbf{x}^*)}{f(\mathbf{x}_k)q(\mathbf{x}^*|\mathbf{x}_k)}\right). \tag{51}$$

3. Accept the new proposal with probability a, by drawing a standard uniform variable, $\mathbf{U}$, and set

$$\mathbf{x}_{k+1} = \begin{cases} \mathbf{x}^* & \text{if } \mathbf{U} < a \\ \mathbf{x}_k & \text{o.w.} \end{cases} \tag{52}$$

Here, stationarity implies that if $\mathbf{x}_k$ is a sample from the target density, then $\mathbf{x}_{k+1}$ will also be a sample from f.

The choice of proposal density, q, will affect the performance of the MH algorithm. In low dimensions, a common proposal density, $\mathrm{q}(\mathbf{x}|\mathbf{x}_k)$, in the MH algorithm is the random walk proposal:

$$\mathbf{x}^* = \mathbf{x}_k + \varepsilon \qquad \varepsilon \in \mathsf{N}(\mathbf{0}, \mathbf{I}\sigma^2) \tag{53}$$

where the step size, ε, is Gaussian with standard error σ. If the step size is to large, the acceptance ratio (51) becomes low, and candidates are seldom accepted, leading to highly dependent samples and slow mixing. On the other hand, a small step size yields high acceptance rates, but a slow exploration of the state space and highly correlated samples.

In the case of a Gaussian target distribution and a Gaussian proposal, the optimal acceptance rate for 1-dimensional problems is about 44%, while for higher dimensional problems this ratio decreases to 23.4% (Gelman et al., 1996, Roberts and Rosenthal, 1998, Roberts et al., 1997, 2001).

Crank Nicholson For high dimensional distributions, it is often hard to find a proposal density for the Metropolis-Hastings algorithm with good mixing properties. Another concern is the computational cost of the MCMC method, which sometimes might be infeasible for high dimensional problems. In Cotter et al. (2013) the authors introduce efficient methods for constructing proposal distributions that are robust to the dimension of the discretized target.

Let us assume a Gaussian latent field and a general observation model:

$$\mathbf{x} \sim \mathsf{N}(\mathbf{0}, \mathbf{Q}^{-1}) \qquad p(\mathbf{y}|\mathbf{x}) \propto \exp\left(-\Phi(\mathbf{x})\right),$$

resulting in a posterior distribution:

$$p(\mathbf{x}|\mathbf{y}) \propto \exp(-\frac{1}{2}\mathbf{x}^\mathsf{T}\mathbf{Q}\mathbf{x} - \Phi(\mathbf{x})). \tag{54}$$

that we want to sample from. The proposals are based on discretizations of the Langevin SDE:

$$\frac{d\mathbf{x}}{dt} = -\mathcal{K}(\mathbf{Q}\mathbf{x} + \nabla\Phi(\mathbf{x})) + \sqrt{2\mathcal{K}}\mathcal{W}, \tag{55}$$

where $\mathcal{W}$ is Brownian white noise and $\mathcal{K}$ is a positive definite matrix acting as preconditioner, suggested by Cotter et al. (2013) to be either the identity matrix, $\mathbf{I}$, or the prior covariance matrix, $\mathbf{Q}^{-1}$. Using an implicit approximation for the linear term ($\mathbf{Qx}$), the SDE is discretized as:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \delta\mathcal{K}\left((1-\nu)\mathbf{Q}\mathbf{x}_k + \nu\mathbf{Q}\mathbf{x}_{k+1} + \nabla\Phi(\mathbf{x}_k)\right) + \sqrt{2\delta\mathcal{K}}\mathcal{W}, \tag{56}$$

where $\nu \in \begin{bmatrix}0 & 1\end{bmatrix}$. Choosing $\nu = 0$, leads to a standard forward Euler approximation of (55), resulting in the commonly used Metropolis-adjusted Langevin algorithm (MALA) proposal (Girolami and Calderhead, 2011, Roberts and Rosenthal, 1998, Roberts and Stramer, 2002).

Instead, choosing $\nu = 1/2$ (Beskos et al., 2008), leads to a Crank Nicholson discretization and update step, with $\mathbf{x}_{k+1}$ obtained by solving:

$$(\mathbf{I} + \tfrac{\delta}{2}\mathcal{K}\mathbf{Q})\mathbf{x}_{k+1} = \left(\mathbf{I} - \tfrac{\delta}{2}\mathcal{K}\mathbf{Q}\right)\mathbf{x}_k - \delta\mathcal{K}\nabla\Phi(\mathbf{x}_k) + \sqrt{2\delta\mathcal{K}}\mathcal{W}. \tag{57}$$

The acceptance rate is computed from (51), with the target, $f(\mathbf{x}, \boldsymbol{\theta})$, being the posterior in (54).

Choosing $\mathcal{K} = \mathbf{I}$, leads to the Crank Nicholson Langevin (CNL) proposal and a corresponding acceptance rate, given by:

$$\mathbf{x}_{k+1} = \left(\mathbf{I} + \tfrac{\delta}{2}\mathbf{Q}\right)^{-1}\left[\left(\mathbf{I} - \tfrac{\delta}{2}\mathbf{Q}\right)\mathbf{x}_k - \delta\nabla\Phi(\mathbf{x}_k) + \sqrt{2\delta}\mathbf{e}\right], \tag{58a}$$

$$\begin{aligned}\log a(\mathbf{x}_{k+1}|\mathbf{x}_k) = {} & \Phi(\mathbf{x}_k) - \Phi(\mathbf{x}_{k+1}) - \tfrac{\delta}{4}\|\nabla\Phi(\mathbf{x}_{k+1})\|_2^2 + \tfrac{\delta}{4}\|\nabla\Phi(\mathbf{x}_k)\|_2^2 \\ & - \tfrac{1}{2}\left[\mathbf{x}_k^\mathsf{T}\left(\mathbf{I} + \tfrac{\delta}{2}\mathbf{Q}\right) - \mathbf{x}_{k+1}^\mathsf{T}\left(\mathbf{I} + \tfrac{\delta}{2}\mathbf{Q}\right)\right]\nabla\Phi(\mathbf{x}_{k+1}) \\ & + \tfrac{1}{2}\left[\mathbf{x}_{k+1}^\mathsf{T}\left(\mathbf{I} + \tfrac{\delta}{2}\mathbf{Q}\right) - \mathbf{x}_k^\mathsf{T}\left(\mathbf{I} + \tfrac{\delta}{2}\mathbf{Q}\right)\right]\nabla\Phi(\mathbf{x}_k),\end{aligned} \tag{58b}$$

where $\mathbf{e} \sim \mathsf{N}(\mathbf{0}, \mathbf{I})$. This proposal is suitable whenever $\left(\mathbf{I} + \frac{\delta}{2}\mathbf{Q}\right)^{-1}\mathbf{x}$ can be easily computed.

Instead, choosing $\mathcal{K} = \mathbf{Q}^{-1}$ leads to the preconditioned Crank Nicholson Langevin (pCNL) proposal and a corresponding acceptance rate:

$$\mathbf{x}_{k+1} = \tfrac{1}{2+\delta}\left[(2-\delta)\mathbf{x}_k - 2\delta\mathbf{Q}^{-1}\nabla\mathit{\Phi}(\mathbf{x}_k) + 2\sqrt{2\delta}\boldsymbol{\varepsilon}\right], \tag{59a}$$

$$\begin{aligned}\log a(\mathbf{x}_{k+1}|\mathbf{x}_k) = {} & \mathit{\Phi}(\mathbf{x}_k) - \mathit{\Phi}(\mathbf{x}_{k+1}) - \tfrac{\delta}{4}\|\nabla\mathit{\Phi}(\mathbf{x}_{k+1})\|^2_{\mathbf{Q}^{-1}} + \tfrac{\delta}{4}\|\nabla\mathit{\Phi}(\mathbf{x}_k)\|^2_{\mathbf{Q}^{-1}} \\ & - \tfrac{1}{4}\left((2+\delta)\mathbf{x}_k^{\mathsf{T}} - (2-\delta)\mathbf{x}_{k+1}^{\mathsf{T}}\right)\nabla\mathit{\Phi}(\mathbf{x}_{k+1}) \\ & + \tfrac{1}{4}\left((2+\delta)\mathbf{x}_{k+1}^{\mathsf{T}} - (2-\delta)\mathbf{x}_k^{\mathsf{T}}\right)\nabla\mathit{\Phi}(\mathbf{x}_k),\end{aligned} \tag{59b}$$

where $\boldsymbol{\varepsilon} \sim \mathsf{N}(\mathbf{0}, \mathbf{Q}^{-1})$. This proposal should be used if we can sample efficiently from the prior.

In both of the Crank Nicholson proposals introduced above, the calculations involved in the proposal and acceptance steps can be speed up by making use of the sparse structure in the precision matrix in a GMRF. In Paper D we use this to construct MCMC based estimation algorithms that avoid some of the computational issues in Paper C.

4 Overview of the papers

Paper A

Parameter optimisation of Terrestrial Ecosystem Models using Gradient Adaptive Stochastic Search

Unn Dahlén, Marko Scholze, Stefan Olin, Andrew McRobert and Johan Lindström

In Paper A, a stochastic optimisation method, Gradient Adaptive Stochastic Search (GASS), is applied and adapted to a parameter estimation problem of a highly non-linear, multimodal and computational expensive vegetation model; LPJ-GUESS. The idea is to demonstrate a new efficient parallellizable optimisation method for application on Dynamical Global Vegetation Models (DGVMs), and investigate the ability of the method to find globally optimal parameter solutions.

A simulation study is performed using two sets of observational networks, one set mirroring the true availability of observations, and the other set representing availability of data in a more ideal set-up. By modifying the standard objective functions used in data assimilation for DGVMs, we are able to evaluate the improvement from additional observations.

Parameters are also optimised using true data of CO_2 fluxes at a single measurement site in Northern of Sweden.

My contribution:

Johan and Marko had the main idea of the project. The implementation and adaptation of the algorithm and the analysis was done by me, as well as the main part of the writing. Johan and Marko provided suggestions for the analysing. Stefan and Andrew helped with the settings up the model to run, and provided suitable prior information on parameters.

Paper B

Using memory-based importance sampling to improve stochastic gradient optimisation of vegetation models.

Unn Dahlén, Johan Lindström and Marko Scholze

In Paper B, we improve a stochastic optimisation algorithm, GASS, for optimisation of non-structured, multimodal, and computational expensive cost functions, motivated by the parameter estimation problem in Paper A.

The main idea is to increase the efficiency of the importance sampling procedure in GASS used for estimating the gradient in the Newton update step of the algorithm. This is done by exchanging a unimodal sample density (instrument distribution) with a multimodal sample model adapted to good candidate solutions observed in the past.

Three types of sample distributions are tested and compared to the standard GASS method. The different methods are evaluated using both test functions and the LPJ-GUESS vegetation model investigated in Paper A.

My contribution:
Me and Johan jointly formulated the project idea. The different sample distributions were mainly developed by me, with minor guidance from Johan. The analysis and algorithm implementation was performed by me. The main writing was done by me, with proofreading by Johan and Marko.

Paper C

Inverse modelling of spatio-temporal CO_2 flux fields using Gaussian Markov Random Fields

Unn Dahlén, Johan Lindström and Marko Scholze

In Paper C, the main idea is to introduce a model for spatio-temporal CO_2 surface fluxes for application in global atmospheric inverse modelling, that allows for a flexible construction of spatial-temporal covariance structures, and efficient calculations for inference.

A GMRF model is introduced for modelling past spatio-temporal CO_2 surface fluxes. In contrast with traditional flux models, the spatial representation of the fluxes on a grid is replaced by a continuously defined basis expansion on the globe, thereby reducing aggregation noise. Unlike previous applications on atmospheric inverse modelling, all unknown model parameters are estimated based on observations of CO_2 concentrations. We demonstrate that this yields the best reconstruction given the statistical models and the available observations.

My contribution:
Johan had the basic idea of the project, but the extension to multiple latent fields and seasonal dependence was my idea. The computational optimisation was developed in close collaboration between Johan and me. The implementation and analysis were done by me with minor inputs from Johan and Marko. The main

writing was done by me, with support from Johan and Marko.

Paper D

An efficient MCMC method for parameter inference in atmospheric inverse modelling of CO_2 using Gaussian Markov Random Fields.
Johan Lindström, Unn Dahlén

In Paper D, the aim is to improve the inference for the atmospheric inverse modelling based on the GMRF model introduced in Paper C.

This is done by assigning a Bayesian Hierarchical model to the inverse problem, and using a MCMC approach for estimating the unknown fluxes and model parameters. By using a version of Crank Nicholson proposals for the latent field we manage to obtain a good proposal for the high dimensional latent field. Moreover, suitable prior distributions for the model parameters are obtained based on theory for penalized complexity-priors.

My contribution:
Johan had the main idea for the project. The implementation was done jointly by Johan and me. The writing was done jointly by Johan and me, with the major methodology written by Johan.

Paper E

Damage Identification in Concrete using Impact Non-Linear Reverberation Spectroscopy.
Unn Dahlén, Nils Rydén and Andreas Jakobsson

In Paper E, an approach for identifying damage in brittle materials such as concrete is developed.

Typically, non-destructive testing are based on the evaluation of the non-linearity from the relative change in frequency and attenuation from a standard impact frequency test. Traditionally, the signal is separated into several shorter time intervals, possible overlapping, wherein the frequency and amplitude is calculated and compared between intervals.

Here, we model the signal over a wider dynamic range as compared to traditional methods by introducing a parametric model with polynomial phase and attenuation, resulting in higher resolved non-linear parameters.

This work is a continuation on the master project by Dahlén (2013).

My contribution:
Nils and Andreas introduced the problem and I had the main idea for the methodology. The implementation of the method was done by me, as well as the main part of the writing.

References

D. Baker, R. M. Law, K. R. Gurney, P. Rayner, P. Peylin, A. Denning, P. Bousquet, L. Bruhwiler, Y.-H. Chen, P. Ciais, et al. Transcom 3 inversion intercomparison: Impact of transport model errors on the interannual variability of regional CO2 fluxes, 1988–2003. *Global Biogeochemical Cycles*, 20(1), 2006.

D. J. Barrett. Steady state turnover time of carbon in the australian terrestrial biosphere. *Global Biogeochemical Cycles*, 16(4):55–1, 2002.

T. Bayes. LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFR S. *Philosophical transactions of the Royal Society of London*, 53: 370–418, 1763.

L. M. Berliner. Hierarchical Bayesian time series models. In *Maximum entropy and Bayesian methods*, pages 15–22. Springer, 1996.

J. Besag. Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society. Series B*, 36:192–236, 1974. URL `www.jstor.org/stable/2984812`.

J. Besag, J. York, and A. Mollié. Bayesian image restoration, with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics*, 43:1–59, 1991. doi: 10.1007/BF00116466.

A. Beskos, G. Roberts, A. Stuart, and J. Voss. MCMC methods for diffusion bridges. *Stochastics and Dynamics*, 8(03):319–350, 2008.

M. Blangiardo and M. Cameletti. *Spatial and spatio-temporal Bayesian models with R-INLA*. John Wiley & Sons, 2015.

R. Byrd, P. Lu, J. Nocedal, and C. Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995. doi: 10.1137/0916069.

R. H. Byrd, H. F. Khalfan, and R. B. Schnabel. Analysis of a symmetric rank-one trust region method. *SIAM Journal on Optimization*, 6(4):1025–1039, 1996.

B. P. Carlin, A. E. Gelfand, and S. Banerjee. *Hierarchical modeling and analysis for spatial data*. Chapman and Hall/CRC, 2014.

P. Ciais, P. P. Tans, J. W. White, M. Trolier, R. J. Francey, J. A. Berry, D. R. Randall, P. J. Sellers, J. G. Collatz, and D. S. Schimel. Partitioning of ocean and land uptake of CO2 as inferred by δ^{13}C measurements from the noaa climate monitoring and diagnostics laboratory global air sampling network. *Journal of Geophysical Research: Atmospheres*, 100(D3):5051–5070, 1995.

S. L. Cotter, G. O. Roberts, A. M. Stuart, D. White, et al. MCMC methods for functions: modifying old algorithms to make them faster. *Statistical Science*, 28 (3):424–446, 2013.

P. M. Cox. Description of the TRIFFID dynamic global vegetation model: Hadley Center Technical Note 24. Technical report, Hadley Centre for Climate Prediction and Research, 2001.

W. Cramer et al. Comparing global models of terrestrial net primary productivity (NPP): overview and key results. *Global Change Biology*, 5(S1):1–15, 1999.

N. Cressie and C. K. Wikle. *Statistics for spatio-temporal data*. John Wiley & Sons, 2015.

U. Dahlén. Damage Identification in Concrete using Impact Non-linear Reverberation Spectroscopy. Master's thesis, Lund University, 2013.

U. Dahlén, J. Lindström, and M. Scholze. Inverse modelling of spatio-temporal CO2 flux fields using Gaussian Markov random fields. Submitted to Environmetrics, 2019.

A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B*, 39 (1):1–38, 1977. URL `http://www.jstor.org/stable/2984875`.

I. Enting. On the use of smoothing splines to filter CO2 data. *Journal of Geophysical Research: Atmospheres*, 92(D9):10977–10984, 1987.

I. Enting, C. Trudinger, R. Francey, and H. Granek. Synthesis inversion of atmospheric CO2 using the GISS tracer transport model. Technical report, CSIRO Australian Division Atmospheric Research, 1993.

I. Enting, C. Trudinger, and R. Francey. A synthesis inversion of the concentration and δ^{13}C of atmospheric CO2. *Tellus B: Chemical and Physical Meteorology*, 47(1-2):35–52, 1995. doi: 10.1034/j.1600-0889.47.issue1.5.x.

R. Fletcher. *Practical Methods of Optimization*. Wiley, second edition, 1987.

A. E. Gelfand, P. Diggle, P. Guttorp, and M. Fuentes. *Handbook of spatial statistics*. CRC press, 2010.

A. Gelman, G. O. Roberts, W. R. Gilks, et al. Efficient metropolis jumping rules. *Bayesian Statistics*, 5(599-608):42, 1996.

A. George and J. W. Liu. *Computer Solution of Large Sparse Positive Definite*. Prentice Hall Professional Technical Reference, 1981. ISBN 0131652745.

W. R. Gilks, S. Richardson, and D. Spiegelhalter. *Markov Chain Monte Carlo in practice*. Chapman and Hall/CRC, 1995.

M. Girolami and B. Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society. Series B*, 73(2): 123–214, 2011.

K. R. Gurney, R. M. Law, A. S. Denning, P. J. Rayner, D. Baker, P. Bousquet, L. Bruhwiler, Y.-H. Chen, P. Ciais, S. Fan, et al. Towards robust regional estimates of CO2 sources and sinks using atmospheric transport models. *Nature*, 415(6872):626, 2002.

K. R. Gurney, R. M. Law, A. S. Denning, P. J. Rayner, D. Baker, P. Bousquet, L. Bruhwiler, Y.-H. Chen, P. Ciais, S. Fan, et al. Transcom 3 CO2 inversion intercomparison: 1. annual mean control results and sensitivity to transport and prior flux information. *Tellus B: Chemical and Physical Meteorology*, 55(2): 555–579, 2003.

K. R. Gurney, R. M. Law, A. S. Denning, P. J. Rayner, B. C. Pak, D. Baker, P. Bousquet, L. Bruhwiler, Y.-H. Chen, P. Ciais, et al. Transcom 3 inversion intercomparison: Model mean results for the estimation of seasonal carbon sources and sinks. *Global Biogeochemical Cycles*, 18(1), 2004.

P. Guttorp and T. Gneiting. Studies in the history of probability and statistics XLIX on the matern correlation family. *Biometrika*, 93(4):989–995, 2006.

D. A. Harville. *Matrix Algebra From a Statistician's Perspective*. Springer, first edition, 1997. doi: 10.1007/b98818.

W. K. Hastings. Monte Carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

M. Heimann and S. Körner. The global atmospheric tracer model tm2. Technical report, Max Plank Institute for Meterology, Jena, Germany, 1996.

M. Heimann and S. Körner. *The global atmospheric tracer model TM3: Model description and user's manual Release 3.8 a*. Max Plank Institute, MPI-BGC, 2003.

J. Hu, P. Hu, and H. S. Chang. A stochastic approximation framework for a class of randomized optimization algorithms. *IEEE Transactions on Automatic Control*, 57(1):165–178, 2012a.

J. Hu, Y. Wang, E. Zhou, M. C. Fu, and S. I. Marcus. A survey of some model-based methods for global optimization. In *Optimization, Control, and Applications of Stochastic Systems*, pages 157–179. Springer, 2012b.

G. Kimeldorf and G. Wahba. A correspondence between Bayesian estimation of stochastic processes and smoothing by splines. *The Annals of Statistics*, 41(2): 495–502, 1970.

S. Kirkpatrick, C. D. Gelatt, M. P. Vecchi, et al. Optimization by simmulated annealing. *Science*, 220(4598):671–680, 1983.

W. Knorr. Annual and interannual CO2 exchanges of the terrestrial biosphere: Process-based simulations and uncertainties. *Global Ecology and Biogeography*, 9(3):225–252, 2000.

N. Y. Krakauer, T. Schneider, J. T. Randerson, and S. C. Olsen. Using generalized cross-validation to select parameters in inversions for regional carbon fluxes. *Geophysical Research Letters*, 31(19), 2004.

J. Lagarias, J. Reeds, M. Wright, and P. Wright. Convergence properties of the Nelder-Mead simplex method in low dimensions. *SIAM Journal of Optimization*, 9:112–147, 1998. doi: 10.1.1.120.6062.

S. L. Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

R. M. Law, Y.-H. Chen, K. R. Gurney, and T.-. Modellers. Transcom 3 CO2 inversion intercomparison: 2. sensitivity of annual mean results to data choices. *Tellus B: Chemical and Physical Meteorology*, 55(2):580–595, 2003.

F. Lindgren, R. Håvard, 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*, 73(4):423–498, 2011.

B. Matérn. *Spatial variation. Stochastic models and their application to some problems in forest surveys and other sampling investigations.* PhD thesis, Stockholm University, Stockholm, Sweden, June 1960.

P. C. McIntosh and G. Veronis. Solving underdetermined tracer inverse problems by spatial smoothing and cross validation. *Journal of Physical Oceanography*, 23 (4):716–730, 1993.

G. McLachlan and T. Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

A. M. Michalak, L. Bruhwiler, and P. P. Tans. A geostatistical approach to surface flux estimation of atmospheric trace gases. *Journal of Geophysical Research: Atmospheres*, 109(D14), 2004.

J. A. Nelder and R. Mead. A simplex method for function minimization. *The computer journal*, 7(4):308–313, 1965.

NOAA. Earth System Research Laboratory. URL `https://www.esrl.noaa.gov/gmd/ccgg/trends/`. (Date last accessed: 2019-07-30).

D. W. Nychka. Spatial-process estimates as smoothers. In M. G. A. Schimek, editor, *Smoothing and Regression: Approaches, Computation, and Application*, pages 393–424. Wiley, New York, USA, 2000.

M. Pelikan, D. E. Goldberg, and F. G. Lobo. A survey of optimization by building and using probabilistic models. *Computational Optimization and Applications*, 21(1):5–20, 2002.

K. B. Petersen and M. S. Pedersen. The matrix cookbook, 2012. Version 20121115.

P. Rayner, I. Enting, R. Francey, and R. Langenfelds. Reconstructing the recent carbon cycle from atmospheric CO2, δ^{13}C and O2/N2 observations. *Tellus B: Chemical and Physical Meteorology*, 51(2):213–232, 1999.

P. Rayner, A. M. Michalak, and F. Chevallier. Fundamentals of data assimilation applied to biogeochemisrty. *Atmospheric Chemistry and Physics*, 2018. doi: 10.5194/acp-2018-1081. In review 2018.

G. O. Roberts and J. S. Rosenthal. Optimal scaling of discrete approximations to langevin diffusions. *Journal of the Royal Statistical Society. Series B*, 60(1): 255–268, 1998.

G. O. Roberts and O. Stramer. Langevin diffusions and Metropolis-Hastings algorithms. *Methodology and Computing in Applied Probability*, 4(4):337–357, 2002.

G. O. Roberts, A. Gelman, W. R. Gilks, et al. Weak convergence and optimal scaling of random walk Metropolis algorithms. *Annals of Applied Probability*, 7 (1):110–120, 1997.

G. O. Roberts, J. S. Rosenthal, et al. Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, 16(4):351–367, 2001.

C. Rödenbeck. Estimating CO2 sources and sinks from atmospheric mixing ratio measurements using a global inversion of atmospheric transport. Technical Report 06, Max Plank Institute for Biochemistry, Jena, Germany, 2005.

C. Rödenbeck, S. Houweling, M. Gloor, and M. Heimann. CO_2 flux history 1982–2001 inferred from atmospheric data using a global inversion of atmospheric transport. *Atmospheric Chemistry and Physics*, 3(6):1919–1964, 2003.

S. Roxburgh, S. Wood, B. Mackey, G. Woldendorp, and P. Gibbons. Assessing the carbon sequestration potential of managed forests: a case study from temperate australia. *Journal of Applied Ecology*, 43(6):1149–1159, 2006.

Y. A. Rozanov. Markov random fields and stochastic partial differential equations. *Mathematics of the USSR-Sbornik*, 32(4):515–534, 1977. doi: 10.1070/SM1977v032n04ABEH002404.

H. Rue and L. Held. *Gaussian Markov random fields: theory and applications*. Chapman and Hall/CRC, 2005.

H. Rue and H. Tjelmeland. Fitting Gaussian Markov random fields to Gaussian fields. *Scandinavian Journal of Statistics*, 29(1):31–49, 2002. doi: 10.1111/1467-9469.00058.

H. Rue, S. Martino, and N. Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society. Series B*, 71(2):319–392, 2009.

K. Sastry, D. Goldberg, and G. Kendall. Genetic algorithms. In *Search methodologies*, pages 97–125. Springer, 2005.

S. Sitch, B. Smith, I. C. Prentice, A. Arneth, A. Bondeau, W. Cramer, J. Kaplan, S. Levis, W. Lucht, M. T. Sykes, et al. Evaluation of ecosystem dynamics, plant geography and terrestrial carbon cycling in the LPJ dynamic global vegetation model. *Global Change Biology*, 9(2):161–185, 2003.

B. Smith, D. Warlind, A. Arneth, T. Hickler, P. Leadley, J. Siltberg, and S. Zaehle. Implications of incorporating N cycling and N limitations on primary production in an individual-based dynamic vegetation model. *Biogeosciences*, 11:2027–2054, 2014. doi: 10.5194/bg-11-2027-2014.

M. L. Stein. *Interpolation of spatial data: some theory for Kriging*. Springer Science & Business Media, 2012.

A. N. Tikhonov and V. I. Arsenin. *Solutions of ill-posed problems*, volume 14. Vh Winston, 1977.

G. Wahba. *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, 1990.

Y.-P. Wang, C. M. Trudinger, and I. G. Enting. A review of applications of model–data fusion to studies of terrestrial carbon fluxes at different scales. *Agricultural and Forest Meteorology*, 149(11):1829–1842, 2009.

P. Whittle. On stationary processes in the plane. *Biometrika*, pages 434–449, 1954.

P. Whittle. Stochastic processes in several dimensions. *Bulletin of the International Statistical Institute*, 40(2):975–994, 1963.

C. K. Wikle, L. M. Berliner, and N. Cressie. Hierarchical Bayesian space-time models. *Environmental and Ecological Statistics*, 5(2):117–154, 1998. doi: 10.1023/A:1009662704779.

M. A. Woodbury. Inverting modified matrices. *Memorandum report*, 42(106): 336, 1950.

E. Zhou and J. Hu. Gradient-based adaptive stochastic search for non-differentiable optimization. *IEEE Transactions on Automatic Control*, 59(7): 1818–1832, 2014.

T. Ziehn, M. Scholze, and W. Knorr. On the capability of monte carlo and adjoint inversion techniques to derive posterior parameter uncertainties in terrestrial ecosystem models. *Global Biogeochemical Cycles*, 26(3), 2012.

M. Zlochin, M. Birattari, N. Meuleau, and M. Dorigo. Model-based search for combinatorial optimization: A critical survey. *Annals of Operations Research*, 131(1-4):373–395, 2004.