Chair of Automatic Control

Technische Universität München

Diego Selle

Adaptive Sampling of Clouds with a Fleet of UAVs

Improving Gaussian Process Regression by Including Prior Knowledge

Master’s Thesis

30 September 2016

Supervisors:

Dr. habil. Simon Lacroix (LAAS-CNRS)
Prof. Dr.-Ing. habil. Boris Lohmann (TUM)
Mikhail Pak (TUM)

Erklärung

Hiermit erkläre ich, die vorliegende Arbeit selbstständig durchgeführt zu haben und keine weiteren Hilfsmittel und Quellen als die angegebenen genutzt zu haben. Mit ihrer unbefristeten Aufbewahrung in der Lehrstuhlbibliothek erkläre ich mich einverstanden.

Garching bei München, den 30. September 2016 ______________________ (Diego Selle)

Dr. habil. Simon Lacroix
Robotics and InteractionS (RIS)
LAAS-CNRS
7, avenue du Colonel Roche
BP 54200
31031 Toulouse cedex 4
France

Chair of Automatic Control (Prof. Boris Lohmann)
Technische Universität München
Boltzmannstraße 15
85748 Garching bei München
Germany

Lehrstuhl für Regelungstechnik (Prof. Boris Lohmann)
Technische Universität München
Boltzmannstraße 15
85748 Garching bei München
Deutschland

Task Description

Cooperative exploration of unknown regions is a typical problem extensively studied in AI and Robotics, with many important applications such as mapping, search and rescue and so on. The main problem for such tasks is to decide ”who goes where”, which comes to selecting the next observation points for each robot with the objective of minimizing the time to complete the exploration. To achieve such decisions, the ability to build and maintain a map of the explored environment online is of primary importance.

The proposed work is in the context of the exploration of clouds by a fleet of Unmanned Aerial Vehicles (UAVs). To devise exploration strategies, the UAVs have to maintain an accurate map of the various parameters that define a cloud: pressure, temperature, humidity, liquid water content (LWC) and wind currents. Building and maintaining a spatial map of these variables is a difficult task, because the environment is dynamic, and the aircrafts can only gather sparse and noisy measurements along their trajectories.

The goal of the Master's Thesis is to address this specific mapping problem. An approach with two hierarchical maps has been suggested: a parametric schematic model of the cloud relates global variables that define the cloud, while local dense models can be obtained using a Gaussian Process Regression (GPR). The work will consist in defining precisely the processes that build these models, and to exploit cloud micro-physics laws that relate the various variables that define the maps.

Abstract

The present Master's Thesis was done as part of the SkyScanner Project. The goal of the SkyScanner project[1], whose team is formed by atmosphere scientists and aerial roboticists, is to conceive and design a fleet of micro UAVs in order to analyze the formation and evolution of low-altitude continental cumulus clouds. In particular, by reasoning in real time on the data gathered during a mission, an *adaptive data collection scheme* that detects areas where additional measures are required can be much more efficient than any predefined acquisition pattern.

To achieve this goal, a proper real-time spatio-temporal map of the environment is essential. The core of this map is created via a GPR, which is particularly suited for mapping based on sparse noisy measurements. Moreover, it is a Bayesian framework and thus has an in-built prediction of uncertainty. This error model is paramount for the adaptive sampling scheme, for it permits to actively plan and execute trajectories that minimize uncertainty.

The current implementation of GPR can be catalogued as "off-the-shelf", which means that it does not exploit all the possibilities that the algorithm has to offer. Moreover, the hyperparameters of GPR's main ingrendient, the covariance function, are determined with online optimizations for each trajectory planning step, which are computationally expensive ($\mathcal{O}(n^3)$). Therefore, this Master's Thesis concentrates on understanding GPR and proposing improvements to the current setup.

These improvements focus on injecting information to GPR to guide it towards the real behavior that one expects from the clouds, in other words, the algorithm is "enhanced" by incorporating prior knowledge. Three types of prior knowledge for GPR were considered, namely, a prior on the mean function, on the covariance structure, and on the correlation structure between output variables.

By implementing an offline approach primarily based on variograms, it was possible to determine a mean function and a prior on the covariance. Thus, three sets of experiments using seven Gaussian Process alternatives, including the "off-the-shelf" one, were implemented. Out of the seven alternatives, one used both options of the determined prior knowledge. Furthermore, this option did not require hyperparameter optimization. Nonetheless, it had a statistically significant improvement of performance (Root-Mean-Square-Error (RMSE)) when compared to the previous "off-the-shelf" alternative.

As explained above, the chosen method to improve GPR was a variogram approach, which is extensively used in the field of *Geostatistics*. The variogram can be viewed as a near relative of the covariance function of GPR, and under certain conditions, they can be used interchangeably. The main difference is the methodology to determine the hyperparameters. While GPR's hyperparameters are calculated via an expensive Bayesian marginal likelihood optimization, the variogram-based approach relies on estimating them from the data through regular curve fitting. This method is much more scalable and thus permitted the calculation of hyperparameter priors with much larger amounts of data as had been possible with GPR.

---

[1] https://www.laas.fr/projects/skyscanner/

Kurzfassung

Diese Masterarbeit wurde im Rahmen des Projekts SkyScanner[2] erstellt. Mit einem Team aus Atmosphärwissenschaftlern und Luftfahrtingenieuren, hat SkyScanner das Ziel, eine Flotte von Mikrodrohnen zu entwickeln, die Cumuluswolken analysieren und erforschen. Es ist gewollt, unter Beachtung von Echtzeitbedingungen, dass die gesammelten Daten während einer Mission für die adaptive Trajektoriegenerierung benutzt werden. Dieses adaptive Schema wird dann die Drohnen zu den Gebieten mit der größten Unwissenheit führen, was die Qualität der Daten deutlich in Bezug auf systematische Strategien verbessern soll.

Um dieses Ziel zu erreichen, ist eine echtzeitfähige Raum-Zeit-Modellierung der Umwelt unabdingbar. Der Kern dieser Modellierung wird mithilfe der Gaussian Process Regression (GPR) implementiert. GPR ist besonders vorteilhaft für die vorhandene Problemstellung, da es gut mit spärlichen verrauschten Daten umgehen kann. Darüber hinaus folgt GPR dem Bayesschen Gedanken, d.h. ein Fehlermodell ist ein Bestandteil der Methode. Dieses Modell der Unwissenheit ist zentral für die adaptive Datensammlung, da das Modell die Gebiete mit der größten Unsicherheit erkennen kann.

Die momentane Implementierung von GPR kann als “off-the-shelf” bezeichnet werden, d.h. eine Standardimplementierung wird benutzt. Dies bedeutet, dass die ganze Palette von Optionen, die Gaussian Process Regression anbieten kann, nicht ausgenutzt wird. Darüber hinaus werden die Hyperparameter des GPRs wichtigsten Bestandteils, der Kovarianzfunktion, durch komplexe online Optimierungen bestimmt. Diese Optimierungen sollen für jeden Trajektorieplanungschritt wiederholt werden und haben eine teure Rechenkomplexität von $\mathcal{O}(n^3)$. Deshalb hat diese Arbeit als Ziel, GPR gründlich zu verstehen und daraus Verbesserungen zur momentanen Implementierung vorzuschlagen.

Die Fokussierung liegt auf der Ergänzung des Algorithmus mit Vorwissen. Dieses Vorwissen soll GPR in die Richtung des wahren Verhaltens von Wolken führen. Diesbezüglich wurden drei Typen von Vorwissen für GPR identifiziert, nämlich, die Mittelwertsfunktion, die Struktur der Kovarianz und die Struktur der Korrelation zwischen Ausgangsvariablen.

Durch die Implementierung eines auf Variogrammen basierten offline Vorgehens konnten Informationen über die Mittelwertsfunktion und die Kovarianzstruktur gewonnen werden. Demnach wurden drei Gruppen von Experimenten implementiert, die sieben Alternativen von GPR getestet haben. Eine der Alternativen, die beide Optionen vom Vorwissen ausgenutzt hat, hatte eine statistisch signifikante Verbesserung der Leistung (RMSE) und brauchte keine Optimierung der Hyperparameter.

Das gewählte auf Variogrammen basierte Vorgehen stammt aus dem Feld der *Geostatistik*. Das Variogramm hat eine enge Beziehung mit der Kovarianzfunktion von GPR, und, unter bestimmten Bedingungen, können beide Konzepte miteinander ausgewechselt werden. Der wichtigste Unterschied ist dabei die Methode, wie die Hyperparameter berechnet werden. Bei GPR wird die Optimierung via Bayessche-Marginale-Likelihood benutzt, wohingegen das Variogramm Vorgehen auf dem Angleichen einer Kurve basiert, die aus den Daten ermittelt wird. Dadurch hat die Ermittlung der Hyperparameter durch Variogramme eine bessere Skalierbarkeit, was die Ermittlung des gewollten Vorwissens mit deutlich mehreren Daten als vorher erlaubt hat.

---

[2] https://www.laas.fr/projects/skyscanner/

Contents

Task Description v

Abstract/Kurzfassung vii

1 Introduction 1

1.1 Motivation . . . 1
1.2 SkyScanner Project . . . 1
1.2.1 Related Work . . . 2
1.2.2 Adaptive Sampling Architecture . . . 4
1.3 Goals of the Thesis . . . 4
1.4 Outline of the Thesis . . . 5

2 Fundamentals 7

2.1 The Gaussian Distribution . . . 7
2.1.1 Conditional Distributions of Joint Gaussians . . . 7
2.2 Gaussian Process Regression . . . 8
2.2.1 Introduction to Gaussian Process Regression . . . 8
2.2.2 Covariance Functions . . . 9
2.2.3 Learning the Hyperparameters . . . 13
2.2.4 Incorporating a Mean Funcction . . . 13
2.2.5 Multitask Gaussian Process Regression . . . 14
2.3 Spatial Statistics . . . 14
2.3.1 Spatial Random Processes . . . 15
2.3.2 Introduction to Kriging . . . 15
2.3.3 The Variogram and Covariogram . . . 16
2.3.4 Empirical Variogram and Covariogram . . . 19
2.3.5 Variogram Models . . . 20
2.3.6 Fitting Variograms with Weighted Least Squares . . . 22
2.3.7 Anisotropy in Spatial Statistics . . . 23

2.3.8 Variogram Interpretation . . . 27
2.4 Spatio-Temporal Processes . . . 30
2.4.1 Stationarity in Time and Space . . . 30
2.4.2 Spatio-Temporal Covariance Functions . . . 30
2.4.3 Spatio-Temporal Kriging . . . 34

3 Research Approach 35
3.1 Critique to the State of the Art and Research Goals . . . 35
3.2 Incorporating Prior Knowledge . . . 35
3.3 Methodology . . . 37
3.3.1 Meso-NH Simulation . . . 37
3.3.2 Sampling Design, Empirical Variogram Estimation and Variogram Fitting . . . 38
3.4 Design of Experiments for the Gaussian Process Regression . . . 39

4 Implementation 43
4.1 Initial Data Exploration . . . 43
4.1.1 Cloud Segmentation and Visualization . . . 44
4.1.2 Cloud Empirical Variograms . . . 50
4.2 Transformation to Cylindrical Coordinates . . . 52
4.2.1 Normalized Radial Trend . . . 53
4.2.2 Normalized Polar Variograms . . . 56
4.3 Experiments . . . 64
4.3.1 Summary of results . . . 66
4.3.2 Static Cross-section . . . 68
4.3.3 Interpretation . . . 71

5 Outlook 75
5.1 Summary . . . 75
5.2 Reflections . . . 77
5.3 Future Work . . . 78

A Appendix to Implementation 79
A.1 Visualization of Clouds Two to Five . . . 79
A.2 Initial Empirical Variograms, Cloud Two to Five . . . 95
A.3 Radial Trends, Cloud Two to Five . . . 97
A.4 Appendix to the Experiments . . . 102
A.4.1 Summary of Results, other Noise Levels . . . 102

A.4.2 Static Cross-section Comparison, other Noise levels 104
A.4.3 Hyperparameter Distribution, best overall Model 109

B Notation 113
B.1 Special operators and symbols 113
B.2 List of symbols 113
B.3 Abbreviations and acronyms 114

C List of Figures 115

D List of Tables 121

E References 123

Chapter 1

Introduction

1.1 Motivation

Atmospheric models still suffer from a gap between ground-based and satellite measurements. As a consequence, the impacts of clouds remain one of the largest uncertainties in the climate General Circulation Model (GCM): for instance the diurnal cycle of continental convection in climate models predicts a maximum of precipitation at noon local time, which is hours earlier compared to observations – this discrepancy being related to insufficient entrainment in the cumulus parameterizations (Del Genio and Wu, 2010).

Despite the continual efforts of cloud micro-physics modelers to increase the complexity of cloud parameterization, uncertainties continue to persist in GCMs and numerical weather prediction (Stevens and Bony, 2013). To alleviate these uncertainties, adequate measurements of cloud dynamics and key micro-physical parameters are required.

It is not only the precision of the instruments that matters, rather it is the way in which the sampling strategy is applied. Fully characterizing the evolution over time of the various parameters (namely pressure, temperature, radiance, 3D wind, liquid water content and aerosols) within a cloud volume requires dense spatial sampling for durations of the order of one hour: a fleet of autonomous lightweight Unmanned Aerial Vehicles (UAVs) that coordinate themselves in real time can fulfill this purpose.

1.2 SkyScanner Project

Atmospheric scientists have been early users of UAVs[1], thanks to which significant scientific results have rapidly been obtained in various contexts: volcanic emissions analysis (Diaz et al., 2010), polar research (Holland et al., 2001; Inoue et al., 2008) and naturally climatic and meteorological sciences (Ramanathan et al., 2007; Corrigan et al., 2008; Roberts et al., 2008). UAVs indeed bring forth several advantages over manned flight to probe atmospheric phenomena: low cost, ease of deployment, possibility to maneuver in high turbulences (Elston et al., 2011), etc.

But a fine understanding of atmospheric phenomena requires numerous synchronized measures over some spatial and temporal extent, which only a *fleet* of UAVs can provide. The objective of the SkyScanner project[2], which gathers atmosphere scientists and aerial roboticists, is to conceive and develop a fleet of micro UAVs to better assess the formation and evolution of low-altitude continental cumulus clouds. The fleet should collect data *within and in the close vicinity* of the cloud, with a spatial and temporal resolution of respectively about 10 m and 1 Hz over the cloud lifespan. In particular, by reasoning in real time on the data gathered so far, an *adaptive data collection scheme* that detects areas where additional measures are required can be much more efficient than any predefined acquisition pattern.

---

[1] Cf the activities of the International Society for Atmospheric Research using Remotely piloted Aircraft – ISARRA, http://www.isarra.org

The SkyScanner project tackles the problem of mapping the relevant thermodynamical variables of a cloud with the following strategy: a fleet of a handful of UAVs is tasked to autonomously gather information in a specified area of the cloud. The UAVs trajectories are optimized using an on-line updated dense model of the variables of interest. The dense model is built on the basis of the gathered data with a Gaussian processes technique, and is exploited to generate trajectories that minimize the uncertainty on the required information, while steering the vehicles within the air flows to save energy.

This strategy was chosen to overcome the following challenges of mapping a cloud with a fleet of UAVs:

- It is a problem with little data available, since on the one hand the UAVs perceive the variables of interest only at the positions they reach (contrary to exteroceptive sensors used in robotics, all the atmosphere sensors perform pointwise measures at their position), and on the other hand these parameters evolve dynamically. The mapping problem in such conditions estimates a 4D structure with serial data acquired along 1D manifolds. Furthermore, even though the coarse schema of air currents within cumulus clouds is known (Fig. 1.1), the definition of laws that relate the cloud dimensions, the inner wind speeds, and the spatial distribution of the various thermodynamic variables is still a matter of research.
- It is a highly constrained problem, as the winds considerably affect the possible trajectories of the UAVs, and their energy consumption – all the more since small sized motor gliders aircrafts (maximum take off weight of 2.0 kg) have been chosen for the task, and the mission duration must be of the order of a cumulus lifespan, that is about 1 hour. Since winds are the most important variables that influence the definition of the trajectories and are mapped as the fleet evolves, mapping the cloud is a specific instance of an "explore vs. exploit" problem.

Therefore the choice of GPR to tackle the sparse mapping, because its map of the vertical winds and its model of uncertainty permit to plan and generate trajectories that minimize energy consumption and maximize information content. Energy efficiency will permit the fleet to sample the cloud for the required timespan of 1 hour, and information efficiency will improve the quality of samples to better characterize the thermodynamical properties of the studied cloud.

1.2.1 Related Work

In the literature, few recent works have tackled this problem considering realistic models. The possibility of using dynamic soaring to extend the mission duration for sampling in

---

[2] https://www.laas.fr/projects/skyscanner/

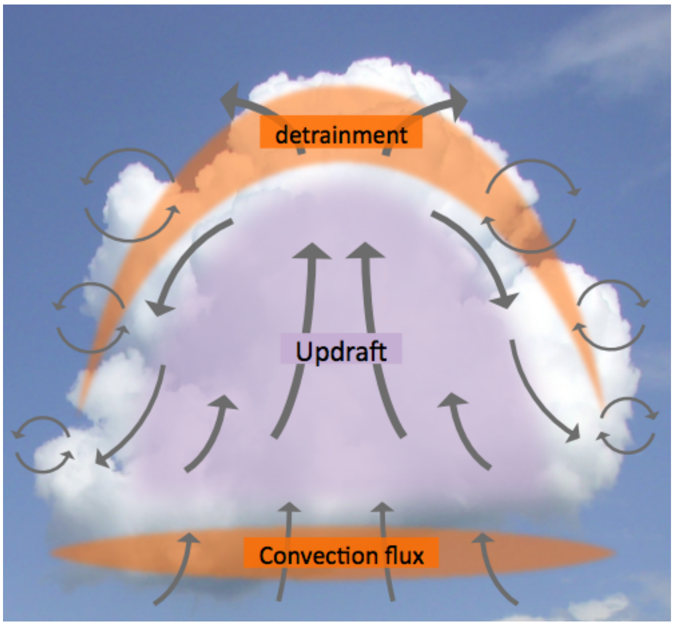


Figure 1.1: Schematic representation of a cumulus cloud. The arrows represent wind velocities, the orange blobs denote areas where mixing is occurring between the cloud and the surrounding atmosphere. This representation is very coarse: for instance the updrafts in the center of the cloud are known to behave as “bubbles” when the cloud is young. The cloud dimensions can vary from 100m to several hundreds of meters.

supercell thunderstorms has been presented in (Elston and Argrow, 2014). In this case, only the energetic part is analyzed, while the gathered information does not affect the planning. In (Lawrance and Sukkarieh, 2011), the authors present a problem very close to ours, where a glider explores a wind field trying to exploit air flows to augment flight duration. This work presents a hierarchic approach for the planning, where a target point is firstly selected and then a trajectory to reach it is generated for every planning cycle.

In a similar scenario, a reinforcement learning algorithm to find a trade-off between energy harvesting and exploration is proposed in (Chung et al., 2015). The problem of tracking and mapping atmospheric phenomena with a UAV is also studied in (Ravela et al., 2013). Even though this latter work does not take into account air currents for the navigation, it is worth to remark that it includes experiments with a real platform – contrary to the previous ones.

Autonomous soaring has also been studied in different scenarios, as in (Nguyen et al., 2013), where a glider has to search for a target on the ground. The goal here is to maximize the probability of detecting the target traveling between thermals with known location. In all the aforementioned work, only the use of a single UAV to achieve the mission is considered, and no cooperative multi-robot strategies are proposed. Autonomous exploration of current fields is not exclusively related to aerial applications: the use of Autonomous Underwater Vehicles for oceanographic studies has been recently investigated (Das et al., 2013; Michini et al., 2014).

1.2.2 Adaptive Sampling Architecture

The current simulation architecture for the entire SkyScanner project is comprised of the five components presented in Fig. 1.2. With a frequency of about $1Hz$ it is expected that the sensors, which are modelled by adding zero mean white noise, sample the ground truth of the wind vector available from the atmospheric simulation. The wind data is thus used to build the Gaussian Process models for each wind direction.

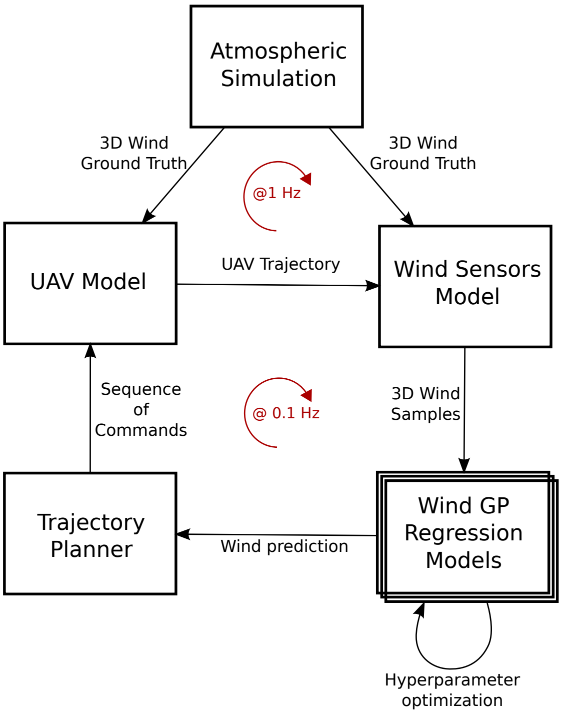

Atmospheric Simulation
3D Wind Ground Truth
3D Wind Ground Truth
@1 Hz
UAV Model
UAV Trajectory
Wind Sensors Model
Sequence of Commands
@ 0.1 Hz
3D Wind Samples
Trajectory Planner
Wind prediction
Wind GP Regression Models
Hyperparameter optimization

Figure 1.2: Simulation architecture.

The dense map of the wind predicted by the Gaussian process in the vicinity of the UAVs is then used to optimize the trajectories in terms of energy and information. The trajectories of the planner are synthetized to control commands, which in conjunction with the winds of the atmospheric simulation and the underlying UAV model generate realistic paths that the UAVs follow. It is expected that the hyperparameters optimization for the Gaussian process and the trajectory planner run on a $0.1Hz$ basis.

1.3 Goals of the Thesis

Maintaining a reliable map of the robots environment is of course of utmost importance for exploration tasks. It is necessary to assess both the feasibility of trajectories and the interesting sampling locations. In the case of atmospheric phenomena, there are numerous variables of interest to the meteorologist. Of particular interest is the 3D wind vector, as it is both one of the most dynamic atmospheric variables and an essential information for planning feasible and energy efficient paths. Therefore, the work will initially focus on the mapping of dynamic 3D wind currents, specially the vertical wind component. Due to the sparsity of the sampling process in a dynamic 3D environment, the Gaussian Process Regression probabilistic framework is particularly adapted for this mapping problem.

This online map should be comprised of two hierarchical models, the first generates a parametric description of the cloud, which should help determine geometry and other global characteristics. The second pertains to a local dense model that maps the atmospherical variables in the proximity of the UAVs, so as to enable real-time decision making regarding trajectory planning and efficient information gathering.

Most of the challenge in this mapping scheme lies in the local dense model computed via a Gaussian Process Regression, for this algorithm requires computationally expensive hyperparameter optimizations for each time step, thus jeopardizing the real time capabilities for the adaptive sampling.

To address this problem, prior knowledge on the hyperparameter probability distribution could greatly accelerate the optimization. Apart from the hyperparameters inference, other prior knowledge could be implemented to decrease computationally complexity, e.g. by computing offline, if existent, spatial trends among the variables and also by detecting and exploiting correlations between variables.

Thus, the goal of the thesis will be to improve the quality and attempt at reducing the computational complexity of the Gaussian process technique by incorporating prior knowledge to the algorithm. The most interesting prior knowledge is the structure of the covariance and its hyperparameter distribution, for this information could accelerate the hyperparameter optimization.

In addition, spatio-temporal trends and correlation between variables are other alternatives that can be exploited as priors to further improve the Gaussian process mapping.

1.4 Outline of the Thesis

After having presented a general introduction to the problematic of cloud mapping related to the SkyScanner project, chapter § 2 will elaborate on the fundamentals necessary to propose meaningful research methodologies. Herein, the emphasis will lie on three main topics:

- Gaussian Process Regression as used in the Machine Learning community
- Spatial Statistics or also colloquially referred to as Geostatistics
- Spatio-Temporal Statistics

The subchapter of GPR will be centered on understanding its theoretical foundation and also on how to implement the algorithm. Furthermore, The Spatial Statistics subchapter will deal with the concepts of spatial stochastic processes and spatial prediction, also called *Kriging*. As for the Spatio-temporal Statistics subchapter, the center of attention will lie on the approaches available to model space-time interactions when dealing with spatio-temporal stochastic processes. The main goal of the “Fundamentals” chapter is to address the different types of prior knowledge available to improve the current “off-the-shelf” implementation, which does not use priors on the Gaussian Process technique.

Chapter § 3 will deal with the “Research Approach”, which will start with a brief critique to the “State of the Art” of the SkyScanner project presented in the introduction. Afterwars, a methodology to improve the current “off-the-shelf” Gaussian process mapping technique will be proposed, which will be based on the theory of the “Fundamentals” chapter. This methodology includes determining the relevant prior knowledge and also an experiment design to test the new approach against the current one. Moreover, the current atmospheric simulation used for the adaptive sampling methodology will be detailed.

“Implementation” will be the topic of chapter § 4, where the objective is to execute the methodologies presented in “Research Approach”. As stated before, this will encompass implementing the methods to extract prior knowledge from the data for the new approach, and also, the new approach will be tested against the previous one.

Lastly, an “Outlook” chapter will be constructed, where the results of the research project will be summarized. Based upon this results, courses of action will be presented to reap the benefits out of the research project. Moreover, a discussion on the technological constraints to permit the implementation of the new approach will be detailed. And to conclude, future research possibilities will be elaborated and scrutinized.

Chapter 2

Fundamentals

This chapter deals with the necessary theory in order to propose research approaches to improve the local dense mapping process implemented via Gaussian Process Regression.

2.1 The Gaussian Distribution

Gaussian Processes are a generalization of the multivariate Gaussian Distribution for continuous random variables, summarized in a p-dimensional vector $\mathbf{x}$:

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{p/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}, \tag{2.1}$$

where $\boldsymbol{\mu}$ is a p-dimensional mean vector, $\boldsymbol{\Sigma}$ is a symmetric $p \times p$ covariance matrix and $|\boldsymbol{\Sigma}|$ refers to the determinant of $\boldsymbol{\Sigma}$.

2.1.1 Conditional Distributions of Joint Gaussians

A key property of two groups of jointly distributed Gaussian variables is that the conditional distribution of one set based on the values of the other is also Gaussian. Furthermore, the marginals of each set of variables will also follow a Gaussian distribution (Bishop, 2006).

Assuming $\mathbf{x}$ is a p-dimensional random vector with distribution $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})$, $\mathbf{x}$ can be partitioned in two disjoint subsets $\mathbf{x}_a$ and $\mathbf{x}_b$, where $\mathbf{x}_a$ comprises the first m variables and $\mathbf{x}_b$ contains the last $p-m$ variables:

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix} \tag{2.2}$$

The corresponding partitions of the mean vector $\boldsymbol{\mu}$ are

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix} \tag{2.3}$$

whereas the partitions of the covariance matrix $\mathbf{\Sigma}$ are given by

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{aa} & \mathbf{\Sigma}_{ab} \\ \mathbf{\Sigma}_{ba} & \mathbf{\Sigma}_{bb} \end{pmatrix}. \tag{2.4}$$

Since $\mathbf{\Sigma}$ is symmetric, $\mathbf{\Sigma}_{ab} = \mathbf{\Sigma}_{ba}^T$ holds and both $\mathbf{\Sigma}_{aa}$ and $\mathbf{\Sigma}_{bb}$ are symmetric.

Under this setup, the marginal distribution for each partition is simply

$$p(\mathbf{x}_a) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_a, \mathbf{\Sigma}_{aa}) \tag{2.5}$$
$$p(\mathbf{x}_b) = \mathcal{N}(\mathbf{x}_b|\boldsymbol{\mu}_b, \mathbf{\Sigma}_{bb}), \tag{2.6}$$

And most importantly, given the values of the subset $\mathbf{x}_b$, the conditional distribution has the form $p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}((\mathbf{x}_a|\boldsymbol{\mu}_{a|b}, \mathbf{\Sigma}_{a|b}))$, with $\boldsymbol{\mu}_{a|b}$ and $\mathbf{\Sigma}_{a|b}$:

$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a + \mathbf{\Sigma}_{ab}\mathbf{\Sigma}_{bb}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b), \tag{2.7}$$
$$\mathbf{\Sigma}_{a|b} = \mathbf{\Sigma}_{aa} - \mathbf{\Sigma}_{ab}\mathbf{\Sigma}_{bb}^{-1}\mathbf{\Sigma}_{ba}. \tag{2.8}$$

2.2 Gaussian Process Regression

2.2.1 Introduction to Gaussian Process Regression

GPR is a very general statistical framework, where an underlying process $y = f(x) : \mathbb{R}^n \to \mathbb{R}$ is modeled as "a collection of random variables, any finite number of which have a joint Gaussian distribution" (Rasmussen and Williams, 2006). One can view this as a way to set a Gaussian prior over the set of all admissible functions: given a location $x \in \mathbb{R}^n$, the values y taken by all admissible functions are distributed in a Gaussian manner. Under the Gaussianity assumption,the process is fully defined by its mean and covariance:

$$m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})], \tag{2.9}$$
$$k(\mathbf{x}, \mathbf{x'}) = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x'}) - m(\mathbf{x'}))] \tag{2.10}$$

In this model, the mean m and covariance k are not learned directly from the data, but given as parameters to the model. In most cases, the process is assumed to have zero mean, so that the only parameter is the covariance function or kernel. The kernel encodes the similarity of the target process f at a pair of given inputs, which in our setting describes the spatio-temporal dependence of the process. Given a set of n samples $(\mathbf{X}, \mathbf{Y})$ and assuming zero mean, the GP prior is fully defined by the $n \times n$ Gram matrix $\Sigma_{X,X} = k(X_i, X_j)$ of the covariances between all pairs of sample locations. Inference of the processes value $y_\star$ at a new location $\mathbf{x}_\star$ is then done by conditioning the joint Gaussian prior distribution on the new samples:

$$\overline{y}_\star = \mathbf{\Sigma}_{\mathbf{x}_\star,\mathbf{X}}\mathbf{\Sigma}_{\mathbf{X},\mathbf{X}}^{-1}\mathbf{Y}, \tag{2.11}$$
$$\mathbb{V}[y_\star] = k(\mathbf{x}_\star, \mathbf{x}_\star) - \mathbf{\Sigma}_{\mathbf{x}_\star,\mathbf{X}}\mathbf{\Sigma}_{\mathbf{X},\mathbf{X}}^{-1}\mathbf{\Sigma}_{\mathbf{x}_\star,\mathbf{X}}^T \tag{2.12}$$

The posterior Gaussian distribution at location $x_\star$ of the values of all admissible functions in the GP model therefore has mean $\overline{y}_\star$ and variance $\mathbb{V}[y_\star]$, which can be used both to predict the value of the function and to quantify the uncertainty of the model at this location. Thanks to the gaussianity assumption, inference has a closed form solution involving only linear algebraic equations. Computing the model is done in $\mathcal{O}(n^3)$, due to the cost of inversion of the Σ matrix and subsequent computation of the posterior are done in $\mathcal{O}(n^2)$. This can be done online using optimized linear algebra software for models of up to a few hundreds of samples.

2.2.2 Covariance Functions

The covariance function is the most important component for implementing Gaussian Process Regression, for it defines the characteristics about the function that will be learned. Furthermore, the covariance function embodies the concept of *similarity*, which is the assumption that points with inputs $\mathbf{x}$ near each other are more likely to have similar target values y.

Covariance functions need to fullfil certain conditions to be valid and moreover these can possess a series of properties that define the type of function that can be generated by the Gaussian Process. These conditions and properties will be presented in the next sections.

2.2.2.1 Properties

Stationarity and Isotropy

Stationarity refers to a covariance function that only depends on $\mathbf{x} - \mathbf{x'}$, which has the consequence that it is invariant to translations. Furthermore, if the covariance function is a function of only $|\mathbf{x} - \mathbf{x'}|$ then it is referred to as *isotropic*. Examples of covariance functions that fulfill these conditions are presented in § 2.2.2.2

If one suspects the process suffers from *Anisotropy*, i.e. the process has different behaviour in different directions, one can still rely on isotropic modeling by using $|\mathbf{x}-\mathbf{x'}|^2 = (\mathbf{x} - \mathbf{x'})^T M (\mathbf{x} - \mathbf{x'})$ for some positive semidefinite M. If M only has diagonal entries, then the model will implement different length-scales on each dimension. This concept can be used to determine the relevance of difference input directions.

Kernels, Symmetry and Positive Definiteness

A function which maps two arguments $\mathbf{x}$ and $\mathbf{x}'$ elements of the input space $\boldsymbol{\mathcal{X}}$ into $\mathbb{R}$ can be called a *kernel*. This term came to existence in the field of integral operators, where the following operator T_k makes use of kernel functions:

$$(T_k f)(\mathbf{x}) = \int_{\mathcal{X}} k(\mathbf{x}, \mathbf{x'}) f(\mathbf{x'}) d\mu(\mathbf{x'}) \tag{2.13}$$

where μ denotes a measure. Moreover, a real kernel is considered *symmetric* if $k(\mathbf{x}, \mathbf{x'}) = k(\mathbf{x'}, \mathbf{x})$, which is a property covariance functions must fulfill.

With a covariance function one can generate the *Gram matrix* Σ for a given set of inputs $\mathbf{x}_i | i = 1, ..., n$, where the entries are $\Sigma_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$. This matrix, also denoted *covariance matrix*, will be positive semidefinite for valid covariance functions, which means that it satisfies the condition $\mathbf{v}^T \Sigma \mathbf{v} \geq 0$ for all vectors $\mathbf{v} \in \mathbb{R}^n$. A symmetric matrix is positive semidefinite if and only if all of its eigenvalues are non-negative. A *kernel* is said to be positive semidefinite if it satisfies the following:

$$(T_k f)(\mathbf{x}) = \int k(\mathbf{x}, \mathbf{x'}) f(\mathbf{x}) f(\mathbf{x'}) d\mu(\mathbf{x}) d\mu(\mathbf{x'}) \geq 0 \tag{2.14}$$

for all $f \in \mathscr{L}_2(\chi, \mu)$. The same applies inverted, i.e. a kernel function that generate positive semidefinite Gram matrices for any amount of input points with real input spaces is positive semidefinite.

Upcrossing rate

In the case of 1D Gaussian processes, the interpretation of the characteristic length scale, which will be presented in section § 2.2.2.2, represents the number of upcrossings of a level u. The expected number of upcrossings $\mathbb{E}[N_u]$ of the level u on the unit interval by a zero mean, stationary, almost surely continuous Gaussian process is defined by:

$$\mathbb{E}[N_u] = \frac{1}{2\pi} \sqrt{\frac{-k''(0)}{k(0)}} \exp\left(-\frac{u^2}{2k(0)}\right) \tag{2.15}$$

Mean Square Continuity and Differentiability

Given a sequence of points $\mathbf{x_1}, \mathbf{x_2}, ...$ and a fixed $\mathbf{x}_* \in \mathbb{R}^D$, so that $|\mathbf{x}_k - \mathbf{x}_*| \to 0$ as $k \to \infty$. Mean square continuity of a process $f(\mathbf{x})$ at point $\mathbf{x}_*$ exists if $\mathbb{E}[|f(\mathbf{x}_k) - f(\mathbf{x}_*)|^2] \to 0$ as $k \to \infty$. If these conditions remains valid for $\mathbf{x}_* \in A$ where A is a subset of $\mathbb{R}^D$ then $f(\mathbf{x})$ can be considered as mean square continuous over A. With regards to the covariance function $k(\mathbf{x}, \mathbf{x'})$ of a process$f(\mathbf{x})$,the mean square continuity of the latter can only happen if and only the former is continuous at the point $\mathbf{x} = \mathbf{x'} = \mathbf{x}_*$. if the covariance function is stationary, then the aforementioned condition is summarized by checking the continuity at $k(\mathbf{0})$.

The process $f(\mathbf{x})$ is mean square differentiable if the following limit exists, i.e the mean square derivative exists:

$$\frac{\partial f(\mathbf{x})}{\partial x_i} = \underset{h \to 0}{l.i.m} \frac{f(\mathbf{x} + h\mathbf{e_i}) - f(\mathbf{x})}{h}, \tag{2.16}$$

where l.i.m refers to the limit in mean square and $\mathbf{e}_i$ is the unit vector in the ith direction. The relationship between the differentiability of the process $f(\mathbf{x})$ and its covariance function $k(\mathbf{x}, \mathbf{x'})$ is given by $\partial f(\mathbf{x})/\partial x_i = \partial^2 k(\mathbf{x}, \mathbf{x'})/\partial x_i \partial x_i'$. This can be extended to higher kth order differentiability in the case of stationary processes via
$\partial^k f(\mathbf{x})/\partial x_{i1}...x_{ik} = \partial^{2k} k(\mathbf{x})/\partial^2 x_{i1} \partial^2 x_{ik}$, where the left-hand-side has to be finite for $\mathbf{x} = \mathbf{0}$ and the right-hand-side has to exist for $\mathbf{x} \in \mathbb{R}^D$ as a mean square limit. Thus, for a stationary process the smoothness is expressed through the properties of covariance function around $\mathbf{0}$.

2.2.2.2 Examples of stationary isotropic covariance functions

In this section some widely used stationary-isotropic covariance functions will be presented. The restriction to stationary isotropic covariance functions is not obligatory but it is the author opinion that their advantages for interpretation of results and easier handling outweigh their limitations. Furthermore, it is expected that the process to be analyzed, for example, the vertical wind inside clouds, has some degree of stationarity.

For all presented covariance functions, σ_f^2 denotes the variance of the process.

Squared Exponential Covariance Function

Probably the most popular covariance function, the *squared exponential*(SE) has the following form:

$$k_{SE} = \sigma^2 \exp\left(\frac{-0.5(\mathbf{x}-\mathbf{x'})^2}{l^2}\right) \tag{2.17}$$

where l is known as the *characteristic length-scale.* This parameter drives the mean number of level-zero upcrossings for 1D squared exponential (SE) process, as can be seen using the definition in (2.15), which yields $(2\pi l)^{-1}$, thus demonstrating its role as a length-scale for the process. As for its differentiability, the SE covariance function is infinitely differentiable, thus a GP with this covariance has all orders of mean square differentiability and consequently has high level of smoothness.

The spectral density of the SE covariance function is $S(s) = \left(2\pi l^2\right)^{D/2} \exp\left(-2\pi^2 l^2 s^2\right)$. It may be considered that the high level of smoothness of the SE covariance is unrealistcal for many natural processes, nonetheless this aspect has not hindered its popularity.

Matérn Class of Covariance Functions

The expression for the *Matérn class* of covariance functions is summarized by the following:

$$k_{Matern}\left(|\mathbf{x}-\mathbf{x'}|\right) = \sigma_f^2 \left(\frac{2^{1-\nu}}{\Gamma(\nu)}\right)\left(\frac{\sqrt{2\nu}|\mathbf{x}-\mathbf{x'}|}{l}\right) K_\nu \left(\frac{\sqrt{2\nu}|\mathbf{x}-\mathbf{x'}|}{l}\right) \tag{2.18}$$

where the parameters ν and l are positive and K_ν is a modified Bessel function of the order ν. The spectral density of this covariance function is given by:

$$S(s) = \frac{2^D \pi^{D/2} \Gamma(\nu + D/2)(2\nu)^2}{\Gamma(\nu)} \left(\frac{2\nu}{l^2} + 4\pi^2 s^2\right)^{-(\nu+D/2)} \tag{2.19}$$

where D corresponds to the dimensions of the inputs. The parameter ν can be seen as a driver for smoothness, the process $f(\mathbf{x})$ is k times mean square differentiable if $\nu > k$. For instance, if $\nu \to \infty$ then the Matérn covariance function becomes the SE covariance function (2.17).The machine learning community has simplified the use of the Matérn class by restricting ν to $\nu = 3/2$ and $\nu = 5/2$, thus making the process respectively once and twice mean square differentiable. These two special cases are specified by:

$$k_{\nu=3/2}\left(|\mathbf{x}-\mathbf{x'}|\right) = \sigma_f^2 \left(1 + \frac{\sqrt{3}|\mathbf{x}-\mathbf{x'}|}{l}\right) \exp\left(\frac{-\sqrt{3}|\mathbf{x}-\mathbf{x'}|}{l}\right), \tag{2.20}$$

$$k_{\nu=5/2}\left(|\mathbf{x}-\mathbf{x'}|\right) = \sigma_f^2 \left(1 + \frac{\sqrt{5}|\mathbf{x}-\mathbf{x'}|}{l} + \frac{5(\mathbf{x}-\mathbf{x'})^2}{3l^2}\right) \exp\left(\frac{-\sqrt{5}|\mathbf{x}-\mathbf{x'}|}{l}\right), \tag{2.21}$$

Values of ν over $7/2$ are difficult to distiguinsh from the SE covariance function and should only be used if explicit prior knowledge about the mean square differentiability exists. The case of $\nu = 1/2$ yields the exponential covariance function, which will be detailed in the following section.

Exponential Covariance Function

The exponential covariance function is given by:

$$k_E\left(|\mathbf{x}-\mathbf{x'}|\right) = \sigma_f^2 \exp\left(\frac{-|\mathbf{x}-\mathbf{x'}|}{l}\right) \tag{2.22}$$

with l a positive parameter which again can be interpreted as a length-scale. Following the argumentation about the role of ν in the previous section, the exponential covariance is mean square continuous but not differentiable.

As an example, the behaviour of the Matérn covariance functions can be seen in Fig. 2.1, where the special cases of $\nu = 1/2$ (Exponential) and $\nu \to \infty$ (SE) are also depicted.

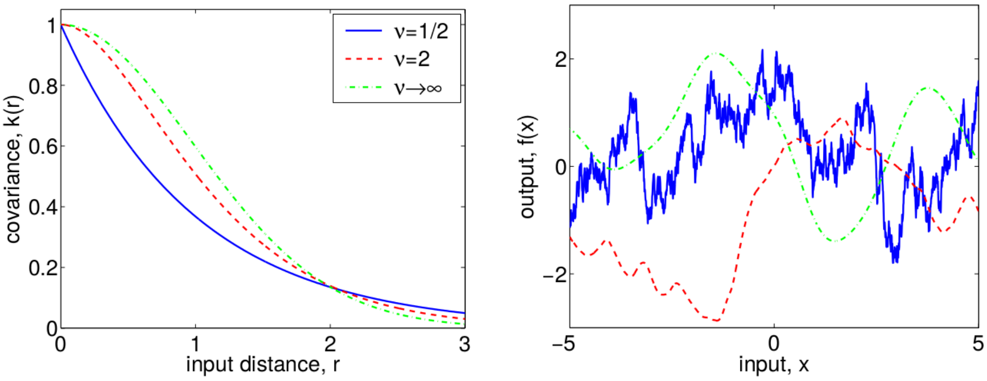


Figure 2.1: Left panel: covariance functions depending on $r = |\mathbf{x}-\mathbf{x'}|$ for different values of ν. Right panel: Random functions obtained by sampling the Gaussian Processes with Matérn covariance functions for different values of ν and $l = 1$. Extracted from (Rasmussen and Williams, 2006)

Building new covariance functions

One of the flexibilities that GPR can offer is that one can build new more complex stationary covariance functions by simply adding or multiplying simpler stationary covariance functions as the ones previously presented.

Thus, model complexity can be increased in an intuitive way to try to better grasp the underlying behaviour of the process being analyzed.

2.2.3 Learning the Hyperparameters

A real life example of a covariance function is e.g. the SE (2.17) with a noise parameter: $k_{SE}(\mathbf{x}, \mathbf{x'}) = \sigma_f^2 \exp\left(-|\mathbf{x} - \mathbf{x'}|/l^2\right) + \delta_{ij}\sigma_m^2$, where δ_{ij} is the Kronecker delta function and σ_n is the Gaussian noise parameter that permits the modeling of e.g. sensor noise. To select the hyperparameters of the aforementioned covariance function, i.e. $\boldsymbol{\theta} = (\sigma_f, l, \sigma_n)$, the most widespread method is maximizing the Bayesian log likelihood of the process:

$$\log p\left(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}\right) = -\frac{1}{2}\mathbf{y}^T\mathbf{\Sigma}^{-1}\mathbf{y} - \frac{1}{2}\log|\mathbf{\Sigma}| - \frac{n}{2}\log 2\pi, \tag{2.23}$$

where the components of the right hand side of the equation play the following role: The first term $-\mathbf{y}^T\mathbf{\Sigma}^{-1}\mathbf{y}/2$ represents the goodnes of fit of the data. The second term $-\log|\mathbf{\Sigma}|/2$ is a complexity penalty which applies only to the covariance matrix, similar to regularization in Linear Regression. The last term $n\log 2\pi/2$ is a normalization constant. The presented marginal likelihood is usually optimized using gradient methods. To this end, the gradients w.r.t the hyperparameters need to be computed:

$$\frac{\partial}{\partial\theta_j}\log p\left(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}\right) = \frac{1}{2}\mathbf{y}^T\mathbf{\Sigma}^{-1}\frac{\partial\mathbf{\Sigma}}{\partial\theta_j}\mathbf{\Sigma}^{-1}\mathbf{y} - \frac{1}{2}tr\left(\mathbf{\Sigma}^{-1}\frac{\partial\mathbf{\Sigma}}{\partial\theta_j}\right) \tag{2.24}$$

This optimization is computationally demanding, since it requires the computation of the inverse Σ^{-1} for each iteration. This translates to a complexity of $\mathcal{O}(n^3)$ for an n by n matrix. The computation of the derivatives, on the other hand , have a complexity of $\mathcal{O}(n^2)$ per hyperparameter. Thus, the computational cost of maximizing the marginal likelihood is dominated by the aforementioned matrix inversion. Furthermore, the presented optimization is non-convex, thus the convergence to global optima is not guaranteed.

2.2.4 Incorporating a Mean Funcction

Even though it is widespread to set the mean function to zero, it is by no means necessary. Nevertheless, assuming a zero mean function is not a drastic restriction, for the mean of the *posterior* process is not limited to zero.

Be that as it may, a non-zero mean function can improve interpretability of the model and it is a medium for incorporating prior information to better model the posterior process (Rasmussen and Williams, 2006).

Given a deterministic mean function $m(\mathbf{x})$ of the process that is known beforehand, the GPR equations with training Data $\mathbf{X}, \mathbf{Y}$ and a new input vector $\mathbf{x}_\star$ become:

$$\bar{y}_\star = m(\mathbf{x}_\star) + \mathbf{\Sigma}_{\mathbf{x}_\star,\mathbf{X}}\mathbf{\Sigma}^{-1}_{\mathbf{X},\mathbf{X}}(\mathbf{Y} - m(\mathbf{X})), \tag{2.25}$$

$$\mathbb{V}[y_\star] = k(\mathbf{x}_\star, \mathbf{x}_\star) - \mathbf{\Sigma}_{\mathbf{x}_\star,\mathbf{X}}\mathbf{\Sigma}^{-1}_{\mathbf{X},\mathbf{X}}\mathbf{\Sigma}^T_{\mathbf{x}_\star,\mathbf{X}}. \tag{2.26}$$

In other words, the model is fitted to the residuals $\mathbf{Y} - m(\mathbf{X})$ and to obtain $\bar{y}_\star$, the Gaussian Process predicts based on the residuals and later this prediction is added to the value of the mean function at $\mathbf{x}_\star$. It can also be seen that the predicted variance does not change with respect to the case with zero mean function.

2.2.5 Multitask Gaussian Process Regression

Multitask GPR offers an elegant way of modeling several functions (tasks) that share the same inputs $\mathbf{X}$ by transfering knowledge from one output to the other through the exploitation of task-relatedness, i.e correlation.

The simplest approach to do so is called the *Intrinsic Correlation Model* (Chai, 2010, § 2.5). Given input data $\mathbf{X}$ with n input vectors and M tasks $y_m = f_m(\mathbf{x}) : \mathbb{R}^n \rightarrow \mathbb{R}, m \in [1, ..., M]$ whose mean functions are $m_m(\mathbf{x})$, the aforementioned model assumes the following family of covariance function:

$$\mathbb{E}[(f_m(\mathbf{x}) - m_m(\mathbf{x}))(f_{m'}(\mathbf{x'}) - m_{m'}(\mathbf{x'}))] = k^f(m, m')k^x(\mathbf{x}, \mathbf{x'}), \tag{2.27}$$

where k^f enbodies the similarity between tasks and k^x relates to the similarity between inputs. By stacking all output vectors of the tasks to a single vector $\mathbf{Y} = (y_1, ..., y_m) \in \mathbb{R}^{n \cdot m}$, the covariance matrix $\mathbf{\Sigma}$ for all tasks becomes:

$$\mathbf{\Sigma} = \mathbf{\Sigma}^f \otimes \mathbf{\Sigma}_{\mathbf{X},\mathbf{X}}, \tag{2.28}$$

where $\otimes$ denotes the *Kronecker product* and $\mathbf{\Sigma}^f = k^f(m_i, m_j)$ and $\mathbf{\Sigma}_{X,X} = k^x(\mathbf{x}_i, \mathbf{x}_j)$ contain the values of the covariances between all pair of tasks and samples locations, respectively. With this expanded covariance matrix, the conditional mean $\hat{\mathbf{Y}}_\star$ for all tasks and the conditional variance $\mathbb{V}[\mathbf{Y}_{star}]$ can be computed in a straightforward manner following equations (2.11) and (2.12). Whereas k^x is usually parameterized and can be learned by optimizing the marginal likelihood, as shown in section § 2.2.3, learning k^f has no undisputed method. Different options are explored by Chai (2010, S 2.5) and the reader is advised to review his work for further details.

One disadvantage of the *Intrinsic Correlation Model* is that it assumes that all tasks have the same covariance structure in terms of the inputs, which may not always be appropriate. To overcome this drawback, the *Linear Model of Coregionalization* extends the *Intrinsic Correlation Model* by assuming a linear combination of P such covariance functions as in (2.27). For further details, the author again advises to review (Chai, 2010, § 2.5).

2.3 Spatial Statistics

The basis for the analysis of spatio-temporal data emerged from the field of Spatial Statistics, which is sometimes colloquially called Geostatistics, for the major advacements in Spatial Analysis were fueled by soil and geological analysis. Thus, the author will introduce the theory of Spatial Statistics before tackling spatio-temporal processes.

2.3.1 Spatial Random Processes

When one has to tackle spatial data, mostly in form of a scalar field, the first abstraction made is to model the process that generates the field as a stochastic process of the form (Cressie, 1993):

$$\mathbf{Z}(\mathbf{s}) : \mathbf{s} \in \mathbb{D}, \tag{2.29}$$

where $\mathbb{D}$ is a fixed subset of $\mathbb{R}^d$ with positive d-dimensional volume. Thus, the spatial index $\mathbf{s}$ is continuous throughout the region $\mathbb{D}$. The goal is to perform spatial prediction at unknown locations $\mathbf{s}_0$ of the stochastic process $\mathbf{Z}(\mathbf{s})$, also known as *kriging*. This is done by quantifying the spatial dependence of the data, which is embodied in the covariance function $C(\mathbf{s}, \mathbf{s'})$

2.3.2 Introduction to Kriging

The goal of this section is to introduce the different forms of performing spatial prediction assuming the variogram or covariogram, also denoted covariance function, are known. When given a data set with n points $\mathbf{Z} \equiv (Z(\mathbf{s}_1), Z(\mathbf{s}_2), ..., Z(\mathbf{s}_n))$ assumed to be generated by (2.29), the model assumption that is made to perform spatial prediction or *kriging* is the following(Cressie, 1993):

$$Z(\mathbf{s}) = Y(\mathbf{s}) + \varepsilon(\mathbf{s}), \tag{2.30}$$

$$Y(\mathbf{s}) = \mu(\mathbf{s}) + \delta(\mathbf{s}), \tag{2.31}$$

where $\varepsilon(\mathbf{s})$ are *iid* $\mathcal{N}(0, \sigma_\varepsilon^2)$ and independent of Y. Moreover, Y is a Gaussian process with mean function $\mu(\mathbf{s})$ and stationary covariance function $C_Y(\mathbf{s}, \mathbf{s'})$, i.e. δ is a zero-mean Gaussian process. The quantity ε models the noise or measurement error of the process Z. Therefore, the covariance function of process $Z(\mathbf{s})$ is given by:

$$C_Z(\mathbf{s}, \mathbf{s'}) = \begin{cases} C_Y(\mathbf{s}, \mathbf{s}) + \sigma_\varepsilon^2, & \mathbf{s} = \mathbf{s'} \\ C_Y(\mathbf{s}, \mathbf{s'}), & \mathbf{s} \neq \mathbf{s'}. \end{cases} \tag{2.32}$$

Cressie and Wikle (2011) differentiate between three types of kriging, *simple*, *ordinary* and *universal*, where the difference in these spatial predictors depends on the assumptions made about the mean function $\mu(\mathbf{s})$. *Simple kriging* refers to the case of a known mean function $\mu_Y(\mathbf{s})$. In this case the predicted value $\hat{Y}(\mathbf{s}_0)$ and predicted variance $\sigma_{Y,sk}^2(\mathbf{s}_0)$ at the new point $\mathbf{s}_0$ have the following form:

$$\hat{Y}(\mathbf{s}_0) = \mu_Y(\mathbf{s}_0) + \mathbf{c}_Y(\mathbf{s}_0)^T \mathbf{C}_Z^{-1}(\mathbf{Z} - \mathbf{1}\mu_Y) \tag{2.33}$$

$$\sigma_{Y,sk}^2(\mathbf{s}_0) = C_Y(\mathbf{s}_0, \mathbf{s}_0) - \mathbf{c}_Y(\mathbf{s}_0)^T \mathbf{C}_Z^{-1} \mathbf{c}_Y(\mathbf{s}_0), \tag{2.34}$$

where $\mathbf{C}_Z \equiv C_Z(\mathbf{s}_i, \mathbf{s}_j)$ is the matrix of covariances of all pairs of points in $\mathbf{Z}$ and $\mathbf{c}_Y(\mathbf{s}_0) \equiv (C_Y(\mathbf{s}_0, \mathbf{s}_1), ..., C_Y(\mathbf{s}_0, \mathbf{s}_n))$ is the covariance vector at the unsampled point $\mathbf{s}_0$. Furthermore, $\mathbf{1}$ is a vector of length n comprised entirely of ones.

As for *ordinary kriging*, it is assumed that the mean is constant, but unknown, i.e. $\mu_Y(\mathbf{s}) = \mu$. Thus, the error optimization includes the estimation of the unknown constant mean, which proves to be akin to a optimal generalized least square estimator that only takes into account the constant parameter or intercept, i.e. the estimated optimal mean results in $\mu = \hat{\mu}_{gls} \equiv (\mathbf{1}^T\mathbf{C}_Z^{-1}\mathbf{Z})/(\mathbf{1}^T\mathbf{C}_Z^{-1}\mathbf{1})$. This results in the following expressions for the predicted value $\hat{Y}(\mathbf{s}_0)$ and predicted variance $\sigma^2_{Y,ok}(\mathbf{s}_0)$ at the new point $\mathbf{s}_0$:

$$\hat{Y}(\mathbf{s}_0) = \hat{\mu}_{gls}(\mathbf{s}_0) + \mathbf{c}_Y(\mathbf{s}_0)^T\mathbf{C}_Z^{-1}(\mathbf{Z} - \mathbf{1}\hat{\mu}_{gls}) \tag{2.35}$$

$$\begin{aligned}\sigma^2_{Y,ok}(\mathbf{s}_0) &= C_Y(\mathbf{s}_0, \mathbf{s}_0) - \mathbf{c}_Y(\mathbf{s}_0)^T\mathbf{C}_Z^{-1}\mathbf{c}_Y(\mathbf{s}_0) \\ &\quad + (1 - \mathbf{1}^T\mathbf{C}_Z^{-1}\mathbf{c}_Y(\mathbf{s}_0))^2/((\mathbf{1}^T\mathbf{C}_Z^{-1}\mathbf{1})),\end{aligned} \tag{2.36}$$

Lastly, the case of *universal kriging*, which can be interpreted as kriging with a trend, considers a mean function of the form of a linear model $\mu(\mathbf{s}) \equiv \mathbf{x}(\mathbf{s})^T\boldsymbol{\beta}$, where $\mathbf{x}(s)$ is a mapping of p basis functions or covariates: $\mathbf{x}(\mathbf{s}) = (x_1(\mathbf{s}), ..., x_p(\mathbf{s}))$. Usual covariates are polynomials, splines, or wavelets and the only condition to be followed is that they have to be defined for all $\mathbf{s} \in \mathbb{D}$. Furthermore, $\boldsymbol{\beta}$ is the p length parameter vector of the linear model, wherein to model constant trends, the covariate $x_1(\mathbf{s})$ should be one. As in ordinary kriging, the parameter vector $\boldsymbol{\beta}$ is obtained automatically within the error optimization. Thus, the universal kriging equation for the predicted value $\hat{Y}(\mathbf{s}_0)$ has the following structure:

$$\hat{Y}(\mathbf{s}_0) = \mathbf{x}(\mathbf{s}_0)^T\hat{\boldsymbol{\beta}}_{gls} + \mathbf{c}_Y(\mathbf{s}_0)^T\mathbf{C}_Z^{-1}(\mathbf{Z} - \mathbf{X}\hat{\boldsymbol{\beta}}_{gls}) \tag{2.37}$$

where $\hat{\boldsymbol{\beta}}_{gls} \equiv (\mathbf{X}^T\mathbf{C}_Z^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{C}_Z^{-1}\mathbf{Z}$ is referred to as a generalized least squares linear parameter vector. As for the predicted variance $\sigma^2_{Y,uk}(\mathbf{s}_0)$, it has the form:

$$\begin{aligned}\sigma^2_{Y,uk}(\mathbf{s}_0) &= C_Y(\mathbf{s}_0, \mathbf{s}_0) - \mathbf{c}_Y(\mathbf{s}_0)^T\mathbf{C}_Z^{-1}\mathbf{c}_Y(\mathbf{s}_0) \\ &\quad + (\mathbf{x}(\mathbf{s}_0) - \mathbf{X}^T\mathbf{C}_Z^{-1}\mathbf{c}_Y(\mathbf{s}_0))^T(\mathbf{X}^T\mathbf{C}_Z^{-1}\mathbf{X})^T(\mathbf{x}(\mathbf{s}_0) - \mathbf{X}^T\mathbf{C}_Z^{-1}\mathbf{c}_Y(\mathbf{s}_0)).\end{aligned} \tag{2.38}$$

It can be noted that as the amount of parameters being estimated rises, so does the predicted variance. In other words, the prediction uncertainty is higher when compared with simple kriging.

2.3.3 The Variogram and Covariogram

The spatial dependence of the stochastic process $Z(\mathbf{s})$ represented by the variogram and covariogram, which are denoted as $2\gamma(\mathbf{s}, \mathbf{s}')$ and $C(\mathbf{s}, \mathbf{s}')$, respectively. Furthermore, the quantity $\gamma(\mathbf{s}, \mathbf{s}')$ is referred to as the *semivariogram.* Herein, $2\gamma(\mathbf{s}, \mathbf{s}')$ quantifies the *dissimilarity* between $\mathbf{s}$ and $\mathbf{s}'$, whereas $C(\mathbf{s}, \mathbf{s}')$ is a quantity of *similarity* between $\mathbf{s}$ and $\mathbf{s}'$.

To formally present the variogram and covariogram (covariance function) it is necessary to review the concept of *stationarity.* It was previously stated in § 2.2.2.1 that *stationarity* refers to the notion that statistical properties of stationary processes are invariant to translations. This assumption for the caractherization of the random process $Z(\mathbf{s})$ is summarized in two types of stationarity, *second order or weak stationarity* and *intrinsic stationarity.*

Second-order or weak Stationarity

According to Cressie (1993), Second order stationarity of $Z(\mathbf{s})$ is satisfied if:

- There exists a constant mean for the process, i.e. $\mathbb{E}(Z(\mathbf{s})) = \mu \in \mathbb{R}, \forall \mathbf{s} \in \mathbb{D}$. This stipulates that Z has an *invariant* probability distribution, i.e. it does not depend on $\mathbf{s}$. Furthermore,

- the covariance function $C(\mathbf{h})$ only depends on the separating vector $\mathbf{h}$ of the two considered locations, i.e. $\mathbf{s}, \mathbf{s}+\mathbf{h} \in \mathbb{D} : \operatorname{cov}(Z(\mathbf{s}_1), Z(\mathbf{s}_2)) = C(\mathbf{s}_1 - \mathbf{s}_2) = C(\mathbf{h})$

This covariance function is equivalent to the notion presented in § 2.2.2. Thus, it has the same properties, e.g. positive semi-definiteness, spectral representation, and so forth.

Intrinsic Stationarity

Cressie (1993) defines *intrinsic stationarity* for the process $Z(\mathbf{s})$ with the following assumptions about the differences $Z(\mathbf{s}) - Z(\mathbf{s}+\mathbf{h})$:

- The mean of the differences is translation invariant in $\mathbb{D}$ and is zero: $\mathbb{E}\left(\mathbf{Z}(\mathbf{s}+\mathbf{h}) - \mathbf{Z}(\mathbf{s})\right) = 0$. This is equivalent to assuming a constant mean of $Z(\mathbf{s})$.

- The variance of the differences is finite and its quantity depends on the displacement $\mathbf{h}$, but not on the position itself: $\operatorname{var}\left(Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s})\right) = 2\gamma(\mathbf{h}) < \infty$.

The variance of the aforementioned differences, $2\gamma(\mathbf{h})$ and half of this variance, i.e. $\gamma(\mathbf{h})$, are denoted the *variogram* and *semivariogram*, respectively. The variogram must fulfill negative semi-defineteness, in analogy to the covariance function. Furthermore, there is another condition for intrinsic stationarity, which has the following form:

$$\lim_{|\mathbf{h}| \to \infty} \frac{\gamma(\mathbf{h})}{|\mathbf{h}|^2} = 0, \tag{2.39}$$

thus, the variogram has to grow slower than quadratically as the lag goes to infinity, or else it is not a valid intrinsic stationary variogram.

Kriging Equations re-expressed with the variogram

While the *simple kriging* can only be represented in terms of the covariogram, the *ordinary kriging* equations can be re-expressed in terms of the variogram $2\gamma_Z(\mathbf{s}, \mathbf{s'})$, where $\gamma_Z(\mathbf{s}, \mathbf{s'}) = \gamma_Y(\mathbf{s}, \mathbf{s'}) + \sigma_\varepsilon^2 I(\mathbf{s} \neq \mathbf{s'})$. Thus, with $\mathbf{\Gamma}_Z \equiv \gamma_Z(\mathbf{s}_i, \mathbf{s}_j)$ an $n \times n$ matrix containing the semivariogram values for each pair of points in $\mathbf{Z}$ and $\tilde{\boldsymbol{\gamma}}_Y(\mathbf{s}_0)$ the semivariogram vector $\tilde{\boldsymbol{\gamma}}_Y(\mathbf{s}_0) \equiv (\gamma_Y(\mathbf{s}_0, \mathbf{s}_1), ..., \gamma_Y(\mathbf{s}_0, \mathbf{s}_n)$, the equations become:

$$\hat{Y}(\mathbf{s}_0) = \hat{\mu}_{gls}(\mathbf{s}_0) + \tilde{\boldsymbol{\gamma}}_Y(\mathbf{s}_0)^T \mathbf{\Gamma}_Z^{-1}(\mathbf{Z} - \mathbf{1}\hat{\mu}_{gls}) \tag{2.40}$$

$$\begin{aligned} \sigma^2_{Y,ok}(\mathbf{s}_0) &= \tilde{\boldsymbol{\gamma}}_Y(\mathbf{s}_0, \mathbf{s}_0)^T \mathbf{\Gamma}_Z \tilde{\boldsymbol{\gamma}}_Y(\mathbf{s}_0, \mathbf{s}_0) \\ &\quad - (1 - \mathbf{1}^T \mathbf{\Gamma}_Z^{-1} \tilde{\boldsymbol{\gamma}}_Y(\mathbf{s}_0))^2 / ((\mathbf{1}^T \mathbf{\Gamma}_Z^{-1} \mathbf{1})), \end{aligned} \tag{2.41}$$

where $\hat{\mu}_{gls} \equiv (\mathbf{1}^T\boldsymbol{\Gamma}_Z^{-1}\mathbf{Z})/(\mathbf{1}^T\boldsymbol{\Gamma}_Z^{-1}\mathbf{1})$ is the equivalent generalized least squares mean.

As in the case of ordinary kriging, the equations of universal kriging can be re-expressed in terms of the semivariogram too, provided $\mathbf{x}(\mathbf{s})^T\boldsymbol{\beta}$ has an intercept term, i.e $x_1(\mathbf{s}) \equiv 1$:

$$\hat{Y}(\mathbf{s}_0) = \mathbf{x}(\mathbf{s}_0)^T\hat{\boldsymbol{\beta}}_{gls} + \tilde{\boldsymbol{\gamma}}_Y(\mathbf{s}_0)^T\boldsymbol{\Gamma}_Z^{-1}(\mathbf{Z} - \mathbf{X}\hat{\boldsymbol{\beta}}_{gls}) \tag{2.42}$$

where in this case, the generalised least squares parameter vector estimate is $\hat{\boldsymbol{\beta}}_{gls} \equiv (\mathbf{X}^T\boldsymbol{\Gamma}_Z^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Gamma}_Z^{-1}\mathbf{Z}$

$$\begin{aligned}\sigma^2_{Y,uk}(\mathbf{s}_0) &= \tilde{\boldsymbol{\gamma}}_Y(\mathbf{s}_0,\mathbf{s}_0)^T\boldsymbol{\Gamma}_Z\tilde{\boldsymbol{\gamma}}_Y(\mathbf{s}_0,\mathbf{s}_0)\\ &\quad - (\mathbf{x}(\mathbf{s}_0) - \mathbf{X}^T\boldsymbol{\Gamma}_Z^{-1}\tilde{\boldsymbol{\gamma}}_Y(\mathbf{s}_0))(\mathbf{X}^T\boldsymbol{\Gamma}_Z^{-1}\mathbf{S})^T(\mathbf{x}(\mathbf{s}_0) - \mathbf{X}^T\boldsymbol{\Gamma}_Z^{-1}\tilde{\boldsymbol{\gamma}}_Y(\mathbf{s}_0))\end{aligned} \tag{2.43}$$

For the derivation of the presented equations, the reader is advised to review (Cressie, 1993, § 3) and(Cressie and Wikle, 2011, § 4). Given the assumption that Y is a Gaussian process, the kriging equations in terms of the covariance function are the same as in § 2.2.4, which is entirely expected. But the alternative of modeling Y in terms of the variogram adds an extra tool which can be exploited in a different way as previously presented in the Gaussian process section.

Perhaps the biggest difference is how the hyperparameters are chosen in spatial statistics. While Maximum Marginal Likelihood is the preferred method to fit the covariance function by Machine Learning practitioners, geostatisticians tend to estimate the variogram empirically from the data and then do regular curve fitting to estimate the variogram model.

This latter procedure will be the center of the following sections.

Equivalence of the Variogram and Covariance Function

It can be noted that the intrinsic stationarity of $Z(\mathbf{s})$ presented in § 2.3.3 is more general than requiring second-order stationarity, since any second-order stationary random process is automatically intrinsically stationary, i.e. the set of all second-order stationary random functions is a subset of the set of all intrinsically stationary functions (Cressie, 1993). Thus, the reversal does not apply. Hence, a variogram function, which only requires intrinsic stationarity, could exist even if there is no covariance function, which requires the stronger assumption of second-order stationarity.

The aforementioned has the consequence that the variogram and covariance have equivalences if $Z(\mathbf{s})$ is second-order stationary (Matheron, 1971):

$$\gamma(\mathbf{h}) = C(\mathbf{0}) - C(\mathbf{h}), \tag{2.44}$$

$$C(\mathbf{h}) = \gamma(\boldsymbol{\infty}) - \gamma(\mathbf{h}), \tag{2.45}$$

which requires a bounded variogram with $\lim_{|h|\to\infty}\gamma(\mathbf{h}) \equiv \gamma(\boldsymbol{\infty}) < \infty$. This directly implies that a bounded variogram is second-order stationary.

As seen in § 2.3.2 and § 2.3.3, the variogram and covariogram are the basis for kriging, for they embody the spatial dependence of the process. The kriging equations require either the knowledge of the variogram or covariance function. Geostatistician estimate these in two phases. First, the empirical variogram or covariogram is calculated from the data taking into account the stationarity assumptions of the process. Secondly, since this empirical quantities do not necessarily have to fulfill the requirement of positive- (covariogram) or negative-semidefinitenes(variogram), model functions that do satisfy these conditions are fitted to the empirical quantity. The model can then be used to perform spatial prediction.

2.3.4 Empirical Variogram and Covariogram

In order to perform kriging, either the variogram or covariogram need to be known. The first step to model a valid variogram or covariogram is to estimate these quantities from the data. In the case of the variogram, this is achieved through the following formula (Cressie, 1993), also known as the *Method of Moments Estimator*:

$$2\hat{\gamma}(\mathbf{h}) \equiv \frac{1}{|N(\mathbf{h})|} \sum_{N(\mathbf{h})} (Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2, \mathbf{h} \in \mathbb{R}^d \tag{2.46}$$

where

$$N(\mathbf{h}) \equiv \{(\mathbf{s}_i, \mathbf{s}_j) : \mathbf{s}_i - \mathbf{s}_j = \mathbf{h}; i, j = 1, ..., n\} \tag{2.47}$$

is the set of points that are at a given lag $\mathbf{h}$ and $|N(\mathbf{h})|$ is the cardinality of said set, i.e. the number of points in the set. This estimator is only valid under the intrinsic stationarity assumption. This can be visualized if one recalls the model structure (2.31). With the aforementioned structure, the variance of the differences of the process implies the following:

$$\mathrm{var}(Z(\mathbf{s}_i) - Z(\mathbf{s}_j)) = 2\gamma(\mathbf{h}) = \mathbb{E}[(Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2] - (\mu(\mathbf{s}_1) - \mu(\mathbf{s}_2))^2 \tag{2.48}$$

Therefore, the empirical estimator, which calculates $\mathbb{E}[(Z(\mathbf{s}_i) - Z(\mathbf{s}_j)^2]$, will actually estimate the variogram correctly if the mean is invariant, i.e. $(\mu(\mathbf{s}_1)-\mu(\mathbf{s}_2))^2 = 0$. Hence, an unbounded empirical variogram that does not satisfy (2.39) may be caused by the false assumption of intrinsic stationarity. This problem may be debugged by detecting and modeling the spatial trend of the process and then estimating the variogram on the residuals.

Moreover, Cressie (1993) also proposes another variogram estimator, which he calls the *robust variogram estimator*, for it should be less prone to contamination due to outliers. This estimator is expressed via:

$$2\hat{\gamma}(\mathbf{h}) \equiv \frac{\left(\frac{1}{|N(\mathbf{h})|} \sum_{N(\mathbf{h})} |Z(\mathbf{s}_i) - Z(\mathbf{s}_j)|^{1/2}\right)^4}{0.457 + 0.494/|N(\mathbf{h})|} \tag{2.49}$$

Apart from the variogram, the covariance function also has an empirical estimator of the form:

$$\hat{C}(\mathbf{h}) \equiv \frac{1}{|N(\mathbf{h})|} \sum_{N(\mathbf{h})} (Z(\mathbf{s}_i) - \bar{Z})(Z(\mathbf{s}_j) - \bar{Z}) \tag{2.50}$$

where $\bar{Z} = \sum_{i=1}^{n} Z(\mathbf{s}_i)/n$ is an estimator of the constant mean μ. This estimator relies on the second-order stationarity assumption, in analogy to the necessity of intrinsic stationarity to estimate the empirical variogram. It can be seen that the estimators do not fulfill $2\hat{\gamma}(\mathbf{h}) \neq \hat{C}(\mathbf{0}) - \hat{C}(\mathbf{h})$. Nonetheless, for $|N(\mathbf{h})|/n$ near 1, the dissimilarity between the two expressions will be negligible.

Empirical Variogram and Covariogram Comparison

Cressie (1993) argues that the empirical variogram should be preferred over the empirical covariogram. Should $Z(\mathbf{s})$ be second-order stationary then the direct relationship in (2.44)(2.45) can be used to obtain the covariogram from the variogram. Moreover, the class of intrinsic stationary process contains the class of second-order stationary process, ergo, the variogram can model processes where the covariogram would be invalid or would be estimating a quantity that does not exist. Lastly, under the intrinsic stationarity assumption, the empirical variogram does not require the mean to calculate the estimator.

2.3.5 Variogram Models

In this section some widely used parametric expressions to model the semivariogram are presented. These are the basis for the model fitting procedure of the next section. Moreover, this current section will only consider isotropic stationary models. The topic of anisotropy in spatial statistics will be dealt with in § 2.3.7. Before presenting the parametric models, the general components that define the semivariogram will be discussed.

Variogram Components

To model bounded stationary semivariograms, the following components that define its general structure can be expected to exist (Curran, 1988):

- The *nugget effect* models Gaussian noise and can be interpreted as a complete lack of spatial correlation. It manifests itself as a constant $c_0 > 0$ so that $\gamma(\mathbf{h}) \rightarrow c_0$ as $\mathbf{h} \rightarrow \mathbf{0}$ (Cressie, 1993). It is usually used to model the measurement error of sensors. Nonetheless, in some geostatistical settings, the nugget effect corresponds to micro-scale variations, e.g. small gold nuggets, that cause a discontiniuty in the origin.

- The *Spatially dependent structural variance* models the variance that actually depends on distance and is thus the center point for the spatial analysis, i.e. this is the component that embodies the dissimilarity between input points based on distance.

- The *sill* is the converging value of the semivariogram:$\lim_{|h| \rightarrow \infty} \gamma(\mathbf{h}) = \gamma(\infty)$, which is the sum of the nugget effect and the maximum of the spatially dependent variance

- The *range* is the distance h at which the semivariogram reaches the sill value for the first time. In case of asymptotical models, the practical range is defined as the distance at which the variogram reaches 95% of the sill. For distances beyond the range, the corresponding random variables are said to be uncorrelated because its associated covariogram equals zero (in case of a bounded variogram)

- Lastly, the *support* is the resolution of the spatial data, i.e. the smallest distance between data points at which the empirical variograms can be calculated.

The aforementioned components can be visualized in Fig. 2.2.

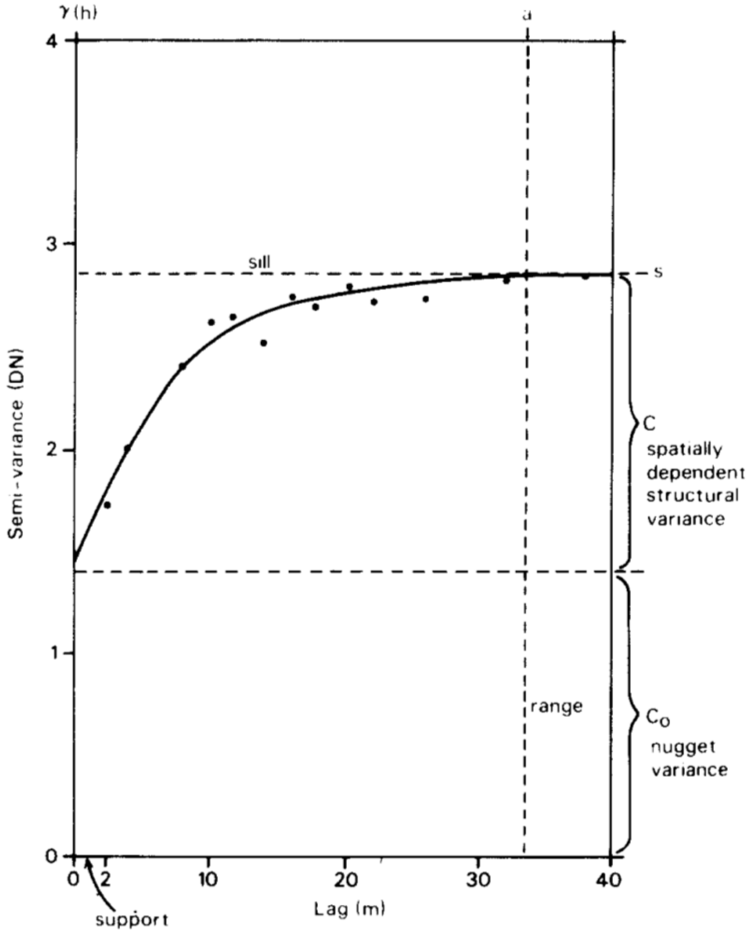


Figure 2.2: Visualization of the semivariogram components as in (Curran, 1988).

To describe the components presented above, parametric models for bounded isotropic semivariograms are available. These are presented in the following segments. Herein, the parameter σ_f^2 represents the sill and l corresponds to the parameter that controls the range. To account for the nugget effect, one simply needs to add a constant c_0 to the parametric models.

The Matérn Family of variogram models

By recalling the expression for the Matérn family of covariance functions (2.18) and the equivalences between second-order stationary covariance functions and variograms (2.44), it is quite straightforward to define the Matérn family of semivariograms:

$$\gamma_{Matern}(\mathbf{h}) = \sigma_f^2 \left(1 - \frac{2^{1-\nu}}{\Gamma(\nu)} \frac{\sqrt{2\nu}|\mathbf{h}|}{l} K_\nu \left(\frac{\sqrt{2\nu}|\mathbf{h}|}{l}\right)\right) \tag{2.51}$$

where the parameters ν and l are positive and K_ν is a modified Bessel function. These have the same interpretation as in (2.18). Thus, for $\nu \to \infty$ one obtains the squared exponential semivariogram:

$$\gamma_{SE}(\mathbf{h}) = \sigma_f^2 \left(1 - \exp\left(\frac{-|\mathbf{h}|^2}{l^2}\right)\right). \tag{2.52}$$

Moreover, for $\nu = 1/2$ one obtains the exponental semivariogram:

$$\gamma_E = \sigma_f^2 \left(1 - \exp\left(\frac{-|\mathbf{h}|}{l}\right)\right). \tag{2.53}$$

Lastly, the simplified Matérn semivariogram for $\nu = 3/2$ and $\nu = 5/2$ have the following form:

$$\gamma_{\nu=3/2}(|\mathbf{h}|) = \sigma_f^2 \left(1 - \left(1 + \frac{\sqrt{3}|\mathbf{h}|}{l}\right) \exp\left(\frac{-\sqrt{3}|\mathbf{h}|}{l}\right)\right), \tag{2.54}$$

$$\gamma_{\nu=5/2}(|\mathbf{h}|) = \sigma_f^2 \left(1 - \left(1 + \frac{\sqrt{5}|\mathbf{h}|}{l} + \frac{5|\mathbf{h}|^2}{3l^2}\right) \exp\left(\frac{-\sqrt{5}|\mathbf{h}|}{l}\right)\right), \tag{2.55}$$

The Spherical Variogram

Another parametric model proposed by Cressie (1993) is the spherical one, which has the following expression:

$$\gamma_{spherical}(|\mathbf{h}|) = \begin{cases} \sigma_f^2 \left(\frac{3}{2}\frac{|\mathbf{h}|}{l} - \frac{1}{2}\frac{|h|^3}{l^3}\right), & 0 < |h| \leq l \\ \sigma_f^2 & |h| > l. \end{cases} \tag{2.56}$$

In contrast to the Matérn family, the covariance of this model becomes exactly zero once the sill is reached, i.e. after the distance of the range, the variogram reaches the sill. This can be observed in the example of Fig. 2.3.

2.3.6 Fitting Variograms with Weighted Least Squares

Cressie (1985) recommends the method of Weighted Least Squares (WLS) to fit variogram models to empirical variograms. Its advantage in comparison to Ordinary Least

Squares (OLS) is that it gives more importance to variogram values at shorter distances through weights that are proportional to the amount of points available at each discrete distance. Naturally, the smaller lags will have more examples available than bigger lags.

Assuming a discrete amount of lags denoted by $h(j)$, which is natural for typical grid designs, the goal of Weighted Least Squares is to minimize the following expression w.r.t. to the hyperparameter vector $\boldsymbol{\theta}$:

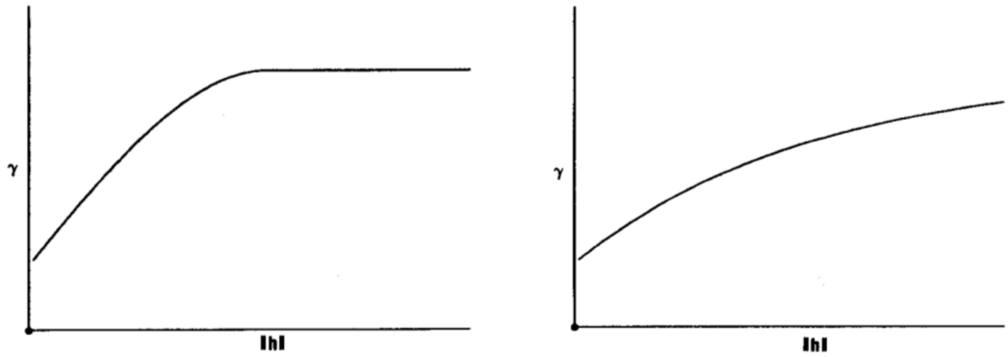
$$\sum_{j=1}^{k} |N(h(j))| \left(\frac{\hat{\gamma}(h(j))}{\gamma(h(j);\theta)} - 1 \right)^2 \tag{2.57}$$

For example, in the case of a SE semivariogram including nugget effect, the hyperparameters vector would be $\boldsymbol{\theta} = (\sigma_f, l, c_0)$. The fitted variogram models can then be utilized to perform spatial predictions via kriging.

2.3.7 Anisotropy in Spatial Statistics

Zimmerman (1993) differentiates between three types of anisotropy in Geostastics, namely *range*, *nugget* and *sill* anisotropy. These can be analyzed in terms of the variogram. To this end, empirical variograms need to be computed in what are expected to be the important directions for the process and then these are compared with one another. If there are not any substantial difference between the variograms in different directions, then the isotropy assumption is valid and it is acceptable to implement kriging with a single isotropic variogram model.

On the other hand, if any of the aforementioned types of anisotropy is present, adjustments to the modeling with isotropic variograms need to be made to enable spatial prediction. Furthermore, certain assumptions, e.g. second-order stationarity, need to be revisited in the presence of certain types of anisotropy.



(a) Spherical with Nugget Effect (b) Exponential with Nugget Effect

Figure 2.3: Examples of Semivariograms Models as in (Cressie, 1993)

2.3.7.1 Sill Anisotropy

When the sill exists, then $Z(\mathbf{s})$ is second-order stationary and thus the relationship (2.44) between the variogram and the covariance holds. Ergo, the sill can be obtained from:

$$\lim_{\alpha \to \infty} \gamma(\alpha \mathbf{h}) = C(\mathbf{0}) - \lim_{\alpha \to \infty} C(\alpha \mathbf{h}) \tag{2.58}$$

for any fixed $\mathbf{h}$. The left-hand side of the equation can be interpreted as the sill in direction $h/|h|$. Therefore, if the sills are different in various directions-assuming isotropic nugget effect-it implies that there are vectors, i.e. directions, $\mathbf{h}_1$ and $\mathbf{h}_2$ such that $\lim_{\alpha \to \infty} C(\alpha \mathbf{h}_1) \neq \lim_{\alpha \to \infty} C(\alpha \mathbf{h}_2)$ and where at least one of these directions fulfill $\lim_{\alpha \to \infty} C(\alpha \mathbf{h}) \neq 0$ for some $\mathbf{h}$. Ergo, there exists directions where the correlation does not vanish as distance increases. An example of sill anisotropy can be visualized in Fig. 2.4. According to Zimmerman (1993), there are two probable interpretations that can cause sill anisotropy. Either the second-order stationarity assumption still holds but some directions do not have dissapearing correlation, or the second-order stationary model is not valid. Zimmerman (1993) is more inclined for the latter option and argues that the usual suspect to cause direction-dependent sills is the presence of trend in the mean, i.e. the mean $\mu(\mathbf{s})$ is not constant and depends on s.

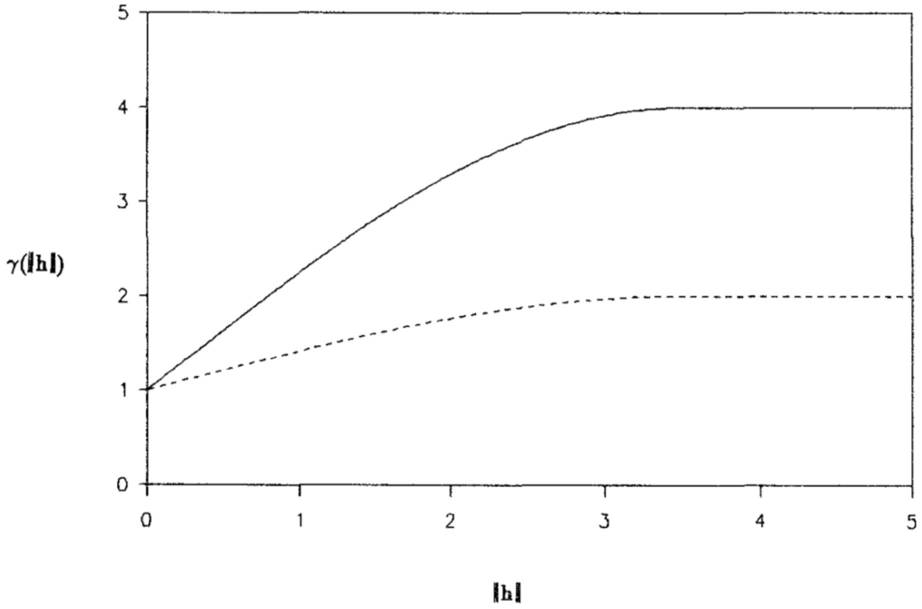


Figure 2.4: Sill Anisotropy in two dimensional data: The solid line corresponds to the E-W direction and the N-S is given by dashed lines Taken from (Zimmerman, 1993)

Be that as it may, common practice to deal with direction-dependent sill is to nest the semivariograms of different directions, i.e building the semivariogram as following:

$$\gamma(\mathbf{h}) = \sum_{i=1}^{m} \gamma_i(|\mathbf{A}_i \mathbf{h}|) \tag{2.59}$$

where $\mathbf{A}_i, ..., \mathbf{A}_m$ are matrices that define proper directions $\mathbf{A}_i/|\mathbf{A}_i\mathbf{h}|$ and $\gamma_i(\cdot)$ are isotropic variograms. For example, in a 2D case with two direction-dependent sills one would have the following:

$$\gamma(h_x, 0) = \sigma_x^2 \gamma_{iso}(|h_x|/l) \tag{2.60}$$

$$\gamma(0, h_y) = \sigma_y^2 \gamma_{iso}(|h_y|/l), \tag{2.61}$$

where γ_{iso} is any of the presented isotropic variogram models and $\sigma_x^2 > \sigma_y^2$. With these direction-dependent variograms the nested model becomes:

$$\gamma(h_x, h_y) = \sigma_y^2 \gamma_{iso}(|\mathbf{h}|/l) + (\sigma_x^2 - \sigma_y^2)\gamma(|h_x|/l) \tag{2.62}$$

It should to be noted that modeling sill anisotropy as previously presented does not tackle its source, i.e. the process $Z(\mathbf{s})$ could violate the second-order stationarity assumption. Furthermore, (2.62) is not the only possibility to implement a nested model, i.e. it comes to the user's discretion to define the nested model that best suits the problem.

2.3.7.2 Range Anisotropy

Range anisotropy refers to bounded directional semivariograms that reach the same sill value at different lag distances while also having the same nugget value. According to Zimmerman (1993) there are two types of range anisotropy, the geometric and the non-geometric, wherein the geometric anisotropy is more common.

Geometric Anisotropy can be modeled through the assumption that there exists a symmetric positive definite coordinate transformation $\gamma(\mathbf{h}) = \gamma((\mathbf{h}^T\mathbf{B}\mathbf{h})^{1/2})$ so that an isotropic model can be used on the new coordinates. To asses this assumption in two dimensional data, the ranges l_{φ_i} in dependence of different directions(angles) φ_i are plotted. If the plot, also referred to as *roseplot*, has an approximate elliptical form, then geometric assumption is valid. Thus, there are two principal directions normal to each other that have the minimal and maximal range and as the angle moves from one to the other direction the range gradually changes from the smaller to the bigger value and viceversa. An example of geometric anisotropy can be visualized in Fig. 2.5

In the case of non-geometric range direction dependency, the range does not follow gradual increase from the dominant to the less dominant direction. This can be viewed in Fig. 2.6.

Thus the approach is to fit isotropic semivariograms in the different relevant directions and then build a nested model as in (2.59).

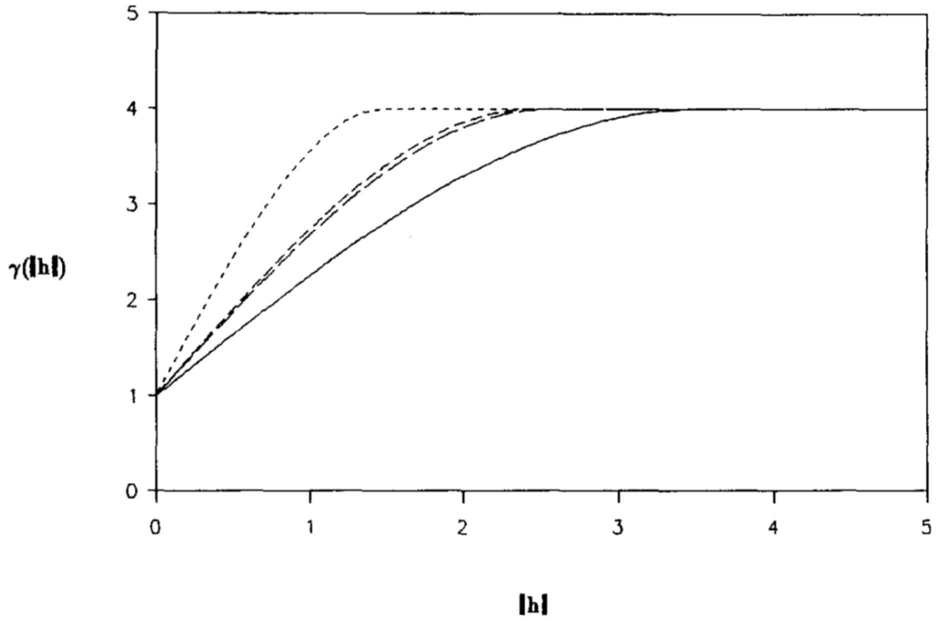


Figure 2.5: Geometric Range Anisotropy in two dimensional data: The solid line corresponds to the E-W direction and the N-S, NW-SE and NE-SW are given by dashed lines with dashes of increasing length. Taken from (Zimmerman, 1993)

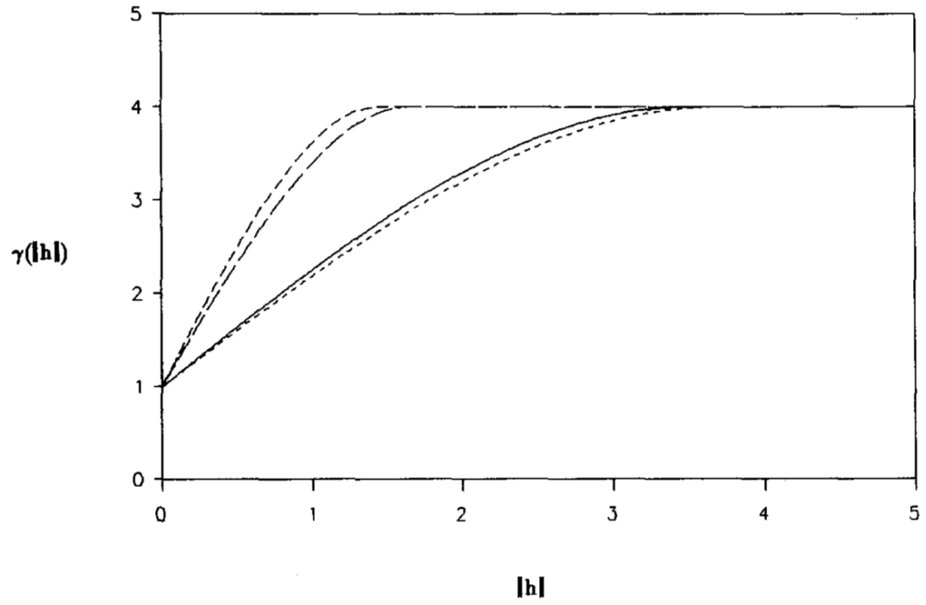


Figure 2.6: Non-Geometric Range Anisotropy in two dimensional data: The solid line corresponds to the E-W direction and the N-S, NW-SE and NE-SW are given by dashed lines with dashes of increasing length. Taken from (Zimmerman, 1993)

2.3.7.3 Nugget Anisotropy

Nugget anisotropy (Fig. 2.7) may indicate that the assumption of white noise measurement error is not appropriate. Ergo, a more general spatially dependent error should be modeled to tackle the problem (Zimmerman, 1993) Moreover, a direction-dependent nugget effect may be caused by the random process $Z(\mathbf{s})$ itself. It is up to modeler to assess which type is present.

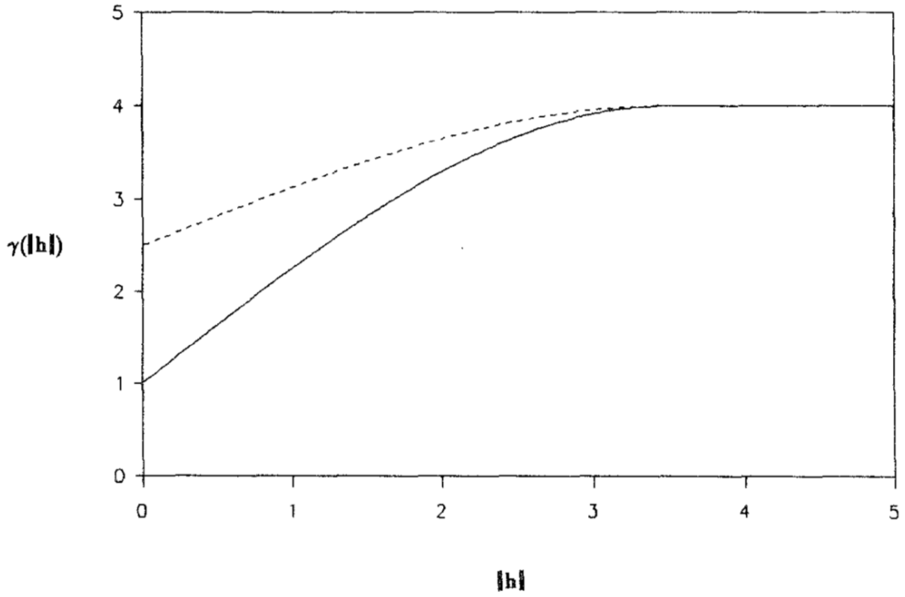


Figure 2.7: Nugget Anisotropy in two dimensional data: The solid line corresponds to the E-W direction and the N-S is given by dashed lines. Taken from (Zimmerman, 1993)

2.3.8 Variogram Interpretation

By recalling the definitions of the empirical variogram (2.46) and the empirical covariogram (2.50) and having in mind the relationships (2.44) and (2.45), then it can be stated that the empirical variogram should converge to the sill $\hat{C}(\mathbf{0}) = \sigma^2$, which corresponds to the variance of $Z(\mathbf{s})$ with respect to its mean.

Gringarten and Deutsch (2001) argue that should the empirical variogram continue to grow above this expected sill, then the assumption of stationarity,specially the assumption of a constant mean across the analyzed domain, is not adequate to the problem at hand. As can be seen in Fig. 2.8, variogram values above the sill correspond to negative correlation between the head values at the end of the lag vector and the tail values at the beginning of the lag. This contradicts the notion that after the distance of the *range*, values of the process $Z(\mathbf{s})$ should be uncorrelated.

The usual suspects to cause the aforementioned behaviour in the variogram are spatial trends, i.e. the mean is not constant across the analyzed domain, it depends on the position $\mathbf{s}$. An example of such a trend can be viewed in Fig. 2.9

Thus, to correct variograms that grow over the sill, the process $Z(\mathbf{s})$ should be analyzed to search for spatial trends. If a spatial trend $\mu(\mathbf{s})$ is present, it should be determined via

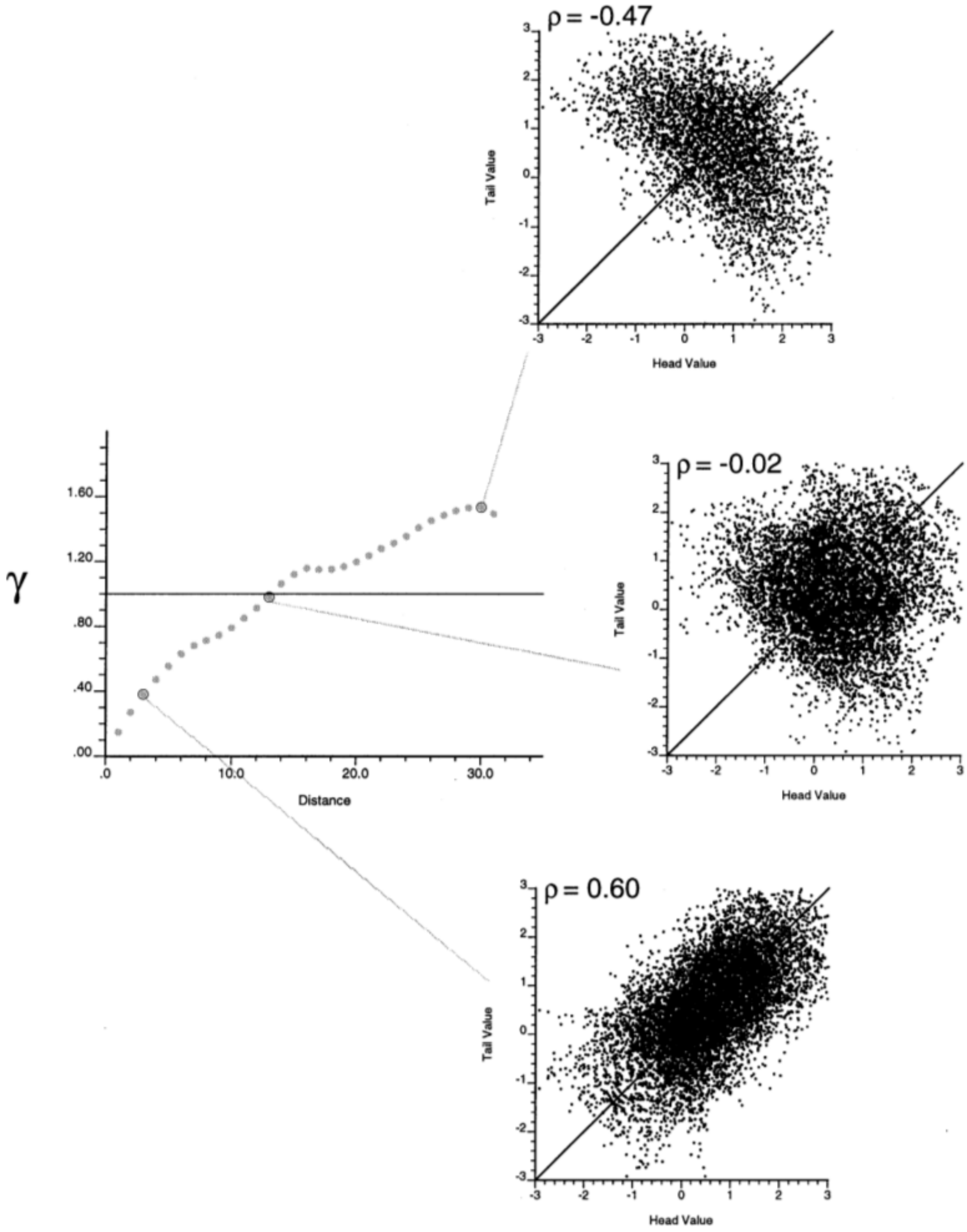


Figure 2.8: Empirical Semivariogram along with three **h**-*scatterplots*: It can be seen that below the sill the values at the end of the lag-vector are highly correlated with the values at the tail of the lag. Moreover, at the sill the values are uncorrelated and above the sill the values start to be negatively correlated. Taken from (Gringarten and Deutsch, 2001)

multidimensional curve fitting or similar methods. Once the trend has been estimated, the empirical variogram can be recalculated with the detrended data, i.e $Z(\mathbf{s}) - \mu(\mathbf{s})$. This new variogram should converge to the expected sill σ^2

The aforementioned process can be visualized in Fig. 2.10. With the detrended variogram

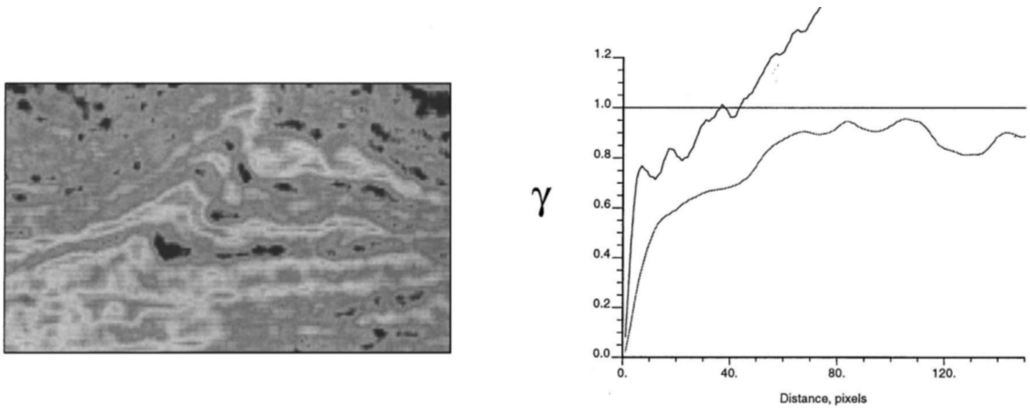


Figure 2.9: Empirical Semivariograms in the vertical and horizontal directions(Right) including the $Z(\mathbf{s})$ process, i.e. the scalar field on the left. The vertical semivariogram grows over the expected sill, which is caused by the trend in the vertical direction. Taken from (Gringarten and Deutsch, 2001)

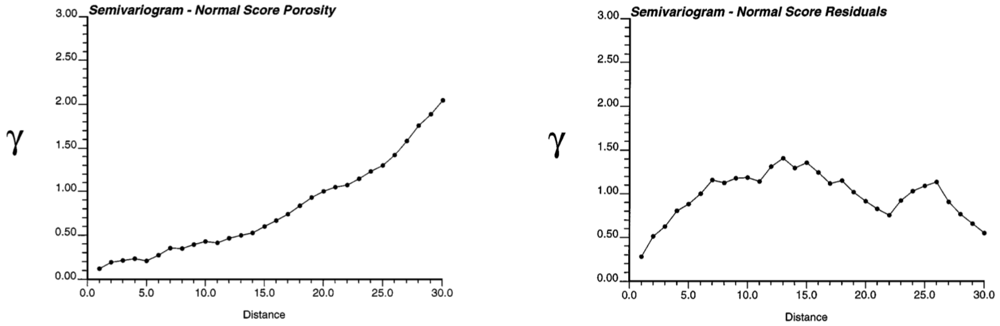


Figure 2.10: Empirical semivariograms before and after detrending. Here, the expected sill is one, since the data has been normalized w.r.t the mean and variance. Taken from (Gringarten and Deutsch, 2001)

and the known trend $\mu(\mathbf{s})$, spatial prediction can be implemented using the simple kriging equations of section § 2.3.2.

2.4 Spatio-Temporal Processes

Following the narrative of spatial processes, spatio-temporal processes can be defined as a stochastic process of the form (Cressie and Wikle, 2011):

$$\mathbf{Y}(\mathbf{s};t) : \mathbf{s} \in \mathbb{D}_{\mathbf{s}} \subset \mathbb{R}^d, t \in \mathbb{D}_t \subset \mathbb{R}, \tag{2.63}$$

where the domain of time is usually that of positive integers $\mathbb{Z}^+$. Again, the goal in this setting is to perform spatio-temporal prediction at an unknown space-time point $(\mathbf{s}_0, t_0)$, i.e spatio-temporal *Kriging.* To this end, some of the concepts presented in the Spatial Statistics section will be revisited

2.4.1 Stationarity in Time and Space

It was previously stated in § 2.3.3 that *stationarity* refers to the notion that statistical properties of stationary processes are invariant to translations. This assumption for the characterization of the random process $Y(\mathbf{s};t)$ is summarized again in two types of space-time stationarity, *second order or weak stationarity* and *intrinsic stationarity.*

As in the spatial setting, *second order stationarity* in time and space assumes that there is a constant mean μ that does not depend on $(\mathbf{s};t)$ and that the covariance of the process can be expressed via a stationary positive-definite function:

$$\operatorname{cov}(Y(\mathbf{s};t), Y(\mathbf{x};r)) = C(\mathbf{s}-\mathbf{x}; t-r) = C(\mathbf{h};\tau), \quad \mathbf{s},\mathbf{x},\mathbf{h} \in \mathbb{R}^d, t, r, \tau \in \mathbb{R}, \tag{2.64}$$

that only depends on the time-space distance, wherein this function is referred to as the covariance function (Cressie and Wikle, 2011).

Accordingly, time-space *intrinsic stationarity* assumes again a constant mean across the domain $(\mathbf{s};t)$, and that the variance of the differences of the process is bounded and only depends on the time-space displacement:

$$\operatorname{var}(Y(\mathbf{s};t) - Y(\mathbf{x};r)) = 2\gamma(\mathbf{s}-\mathbf{x}; t-r) = \gamma(\mathbf{h};\tau), \quad \mathbf{s},\mathbf{x},\mathbf{h} \in \mathbb{R}^d, t, r, \tau \in \mathbb{R}, \tag{2.65}$$

wherein γ is called the variogram. Similarly to the spatial case, these two functions share the relationship:

$$\gamma(\mathbf{h};\tau) = C(\mathbf{0};0) - C(\mathbf{h};\tau) \quad \mathbf{h} \in \mathbb{R}^d, \tau \in \mathbb{R}, \tag{2.66}$$

when the process is second order stationary in time and space. These two functions can be estimated from the data the same way as explained in § 2.3.4 by including the time coordinate. Furthermore, other properties such as *isotropy* can be naturally extended to the spatio-temporal case.

2.4.2 Spatio-Temporal Covariance Functions

While in spatial statistics the variogram is the preferred tool for modeling the behaviour of a spatial random process, in the spatio-temporal community covariance functions have established themselves as the modeling tool of preference. There is a wide range of covariance functions for spatio-temporal processess that focus on the different options to model the interaction between time and the spatial coordinates.

To ease the interpretation of the different options, Cressie and Wikle (2011) introduce spatio-temporal covariance functions by presenting that dynamic spatio-temporal stochastic processes, which are usually described with stochastic partial differential equation (SPDE), can also be reproduced via spatio-temporal covariance functions.

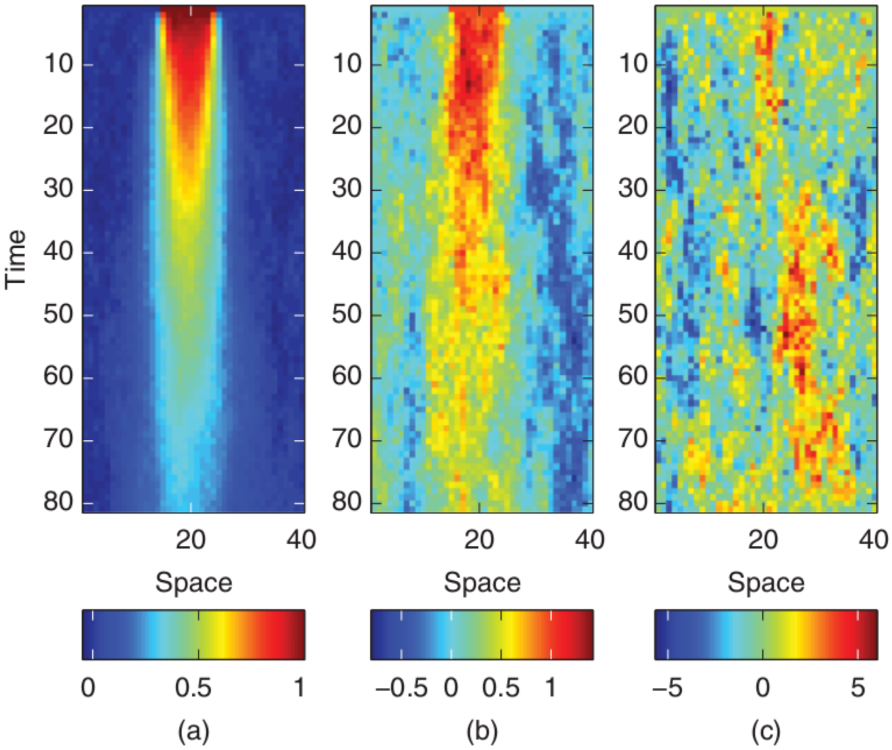


Figure 2.11: Three simulations based on (2.67) with the same $\alpha = 1$,$\beta = 20$ and starting values $Y(s, 0) = 1$ for $15 \leq s \leq 24$, and δ is mean-zero white noise with **(a)** $\sigma_\delta = 0.01$, **(b)** $\sigma_\delta = 0.1$, **(c)** $\sigma_\delta = 1$. Taken from (Cressie and Wikle, 2011)

Based on the work ofHeine (1955), Cressie and Wikle (2011) simulate the stochastic diffusion model of Fig. 2.11 by applying the stochastic PDE numerically, which in the one dimensional case has the structure:

$$\frac{\partial Y(s;t)}{\partial t} - \beta \frac{\partial^2 Y(s;t)}{\partial s^2} + \alpha Y(s;t) = \delta(s;t), \tag{2.67}$$

where $\alpha, \beta > 0$ and $\delta(s;t)$ is an error process that accounts for the non-deterministic behaviour, usually assumed as white noise. The aforementioned dynamical process has the following covariance function:

$$\begin{aligned} C_Y(h,\tau) = \frac{1}{2}\sigma^2 \Bigg\{ & e^{-h(\alpha/\beta)^{1/2}} Erfc\left(\frac{2\tau(\alpha/\beta)^{1/2} - h/\beta}{2(\tau/\beta)^{1/2}}\right) \\ & + e^{h(\alpha/\beta)^{1/2}} Erfc\left(\frac{2\tau(\alpha/\beta)^{1/2} + h/\beta}{2(\tau/\beta)^{1/2}}\right) \Bigg\} \end{aligned} \tag{2.68}$$

where $Erfc$ is the *complementary error function*. Using the aforementioned covariance function, Cressie and Wikle (2011) simulated the stochastic process via a zero-mean Gaussian process, which yielded the results seen in Fig. 2.12.

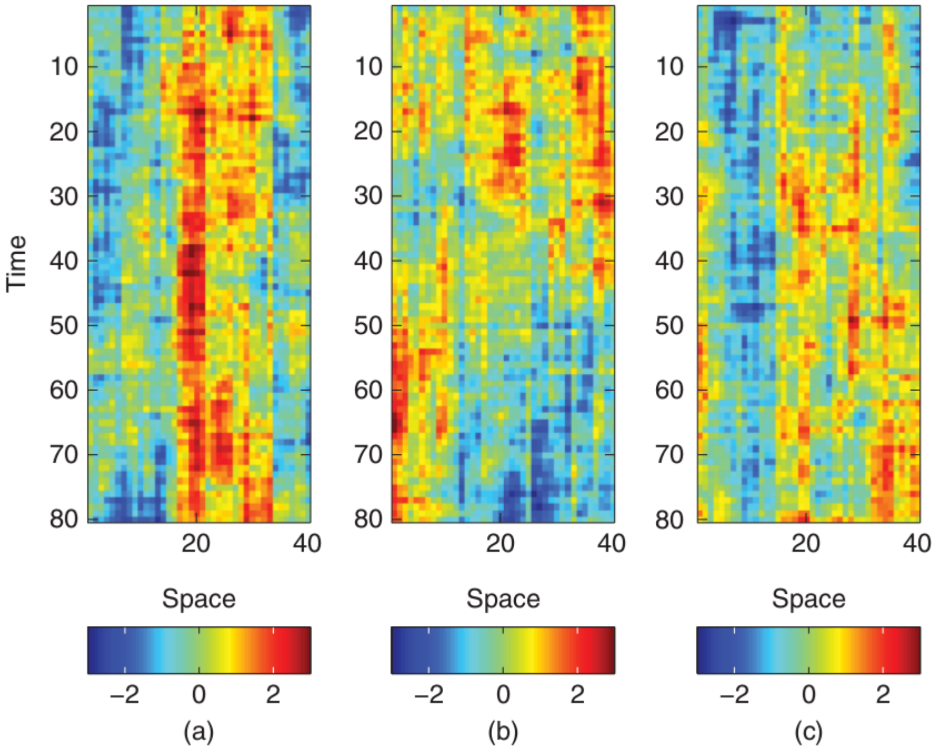


Figure 2.12: Three independent simulations drawn from a zero-mean Gaussian process with covariance function (2.68) and the same $\alpha = 1$, $\beta = 20$ as in Fig. 2.11. Taken from (Cressie and Wikle, 2011)

It can be seen that the structure of the process in Fig. 2.12 is quite similar to the one observed in Fig. 2.11 for the cases with larger white noise.

Having the aforementioned example in mind, the following segments will present the different ways avalaible to model space-time interactions via the covariance function.

Time as a Metric

The most straightforward space-time interaction model is to handle time as another spatial dimension, i.e. the time displacement is modelled as a spatial distance (Montero et al., 2015). This reduces to adding $|\tau|$ to the covariance function argument of the stationary spatial case:

$$C(\mathbf{h}; \tau) = C(|\mathbf{h}| + |\tau|) \quad \mathbf{h} \in \mathbb{R}^d, \tau \in \mathbb{R}. \tag{2.69}$$

This simplification may not always be appropriate but it permits the handling of common problems, such as anisotropy, with the established approaches of Geostatistics.

Separable Covariance Functions

A separable spatio-temporal covariance function assumes that the random process has the following form of covariance function (Cressie and Wikle, 2011):

$$\mathrm{cov}(Y(\mathbf{s};t),Y(\mathbf{x};r)) = C(\mathbf{h};\tau) = C^{(s)}(\mathbf{h}) \cdot C^{(t)}(\tau), \quad \mathbf{h} \in \mathbb{R}^d, \tau \in \mathbb{R}, \tag{2.70}$$

where $C^{(s)}$ and $C^{(t)}$ are stationary spatial and temporal covariance functions, respectively. Therefore, the computational advantage of assuming a separable process is that the spatial and temporal structures can be modeled separately. Moreover, in the case of the Square Exponential and Exponential covariance function, modeling the space-time interaction with the method of *Time as a Metric* or with the *Separability* assumption yields the same results.

The downside of separable covariance functions is that they exhibit properties that may not be appropriate for many spatio-temporal settings, e.g. the following holds: $C(\mathbf{h}_1;\tau) \propto C(\mathbf{h}_2;\tau)$ and likewise: $C(\mathbf{h};\tau_1) \propto C(\mathbf{h};\tau_2)$. This means that each time-series shares *exactly the same* cross-correlational properties with any other time-series at any displaced location and vice-versa for the spatial properties.

Sums and Products of Covariance Functions

As a natural extension of the separable case, new covariance functions can be created by adding a separable spatio-temporal covariance function with more spatials or temporal covariance functions, for example:

$$C(\mathbf{h};\tau) = pC^{(s_1)}(\mathbf{h}) \cdot C^{(t_1)}(\tau) + qC^{(s_2)}(\mathbf{h}) + rC^{(t_1)}(\tau), \quad \mathbf{h} \in \mathbb{R}^d, \tau \in \mathbb{R}. \tag{2.71}$$

This type of covariace functions, along with the separable ones, are part of the family of *symmetric covariance functions*, which fulfill the following condition:

$$\mathrm{cov}(Y(\mathbf{s};t),Y(\mathbf{x};r)) = \mathrm{cov}(Y(\mathbf{s};r),Y(\mathbf{x};t)), \quad \mathbf{s},\mathbf{x} \in \mathbb{R}^d, t,r \in \mathbb{R}. \tag{2.72}$$

Imagining Y as a maximum-minimum daily temperature variable, symmetry assumes that the covariance between yesterday's temperature in Toulouse and today's temperature in Paris is equivalent to the covariance between today's temperature in Toulouse and yesterday's temperature in Paris. This is unlikely to be the case, considering the difference in latitude between both cities.

Non-Separable Covariance Functions

Non-separable covariance functions do not fulfil condition (2.70). An example of such a covariance function is (2.68) and the following one proposed by Cressie and Wikle (2011):

$$C(\mathbf{h};\tau) = \sigma^2 \exp\left\{-b^2|h|^2/(a^2\tau^2+1)\right\}/(a^2\tau^2+1)^{d/2} \quad \mathbf{h} \in \mathbb{R}^d, \tau \in \mathbb{R}, \tag{2.73}$$

where $a, b \geq 0$ and $\sigma^2 = C(\mathbf{0}; 0)$. This type of covariance function try to overcome the shortcomings of the assumption of *separability* and *symmetry*, but increase model complexity. A comparison between a separable and non-separable covariance function can be visualized in Fig. 2.13.

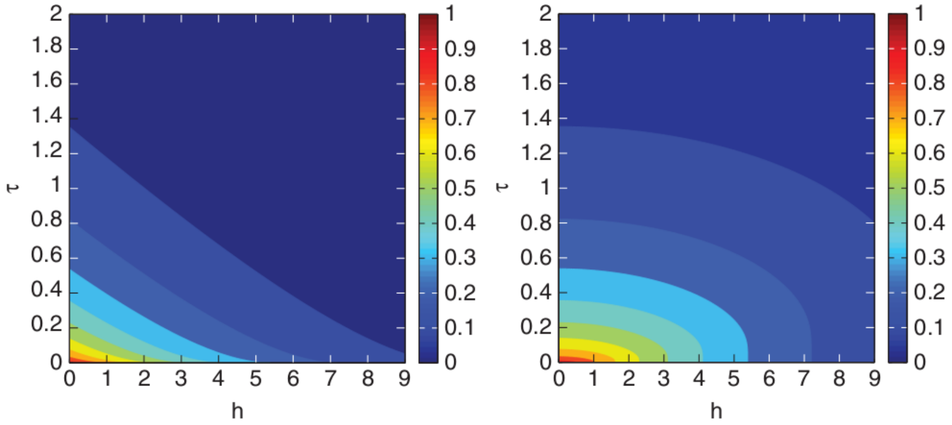


Figure 2.13: Comparison of a separable(left) and non-separable(right) covariance function. The non-separable case corresponds to (2.68). Taken from (Cressie and Wikle, 2011)

Cressie and Wikle (2011, § 6) added that there have been several proposals to test for *symmetry* or *separability* and recommends these types of tests to assess the properties of the covariance function with which to model the spatio-temporal process at hand.

Taylor's Hypothesis

A spatio-temporal covariance function satisfies *Taylor's Hypothesis* if there exists a velocity vector $\mathbf{u}_0$ such that:

$$C(\mathbf{0}; \tau) = C(\mathbf{u}_0\tau; 0), \quad \tau \in \mathbb{R}, \tag{2.74}$$

is fulfilled. Hence, a stationary spatio-temporal covariance function of the form $C(\mathbf{h}; \tau) = C^{(s)}(\mathbf{h} - \mathbf{u}_0\tau)$, where $C^{(s)}$ is a purely spatial covariance, is a valid spatio-temporal covariance function and it is referred to as *Taylor's frozen field*, because it models the process as if it were a frozen spatial field that is moving with constant velocity. One can see that the aforementioned covariance function is non-separable.

Moreover, Cressie and Wikle (2011) show that spatio-temporal covariance functions built with the method *Time as a Metric* or by assuming *separability* satisfy Taylor's hypothesis for some idealized velocity vectors $\mathbf{u}_0$ that have the same space-time interaction as in the mentioned modeling methods.

2.4.3 Spatio-Temporal Kriging

As a natural extension to the spatial setting, variogram models can be fitted to the empirical ones obtained from the spatio-temporal data. Once the spatio-temporal covariance or variogram model is known, *Kriging* can be implemented following the same line as in § 2.3.2 in order to predict $Y(\mathbf{s}; t)$ at unknown locations $(\mathbf{s}_0; t_0)$

Chapter 3

Research Approach

3.1 Critique to the State of the Art and Research Goals

The current implementation of the dense cloud mapping of thermodynamical variables used in the SkyScanner project can be considered as a standard “out of shelf” Gaussian process regression, which consists of assuming a zero-mean process and using the widespread squared exponential covariance function (2.17) with added white noise parameter. The hyperparameters of this model are learned by optimizing the Bayesian marginal likelihood, as presented in § 2.2.3.

This approach certainly does not exploit the full machinery that Gaussian processes can offer. Furthermore, the hyperparameter optimization is computationally expensive and jeopardizes the real-time constraint put on the mapping, which expects new mappings of the vicinity of the drones at a rate of one per 10 seconds.

The author considers that the aforementioned approach can be improved by incorporating prior knowledge to the Gaussian process scheme and thus, the goal of the research project will concentrate on improving Gaussian Process Regression through the incoporation of prior knowledge. The types of prior knowledge and how these priors could be determined will be discussed in the following segment.

3.2 Incorporating Prior Knowledge

The author considers that there are three distintive ways to incorporate prior knowledge to the Gaussian process, which are:

- Determining type, hyperparameter distribution and space-time interaction of the covariance function.
- Incorporating a mean function.
- Exploiting correlation between output variables.

These approaches will be briefly discussed in the following segments.

Prior knowledge of the covariance structure

To determine prior knowledge about the structure of the covariance function, the author recognizes the possibility of implementing two separate techniques. The first consists of a “brute force” approach, where experiments of the mapping, including hyperparameter optimization via Marginal Likelihood, would be performed on several types of covariance functions. The best model would be chosen on the basis of cross-validating the RMSE of the predictions across the various experiments performed on different clouds.

The second approach would be variogram based. The data of the clouds in the atmospheric simulation would be used to determine empirical variograms in the four main directions t, z, x, y and these variograms would later be fitted through WLS § 2.3.6. The hyperparameters of the variogram model with the least amount of fitting error in all directions would later be used to test the Gaussian process along with the current “off-the-shelf” implementation to compare the performance in terms of prediction RMSE.

The author favors this latter approach, for the calculation of the empirical variograms itself permits to assess graphically several assumptions, as elaborated in § 2.3.8. Of particular interest is the assumption of second-order stationarity, which requires the absence of a spatio-temporal trend, i.e. the absence of a spatio-temporal mean function.

Therefore, this work will focus on implementing the variogram based approach, where the goal is to test several types of variogram models to determine which one is best suited. Moreover, another aim will be to attempt to determine a prior distribution of the hyperparameters for the best model. With this distribution, the hyperparameter optimization with marginal likelihood could be accelerated in comparison to the current implementation, which uses a uniform prior.

Incorporating a Mean Function

By following the concept in § 2.2.4, a mean function can be included in GPR. Nonetheless, determining the mean function itself, if any, will require some ingenuity and keen eye from the author when exploring the data.

Non-converging empirical variograms, as explained in § 2.3.8, can also be an indication to further explore the possibility that the process possesses a mean function. Thus, the variogram analysis advocated in the previous section can be complemented by the search of a mean function, if the variograms were to justify such a search.

Exploiting correlation between output variables

As presented in § 2.2.5, output correlations could be exploited to improve the mapping performance. Furthermore, retrieving priors for output correlations could greatly flexibilize the mapping task by reducing sensor payload for the fleet of drones. This payload reduction would arise from the fact that each drone could specialize in sampling a single atmospheric variable but thanks to the priors of the output correlations a map of all atmospheric variables could still be better estimated than in the case without exploiting the correlation.

3.3 Methodology

The author considers that the core of the research project should lie on obtaining the prior knowledge related to the covariance structure, since the covariance function is the main ingredient to implement Gaussian process regression.

Furthermore, the estimation of the variogram for this end will deliver valuable insights of the process being analyzed, such as validation for some of the assumptions made to perform Gaussian process mapping. Initially, the author will concentrate on the vertical wind component inside of clouds, as it is the most relevant variable for the planning algorithm and the one with the largest amplitude.

3.3.1 Meso-NH Simulation

To validate the mapping and planning algorithms for the SkyScanner project, realistic cloud simulations are required: this is provided by atmospheric models, that can simulate the microphysical and dynamical properties of clouds.

The atmospheric model used for the current work is Meso-NH Lafore et al. (1998). This model is the result of the joint collaboration between the national center of meteorological research (CNRM, Météo-France) and Laboratoire d'Aéorologie (LA, UPS/CNRS). Meso-NH is a Non-Hydrostatic Model with the flexibility to simulate atmospheric phenomena at a wide range of resolutions that extends from one meter up to tens of kilometers. For this research project, non-precipitating shallow cumulus clouds over land are simulated with the LES (Large-eddy simulations) version of Meso-NH, with resolutions down to ten meters. The simulation was driven by realistic initial conditions obtained on June 21, 1997 from meteorological measurements at the Southern Great Plains site in Oklahoma, U.S.A Brown et al. (2002). This site is the first field measurement site established by the Atmospheric Radiation Measurement (ARM) Program.

To capture more details about clouds and their surroundings, it is preferable to set the atmospheric model at its highest resolution. The considered simulation domain is a cube of 400x400x161 grid points representing a volume of $4\,\mathrm{km} \times 4\,\mathrm{km} \times 4\,\mathrm{km}$ with horizontal resolutions of $dx = dy = 10\,\mathrm{m}$, vertical resolutions from $dz = 10\,\mathrm{m}$ to $100\,\mathrm{m}$ and a time-step of $0.2\,\mathrm{s}$. This setup is a compromise between the desired high resolutions and a reasonable simulation computation time.

The 161 vertical levels have a high resolution of $10\,\mathrm{m}$ in both convective cloud and surface layers; in the upper cloud-free troposphere, the domain has stretched resolutions from $10\,\mathrm{m}$ up to $100\,\mathrm{m}$. The upper five layers of the simulation domain act as a sponge layer to prevent wave reflection. In addition, the horizontal boundary conditions are cyclic with a periodicity equal to the horizontal width of the simulation domain.

The simulation estimates the following atmospheric variables: cloud LWC, water vapor, pressure, temperature, and the three components of wind. Fig. 3.1 illustrates the 3D cloud water content of convective cumulus clouds at a given time. The overall simulation covers a time period of 15 hours, but variables of interest have been saved every second only during one hour that corresponds to the maximum of surface fluxes.

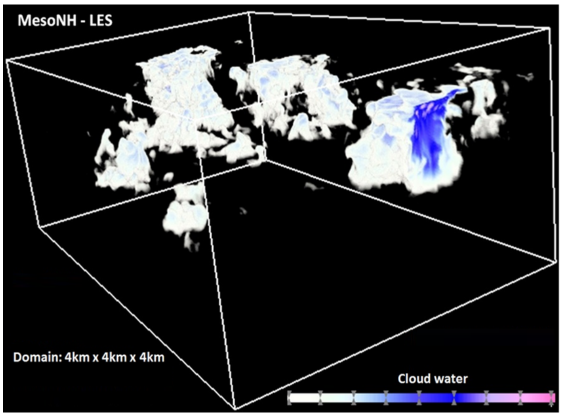


Figure 3.1: Meso-NH LES simulation: liquid cloud water content of the cumulus formed at 1h30 PM (ARM Southern Great Plains, June 21, 1997 conditions)

3.3.2 Sampling Design, Empirical Variogram Estimation and Variogram Fitting

Since the Meso-NH simulation results are stored in equally spaced data grids across the relevant domain to be analyzed, estimating the empirical variogram in each of the four directions can be done in a straightforward manner. Since the resolution across the relevant domain of clouds is 10 m for spatial dimensions and 1 s for time, the empirical variogram can be estimated by taking the data points that are displaced by integer steps of the resolution in each direction, in other words, by exploiting the inherent structure of the indexing of the data in the four directions using the equations presented in § 2.3.4. Therefore the variogram support for the spatial dimensions will be 10 m and the support of time will be 1 s.

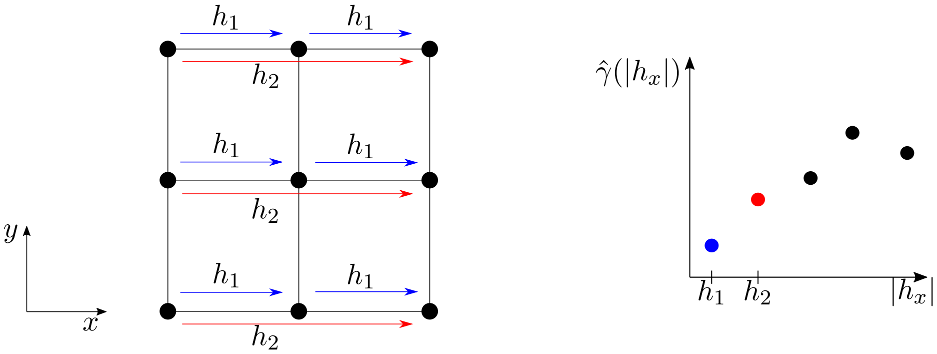


Figure 3.2: Determining the Empirical Variogram in a equally spaced data grid. This can be easily extended for each main direction of the data.

The aforementioned approach is illustrated in Fig. 3.2. Furthermore, with the information of the LWC, the analyzed variable can be filtered to only take into account values that are inside of the cloud.

The goal is to estimate empirical variograms for several clouds and later fit them with models through WLS § 2.3.6. For this introductory work, the following covariance functions and type of space-time interaction will be considered (For a review, consult § 2.3.5,§ 2.4.2):

- Separable space-time Squared Exponential covariance function
- Separable space-time Exponential covariance function
- Time as a Metric space-time Matérn covariance function with $\nu = 3/2$
- Time as a Metric space-time Matérn covariance function with $\nu = 5/2$

The fitting error of this four alternatives will be compared and the best option along with the hyperparameter distribution will be used to test this new covariance function against the “off the shelf” one in different experimental settings.

Requirements

Estimating the empirical variogram and performing the experiments as explained in the previous section requires the implementation of a cloud segmentation algorithm that permits to track and obtain relevant geometrical information of clouds, such as center of geometry, center of mass of relevant thermodynamical variables, cloud skeleton, among others. The basis for obtaining this information is the LWC variable of the Meso-NH simulation.

3.4 Design of Experiments for the Gaussian Process Regression

The new covariance function obtained with the variograms approach will be tested against the “off the shelf” option in three experiments using the atmospheric simulation. To reduce complexity, the trajectories considered for these tests will be systematic, i.e. the adaptive sampling scheme of the SkyScanner project will be turned off.

The conditions for the three experiments considered will be explained in the next segments.

Experiment One: Static Cross-sections

The first experiment will consist of five drones following circular trajectories along several static cross-sections of a cloud(Example in Fig. 3.3), i.e. time is static, the vertical wind scalar field will be 2-dimensional, since the drones are flying at the same height, and furthermore, the scalar field remains unchaged while being sampled by the drones. It will be assumed that the drones sample a point of the scalar field each second and that its speed will be $15\,m/s$ along the circle. The sensor noise will be modeled by adding mean-zero Gaussian white noise to the samples.

The sampled data including noise will be the training basis to build the models for the different covariance functions. The idea is to compare the “off the shelf” covariance function against three variants of the new covariance function obtained with the variograms, including mean function, if present.

The first alternative will have frozen hyperparameters, namely, the ones obtained with the variogram. The second option will take the same structure as the new covariance function, but the hyperparameters will be optimized without any prior knowledge. The

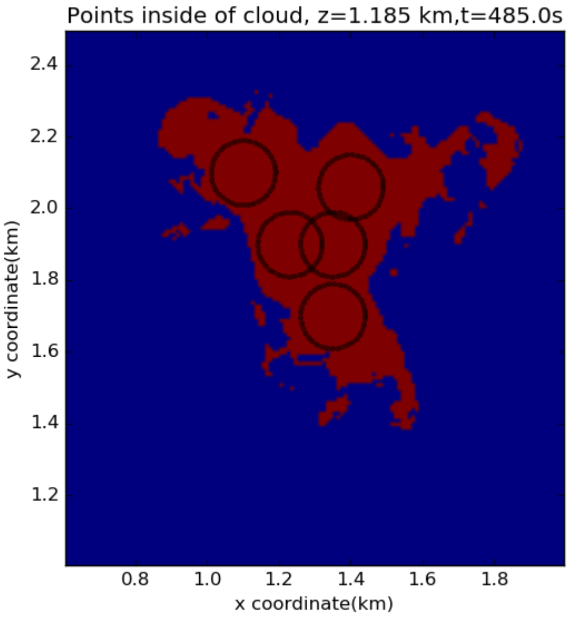


Figure 3.3: Exemplary circular trajectories to sample data of a cloud cross-section.

third and last option will be optimized as in the second option, but will use the prior knowledge about the hyperparameters obtained from the variogram analysis.

Thus, the obtained models will be tested against the values of all points inside the analyzed frozen cross-section to determine RMSE and other metrics. Furthermore, the effect of the standard deviation of the white noise will also be studied.

Experiment Two: Dynamic Cross-section

The second experiment will also consist of five drones doing circular trajectories along several cross-sections, but in this case the cross-section will be dynamic. It will be again assumed that the drones sample a point of the scalar field each second and that its speed will be $15\,m/s$ along the circle. The sensor noise will as before be modeled by adding mean-zero Gaussian white noise to the samples of said circular trajectories.

Moreover, the rest of the conditions will be the same as explained in the static case, thus having the goal of comparing the “off the shelf” alternative with the three versions of the covariance function based on the variogram analysis.

The obtained models will be tested against the vertical wind values of all points inside the cross-section at the time point of the last samples of the five drones. Under this configuration, RMSE and other metrics will be computed for all tested models. Herein, the effect of the standard deviation of the white noise will be studied as in the static case.

Experiment Three: Entire Cloud

The third experimental setting will be comprised of five drones doing helicoidal trajectories to sample vertical wind values across the cloud in its entirety, under realistic dynamical conditions. It will be assumed again that the drones can sample a data point each second. As for the helix trajectory, each drones will have a vertical climb rate of $2\,m/s$ and a horizontal velocity of $10\,m/s$ along the "projected" circle of the helix.

The sensor noise will be idealized again as mean-zero Gaussian white noise that is added to the samples of the drones. Moreover, the rest of the conditions remain the same as explained in the previous two experiments, therefore having the goal of comparing the "off the shelf" alternative with the three versions of the covariance function based on the variogram analysis.

The obtained models will be tested against the vertical wind values of all points inside the height bounds of the cloud defined by the lowest and highest data point of sampled trajectories. Only the time point of the last samples of the five drones will be considered for the test. Under this configuration, RMSE and other metrics will be computed for all tested models. Herein, the effect of the standard deviation of the white noise will be studied as in the two previous cases.

Chapter 4

Implementation

As presented in the Research Approach, this chapter deals with the results of determining prior knowledge in terms of the covariance structure. It includes the initial exploration to find approapriate clouds, the corresponding variogram analysis and the experiments to compare the new approach versus the “off the shelf” Gaussian Process approach. As stated before, the work will focus on the vertical wind component.

Moreover, the work of the following sections was performed using the programming language *Python* version 3.5, by using packages such as Numpy and Scipy and software developed internally by the SkyScanner Project to manipulate the simulation data, which is stored in NetCDF format, a data format designed to store grid-form scientifical data.

4.1 Initial Data Exploration

During the initial data exploration, the goal was to find useful clouds among the three terabytes of simulation comprising one hour across the 4 km cube. The version of the simulation used was configured to have zero-mean horizontal winds, i.e. the clouds did not move significantly in the horizontal directions across the cube.

The liquid water content variable determines the presence or not of a cloud. If this variable exceeds the value of 10^{-5} kilograms of liquid water per kilogram of air, then the corresponding data point can be considered as part of a cloud.

With this knowledge, the liquid water content was binarized to signal the presence or not of clouds and thus it could be determined that the majority of the clouds across the 4 km cube started at a height of around 1km. Accordingly, the entire one hour of simulation was animated at this height to visualize the cross-sections of the clouds available.

The animation enabled determining the rough bounding boxes of five clouds that had 400m of diameter or greater. These bounding boxes served as a restriction for the segmentation algorithm, with which all points inside the five clouds could be determined. The next section will present some of the properties of these clouds.

4.1.1 Cloud Segmentation and Visualization

Cloud One

The rough bounding box of cloud one has the following coordinates:

- $x_{start} = 0.605\,\text{km}$ and $x_{end} = 2.005\,\text{km}$
- $y_{start} = 1.105\,\text{km}$ and $y_{end} = 2.505\,\text{km}$
- $z_{start} = 0.985\,\text{km}$ and $z_{end} = 1.485\,\text{km}$.

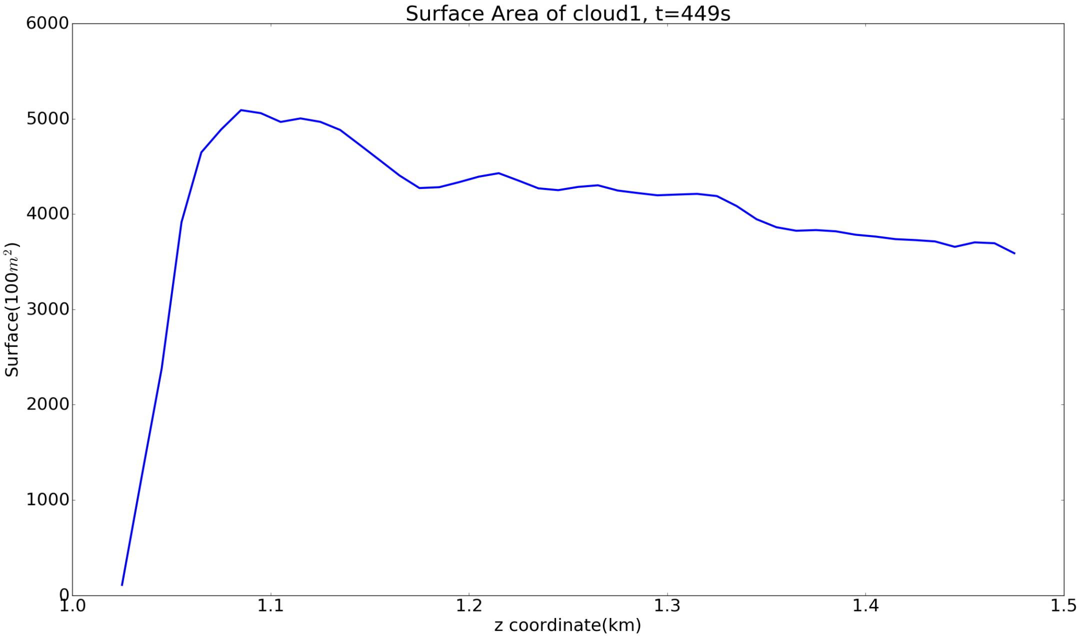
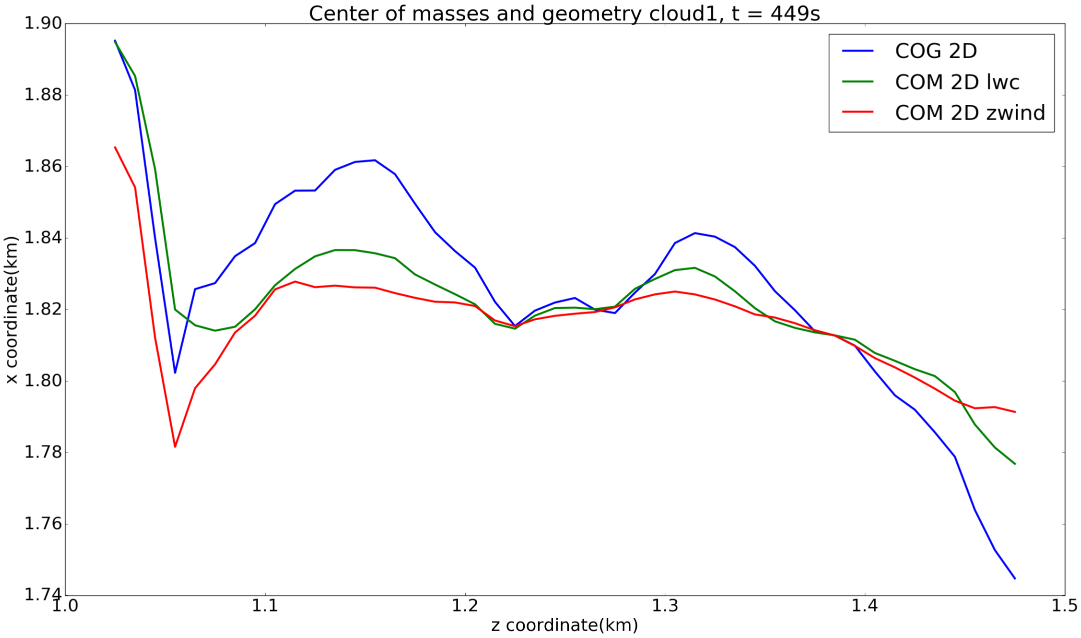


Figure 4.1: Surface area of cloud one per each cross-section, at time point $t = 449\,\text{s}$.

Herein, the cloud actually started at a height of about 1.035 km (Fig. 4.1) and spanned further than the end height selected. This end height was selected since it corresponds to the last grid point of the simulation where the resolution remains 10m for the z-direction. Since the variogram analysis uses the inherent grid structure, the variable resolutions above $z_{end} = 1.485\,km$ cannot be handled. The starting time point for the cloud segmentation algorithm started at $t = 449\,\text{s}$ and ended 150s later.



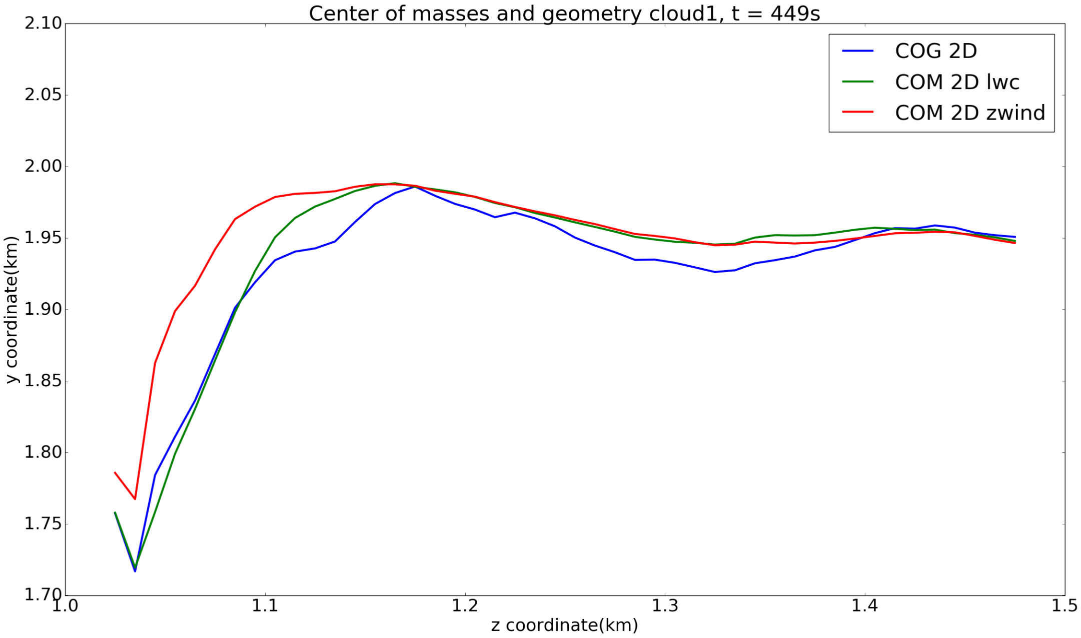


Figure 4.2: cloud one. Plots of the x and y coordinates of the center of masses of vertical wind and LWC. Also included the center of geometry in dependence of the height(Cloud Skeleton). These were all calculated at the same time point $t = 449\,\text{s}$.

Thanks to the segmentation algorithm, all points inside cloud one during the 150s timespan could be determined. Using these points, the cloud skeleton, i.e the centers for each cross-section of the cloud, can be calculated. There are three alternatives to define the aforementioned skeleton:

- **Center of geometry**: all coordinates of points inside each cross-section are averaged without weights, i.e. the mass at each point is one.

- **Center of liquid water content**: all coordinates of points inside each cross-section are used to compute an average weighted by the values of the liquid water content at each of the points. This can be viewed as the actual center of mass of the cloud in terms of water mass.
- **Center of vertical wind**: all coordinates of points inside each cross-section are used to compute an average weighted by the values of the vertical wind at each of the points. This corresponds to a center of "mass" in terms of the vertical wind.

Using the notions explained above, an example of the cloud skeleton in terms of each of the three centers is shown in Fig. 4.2 for time point $t = 449\,s$. Moreover, an instance of the three centers can be visualized in Fig. 4.3. It can be seen that all three alternatives to compute the centers follow similar behaviour, specially the center of masses of the vertical wind and the liquid water content.

Furthermore, using the geometric information of the cloud acquired with the segmentation and tracking algorithm, the vertical wind field inside the cloud can be visualized. An example of two cross-sections at time point $t = 449\,\mathrm{s}$ can be viewed in Fig. 4.3.

The previous vertical wind cross-sections permit to visualize the behaviour of the wind in the xy-plane. In order to visualize the behaviour of the vertical wind in the z-direction, the author calculated the mean and maximum of each cross-section. This behaviour is portrayed in Fig. 4.4 for time point $t = 449\,\mathrm{s}$.

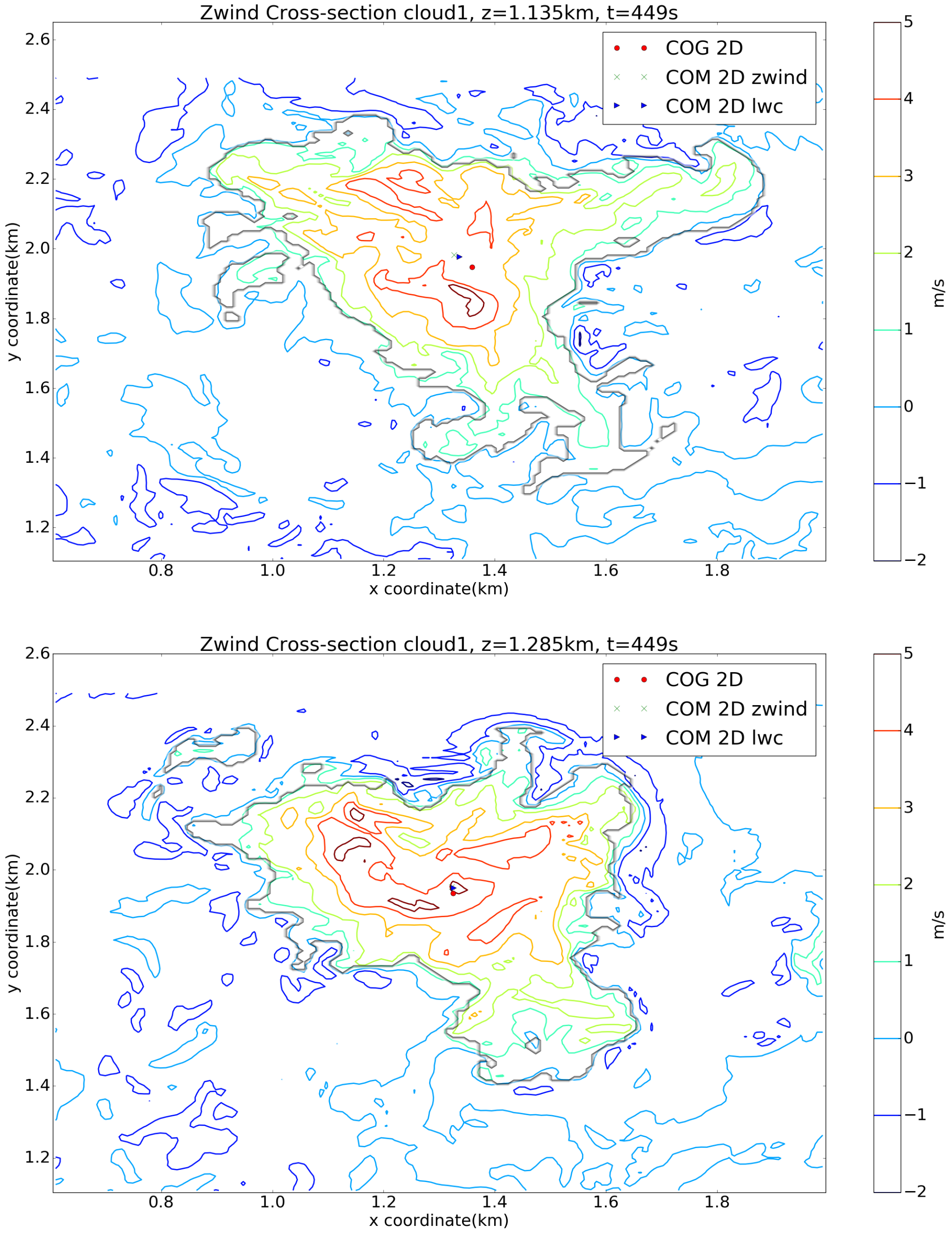


Figure 4.3: Two contour plots of the vertical wind of two cross-sections of Cloud1, 150m apart in height and at the same time point $t = 449\,\mathrm{s}$. The outer contour corresponds to the boundaries of the cloud. Furthermore, the center of geometry and the center of masses of the vertical wind and liquid water content are included.

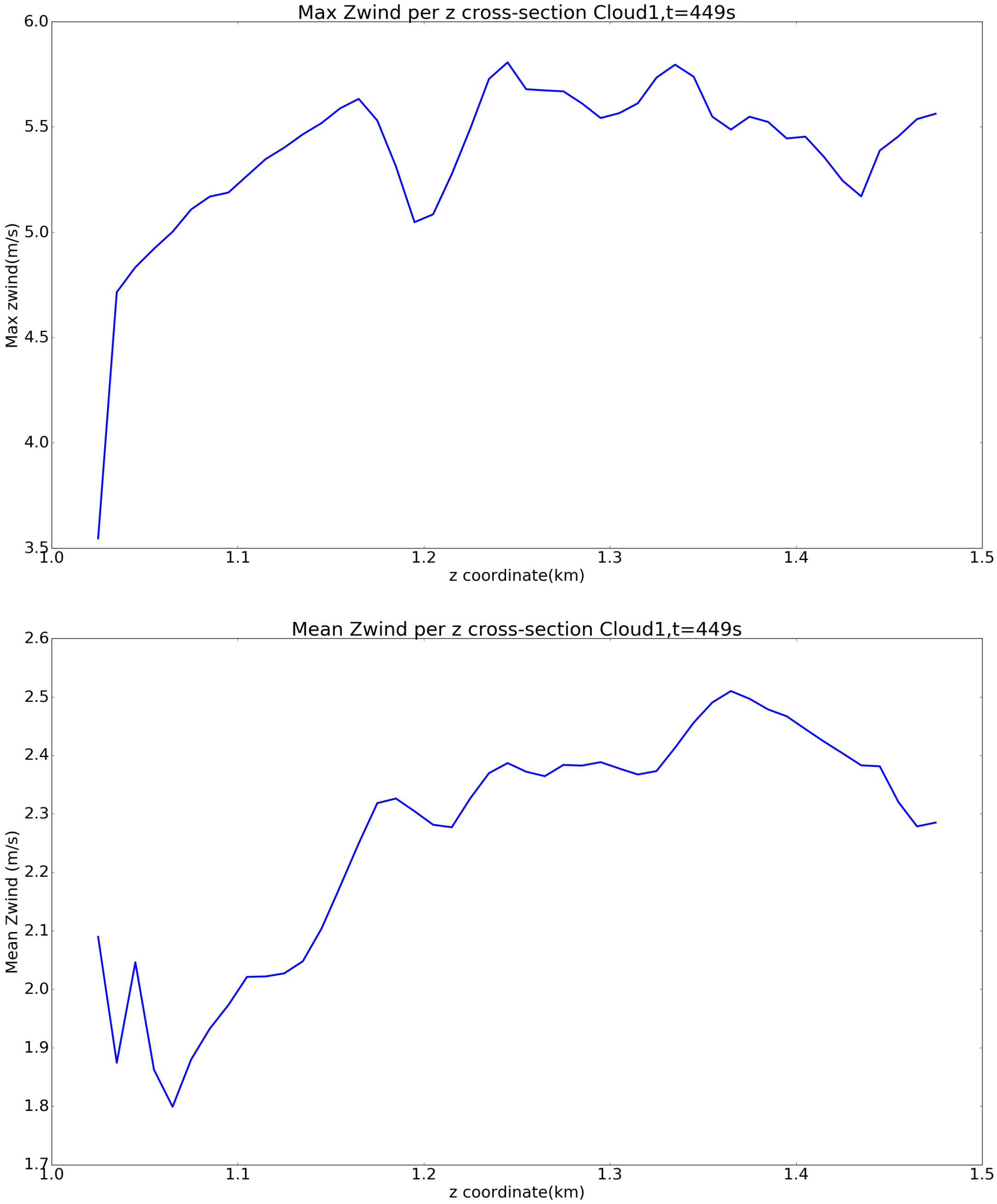


Figure 4.4: Plots of the evolution of the mean and maximum of the vertical wind for each cross-section , at time point $t = 449\,\mathrm{s}$.

Cloud Two

The bounding box of cloud two that served as the input for the segmentation algorithm has the following proportions:

- $x_{start} = 1.255\,\text{km}$ and $x_{end} = 2.505\,\text{km}$
- $y_{start} = 2.705\,\text{km}$ and $y_{end} = 3.705\,\text{km}$
- $z_{start} = 0.985\,\text{km}$ and $z_{end} = 1.485\,\text{km}$.

The starting time point for the cloud segmentation algorithm was set at $t = 929\,\text{s}$ and ended 150s later. The same type of analysis as in cloud one was repeated for cloud two. The corresponding plots can be viewed in appendix § A.1.

Cloud Three

The bounding box of cloud three that was used as the input for the segmentation algorithm has the following coordinates:

- $x_{start} = 0.005\,\text{km}$ and $x_{end} = 1.155\,\text{km}$
- $y_{start} = 0.505\,\text{km}$ and $y_{end} = 2.005\,\text{km}$
- $z_{start} = 0.985\,\text{km}$ and $z_{end} = 1.485\,\text{km}$.

The starting time point for the cloud segmentation algorithm started at $t = 1529\,\text{s}$ and ended 150s later. The same line of analysis as in cloud one was repeated for cloud three. The resulting plots can be viewed in the appendix § A.1.

Cloud Four

The bounding box of cloud four utilized as the input for the segmentation algorithm has the following coordinates:

- $x_{start} = 3.105\,\text{km}$ and $x_{end} = 3.995\,\text{km}$
- $y_{start} = 0.005\,\text{km}$ and $y_{end} = 1.255\,\text{km}$
- $z_{start} = 0.985\,\text{km}$ and $z_{end} = 1.485\,\text{km}$.

The starting time point for the cloud segmentation algorithm started at $t = 2789\,\text{s}$ and ended 150s later. The same line of analysis as in cloud one was repeated for cloud four. The resulting plots can be viewed in the appendix § A.1.

Cloud Five

The rough bounding box of cloud five utilized as the input for the segmentation algorithm has the following coordinates:

- $x_{start} = 0.005\,\text{km}$ and $x_{end} = 2.005\,\text{km}$
- $y_{start} = 2.005\,\text{km}$ and $y_{end} = 3.995\,\text{km}$
- $z_{start} = 0.985\,\text{km}$ and $z_{end} = 1.485\,\text{km}$.

The starting time point for the cloud segmentation algorithm started at $t = 3389\,\text{s}$ and ended 150s later. The same line of analysis as in cloud one was repeated for cloud five. The resulting plots can be viewed in the appendix § A.1.

Summary of Cloud visualization

The cloud segmentation algorithm enabled determining the geometry of five different clouds. With this information, some of the characteristics of the cloud could be visualized, such as their skeleton in terms of geometry, liquid water mass and vertical wind.

Moreover, typical cross-sections of the five clouds including their vertical wind fields among other interesting information such as the evolution of surface area in dependence of height, could be visualized.

Although the following sections will perform analysis on the five clouds presented above, for the sake of readability only the results of cloud one will be presented in the main document. The rest of the analysis on the other clouds can be reviewed in the appendix § A.

4.1.2 Cloud Empirical Variograms

With the geometric information of the cloud determined with the segmentation algorithm, it was possible to calculate the vertical wind empirical variograms in the four directions using a lagrangian hypercube that was placed in the center of the real bounding box of the cloud at each time-point. In other words, the biggest bounding box during the 150s was centered each second at the middle point of the momentary bounding box of the analyzed time-point. As a clarification, these bounding boxes at each time-point are the actual geometric limits of the cloud and should not be confused with the “rough" version in the previous section, which served as the input for the segmentation algorithm

Thus, when calculating the variograms, the points compared with each other followed the slight horizontal movement of the clouds. Furthermore, with the assistance of the liquid water content threshold, the empirical variograms only took into account a pair of points if both were inside the cloud when calculating the squared differences.

Under these conditions, the empirical variograms of the five clouds were determined. The empirical variogram for cloud one and can be viewed in Fig. 4.5, while the rest are depicted in appendix § A.2.

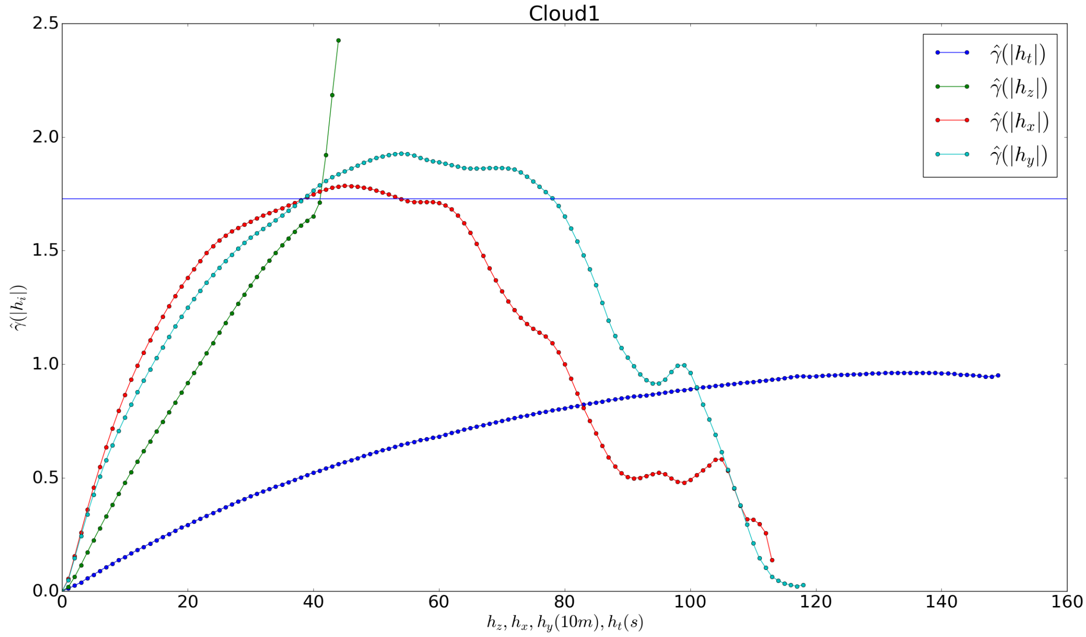


Figure 4.5: Empirical variograms in the four directions for cloud one.

Interpretation

Although the time variogram of every cloud exhibits the desired behaviour of converging near the theoretical sill, defined by the variance σ^2 of the vertical wind data inside the cloud, the rest of the directions at all clouds have erratic behaviour that differs from the expected by theory.

Most of the z-variograms grow over the sill and do not show signs of convergence. This could be caused by a trend of the vertical wind in the z-direction. The evolutions of the mean and maximum of the vertical wind in terms of height for each cloud shown in the visualization section § 4.1.1 do show a slight positive trend for some clouds, but these plots do not permit to draw clear conclusions of whether there is a vertical trend or not.

Moreover, at large distances, the z-variograms show very erratic step-like changes in values. The explanation for this lies in the fact that at these large distances the point pairs being used to compute the squared differences are probably part of the irregular parts of the cloud that are below the actual well-formed cloud base. This cloud parts pertain to the first cross-sections seen in Fig. 4.1 before the height of around 1.075 km.

As for the x-and y-variograms, it is clear that all variograms start to descend consistently after a certain distance. Since at larger distances the pair of points compared in the squared differences are both nearer and nearer to the boundaries of the clouds, it is natural that at these distances the vertical wind is more correlated. Since inside the clouds the values of the vertical wind are normally positive and outside of the cloud the values are usually below zero, vertical wind values at boundaries that are far from being adjacent will still be highly correlated.

Futhermore, the previous affirmation along with the wind field structure and geometric structure that can be seen in Fig. 4.3 suggest that a polar representation of the xy-plane seems more appropriate. In addition, the fact that at the boundaries of the clouds there is a gradient from positive to negative vertical wind values also indicates that a radial trend with higher values at the center and gradually lower values as the distance to the center is increased may be present.

The aforementioned radial trend may also explain why most xy-variograms of the different clouds go over the sill, thus suggesting the presence of trend in the xy-plane.

4.2 Transformation to Cylindrical Coordinates

Given the behaviour of the xy-variograms presented in the previous section, and also, thanks to the visualizations of the vertical wind cross-sections, the author deemed necessary to transform each cross-section to polar coordinates that are based on the center of the LWC for each cross-section. Furthermore, to be able to compare the radial behaviour of each cloud with each other, a normalized radius was used in the polar coordinates.

These conditions led to a vertical wind field and geometric cloud structure such as the one presented in Fig. 4.6

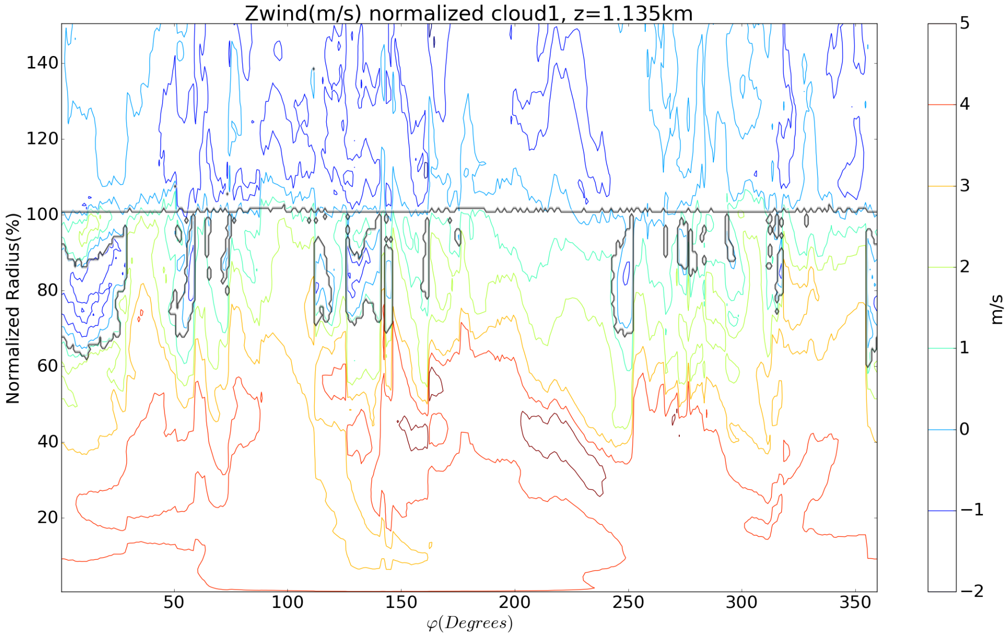


Figure 4.6: Contour plot of the radially normalized wind field of a cross-section of cloud one. The outer highlighted contour pertains to the boundaries of the cloud. The normalized radius coordinate goes up to 150% of the radius at given direction *phi*

The goal of the coordinates transformation is to determine a radial trend that can later be used to detrend the vertical wind data of a cloud. Thus, the empirical variograms can be calculated on the detrended data.

Be that as it may, since the coordinates transformation warps the cloud to become a cylinder where the center of the circle at each cross-section corresponds to the LWC center of the cross-sections of the actual clouds, the behaviour of the variograms in the t

and z will be affected due to this warping. Thus, it is expected that the behaviour seen in section § 4.1.2 will be altered.

4.2.1 Normalized Radial Trend

Using the normalized radius and a resolution of one angular degree in the polar representation, tens of thousands of radial trends could be extracted from each cloud. Herein, about 6 seconds of the simulation per cloud and at least 200m in terms of height were part of the analysis.

Moreover, the trend normalized by the vertical wind value measured at the center, in our case the center of mass of LWC, was also of interest. This trend would actually enable an initial rough modeling of the vertical wind field independent of the cloud. It would only suffice a measurement of the vertical wind value at the center of the analyzed cross-section to scale the trend accordingly and have a rough model of the entire cross-section or even the whole cloud.

The results of the analysis explained above can be viewed in Fig. 4.7. The trend of the other four clouds are depicted in § A.3.

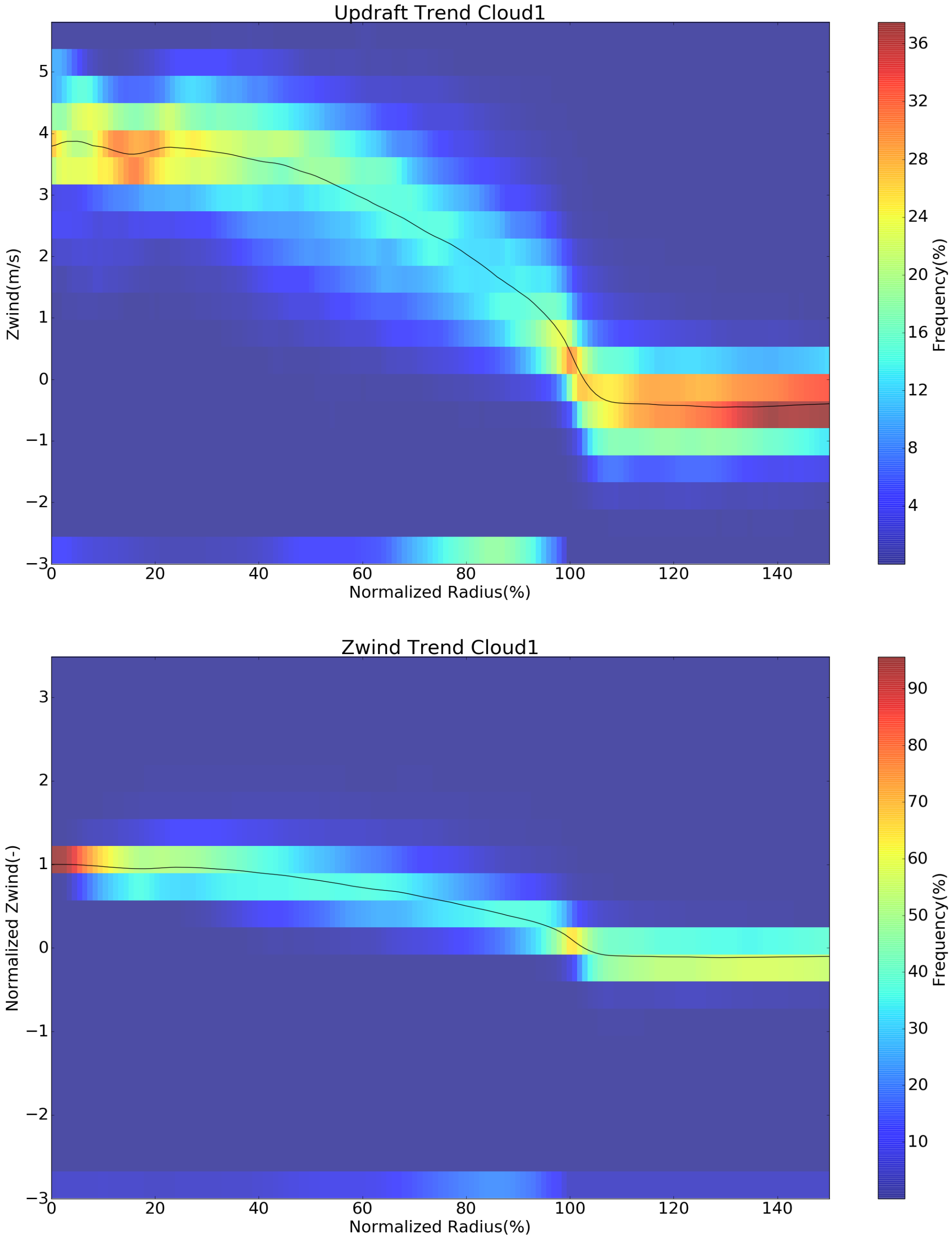


Figure 4.7: 2D Histograms of the radial trend for cloud one. Top: normalized radius up to 150%. Bottom: Normalized radius up to 150% and all trend wind values normalized by the value measured at the center of mass of the cross-section

The bar at the bottom of the trends, i.e. at value -3, corresponds to the default value chosen for the frequency of *NANs* at a given position of the normalized radius. The *NANs* encode the absence of cloud at that given normalized radius, which happens, for instance, when some angular directions have cloud holes.

For this reason, the median trends calculated do not take into account *NANs* entries.

Global Trend

Over 300.000 trends with both wind and radius normalized of the five clouds were joined and with this global trend, the 2D histogram in Fig. 4.8 was created. The goal now is to find an adequate analytical representation of this new acquired median global trend, so that afterwards the vertical wind data of the clouds can be detrended in order to determine the empirical variograms in the new polar detrended representation.

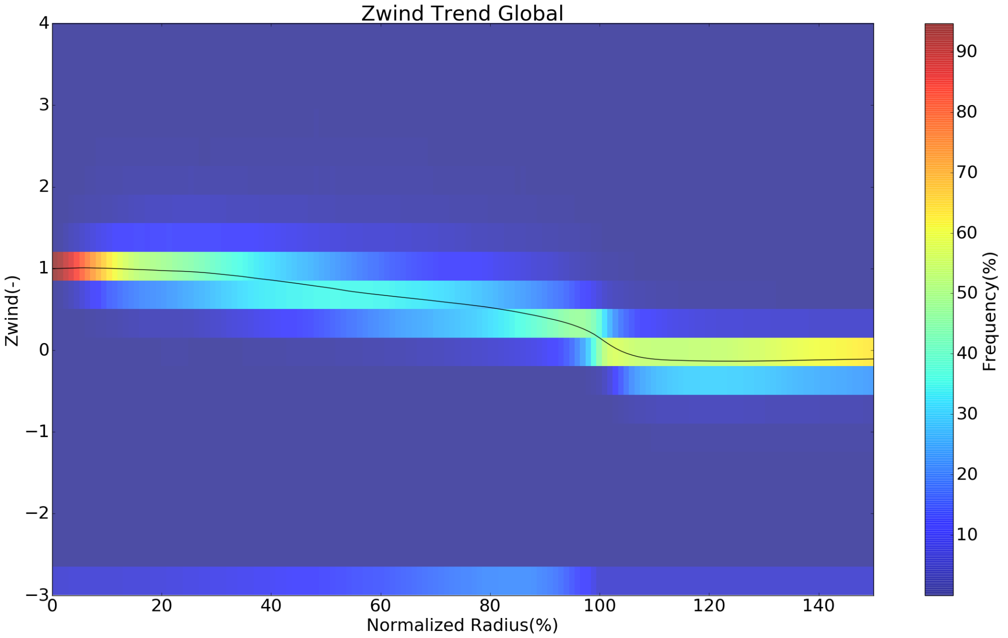


Figure 4.8: 2D Histogram of the radial trend of all clouds put together.

Fitting Global Median Trend with Analytical Function

Several analytical functions were fitted to the global median trend using the Least Squares optimizer offered in the *scipy* package. These functions are:

- Generalized Logistic Function: $trend(r) = A + (K - A)/(1 + Q\exp(-Br))^{1/nu}$
- 5th order Polynomial: $trend(r) = (\theta_0 + \theta_1 r + \theta_2 r^2 + \theta_3 r^3 + \theta_4 r^4 + \theta_5 r^5)$
- Inverse Quadratic: $trend(r) = A + B/(1 + Cr^2)$
- Gaussian-like Radial Basis Function: $trend(r) = A + B\exp(-Cr^2)$

- Tangens-hyperbolicus: $trend(r) = A + B\tanh(Cr + D)$

The results can be viewed in Fig. 4.9. The Analytical function with the lowest fitting error was the *Generalized Logistic Function*, but the problem with this function is that parameter Q was on the order of magnitude of 10^{15}, which certainly is not desirable.

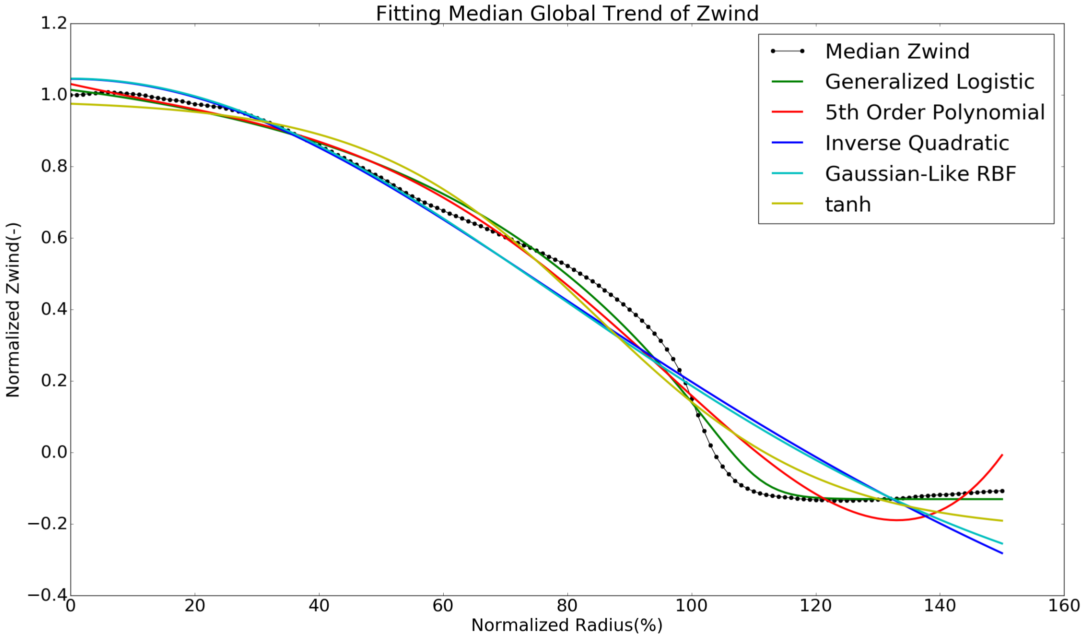


Figure 4.9: Fitting the median radial trend with analytical functions

Seeing that the other options did not fit well and increasing the order of the polynomial is certainly not elegant, the author decided to use the trend "as is" for detrending the data to calculate the empirical variograms. Since applying this median trend only consists of performing linear interpolation of the values of an array with length of 151, i.e. one index for each radial procentual step, computational considerations can be set aside, for this is a simple operation performed on a very small dataset.

4.2.2 Normalized Polar Variograms

With the help of the median radial trend normalized both in radius and in vertical wind Fig. 4.8, henceforth called the trend, new empirical variograms with the detrended vertical wind can be calculated.

Due to memory issues during the preprocessing to transform to polar coordinates, the 150s of the different clouds could not be used in its entirety, as was the case in the calculation of the empirical variograms in the cartesian t, z, x, y representation. Thus, to obtain the empirical variograms along the z, φ, r directions, several 5 seconds partitions of each cloud were used. This analysis resulted in 24 empirical variograms along the said directions, where all clouds were as evenly represented as possible, i.e. almost 5 variograms per cloud.

As for the empirical variograms in the t-direction, they were calculated with a timespan of 150s as in the cartesian case, but since the whole cloud could not be used, several cross-sections per cloud were used. Thus, 32 empirical variograms were obtained, again with the intention of having each cloud as evenly represented as possible.

Detrending was done by scaling the trend (Fig. 4.8) with the average of several values of the vertical wind measured at the center of the cross-section at different time points. In the cases where the whole cloud was used to calculate the variograms, the vertical wind values at the center of several cross-sections were used, wherein the measurements took place at the same time point.

Some of the details and technicalities of the above presented anaysis will be elaborated in the next section.

Empirical Variograms

The quality of the empirical variograms, i.e. having the variograms behave as expected by theory, depended considerably on the performance of the trend. When the trend had a good performance and was unbiased, i.e. it overestimated as much as it underestimated the values of the entire cloud or the cross-section being analyzed, then the empirical variograms behaved near to what was expected by theory. This can be visualized in Fig. 4.10, wherein the trend was unbiased (see residuals in Fig. 4.11) and had an $R^2 = 0.44$, namely, it was a trend with good performance.

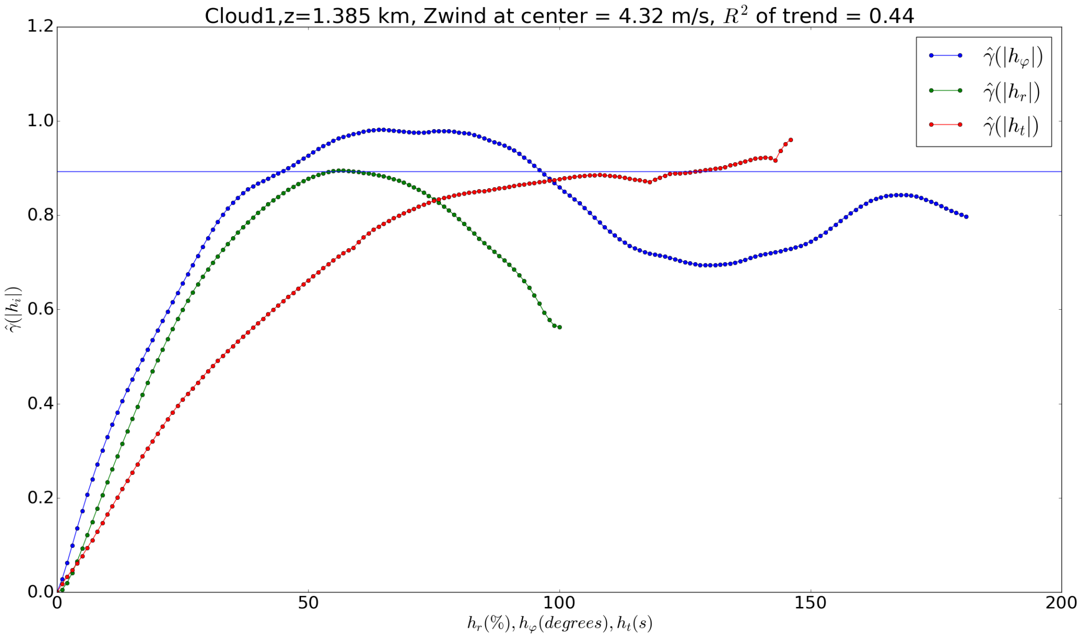


Figure 4.10: Empirical variograms of a cross-section of cloud one, analyzed over a span of 150s. Herein, the dimensions r, φ and t were calculated.

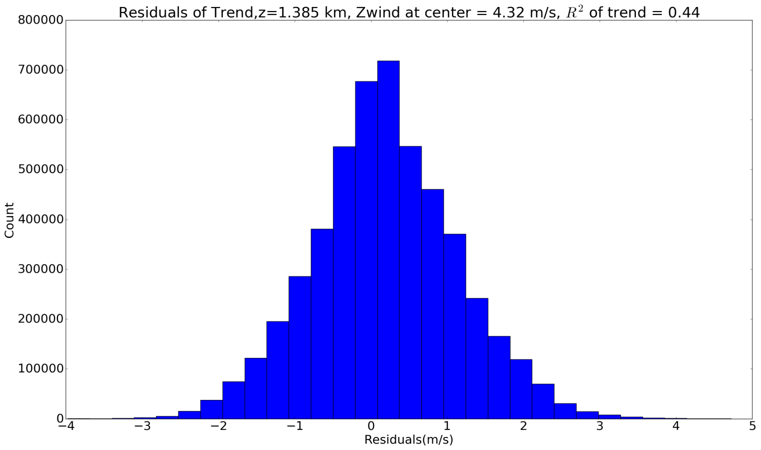


Figure 4.11: Residuals of the Trend inside the cross-section used to calculate the variogram in Fig. 4.10

In the case of the variogram in Fig. 4.12, the trend has less performance, as can be seen in Fig. 4.13. The trend slightly understimates the values inside the cross-section being analyzed. The corresponding variograms in the three directions do not seem to have the inclination to converge and also go well over the theoretical expected sill, specially the t-variogram.

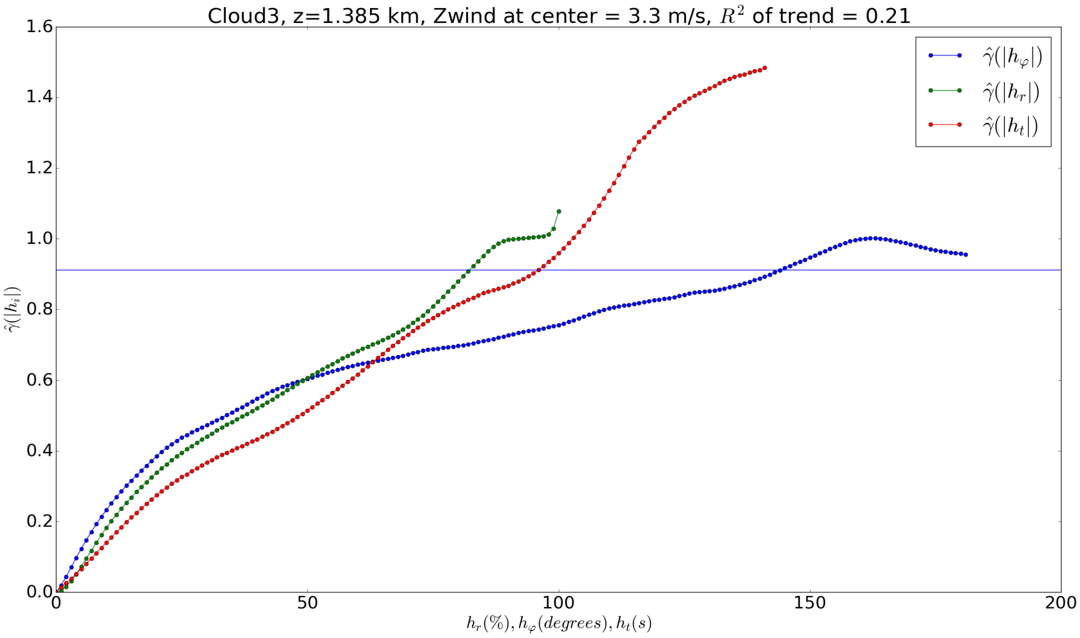


Figure 4.12: Empirical variograms of a cross-section of cloud three, analyzed over a span of 150s. Herein, the dimensions r, φ and t were calculated.

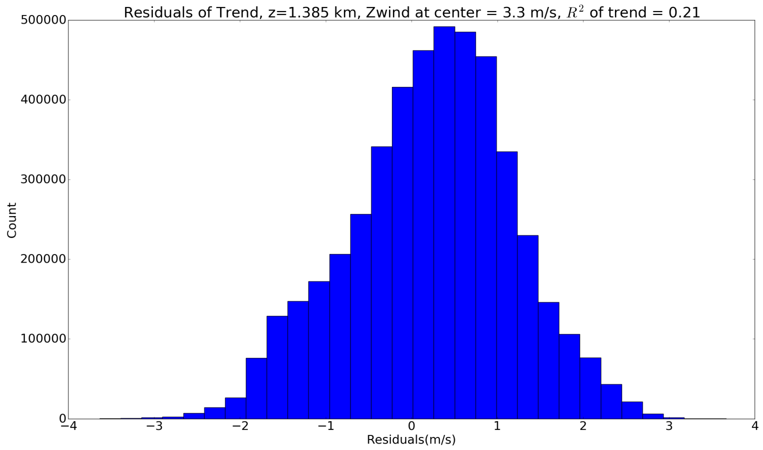


Figure 4.13: Residuals of the Trend inside the cross-section used to calculate the variogram in Fig. 4.12

Lastly, the variogram of Fig. 4.14 was calculated on the detrended data of one of the trends with the worst perfomance($R^2 = -0.41$, Fig. 4.15). This trend clearly underestimates the values of the whole cloud and its prediction error inside the cloud is significantly higher than the prediction error obtained, if one were to predict with the mean of the vertical wind inside the cloud. This is mainly caused by the value of $1.74\,m/s$ measured at the center, which is clearly too low. Thus, the center was not near the updraft bubble that had the actual higher values that are expected.

As for the variogram behaviour, it can be seen that the downward behaviour of the variogram in the normalized radial direction is even more accentuated than in Fig. 4.10. This downward behaviour at larger normalized radii conveys the message that detrending creates similar values at shorter and larger distances from the center of the clouds. This means that the trend is similarly biased at shorter and larger distances from the center, i.e. it probably understimates both the values near the center and near the boundaries. Moreover, the z-variogram shows no signs of corvergence, thus implying that there may be some form of trend or other causes for the non-stationary behaviour.

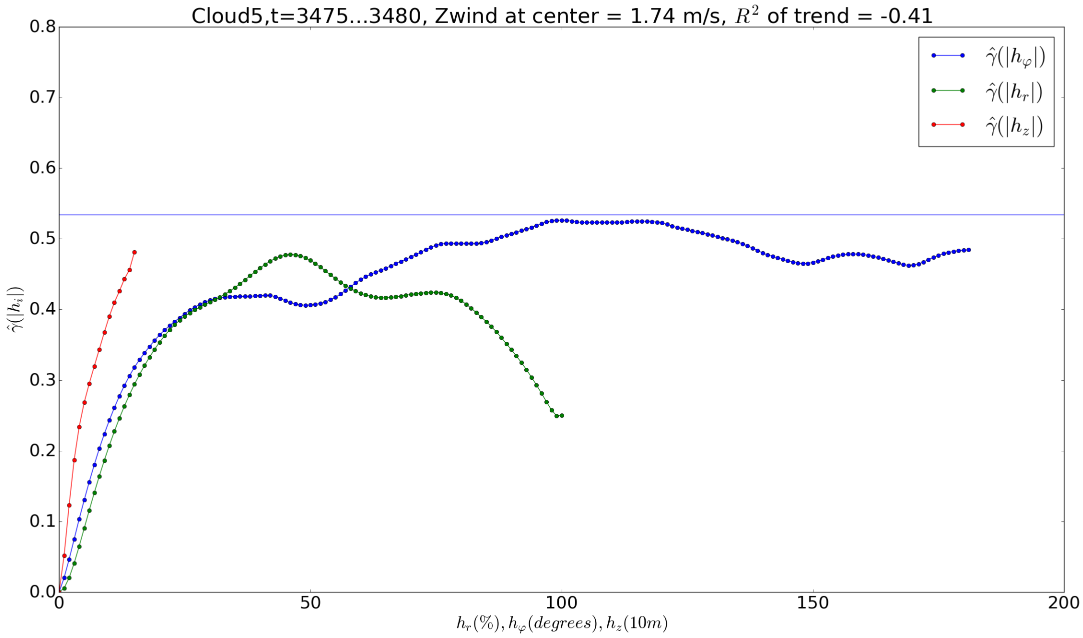


Figure 4.14: Empirical variograms of the entire cloud five, analyzed over a span of 5s. Herein, the dimensions r, φ and z were calculated.

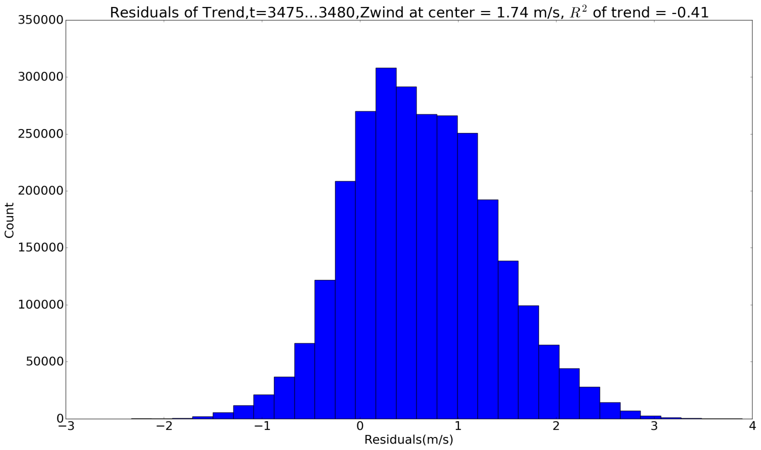


Figure 4.15: Residuals of the Trend inside the cloud used to calculate the variogram in Fig. 4.14

Summary of Behaviour

The behaviour of the z-variogram, as seen in Fig. 4.14, was common across all z-variograms. As stated before, the behaviour in the z-direction should be analyzed to determine countermeausures, so as to make the variograms in this direction more stationary.

In addition, even though most of the trends were of good quality, as in the case of Fig. 4.11, most of the variograms in the radial-direction exhibited the downward behaviour after a certain normalized radial distance. As mentioned previously, the most probable cause is the trend being biased in the same way, underestimating or overestimating both near the center and near the boundaries of the cloud.

As for the variograms in the φ-direction, they behaved as expected by theory in most of the cases. This is surprising considering that angular distances are bigger as the radius increases, thus the expectation of finding undesired behaviour.

Lastly, it should be added that most of the t-variograms did not follow the behaviour of Fig. 4.10. Although most of them did converge, the majority did so over the sill. This in contrast with Fig. 4.5 and the variograms of the other clouds (§ A.2), where the majority converged to the sill or under it. This suggests that both the radial detrending or the coordinates transformation may change the temporal behaviour of the vertical wind.

Fitting the Variograms

As stated in the previous segment, due to memory problems during the preprocessing to transform to polar coordinates, the entire 150s of the different clouds could not be used in its entirety, as was the in the cartesian case. Thus, the empirical variograms in the z, φ, r directions were calculated separatedly from the empirical variograms in the t-direction.

This resulted in 24 empirical variograms in the z, φ, r directions where most of the cloud was used but in short timespans of five seconds. Moreover, regarding the t-direction, 32 empirical variograms were calculated, wherein a single cross-section was used, but using a timespan of 150 seconds.

The empirical variograms were fitted with 4 different models for each direction using weighted-least-squares((2.57)), as planned in § 3.3.2, so as to obtain prior knowledge about the covariance structure. Herein, the objective is to determine which of the covariance functions had the least amount of fitting error across all variograms in all directions. Furthermore, once the best covariance function out of the four studied is determined, its hyperparameter distribution will be analyzed.

As explained above, the fitting procedure was implemented with all available empirical variograms. This was accomplished by using the "optimize" subpackage The resulting error metrics for each four coordinates in the normalized polar representation can be viewed in Table 4.1.

Thus, the exponential covariance function had the best overall performance, being the best fit for the coordinates t, z, φ. Along with the fitting performance, the distribution of hyperparameters of the exponential covariance function can be useful to improve the hyperparameter optimization of the Gaussian Process Regression. Furthermore,

Coord.\Model	Sq. Exp.	Exp.	Matérn, $\nu = 3/2$	Matérn, $\nu = 5/2$
$\sum t, \quad n = 32$	$8.434 \cdot 10^8$	$\mathbf{2.737 \cdot 10^7}$	$6.879 \cdot 10^8$	$7.807 \cdot 10^8$
$\sum z, \quad n = 24$	$5.012 \cdot 10^7$	$\mathbf{5.250 \cdot 10^6}$	$2.4 \cdot 10^7$	$3.579 \cdot 10^7$
$\sum \varphi, \quad n = 24$	$3.467 \cdot 10^8$	$\mathbf{4.332 \cdot 10^7}$	$2.39 \cdot 10^8$	$2.892 \cdot 10^8$
$\sum r, \quad n = 24$	$7.376 \cdot 10^7$	$1.172 \cdot 10^8$	$\mathbf{1.402 \cdot 10^7}$	$3.673 \cdot 10^7$
$\sum t, z, \varphi, r$	$1.314 \cdot 10^9$	$\mathbf{1.931 \cdot 10^8}$	$9.649 \cdot 10^8$	$1.151 \cdot 10^9$

Table 4.1: Weighted Sum of Squared Errors of the fitting of empirical variograms. Minimum values for each row are boldened

the model with the median hyperparameters for each direction should be analyzed to determine which type of anisotropy to take into account in the modeling.

Therefore, the histograms of all hyperparameters obtained by the fitting procedure will be illustrated in the following figures.

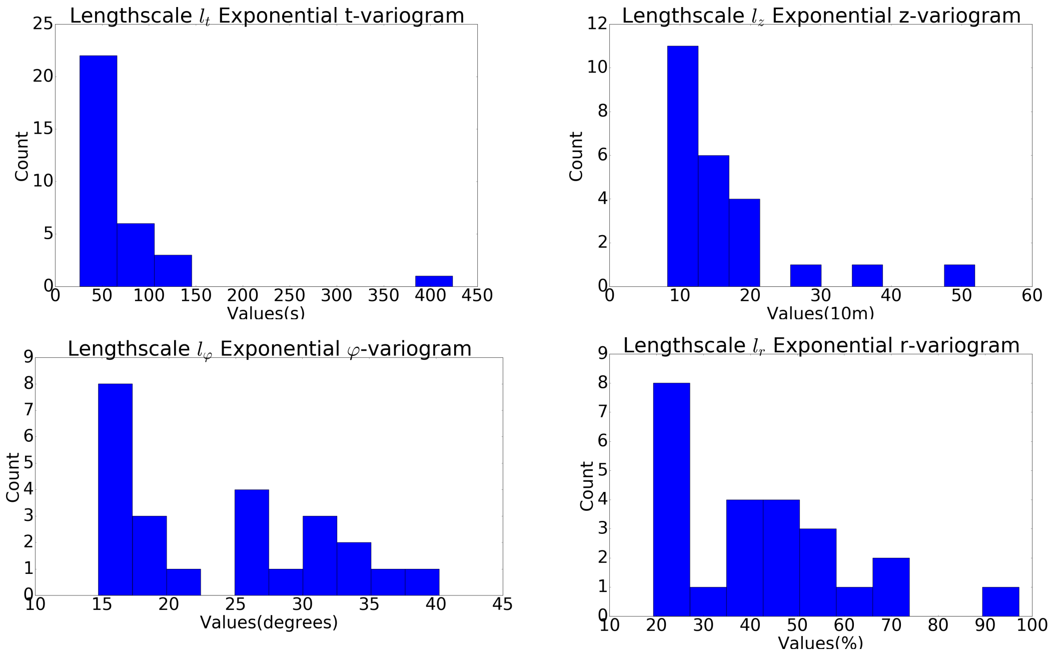


Figure 4.16: Histograms of the lengthscales obtained by the fitting procedure in in all directions of the polar representation.

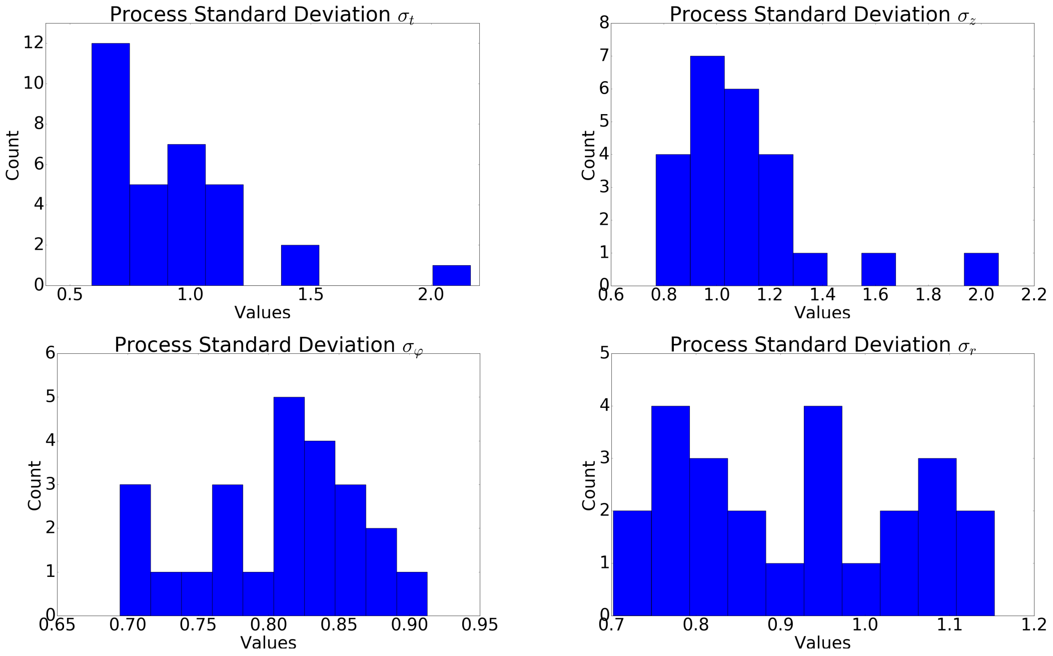


Figure 4.17: Histograms of the process standard deviations obtained by the fitting procedure in all directions of the polar representation.

The above presented distributions of hyperparameters regrettably do not present any accentuated probabilistic distribution. This is probably due to the small amount of variograms that were available for fitting. Nonetheless, the median hyperparameters will be useful to set the starting point for the hyperparameter optimization when later implementing Gaussian Process mapping. Moreover, the performance in mapping of the aforementioned medians will be analyzed, for it would be desirable to have default parameters that do not need optimization.

The median hyperparameters obtained for each direction can be visualized from Table 4.2.

Median Hyperparameter	Value
$l_t(s)$	51.027
$l_z(m)$	130.550
$l_\varphi(degrees)$	23.026
$l_r(\%)$	40.199
σ_t	0.871
σ_z	1.033
σ_φ	0.818
σ_r	0.931

Table 4.2: Median Hyperparameters of the fitted Exponential covariance function in all directions

With these median hyperparameters, the variogram models for each direction can be visualized in Fig. 4.18 to inspect for anisotropy.

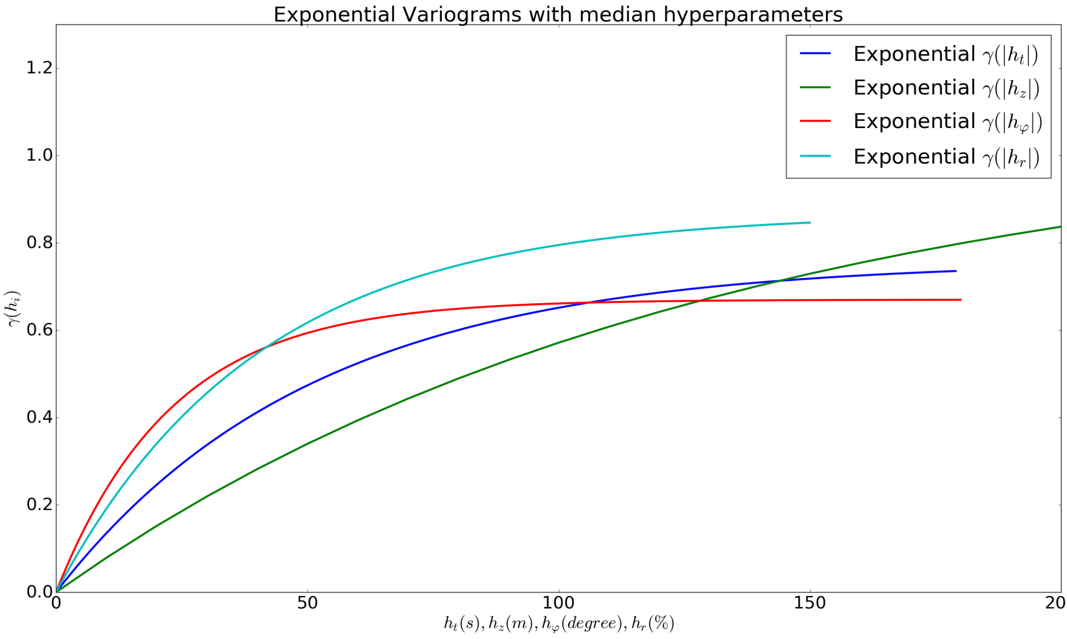


Figure 4.18: Modeled Variograms using median hyperparameters for each direction in the polar normalized representation.

It can be seen that there is an accentuated presence of range anisotropy, but the sill anisotropy is negligible. This can also be observed directly in the values of each process standard deviation(Table 4.2) where the difference between the lowest and highest is about 0.2.

Thus, for the experiments, the author deems reasonable to model the process with a single covariance function that models range anisotropy and uses a variance of $mean((median(\sigma_i)^2)) = 0.841$. In contrast, the alternative would be to create a nested model that sums the effect of each direction, automatically increasing model complexity.

4.3 Experiments

As explained in the research approach (§ 3.4), three different sets of experiments will be carried out to test the obtained prior knowledge about the vertical wind process: mapping of a static cross-section, mapping of a dynamic cross-section, and mapping of an entire dynamic cloud. The goal is to compare the “off-the-shelf” variant of the Gaussian Process that does not use any prior knowledge against a combination of variants that do include the prior knowledge calculated in the previous sections of the current chapter.

The following variants were implemented:

1. Exponential covariance function that will be optimized using the corresponding sample trajectories in cartesian coordinates and the sampled “as is” vertical wind, i.e. without any prior knowledge.

2. Squared Exponential covariance function that will be optimized using the corresponding sample trajectories in cartesian coordinates and the sampled “as is” vertical wind, i.e. without any prior knowledge. This is the actual “off-the-shelf” alternative.

3. Exponential covariance function using the hyperparameters obtained from the variogram approach. The input trajectories will be transformed to the normalized polar coordinates and the sampled vertical wind will be detrended. The optimization will be turned off. This would be the version that uses both the priors on the covariance structure (without optimizing) and on the mean function.

4. Exponential covariance function. The input trajectories will be transformed to the normalized polar coordinates and the sampled vertical wind will be detrended. The model will be optimized without default starting points. This model will serve to compare the quality of the obtained hyperparameters versus optimizing with default settings.

5. Exponential covariance function. The input trajectories will be transformed to the normalized polar coordinates and the sampled vertical wind will be detrended. The model will be optimized using the hyperparameters obtained with the variogram approach as starting points. This model will serve to determine if using the obtained hyperparameters as starting points helps the optimization to reach better optima.

6. Squared Exponential covariance function. The input trajectories will be in cartesian coordinates, but the sampled vertical wind will be detrended. The hyperparameters will be optimized. This corresponds to using the prior on the mean function, but not using the prior on the covariance structure.

7. Exponential covariance function. The input trajectories will be in cartesian coordinates, but the sampled vertical wind will be detrended. The hyperparameters will be optimized. This also corresponds to using the prior on the mean function, but not using the prior on the covariance structure.

Along with the above presented Gaussian Process alternatives, the performance of the trend will also be assessed. Moreover, the trivial model of predicting with the mean of the sampled wind will serve as benchmark to compare the Gaussian Process alternatives and the trend.

As presented in § 3.4, three sets of experiments were carried out, one on static cross-sections of clouds, the second set on dynamic cross-sections and the third set of experiments pertain to the entire cloud. The data used for the experiments was all related to cloud one. Other clouds were not targeted for experiments due to contraints in time.

The first two set of experiments used circular trajectories of five drones to sample data for training (view Fig. 3.3). The five drones followed circles at five different locations of the cross-sections and sampled the wind each second. Eight different cross-sections were considered and, in order to have more realistic training data, the sampled winds were contaminated with Gaussian white noise. Moreover, to test the impact of the level of noise and the randomness of the noise itself, 5 noise levels were tested, wherein 20 trials per noise level were implemented. The standard deviations for the noise were $0.001, 0.1, 0.25, 0.5$ and $0.75\, m/s$. The sample time was of 75 seconds and thus there were about 350-385 data points inside the cloud cross-section for training, the actual amount of data depending on each cross-section.

For the test on the entire cloud five helicoidal trajectories were used to sample the wind data. Herein, the wind data was also contaminated with additive Gaussian white noise and as in the experiments with the cross-sections, 20 trials for each of the five noise levels presented in the previous paragraph were implemented. The sample time was again of 75 seconds and thus there were around 375 points available inside the cloud for training.

Apart of the sampling process, additive noise was also added to the value measured at the center of the cross-section for detrending. Moreover, in the case of the experiment on the entire cloud, only the value measured at the center of one cross-section was used to detrend the entire cloud.

It should also be added that the seven covariance functions mentioned above will have an added noise covariance function, i.e a noise hyperparameter σ_n. It will be assumed that covariance model three, whose hyperparameters are not optimized, has a perfect noise model, i.e. the noise hyperparameter is set equal to the actual added noise. Moreover, covariance model five will also have a perfect noise model as a starting point for the hyperparameter optimization. For the rest of the covariance functions, the noise hyperparameter will be obtained through the hyperparameter optimization.

Under the aforementioned conditions, the results of the experiments will be presented in the following segments. Herein, only the noise level $\sigma_{noise} = 0.25\, m/s$ will be dealt with. For the rest of the results, the reader is advised to check § A.4.

4.3.1 Summary of results

The summary of the results for the three scenarios for noise level $\sigma_{noise} = 0.25\, m/s$ can be visualized in Fig. 4.19. These include the RMSE and predicted standard deviation of the seven covariance functions, and also the RMSE of the trend and the trivial mean prediction model. The seven covariance function explained above are depicted in the same order from left to right as in the list. It can be seen that model seven had the best RMSE performance in the scenarios of the static cross-section and the entire cloud. As for the experiments on the dynamic cross-section, the best model was model one.

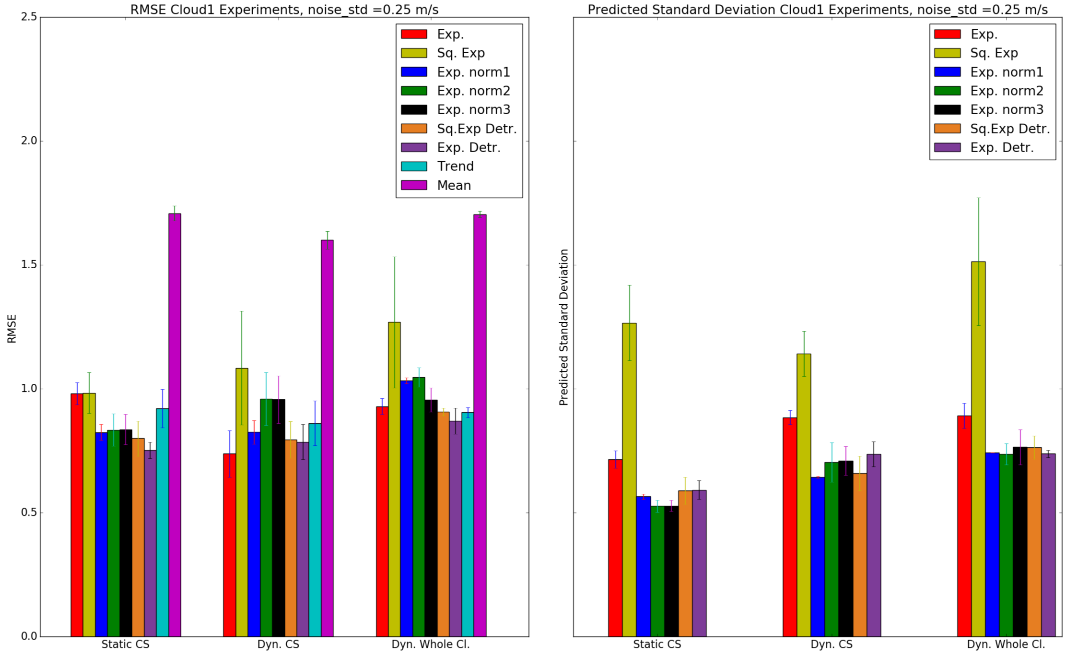


Figure 4.19: Summary of RMSE and predicted standard deviation for noise $\sigma_{noise} = 0.25 m/s$, all mapping variants for the three scenarios.

As for the performance of models three, four and five, which used both the prior knowledge on the covariance structure in normalized polar coordinates and the prior on the mean function, it can be seen that they outperform the “off-the-shelf” variant, i.e. model two, in all scenarios.

Nonetheless, in both dynamic scenarios, model four and five have more predictive error than the trend, which is counterintuitive. Moreover, model three, which does not undergo hyperparameter optimization, also worsens the prediction of the trend in the scenario with the entire cloud, but not in the dynamic cross-sections experiments. When comparing the performance of these three models, it can be seen that their performance is almost identical in the static experiment, but they differ significantly in the dynamic scenarios. Particularly suprising is that model three outperforms model four and five in the dynamic cross-section experiment.This reflects the quality of the hyperparameters obtained with the variogram approach and also the importance of having a good noise model for the covariance functions.

In addition, it can be stated that the predictions created by models three, four and five, which use the residuals of the trend for training, when put on top the trend, worsen the prediction of the trend itself. This behaviour was in contrast with model seven, which actually improves the trend prediction, albeit slightly in the dynamic scenarions, as is expected by common sense.

Moreover, almost all Gaussian Process mappings that were based on the residuals of the trend (Models three to seven) underestimate the actual error in the static cross-sections and dynamic whole cloud scenarios, as can be seen by comparing the RMSE with the respective predicted standard deviation. Be that as it may, the predicted standard deviation of model six and seven are more congruent to their RMSE than the other three. In the experiments pertaining to the dynamic cross-sections, model three, four and five seem again to understimate their error, whereas models six and seven appear to have consistent error predictions, as far RMSE and mean of standard deviation goes.

Of particular interest is also model one, which is the best one in terms of RMSE for the experiments on the dynamic cross-section and also outperforms models three, four and five in the entire cloud scenario. This result was unforeseen, considering that this model does not make use of neither options of the prior knowledge that were determined in the precedent analysis. Furthermore, its error prediction in the dynamic scenarios seems to be consistent when comparing RMSE and the mean of the predicted standard deviation.

4.3.2 Static Cross-section

In order to better grasp the results that the different Gaussian Process mappings are producing, one trial at noise level $\sigma_{noise} = 0.25\,m/s$ of the static cross-sections scenario will be analyzed. In Fig. 4.20, an instance of the trend can be visualized, including its prediction error. This prediction error, i.e. the residuals of the trend, are sampled in the locations of the five circular trajectories to serve as output for the training phase of models three to seven.

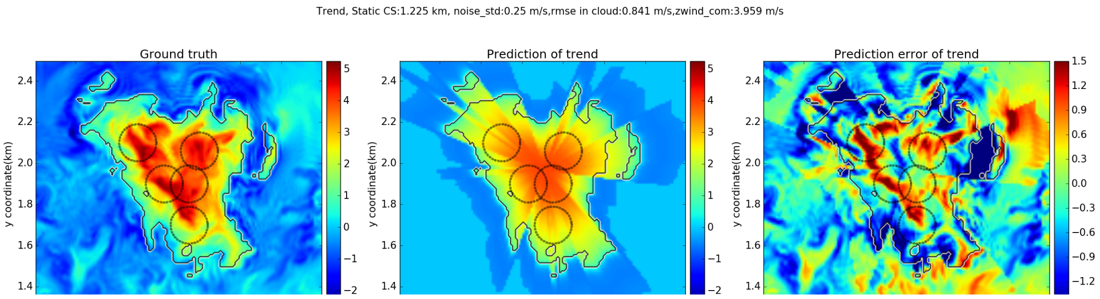


Figure 4.20: Mapping of the cross-section by implementing the trend scaled by the wind measured at the center. Noise level $\sigma_{noise} = 0.25\,m/s$.

The trend does a good job at depicting the essence of the wind field inside of the cloud, as is shown by the preponderance of light green in the prediction error. Nonetheless, due to the non-convex geometry and the non-connected sections outside of the contiguous central part of the cloud, the trend seems in general to overestimate the values near the boundaries, as can be seen by the predominance of blue in the prediction error near the outer contour. Moreover, the residuals at the core of the cloud are dominated by the red color, demonstrating that the bubbles of maximum vertical wind are not exactly at the center of the cross-sections. Be that as it may, the notion that the vertical wind dimishes as one increases the distance from the center certainly holds.

As for the predictions done by the Gaussian Process alternatives, model two and model three are portrayed in Fig. 4.21, along with the trend. This figure depicts the benchmark to be beaten, i.e. the "off-the-shelf" version without any prior knowledge, and the mapping of the covariance function created using the hyperparameters determined through the variogram approach, without optimizing.

It can be seen that model three improves the prediction of the trend slightly (RMSE $0.831\,m/s$ vs. $0.841\,m/s$), and thus the predominance of red in the prediction error is reduced. Furthermore, model three clearly outperforms the "off-the-shelf" mapping, as was expected due to the summary of the results shown previously. Moreover, most of

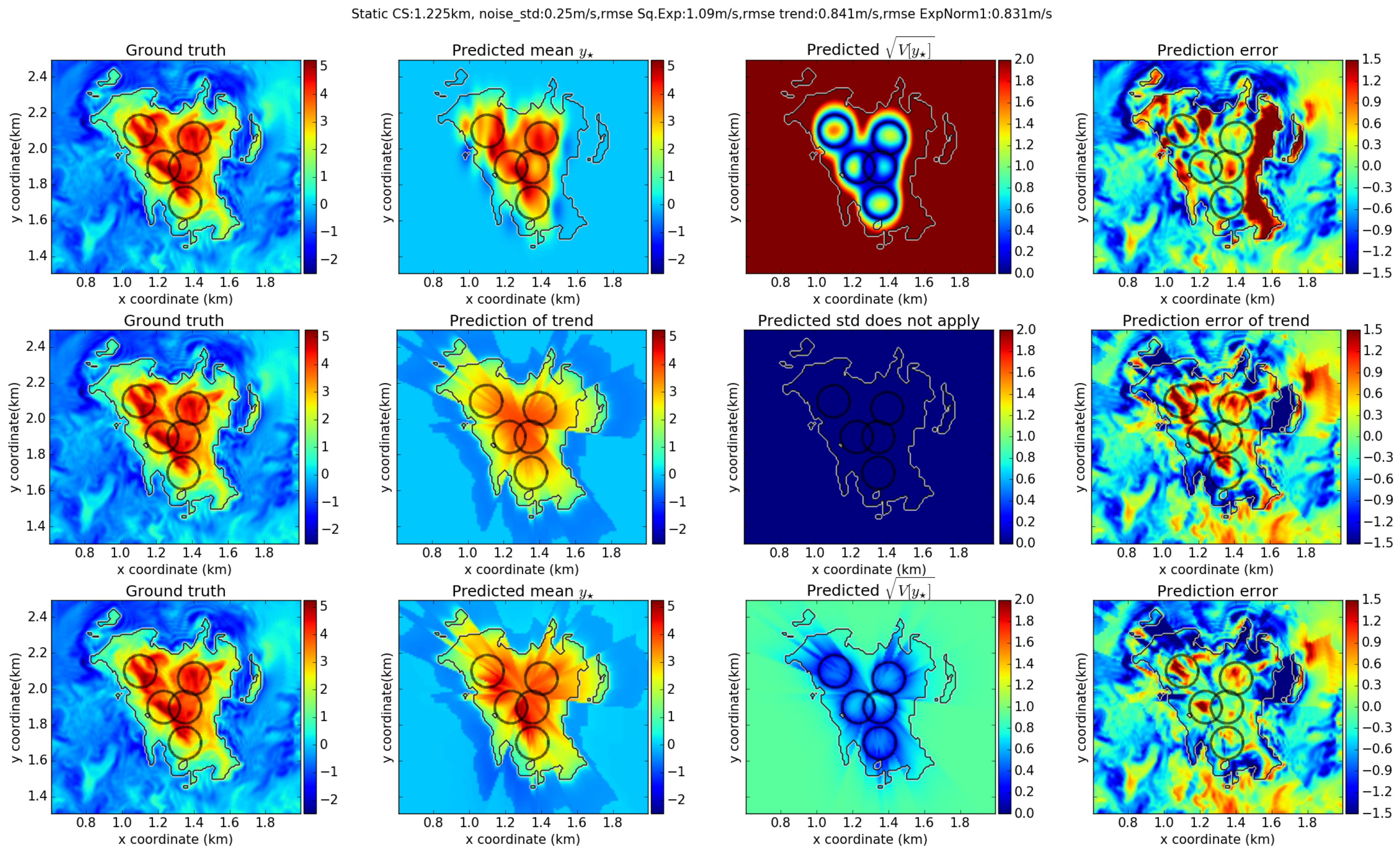


Figure 4.21: Gaussian Process mapping of a static cross-section of the cloud. The top row corresponds to the “off-the-shelf” squared exponential variant optimized with inputs in cartesian coordinates and “as is” sampled vertical wind contaminated with noise. The middle row depicts the behaviour of the radial trend. The bottom row shows the behaviour of the exponential covariance function with inputs in normalized polar coordinates using the hyperparameters obtained with the variogram approach and detrended vertical wind with additive white noise. Noise level $\sigma_{noise} = 0.25\,m/s$.

the predicted error of the “off-the-shelf” alternative is done at the right side of the cross-section, which contrasts a lot with the structure of the prediction error of the trend and the mapping of model three. This behaviour is caused due to the combined effect of being at a large distance to the sample points of the circular trajectories and the effect of the mean zero assumption, thus predicting near zero vertical wind at the right hand of the cloud. This is corroborated by the median of the lengthscales of this mapping, which were of around 0.05 km for both x and y.

Furthermore, both Gaussian Process mappings show similar behaviour in terms of their predicted standard deviation. The models are only certain of their predictions in the immediate vicinity of the sample points. Thus, this in an indication that the test domain, i.e. the whole cross-section, is quite large in comparison to the train domain used to get the models. This is regrettably the nature of the problem at hand, where 1D sample trajectories are used to infer the values inside a cloud, whose size is several orders of magnitude larger than the training set size available.

Visualizing Model with best Performance

Of particular interest is the visualization(Fig. 4.22) of model seven, which had the best performance. This superior performance is quite visible in the prediction error field, where the predominance of light green is clear. Herein, the prediction based on the residuals of the trend substantially diminished the error in comparison to the actual trend, which was the behaviour expected in the models two to five.

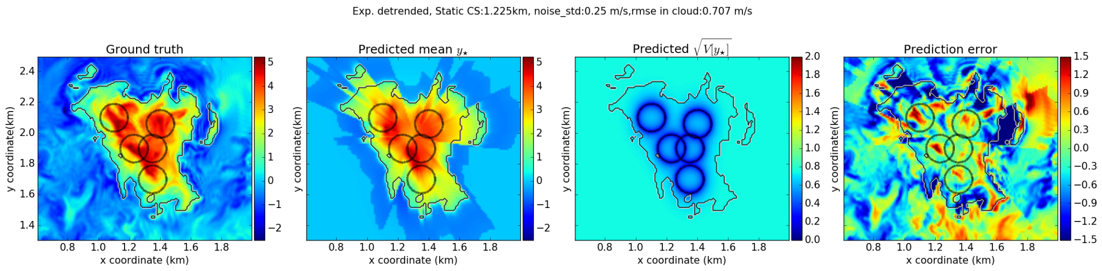


Figure 4.22: Gaussian Process mapping of a static cross-section of the cloud. Herein, an exponential covariance function was used with inputs in cartesian coordinates of the five circle trajectories. The output for training was the detrended vertical wind under the effect of white noise. Noise level $\sigma_{noise} = 0.25\,m/s$.

The hyperparameter distribution of this model may be interesting for future work, and thus it is presented for all three scenarios and noise level $\sigma_{noise} = 0.25\,m/s$ in appendix § A.4.3. The medians of these distributions can be extracted from Table 4.3.

Median Hyper.\Scenario	Static CS	Dyn. CS	Dyn. whole Cl.
$l_t\,(s)$	-	31.640	46.974
$l_z\,(km)$	-	-	0.811
$l_x\,(km)$	0.080	0.075	0.095
$l_y\,(km)$	0.112	0.061	0.090
$\sigma_f\,(m/s)$	0.840	0.857	0.897
$\sigma_n\,(m/s)$	0.184	0.018	0.156

Table 4.3: Median Hyperparameters for all scenarios of the exponential covariance function with inputs in cartesian coordinates of the five circle/helicoidal trajectories. The output for training was the detrended vertical wind under the effect of white noise. Noise level $\sigma_{noise} = 0.25\,m/s$.

4.3.3 Interpretation

The results of the seven tested covariance functions across the three experiments showed valuable insights, albeit also unexpected ones. The visualization of the static cross-section Gaussian Process mappings revealed that the prior on the mean function gives a head start to the models that use the residuals for predicting, i.e. models three to seven. This head start permits the proposed models that also make use of the prior knowledge on the covariance structure (models three to five) to outperform the benchmark, i.e. the "off-the-shelf" alternative with squared exponential covariance function and no mean function.

Nonetheless, in the dynamic experiments, models one, six and seven, which used cartesian coordinates for their inputs, perform better than the three models that have their inputs in the normalized cylindrical coordinates. The latter models even worsen the performance of the trend with the prediction they add to it, which should not happen. Moreover, it is noteworthy that in the static cross-sections experiments models three, four and five perform well and as expected, i.e. they improve the predictions of the trend.

This is a hint that the transformation to normalized polar coordinates, which warps the cloud to be represented as a cylinder with radius one, distorts too much the behaviour in all directions, and thus it is more difficult for the hyperparameter optimization to grasp the underlying process. Since models three to five have undesired behaviour (Trend prediction worsening) in the dynamic experiments and not in the static cross-sections scenario, the distortion appears to affect the t- and z-direction the most.

On the other hand, the empirical variograms calculated with cartesian coordinates, as can be seen in Fig. 4.5, show favorable variograms in the t-direction, thus demonstranting that cartesian coordinates may be more natural in terms of the temporal aspect of the underlying process.

Be that as it may, it has to be added that all of the experiments suffered from the setback that the test domain may be too large to indisputably compare models with each other, specially the dynamic cases. Considering that a typical cross-section of the analyzed cloud has about 4000-4500 grid points for a given second, sampling 350 to 375 points during 75 seconds corresponds to using about 0.1 % of the available data for training, well below the usual 50-70% that is common practice for machine learning problems. This ratio of training data vs. total data is even worse for the entire cloud, where it is approximately 0.01%. Therefore, the results of the dynamic experiment should be seen as introductory, for they do deliver valuable information, but the task at hand was probably too daunting for the capabilities of the Gaussian Process Regression algorithm.

Notwithstanding, the dynamic experiments still portray a plausible approximation, i.e. it truly corresponds to the problem that will be faced in reality. But the problem of defining a more appropriate test domain still should be analyzed in more detail. For example, when comparing the exponential and the squared exponential covariance functions that use neither options of prior knowledge available, namely, models one and two, it is insightful to glimpse at their median hyperparameters (Table 4.4), particularly the lenghtscales, for the dynamic cross-sections scenarios.

Median Hyper.\Model	Exp. cart. no Priors	Sq. Exp. cart. no Priors
$l_t\,(s)$	197.5	5.340
$l_x\,(km)$	0.435	0.120
$l_y\,(km)$	0.410	0.210
$\sigma_f\,(m/s)$	1.911	2.330
$\sigma_n\,(m/s)$	0.002	0.330

Table 4.4: Comparison of median hyperparameters for the dynamic cross-section scenario between models exponential and square exponential with no priors, i.e. model one and two of the list at the introduction of this chapter. Noise level $\sigma_{noise} = 0.25\,m/s$.

The much larger lengthscales of model one thus lets it extrapolate the vertical wind more smoothly in a broader range than model two. Since both of these Gaussian Process mappings have the zero-mean assumption, the exponential model with larger lengthscales will reach the expected zero-mean at much larger spatio-temporal distances from the sample points, thus extrapolating the measured values of the circles across bigger domains. On the other hand, the “off-the-shelf” alternative with smaller lengthscales will reach the expected zero-mean more rapidly, which leads to near zero vertical wind predictions for cloud parts that are not in the vicinity of the sample trajectories.

This behaviour considerably penalizes model two in terms of RMSE when considering the entire cloud. Nonetheless, shorter lengthscales also means “faster” reaction to adapt to the variability of the data of the training set. Therefore it is well possible, were the test domain smaller, e.g. a patch inside the cloud in front of the UAVs, that model two outperforms model one.

Thus, it should be stated that the findings of all experiments should be contextualized, i.e. the goal was to map the vertical wind field of entire cross-sections, or even the whole cloud, with data sampled along five systematic circular trajectories over a short timespan. It may be that reframing the problem to smaller scales, larger timespans or a different sampling scheme yields different insights.

To conclude, the “off-the-shelf” benchmark model was beaten by the three models that incorporated prior knowledge in terms of covariance structure and mean function, even the alternative without hyperparameter optimization. These three models require a series of preprocessing steps, such as transforming coordinates to a normalized cylindrical representation and detrending, and map the cloud using inputs in the aforementioned polar coordinates and the detrended vertical wind as output. These results validate the research approach based on the variograms, albeit some of the results were not expected, such as worsening the prediction of the trend, even though the trend is part of their prediction.

Regardless, there were other covariance functions that were based on inputs in cartesian coordinates and only used prior knowledge on the mean function, i.e. the trend, and outperformed the models three to five in all scenarios. Equally surprising was the performance of the exponential covariance function that used neither options of prior knowledge and still performed better in the dynamic scenarios than the models using both priors.

This evidence suggests that the normalized polar coordinates are useful to determine the trend and thus detrend the wind, but are not well suited for the Gaussian Process technique, because the warping in all directions distorts the underlying spatio-temporal behaviour.

Chapter 5

Outlook

5.1 Summary

The present research project concentrated on improving a current “off-the-shelf” implementation of Gaussian Process Regression to better map atmospherical variables in the context of the SkyScanner project (§ 1.2). The Gaussian Process technique is particularly suited for the goal of SkyScanner, which is to sample clouds adaptively by minimizing energy consumption and information uncertainty. This energy and information efficient adaptive sampling strategy will permit longer sampling missions that inherently sample more and better data than a systematic aquisition pattern would.

Information efficiency is possible thanks to the in-built error model of the Gaussian Process, which quantifies uncertainty and can thus be optimized to generate trajectories that aim at sampling regions with the biggest uncertainty. Energy efficiency can be achieved by sampling and mapping the vertical wind currents, thus exploiting positive vertical winds to soar with low energy consumption.

Therefore, to succesfully implement the aforementioned adaptive sampling strategy, the quality of the enviroment’s map, in particular the vertical wind currents, is of paramount importance. The current “off-the-shelf” implementation of GPR does not exploit all the machinery that the algorithm has to offer.

Of particular interest is the possibility of incorporating prior knowledge to condition the outputs of the mapping, which gives the models a head start towards the expected behaviour or belief one has about the process at hand before performing the modeling step (Hyperparameter Optimization), thus improving the mapping.

Considering that the daunting problem at hand is to map 4D enviroments (time and the three spatial coordinates) using noisy samples along 1D manifolds (drone trajectories), this inclusion of prior knowledge may be indispensable to alleviate the setbacks of modeling with such sparse noisy data.

Consequently, the current research work was based on this notion of prior knowledge and concentrated on the types of prior knowledge available and the methods to determine and inject the prior knowledge to the Gaussian Process technique. In order to do so, the fundamentals chapter (§ 2) reviews the relevant theory and literature to accomplish the aforementioned goal, which requires the use of methods related to the fields of *Gaussian Process for Machine Learning*, *Geostatistics* and *Spatio-Temporal Statistics.*

Assisted by this theory, it can be stated that GPR models a stochastic process $y = f(\mathbf{x})$ by conditioning a distribution over functions to adapt to the data available of said process. The two main ingredients to do so are the mean function $m(\mathbf{x})$ and the covariance function $k(\mathbf{x}, \mathbf{x'})$. The mean function defines the first moment of the distribution over functions and thus can be viewed as the center around which the functions vary. The covariance function defines the second moment characteristics of the process, which directly affects properties such as smoothness and mean square differentiability. Moreover, it embodies the idea of "spatial" regularity, i.e. if inputs $\mathbf{x}$, $\mathbf{x'}$ are similar, so should the outputs.

In the "off-the-shelf" alternative, the mean function is set to zero and the covariance function is the Squared Exponential, which is parameterized as in (2.17). These hyperparameters are adjusted through an optimization (§ 2.2.3), using a bounded uniform prior. This optimization is non-convex and thus is not guaranteed to converge to global optima.

To improve this "off-the-shelf" Gaussian Process, the author identified three types of prior knowledge that can be injected to the mapping. These are:

1. Determining type, hyperparameter distribution and space-time interaction of the covariance function.
2. Incorporating a mean function.
3. Exploiting correlation between output variables.

The methodology to determine these priors can be reviewed in § 3.2. The implementation of this methodology was entirely centered on the vertical wind variable. Of the three types of priors, the author was succesful at extracting a mean function and the covariance structure. The chosen method to do so was based on a variogram based approach, which is explained in more detail in § 3.3.2. The variogram can be viewed as a near relative of the covariance function, and under certain conditions, they can be used interchangeably (§ 2.3.3). The difference is that the hyperparameters are adjusted estimating the variogram empirically from the data and then regular curve fitting is done.

The realization of said variogram approach can be recapped at the beginning of the Implementation chapter § 4. It should be noted that all results of the present research project are based on realistic atmospherical simulations. The details of this simulation can be reviewed in § 3.3.1.

Once the prior knowledge was determined, experiments were designed (§ 3.4) and implemented (§ 4.3) to compare performance between the "off-the-shelf" alternative, the new Gaussian Process mappings that exploited both options of prior knowledge, and other variants of Gaussian Process mapping that either only used the mean function or none at all. The task of these experiments was to map an entire cross-section or the whole cloud with the samples gathered systematically by five drones during a timespan of 75 seconds.

It was determined that the new mapping with both options of prior knowledge indisputably outperformed the "off-the-shelf" benchmark in terms of RMSE. Nonetheless, other alternatives that only used the mean function but not the prior on the covariance structure had better performance than the approach making use of both priors. Thus, further analysis needs to be done to thoroughly understand why this behaviour happens.

5.2 Reflections

The covariance structure prior and mean function prior that were determined in the present research project are expressed in cylindrical coordinates that have normalized radius. The center for each cross-section of the obtained cylinder after transforming each cloud is based on the center of liquid water mass for each of the cloud's cross-sections.

Thus, to implement the Gaussian Process mapping following the approach presented in this thesis, a liquid water radar is of paramount importance. This radar has to provide the relevant geometrical information of the cloud, such as the cloud skeleton and cloud boundaries, at a frequency of around 1 Hz to conform to the mapping strategy of this research work.

Aside from this important requirement, the author considers that there is some work that was left undone that could be implemented without excessive effort. The experiments showed that the mapping alternatives using simple covariance functions such as the exponential and squared exponential with inputs in cartesian coordinates, but making use of the mean function, outperformed the proposed alternatives that also used the prior on the covariance structure, which requires inputs in normalized cylindrical coordinate.

Given this reality, it seems that the mean function, which also needs the coordinates transformation, has a better impact on performance than the prior on the covariance structure represented in the polar coordinates, as can be seen in § 4.3.1. Thus, it would be interesting to redo the variogram analysis on the clouds in cartesian coordinates after detrending the vertical wind, to visualize its behaviour.

Moreover, to further understand why the models that use covariance functions in polar coordinates worsen the prediction of the trend, whereas the covariance functions in cartesian coordinates improve it, the predictions of these Gaussian Process mappings should be further analyzed. Of particular interest is to visualize the prediction that goes on top of the trend, in order to understand the actual improvement or worsening caused by the models mentioned.

Furthermore, the work done with the static experiment, i.e. doing a thorough analysis of a single instance of the experiment, as performed in § 4.3.2, should be repeated for instances of both dynamic experiments. In addition, analyzing the hyperparameter distributions of all models implemented in the three experimental scenarios, and not just a selected subset, may help to better interpret the results.

Another interesting possibility would be to significantly increase the duration of the sampling to have more data available for training. Also changing the type of the systematic sampling trajectory, e.g. to more exhaustive grid trajectories, may improve the performance or change the nature of the results.

Lastly, it should be added that the error prediction consistency needs to be verified. The author already began this line of work by storing summary statistics of $(\sigma - |e|)/\sigma$, where σ is the predicted standard deviation of the models at a given test point and e is the actual error $y - y_\star$ at the same point. Assuming normal gaussianity, the error prediction would be consistent if the above expressed quantity has a distribution where the 5% quantile lies on around -1, and the 32% quantile is close to 0.

5.3 Future Work

Long term future work should focus on incorporating the third type of prior knowledge suggested in this research project, i.e. exploiting correlations between output variables. Of particular interest is the probable correlation between the liquid water content and the vertical wind. A small glimpse of this correlation can be viewed in Fig. 4.2 and Fig. 4.3, where it can clearly be seen that the center of masses in terms of liquid water content and vertical wind share almost identical behaviour. This suggests similar random fields for both variables. In order to exploit these correlations, an introduction to the available methods is delivered in § 2.2.5 and the reader is advised to do an in-depth review of (Chai, 2010).

One of the explicit advantages of obtaining this prior knowledge in terms of correlation between output variables is that each drone of the fleet can be specialized to a certain subset of the variables of interest in regards to sensor payload, but still be able to map the other variables not directly sensed thanks to the correlations.

Apart from exploiting variable correlations, a deeper analysis of the random field of the vertical wind in the z-direction is necessary. The variograms in both cartesian (§ 4.1.2) and normalizad polar coordinates (§ 4.2.2) showed non-convergence, suggesting that there may be a trend or other sources of non-stationary behaviour.

Regarding this topic, it was suggested by the meteorological sub-team of the SkyScanner project that the behaviour of the relevant atmospherical variables becomes more chaotic as one nears the upper boundaries of the clouds, since there is where the detrainment-entrainment process happens, as schematically portrayed in Fig. 1.1. To assess this behaviour one could analyze the evolution of the variance of the data as the height is increased, and also, analyzing the lengthcales of the covariance structure of cross-sections at different heights may be informative. More chaotic behaviour should result in shorter lengthscales and bigger variances as the height is increased.

Another line of future work that may be fruitful is to analyze the spatio-temporal behaviour with more scrutiny, in order to better identify the nature of the space-time interaction that is present (See § 2.4.2 for a review). By seeing the experiments with the exponential and squared exponential covariance functions that did not make use of neither options of prior knowledge, one may conclude that the process may have short-scale and large-scale behaviour, as was showed by the large lengthscales of the exponential covariance function and short ones of the squared exponential. Thus, it may be purposeful to use a sum of these covariances to grasp this behaviour.

Lastly, it is the author's opinion that effort should be invested in integrating the results and methods of this thesis in the existing adaptive sampling architecture, e.g. the coordinates transformation and detrending scheme, so as to implement new experiments to further assess the quality of the Gaussian Proces mapping. This would enable to draw indisputable conclusions about the performance of the different mapping alternatives presented on this research project, for they would be tested under the exact conditions and with the same sampling strategy that are expected to be implemented in reality.

Appendix A

Appendix to Implementation

A.1 Visualization of Clouds Two to Five

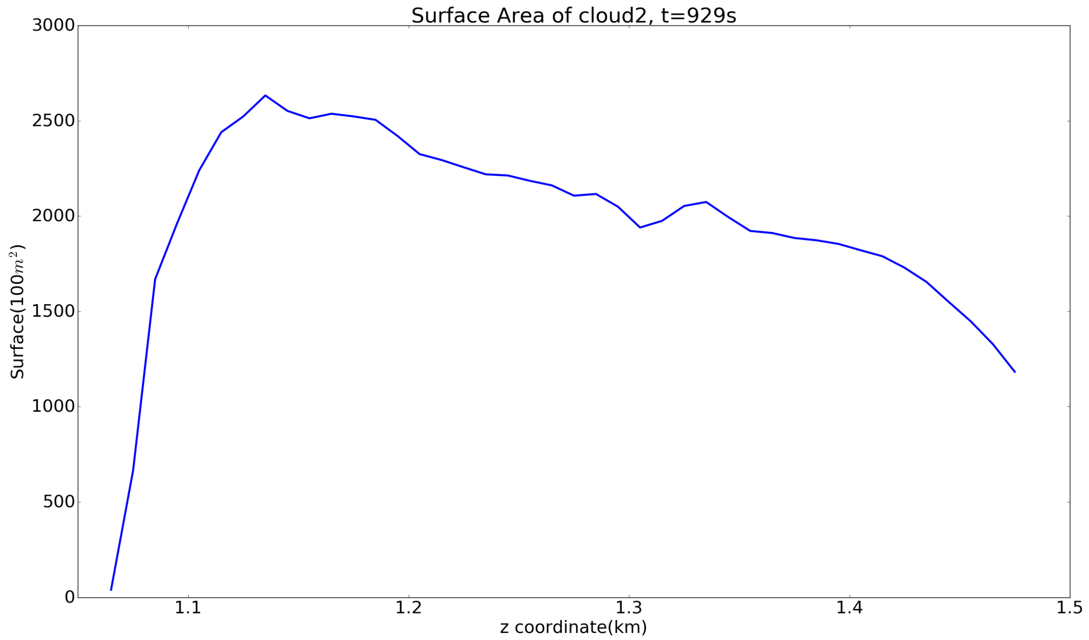


Figure A.1: Surface area of cloud two per each cross-section, at time point $t = 929\,\mathrm{s}$.

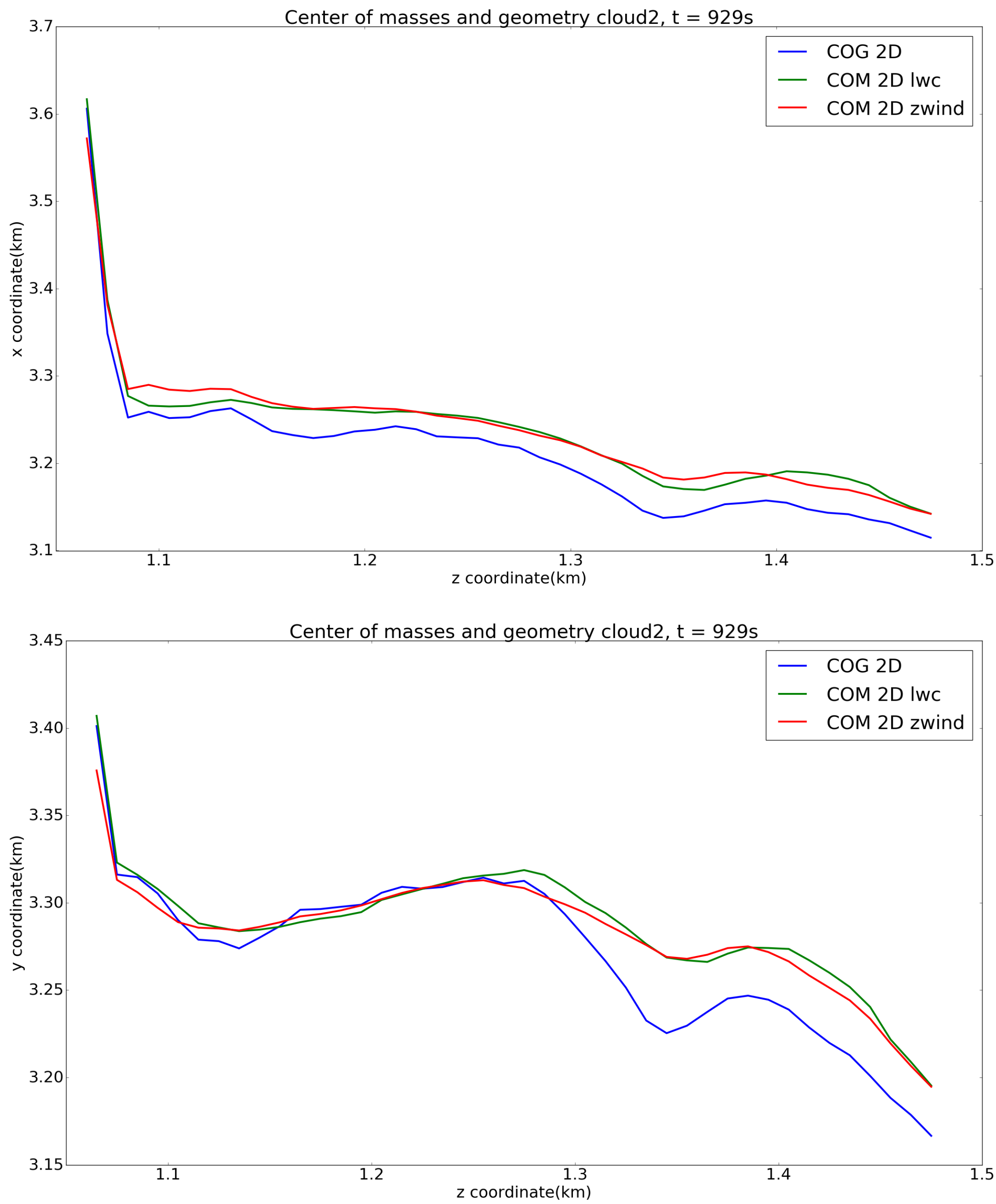


Figure A.2: Plots of the x and y coordinates of the center of masses of vertical wind and LWC. Also included the center of geometry in dependence of the height(Cloud Skeleton). These were all calculated at the same time point $t = 929\,\mathrm{s}$.

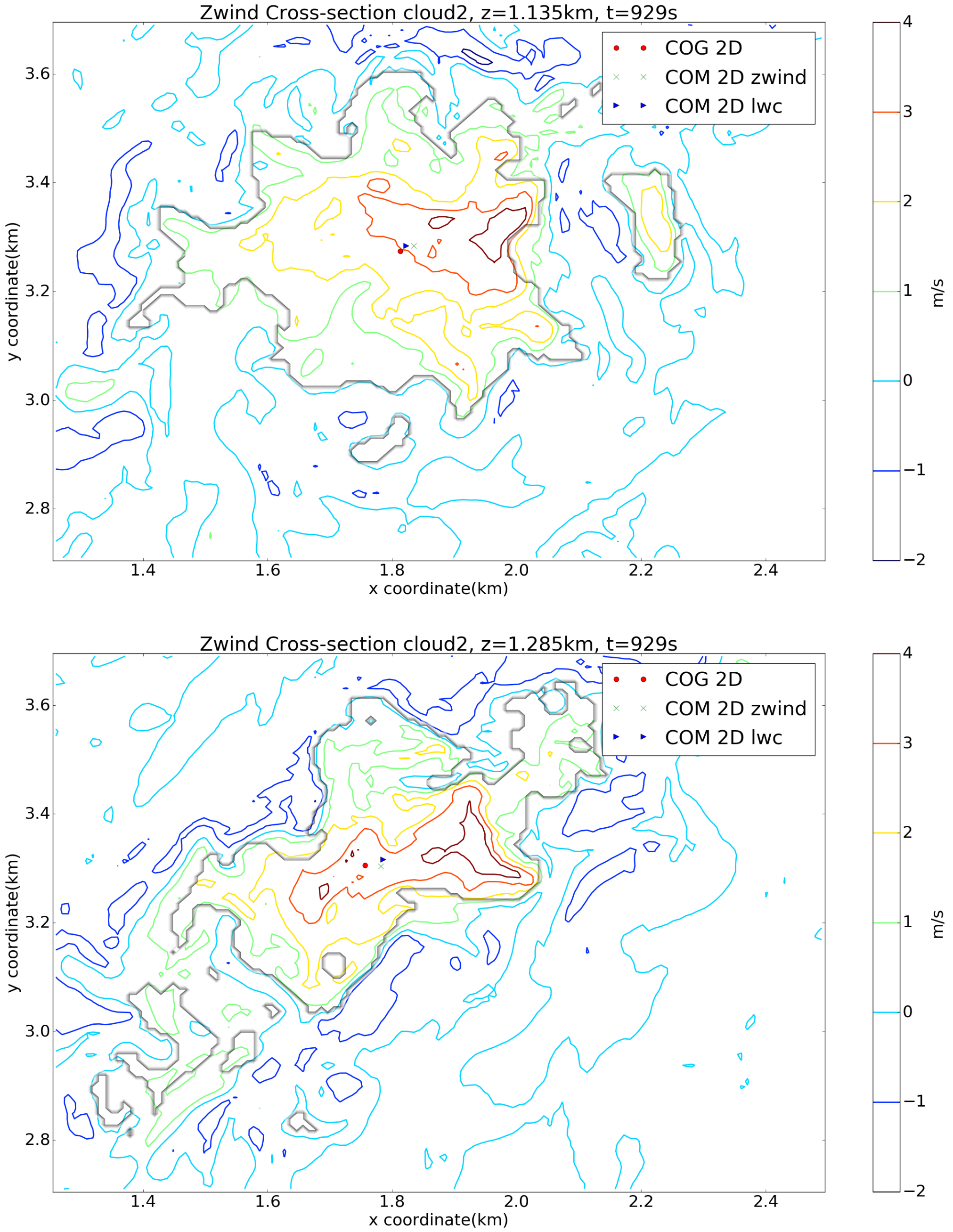
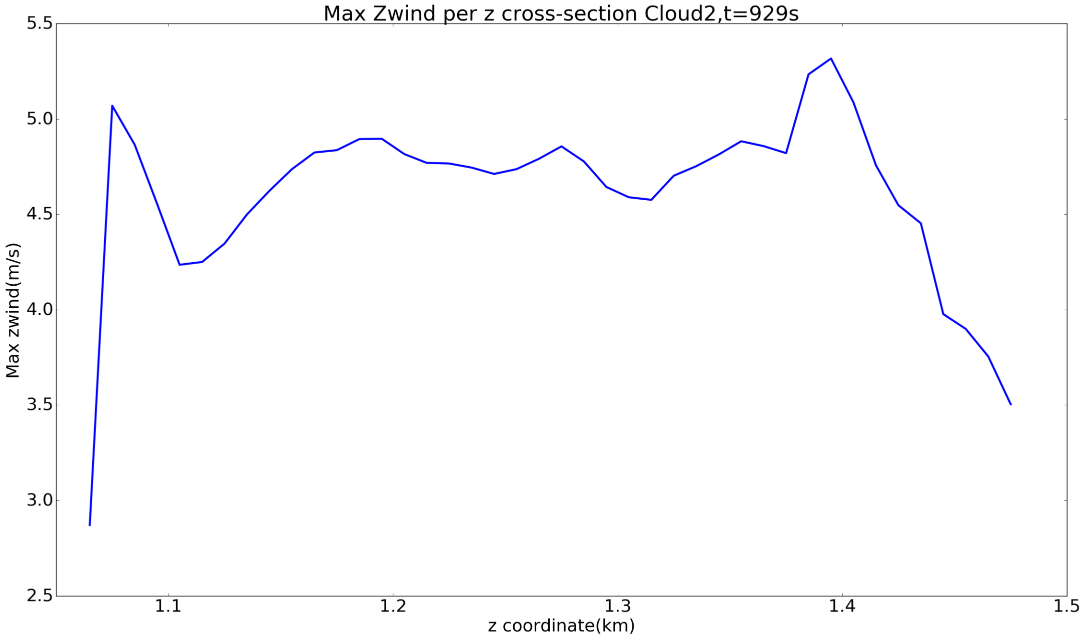


Figure A.3: Two contour plots of the vertical wind of two cross-sections of Cloud two, 150m apart in height and at the same time point $t = 929\,\mathrm{s}$. The outer contour corresponds to the boundaries of the cloud. Furthermore, the center of geometry and the center of masses of the vertical wind and liquid water content are included.



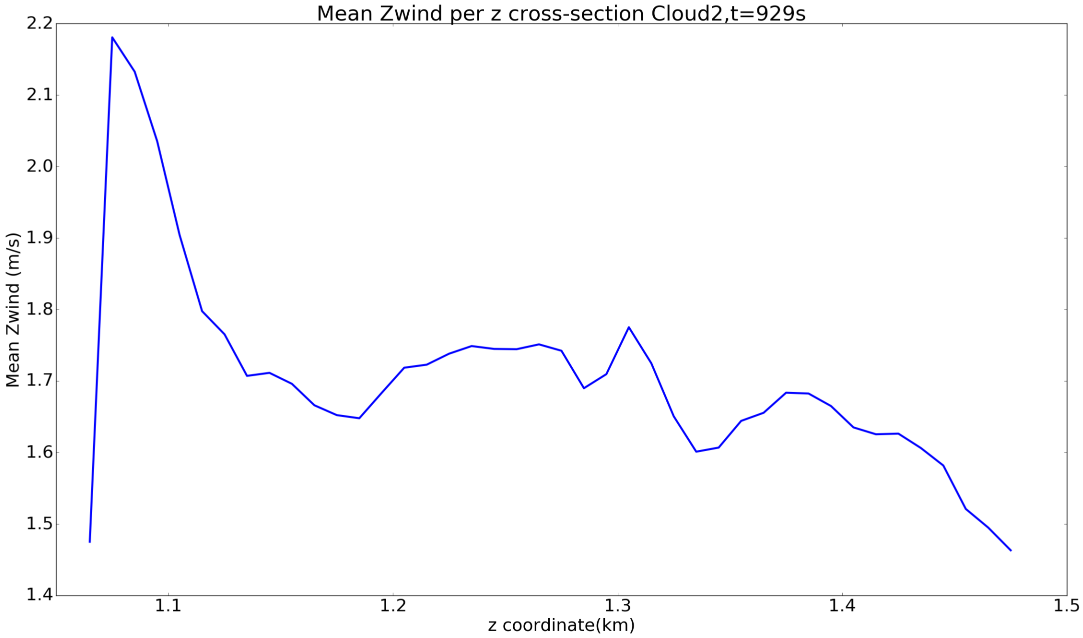


Figure A.4: Plots of the evolution of the mean and maximum of the vertical wind for each cross-section , at time point $t = 929\,\text{s}$.

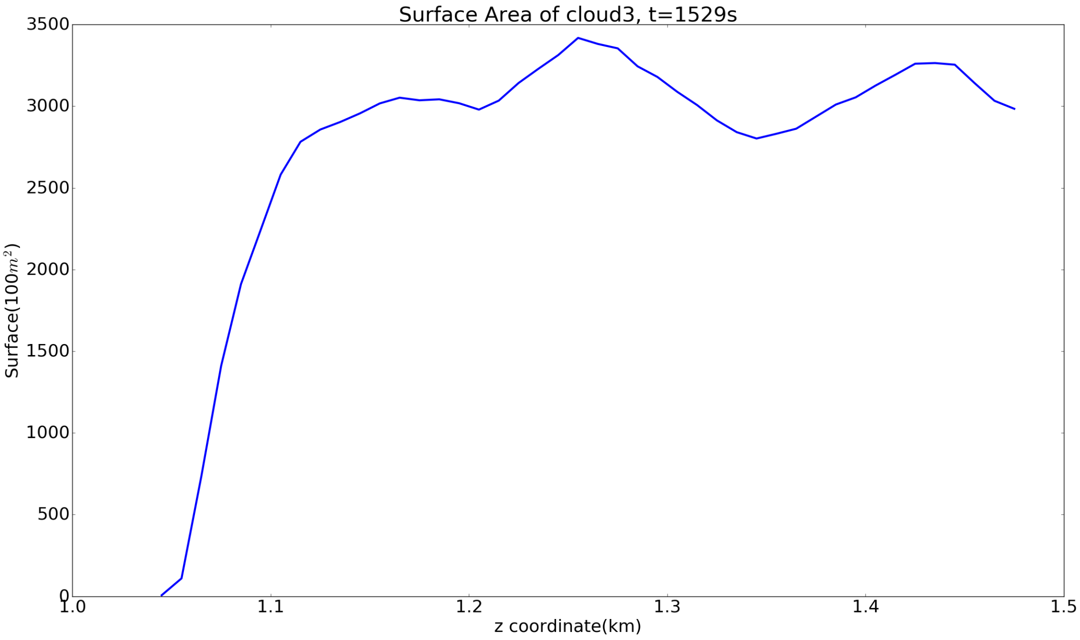
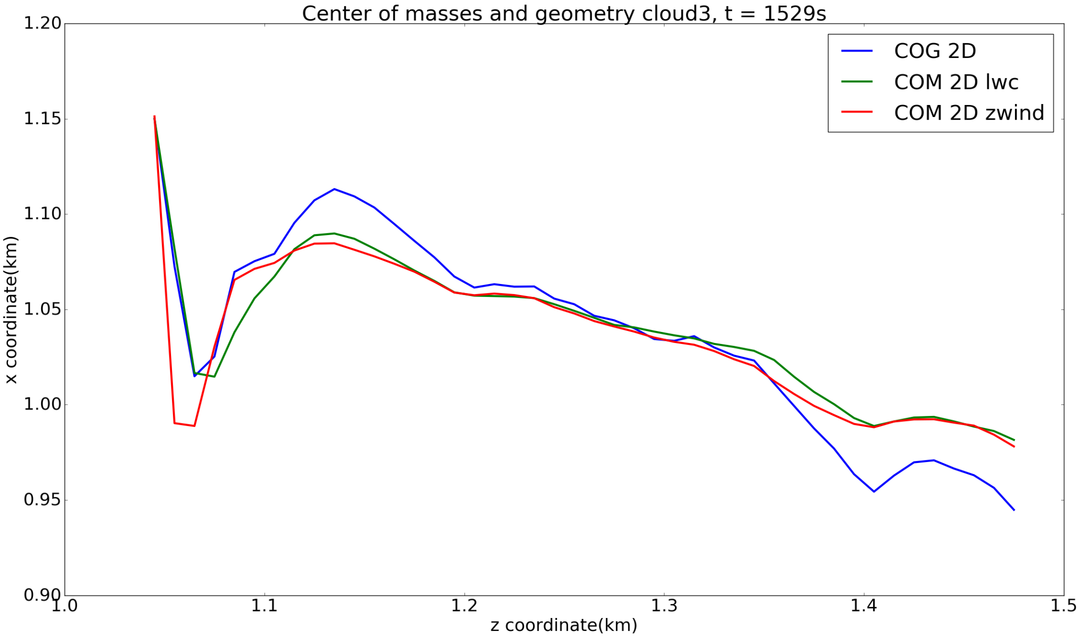


Figure A.5: Surface area of cloud three per each cross-section, at time point $t = 1529\,\mathrm{s}$.



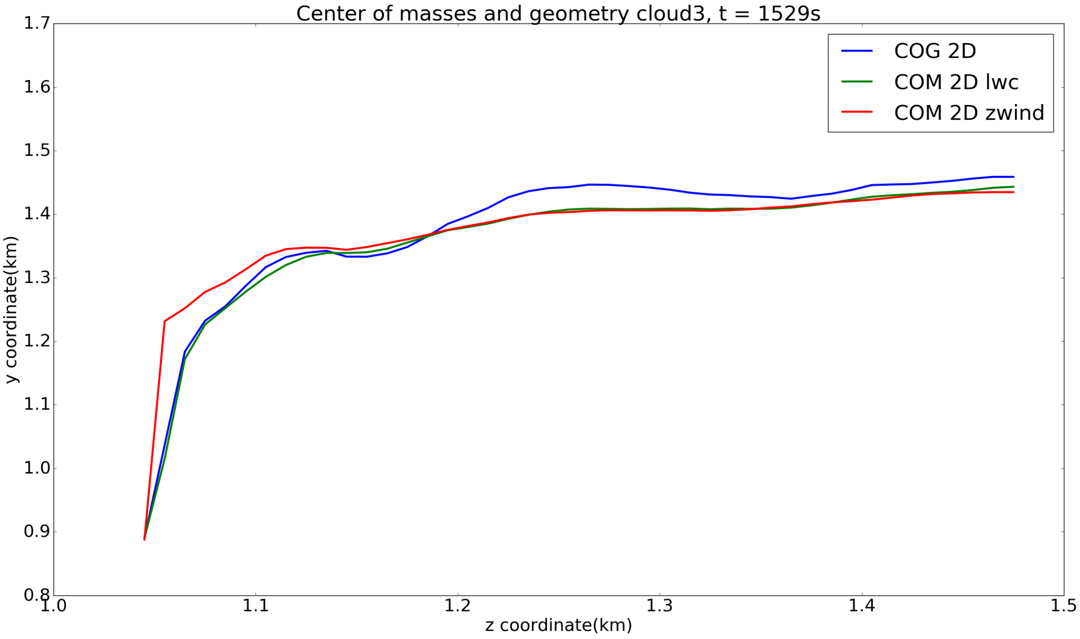


Figure A.6: Cloud three. Plots of the x and y coordinates of the center of masses of vertical wind and LWC. Also included the center of geometry in dependence of the height(Cloud Skeleton). These correspond to the same time point $t = 1529\,\mathrm{s}$.

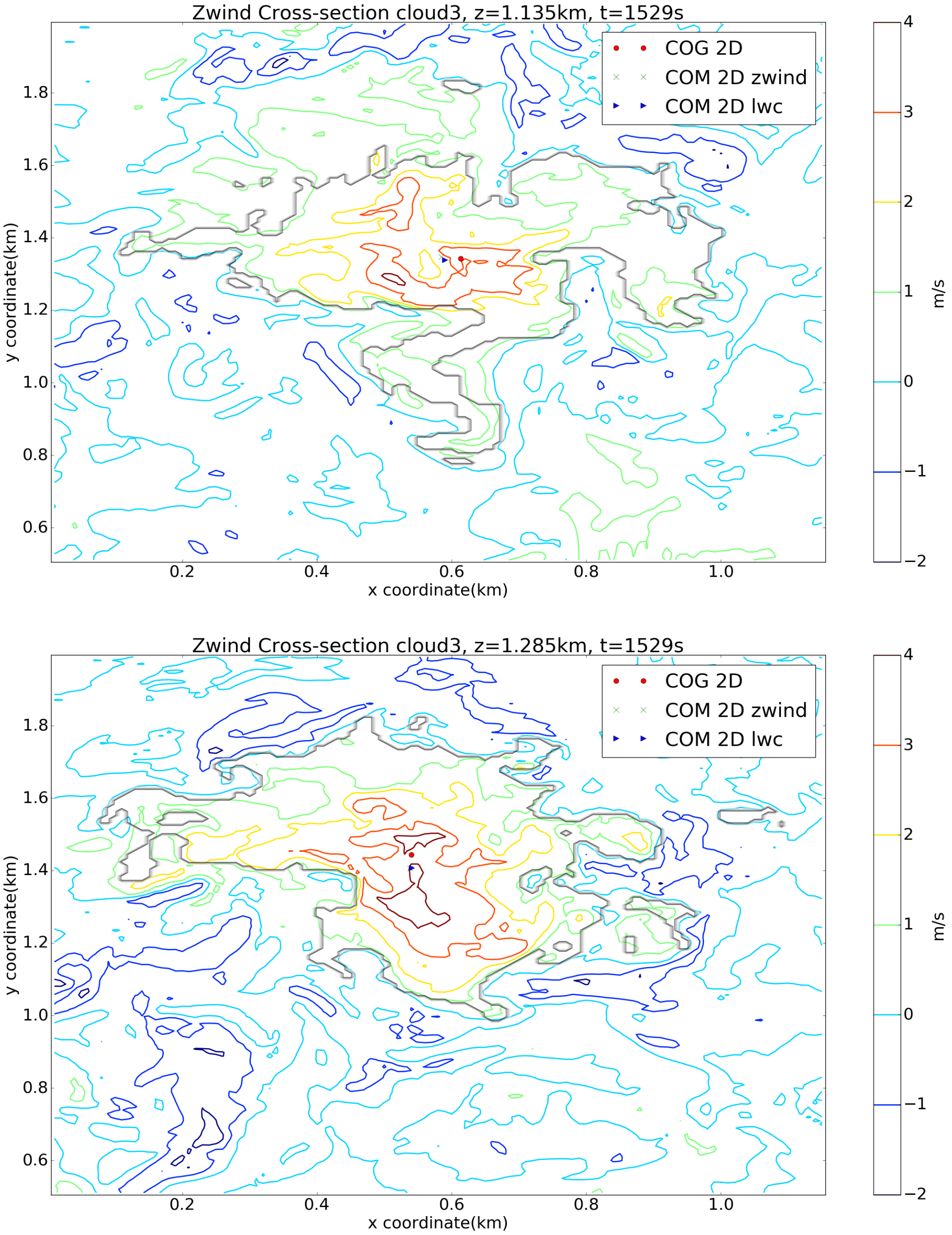


Figure A.7: Two contour plots of the vertical wind of two cross-sections of Cloud three, 150m apart in height and at the same time point $t = 1529\,\mathrm{s}$. The outer contour corresponds to the boundaries of the cloud. Furthermore, the center of geometry and the center of masses of the vertical wind and liquid water content are included.

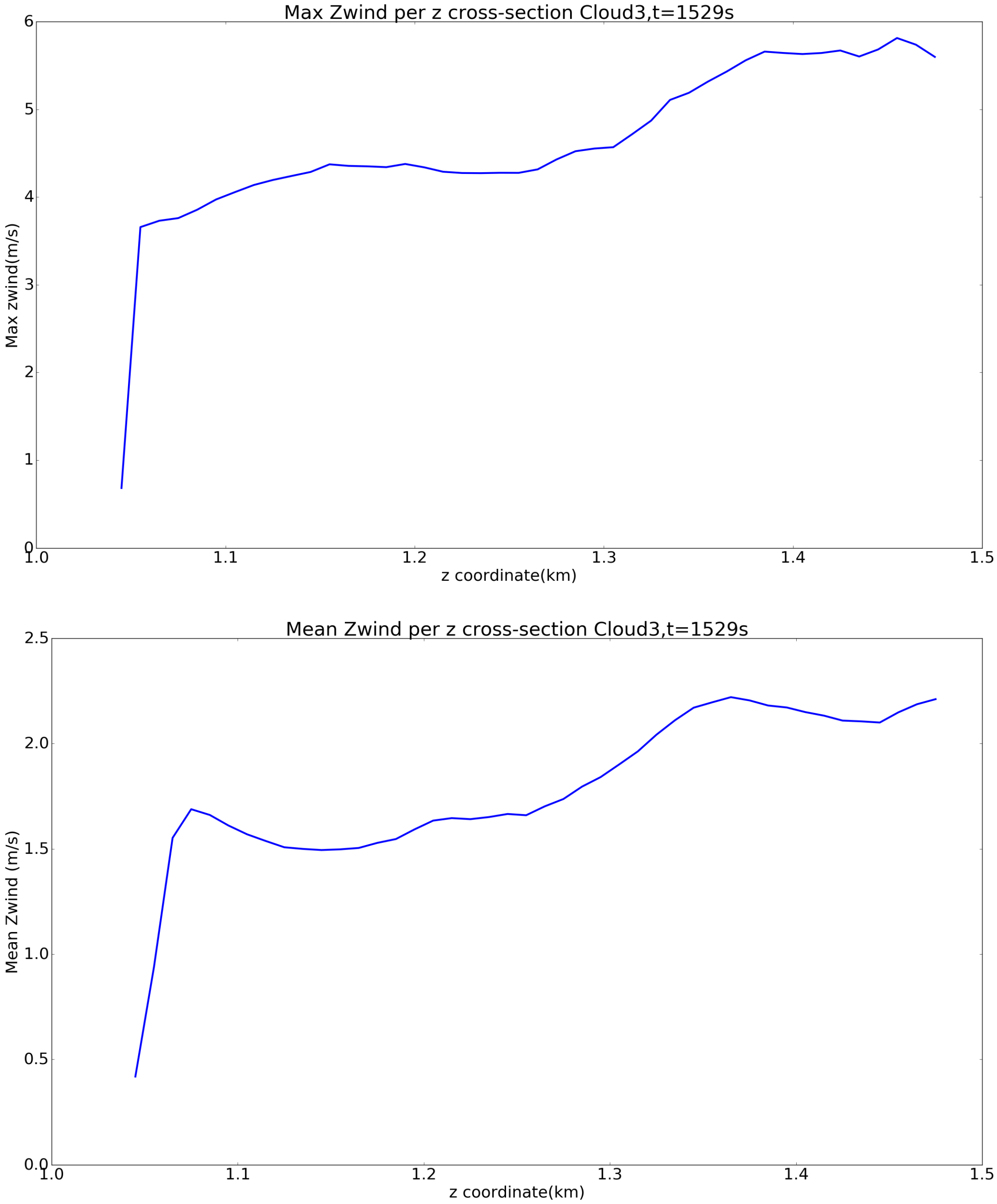


Figure A.8: Cloud three. Plots of the evolution of the mean and maximum of the vertical wind for each cross-section , at time point $t = 1529\,\mathrm{s}$.

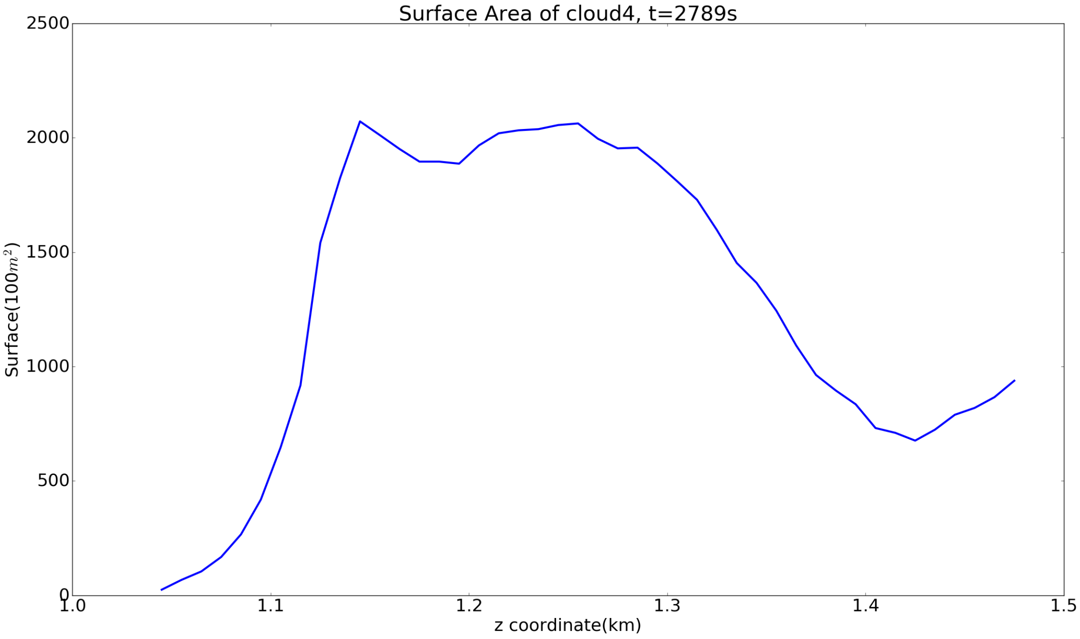


Figure A.9: Surface area of cloud four per each cross-section, at time point $t = 2789\,\mathrm{s}$.

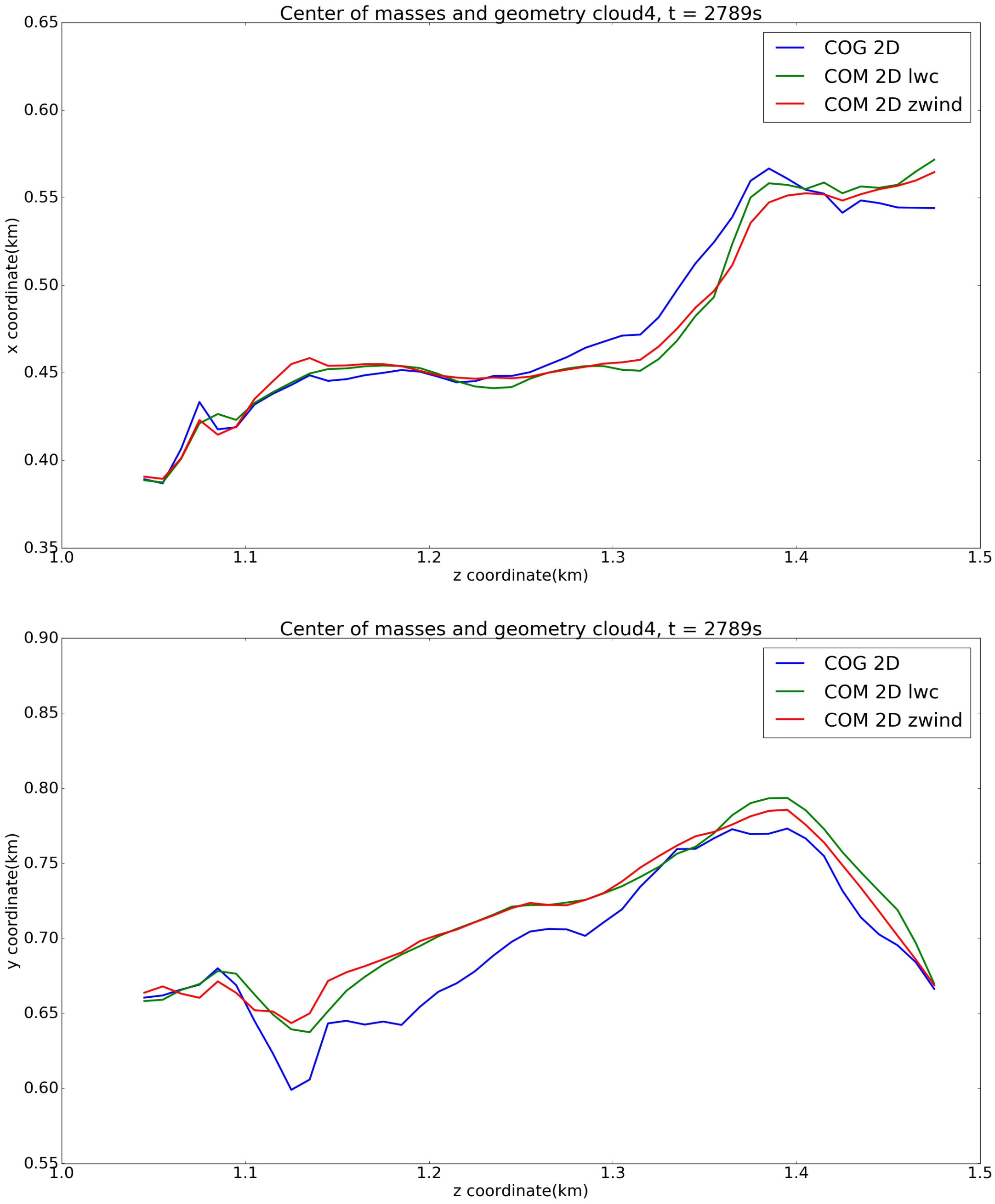


Figure A.10: Cloud four. Plots of the x and y coordinates of the center of masses of vertical wind and LWC. Also included the center of geometry in dependence of the height(Cloud Skeleton). These correspond to the same time point $t = 2789\,\mathrm{s}$.

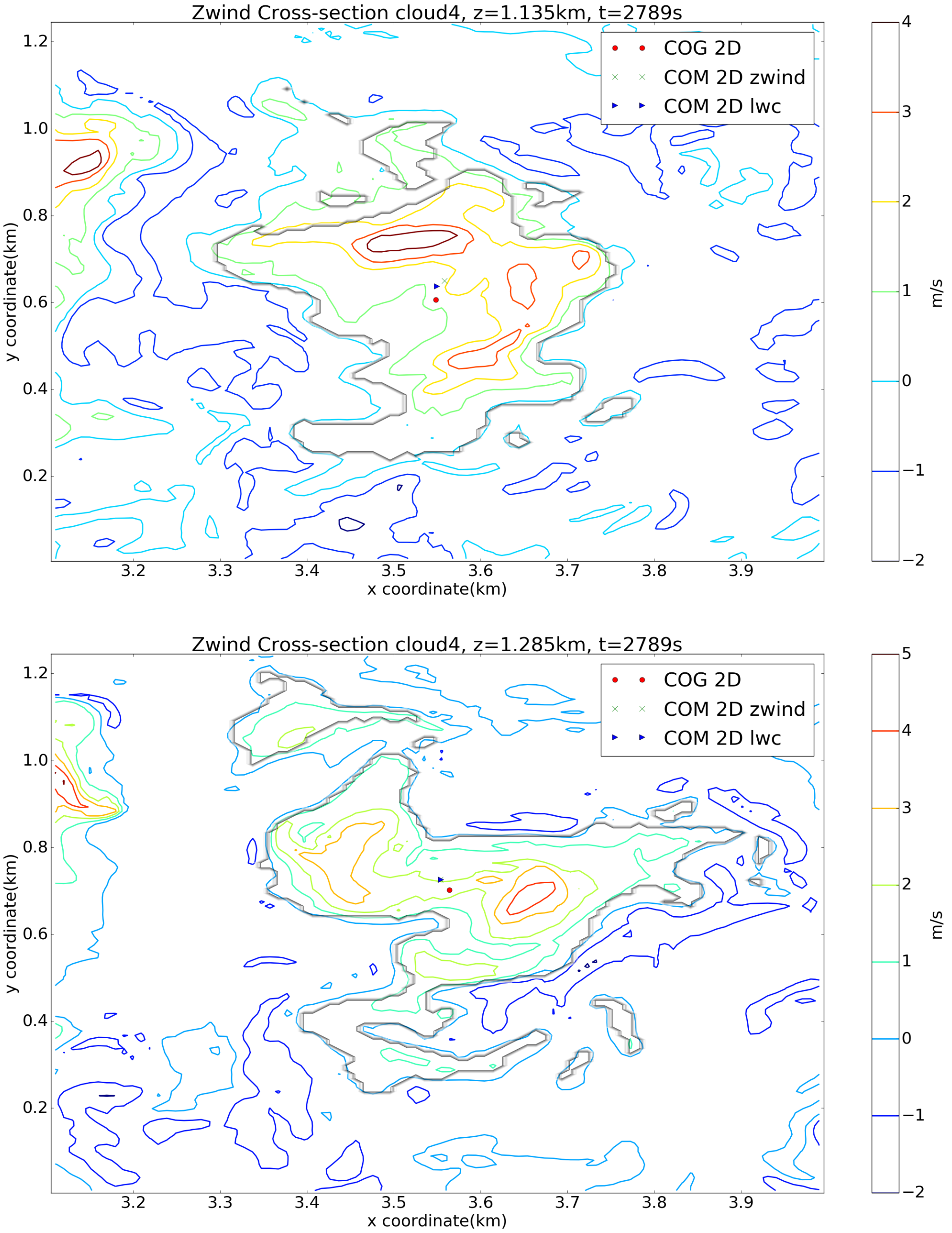


Figure A.11: Two contour plots of the vertical wind of two cross-sections of Cloud four, 150m apart in height and at the same time point $t = 2789\,\text{s}$. The outer contour corresponds to the boundaries of the cloud. Furthermore, the center of geometry and the center of masses of the vertical wind and liquid water content are included.

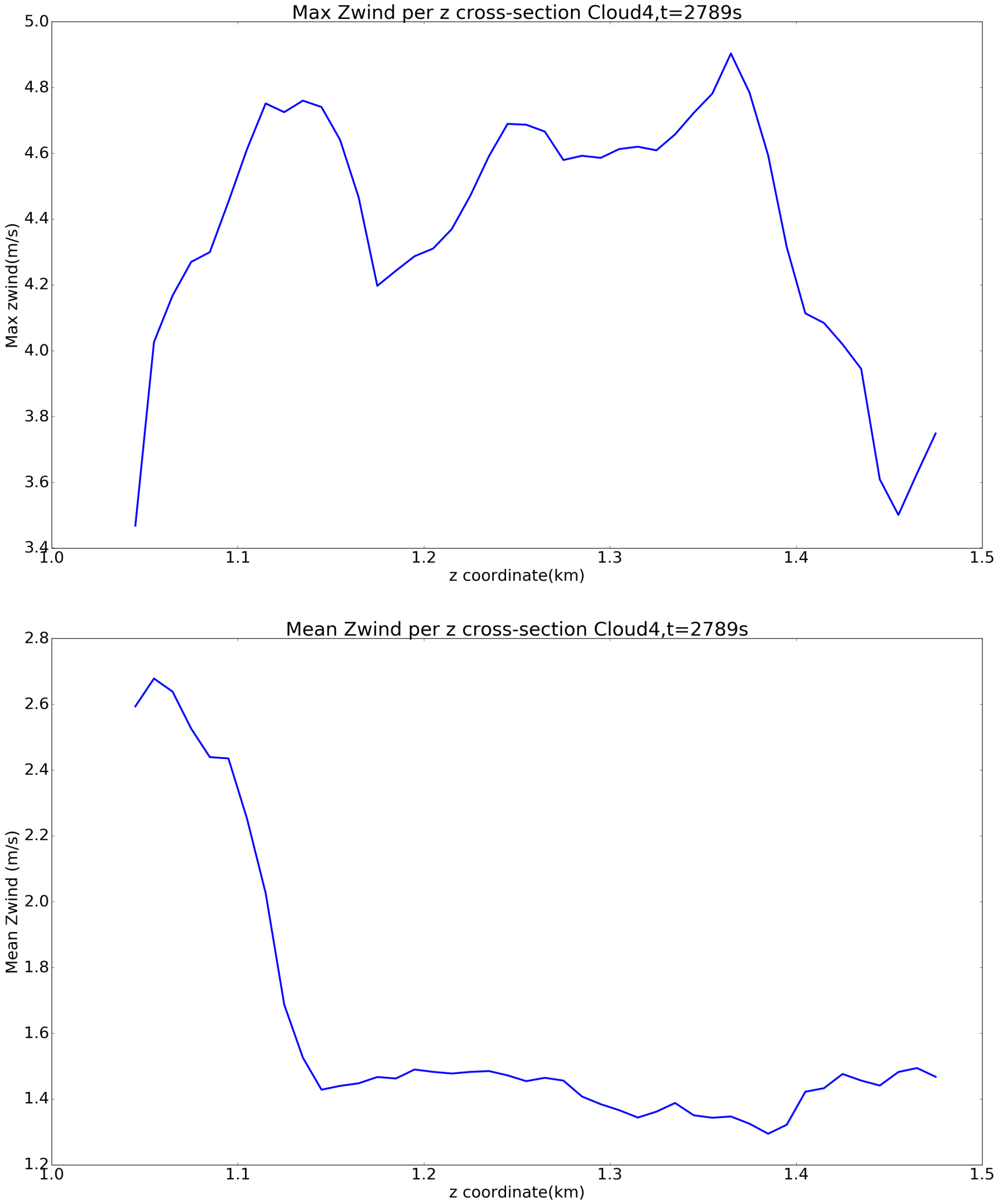


Figure A.12: Cloud Four. Plots of the evolution of the mean and maximum of the vertical wind for each cross-section , at time point $t = 2789\,\text{s}$.

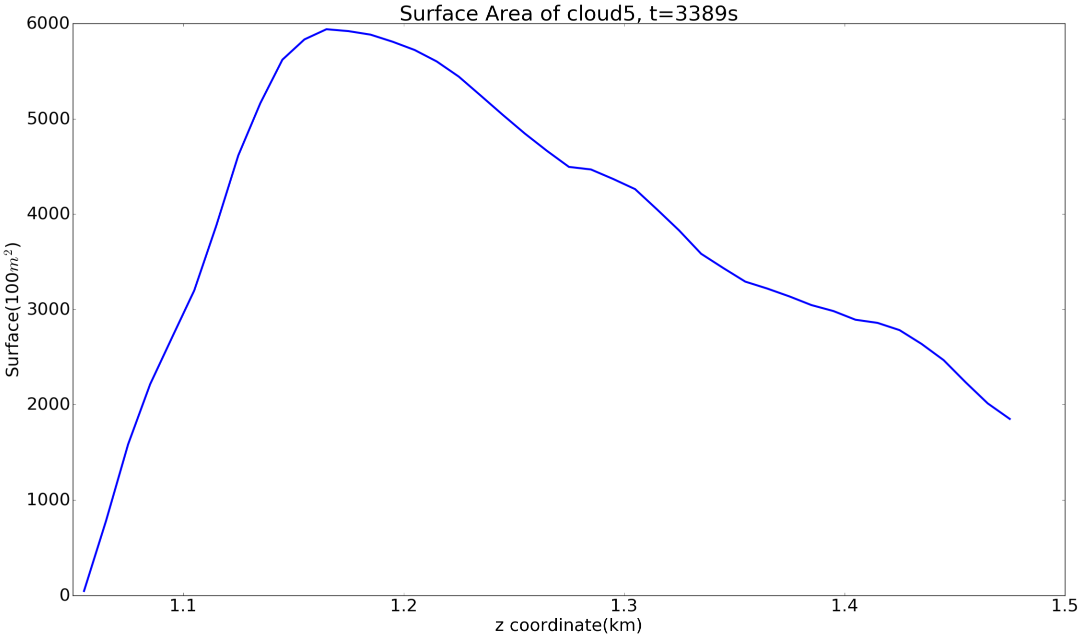
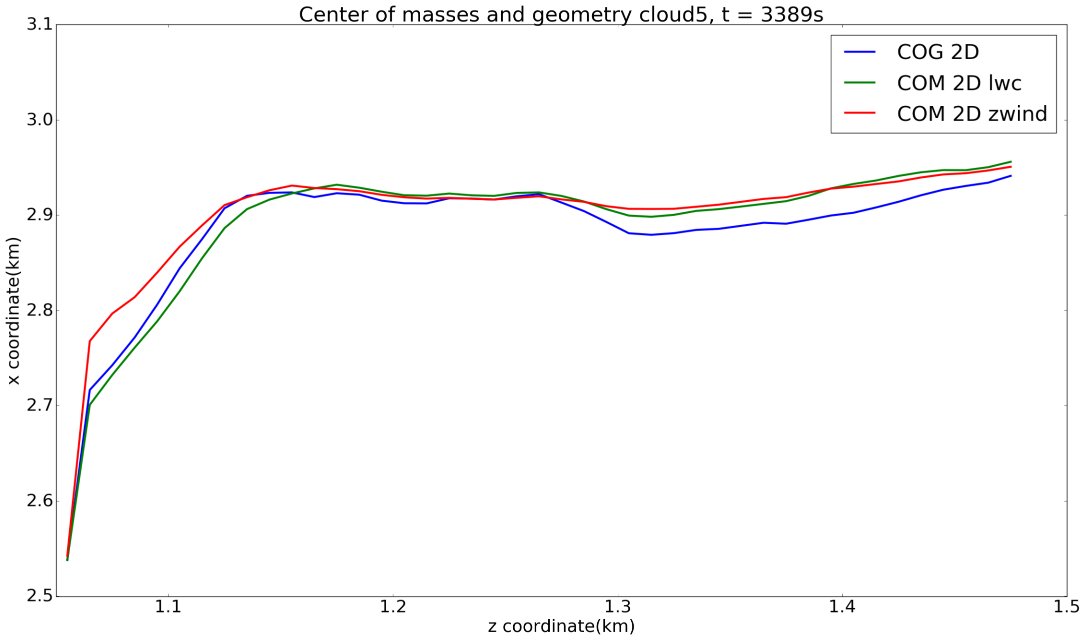


Figure A.13: Surface area of cloud five per each cross-section, at time point $t = 3389\,\mathrm{s}$.



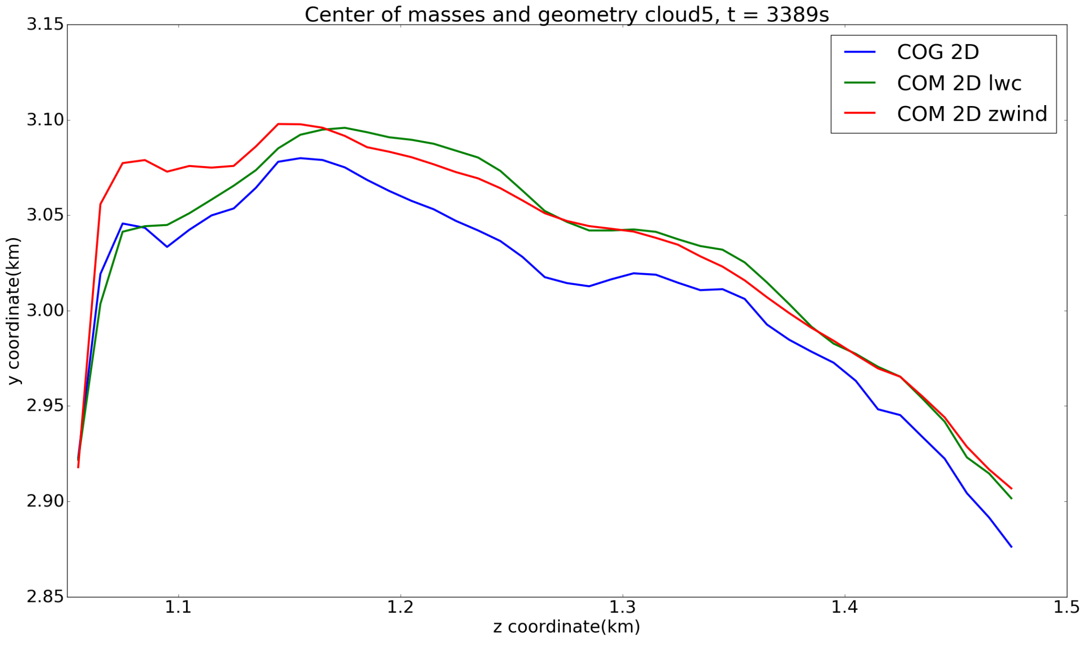
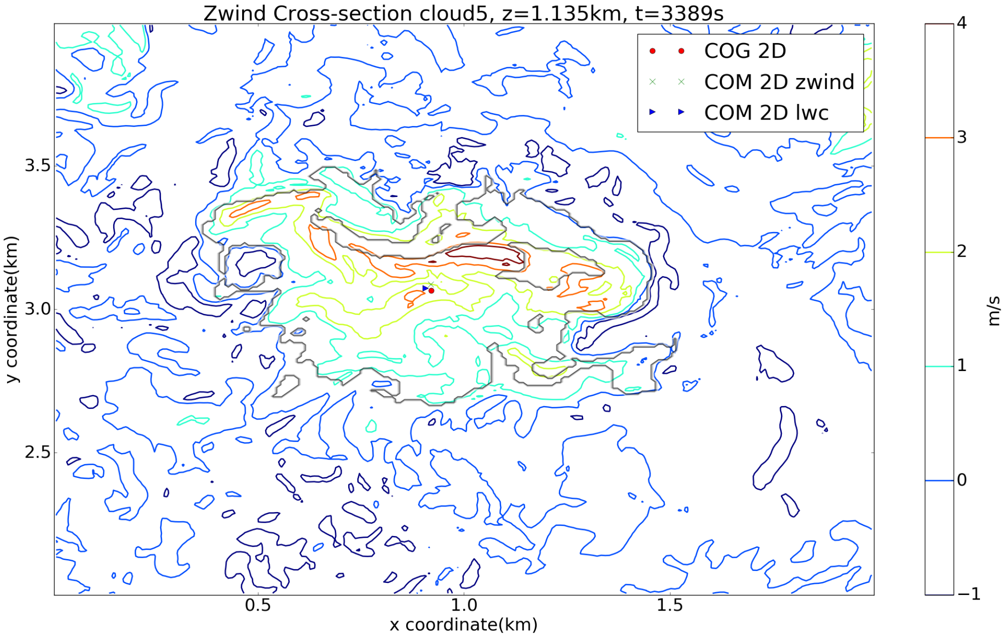


Figure A.14: Cloud five. Plots of the x and y coordinates of the center of masses of vertical wind and LWC. Also included the center of geometry in dependence of the height(Cloud Skeleton). These correspond to the same time point $t = 3389\,\mathrm{s}$.



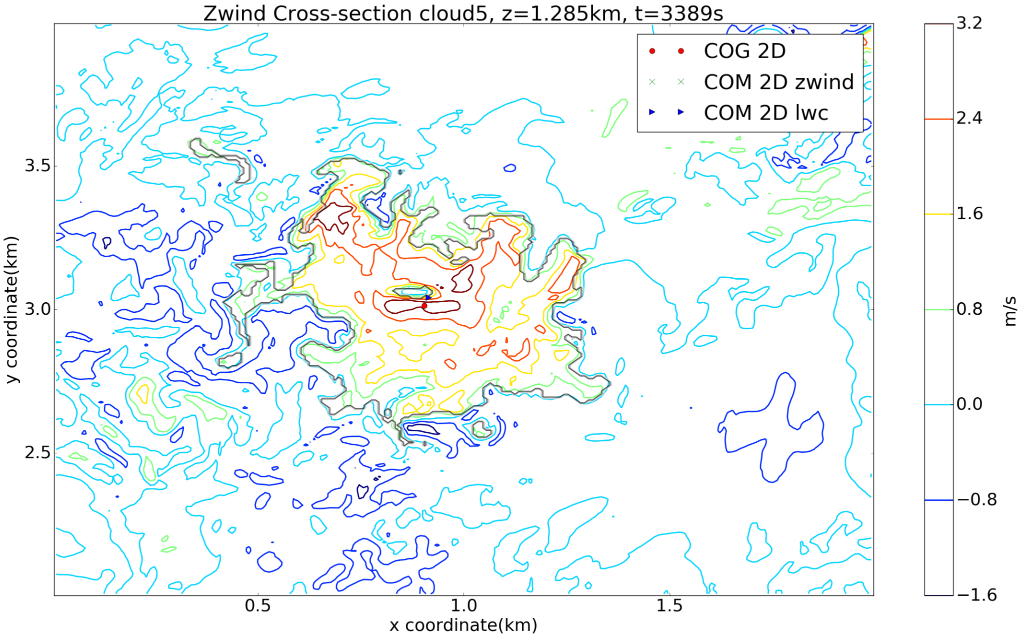


Figure A.15: Two contour plots of the vertical wind of two cross-sections of Cloud five, 150m apart in height and at the same time point $t = 3389\,\mathrm{s}$. The outer contour corresponds to the boundaries of the cloud. Furthermore, the center of geometry and the center of masses of the vertical wind and liquid water content are included.

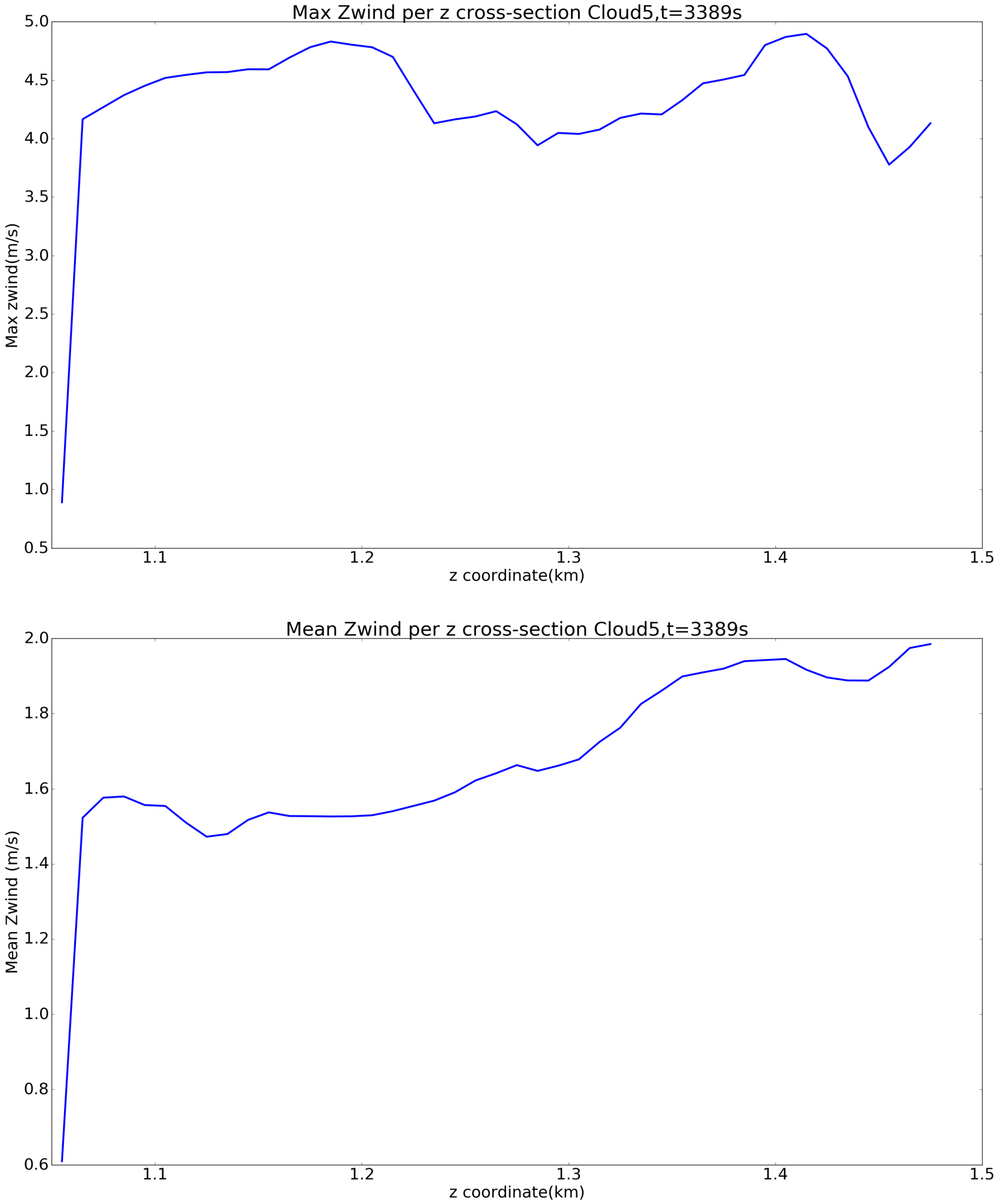


Figure A.16: Cloud Five. Plots of the evolution of the mean and maximum of the vertical wind for each cross-section , at time point $t = 3389\,\mathrm{s}$.

A.2 Initial Empirical Variograms, Cloud Two to Five

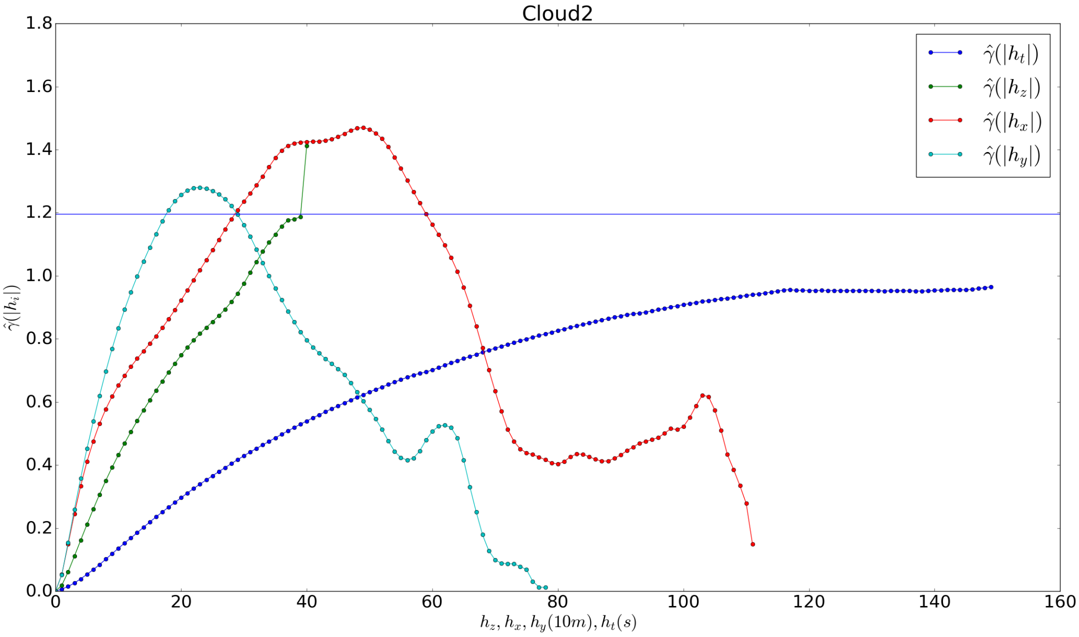


Figure A.17: Empirical variograms in the four directions for cloud two.

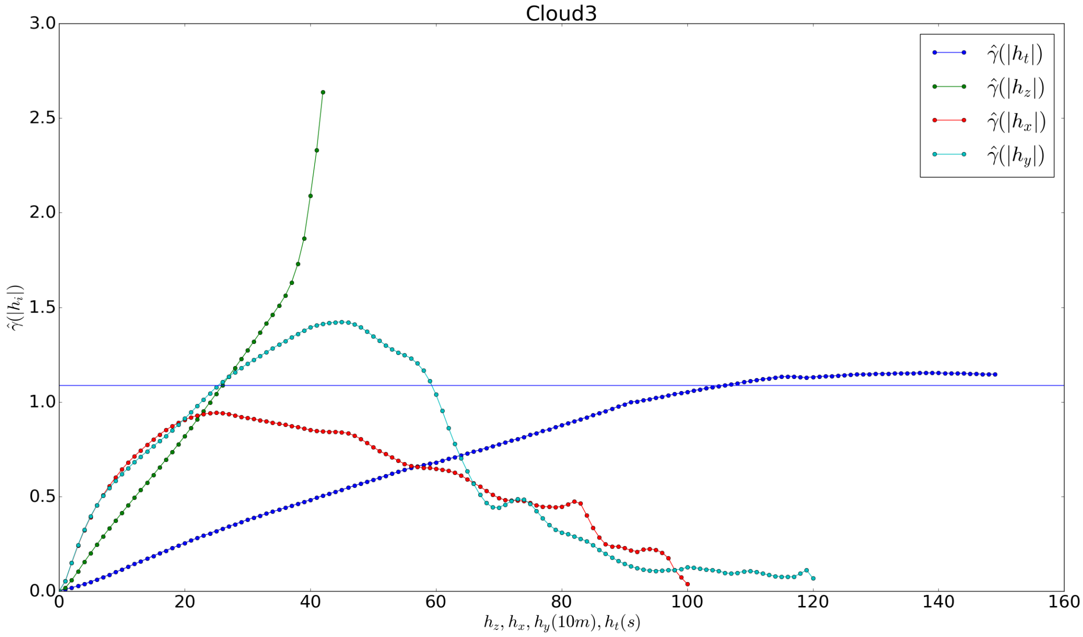


Figure A.18: Empirical variograms in the four directions for cloud three.

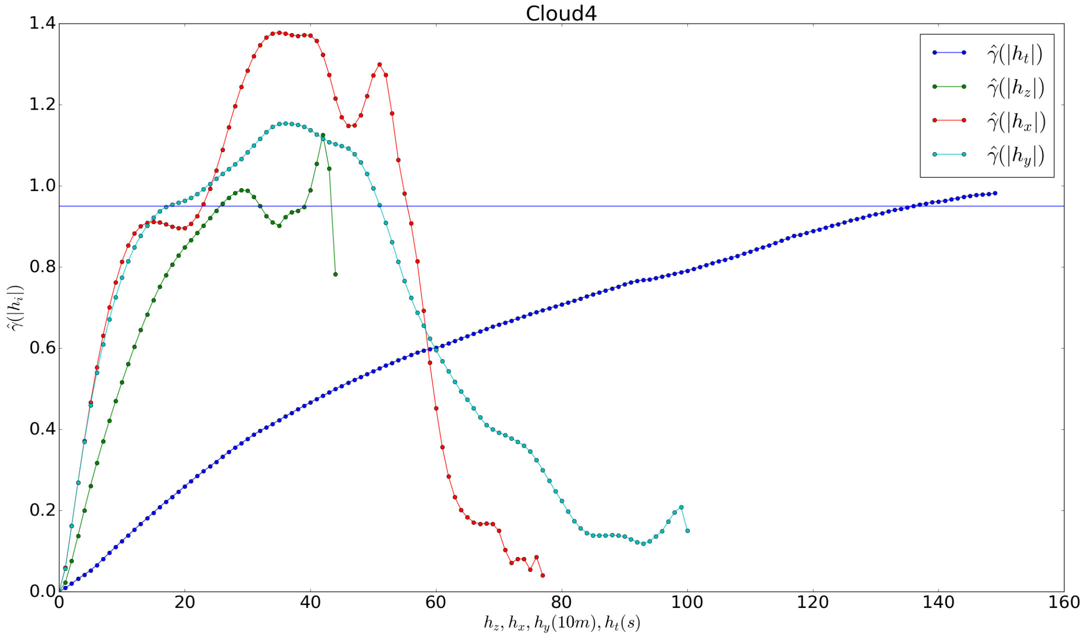


Figure A.19: Empirical variograms in the four directions for cloud four.

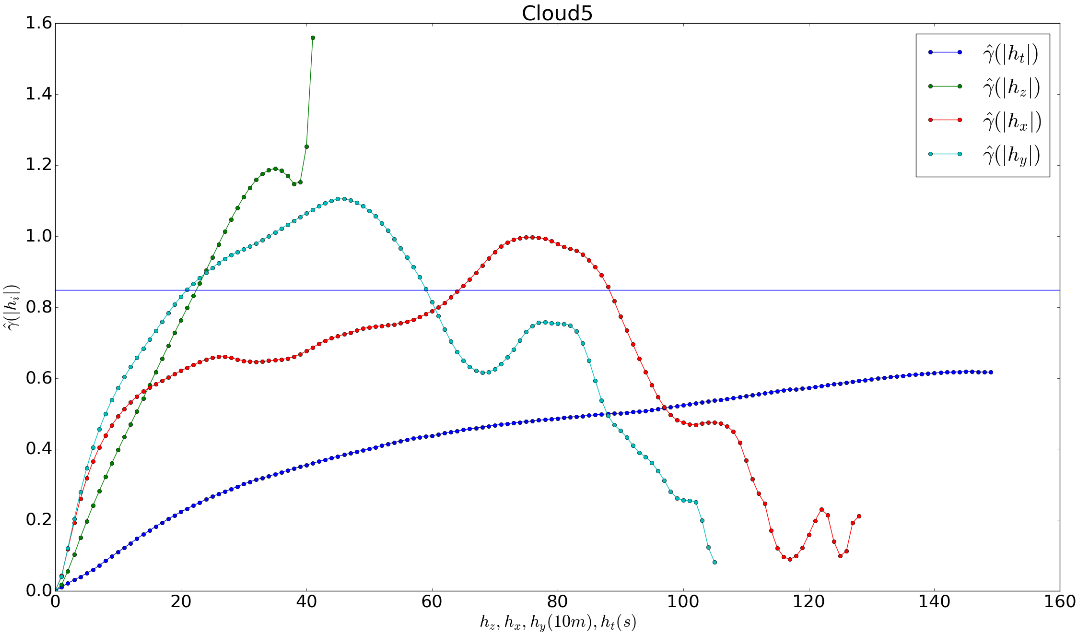


Figure A.20: Empirical variograms in the four directions for cloud five.

A.3 Radial Trends, Cloud Two to Five

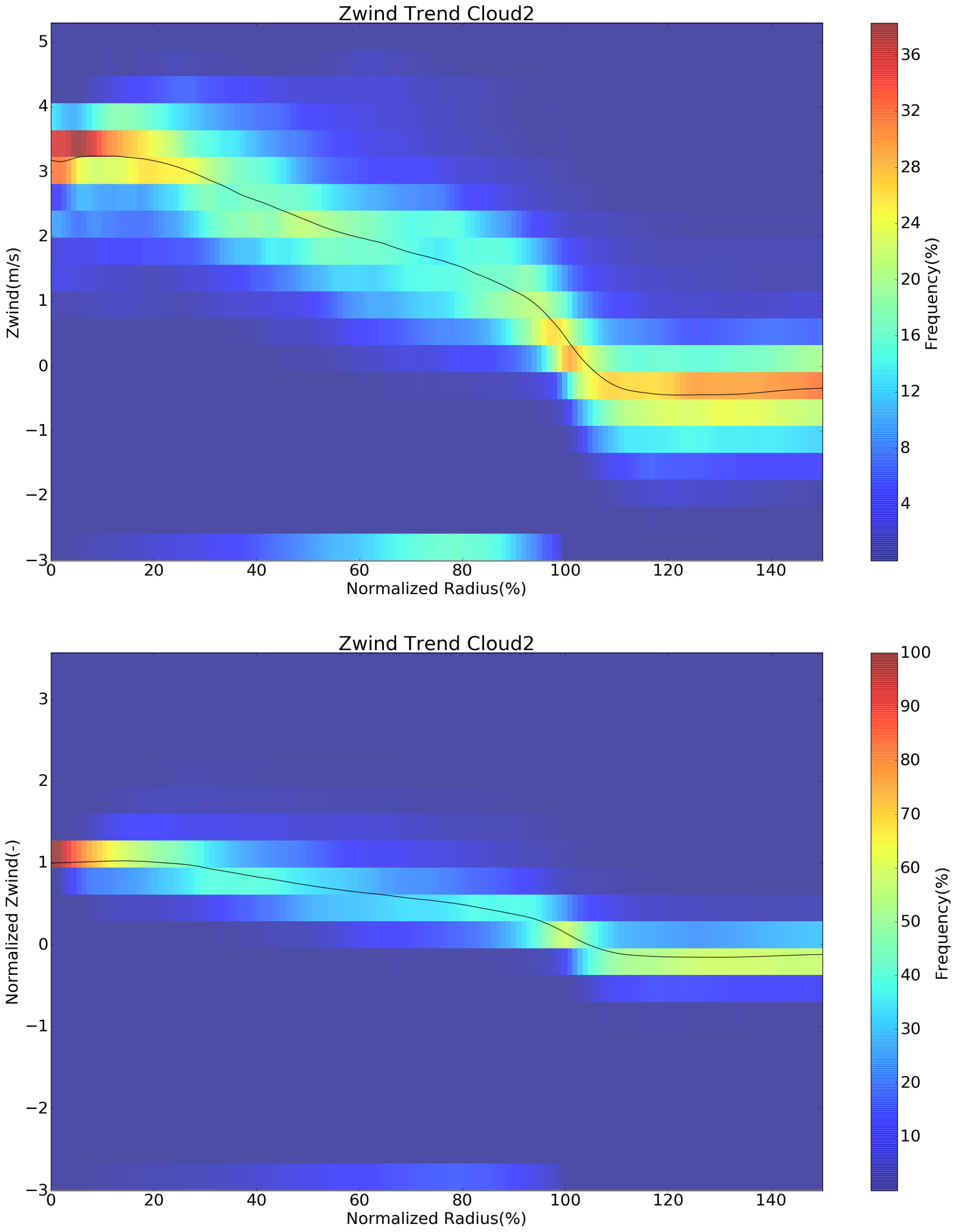
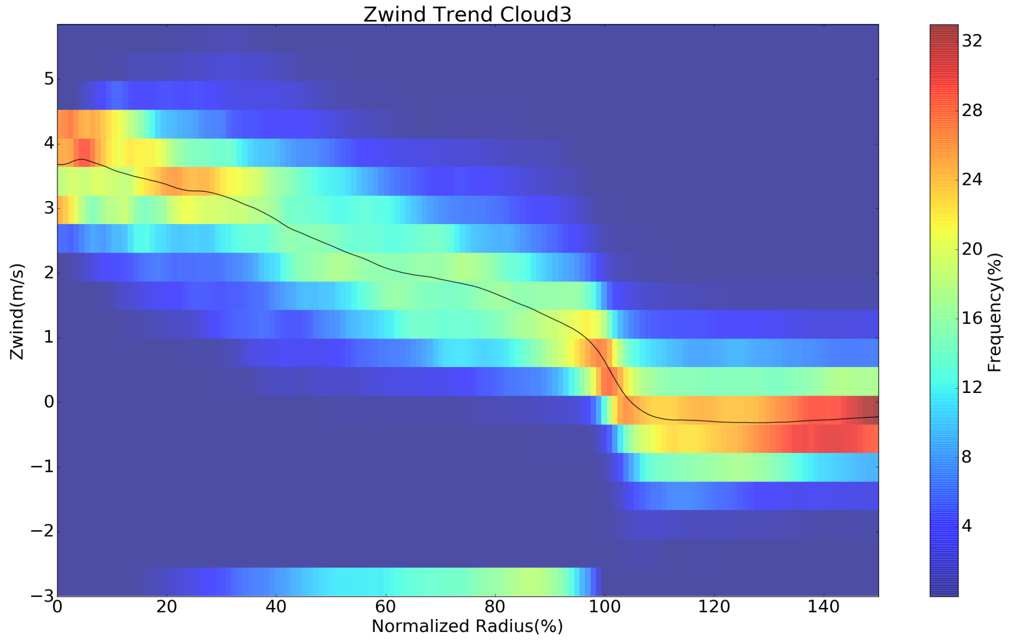


Figure A.21: 2D Histograms of the radial trend for cloud two. Top: normalized radius up to 150%. Bottom: Normalized radius up to 150% and all trend wind values normalized by the value measured at the center of mass of the cross-section



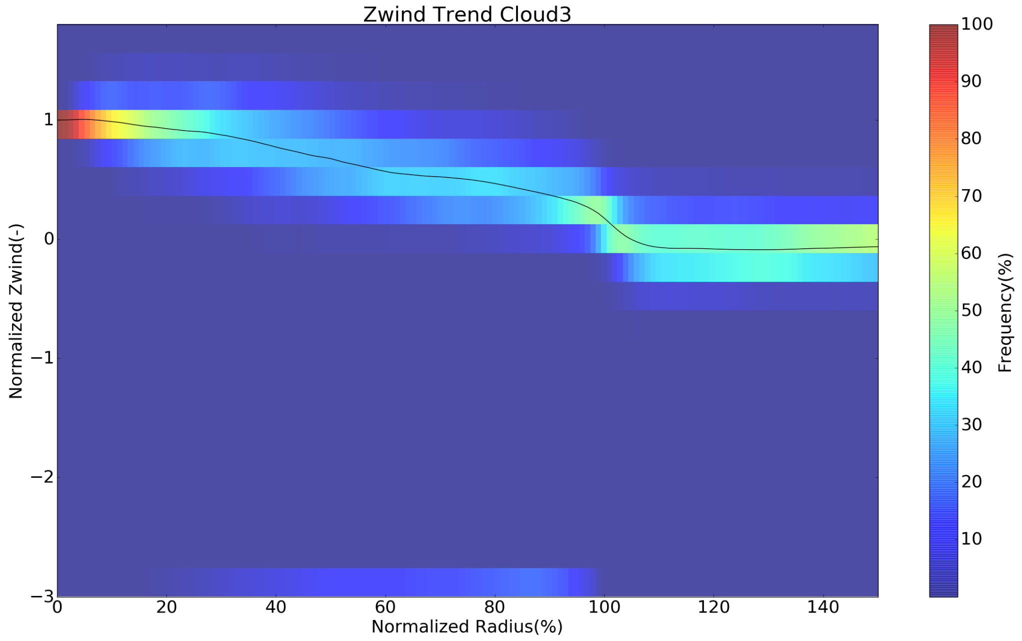


Figure A.22: 2D Histograms of the radial trend for cloud three. Top: normalized radius up to 150%. Bottom: Normalized radius up to 150% and all trend wind values normalized by the value measured at the center of mass of the cross-section

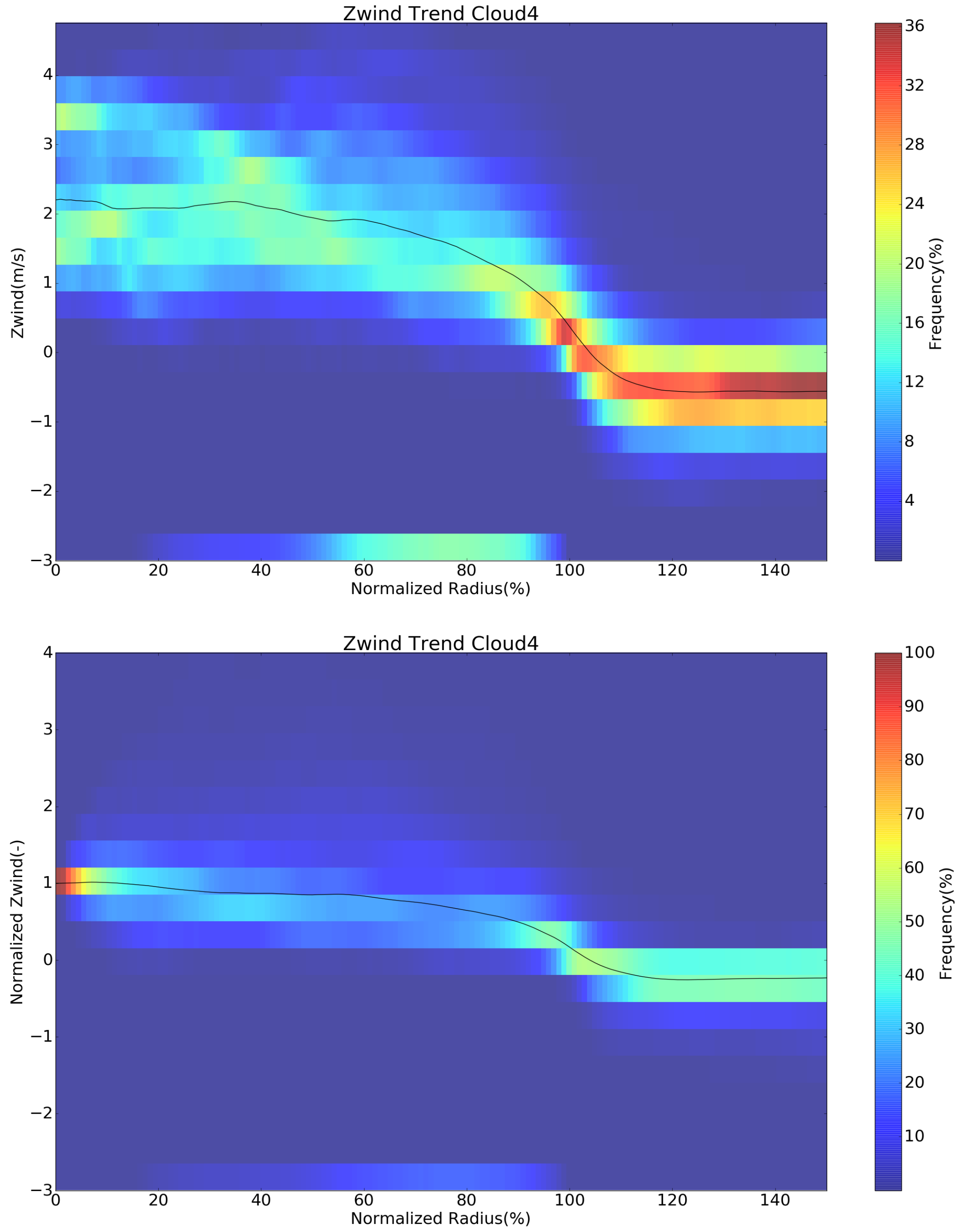
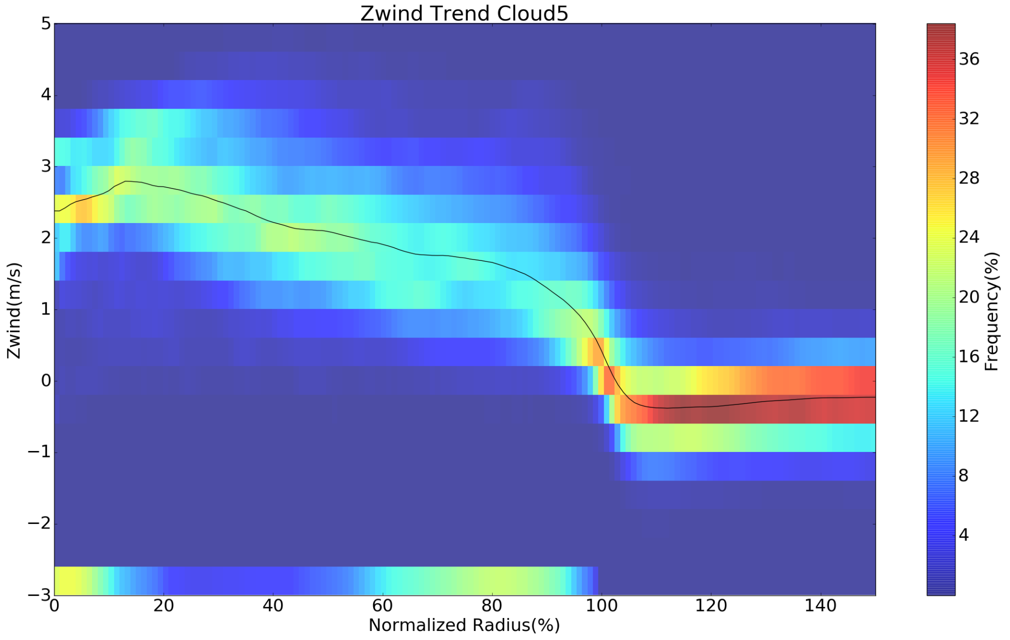


Figure A.23: 2D Histograms of the radial trend for cloud four. Top: normalized radius up to 150%. Bottom: Normalized radius up to 150% and all trend wind values normalized by the value measured at the center of mass of the cross-section



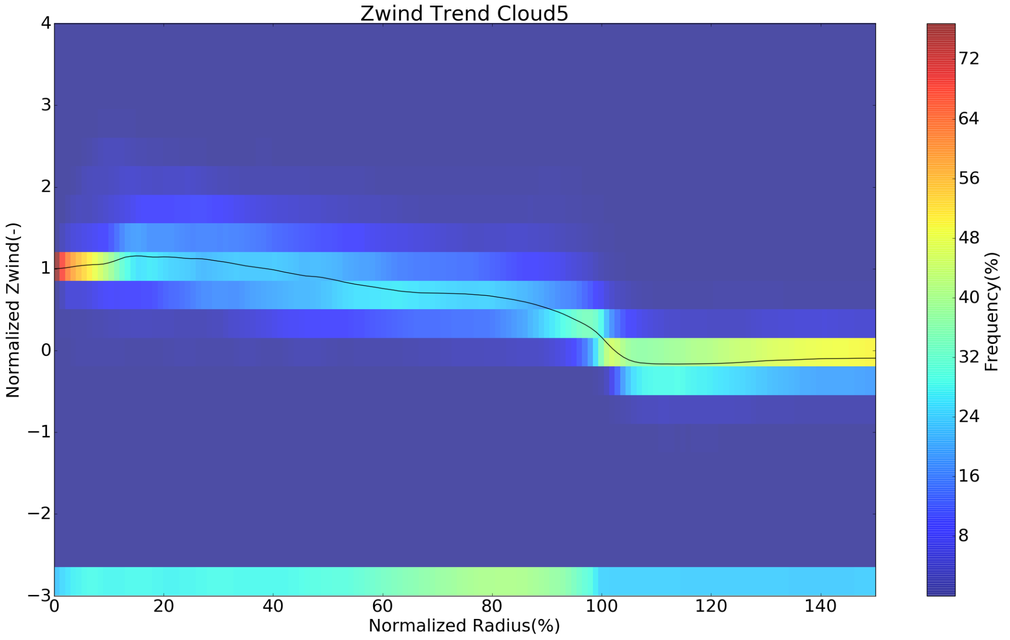


Figure A.24: 2D Histograms of the radial trend for cloud five. Top: normalized radius up to 150%. Bottom: Normalized radius up to 150% and all trend wind values normalized by the value measured at the center of mass of the cross-section

A.4 Appendix to the Experiments

A.4.1 Summary of Results, other Noise Levels

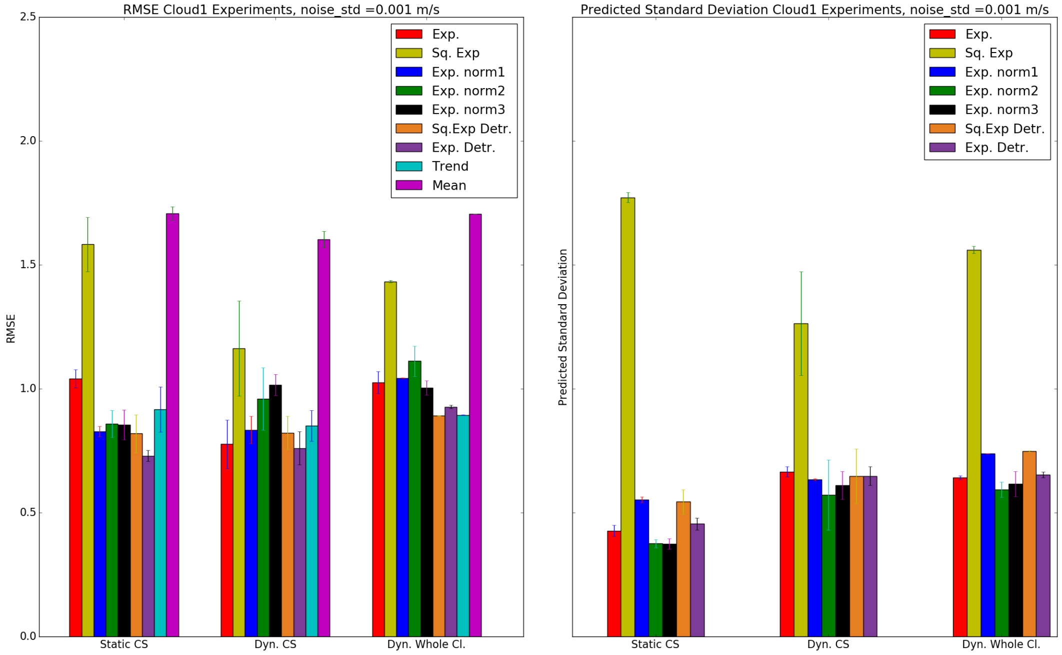


Figure A.25: Summary of RMSE and predicted standard deviation for actual noise level $\sigma_{noise} = 0.001\, m/s$, all mapping variants for the three scenarios.

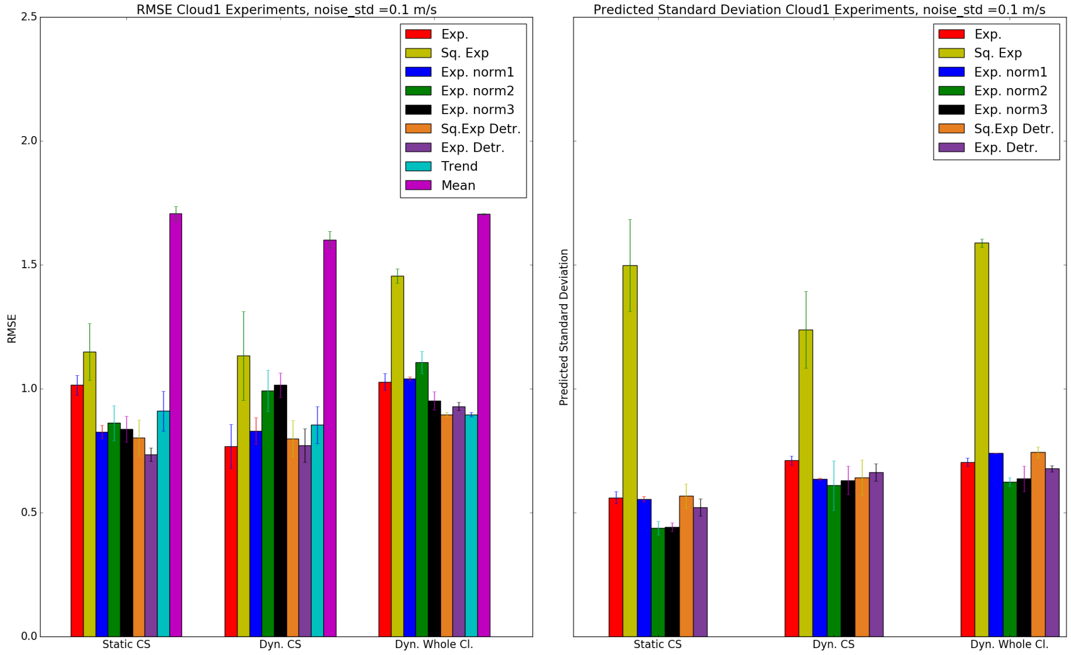


Figure A.26: Summary of RMSE and predicted standard deviation for actual noise level $\sigma_{noise} = 0.1\, m/s$, all mapping variants for the three scenarios.

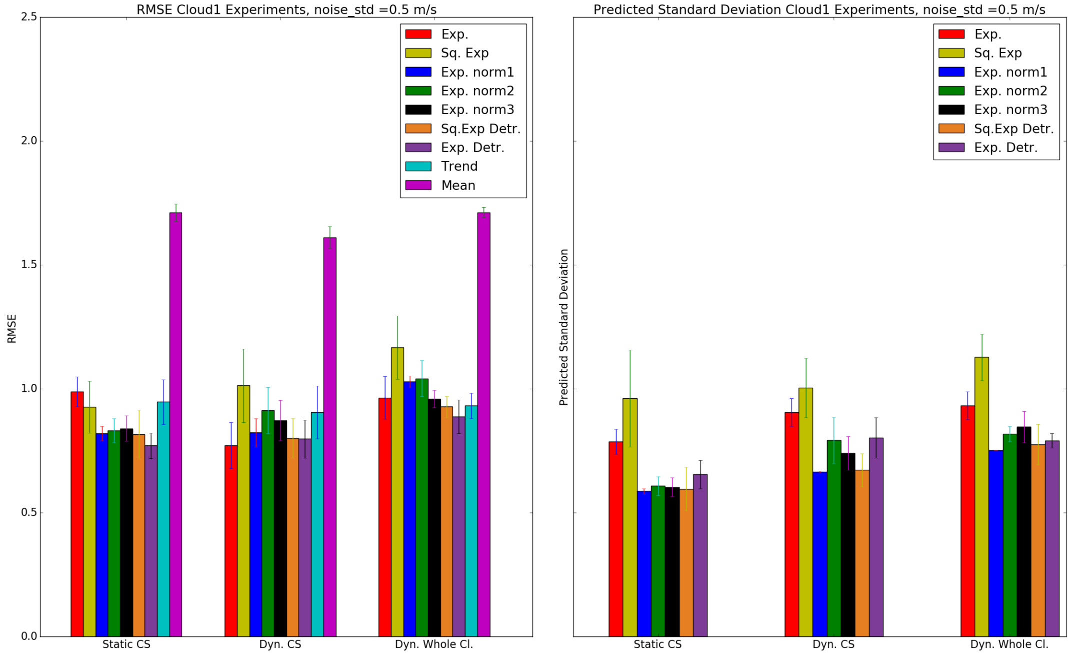


Figure A.27: Summary of RMSE and predicted standard deviation for actual noise level $\sigma_{noise} = 0.5\, m/s$, all mapping variants for the three scenarios.

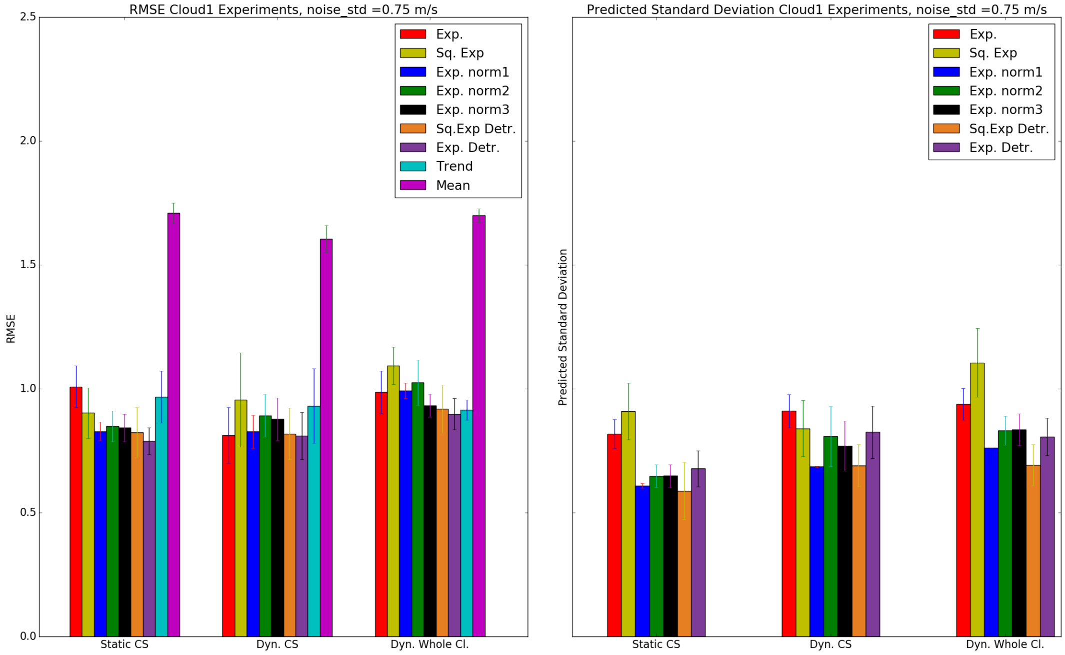


Figure A.28: Summary of RMSE and predicted standard deviation for actual noise level $\sigma_{noise} = 0.75\, m/s$, all mapping variants for the three scenarios.

A.4.2 Static Cross-section Comparison, other Noise levels

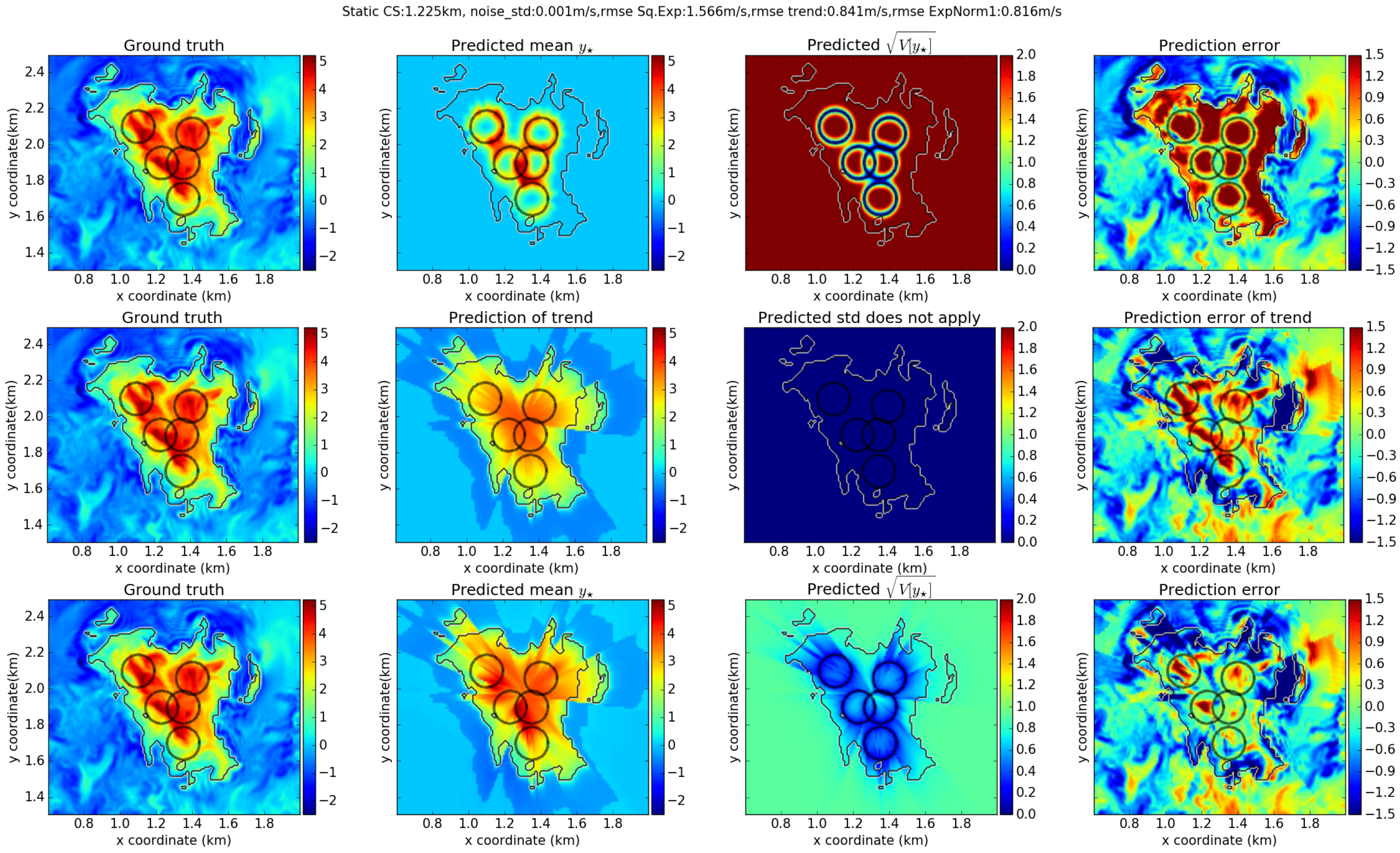


Figure A.29: Gaussian process mapping of a static cross-section of the cloud. The top row corresponds to the “off-the-shelf” squared exponential variant optimized with inputs in cartesian coordinates and “as is” sampled vertical wind contaminated with noise. The middle row depicts the behaviour of the radial trend. The bottom row shows the behaviour of the exponential covariance function with inputs in normalized polar coordinates using the hyperparameters obtained with the variogram approach and detrended vertical wind with additive white noise. Noise level $\sigma_{noise} = 0.001\,m/s$.

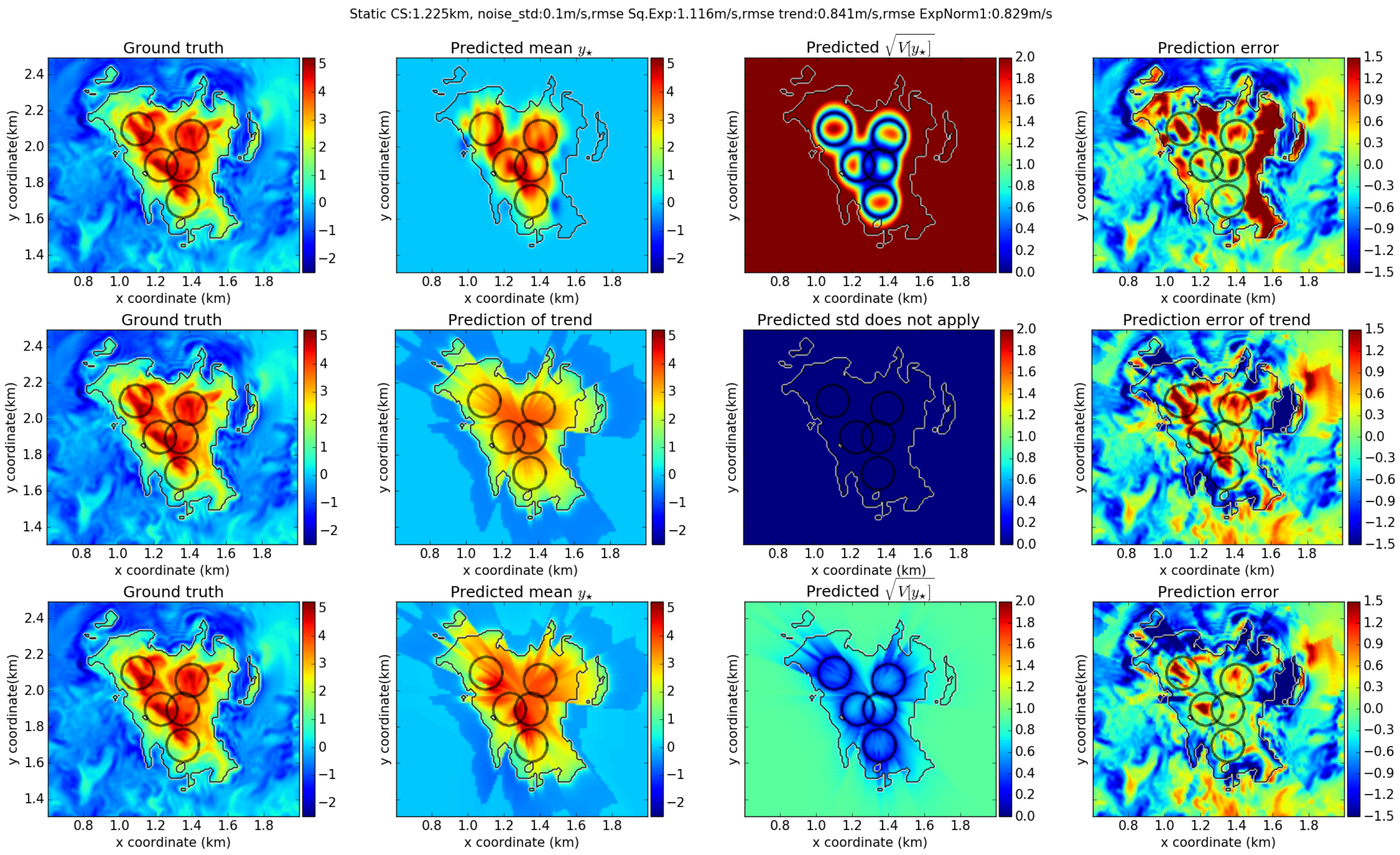


Figure A.30: Gaussian process mapping of a static cross-section of the cloud. The top row corresponds to the “off-the-shelf” squared exponential variant optimized with inputs in cartesian coordinates and “as is” sampled vertical wind contaminated with noise. The middle row depicts the behaviour of the radial trend. The bottom row shows the behaviour of the exponential covariance function with inputs in normalized polar coordinates using the hyperparameters obtained with the variogram approach and detrended vertical wind with additive white noise. Noise level $\sigma_{noise} = 0.1\,m/s$.

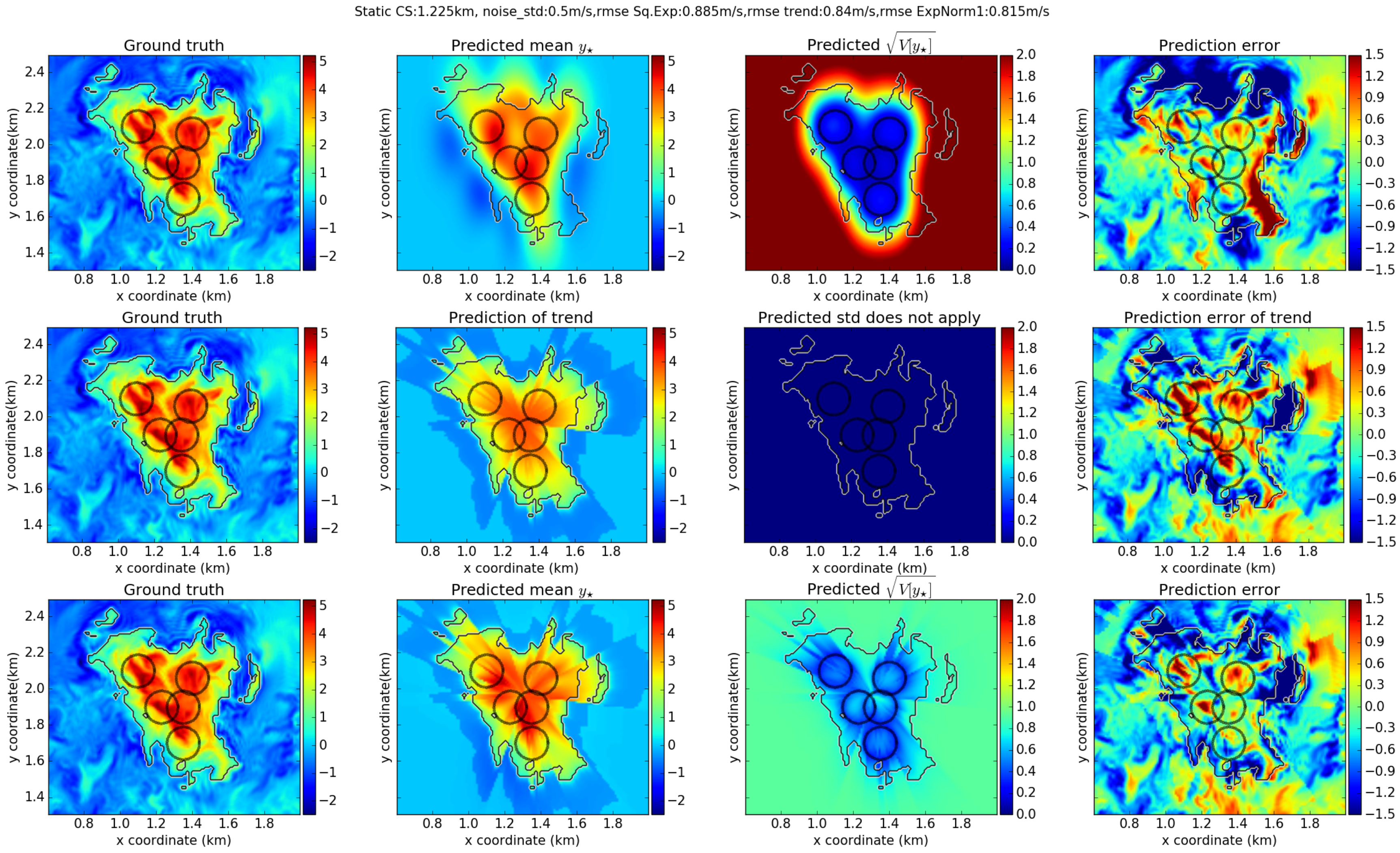


Figure A.31: Gaussian process mapping of a static cross-section of the cloud. The top row corresponds to the “off-the-shelf” squared exponential variant optimized with inputs in cartesian coordinates and “as is” sampled vertical wind contaminated with noise. The middle row depicts the behaviour of the radial trend. The bottom row shows the behaviour of the exponential covariance function with inputs in normalized polar coordinates using the hyperparameters obtained with the variogram approach and detrended vertical wind with additive white noise. Noise level $\sigma_{noise} = 0.5\,m/s$.

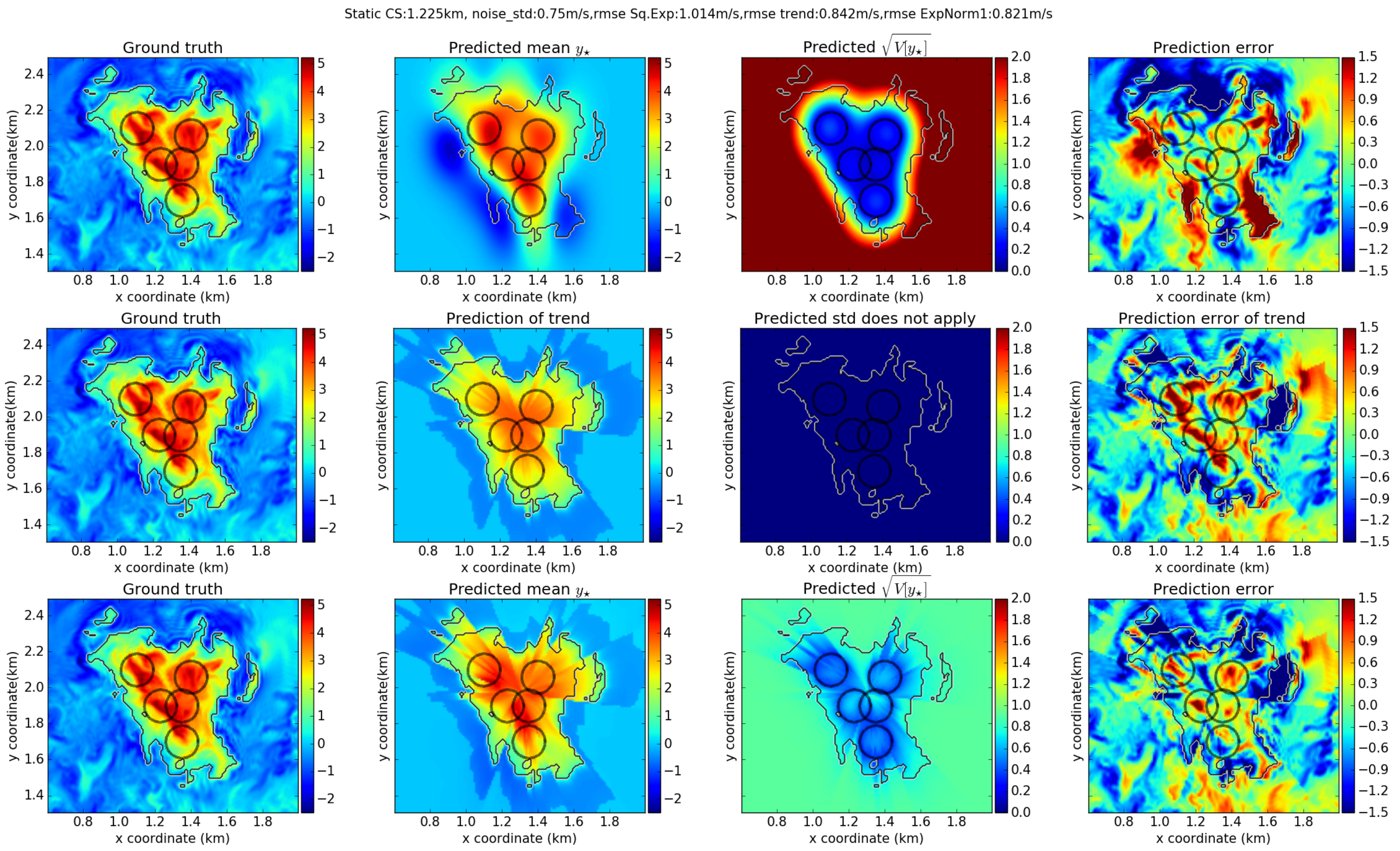


Figure A.32: Gaussian process mapping of a static cross-section of the cloud. The top row corresponds to the “off-the-shelf” squared exponential variant optimized with inputs in cartesian coordinates and “as is” sampled vertical wind contaminated with noise. The middle row depicts the behaviour of the radial trend. The bottom row shows the behaviour of the exponential covariance function with inputs in normalized polar coordinates using the hyperparameters obtained with the variogram approach and detrended vertical wind with additive white noise. Noise level $\sigma_{noise} = 0.001\, m/s$.

A.4.3 Hyperparameter Distribution, best overall Model

Exponential with cartesian inputs and detrended wind

The hyperparameter distribution was only determined for noise level $\sigma_{noise} = 0.25\,m/s$

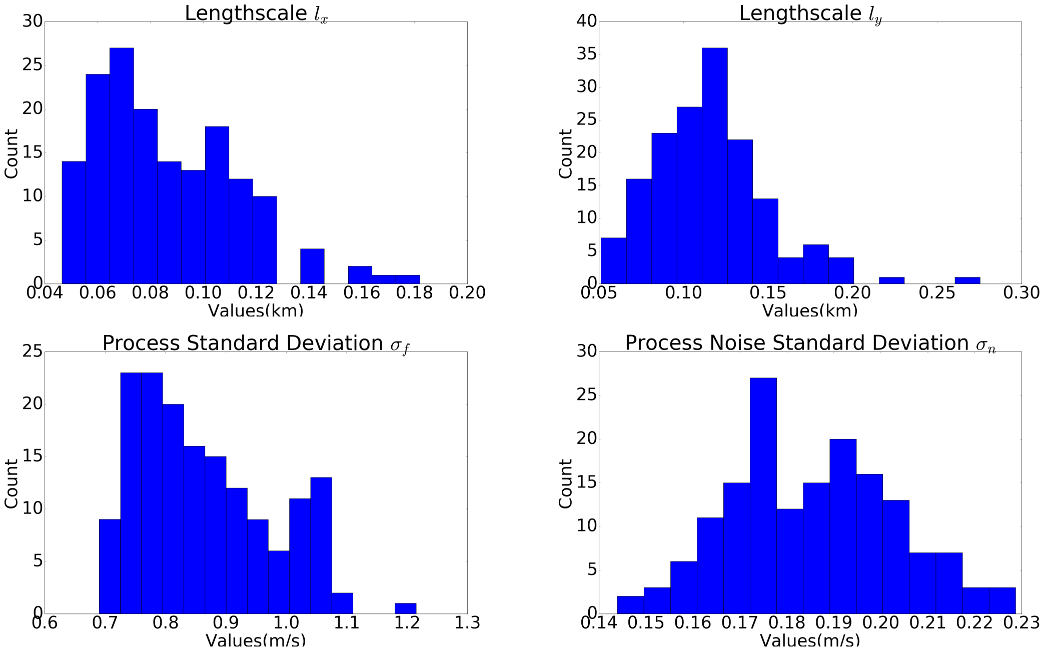


Figure A.33: Histograms of the hyperparameters of the static experiment on different cross-sections of cloud one.

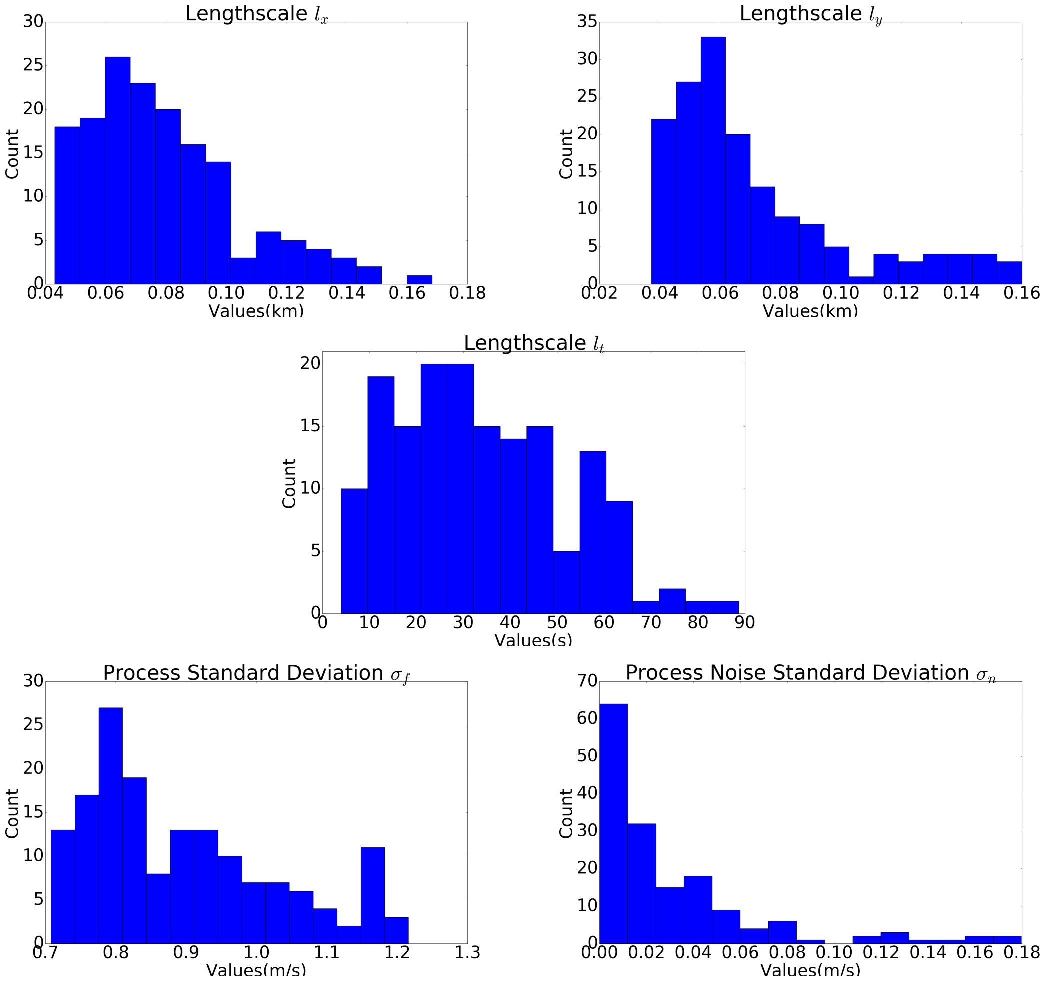


Figure A.34: Histograms of the hyperparameters of the dynamic experiment on different cross-sections of cloud one.

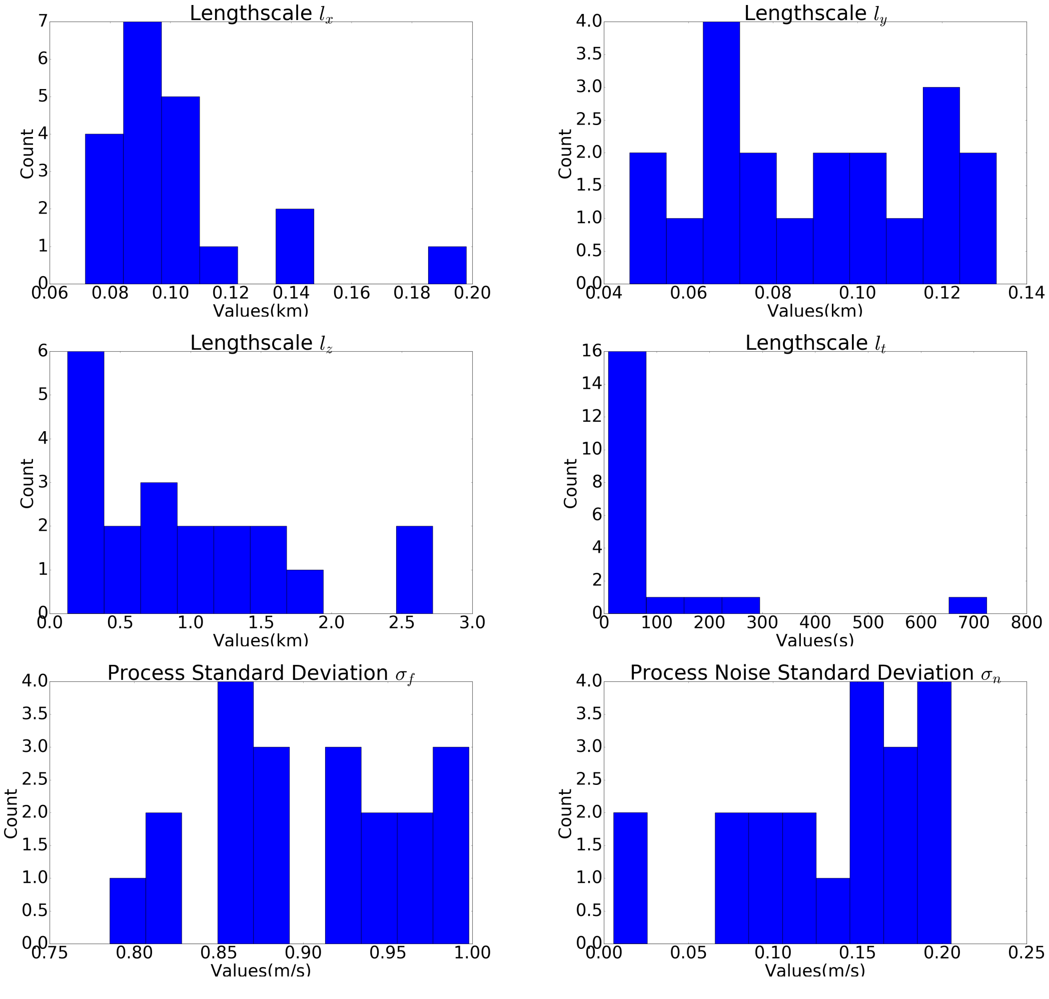


Figure A.35: Histograms of the hyperparameters of the dynamic experiment of the entire cloud one

Appendix B

Notation

B.1 Special operators and symbols

Kronecker product

The Kronecker product $\mathbf{A} \otimes \mathbf{B}$ between an $m \times p$ matrix $\mathbf{A}$ and an $n \times q$ matrix $\mathbf{B}$ generates an $mn \times qp$ block matrix of the following form:

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & \cdots & a_{1p}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & \cdots & a_{mp}\mathbf{B} \end{bmatrix}. \tag{B.1}$$

Inner product

A bilinear form $\langle \cdot , \cdot \rangle_{\mathbf{A}} : \mathbb{C}^n \times \mathbb{C}^n \mapsto \mathbb{C}$ named $\mathbf{A}$-inner product is defined as $\langle \mathbf{u}, \mathbf{v} \rangle_{\mathbf{A}} = \mathbf{u}^* \mathbf{A} \mathbf{v}$. In the special case of $\mathbf{A} = \mathbf{I}$, which is denoted by simply $\langle \cdot , \cdot \rangle$, the inner product is equivalent to the dot product of two vectors: $\langle \mathbf{u}, \mathbf{v} \rangle_{\mathbf{I}} = \langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^* \mathbf{v}$.

B.2 List of symbols

Typographical symbols

§ chapter, section, subsection

Mathematical symbols

In mathematic equations, non-boldened lowercase letter depict scalars, boldened lowercase letters depict vectors and boldened capital letters depict matrices. Furthermore:

j imaginary unit

$\mathbf{I}$ identity matrix

$\mathbf{0}$ zero matrix

$\mathbf{P}$ permutation matrix

$\mathbf{Q}$, $\mathbf{U}$ unitary (orthogonal) matrix

$\mathbf{T}$ transformation matrix

$\boldsymbol{R}(\varphi)$ rotation matrix

$\kappa(\mathbf{A})$ condition number of $\mathbf{A}$

$\lambda(\mathbf{A})$ eigenvalue of $\mathbf{A}$

$\sigma(\mathbf{A})$ singular value of $\mathbf{A}$

$\mathscr{H}_p$ Hardy p-norm

$\mathscr{L}_p$ Lebesgue p-norm

B.3 Abbreviations and acronyms

GCM General Circulation Model

GPR Gaussian Process Regression

LWC liquid water content

OLS Ordinary Least Squares

RMSE Root-Mean-Square-Error

SE squared exponential

SPDE stochastic partial differential equation

UAVs Unmanned Aerial Vehicles

WLS Weighted Least Squares

Appendix C

List of Figures

1.1 Schematic representation of a cumulus cloud. The arrows represent wind velocities, the orange blobs denote areas where mixing is occurring between the cloud and the surrounding atmosphere. This representation is very coarse: for instance the updrafts in the center of the cloud are known to behave as “bubbles” when the cloud is young. The cloud dimensions can vary from 100m to several hundreds of meters. 3

1.2 Simulation architecture. 4

2.1 Left panel: covariance functions depending on $r = |\mathbf{x} - \mathbf{x'}|$ for different values of ν. Right panel: Random functions obtained by sampling the Gaussian Processes with Matérn covariance functions for different values of ν and $l = 1$. Extracted from (Rasmussen and Williams, 2006) 12

2.2 Visualization of the semivariogram components as in (Curran, 1988). . . . 21

2.3 Examples of Semivariograms Models as in (Cressie, 1993) 23

a Spherical with Nugget Effect 23

b Exponential with Nugget Effect 23

2.4 Sill Anisotropy in two dimensional data: The solid line corresponds to the E-W direction and the N-S is given by dashed lines Taken from (Zimmerman, 1993) 24

2.5 Geometric Range Anisotropy in two dimensional data: The solid line corresponds to the E-W direction and the N-S, NW-SE and NE-SW are given by dashed lines with dashes of increasing length. Taken from (Zimmerman, 1993) 26

2.6 Non-Geometric Range Anisotropy in two dimensional data: The solid line corresponds to the E-W direction and the N-S, NW-SE and NE-SW are given by dashed lines with dashes of increasing length. Taken from (Zimmerman, 1993) 26

2.7 Nugget Anisotropy in two dimensional data: The solid line corresponds to the E-W direction and the N-S is given by dashed lines. Taken from (Zimmerman, 1993) 27

2.8 Empirical Semivariogram along with three **h**-*scatterplots*: It can be seen that below the sill the values at the end of the lag-vector are highly correlated with the values at the tail of the lag. Moreover, at the sill the values are uncorrelated and above the sill the values start to be negatively correlated. Taken from (Gringarten and Deutsch, 2001) 28

2.9 Empirical Semivariograms in the vertical and horizontal directions(Right) including the $Z(\mathbf{s})$ process, i.e. the scalar field on the left. The vertical semivariogram grows over the expected sill, which is caused by the trend in the vertical direction. Taken from (Gringarten and Deutsch, 2001) . . . 29

2.10 Empirical semivariograms before and after detrending. Here, the expected sill is one, since the data has been normalized w.r.t the mean and variance. Taken from (Gringarten and Deutsch, 2001) 29

2.11 Three simulations based on (2.67) with the same $\alpha = 1$,$\beta = 20$ and starting values $Y(s, 0) = 1$ for $15 \leq s \leq 24$, and δ is mean-zero white noise with **(a)** $\sigma_\delta = 0.01$, **(b)** $\sigma_\delta = 0.1$, **(c)** $\sigma_\delta = 1$. Taken from (Cressie and Wikle, 2011) 31

2.12 Three independent simulations drawn from a zero-mean Gaussian process with covariance function (2.68) and the same $\alpha = 1$, $\beta = 20$ as in Fig. 2.11. Taken from (Cressie and Wikle, 2011) 32

2.13 Comparison of a separable(left) and non-separable(right) covariance function. The non-separable case corresponds to (2.68). Taken from (Cressie and Wikle, 2011) 34

3.1 Meso-NH LES simulation: liquid cloud water content of the cumulus formed at 1h30 PM (ARM Southern Great Plains, June 21, 1997 conditions) 38

3.2 Determining the Empirical Variogram in a equally spaced data grid. This can be easily extended for each main direction of the data. 38

3.3 Exemplary circular trajectories to sample data of a cloud cross-section. . . 40

4.1 Surface area of cloud one per each cross-section, at time point $t = 449\,\mathrm{s}$. . 44

4.2 cloud one. Plots of the x and y coordinates of the center of masses of vertical wind and LWC. Also included the center of geometry in dependence of the height(Cloud Skeleton). These were all calculated at the same time point $t = 449\,\mathrm{s}$. 45

4.3 Two contour plots of the vertical wind of two cross-sections of Cloud1, 150m apart in height and at the same time point $t = 449\,\mathrm{s}$. The outer contour corresponds to the boundaries of the cloud. Furthermore, the center of geometry and the center of masses of the vertical wind and liquid water content are included. 47

4.4 Plots of the evolution of the mean and maximum of the vertical wind for each cross-section , at time point $t = 449\,\mathrm{s}$. 48

4.5 Empirical variograms in the four directions for cloud one. 51

4.6 Contour plot of the radially normalized wind field of a cross-section of cloud one. The outer highlighted contour pertains to the boundaries of the cloud. The normalized radius coordinate goes up to 150% of the radius at given direction *phi* 52

4.7 2D Histograms of the radial trend for cloud one. Top: normalized radius up to 150%. Bottom: Normalized radius up to 150% and all trend wind values normalized by the value measured at the center of mass of the cross-section 54

4.8 2D Histogram of the radial trend of all clouds put together. 55

4.9 Fitting the median radial trend with analytical functions 56

4.10 Empirical variograms of a cross-section of cloud one, analyzed over a span of 150s. Herein, the dimensions r, φ and t were calculated. 57

4.11 Residuals of the Trend inside the cross-section used to calculate the variogram in Fig. 4.10 58

4.12 Empirical variograms of a cross-section of cloud three, analyzed over a span of 150s. Herein, the dimensions r, φ and t were calculated. 58

4.13 Residuals of the Trend inside the cross-section used to calculate the variogram in Fig. 4.12 59

4.14 Empirical variograms of the entire cloud five, analyzed over a span of 5s. Herein, the dimensions r, φ and z were calculated. 60

4.15 Residuals of the Trend inside the cloud used to calculate the variogram in Fig. 4.14 60

4.16 Histograms of the lengthscales obtained by the fitting procedure in in all directions of the polar representation. 62

4.17 Histograms of the process standard deviations obtained by the fitting procedure in all directions of the polar representation. 63

4.18 Modeled Variograms using median hyperparameters for each direction in the polar normalized representation. 64

4.19 Summary of RMSE and predicted standard deviation for noise $\sigma_{noise} = 0.25m/s$, all mapping variants for the three scenarios. 67

4.20 Mapping of the cross-section by implementing the trend scaled by the wind measured at the center. Noise level $\sigma_{noise} = 0.25\,m/s$. 68

4.21 Gaussian Process mapping of a static cross-section of the cloud. The top row corresponds to the “off-the-shelf” squared exponential variant optimized with inputs in cartesian coordinates and “as is” sampled vertical wind contaminated with noise. The middle row depicts the behaviour of the radial trend. The bottom row shows the behaviour of the exponential covariance function with inputs in normalized polar coordinates using the hyperparameters obtained with the variogram approach and detrended vertical wind with additive white noise. Noise level $\sigma_{noise} = 0.25\,m/s$. . . 69

4.22 Gaussian Process mapping of a static cross-section of the cloud. Herein, an exponential covariance function was used with inputs in cartesian coordinates of the five circle trajectories. The output for training was the detrended vertical wind under the effect of white noise. Noise level $\sigma_{noise} = 0.25\, m/s$. 70

A.1 Surface area of cloud two per each cross-section, at time point $t = 929\,\mathrm{s}$. . 79

A.2 Plots of the x and y coordinates of the center of masses of vertical wind and LWC. Also included the center of geometry in dependence of the height(Cloud Skeleton). These were all calculated at the same time point $t = 929\,\mathrm{s}$. 80

A.3 Two contour plots of the vertical wind of two cross-sections of Cloud two, 150m apart in height and at the same time point $t = 929\,\mathrm{s}$. The outer contour corresponds to the boundaries of the cloud. Furthermore, the center of geometry and the center of masses of the vertical wind and liquid water content are included. 81

A.4 Plots of the evolution of the mean and maximum of the vertical wind for each cross-section , at time point $t = 929\,\mathrm{s}$. 82

A.5 Surface area of cloud three per each cross-section, at time point $t = 1529\,\mathrm{s}$. 83

A.6 Cloud three. Plots of the x and y coordinates of the center of masses of vertical wind and LWC. Also included the center of geometry in dependence of the height(Cloud Skeleton). These correspond to the same time point $t = 1529\,\mathrm{s}$. 84

A.7 Two contour plots of the vertical wind of two cross-sections of Cloud three, 150m apart in height and at the same time point $t = 1529\,\mathrm{s}$. The outer contour corresponds to the boundaries of the cloud. Furthermore, the center of geometry and the center of masses of the vertical wind and liquid water content are included. 85

A.8 Cloud three. Plots of the evolution of the mean and maximum of the vertical wind for each cross-section , at time point $t = 1529\,\mathrm{s}$. 86

A.9 Surface area of cloud four per each cross-section, at time point $t = 2789\,\mathrm{s}$. 87

A.10 Cloud four. Plots of the x and y coordinates of the center of masses of vertical wind and LWC. Also included the center of geometry in dependence of the height(Cloud Skeleton). These correspond to the same time point $t = 2789\,\mathrm{s}$. 88

A.11 Two contour plots of the vertical wind of two cross-sections of Cloud four, 150m apart in height and at the same time point $t = 2789\,\mathrm{s}$. The outer contour corresponds to the boundaries of the cloud. Furthermore, the center of geometry and the center of masses of the vertical wind and liquid water content are included. 89

A.12 Cloud Four. Plots of the evolution of the mean and maximum of the vertical wind for each cross-section , at time point $t = 2789\,\mathrm{s}$. 90

A.13 Surface area of cloud five per each cross-section, at time point $t = 3389\,\mathrm{s}$. 91

A.14 Cloud five. Plots of the x and y coordinates of the center of masses of vertical wind and LWC. Also included the center of geometry in dependence of the height(Cloud Skeleton). These correspond to the same time point $t = 3389\,\mathrm{s}$. 92

A.15 Two contour plots of the vertical wind of two cross-sections of Cloud five, 150m apart in height and at the same time point $t = 3389\,\mathrm{s}$. The outer contour corresponds to the boundaries of the cloud. Furthermore, the center of geometry and the center of masses of the vertical wind and liquid water content are included. 93

A.16 Cloud Five. Plots of the evolution of the mean and maximum of the vertical wind for each cross-section , at time point $t = 3389\,\mathrm{s}$. 94

A.17 Empirical variograms in the four directions for cloud two. 95

A.18 Empirical variograms in the four directions for cloud three. 96

A.19 Empirical variograms in the four directions for cloud four. 96

A.20 Empirical variograms in the four directions for cloud five. 97

A.21 2D Histograms of the radial trend for cloud two. Top: normalized radius up to 150%. Bottom: Normalized radius up to 150% and all trend wind values normalized by the value measured at the center of mass of the cross-section 98

A.22 2D Histograms of the radial trend for cloud three. Top: normalized radius up to 150%. Bottom: Normalized radius up to 150% and all trend wind values normalized by the value measured at the center of mass of the cross-section 99

A.23 2D Histograms of the radial trend for cloud four. Top: normalized radius up to 150%. Bottom: Normalized radius up to 150% and all trend wind values normalized by the value measured at the center of mass of the cross-section 100

A.24 2D Histograms of the radial trend for cloud five. Top: normalized radius up to 150%. Bottom: Normalized radius up to 150% and all trend wind values normalized by the value measured at the center of mass of the cross-section 101

A.25 Summary of RMSE and predicted standard deviation for actual noise level $\sigma_{noise} = 0.001\,m/s$, all mapping variants for the three scenarios. 102

A.26 Summary of RMSE and predicted standard deviation for actual noise level $\sigma_{noise} = 0.1\,m/s$, all mapping variants for the three scenarios. 103

A.27 Summary of RMSE and predicted standard deviation for actual noise level $\sigma_{noise} = 0.5\,m/s$, all mapping variants for the three scenarios. 103

A.28 Summary of RMSE and predicted standard deviation for actual noise level $\sigma_{noise} = 0.75\,m/s$, all mapping variants for the three scenarios. 104

A.29 Gaussian process mapping of a static cross-section of the cloud. The top row corresponds to the “off-the-shelf” squared exponential variant optimized with inputs in cartesian coordinates and “as is” sampled vertical wind contaminated with noise. The middle row depicts the behaviour of the radial trend. The bottom row shows the behaviour of the exponential covariance function with inputs in normalized polar coordinates using the hyperparameters obtained with the variogram approach and detrended vertical wind with additive white noise. Noise level $\sigma_{noise} = 0.001\, m/s$. . . 105

A.30 Gaussian process mapping of a static cross-section of the cloud. The top row corresponds to the “off-the-shelf” squared exponential variant optimized with inputs in cartesian coordinates and “as is” sampled vertical wind contaminated with noise. The middle row depicts the behaviour of the radial trend. The bottom row shows the behaviour of the exponential covariance function with inputs in normalized polar coordinates using the hyperparameters obtained with the variogram approach and detrended vertical wind with additive white noise. Noise level $\sigma_{noise} = 0.1\, m/s$. . . . 106

A.31 Gaussian process mapping of a static cross-section of the cloud. The top row corresponds to the “off-the-shelf” squared exponential variant optimized with inputs in cartesian coordinates and “as is” sampled vertical wind contaminated with noise. The middle row depicts the behaviour of the radial trend. The bottom row shows the behaviour of the exponential covariance function with inputs in normalized polar coordinates using the hyperparameters obtained with the variogram approach and detrended vertical wind with additive white noise. Noise level $\sigma_{noise} = 0.5\, m/s$. . . . 107

A.32 Gaussian process mapping of a static cross-section of the cloud. The top row corresponds to the “off-the-shelf” squared exponential variant optimized with inputs in cartesian coordinates and “as is” sampled vertical wind contaminated with noise. The middle row depicts the behaviour of the radial trend. The bottom row shows the behaviour of the exponential covariance function with inputs in normalized polar coordinates using the hyperparameters obtained with the variogram approach and detrended vertical wind with additive white noise. Noise level $\sigma_{noise} = 0.001\, m/s$. . . 108

A.33 Histograms of the hyperparameters of the static experiment on different cross-sections of cloud one. 109

A.34 Histograms of the hyperparameters of the dynamic experiment on different cross-sections of cloud one. 110

A.35 Histograms of the hyperparameters of the dynamic experiment of the entire cloud one 111

Appendix D

List of Tables

4.1 Weighted Sum of Squared Errors of the fitting of empirical variograms. Minimum values for each row are boldened 62

4.2 Median Hyperparameters of the fitted Exponential covariance function in all directions 63

4.3 Median Hyperparameters for all scenarios of the exponential covariance function with inputs in cartesian coordinates of the five circle/helicoidal trajectories. The output for training was the detrended vertical wind under the effect of white noise. Noise level $\sigma_{noise} = 0.25\, m/s$. 71

4.4 Comparison of median hyperparameters for the dynamic cross-section scenario between models exponential and square exponential with no priors, i.e. model one and two of the list at the introduction of this chapter. Noise level $\sigma_{noise} = 0.25\, m/s$. 72

Appendix E

References

Christopher M Bishop. *Pattern Recognition and Machine Learning.* Springer, 2006. ISBN 0-387-31073-8.

A. R. Brown, R. T. Cederwall, A. Chlond, P. G. Duynkerke, J.-C. Golaz, M. Khairoutdinov, D. C. Lewellen, A. P. Lock, M. K. MacVean, C.-H. Moeng, R. A. J. Neggers, A. P. Siebesma, and B. Stevens. Large-eddy simulation of the diurnal cycle of shallow cumulus convection over land. *Quarterly Journal of the Royal Meteorological Society*, 128(582):1075–1093, 2002.

Kian Ming Chai. *Multi-task learning with gaussian processes.* PhD thesis, The University of Edinburgh, 2010.

Jen Jen Chung, Nicholas RJ Lawrance, and Salah Sukkarieh. Learning to soar: Resource-constrained exploration in reinforcement learning. *The International Journal of Robotics Research*, 34(2):158–172, 2015.

C. E. Corrigan, G. C. Roberts, M. V. Ramana, D. Kim, and V. Ramanathan. Capturing vertical profiles of aerosols and black carbon over the indian ocean using autonomous unmanned aerial vehicles. *Atmospheric Chemistry and Physics*, 8(3):737–747, 2008.

Noel Cressie. Fitting variogram models by weighted least squares. *Journal of the International Association for Mathematical Geology*, 17(5):563–586, 1985.

Noel Cressie. *Statistics for spatial data.* John Wiley & Sons, 1993. ISBN 978-1-119-11461-1.

Noel Cressie and Christopher K Wikle. *Statistics for spatio-temporal data.* John Wiley & Sons, 2011. ISBN 978-0-471-69274-4.

Paul J Curran. The semivariogram in remote sensing: an introduction. *Remote Sensing of Environment*, 24(3):493–507, 1988.

J. Das, J. Harvey, F. Py, H. Vathsangam, R. Graham, K. Rajan, and G.S. Sukhatme. Hierarchical probabilistic regression for AUV-based adaptive sampling of marine phenomena. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 5571–5578, 2013.

Anthony D. Del Genio and Jingbo. Wu. The role of entrainment in the diurnal cycle of continental convection. *Journal of Climate*, 23(10):2722–2738, 2010.

J.A. Diaz, D. Pieri, C. R. Arkin, E. Gore, T.P. Griffin, M. Fladeland, G. Bland, C. Soto, Y. Madrigal, D. Castillo, E. Rojas, and S. Achí. Utilization of in situ airborne ms-based instrumentation for the study of gaseous emissions at active volcanoes. *International Journal of Mass Spectrometry*, 295(3):105–112, 2010.

J. Elston and B. Argrow. Energy efficient UAS flight planning for characterizing features of supercell thunderstorms. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 6555–6560, 2014.

Jack S Elston, Jason Roadman, Maciej Stachura, Brian Argrow, Adam Houston, and Eric Frew. The tempest unmanned aircraft system for in situ observations of tornadic supercells: design and VORTEX2 flight results. *Journal of Field Robotics*, 28(4): 461–483, 2011.

Emmanuel Gringarten and Clayton V Deutsch. Teacher's aide variogram interpretation and modeling. *Mathematical Geology*, 33(4):507–534, 2001.

V Heine. Models for two-dimensional stationary stochastic processes. *Biometrika*, 42 (1-2):170–178, 1955.

G. J. Holland, P. J. Webster, J. A. Curry, G. Tyrell, D. Gauntlett, G. Brett, J. Becker, R. Hoag, and W. Vaglienti. The aerosonde robotic aircraft: A new paradigm for environmental observations. *Bulletin of the American Meteorological Society*, 82(5): 889–901, 2001.

J. Inoue, J. A. Curry, and J. A. Maslanik. Application of aerosondes to melt-pond observations over arctic sea ice. *Journal of Atmospheric and Oceanic Technology*, 25 (2):327–334, 2008.

J. P. Lafore, J. Stein, N. Asencio, P. Bougeault, V. Ducrocq, J. Duron, C. Fischer, P. Héreil, P. Mascart, V. Masson, J. P. Pinty, J. L. Redelsperger, E. Richard, and J. Vilà-Guerau de Arellano. The Meso-NH Atmospheric Simulation System. Part I: adiabatic formulation and control simulations. *Annales Geophysicae*, 16(1):90–109, 1998.

Nicholas RJ Lawrance and Salah Sukkarieh. Autonomous exploration of a wind field with a gliding aircraft. *Journal of Guidance, Control, and Dynamics*, 34(3):719–733, 2011.

Georges Matheron. *The Theory of Regionalized Variables and its Applications*, volume 5. École national supérieure des mines, 1971.

Matthew Michini, M Ani Hsieh, Eric Forgoston, and Ira B Schwartz. Robotic tracking of coherent structures in flows. *Robotics, IEEE Transactions on*, 30(3):593–603, 2014.

José-María Montero, Gema Fernández-Avilés, and Jorge Mateu. *Spatial and Spatio-Temporal Geostatistical Modeling and Kriging.* John Wiley & Sons, 2015. ISBN 978-1-118-41318-0.

J. Nguyen, N. Lawrance, R. Fitch, and S. Sukkarieh. Energy-constrained motion planning for information gathering with autonomous aerial soaring. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 3825–3831, 2013.

Veerabhadran Ramanathan, Muvva V Ramana, Gregory Roberts, Dohyeong Kim, Craig Corrigan, Chul Chung, and David Winker. Warming trends in Asia amplified by brown cloud solar absorption. *Nature*, 448(7153):575–578, 2007.

Carl Edward Rasmussen and Christopher KI Williams. *Gaussian processes for machine learning.* the MIT Press, 2006. ISBN 0-262-18253-X.

S. Ravela, T Vigil, and I Sleder. Tracking and Mapping Coherent Structures. In *International Conference on Computational Science (ICCS)*, 2013.

C. G. Roberts, M.V. Ramana, C. Corrigan, D. Kim, and V. Ramanathan. Simultaneous observations of aerosol–cloud–albedo interactions with three stacked unmanned aerial vehicles. *Proceedings of the National Academy of Sciences of the United States of America*, 105(21):7370–7375, 2008.

Bjorn Stevens and Sandrine Bony. What are climate models missing? *Science*, 340 (6136):1053–1054, 2013.

Dale L Zimmerman. Another look at anisotropy in geostatistics. *Mathematical Geology*, 25(4):453–470, 1993.