

Merethe Tørresen

# Multivariate Analysis of Ocean Currents in the Barents Sea

Multivariabel Analyse av Havstrømmer i Barentshavet

Master's thesis in Marine Technology

Supervisor: Roger Skjetne & Francesco Scibilia

June 2019



Merethe Tørresen

# Multivariate Analysis of Ocean Currents in the Barents Sea

Multivariabel Analyse av Havstrømmer i  
Barentshavet

Master's thesis in Marine Technology  
Supervisor: Roger Skjetne & Francesco Scibilia  
June 2019

Norwegian University of Science and Technology  
Faculty of Engineering  
Department of Marine Technology



Norwegian University of  
Science and Technology





NTNU Trondheim  
Norwegian University of Science and Technology  
Department of Marine Technology

## MSC THESIS DESCRIPTION SHEET

**Name of the candidate:** Tørresen, Merethe  
**Field of study:** Marine control engineering  
**Thesis title (Norwegian):** Multivariabel Analyse av Havstrømmer i Barentshavet  
**Thesis title (English):** Multivariate Analysis of Ocean Currents in the Barents Sea

### Background

Ocean circulation is a seasonal, primarily horizontal and layered, water movement that derives its energy either by wind stress on the sea surface, inducing a momentum exchange, or by variations in water density – imposed at the sea surface by exchange of ocean heat and water with the atmosphere. Ocean currents are also driven and influenced by different effects like pressure gradients, Coriolis effects, and tides. Seabed topography, shoreline configuration, and interaction with other currents also influence the current velocity. The wind-driven circulation is strongest in the surface layer, which has a depth of about 100 to 150 m. The coupling between wind and surface ocean currents is described in idealized conditions by the Ekman spiral. However, the real ocean may vary significantly from the idealized conditions.

Knowledge about local current regimes is important for offshore facility design and operation. Particularly, information of the currents in the upper 200 m of the water column can be of significant help in, for instance, the prediction of icebergs drifts in areas as offshore Newfoundland (Canada) and Barents Sea.

The hypothesis to be investigated in this thesis is that, given a specific location, there is a correlation between wind and surface top currents in relation to deeper currents. Furthermore, based on historic data, can the current can be predicted by using multivariate modelling techniques?

### Work description

1. Perform a background and literature review to provide information and relevant references on:
  - Relevant oceanographic processes in the ocean.
  - Principal Component Analysis (PCA).
  - Partial Least Squares Regression (PLSR)
  - Relevant nonlinear regression methods and/or machine learning methods.Write a list with abbreviations and definitions of terms, relevant to the literature study and assignment (as part of the front matter).
2. Review important oceanographic processes in the ocean to get understanding of the main underlying phenomenon of ocean processes.
3. Explore dominating effects in wind and current by performing Principle Component Analysis along depth- and horizontal spatial dimensions of the dataset (continuing work from MSc project). Discuss correlation between wind and current, and between current and current.
4. Make Partial Least Squares Regression for estimating deeper currents based on wind and top currents. Discuss findings and consider if other model identification methods are necessary to improve the analysis and compare the results.
5. Make Partial Least Squares Regression for forecasting current based on historic data. Discuss findings and consider if other model identification methods are necessary to improve the analysis and compare the results.



**Tentatively:**

6. Apply different model identification method to improve the analysis, presumably a nonlinear regression method.
7. Compare linear and nonlinear regression methods, as used on the original dataset, and discuss pros and cons.

**Specifications**

The scope of work may prove to be larger than initially anticipated. By the approval from the supervisor, described topics may be deleted or reduced in extent without consequences with regard to grading.

The candidate shall present personal contribution to the resolution of problems within the scope of work. Theories and conclusions should be based on mathematical derivations and logic reasoning identifying the various steps in the deduction.

The report shall be organized in a logical structure to give a clear exposition of background, results, assessments, and conclusions. The text should be brief and to the point, with a clear language. Rigorous mathematical deductions and illustrating figures are preferred over lengthy textual descriptions. The report shall have font size 11 pts., and it is not expected to be longer than 70 A4-pages, 100 B5-pages, from introduction to conclusion, unless otherwise agreed upon. It shall be written in English (preferably US) and contain the following elements: Title page, abstract, acknowledgements, thesis specification, list of symbols and acronyms, table of contents, introduction with objective, background, and scope and delimitations, main body with problem formulations, derivations/developments and results, conclusions with recommendations for further work, references, and optional appendices. All figures, tables, and equations shall be numerated. The original contribution of the candidate and material taken from other sources shall be clearly identified. Work from other sources shall be properly acknowledged using quotations and a Harvard citation style (e.g. *natbib* Latex package). The work is expected to be conducted in an honest and ethical manner, without any sort of plagiarism and misconduct. Such practice is taken very seriously by the university and will have consequences. NTNU can use the results freely in research and teaching by proper referencing, unless otherwise agreed upon.

The thesis shall be submitted with an electronic copy to the main supervisor and department according to NTNU administrative procedures. The final revised version of this thesis description shall be included after the title page. Computer code, pictures, videos, dataseries, etc., shall be included electronically with the report.

**Start date:** January, 2019                      **Due date:** As specified by the administration.

**Supervisor:** Roger Skjetne  
**Co-advisor(s):** Francesco Scibilia (Equinor), Lars Erik Holmedal, and Leif Erik Andersson (ITK)

**Trondheim, 27.05.2019**

Digitally signed by Roger Skjetne  
Date: 2019.05.27 15:51:52 +02'00'

---

**Roger Skjetne**  
Supervisor

---

# Summary

Analyses for parameter reduction and prediction are performed on a hindcast dataset from a specific location in the Barents Sea. Three methods, PCA, PLSR and GPR are applied to different scenarios, and a selection of data amounts. Principal Component Analysis (PCA) are applied to data from two different spatial dimensions in order to explore parameter reduction and relations in the wind- and current data. It shows that current data through the water depth are highly correlated and well suited for parameter reduction, while data in the horizontal direction are not. Partial Least Squares Regression (PLSR) is an analysis tool that is very similar to PCA, but the algorithm is trained with more consideration towards specified response variables. The method is performed on datasets including different amounts of data with measurements every hour for 2, 5 and 10 years. Two techniques are used, estimation and forecasting. Estimation of currents through the water column are done based on the predictor variables wind and currents at 0m, 10m and 20m. The resulting estimate is quite good in upper current layers, while its quality decrease fast towards the bottom. This is clearly seen in timeseries of reconstructed data and Mean Squared Error (MSE). At a depth of around 100m to 150m, the estimation is quite poor. However, as icebergs are floating in the surface, the estimation might be good enough to help with iceberg drift predictions. PLSR is also used for forecasting, which is based on historic data through the entire water column. The forecast is made for one hour into the future, and as currents do not change very much in an hour, the results from these analyses are very good. They show that the number of years of training data included has a smaller effect than expected on the MSE and the resulting timeseries. A third method, Gaussian Process Regression (GPR) is applied to the data. This method is quite different from PCA and PLSR, as it is non-linear and it lets the data speak more for itself. GPR is used for estimation, and gives a much better result than PLSR. However, bottom currents still have a moderate error and the computation time is significant, even for small datasets. Combining methods by performing PCA and then GPR to reduce computation time is proposed.

---

# Sammendrag

Analysen for parameterreduksjon og prediksjon er utført på et hindcast datasett fra et bestemt område i Barentshavet. Tre analysemetoder, PCA, PLSR og GPR utføres på et utvalg scenarier og datamengder. Principal Component Analysis (PCA) brukes på data fra to forskjellige romlige dimensjoner for å utforske parameterreduksjon og relasjoner mellom vind- og strømdata. De viser at havstrøm gjennom vannkolonnen er sterkt korrelert og velegnet for parameterreduksjon. Partial Least Squares Regression (PLSR) er et analyseverktøy som ligner PCA, men algoritmen er trent med hensyn til en spesifisert respons. Metoden utføres på datasett med ulike mengder data, hvor målinger er tatt hver time i 2, 5 og 10 år. To PLSR-teknikker brukes, estimering og prognoser. Estimering av strøm gjennom vannkolonnen utføres basert på prediktorvariabler som inneholder vind og havstrømmålinger på 0m, 10m og 20m. Det resulterende estimatet er ganske bra i de øverste lagene, men kvaliteten reduseres raskt mot bunnen. Dette kan observeres tydelig i tidsserier av rekonstruerte data og Mean Squared Error (MSE). Fra en dybde på rundt 100-150m og til bunnen, er estimatet ganske dårlig. Men det kan kanskje være godt nok til å brukes til prediksjon av drift av isberg, siden de flyter i overflaten og vanligvis har kjøle på max 200m. PLSR brukes også til prognoser. Disse er basert på historiske data gjennom hele vannsøylen. Prognosen er utført for en time fram i tid, og siden strømmen ikke endres veldig mye i løpet av en time, er resultatene fra disse analysene veldig gode. De viser at antall år med treningsdata inkludert, har en mindre enn forventet effekt på MSE og de resulterende tidsseriene. En tredje metode, Gaussian Process Regression (GPR) blir også utført på dataene. Denne metoden er ganske forskjellig fra PCA og PLSR, da den er ulineær og den lar dataene i mye høyere grad snakke for seg selv. GPR er brukt til estimering, og gir et mye bedre resultat enn ved PLSR. Havstrømmer mot bunnen får imidlertid fortsatt en moderat feil, og beregningstiden er betydelig, selv for relativt små datasett. Å kombinere metoder ved å utføre PCA og deretter GPR for å redusere beregningstiden foreslås.



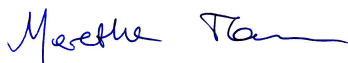
---

# Preface

The work presented in this thesis completes the course TMR4930 Marine Technology, Masters's Thesis and a two year Master programme in Marine Technology. The thesis is written during spring 2019 at the Department of Marine Technology in cooperation with the Department of Engineering Cybernetics at the Norwegian University of Science and Technology, and Equinor ASA. Preliminary work were carried out in a project thesis in the course TMR4510 Marine Control Systems, Specialization Project.

I would like to express my gratitude towards my supervisors Professor Roger Skjetne and Professor Francesco Scibilia for their guidance and advice through the year. Thank you for inspiring me with possibilities, while trusting me with the freedom to let the work follow my own interests and ideas. I would also like to thank co-supervisor Leif Erik Andersson for always taking the time to help me and his contributions to my project through guidance hours and good advise. I particularly appreciate help and patience with my understanding of methodologies. I would like to thank Professor Lars Erik Holmedal for his tutoring lectures in Oceanography and his general interest in the project. Finally, I would like to thank my friends and my wonderful family for their support and encouragement. A special thank you goes to my office mates at A1.007. We have had a lot of coffee and even more fun.

June 11<sup>th</sup> 2019  
Trondheim



Merethe Tørresen, candidate number 10109

---

# Table of Contents

<b>Summary</b>	<b>i</b>
<b>Sammendrag</b>	<b>ii</b>
<b>Preface</b>	<b>iii</b>
<b>Table of Contents</b>	<b>vii</b>
<b>List of Figures</b>	<b>xii</b>
<b>List of Tables</b>	<b>xiii</b>
<b>Abbreviations</b>	<b>xiv</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Problem definition . . . . .	1
1.2 Literature review . . . . .	2
1.3 Hindcast data . . . . .	3
1.4 Ocean currents . . . . .	4
1.5 Analysis tools . . . . .	5
<b>2 Driving mechanisms in the ocean</b>	<b>7</b>
2.1 The oceanic heat budget . . . . .	7
2.2 Tides . . . . .	9
2.2.1 Solar effects on the tides . . . . .	13
2.3 Coriolis Force . . . . .	13
2.4 Boundary layer . . . . .	14
2.4.1 Boundary layer equations . . . . .	15

---

2.4.2	Governing equations . . . . .	16
2.4.3	Rosby approximation . . . . .	17
2.4.4	Geostrophic flow . . . . .	18
2.5	Wind . . . . .	18
2.5.1	The Ekman layer . . . . .	19
<b>3</b>	<b>Pre-processing</b>	<b>21</b>
3.1	Exploring relations in the data . . . . .	21
3.2	Bathymetry . . . . .	22
3.3	Statistical properties . . . . .	23
3.3.1	Stationarity . . . . .	24
3.3.2	Normalization, covariance and correlation . . . . .	24
3.3.3	Degree of Compression (DoC) and Mean Squared Error (MSE) . . . . .	25
3.4	Choosing data to analyze . . . . .	26
3.5	Organizing data . . . . .	27
<b>4</b>	<b>Multivariate analysis</b>	<b>29</b>
4.1	Principal Component Analysis methodology . . . . .	29
4.1.1	Finding the Principal Components . . . . .	30
4.1.2	Singular Value Decomposition . . . . .	31
4.2	Partial Least Squares Regression methodology . . . . .	32
4.2.1	SIMPLS algorithm . . . . .	32
4.2.2	Prediction and response matrices . . . . .	33
4.3	Gaussian Process Regression methodology . . . . .	34
<b>5</b>	<b>Principal Component Analysis</b>	<b>37</b>
5.1	First scenario: Twelve variables through the water column . . . . .	37
5.1.1	Timeseries and normalizing . . . . .	38
5.1.2	Variance . . . . .	38
5.1.3	Reconstruction . . . . .	41
5.2	Second scenario: Horizontal spatial dimension . . . . .	44
5.2.1	Variance . . . . .	44
5.2.2	Correlation and error . . . . .	46
<b>6</b>	<b>Partial Least Squares Regression</b>	<b>49</b>
6.1	Estimating deeper currents based on surface currents . . . . .	49
6.1.1	Six weeks of training data . . . . .	49
6.1.2	2, 5 and 10 years of training data . . . . .	51
6.2	Forecasting current based on historic water column measurements	60

---

---

<b>7</b>	<b>Gaussian Process Regression</b>	<b>73</b>
<b>8</b>	<b>Discussion</b>	<b>75</b>
8.1	PCA . . . . .	75
8.2	PLSR . . . . .	78
8.3	GPR . . . . .	81
<b>9</b>	<b>Conclusion and further work</b>	<b>83</b>
9.1	Further work . . . . .	84
	<b>Bibliography</b>	<b>87</b>
	<b>Appendix</b>	<b>91</b>

---

# List of Figures

2.1	Net heat flux in the x-direction of a cubic element. Illustration: Myrhaug (2012). . . . .	8
2.2	The principle of the Earth, Moon and Sun system. The center of mass T is in reality positioned on Earth. Illustration: Myrhaug (2012)	10
2.3	Un-scaled derivation of the centrifugal, gravitational and resulting tide-producing forces. Illustration: Brown et al. (1999) . . . . .	12
2.4	Illustration of the effect of the fictive Coriolis force when a particle moves on the surface of the Earth. Illustration: Morales (2018) . .	14
2.5	Illustration of the typical velocity profile in a boundary layer for fully developed flow. $L$ must be sufficiently long to find a change in $u(z)$ . Illustration from Myrhaug (2012) . . . . .	15
2.6	The Ekman spiral describing how the horizontal wind sets the surface waters in motion. Illustration from Ocean Motion, NASA (2018) . . . . .	20
3.1	$C_s$ related to wind at different depths and $C_s$ through the water column at different times. It should be noted that the current are expressed in cm/s and the wind in m/s. . . . .	22
3.2	Seabed topography of the area where the hindcast data originates from. . . . .	23
3.3	Name and configuration of the points. . . . .	26
4.1	The ocean dimensions, where a is North, b is East and c is depth. .	33
5.1	The timeseries from year 2004 of wind and through the watercolumn from surface to 100m . . . . .	38
5.2	Biplot of $\mathbf{X}$ . . . . .	40

---

5.3	Percentage of variance explained in each PC and degree of compression of the data . . . . .	40
5.4	Similarity between the reconstructed data (x-axis) and the original timeseries data (y-axis) in each variable, for reconstruction of rank 1	42
5.5	Similarity between the reconstructed data (x-axis) and the original timeseries data (y-axis) in each variable, for reconstruction of rank 3	42
5.6	The squared error in each PC in reconstructions of different ranks	43
5.7	Biplot of $\mathbf{X}$ at 0m depth . . . . .	45
5.8	Biplot of $\mathbf{X}$ at 200m depth . . . . .	45
5.9	Percentage of variance explained in each PC for current in the horizontal plane . . . . .	46
5.10	Reconstructed data (x-axis) and original data (y-axis) in each variable, for reconstruction of rank 1 . . . . .	47
5.11	Reconstructed data (x-axis) and original data (y-axis) in each variable, for reconstruction of rank 2 . . . . .	48
5.12	The squared error in each PC in reconstructions of different ranks for om depth . . . . .	48
6.1	Timeseries the first 100 hours of x- and y-velocity for current at 30m depth, 80m, 150m and 300m. the reconstruction is done using full rank 32. . . . .	50
6.2	MSE in each variable for x- and y-velocity . . . . .	51
6.3	Variance explained with different amount of data, depending on number of PLS components included. . . . .	52
6.4	Timeseries of estimated current and real current using 2 years of testing data. All 32 components are used in the reconstruction (rank 32). . . . .	53
6.5	Timeseries of estimated current and real current using 2 years of testing data. The rank is at its minimum in this reconstruction (rank 4). . . . .	54
6.6	Timeseries of rank 32 of estimated current and real current from testing set with 5 years of training data. . . . .	55
6.7	Timeseries of rank 4 of estimated current and real current from testing set with 5 years of training data. . . . .	56
6.8	Timeseries of rank 32 of estimated current and real current from testing set with 10 years of training data. . . . .	57
6.9	Timeseries of rank 4 of estimated current and real current from testing set with 10 years of training data. . . . .	58
6.10	Mean Squared (sum) Error [cm/s] for testing set at rank 32. . . . .	59
6.11	Mean Squared (sum) Error [cm/s] for testing set at rank 4. . . . .	59

---



---

6.12	Variance explained in forecasting depending on number of PLS components included. Graphs for all three training data amounts are shown. . . . .	60
6.13	Timeseries of rank 32 of forecasted current and real current from testing set with 2 years of training data. . . . .	62
6.14	Timeseries of rank 18 of forecasted current and real current from testing set with 2 years of training data. . . . .	63
6.15	Timeseries of rank 5 of forecasted current and real current from testing set with 2 years of training data. . . . .	64
6.16	timeseries of rank 32 of forecasted current and real current from testing set with 5 years of training data. . . . .	65
6.17	timeseries of rank 18 of forecasted current and real current from testing set with 5 years of training data. . . . .	66
6.18	timeseries of rank 5 of forecasted current and real current from testing set with 5 years of training data. . . . .	67
6.19	Timeseries of rank 32 of forecasted current and real current from testing set with 10 years of training data. . . . .	68
6.20	Timeseries of rank 18 of forecasted current and real current from testing set with 10 years of training data. . . . .	69
6.21	Timeseries of rank 5 of forecasted current and real current from testing set with 10 years of training data. . . . .	70
6.22	Mean Squared (sum) Error [cm/s] for testing set at rank 32. . . . .	71
6.23	Mean Squared (sum) Error [cm/s] for testing set at rank 5. . . . .	71
7.1	Timeseries the first 100 hours of x- and y-velocity for current at 30m depth, 80m, 150m and 300m. . . . .	74
7.2	MSE in each variable for x- and y-velocity . . . . .	74
9.1	Similarity between the reconstructed data (x-axis) and the original timeseries data (y-axis) in each variable, for reconstruction of rank 2	91
9.2	Similarity between the reconstructed data (x-axis) and the original timeseries data (y-axis) in each variable, for reconstruction of rank 4	92
9.3	Similarity between the reconstructed data (x-axis) and the original timeseries data (y-axis) in each variable, for reconstruction of rank 3	93
9.4	Similarity between the reconstructed data (x-axis) and the original timeseries data (y-axis) in each variable, for reconstruction of rank 4	93
9.5	Timeseries of estimated current and real current from training set with 2 years of training data. ncomp = 32 . . . . .	94
9.6	Timeseries of estimated- and real current strength for test set with 2 years of training data. ncomp = 32 . . . . .	95

---

---

9.7	Current directions observed for estimated- and real current strength for test set with 2 years of training data. The estimation of of full rank 32. . . . .	95
9.8	A set of four subfigures. . . . .	97
9.9	Timeseries of forecasted current and real current from training set with 2 years of training data. ncomp = 32 . . . . .	98
9.10	Timeseries of forecasted- and real current strength for test set with 2 years of training data. ncomp = 32 . . . . .	98
9.11	Current directions observed for forecasted- and real current strength for test set with 2 years of training data. ncomp = 32 . . . . .	99

# List of Tables

3.1	The used amounts of data and what data are included. . . . .	27
-----	--	----

---

# Abbreviations

$\rho_0$	=	water density
$k$	=	thermal conductivity of the water
$\Theta$	=	potential temperature
$Q, q$	=	heat, heat flux
$C_p$	=	specific heat capacity for pure water
$S$	=	salinity
$\vec{\omega}, \dot{\vec{\omega}}$	=	angular velocity and angular acceleration
$\vec{v}, \vec{a}$	=	velocity and acceleration
$G$	=	gravitational constant
$\vec{\rho}$	=	vector from Moon centre of gravity to a point $P$
$\vec{T}$	=	tide producing force
$\vec{S}_0$	=	centrifugal force in the Earth-Moon system
$u(z), U_0$	=	fluid velocity, free stream velocity in x-direction
$\tau$	=	Shear stress
$\mu$	=	fluid viscosity (oceanography), mean function (statistics)
$Re$	=	Reynolds number
$\nu$	=	kinematic fluid viscosity
$\delta$	=	boundary layer thickness
$R$	=	Rossby number
$X_{n \times p}$	=	matrix with n rows (observations) and p columns (variables)
$Y$	=	response matrix
$Z$	=	matrix of normalized data
$E[\cdot], Var(\cdot)$	=	expected value, variance
$\sigma$	=	standard deviation
$\Sigma, Cov[\cdot, \cdot]$	=	covariance
$\rho$	=	correlation
$\sigma_f^2$	=	maximum allowable covariance
$\sigma_m^2$	=	length parameter
$\sigma^2$	=	noise variance
DoC	=	Degree of Compression
MSE	=	Mean Squared (sum) Error
PCA	=	Principal Component Analysis
PLSR	=	Partial Least Squares Regression
GPR	=	Gaussian Process Regression
POD	=	Proper Orthogonal Decomposition
SLC	=	Standardized Linear Combination
SVD	=	Singular Value Decomposition

# Introduction

The Ocean have been studied for centuries, as it has always been an important way for humans to travel and explore the world. Therefore, we already know a lot about the ocean dynamics and the processes that affects the water. Due to high activity in offshore industries such as oil and shipping, these research fields are still pushing the limits of our present knowledge every day. With ships operating mainly on the ocean surface, waves can be considered the most complex and important component of ocean dynamics to be able to handle. But in modern industry applications, ocean currents are at least an equally important component. The oil industry conveys the impression that further oil exploration will take place in deeper waters and in Arctic areas. Whilst persistently being pressured to deliver oil and gas in a more cost effective manner. Consequently, the industry is introduced to new challenges.

A step towards addressing such challenges are to obtain even more knowledge of ocean current dynamics. This is useful information in many applications. In Arctic environments, knowledge of the surface current movement can be used in order to determine and predict iceberg movement. This can be valuable information in determining if an iceberg poses a risk on an installation, and as a result of this be able to maximize the production time of the installation. In deeper water applications, knowledge of the entire water column circulation is valuable in applications where installations stretches partially or entirely through the water column. This is useful in the design of risers, umbilicals and when planning lengthy marine operations.

## 1.1 Problem definition

The work in this thesis aims to explore the relation between wind and ocean currents at different depths. Modern regression techniques are used for compressing the dataset, while still keeping enough information for further processing, and

mainly for two purposes. The first is for forecasting current based on historic data from the entire water column and wind. The other is for estimating deeper currents based on wind and currents from the surface to a depth of 20m, defined as top currents. The work is based on a project thesis on the same topic written during autumn 2018.

The hypothesis to be investigated is that, given a specific location, is there a correlation between wind and surface top currents in relation to deeper currents? Furthermore, based on historic data, can the current be predicted by using multivariate modelling techniques?

To explore this, the objectives of the thesis are:

- Review important oceanographic processes in the ocean to get a understanding of the main underlying phenomenon of ocean processes.
- Explore dominating effects in wind and current by performing Principle Component Analysis along depth- and horizontal spatial dimensions of the dataset. Find correlation between current/current and wind/current.
- Make Partial Least Squares Regression for estimating deeper current based on wind and top currents.
- Make Partial Least Squares Regression for forecasting water column based on historic data.
- Explore other analysis using different method and compare the findings. Discuss pros and cons.

## 1.2 Literature review

Common methodologies used when researching dynamic processes in the ocean often follows a similar procedure. A simulation model of the flow is made, based on established mathematical models or state-of-the-art. After the simulations are performed, the results are compared with field measurements in order to make a conclusion. A different approach might be to use field measurements directly in analyses to find trends.

Ocean waves are widely studied, and although they are closely connected with other ocean processes, in many ways it can be considered its own field. Waves are usually modelled using potential theory, disregarding viscous and turbulent effects (Holmedal, 2002). Although ocean waves and currents are closely connected, ocean currents are the main interest in this project. Ocean currents are hard to study

because it is usually complicated to distinguish one phenomena acting on the water masses from another. Furthermore, the ocean current acts in the entire water column, where the surface layer is strongly driven by wind and waves. How this influences water through the water column, referred to as the deeper layers, can be more complicated to determine. The influence of wind on the current, is particularly interesting. There are many processes influencing the ocean circulation, such as heat flux, gravitational effects from the Moon and Sun, the Earth's rotation, and meteorology (Holmedal, 2002). It is also evident that water particles relates mutually when they are exposed to a disturbance.

Many years of offshore industrial activity have resulted in correspondingly extensive data sampling. Both current speed and corresponding direction, referred to as the current vector, and the wind vector (speed and corresponding direction) are parameters that needs to be studied vigorously in offshore operations. They contain valuable information on tendencies in the ocean. Year around measurements on the ocean current vector include trends regarding seasonal changes and the time-dependent current vector as a result of all the mechanisms acting in the ocean. Wind measurements are sampled equally extensively as the current. Analyzing this data is challenging due to the vast amount of data points, making it hard to grasp and consequently to distinguish abnormal behavior. Modern data analysis offers a range of methods applicable to process timeseries datasets. Many methods are capable of handling multiple dimensions by analyzing correlation between the data points. Methodologies relevant in ocean current analysis might be for the purpose of dimensional reduction.

With such large datasets, many analysis tools do not suffice in analyzing several dimensions at the same time. Nor do they have the capability to describe how all the data correlates. This is where multivariate analysis methods can be a useful tool. In general, methods used in multivariate analysis finds variables that contribute most to the variance in the system, as well as isolating variables that are correlated. In that sense, the method reduces the datasets to a few components that describes a specific behavior. The results are presented in a graphical manner typically used for parameter reduction applications in order to interpret the result, in spite of many variables (Martens and Martens, 2001).

### **1.3 Hindcast data**

The analysis is based provided data from Equinor (Røed et al., 2015a,b). It includes extensive, long-term hindcast data from 1985 to 2012 of wind- and current-strength and direction through the water column. The wind measurement referred

to, corresponds to one point at a height of 10m above the sea surface. The current is measured in several points through the water column. Through the deep, there are one measurement for each 10th meter for the first 100m (including the surface). From 100m to 200m depth, there are one measurement for each 25th meter and finally, from 200m to 300m, there are 50m between each measurement. The difference in distance between measurement points is because the variation in the deeper layers are usually less significant than in the surface layers.

The sets are developed from raw data timeseries originating at moored current measurement stations. The raw data are prone to disturbances and measurement failures providing non-continuous timeseries that are limited in time (Røed et al., 2015a,b). Consequently, the data is thoroughly validated against other measured data. The validation is based on an eddy-permitting, coupled ice-ocean circulation model where the hindcast timeseries are developed by extensive numerical methods and by validating the data with previous years (Røed et al., 2015a). This is a cost efficient method of producing large scale data, even though it is computationally demanding. The resulting model gives a well represented wind and current, but it does not provide the correct values at a specific time. These Equinor Metocean datasets are considered the best available and are currently used when establishing design criteria for the entire Norwegian Continental Shelf.

In this application it is advantageous to have a complete structure in the dataset for simplicity in regards to data handling. Furthermore, accurate data in specific points are not required, as long as the data represents the dynamics, which are obtained from the combination of measurements and numerical validation. Therefore, hindcast data are considered the best choice for the analysis. The wind data are from the NORA10 model and the current from the BaSIC4 hindcast model with 4x4 km grid in the horizontal plane (Røed et al., 2015a). BaSIC4 is a recently improved hindcast model with a high horizontal resolution grid which allows for a more detailed assessment, as the BaSIC4 model is eddy-permitting (Røed et al., 2015a). Eddies are related to mixing in the ocean as a result of turbulent or spinning flows (Stewart, 2008). The data are from a relevant area in the Barents Sea, with an average water depth of 300m. For the purpose of this work, the specific location is not important and will not be revealed.

## 1.4 Ocean currents

Ocean currents are circulatory systems of vertical and horizontal components produced by variations in gravity, wind, friction, and density across the ocean. Phenomena such as heat exchange on the surface and salinity differences makes for ad-



ditional, globally continuous sea current components that combines the five world oceans. It is difficult to distinguish how these components contribute to the ocean current individually. In Myrhaug (2012), the phenomena regarded as current are defined as fluctuations with periods larger than 10 minutes. The following forces generate the ocean current:

- Wind forces.
- Gravitational forces between Earth, Moon and Sun.
- Variations in air pressure.
- Forces due to density variations.
- And more...

These forces alters the current after it is formed:

- Coriolis force (rotation of the Earth)
- Friction (turbulence and viscosity)
- Gravitational forces (buoyancy)
- Topography

An advantage of fluid dynamics is that one governing set of equations, the Navier-Stokes equations, completely describes all effects in the flow field of the fluid. The disadvantage of these equations are that the fully developed Navier-Stokes equations are very complex, especially in regards to friction forces, and require significant computational time. Therefore, they are usually simplified in practice (Holmedal, 2002).

## **1.5 Analysis tools**

During the last years, words such as big data and machine learning have been associated with modern multivariate data analytic methods and are very popular. But the meaning behind these terms includes a range of concepts, including many that are not necessarily new. Still, a common denominator are that they are closely associated with the concept of extracting valuable information from large and/or complex sets of data.

The two linear multivariate data modelling techniques applied in this work, are Principal Component Analysis (PCA) and Partial Least Squares Regression (PLSR).

Both methods are well established and widely used in problems of classification, discrimination, prediction and modelling of data. PCA was first introduced in 1901 by Pearson (Pearson, 1901), and further developed by Hotelling around 1933. This makes it one of the oldest and most recognized techniques of multivariate analysis. In modern applications, Jolliffe (2002) is one of the most popular references, where it is obvious that the more than 100-year-old method is both relevant and widely used in many modern fields and applications. Naturally, this leads to a variety of methods that are either very closely related to PCA or entirely equal. One of these are PLSR, which is considered a relatively new method derived from PCA, but with clear variations. PLSR were first developed by S.Wold et al. (1984) around 1984, and has developed into a useful tool in environmental exploration (Mudge, 2015). A non-linear method that is (briefly) explored in this work is Gaussian Process Regression. This method is considered a parametric supervised learning technique and is used in regression and in classification. Once again, this is not a new method, and the work can be dated back to at least the 1940's. The method is considered well known in meteorology, mainly from Thompson in 1956 and Daley in 1991 (Rasmussen and Williams, 2006).

These methods are chosen because they are not very complex to apply, and expected to be well equipped for these kinds of data. They are also recommended by scholars well familiar with what is available. Still, as always, there is no guarantee that the methods will give the wanted results for these data. As always when working with computers and computational tools, the method can only be as good as the data that is given.

## Driving mechanisms in the ocean

Driving mechanisms in the ocean are divided between phenomena that produces, and phenomena that alters oceanographic forces. Most of the oceanographic theory are from Myrhaug (2012).

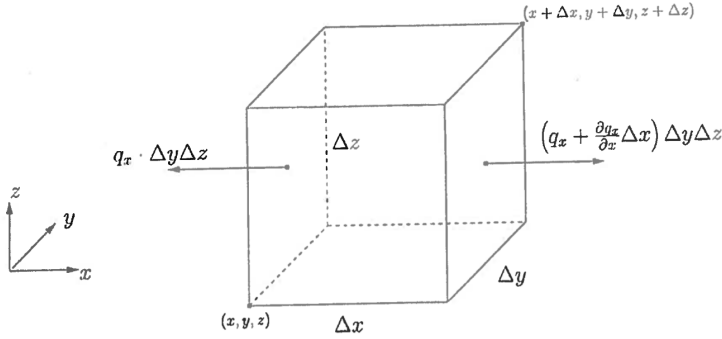
### 2.1 The oceanic heat budget

The heat budget is defined as the sum of the changes in heat fluxes into or out of a volume of water. In the upper layers of the ocean, variations in stored heat are caused by a local imbalance between input and output through the sea surface. The heat flux through the surface is usually much larger than to deeper layers. Globally, the heat flux must balance. The most important terms in the sea surface heat budget are:

- $Q_{SW}$ , flux of sunlight into the sea
- $Q_{LW}$ , net flux of infrared radiation from the sea
- $Q_S$ , flux of heat through surface due to convection
- $Q_L$ , flux of heat carried by evaporated water
- $Q_V$ , heat carried by currents.

Denoting the resulting heat gain or loss as  $Q_{Total}$ , the equation for conservation of heat reads:

$$Q_{Total} = Q_{SW} + Q_{LW} + Q_S + Q_L + Q_V \quad (2.1)$$



**Figure 2.1:** Net heat flux in the x-direction of a cubic element. Illustration: Myrhaug (2012).

Looking at the total heat transport through a cubic fluid element illustrated in figure 2.1, the net flux in each direction is:

$$\text{x-dir: } \frac{\partial q_x}{\partial x} \Delta x \Delta y \Delta z \quad (2.2a)$$

$$\text{y-dir: } \frac{\partial q_y}{\partial y} \Delta x \Delta y \Delta z \quad (2.2b)$$

$$\text{z-dir: } \frac{\partial q_z}{\partial z} \Delta x \Delta y \Delta z \quad (2.2c)$$

Which leads to a total net flux out of the cube (for each cube) as in equation 2.3,

$$\rho_0 C_p \left( \frac{\partial \Theta}{\partial t} + \vec{v} \Delta \Theta \right) = - \frac{\partial q_x}{\partial x} - \frac{\partial q_y}{\partial y} - \frac{\partial q_z}{\partial z} \quad (2.3)$$

where  $\Theta$  denotes potential temperature,  $C_p$  is specific heat capacity for pure water and  $\rho_0$  is the water density. Fourier's law is used as stated in equation 2.4 where  $k$  denotes the thermal conductivity of the water.

$$(q_x, q_y, q_z) = - \left( k \frac{\partial \Theta}{\partial x}, k \frac{\partial \Theta}{\partial y}, k \frac{\partial \Theta}{\partial z} \right) \quad (2.4a)$$

$$\vec{q} = -k \Delta \Theta \quad (2.4b)$$

Finally, combining these equations leads to the heat equation in equation 2.5

$$\frac{\partial \Theta}{\partial t} + \vec{v} \Delta \Theta = k \Delta^2 \Theta \quad (2.5)$$

A similar reasoning as for the heat equation is used for developing the salinity equation, where  $S$  is salinity and  $D_s$  is salinity diffusivity, and yields equation 2.6.

$$\frac{\partial S}{\partial t} + \vec{v} \Delta S = D_s \Delta^2 S \quad (2.6)$$

The equations describing heat flux and salinity flux are the governing contributions to the deep ocean circulation, often referred to as thermohaline circulation. This is a large scale current induced by density and salinity differences in the ocean and the dominant form of water flow globally. Thermohaline circulation occurs far from land and mainly in the subsurface part of the water column (Pinet, 2016).

## 2.2 Tides

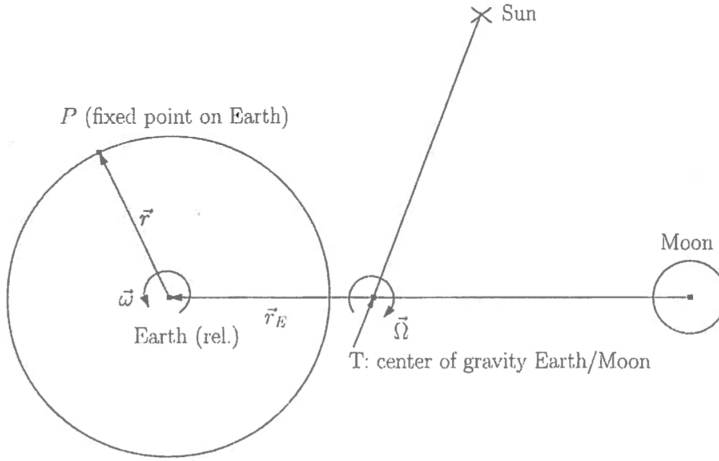
Tides are recognized by a periodic, vertical rise and fall of the sea surface. They are caused by well known gravitational interactions between the Earth, Moon, and Sun. In spite of the Sun having a much higher mass than the Moon, the Moon has almost twice the effect of the Sun towards the tides. This is due to Newton's Law of universal gravitation, where the gravity forces decay by the square distance between the masses. A balance in the gravitational effect between the masses and the centrifugal force caused by a rotating body, are the main forces causing tides.

When describing the physics of the tides, it is common to use the equilibrium model developed by Isaac Newton in the 17<sup>th</sup> century as a simplified model. This includes some assumptions, as the model is based on an idealized world (Pinet, 2016):

- The surface of the Earth is considered completely covered in water and has an infinite water depth implying no interacting effect with landmasses or the seabed
- The Earth is a perfect sphere, making the gravitational force on the surface of the Earth equal in all points  $P$
- The associated tidal waves are considered progressive waves

- At all times, there is an equilibrium between the seawater and the tide-generating forces

The Earth and Moon rotates about a common center of mass T as shown in figure 2.2. As opposed to what shown in the figure, T is located on Earth due to the difference in mass between the two bodies.



**Figure 2.2:** The principle of the Earth, Moon and Sun system. The center of mass T is in reality positioned on Earth. Illustration: Myrhaug (2012)

An Earth-Moon system is considered rotating with respect to the common axis T through the center of mass of the Earth-Moon system. A point P is fixed on the surface of the Earth as in figure 2.2. The acceleration of P can be written as in equation 2.7.

$$a_{P,abs} = a_{P,rel} + 2\vec{\omega} \times v_{P,rel} + \vec{a}_E + \dot{\vec{\omega}} \times \vec{r} + \vec{\omega} \times \vec{\omega} \times \vec{r} \quad (2.7)$$

The  $\vec{\omega}$  and  $\dot{\vec{\omega}}$  are the angular velocity and the angular acceleration of the Earth's rotation, respectively.  $\vec{v}_{P,rel}$  and  $\vec{a}_{P,rel}$  are the velocity and acceleration of the point P in the relative reference frame, respectively.  $\vec{a}_E$  is acceleration of the origin in the relative reference system, and  $\vec{r}_P$  is the position of P in the relative reference system. As P is fixed to the Earth surface, the two first right hand terms of equation 2.7 are equal to zero. Since the Earth rotates very slow,  $\vec{\omega}$  is small and  $\dot{\vec{\omega}} \times \vec{r} \approx 0$ . The last right hand side term is equal all over the Earth surface and normal to the Earth axis. Consequently, it does not contribute to tides (Myrhaug, 2012). Thus, equation 2.7 reduces to equation 2.8.

$$a_{P,abs}\vec{r} = a_T\vec{r} + \vec{\Omega} \times \vec{\Omega} \times r_E\vec{r} \quad (2.8)$$

Where  $\vec{\Omega}$  is the angular velocity of the Earth-Moon system which rotates about the common Earth-Moon centre of mass axis,  $r_E\vec{r}$ .  $a_T\vec{r}$  represents the motion around the Sun. Because the Moon effect has a much larger magnitude than the effect of the Sun, this term can optionally be neglected and thereby, only the Earth-Moon interactions are considered.

Inserting equation 2.8 to Newton's second law yields equation 2.9.

$$\frac{\Sigma F}{m} = a_{P,abs}\vec{r} \quad (2.9a)$$

$$\frac{\Sigma F}{m} - \vec{\Omega} \times \vec{\Omega} \times r_E\vec{r} - a_T\vec{r} = 0 \quad (2.9b)$$

The term  $-\vec{\Omega} \times \vec{\Omega} \times r_E\vec{r}$  is interpreted as a centrifugal force in the relative reference system caused by eccentric motion when Earth revolves around the Earth-Moon common center of mass. This should not be confused with the centrifugal force produced from Earth's spin around its own axis. Due to the concentric property, an arbitrary point on the Earth surface follows a circular path, with a radius that will be equal for all points. Thus, all points are subjected to the same angular velocity. This leads to equal acceleration and centrifugal force in all points (Brown et al., 1999). Neglecting the effect of the Sun for now, the centrifugal force  $\vec{S}_0$  in the Earth-Moon system is equal in all points  $P$  on Earth and denoted as:

$$\vec{S}_0 = -m\vec{\Omega} \times \vec{\Omega} \times r_E\vec{r} \quad (2.10)$$

The physical gravitation force acting on a point  $P$  on the Earth is given by (Myrhaug, 2012):

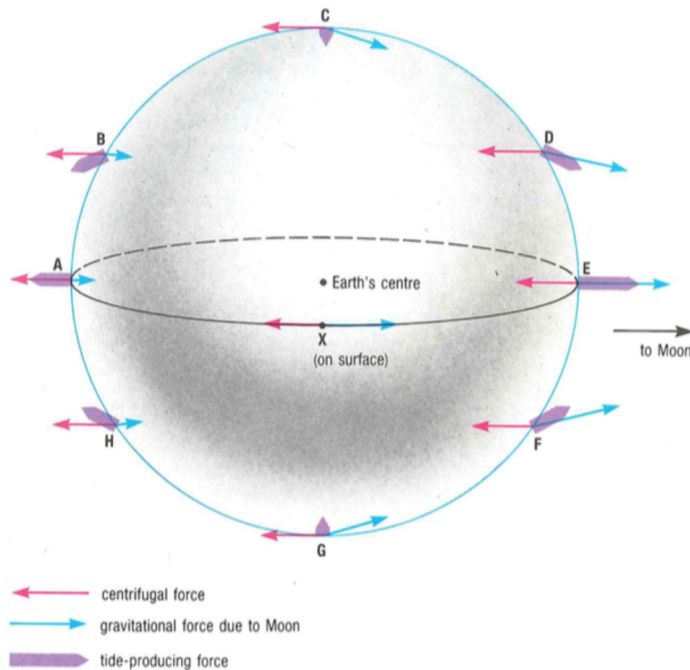
$$\vec{F} = \vec{G} = -G \frac{mM}{\rho^2} \frac{\vec{\rho}}{\rho} \quad (2.11)$$

where  $m$  and  $M$  are the mass of the Earth and Moon, respectively.  $G$  is the gravitational constant,  $\vec{\rho}$  is a vector from the Moon centre of gravity to the point  $P$  and  $\rho$  is the length of  $\vec{\rho}$ . The position of  $P$  determines the magnitude of  $\vec{G}$ , since the distance between the Earth and the Moon is dependent on the position of  $P$ . As a result, the gravitational force from the Moon varies along the surface of the Earth.

Considering the total effect of both the centrifugal force and the gravitational effect of the Moon on the point  $P$  causes a tide producing force with equilibrium in the center of the Earth, and a varying effect in the point  $P$  on the surface of the Earth. The tide producing force is denoted as  $\vec{T}$ :

$$\vec{T} = \vec{S}_0 + \vec{G} \quad (2.12)$$

$\vec{T}$  is the only force contributing to the tidal force, and its direction depends on the position of the Earth surface in relation to the position of the Moon (Brown et al., 1999). This causes the periodically varying tidal force in different points  $P$  on the Earth to be as in figure 2.3. It shows the (un-scaled) magnitude and direction of the tide-producing force for different points on the surface of the Earth with respect to the Moon. The tides rises and falls periodically, with maximum and minimum values every sixth hour, resulting in two low tides during 24 hours. The high tides are moved 50 minutes every day due to Earth rotation, but this variation is highly regular and can easily be predicted for the near future.



**Figure 2.3:** Un-scaled derivation of the centrifugal, gravitational and resulting tide-producing forces. Illustration: Brown et al. (1999)



### 2.2.1 Solar effects on the tides

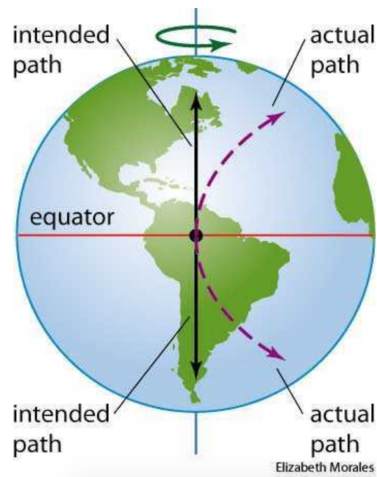
Although the effect of the Sun have been neglected in the previous derivation of the tidal-producing forces  $\vec{T}$ , its effect on the tides will contribute when working with data from the real world. Specially in regards to minimum and maximum tidal-producing force, known as spring- and neap-tides. The tidal force from the Sun is deducted in the same way as for the Moon. But the Earth-Moon-Sun system will have a different oscillation period due to the Earth and Moon orbiting the Sun with a different period than the Earth-Moon rotation. When the Earth, Moon and Sun are aligned, the tidal force from the Moon and sun works in the same direction producing a maximum force called spring tide. On the other hand, when the Earth, Moon and Sun forms a  $90^\circ$  angle, the Moon and Sun produces opposite tide-raising forces which causes destructive interference resulting in a minimum tidal force, called neap tides. Spring and neap tides occurs twice a month (Brown et al., 1999; Myrhaug, 2012).

## 2.3 Coriolis Force

The Coriolis force is the effect of the Earth rotating about its own axis when considering a point  $P$  moving on the surface of the Earth. This is considered a fictive force, because although Newton's first law states that there is no change in motion of a body unless a resulting force acts on it, for an observer in a rotating reference system it will be perceived as a force (Stewart, 2008). Considering the acceleration of the same point  $P$  from equation 2.7 in section 2.2

$$a_{P,abs} = a_{P,rel} + 2\vec{\omega} \times v_{P,rel} + a_E + \dot{\vec{\omega}} \times \vec{r} + \vec{\omega} \times \vec{\omega} \times \vec{r}$$

the term  $2\vec{\omega} \times v_{P,rel}$  represents the Coriolis force. When the point  $P$  is fixed to the Earth, this term is negligible, but if  $P$  is unfixed, the Coriolis force can be important as it will affect all particles moving on the Earth surface, including wind and currents (Myrhaug, 2012). Physically, the Coriolis effect is observed as a deflection of the intended path, with the rotation of the Earth as illustrated in figure 2.4. In the northern hemisphere the path defects towards the right, in accordance with the Earth rotation, and correspondingly towards the left in the southern hemisphere (Stewart, 2008).



**Figure 2.4:** Illustration of the effect of the fictive Coriolis force when a particle moves on the surface of the Earth. Illustration: Morales (2018)

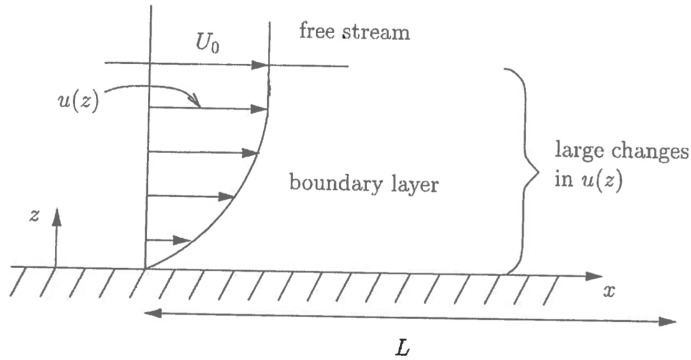
## 2.4 Boundary layer

The boundary layer is a velocity field where there is a transition in fluid velocity due to friction. Normally, two assumptions are made for boundary conditions in fluid mechanics (Stewart, 2008).

- 2-dimensional fluid flow, which means no velocity normal to a boundary.
- No slip condition, which means no flow parallel to a solid boundary.

In its simplest form, a low viscosity fluid flowing with a high Reynolds number over a flat plate is considered. The 2-D fluid field will look like in figure 2.5, where the velocity of the fluid  $u(z)$  is zero on the plate due to the no-slip condition. Through the boundary layer, the velocity  $u(z)$  is much lower close to the wall than closer to end of the boundary layer where the free stream has velocity  $U_0$  (Myrhaug, 2012). The velocity gradient is caused by shear forces between the water particles, holding them back on and close to the plate. As the particles are further away from the plate, the shear forces from the water particles gradually holds the next layer less.

In the ocean, there is a boundary layer both in the top and bottom of the water column. In the bottom of the water column, the flat seabed is holding back the flow. In the top of the water column, there is an equivalent boundary layer where the surface phenomena drives the flow. A third interacting boundary layer is found on the ocean surface between the surface the wind where the geostrophic wind is



**Figure 2.5:** Illustration of the typical velocity profile in a boundary layer for fully developed flow.  $L$  must be sufficiently long to find a change in  $u(z)$ . Illustration from Myrhaug (2012)

considered the free stream. As will be described later, this is not the full truth. In practice, boundary layers are often turbulent, making it harder to distinguish the boundary layer than with laminar flow. Moreover, other phenomena are usually present in the ocean.

### 2.4.1 Boundary layer equations

The boundary layer equations can be derived from the Navier-Stokes equations by making some additional assumptions. The boundary layer equations are not derived in this thesis, but the governing assumptions are presented.

The friction in the fluid in the form of shear stress are causing the change in  $u(z)$  in the boundary layer of a flowing fluid. This can be described as in equation 2.13.

$$\tau = \mu \frac{du}{dz} \quad (2.13)$$

where  $\mu$  is the viscosity of the fluid and  $\frac{du}{dz}$  is the velocity gradient through the boundary layer. In order to describe the flow of a fluid, Reynolds number ( $Re$ ) is used.  $Re$  describes the ratio of internal forces in relation to viscous forces, and can be found as in equation 2.14.

$$Re = \frac{UL}{\nu} \quad (2.14)$$

where  $U_0$  is the free stream velocity,  $L$  is the length of the body and  $\nu$  is the kinematic viscosity of the fluid. Flows with equal  $Re$  are considered flows with mechanical similarity. Lower  $Re$  suggests laminar flow, which is characterized by a orderly flow, and higher  $Re$  suggests turbulent flow which is characterized by a chaotic flow.

The boundary layer thickness depends on the fluid viscosity and the free stream velocity  $U_0$ . Consequently, a turbulent flow usually has a smaller boundary layer than a laminar flow. According to Schlichting and Gersten (2017), the thickness of the boundary layer  $\delta$  is proportional to the square root of the kinematic viscosity as in equation 2.15. In the deduction of the boundary layer equations, a key assumption is a small  $\delta$ .

$$\delta \propto \sqrt{\nu} \quad (2.15)$$

These relations, combined with the scaling of the ocean described next, are used to simplify the Navier-Stokes equations. When looking at the entire ocean as a whole, the width of the ocean is considered much larger than its depth. A commonly used analogy found in literature from Pinet (2016), states that if the entire ocean is scaled down so that the ocean width (length,  $L$ ) is 20 cm, its depth (height  $H$ ) would only be about the thickness of paper. As a result of the ocean scaling, assuming a small  $\delta$  compared to the characteristic length of the body (Schlichting and Gersten, 2017) is considered a valid simplification for a typical ocean basin.

## 2.4.2 Governing equations

The governing equations describing boundary layers are simplified versions of the Navier-Stokes equations. Initially, it is assumed a homogenous, isotropic and Newtonian fluid and an incompressible flow. A reference system analogue to in figure 2.5, where the  $z$ -axis is pointing up is used together with Einstein summation convention,  $x_i = (x, y, z)$  and  $u_i = (u, v, w)$  where  $i = (1, 2, 3)$  and equal indices are repeated (Holmedal, 2002).

This results in the continuity equation shown in equation 2.16 and the Navier-Stokes equations corresponding to equation 2.17.

$$\frac{\partial u_i}{\partial x_i} = 0 \quad (2.16)$$

$$\frac{\partial u_i}{\partial t} + \frac{\partial u_i u_j}{\partial x_j} = -\frac{1}{\rho} \frac{\partial P}{\partial x_j} + \nu \frac{\partial^2 u_i}{\partial x_j \partial x_j} \quad (2.17)$$

A deduction of the boundary layer equations will not be done in this thesis, but rather a brief explanation of how they are deduced. The level of complexity of the Navier-Stokes equations makes them hard to solve and requires significant computational time. Because of this they are usually found in simplified versions, depending on the application. The deductions into the boundary layer equations are based on finding dominating balance in the Navier-Stokes equations. This means that the terms are individually scaled, and based on their relative size the less important terms are neglected. To deduce the boundary layer equations, important simplifications are made in correspondence with the relation where  $\frac{H}{L} \ll 1$ :

$$H \ll L$$

$$w \ll u$$

This results in the boundary layer equations in 2.18. These are much easier to solve than the full Navier-Stokes equations Holmedal (2002).

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial t} + w \frac{\partial u}{\partial z} = -\frac{1}{\rho} \frac{\partial P}{\partial x} + \nu \frac{\partial^2 u}{\partial z^2} \quad (2.18a)$$

$$\frac{\partial P}{\partial z} = 0 \quad (2.18b)$$

$$\frac{\partial u}{\partial x} + \frac{\partial w}{\partial z} = 0 \quad (2.18c)$$

An important property evident from the deduction, is that the pressure does not change through the boundary layer as seen in equation 2.18b.

### 2.4.3 Rossby approximation

The Rossby approximation is an other simplification technique which can be applied to the boundary layer equations. The Rossby number  $R$  compares the importance of the terms related to local flow acceleration with the Coriolis acceleration and it is defined as in equation 2.19.

$$R = \frac{U_0}{Lf} \quad (2.19)$$

Large scale motions in the ocean typically yields a low Rossby number, caused by the Coriolis acceleration term dominating over the acceleration of the flow. This implies that the acceleration terms of the flow can be neglected in the boundary layer equations. Similarly, for a high Rossby number, the flow acceleration dominates over the Coriolis acceleration implying that terms related to Coriolis acceleration can be neglected. In the cases where  $R \approx 1$ , the full boundary layer equations must be solved (Myrhaug, 2012).

#### 2.4.4 Geostrophic flow

A geostrophic flow is considered the flow in a region between two boundary layers. In this area all particles are assumed to have the velocity of a free stream  $U_0$ , if a frictionless flow is assumed. Geostrophic flow is found both in the sea and in the air. Due to horizontally isobaric conditions, the flow follows a straight path along the isobars.

The atmospheric boundary layer is considered the atmosphere within 100m of the sea surface. This layer is influenced by heat flux through the sea surface as well as turbulence and drag from the wind (Stewart, 2008). Because there are different boundary layers on both sides of the sea-air interface, this area is highly influenced by mixing due to turbulence both in the wind and at the sea surface. Consequently, the height of wind measurements is important to minimize the effect of the mixing.

### 2.5 Wind

When wind blows over the ocean surface, currents are generated starting at the top of the surface. This is caused by shear forces between the wind and the water. The motion due to the wind propagates down in the deeper layers of the ocean due to friction forces also between the layers. The most common type of current found in the ocean are referred to as inertial currents which are generated as a response to an impulse that sets the water in motion. The impulse can be a rapid change in wind blowing over the surface or strong wind blowing for a few hours. The driving force moving the water is solely Coriolis force. If frictionless flow is assumed, the motion of the water particles acts very similar to harmonic oscillators (Stewart, 2008).

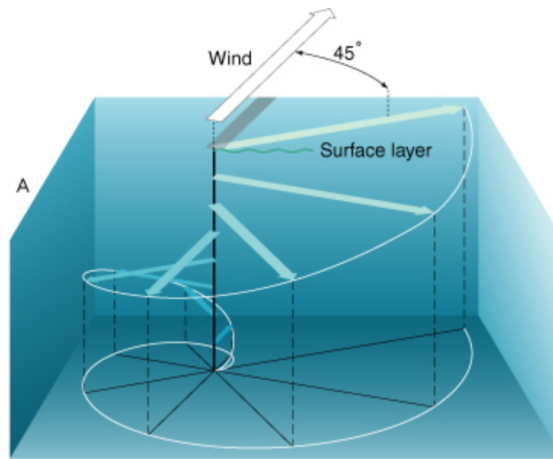
The main driving force of the wind are pressure gradients in the air. In the case of pressure differences, a pressure gradient occurs when a fluid flows from areas with higher pressure towards areas with lower pressure. A steep pressure gradient

indicates a large change in pressure, while a shallow pressure gradient indicates a smaller change. Air moving as a result of pressure gradients are influenced by Coriolis force as described in Section 2.3. In the Northern hemisphere, the Coriolis force acts normally and to the right of the velocity vector from the wind, until a balance is obtained between the Coriolis force and the wind force. This balance is referred to as geostrophic balance, and introduces geostrophic winds (Pinet, 2016).

### **2.5.1 The Ekman layer**

One may think that the wind-generated surface current moves in the same direction as the wind. In 1898, Fridtjof Nansen observed that the icebergs moves at an angle of the wind, which he thought was peculiar. Nansen himself was on expedition, but told his friend about his observations and asked him to look in to the phenomenon. In 1902, V. W. Ekman presented his doctoral thesis describing the dynamics of the phenomena thereafter known as the Ekman Layer (Pinet, 2016; Stewart, 2008).

The Ekman Layer describes a thin layer of typically around 100m to 150m, close to the surface. It is a result of a balance between the Coriolis force, friction forces described as drag and pressure gradient forces described in Section 2.5. The Ekman layer includes a surface velocity vector rotated 45° clockwise of the wind in the Northern hemisphere, when looking downwind as shown in figure 2.6. This is a result of the surface layer dragging the next layer of water into motion in each successive layer through the Ekman layer. In the surface, due to the Coriolis effect, the surface layer is continuously rotated each successive layer due to the wind. This creates a spiral through the Ekman layer known as the Ekman spiral and can be seen in figure 2.6.



**Figure 2.6:** The Ekman spiral describing how the horizontal wind sets the surface waters in motion. Illustration from Ocean Motion, NASA (2018)

The relation between the idealized and real deflection between wind and current direction in the Ekman layer deviates. An angle of  $45^\circ$  are found in idealized conditions and an angle between  $20^\circ$  to  $40^\circ$  in reality (Stewart, 2008). This is mainly due to the assumptions of a steady, homogenous, horizontal flow with friction being violated, as in many real world processes.



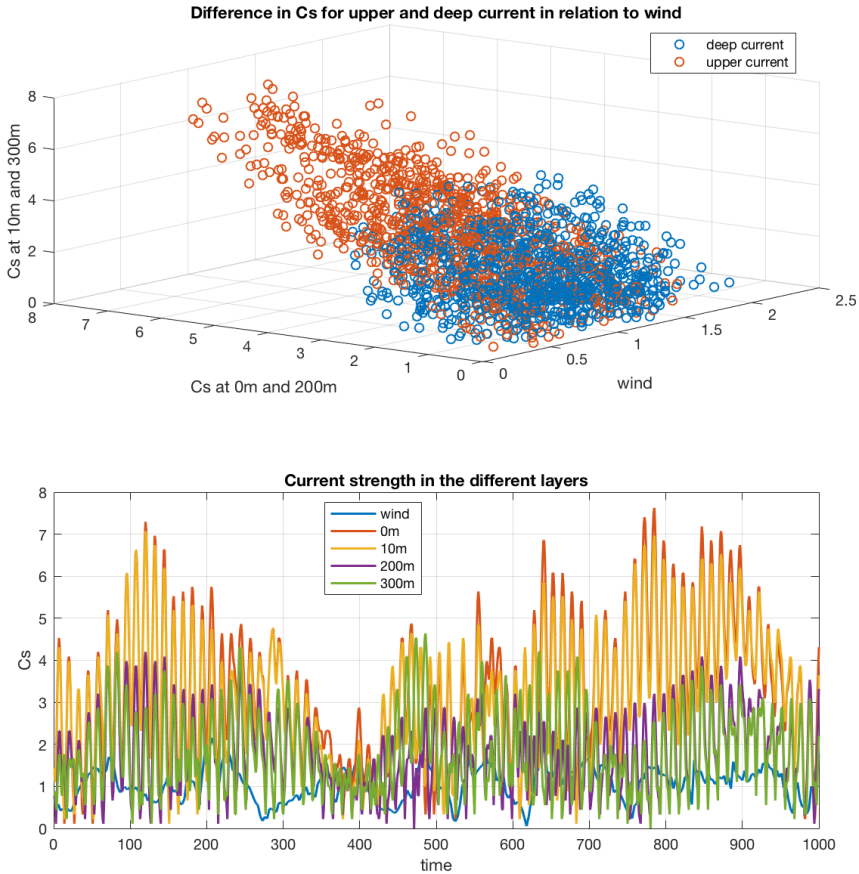
## Pre-processing

Before analyzing the data, it is advantageous to establish some familiarity with the data. That includes exploring relations in the data before they are processed, to get familiar with how the data points relate with each other. Just as important as the interactions within the water column, are the relations with the seabed, which will influence the currents. Therefore, the seabed needs to be explored. Also some statistical properties used in describing the data results later are introduced. Furthermore, before applying analysis algorithms to the data, it is beneficial to have them of a certain structure.

### 3.1 Exploring relations in the data

Before analyzing the data, it is useful to have some ideas of how the current is represented through the water column, including how the wind is represented. Because of the large amount of data available, it is not relevant or even feasible to present all observed relations. Rather, an example of frequently observed relations in the current, as compared with the wind, are introduced. This is shown in the upper figure in 3.1. The top figure shows at what wind- and current velocities observations are made. Blue points represent deep current, where current velocities at 200m and 300m are shown. Orange points represent upper current velocity observations, at 0m and 10m. It is obvious that upper currents are observed at higher velocities than deeper currents, and often at a relatively high wind velocity. Moreover, for top currents the range of observed current strength are close to normally distributed, and it is strongly influenced by wind velocity. For deep currents, it is clearly seen that the observed velocities are generally lower than in the surface, but they are largely influenced by one another. The wind has no visually obvious relation with the deeper currents before the data are analyzed. The lower figure in 3.1 shows the difference in current strength in the different layers through the water column.

Generally, the current strength decreases through the water column, as seen in the upper figure. The wind stands out from the currents, and is expressed as m/s, while the current are expressed as cm/s, to be able to include as many details as possible.

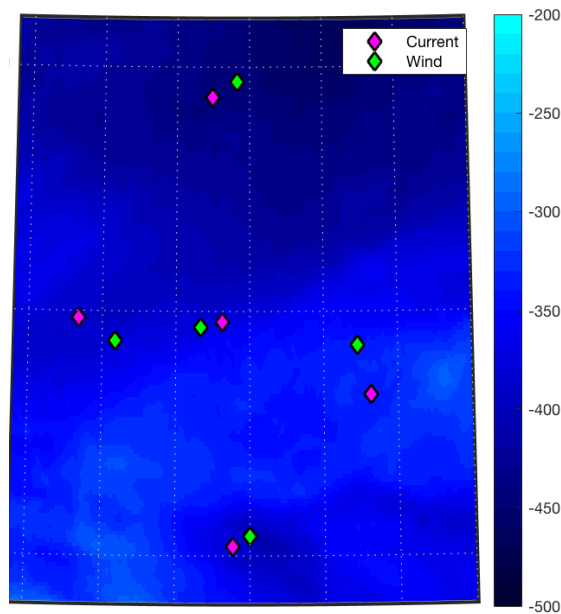


**Figure 3.1:**  $C_s$  related to wind at different depths and  $C_s$  through the water column at different times. It should be noted that the current are expressed in cm/s and the wind in m/s.

### 3.2 Bathymetry

The variations in the seabed will influence propagation and strength of the currents in the ocean. To make the analysis as undisturbed by unwanted effects as possible, it is important to make sure the seabed is fairly even. In this work, that is done by looking at gridded bathymetry data from the General Bathymetric Chart of

the Oceans (GEBCO). This is a non-profitable project organization that provides publicly-available bathymetry data and works towards complete seafloor mapping (British Oceanographic Data Centre (BODC) / General Bathymetric Chart of the Oceans (GEBCO), 2018). Figure 3.2 shows the seabed of the unnamed location where the hindcast data originates from. As the figure shows, the seabed is fairly even and has no sudden changes in depth, even though the northernmost part of the map is moderately deeper than the rest of the area. That might lead to some minor differences in the analysis for the northernmost point in deeper currents, but these are not expected to be crucial.



**Figure 3.2:** Seabed topography of the area where the hindcast data originates from.

### 3.3 Statistical properties

Some statistical properties of the analysis tools used in this project requires to be addressed. This entails both assumptions about the data as well as well known and self defined statistical terms.

### 3.3.1 Stationarity

The assumption of stationarity includes a stochastic process with a probability distribution that is independent of time. This assumption is arguable for the BaSIC hindcast dataset, mainly due to seasonal effects being non-stationary (Newland, 1993). Stationarity is an assumption in all the explored statistical methods in this thesis. One way to practically address this in the analysis, is to remove the mean of the timeseries data. Also, when possible, choosing data from the same time of the year when comparing the years against each other. This reduces the risk of violating the assumption of a stationary process required for many statistical applications. In this work, this is handled in two ways. For analyzes where an integer amount of years are not chosen, the amount of time included is 6 weeks, which is a short enough period to assume that seasonal effects is not effecting the data internally. For other analyzes, an integer amount of years are chosen, to minimize the internal effect of seasonal differences. In all analyzes, the data is normalized before analyzing, and reversed after.

### 3.3.2 Normalization, covariance and correlation

The `pca()` function in Matlab normalize the model by default. In order to do equivalent analyses with different tools, all data are normalized before they are analyzed. Therefore, as a part of data preparation, it is normalized by standard deviation on standard score as in equation 3.1, and reversed back after the analysis.

$$Z = \frac{X - E[X]}{\sigma(X)} \quad \text{where} \quad \sigma(X) = \sqrt{Var(X)} \quad (3.1)$$

where  $Z$  are the normalized data,  $X$  are the dataset,  $E[\cdot]$  is expected value,  $\sigma$  is standard deviation and  $Var(\cdot)$  is variance. In the first scenario using PCA, the data are also scaled in the range between -1 and 1. That step is later omitted in the other analyzes to make interpretation more intuitive.

Often the aim of an analysis is to capture the way the stochastic variables differs in relation to each other. The covariance  $\Sigma = Cov[i, k]$  is a measure of the linear association between a pair of variables, such as  $\mathbf{X}_i$  and  $\mathbf{X}_k$ . The covariance is found by equation 3.2.

$$\Sigma = Cov[i, k] = E[(X_i - \mu_i)(X_k - \mu_k)] \quad (3.2)$$

where  $E(\cdot)$  is the expected value, and  $\mu$  denotes the expected value for each of  $i$  and  $k$  (Johnson and Wichern, 2007; Løvås, 2013). In GPR, the covariance is specified using a kernel function which will be introduced later.

In order to interpret covariance, the correlation between the variables are calculated. It is represented by a correlation  $\rho$ , which is a value between  $[-1, 1]$  (Løvås, 2013). This is somewhat similar to normalizing covariance by standard deviation, and specified in equation 3.3.

$$\rho_{ik} = \frac{Cov[i, k]}{\sqrt{Var[i]Var[k]}} \quad (3.3)$$

The magnitude of  $\rho_{ik}$  indicates the strength of the linear correlation between the variables  $i$  and  $k$ .  $\rho_{ik} = 0$  indicates uncorrelated variables while  $\rho_{ik} = 1$  indicates a perfect correlation. The sign of  $\rho_{ik}$  indicates the direction of the correlation, i.e. whether the variables are positively or negatively correlated with one another (Løvås, 2013).

Autocorrelation is another concept closely connected to correlation, and relevant for the datasets in this project. While correlation indicates how similar two time-series are, autocorrelation compares the timeseries with itself at different lags. Consequently, a autocorrelation matrix indicates at what lag-values the timeseries is similar to itself (Jolliffe, 2002; Newland, 1993). In the hindcast data structure, this means comparing how similar the variables are at different observations. Furthermore, multi-colinearity is another closely related property which refers to the same dynamics being explained in multiple variables (Johnson and Wichern, 2007). This is highly relevant in oceanic data applications, where every water particle is influenced by surrounding particles, and they influence one another interchangeably.

### 3.3.3 Degree of Compression (DoC) and Mean Squared Error (MSE)

Two terms that are useful when describing the results are Degree of Compression (DoC) and Mean Squared Error (MSE). DoC is used to describe how much a dataset is compressed by in an analysis, as compared to the original data structure from the hindcast data. It is relevant in PCA and PLSR, and defined in this work as:

$$DoC = 100 - \frac{rank}{p} \quad (3.4)$$

where rank is the number of components included and  $p$  is total number of variables. Mean squared error (MSE) is useful for describing the error of a variable

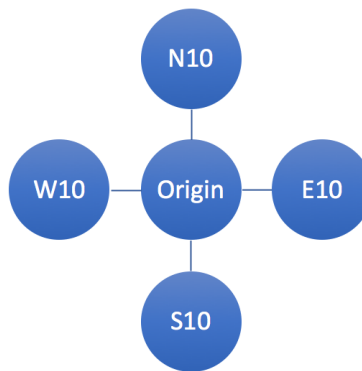
after the data are analyzed (Løvås, 2013). This is relevant when comparing PLSR and GPR, as well as for comparing PLSR with different amounts of training data. It includes the summed error in each variable, and is defined as:

$$MSE = \sum_{n=1}^N \left( \frac{(Y_{n,real} - Y_{n,fit})^2}{N - 1} \right) \quad (3.5)$$

where  $N$  is the number of observations,  $Y_{real,n}$  is the measured response and  $Y_{fit,n}$  is the fitted response in the same specific observation.

### 3.4 Choosing data to analyze

The lowest amount of data used in analyzes includes 1000 points, and are taken from one single point in the horizontal plane, referred to as the origin. In most analyzes where a larger number of observations are included, the analysis is expanded in the horizontal plane as well. The horizontal plane points are distributed and referred to as shown in figure 3.3. The points surrounding the origin are linearly interpolated to a distance of 10km from the origin to the closest found point. This is done for both wind and current data. The closest found points for wind and current measurements are shown in figure 3.2. The projection of the Earth has not been accounted for in this interpolation. The resulting relations are used in both PCA and PLSR.



**Figure 3.3:** Name and configuration of the points.

Based on what is observed after exploring the data, the data that is most equipped for the analysis are chosen. There are a few criteria to consider when choosing

data. The data should, as far as possible, represent the overall data well, to be able to reliably check the capabilities of the methods. While not representing extreme cases which will complicate the interpretation of the results. Another consideration is that the starting date has to be early enough so that the succeeding data includes a sufficient amount to produce both a training set and a testing set, i.e. be twice as long in time as the specified period. As presented in 3.3, the seasonal changes also has to be considered when the amount of data is less than an integer amount of years.

For analyzes using more than a year of data, three periods are used. These periods, referred to as their respective amounts of years, are used in horizontal-plane PCA and in PLSR, and include data from all points shown in figure 3.3. The amounts of years are 2 years, 5 years and 10 years, and are taken from time intervals specified in table 3.1.

Years training	Time interval [dd-mm-yyyy hh:mm:ss]
2	08-March-1990 04:00:00 to 08-March-1992 07:00:00
5	08-March-1986 04:00:00 to 08-March-1991 07:00:00
10	08-March-1985 04:00:00 to 08-March-1995 07:00:00

**Table 3.1:** The used amounts of data and what data are included.

### 3.5 Organizing data

Both wind- and current data are imported into Matlab on their original format from provided hindcast datasets. Their configuration is similar to the preferred structure. The wind is scaled so both wind and current strength is given in cm/s. The directions of wind and current are defined in the same coordinate system, and does not need to be corrected (Røed et al., 2015a). The wind is measured every third hour and not every hour, such as the current, therefore the wind vector is linearly interpolated from having measurements every third hour to every hour.

In all analyzes, rows represents observations for every hour, and the columns represents the variables. The variables are arranged in a chronological structure as the matrix seen in 3.6. When the data matrix is on this format, selecting the different amounts of data is easy.

$$X_{n \times p} = \begin{bmatrix} date & wind & C_0 & C_{10} & C_{20} & C_{30} & C_{40} & C_{50} & C_{60} & C_{70} \\ & & C_{80} & C_{90} & C_{100} & C_{125} & C_{150} & C_{175} & C_{200} & C_{250} & C_{300} \end{bmatrix} \quad (3.6)$$

To prepare for PLSR analysis, each of the variables are further decomposed from polar coordinates ( $C_d$  and  $C_s$ ) into Cartesian coordinates (x-velocity and y-velocity) before being analyzed, to include direction and velocity in the same variables. The resulting matrix is almost twice as large as 3.6, including an x- and y-velocity component arranged couple-wise for every  $C$ -element.



# Multivariate analysis

In modern analysis it is not the lack of data that is the problem, but rather how to process large amounts of it to extract the required information. Furthermore, the interpretation is hard to grasp and thereby it can be difficult to make conclusions based on large amounts of material. In these situations the analysis tools have to be able to handle data of many dimensions, efficiently. Moreover, it is equally as important to present the results in an intuitive manner. Then it is easy to determine the utility of the analyze and make a corresponding conclusion. This is the purpose of multivariate analysis.

## 4.1 Principal Component Analysis methodology

PCA is a well known modal decomposition method widely used in several different fields, from neuroscience to fluid dynamics (Johnson and Wichern, 2007). It is considered a simple statistical method for reducing complex datasets to reveal a simplified structure, while still maintaining most of the dynamics. This is done by identifying the most important modes (or most meaningful basis in linear theory) in order to reconstruct a dataset. By being able to reconstruct the dataset with less modes, the system gets less complicated. It is important to use as few modes as possible as it will both simplify the interpretation and stabilize the results against data errors (Martens and Martens, 2001). At the same time, every mode includes a different part of the process dynamics, so enough modes need to be included so that the system is well reconstructable in the reduced model.

While PCA is considered a modern method in data analysis, it is not a new method. It can be traced back as far as to 1901 (Pearson, 1901), and since then the method has been widely used and developed. Nowadays, it is even known by different names in different fields. For instance, Proper Orthogonal Decomposition (POD)

is used in the field of fluid dynamics. The methods are considered equivalent (Tu, 2013).

PCA has some limitations in regards to time-series analysis. The method is considered linear and thereby not optimal for describing nonlinear dynamics. Still, it can yield good results in such applications as stated by both Vautard and Ghil (1989) and Johnson and Wichern (2007). The modes are not time-dependent, and thereby do not change if the data is reordered.

The basic concept of PCA is to explain the variance-covariance of a selection of variables through *a few* linear combinations between them. In other words, it reduces the number of variables in order to solve the problem of multi-collinearity. Multi-collinearity is the problem when the variables are subject to a high degree of correlation which increases the variance in the regression parameters. That means that large datasets, when many variables are used as independent variables, the variables might be measuring the same characteristics, making the regression display a high degree of correlation (Johnson and Wichern, 2007). Hence, the objectives of PCA are data reduction and interpretation. When the number of predicative numbers are reduced, it is important to make sure that not too much information is lost. Therefore, the principal components (PCs) have to be chosen carefully.

### 4.1.1 Finding the Principal Components

The basic idea of PCA is to orthogonally rotate the axes of an original variable. It is rotated towards the direction of maximum variance of all the original observations, in order to have them coincide. This is done several times, rendering Principal Components (PCs) of different ranks. The successive principal axes are determined with the property of being orthogonal to the previous principal component, and that they maximize the variation of the the projected points (subject to these constraints). The projected values corresponding to this directions of maximum variation are called principal component scores (Johnson and Wichern, 2007).

The procedure of finding the PCs for further analysis are based on the procedures from Johnson and Wichern (2007) and Jolliffe (2002). The data matrix  $n \times p$  is denoted  $\mathbf{X}_{n \times p}$ . It includes the variables where each row  $n$  denotes the observations of each variable  $p$ . Again, the problem is to create a subset of the variables that holds most of the information. In accordance with Section 3.3.2,  $Cov[i, k]$  denotes the covariance between  $X_i$  and  $X_j$ , and  $\Sigma$  denotes the matrix of  $Cov[i, k]$ . The corresponding correlation matrix is denoted  $\rho$ . Both  $\Sigma$  and  $\rho$  are  $p \times p$  symmetric and square matrices.

The expression in Equation 4.1 is a linear combination of the data matrix  $\mathbf{X}_{n \times p}$ .

$$\text{Var}(Y_p) = \Sigma \alpha_i \mathbf{X}_i \quad (4.1)$$

where  $\alpha_i$  are scalars and  $i$  is in the interval  $[1, p)$ . PCs are determined as the uncorrelated linear combinations  $Y_p$  which maximizes the variance in Equation 4.1 (Johnson and Wichern, 2007). In order to eliminate interdeterminacy, a normalization of the expression known as the standardized linear combination (SLC) is introduced. SLC is true if the sum of absolute values are equal to 1, expressed as  $\Sigma \|\alpha_i\| = 1$ . Each PC is chosen as a linear combination that maximizes the variance, subject to this constraint (Johnson and Wichern, 2007).

As in Section 3.3.2, the correlation between variables is, statistically, a measure of the linear dependence between them. A high correlation indicates a high linear dependence between the variables. The rank of a matrix, which is the maximum number of linearly independent rows or columns of a matrix, can be a useful tool for finding the linear dependence. Rank of a matrix should not be confused with rank of the reconstruction, which will be used in interpretation of the results, where rank corresponds with the number of PCs used.

### 4.1.2 Singular Value Decomposition

When the command `PCA()` is implemented in MATLAB, the default algorithm of the function includes the Singular Value Decomposition (SVD) as the eigenvalue decomposition method. SVD methodology is based on factorization. This is the most exact algorithm option in MATLAB for this application (MathWorks, 2018). In the analysis in this thesis, this is referred to as PCA analysis.

In order to explain SVD,  $\mathbf{X}$  is considered the timeseries matrix  $n \times p$  of real numbers. There exists an  $n \times n$  matrix  $\mathbf{U}$  and an  $p \times p$  matrix  $\mathbf{V}$ , both with orthogonal columns such that

$$\mathbf{X} = \mathbf{U} \mathbf{\Lambda} \mathbf{V}' \quad (4.2)$$

where the  $n \times p$  matrix  $\mathbf{\Lambda}$  has positive constants  $\lambda_i$  on the  $i$ th diagonal value of its principal diagonal and the other entries are zero (Johnson and Wichern, 2007). The positive constants in  $\mathbf{\Lambda}$  are called the singular values of  $\mathbf{X}$ . The matrix  $\mathbf{\Lambda}$  can be represented in three standard forms where the difference is the matrix dimensions. It can be presented with the same dimensions as  $\mathbf{X}$ , as a square matrix or on a reduced SVD form which is also square (Hogben et al., 2007).

## 4.2 Partial Least Squares Regression methodology

Partial Least Squares Regression (PLSR) is a linear regression method that predicts a response based on a collection of predictor variables. These are given as the training data matrix  $\mathbf{X}$  (Johnson and Wichern, 2007), and has many elements to it that is similar to PCA. In PLSR the new response variables are chosen to satisfy three conditions simultaneously, as stated in Glen et al. (1989). The first condition is that they are highly correlated with the response ( $\mathbf{Y}$ ). The second condition is that they model as much of the variance in  $\mathbf{X}$  as possible. The idea behind this is that the components with the largest variance corresponds to the largest signal to noise ratio. The third condition is that they are uncorrelated with each other, which minimizes the redundancy of information (multi-collinearity) and the necessary number of variables. The main element that separates PLSR from PCA is condition one, as opposed to PCA which chooses its components based on the highest variance distinguished from the mean, without particular regards to a response. The technique aims to find a linear decomposition of  $\mathbf{X}$  and  $\mathbf{Y}$  such that,

$$\mathbf{X} = \mathbf{TP}^T + \mathbf{E} \quad (4.3)$$

$$\mathbf{Y} = \mathbf{UQ}^T + \mathbf{F} \quad (4.4)$$

where

$$\begin{aligned} \mathbf{T} &= \mathbf{X} - \text{scores} & \mathbf{U} &= \mathbf{Y} - \text{scores} \\ \mathbf{P} &= \mathbf{X} - \text{loadings} & \mathbf{Q} &= \mathbf{Y} - \text{loadings} \\ \mathbf{E} &= \mathbf{X} - \text{residuals} & \mathbf{F} &= \mathbf{Y} - \text{residuals.} \end{aligned}$$

When the covariance is maximized between  $\mathbf{T}$  and  $\mathbf{U}$ , the decomposition is finalized. This can be achieved with many different algorithms, where all of them follows an iterative approach to extract  $\mathbf{X}$ -scores and  $\mathbf{Y}$ -scores (Maitra and Yan, 2008).

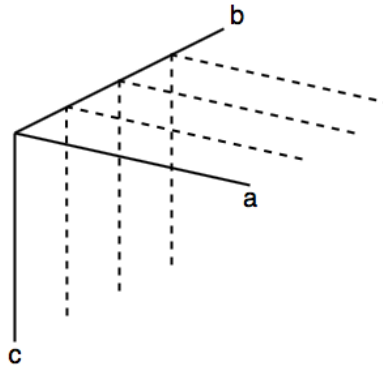
### 4.2.1 SIMPLS algorithm

On its classical form, the nonlinear iterative partial least squares (NIPALS) algorithm is used to find weights for maximizing the covariance between the components. In MATLAB, SIMPLS algorithm is used for calculation of the weights. This is a relatively new method developed by de Jong (1993) (MathWorks, 2019b),

which is faster when implemented in the MATLAB Statistics toolbox. This is because it does not involve a breakdown of the  $\mathbf{X}$ -matrix and yields only the minimal output required for estimating prediction errors during validation. More details about the algorithm can be found in the publication about the novel algorithm by de Jong (1993).

#### 4.2.2 Prediction and response matrices

Before applying the PLSR algorithm, it is important to define the predictor  $\mathbf{X}$ - and response  $\mathbf{Y}$ -matrices correctly. These are defined as shown below in equation 4.5, where  $xy$  is a shortened term for two separate points, one for  $x$  and one for  $y$ , with the same inputs. The corresponding meaning of the other input variables are shown in figure 4.1.



**Figure 4.1:** The ocean dimensions, where a is North, b is East and c is depth.

When estimating data and forecasting data, the main practical difference is how the  $\mathbf{X}$ - and  $\mathbf{Y}$ -matrices are set up, which again determines how the algorithm is trained. It should be noted that in this part of the analysis, the training data are used. When variables are estimated, the predictor matrix  $\mathbf{X}_{n \times p}$  includes the top currents that are measured and which the estimation thereby is based on.  $\mathbf{Y}_{n \times m}$  includes the observed responses through the rest of the water column, at the same time instances  $n$  as in  $\mathbf{X}_{n \times p}$ . In forecasting, data from the entire water column are used in both  $\mathbf{X}_{n \times p}$  and  $\mathbf{Y}_{n \times m}$ , but where  $n$  corresponds to different times where

$\mathbf{Y}_{n \times m}$  includes the consecutive data from  $\mathbf{X}_{n \times p}$ .

$$\mathbf{X}_{n \times p} = \begin{bmatrix} xy(a, b, c, t_0) & xy(a, b, c, t_{-1}) & xy(a, b, c, t_{-2}) \\ xy(a, b, c, t_1) & xy(a, b, c, t_0) & xy(a, b, c, t_{-1}) \\ xy(a, b, c, t_2) & xy(a, b, c, t_1) & xy(a, b, c, t_0) \end{bmatrix} \quad (4.5a)$$

$$\mathbf{Y}_{n \times m} = \begin{bmatrix} xy(a, b, c, t_0) & xy(a, b, c, t_1) & xy(a, b, c, t_2) \\ xy(a, b, c, t_1) & xy(a, b, c, t_2) & xy(a, b, c, t_3) \\ xy(a, b, c, t_2) & xy(a, b, c, t_3) & xy(a, b, c, t_4) \end{bmatrix} \quad (4.5b)$$

### 4.3 Gaussian Process Regression methodology

Gaussian process regression (GPR) is a regression method that is quite different from PCA and PLSR. It is a nonlinear, parametric method for modelling of dependent data. The method is popular, and two essential properties makes it favored. First, the mean and covariance functions completely determines a Gaussian process. This means that to do model fitting, only the first- and second-order moments of the process requires specification, as well as the assumption of a Gaussian process. Second, solving the prediction problem is relatively straight-forward, normally computed by using recursive formulas (Rasmussen and Williams, 2006). The main limitation of GPR is that computational demands increase fast with increased training data.

The method works by defining that a real-valued stochastic process  $\{X_t, t \in T\}$  where  $T$  is an index set, is a Gaussian process if all the finite-dimensional distributions have a multivariate normal distribution. That is, for any choice of distinct values,  $t_1, \dots, t_k \in T$ , the random vector  $\mathbf{X} = (X_{t_1}, \dots, X_{t_k})'$  has a multivariate normal distribution with mean vector  $\mu = E[\mathbf{X}]$  and covariance matrix  $\Sigma = Cov(\mathbf{X}, \mathbf{X})$ , denoted by  $\mathbf{X} \sim N(\mu, \Sigma)$  (Davis, 2001). The method aims to model the response variables as a function of the predictor variables and an error  $\epsilon$ , as  $y = f(x) + \epsilon$ . As new points are observed, GP finds a distribution over these functions by using a Bayesian probabilistic approach to find the best fitted function. The function is updated as new points are observed (Quiñonero-Candela and Rasmussen, 2005).

The toolbox for GPR in Matlab includes the function `fitrgp()`. A great advantage of using this function is that the input parameters for the dataset ( $\mathbf{X}$  and  $\mathbf{Y}$ ) are on the same format as when doing PLSR analysis. Thus, the training data are still on the form  $\mathbf{X}_{n \times p}$ , where  $n$  corresponds to the number of observations and  $p$  to

the number of variables. In the GPR model, input values are chosen to specify the method for this application. Some of the most important inputs are stated. For the Kernel (covariance) function, the option 'ardsquardexponential' is chosen as the covariance function form. This makes the covariance function a squared exponential kernel function with a separate length scale for each predictor defined as:

$$k(x_i, x_j | \theta) = \sigma_f^2 \exp \left[ -\frac{1}{2} \sum_{m=1}^d \frac{(x_{im} - x_{jm})^2}{\sigma_m^2} \right] \quad (4.6)$$

By choosing this kernel function, the maximum allowable covariance is defined as  $\sigma_f^2$  which will be high for functions that covers large ranges of the  $y$ -axis, which again will give a highly correlated regression and a smooth signal.  $\sigma_m^2$  is defined as the length parameter, which influences if observations far from the original point might have negligible effect. The following includes specified parameters for the analysis, and is based largely on the documentation found in MathWorks (2019a). The fit method and the predict method are both chosen as exact Gaussian process regression. This is the default choice when the  $\mathbf{X}$ -matrix includes up to 2000 and 10 000 observations, respectively, which will be the case in this analysis. The hyperparameters, which are parameters that has to be deduced and/or guessed, are optimized by attempting to minimize the cross-validation loss (error) for the function `fitrgp()` by varying the parameters. The option 'auto' optimizes these based on a Bayesian update procedure by using different values of  $\sigma$ , where  $\sigma^2$  is the noise variance which should be minimized. More details on this can be found in the MATLAB documentation in MathWorks (2019a) and in Ebden (2008). The referred literature also includes details on the hyperparameters optimization options that are chosen in correspondence with the optimization method. The most important aspect is that a method that yield reproducible results are chosen, which is important because the optimization depends on the runtime of the objective function.





# Principal Component Analysis

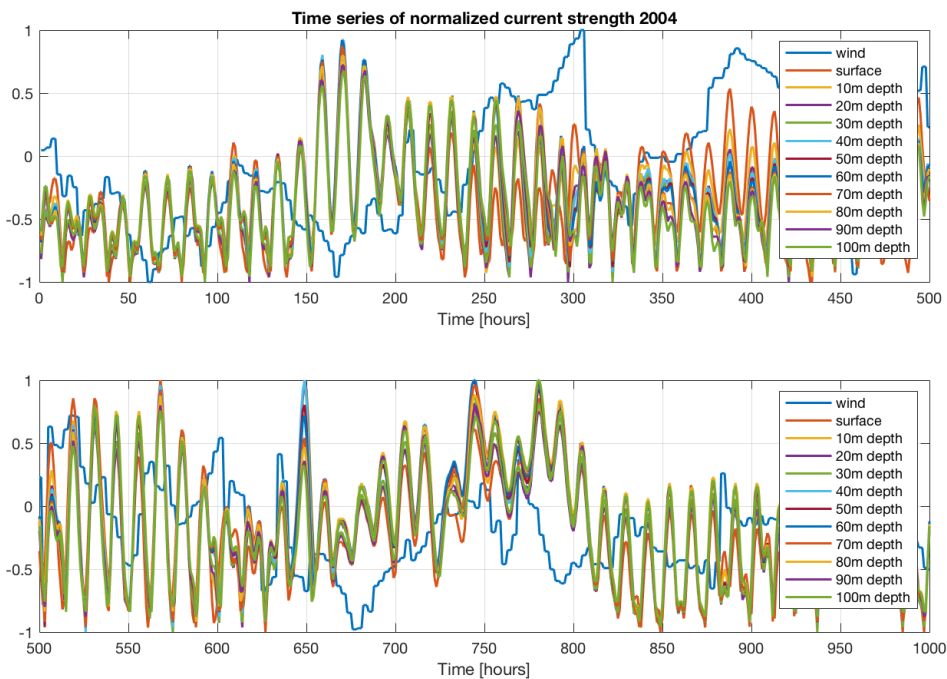
Building on analyses done in the project thesis on the same topic (Tørresen, 2018), larger datasets with multiple dimensions are explored. 1000 observations are used in the first analysis, corresponding to a period of 6 weeks from year 2004. The first scenario includes 12 variables, where the first variable is wind and the other variables are sampled through the upper 100m of the water column. These are high resolution measurements, sampled for each 10th meter. The variables corresponding to different water depths are referred to as layers through the water column. Wind is included to explore relations in the dataset with an expected weaker correlation. This is a result of the wind generally having a different magnitude and being more fluctuating compared to the current. The second analysis includes exploration of the spatial horizontal directions in the dataset. Five different points at two depths are explored. The intention is to see the influence of each point compared to one another. As analysis tool for PCA, the `pca()` function in MATLAB is used and the results are visually presented by using graphs associated with PCA interpretation.

## 5.1 First scenario: Twelve variables through the water column

The first PCA is done by looking at the current strength in twelve different depth points through the water column, including wind. The different layers are referred to by their respective depths. The selection of data points are based on a six week period from year 2004, corresponding to 1000 observations. This produces the data matrix  $\mathbf{X}_{1000 \times 12}$ .

### 5.1.1 Timeseries and normalizing

A normalized and scaled timeseries of the data before they are analyzed at the different depths is shown in figure 5.1. It includes a lot of information and it can be hard to read exactly the intended information in a graphical representation like this. This underlines the need for a simplification and parameter reduction. Nevertheless, one can see that the current strength is highest in the top layer of the water column and reduces through the water column down towards the bottom. This is in accordance with chapter 2.4. The wind has a very distinct behavior compared to the current, and it is obvious that it has a weaker correlation than the layers in the water column. However, they might still be correlated, which in such case will be evident in the analysis.



**Figure 5.1:** The timeseries from year 2004 of wind and through the watercolumn from surface to 100m

### 5.1.2 Variance

Figure 5.2 shows a biplot of the analyzed data. It includes a scatter plot, represented by red dots, and the system eigenvectors for each variable. The axes rep-

resents the three first PCs, which means that the variable eigenvectors show how much of each variable is explained in each PC. In accordance with the principle of PCA explained in chapter 4.1, the dimension of largest variance is represented by the Component 1-axis (which is equivalent with PC1-axis). The eigenvectors of the current-variables are directed in both the positive- and negative directions of the PC1 axis, corresponding with positive and negative correlation, respectively. The figure shows that the layers through the water column are highly correlated, as they are explained as a combination of PC1 to PC3. The variable representing wind also has variance along PC1, meaning it is somewhat explained by PC1. But the variance in the wind-variable stands out being mainly explained in PC2, which some of the deeper currents are negatively correlated with. PC3 also includes a small portion of the wind variance, but mainly the surface current-variance is included in PC3. This means that the largest variance is found both in current and in the wind. The second largest variance is found mainly in the wind and in deeper currents, and the third largest variance is found mainly in the surface currents. When doing PCA on all measurements through the water column, a very similar composition is observed.

What this biplot generally indicates, is that by making a reconstruction of rank 1, equivalent with using only PC1, some of the variance from all variables are included. From looking at figure 5.3 in combination with figure 5.2, it is seen that a reconstruction of rank 1 includes a little less than 90% of the total variance in the data. Increasing to rank 2 improves the reconstruction of the wind parameter and the deeper currents, as their variance are largely explained in PC2. Moreover, this means that the reconstruction explains more than 95% of the total variance. Further increasing to rank 3 largely improves the surface currents reconstruction, and the total variance explained is very close to 99%. A reconstruction of rank 3 includes only 25% of the amount of data from the original  $\mathbf{X}$  matrix which means the degree of compression (DoC) is 75%, and still represents nearly all of the system variance.

Figure 5.3 shows how much of the total variance is explained in each rank, cumulatively. As the theory in chapter 4.1.1 indicates, the first PCs includes most of the system variance. The DoC is also shown in the same figure. It indicates the amount that the data are compressed by, as compared with the original dataset in  $\mathbf{X}$ .

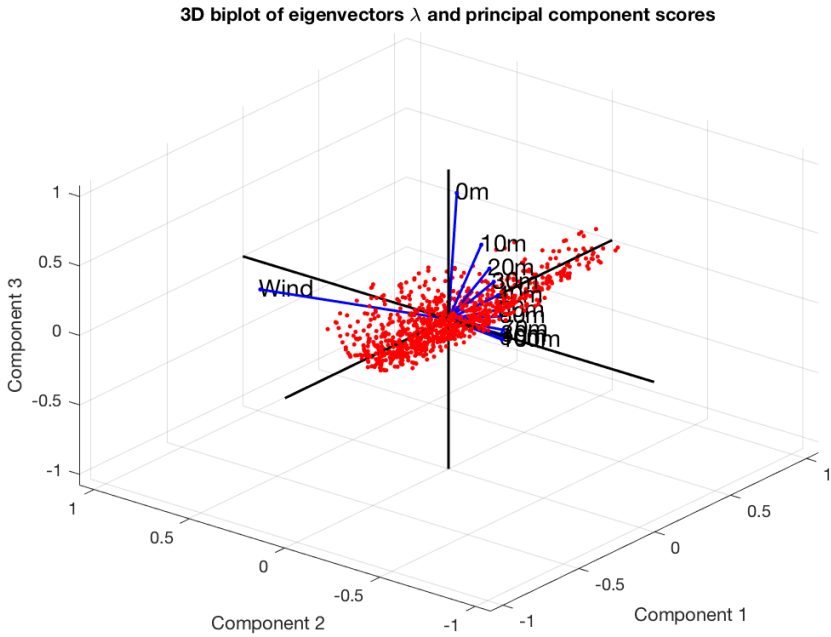


Figure 5.2: Biplot of X

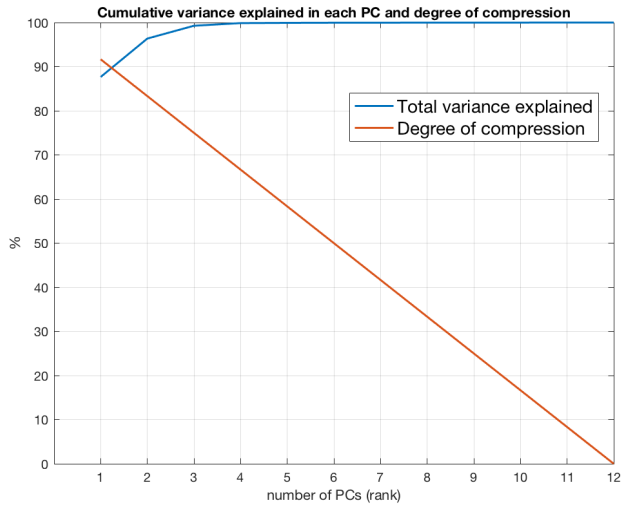


Figure 5.3: Percentage of variance explained in each PC and degree of compression of the data

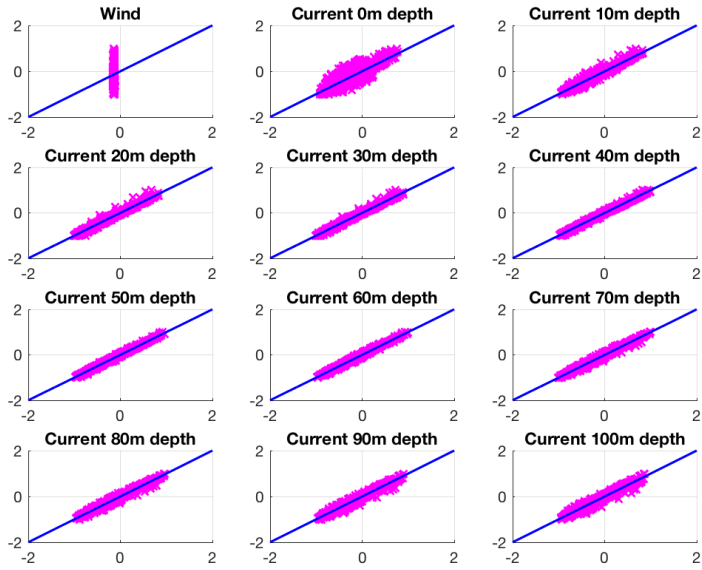
### 5.1.3 Reconstruction

For a dataset such as  $\mathbf{X}$  with twelve variables, there is not much useful information to be gained from showing graphs of the reconstructed timeseries, as the lines are very close to being identical. Furthermore, the variables are hard to distinguish, and many details are lost. Instead, quantile-quantile-plots (qq-plots) of each variable are presented to illustrate how well the reconstructions of the compressed data resembles the original timeseries data. It is shown for two cases, a reconstruction of rank 1 in figure 5.4 and of rank 3 in figure 5.5. Reconstructions of rank 2 and rank 4 can be found in Appendix A1. The x-axis in all following figures represents the reconstruction and the y-axis represents the original timeseries data. The blue line represents equality between the two data structures.

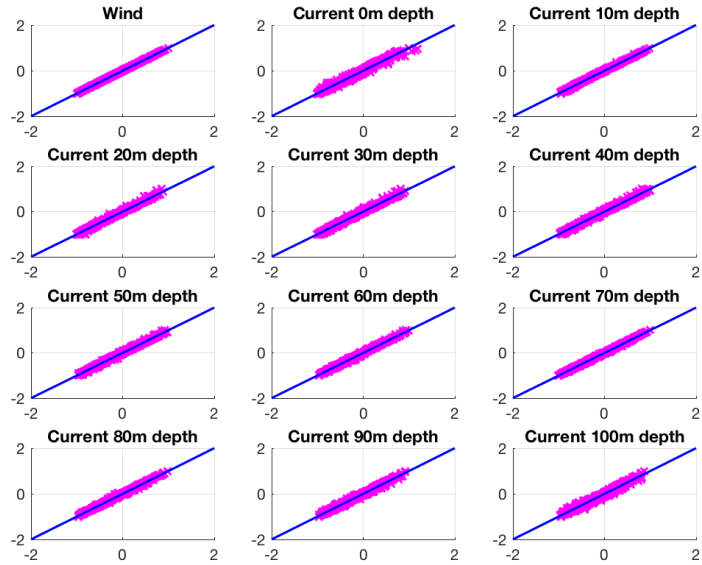
Figure 5.4 shows the reconstruction of rank 1, i. e. reconstructing the data using PC1 alone, for all the variables. As indicated from figure 5.2 and figure 5.3, a rank 1 reconstruction provides a fairly good result in all variables, except for the wind. The wind also stands out in the biplot, making it a compatible result. Because the surface currents have the highest correlation with the wind out of all the variables, its reconstruction is therefore poorer than for the other variables. Still, a lot of the dynamics are well reconstructed.

Figure 9.1 in Appendix A1 shows the reconstruction of rank 2. As expected from figure 5.2, including PC2 significantly improves the representation of the wind, but the surface current representation is only slightly improved. Looking at the biplot, it is clear that the surface currents variance is largely explained in PC3, which is further seen in the reconstruction of rank 3 in figure 5.5. This reconstruction yields good results in all variables, and from figure 5.3 it is seen that close to 99% of the total variance is explained.

It is hard to explain exactly what happens if the rank is further increased, because in a 3D world this cannot be graphically shown in a biplot such as for PC1 to 3 in figure 5.2. But looking at the reconstruction found in Appendix A1, it is seen that it is further improved, although the improvements are marginal. Particularly variance in the 0m-variable is improved. The quality of these results have to be evaluated depending on the application of the analyzed data. Generally, reconstructions of rank 3 and rank 4 include close to 99% of the total variance, with DoF of about 85% and 68%, respectively. Both of these results are examples of significant data amount reductions, and it is expected that ranks with similar qualities can be used in further analyzes.

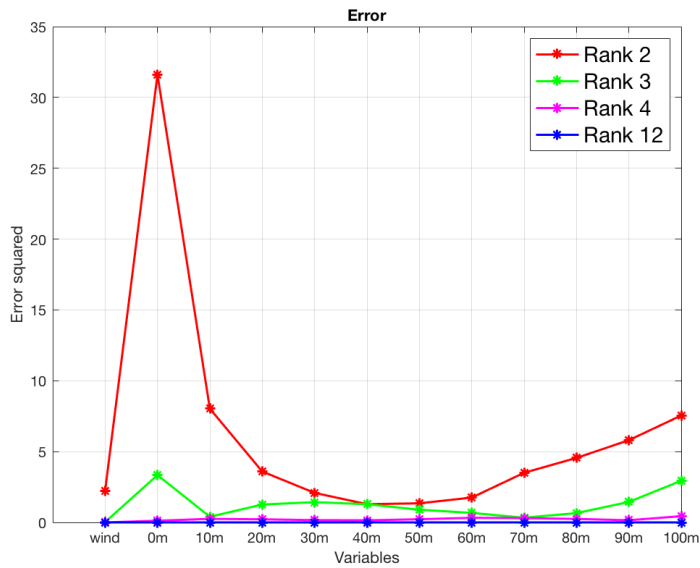


**Figure 5.4:** Similarity between the reconstructed data (x-axis) and the original timeseries data (y-axis) in each variable, for reconstruction of rank 1



**Figure 5.5:** Similarity between the reconstructed data (x-axis) and the original timeseries data (y-axis) in each variable, for reconstruction of rank 3

A final way of interpreting the results are by looking at the squared error as in figure 5.6. Here, rank 1 is omitted because of the high error dominating over the other ranks. For rank 3, the highest error is observed in the 0m-variable, while for rank 4 the error is very low in all variables. That is compatible with the observed improvement when increasing from rank 3 to rank 4 in the reconstructed variables.



**Figure 5.6:** The squared error in each PC in reconstructions of different ranks

## 5.2 Second scenario: Horizontal spatial dimension

In this scenario, the the 2D horizontal spatial dimension is explored by using PCA. The chosen approach is to do the analysis at two depths (0m and 200m), where the  $\mathbf{X}_{245808 \times 5}$  matrix includes timesteps as observations and the horizontal locations as variables. At these two depths, the differences in their respective dynamics are expected to be large. The number of observations are significantly increased and now corresponds to two years. The horizontal PCA is done in two separate analyzes using current strength and current direction as variables. The result show such high similarities between them that the results presented are valid for both wind and current strength and - direction, unless otherwise mentioned. In this case, the interesting information is the variance and correlation between the points and not necessarily the quality of the reconstruction. The data is normalized to prepare for the PCA analysis. However, in this scenario it is changed back to original values before the reconstructed data is graphed as explained in chapter 3.3.2.

### 5.2.1 Variance

The most interesting element of this analysis is to see how the dynamics in the points referred to as origin, North, East, South and West relates to one another. Figure 5.7 and figure 5.8 shows resulting biplots from two PCA analyses at depths of 0m and 200m, respectively. These figures show very similar trends with regards to the distribution of the observations, and the corresponding directions of the variable eigenvectors. The scattered observations shows a clear trend along the PC1-axis, with six surrounding clusters with lower density of points in the PC1-PC3-plane, stretching along the PC2-axis. The eigenvectors of the variables shows an interesting behavior. In the PC1-PC2-plane, all variables except E10 are close to perfectly correlated positively along the PC1-axis. E10 are positively correlated along the PC2 axis, and it appears that the PC2-axis is single handily explaining the variance in E10, while all the other variables are explained in the PC1-PC3-plane. In the PC1-PC3-plane, figure 5.7 shows that at a depth of 0m, the current strength variance of the origin variable has a very similar eigenvector to the N10 variable, being represented in the same quadrant. Furthermore, a similar behavior is seen for the current strength variance of the S10 and W10 variables. However, these are expressed in the positive direction of PC3. In the PC1-PC3-plane, figure 5.8 shows that at a depth of 200m, there is a very similar behavior between the variables.



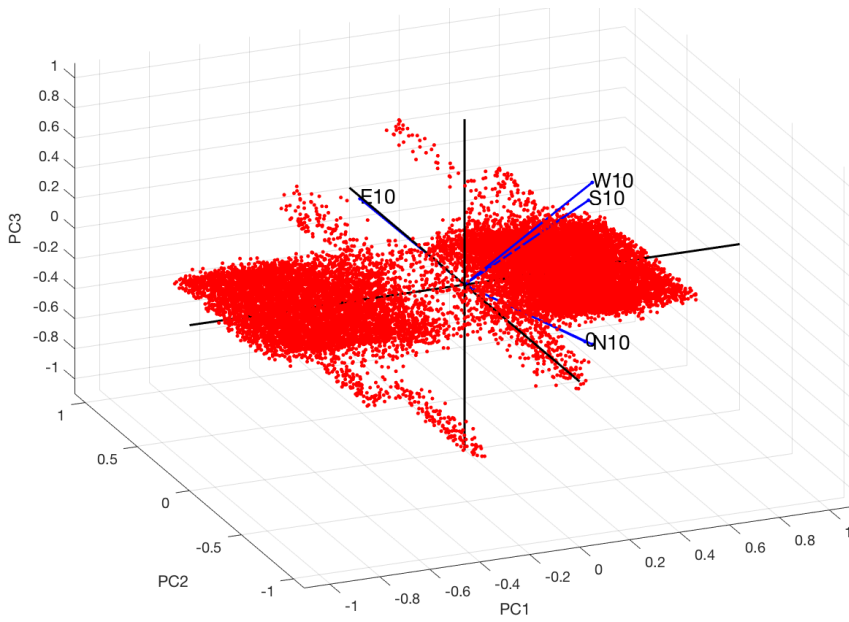


Figure 5.7: Biplot of X at 0m depth

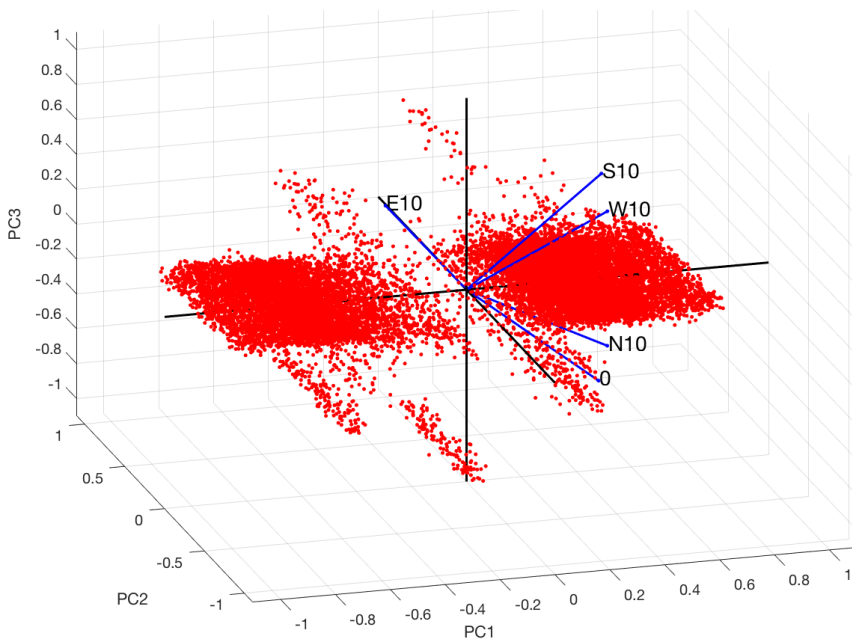
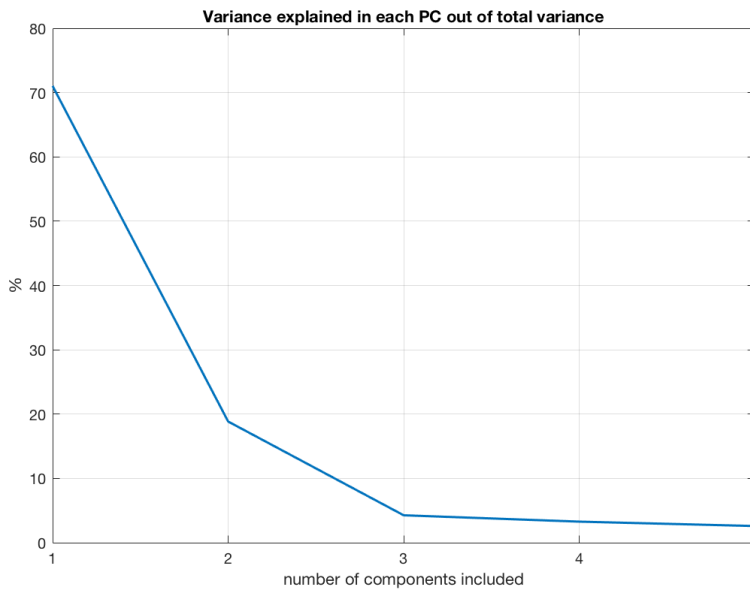


Figure 5.8: Biplot of X at 200m depth

In general, the results show that the origin and the N10 variables, and the S10 and W10 variables have a very similar variance between the pairs. The E10 variable separates itself entirely from the others, being represented by PC2.

Figure 5.9 shows how much of the total variance is explained in each PC. There is an indistinguishable difference of less than 1% in either variable, between the two water depths of 0m and 200m. Therefore, only one graph is shown in the figure, representing both water depths. The figure shows that a high percentage of the total system variance is expressed in PC1 of 70%. PC2, which the biplots indicated that included close to all variance in the E10 point, includes 20% of the total variance, which seems reasonable as there are five variables. The following PCs includes less than 5% of the total variance, indicating that the PC3 component of the biplot does not include high values.

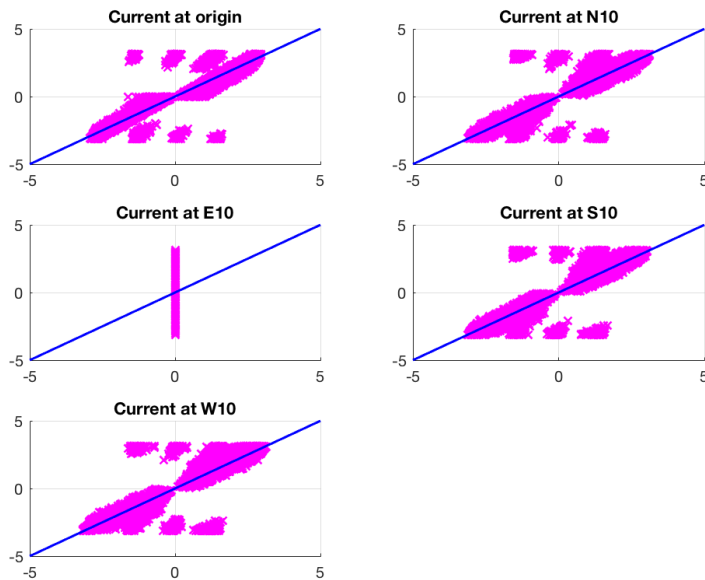


**Figure 5.9:** Percentage of variance explained in each PC for current in the horizontal plane

## 5.2.2 Correlation and error

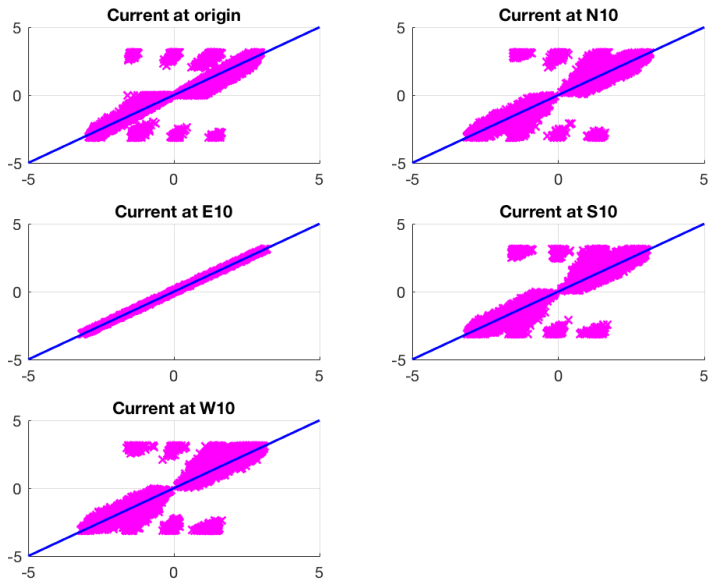
In figure 5.10 and figure 5.11, the reduced reconstructed data (x-axis) are shown with the original data (y-axis) of the five spatial variables. The biplot shows that PC2 completely describes the variable E10, which is shown in the reconstruction as the E10 variable is completely restored when expanding the reconstruction to rank 2. This indicates that the variable E10 has the most distinct variance out of all

the points. The other variables have a slightly more complicated composition, but follows the description given in the last section. Reconstructions of rank 3 and rank 4 can be found in Appendix A2. They show that the reconstructions are gradually improving as the rank of the analysis is increased. However, the reconstructions does not reach complete reconstruction in all variables before the full rank 5.

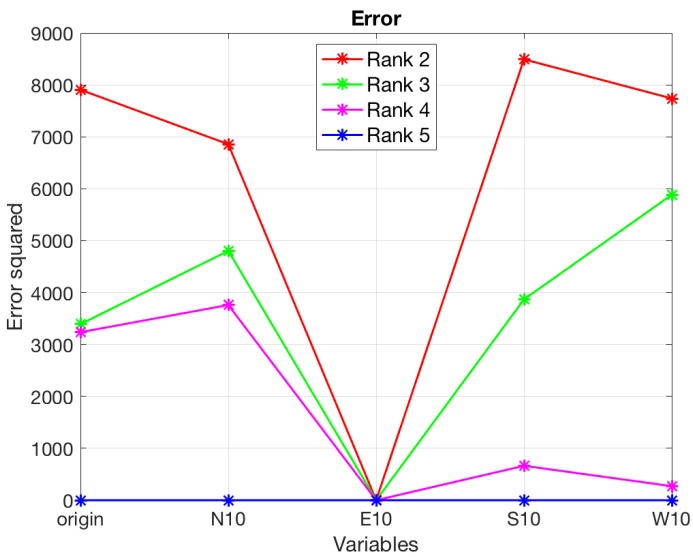


**Figure 5.10:** Reconstructed data (x-axis) and original data (y-axis) in each variable, for reconstruction of rank 1

The error is shown in figure 5.12 for rank 2 to rank 5. Error for rank 1 is excluded because it dominates the picture as it follows a similar form as the other graphs, but with a much larger value. Corresponding error plot for 200m depth shows a very similar behavior with only small deviations, and errors in the same magnitude. Therefore, the shown plot represents the error for both depths well. The error in E10 is very small compared to the rest of the variables, likely due to the variable being close to entirely explained in PC2, but it not zero as it appears to be in the plot.



**Figure 5.11:** Reconstructed data (x-axis) and original data (y-axis) in each variable, for reconstruction of rank 2



**Figure 5.12:** The squared error in each PC in reconstructions of different ranks for om depth

# Partial Least Squares Regression

Partial Least Squares Regression (PLSR) are used as a tool in three analyzes, including two different types of regression techniques, estimation and prediction. The first sections includes results from estimating deeper currents based on wind and surface current. The first estimation is done on a relatively small dataset, and will be compared with GPR later. The second section is based on a similar analysis, but with a much larger dataset. The last section includes results from forecasting future current based on historic data.

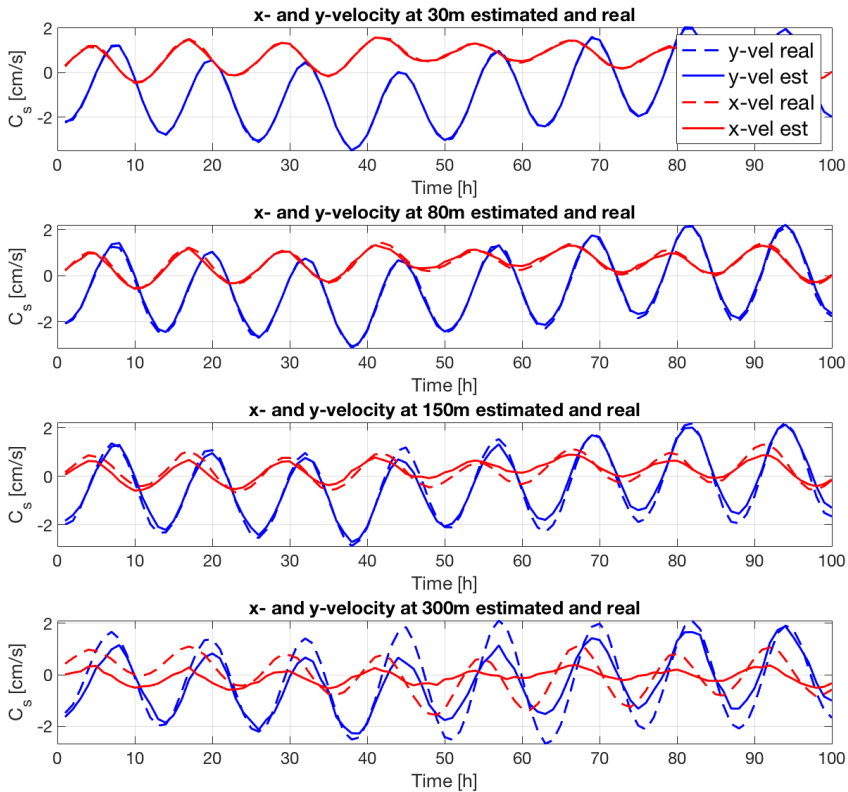
## **6.1 Estimating deeper currents based on surface currents**

Estimation in this work refers to performing regression on data where wind and top currents are used for training, and then the deeper currents - at the same time interval - are used as response variables. When selecting data, the succeeding data in all analyzes are chosen as a separate and equivalent dataset in this work, and used as a testing set to validate the estimation. The difference between the estimated and the real values are referred to as error. PLSR analyzes includes different ranks, in a very similar manner as for PCA. The rank largely influences the results, and are addressed in all PLSR analyzes.

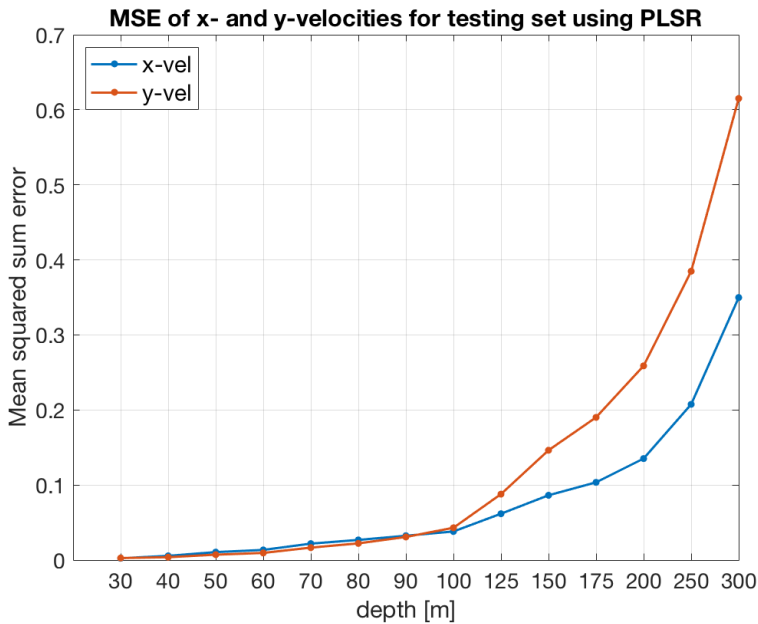
### **6.1.1 Six weeks of training data**

The first PLSR includes 1000 observations in the analysis, which is equivalent with about 6 weeks in time. 28 variables are included, they are x- and y-velocity components through the water column from 30m as seen in section 3.5. Figure 6.1 shows how the regression performs as compared with the actual data. An example of 100 observations are chosen from the results and presented, in order to be able

to see the difference. It is seen that for the uppermost currents, the estimation works quite well, and then the quality gets increasingly poorer down through the water column. In this selection of data, the x-velocity component is very poorly estimated in particular towards the bottom of the water column. This is further seen in a error plot of the summed MSE for all the variables in figure 6.2. It shows that the estimation works well for the first 100m of the water column, and then the error increases for deeper currents. From 200-300m depth, the error is quite large and increases considerable more rapidly than in the top part of the water column.



**Figure 6.1:** Timeseries the first 100 hours of x- and y-velocity for current at 30m depth, 80m, 150m and 300m. the reconstruction is done using full rank 32.

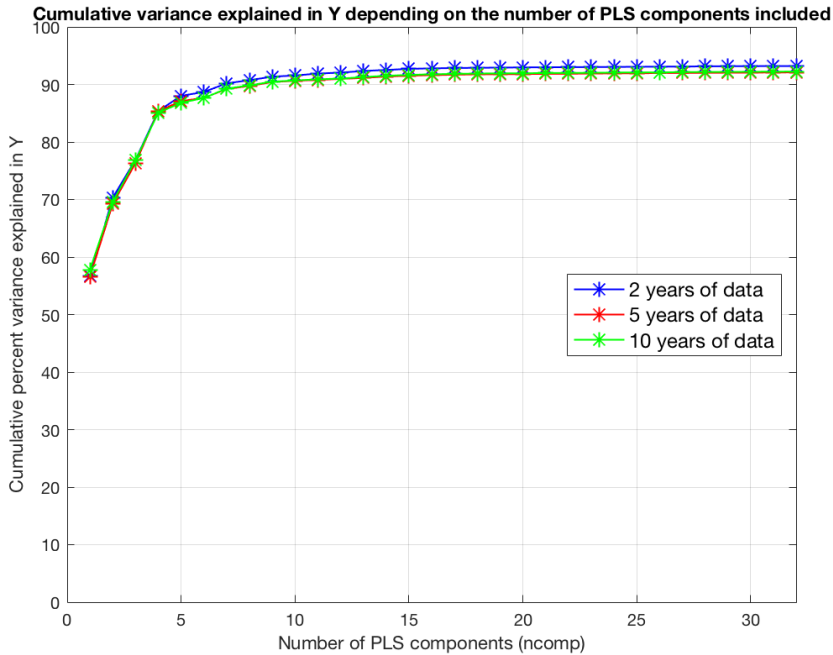


**Figure 6.2:** MSE in each variable for x- and y-velocity

### 6.1.2 2, 5 and 10 years of training data

The second PLSR includes three analyzes done for amounts of training data equivalent with 2, 5 and 10 years. The intention behind these analyzes are to check the influence from the amount of training data. For the estimation, three different amounts of data are used. What data are included are shown in table 3.1.

As in PCA, the rank of the analysis has to be specified in PLSR. This will influence the quality of the analysis in a similar manner as for PCA, as each component includes a part of the system variance. In figure 6.3, the cumulative variance explained are plotted for all three amounts of training data. It does not differ much for different amounts of training data. Based on this figure, it is chosen that the minimum rank is 4, and the full rank is the 32 components model. Analyzes are done for both cases, to compare. As opposed to when using PCA, the full analysis at rank 32 does not explain 100% of the variance because, as explained in section 4.2, PLSR strives to emulate maximum variance in a set of specified predictor variables, not only within the training dataset. This means that the figure refers to how much of the total variance in the predictor variables are explained for each rank.



**Figure 6.3:** Variance explained with different amount of data, depending on number of PLS components included.

### Timeseries for 2 years of training data

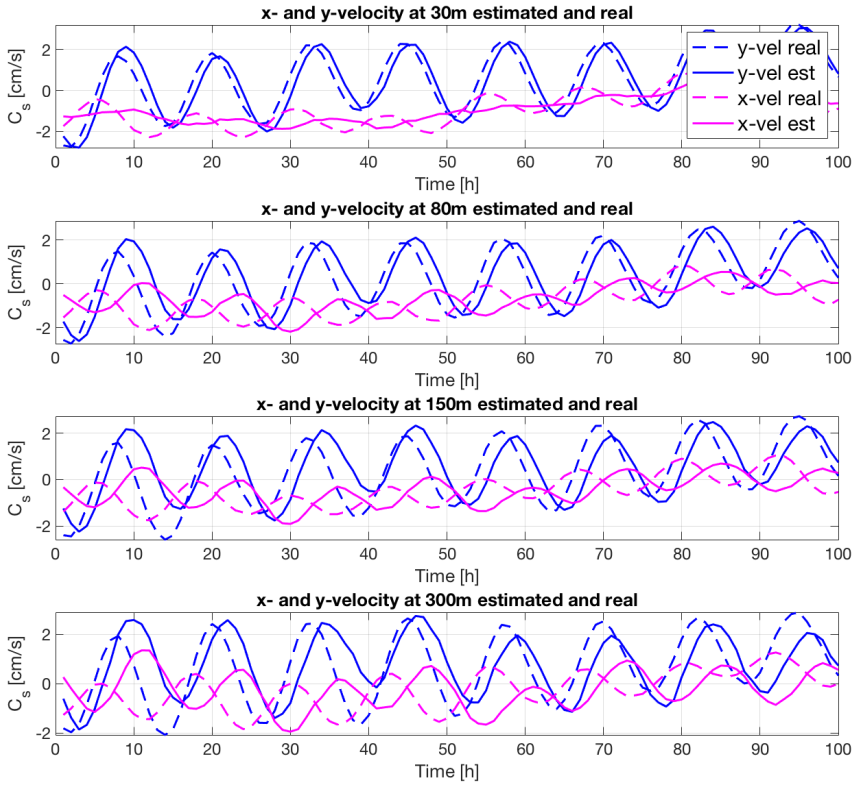
Figure 6.4 shows estimated and real current for an interval of the resulting time-series from an analysis of training data equivalent of 2 years, when using the full 32 components (rank 32). A few water depths distributed evenly through the water column are shown. As seen earlier, the estimation yields better results in the upper part of the water column than the lower part. The y-velocity component has large oscillation along the timeseries, which are somewhat well estimated by the algorithm, particularly in top currents. As the water depth increase, the y-velocity estimate gets progressively poorer, and a phase-shift between the estimate and real value gets progressively larger. In the x-velocity component, the estimate and the real value are mostly in anti-phase with each other along the timeseries. Furthermore, the oscillations in x-velocity has a much lower amplitude than y-velocity, and therefore its error has a lower magnitude, although the oscillations of the estimate and the real value are in anti-phase with each other. In figure 6.5, the same is shown when using 4 components. Here, it is seen that offsets are significantly larger and occurring more frequently. However, for the x-velocity component, anti-phase is not observed in the top currents any more. Moreover, while the esti-



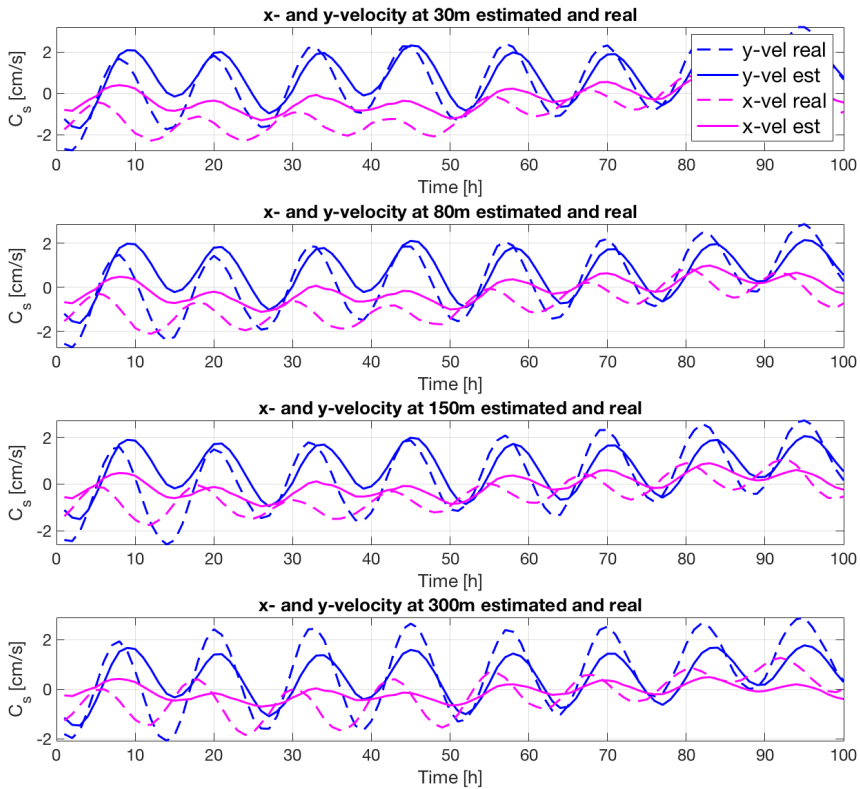
---

## 6.1 Estimating deeper currents based on surface currents

mate is not oscillating with the same amplitude as the real data in deeper currents, it follows the trend well.



**Figure 6.4:** Timeseries of estimated current and real current using 2 years of testing data. All 32 components are used in the reconstruction (rank 32).

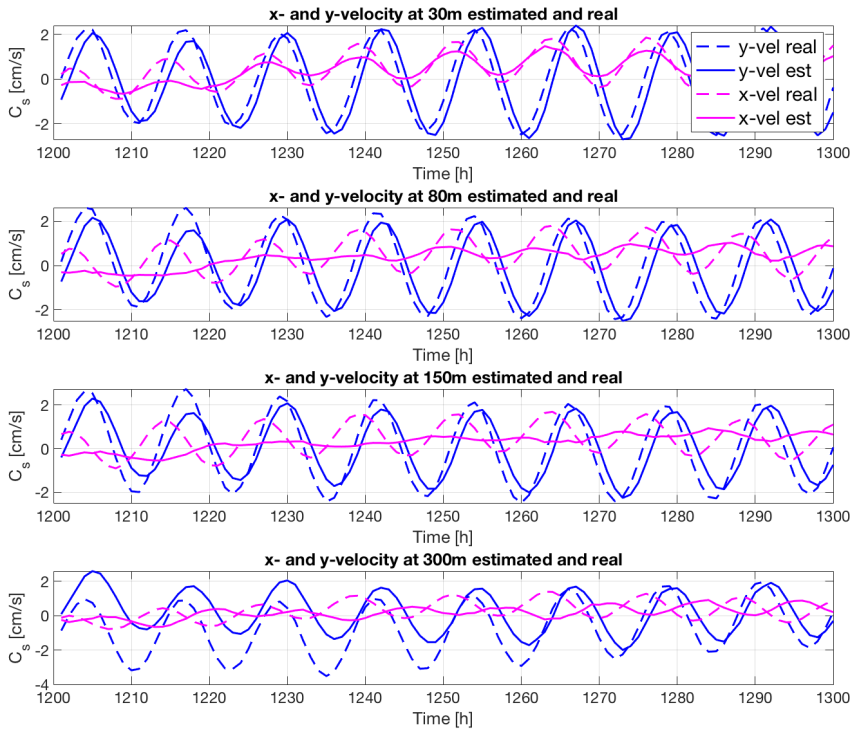


**Figure 6.5:** Timeseries of estimated current and real current using 2 years of testing data. The rank is at its minimum in this reconstruction (rank 4).

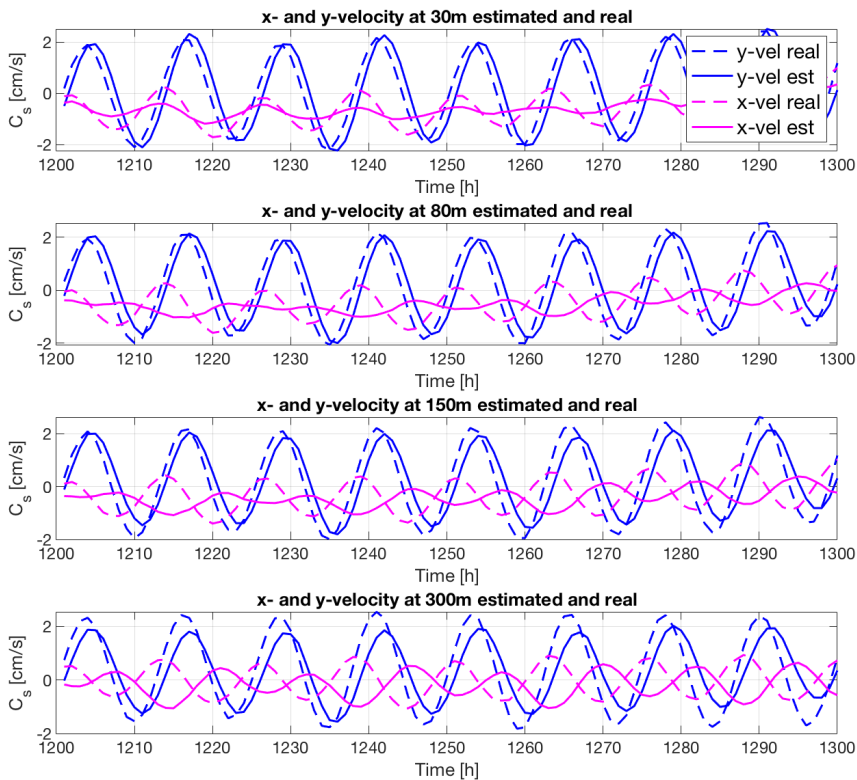
### Timeseries for 5 years of training data

Figure 6.6 shows an equivalent timeseries using 5 years of training data and the full rank 32. Once again, the estimation is better for upper currents than for currents in the lower part of the water column. The estimated variables does also seem to fit better with the real data, although there are still significant errors. This is also seen in figure 6.7, where similar relations between the minimal and full ranks as when using 2 years of data are observed. The estimation generally fits quite well for 5 years of training data also. As the rank is decreased, the error in the estimated variables increases, which is more distinct in larger water depths than closer to the surface. Moreover, in the bottom currents, the x-velocity oscillations are in anti-phase at both ranks.

## 6.1 Estimating deeper currents based on surface currents



**Figure 6.6:** Timeseries of rank 32 of estimated current and real current from testing set with 5 years of training data.

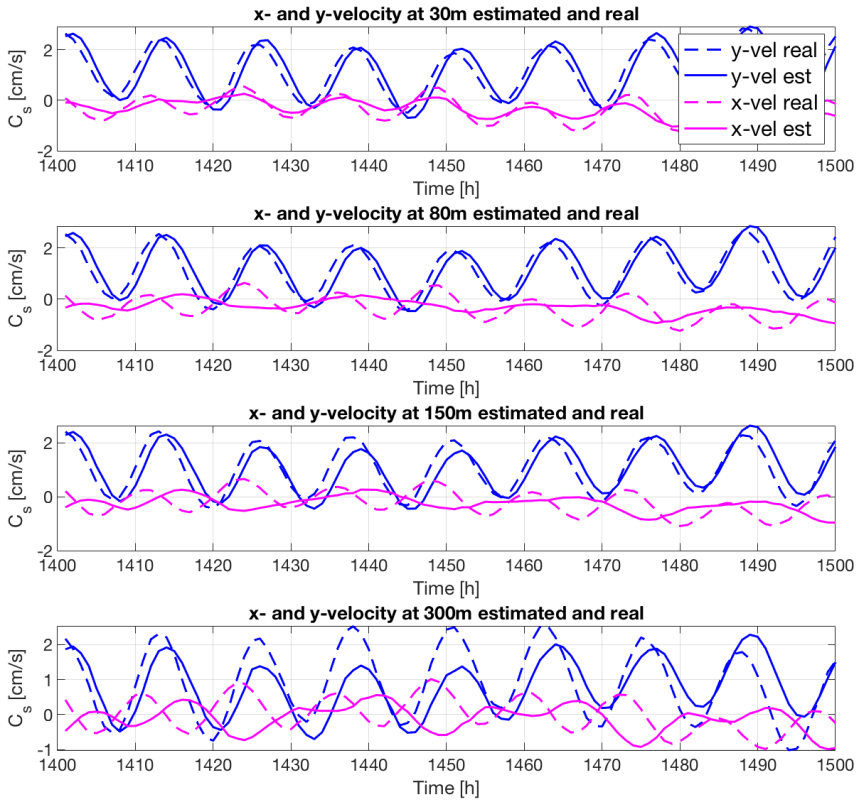


**Figure 6.7:** Timeseries of rank 4 of estimated current and real current from testing set with 5 years of training data.

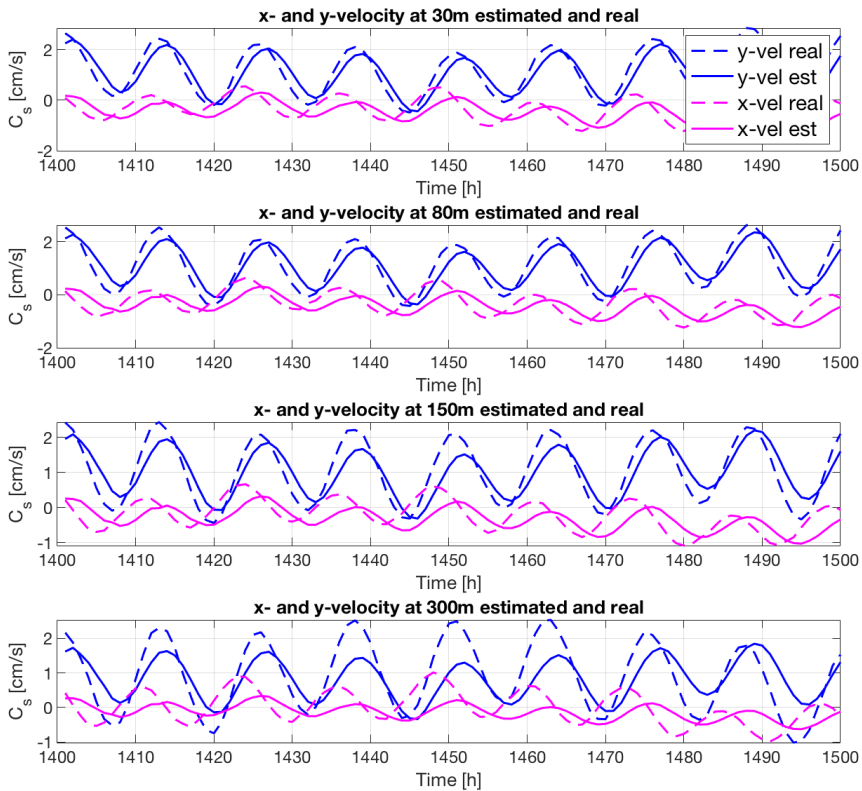
### Timeseries for 10 years of training data

Also when using 10 years of training data, the current in the top part of the water column is better estimated than deeper, as seen in figure 6.4. In this rank 32 reconstruction, the current is quite well predicted in the upper part of the water column. The estimation is not perfect, and still contains a certain error, particularly in the x-velocity component. At 300m, the estimation is considered quite poor. There are large deviations between estimated and real values in both y-velocity and x-velocity, which once again are in anti-phase. The lower rank reconstruction timeseries in figure 6.7 also have large errors, but this model performs quite well, as it is quite similar to the full-rank model in many variables.

## 6.1 Estimating deeper currents based on surface currents



**Figure 6.8:** Timeseries of rank 32 of estimated current and real current from testing set with 10 years of training data.

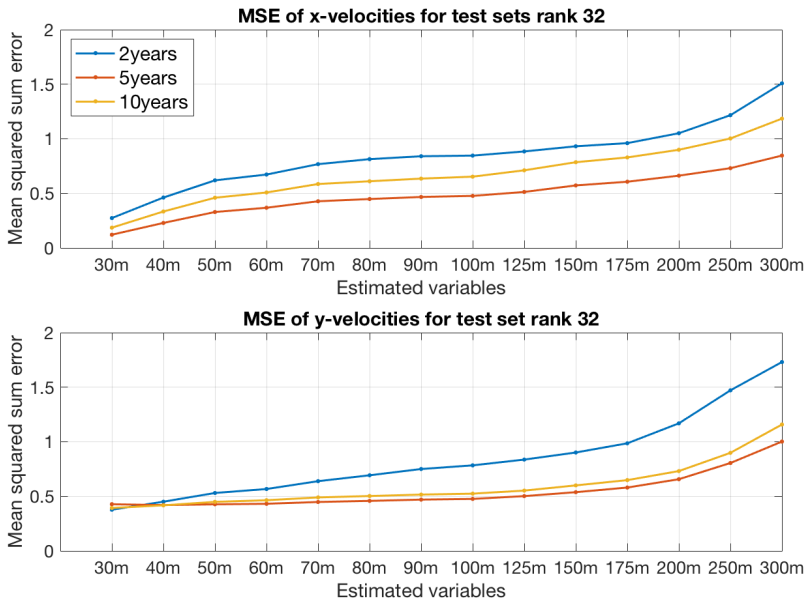


**Figure 6.9:** Timeseries of rank 4 of estimated current and real current from testing set with 10 years of training data.

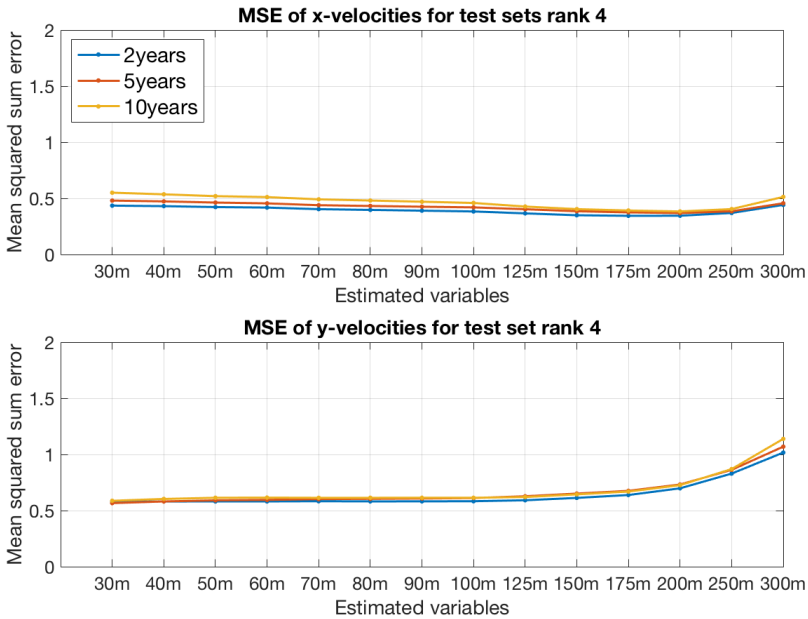
Error between the estimated values and the real current are further explored by using MSE in figure 6.10, where the estimation of rank 32 is seen, and figure 6.11 where rank 4 is seen. Figure 6.10 show a trend such as the one expected, where the error increases through the water column. Furthermore, the estimation improves when larger datasets are explored, both in x- and y-velocity. However, figure 6.11 show an error that is close to constant, in both velocity components.

In Appendix B.1, figures of the reconstructed currents in polar coordinates can be found, together with directional data. They provide an alternative representation of the results from the PLSR analysis, as well as an example for trained and tested results.

## 6.1 Estimating deeper currents based on surface currents



**Figure 6.10:** Mean Squared (sum) Error [cm/s] for testing set at rank 32.

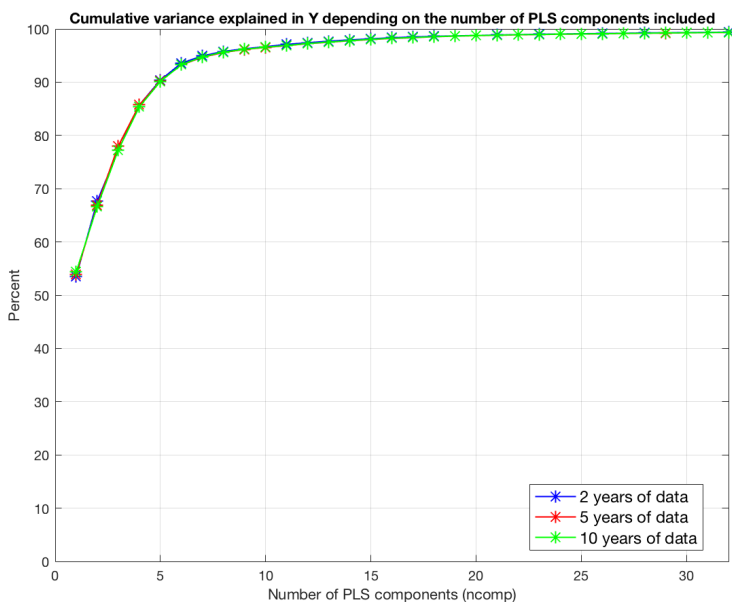


**Figure 6.11:** Mean Squared (sum) Error [cm/s] for testing set at rank 4.

## 6.2 Forecasting current based on historic water column measurements

The full 2, 5 and 10 years of datasets are used in this part of the analyzes. The main difference between estimation and forecasting is that in forecasting the next time step is predicted. Now both the training and the testing data includes measurements from the whole water column, resulting in a much more demanding analysis computationally, than when doing estimation. As in the estimation problem, the testing set is chosen as the data right after the training set. Moreover, results using different ranks, including minimal and full rank are explored.

From figure 6.12, the rank of the PLSR for forecasting is seen. In this case, the full rank is still 32, and the new minimal rank is determined as 5, where about 90% of the total variance is explained. From the figure, it is seen that at rank 32, very close to 100% of the total variance is explained. This indicates that forecasting is better suited for PLSR than estimation is. As a high percentage of total variance explained is reached at a lower rank than 32, this is also reconstructed. Rank 18 is determined to be the rank where the model is very close to as good as it becomes, and reconstructions at this rank are shown for the three amounts of training data.

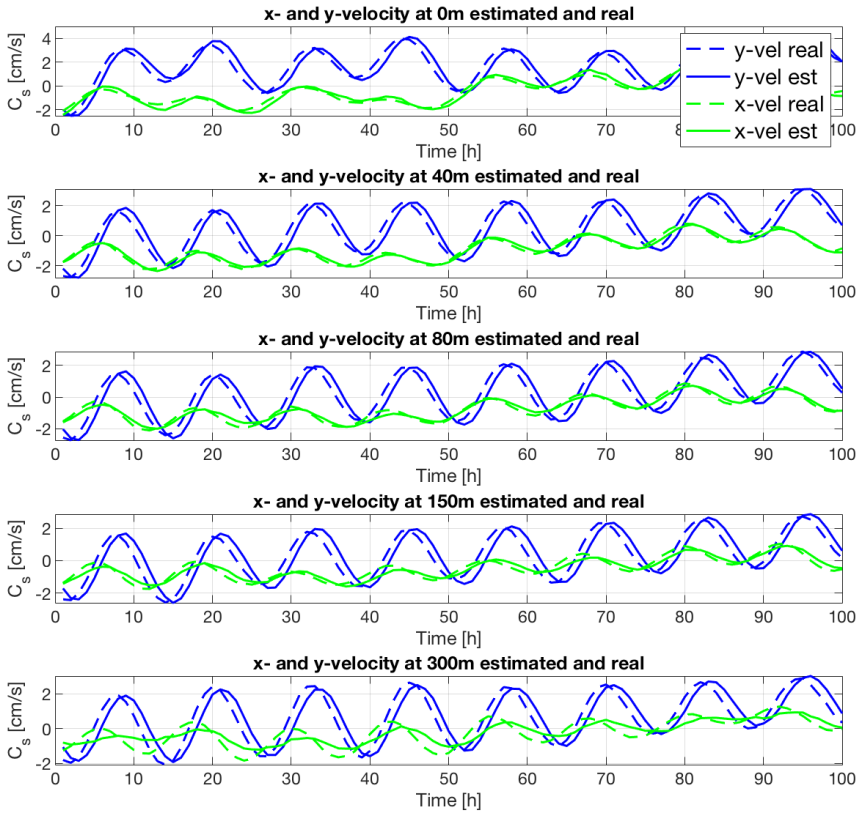


**Figure 6.12:** Variance explained in forecasting depending on number of PLS components included. Graphs for all three training data amounts are shown.



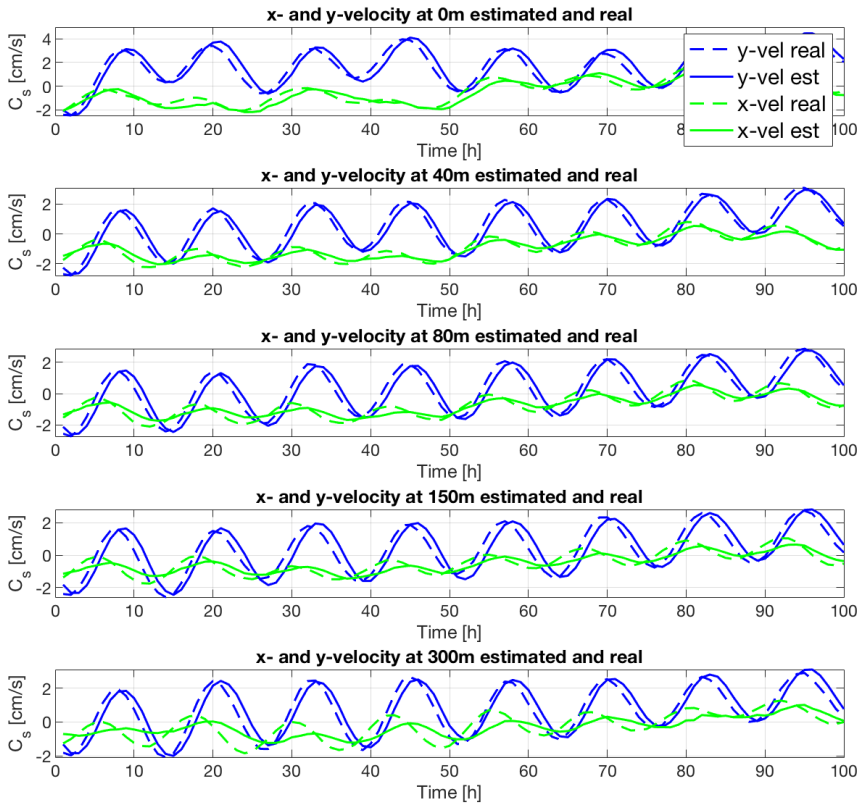
### **Timeseries forecasting with 2 years of training data**

Figure 6.13 and figure 6.14 show reconstructions of rank 32 and 18, respectively, for 2 years of training data. From the two representations, it is clear that the model is well forecasted, and gives a well reconstructed current. The two ranks give quite similar results, as indicated from figure 6.12, but there are a few marginal differences. The y-velocity component is well predicted at both ranks, and gives almost identical results. This component is also responsible for the highest variance along the timeseries, and is therefore most likely highly represented in the first PLS components. Some differences between the ranks are seen in the x-velocity component. The forecasted values does not follow the oscillations of the real data as well at rank 18. Still, they are quite similar. These oscillations have a quite low amplitude, which means that most likely the deeper currents dynamics are lost when the model is reduced, thus the most significant difference between these ranks are in the low-variance velocity components. Also at rank 5, shown in figure 6.15 the currents in the upper part of the water column gives better results for both x- and y-velocity components. However, there are larger error between forecasted and real values for both x- and y-velocity components. Moreover, in deeper currents, the errors are quite large.

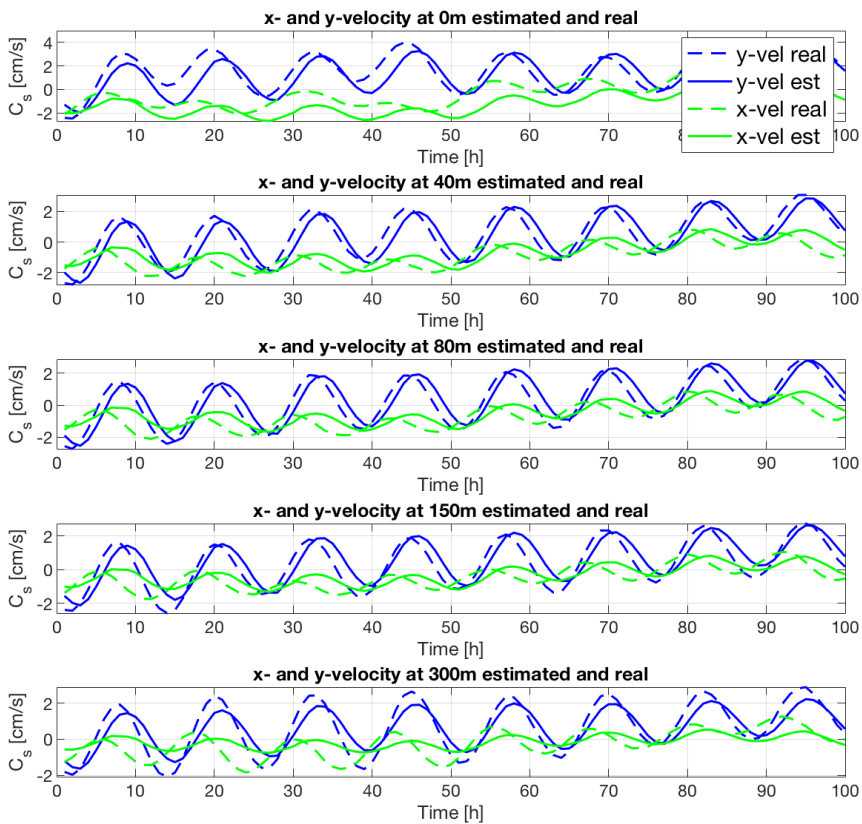


**Figure 6.13:** Timeseries of rank 32 of forecasted current and real current from testing set with 2 years of training data.

## 6.2 Forecasting current based on historic water column measurements



**Figure 6.14:** Timeseries of rank 18 of forecasted current and real current from testing set with 2 years of training data.

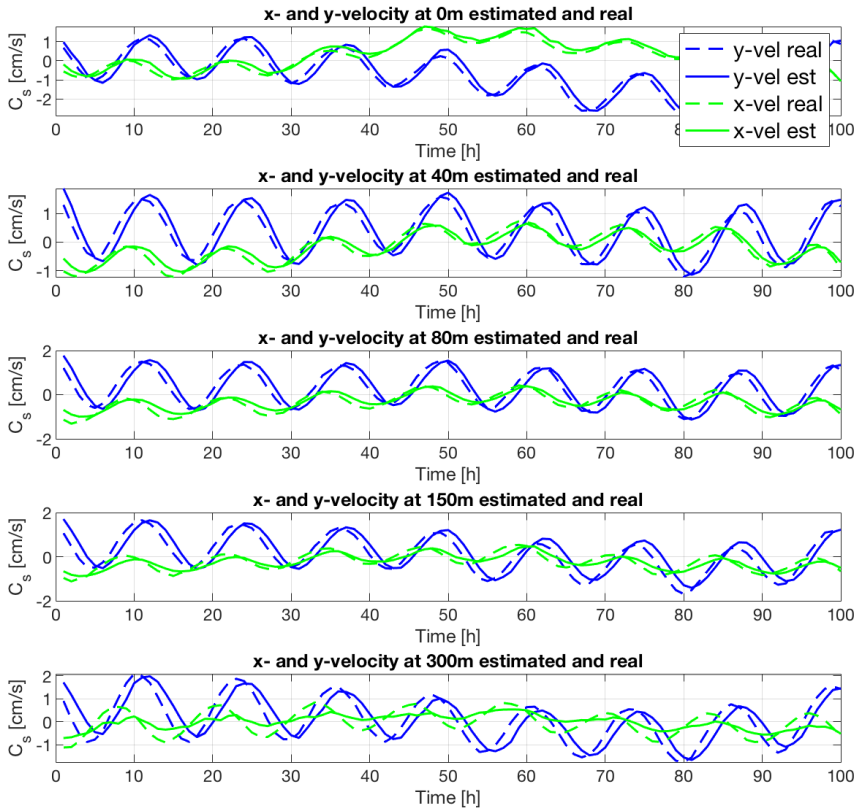


**Figure 6.15:** Timeseries of rank 5 of forecasted current and real current from testing set with 2 years of training data.

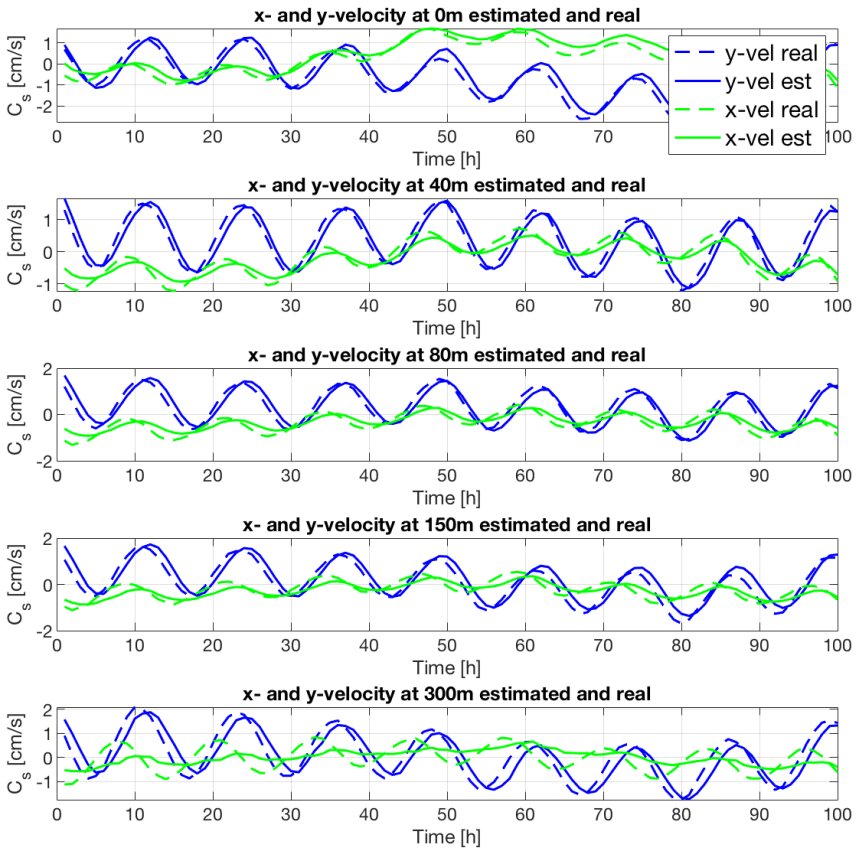
### Timeseries forecasting with 5 years of training data

Also when using 5 years of training data, forecasting using rank 32 and rank 18, as seen in figure 6.16 and figure 6.17, give quite similar quality results. However, some differences can be observed also here by through inspection. For 5 years of training data, the difference is distributed in all depths and in both  $x$ - and  $y$ -velocity components, rather than being very obvious in some depths or velocity components. At rank 5, shown in figure 6.18, the current seems to be well predicted down to a depth of approximately 100m. Current that lies deeper than this give large errors in both  $x$ - and  $y$ -velocity components.

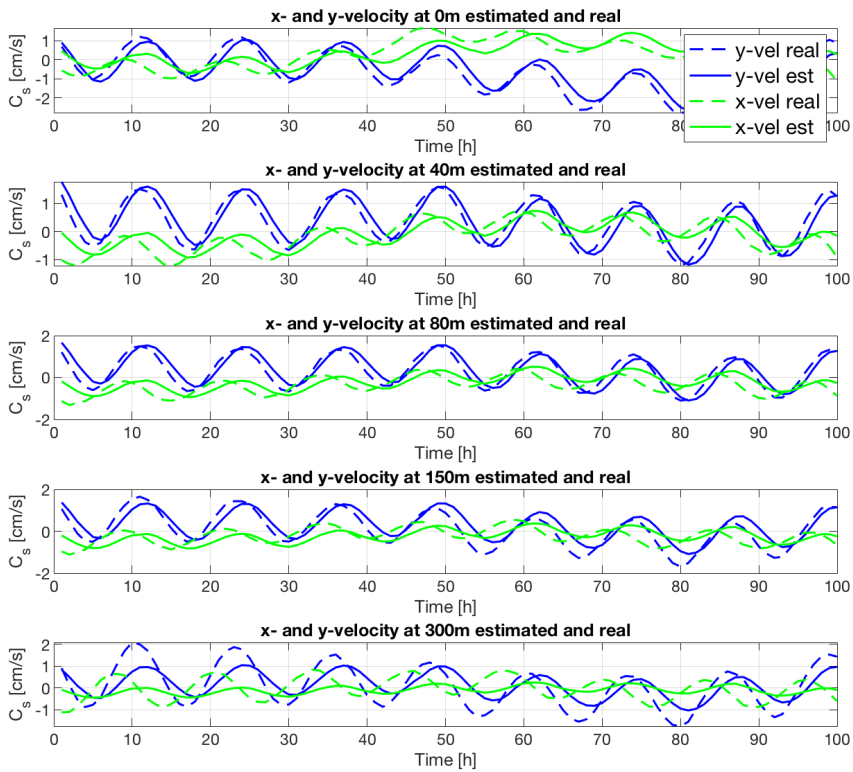
## 6.2 Forecasting current based on historic water column measurements



**Figure 6.16:** timeseries of rank 32 of forecasted current and real current from testing set with 5 years of training data.



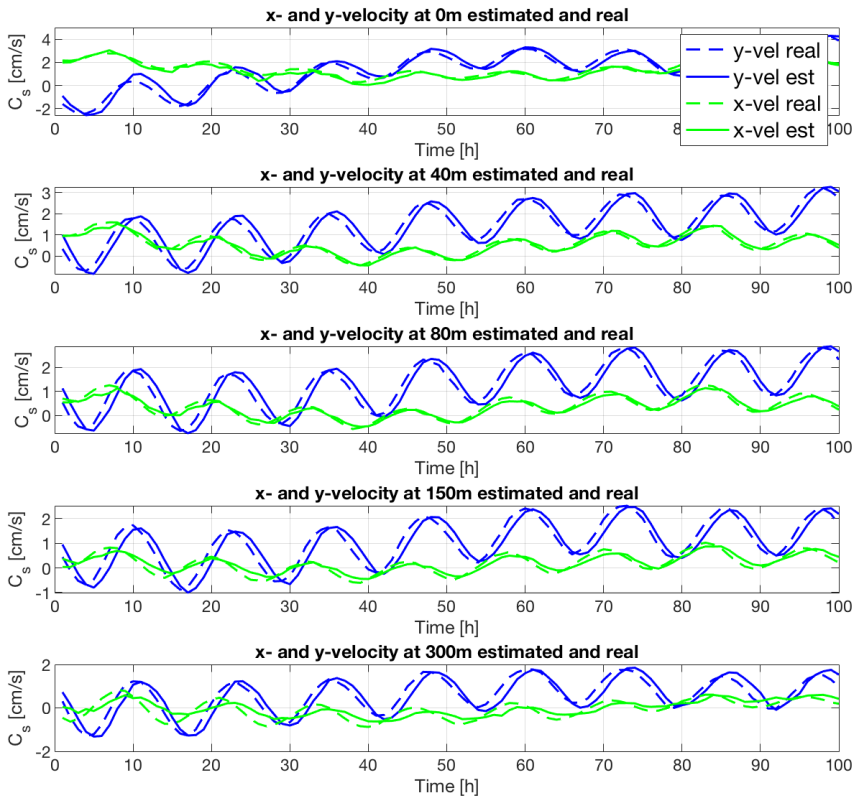
**Figure 6.17:** timeseries of rank 18 of forecasted current and real current from testing set with 5 years of training data.



**Figure 6.18:** timeseries of rank 5 of forecasted current and real current from testing set with 5 years of training data.

### Timeseries forecasting with 10 years of training data

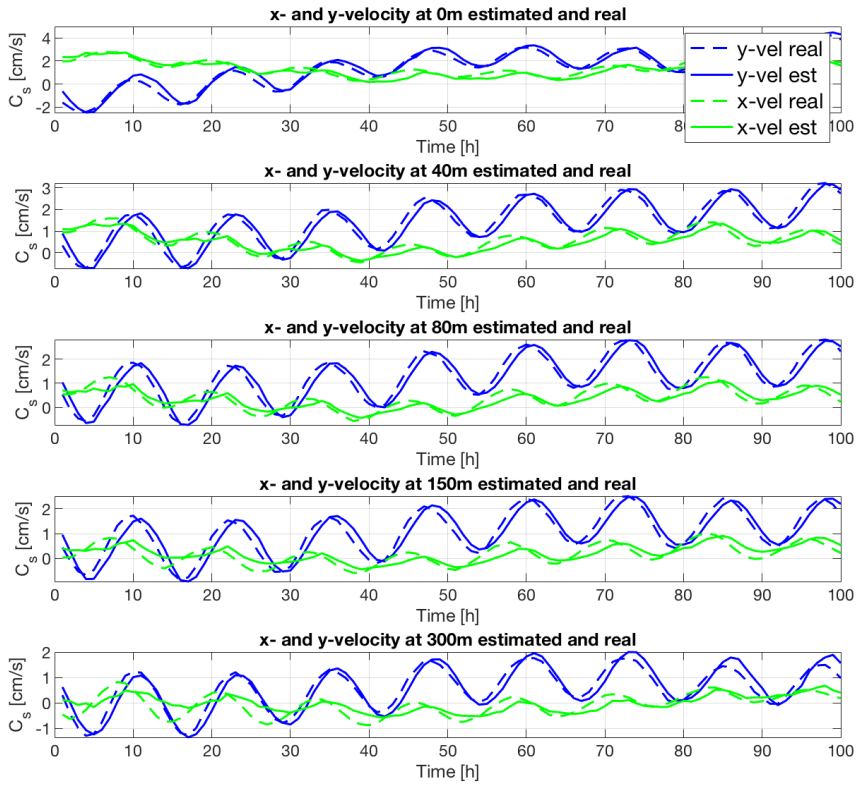
Figure 6.19 shows the forecasted full rank 32 model from the analysis. It performs very well for these data, and gives an almost exact reconstruction in most of the water column. In the deeper part of the water column, the prediction is somewhat poorer, but still the best prediction for deep currents out of all observed timeseries. By looking at figure 6.12, it is seen that from about rank 18, there is almost no change in total variance explained. Therefore, results from rank 18 is shown in figure 6.20. Also at rank 18, the forecasted variables are very well predicted, with only slightly larger error in some parts of the timeseries in x-velocity component through the entire water column. Furthermore, the minimal rank is shown in figure 6.21. For this minimal rank, the x-velocity component is quite poorly forecasted through the entire water column, and the y-velocity component gives a relatively large error from a depth of about 80m to the bottom measurement.



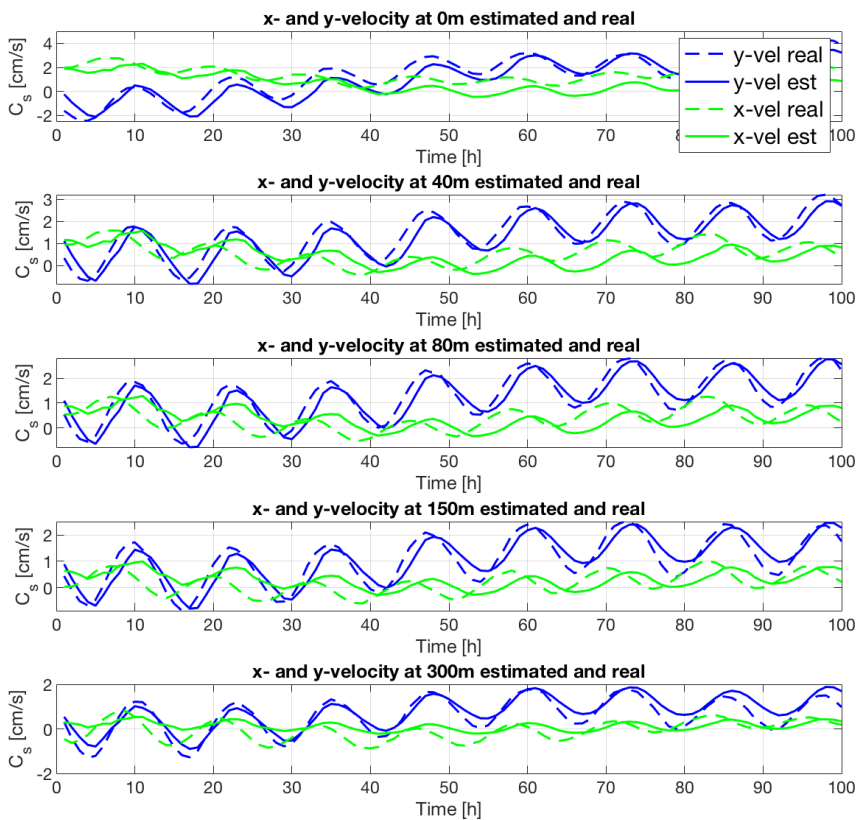
**Figure 6.19:** Timeseries of rank 32 of forecasted current and real current from testing set with 10 years of training data.



## 6.2 Forecasting current based on historic water column measurements



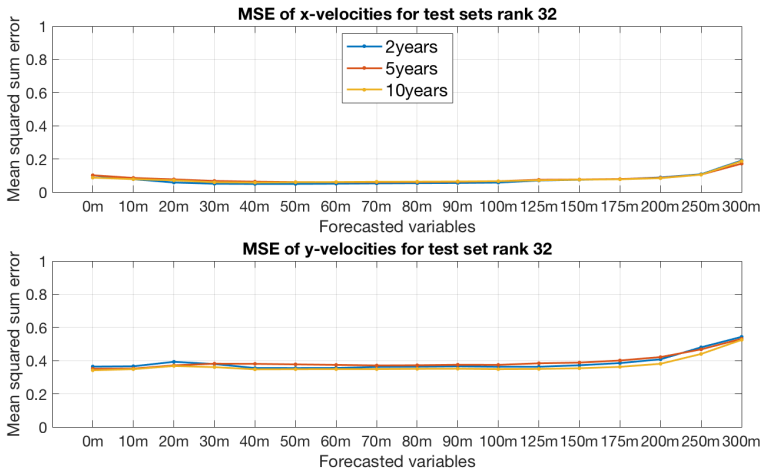
**Figure 6.20:** Timeseries of rank 18 of forecasted current and real current from testing set with 10 years of training data.



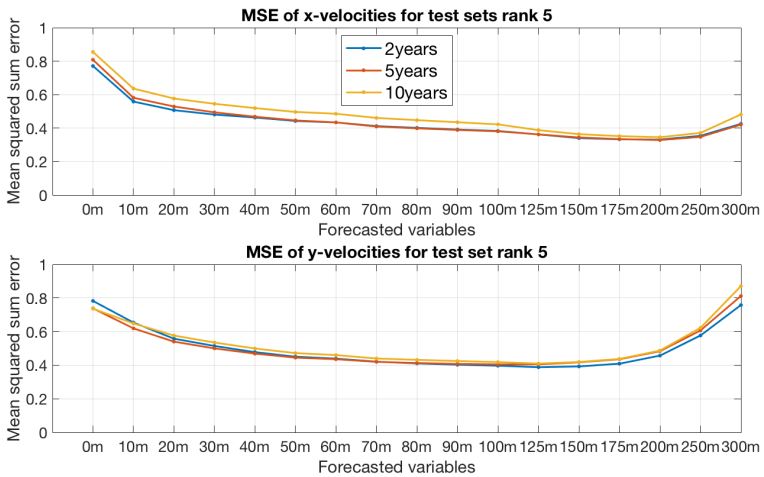
**Figure 6.21:** Timeseries of rank 5 of forecasted current and real current from testing set with 10 years of training data.

As seen in the reconstructed timeseries for the different amounts of training data, the error is quite low in total. Error from rank 32 is seen in figure 6.22, and the error in all variables are very low in x-velocity component for all amounts of training data. For y-velocity component, the error is higher, but it is still considered quite low. It is also seen that MSE slightly increase in the surface and bottom currents for x-velocity component, and in bottom currents for y-velocity. The error when rank 5 is used generally gives a higher MSE, as seen in figure 6.23. MSE in the x-velocity component is considerably higher in all variables, and much higher in surface and bottom currents. Also MSE for y-velocity component is higher for rank 5 than for the full rank 32 in surface and bottom current. The middle part of the water column, from about 50m to 175m, the error is approximately the same as for full rank. The magnitude of the error does not vary particularly much between the different amounts of training data, throughout the forecasting analysis using PLSR.

## 6.2 Forecasting current based on historic water column measurements



**Figure 6.22:** Mean Squared (sum) Error [cm/s] for testing set at rank 32.



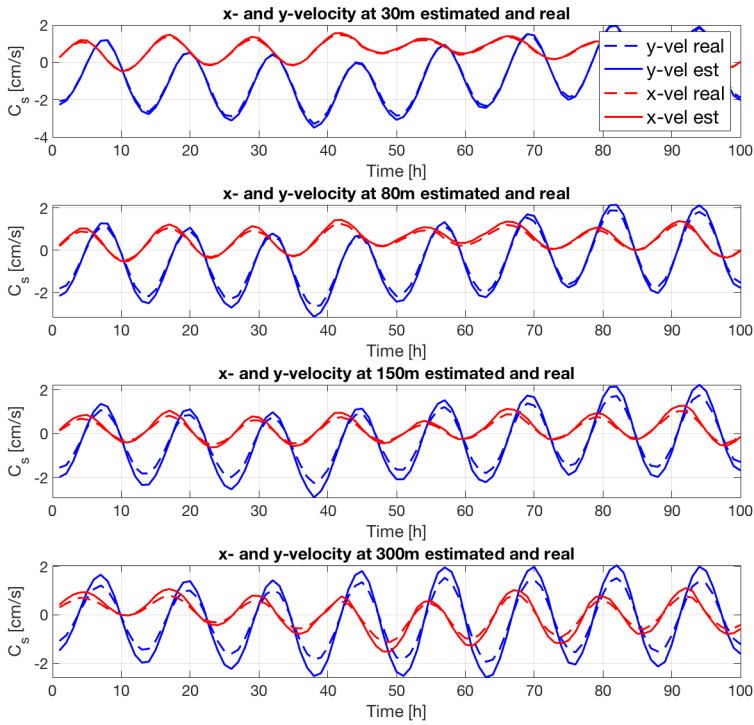
**Figure 6.23:** Mean Squared (sum) Error [cm/s] for testing set at rank 5.



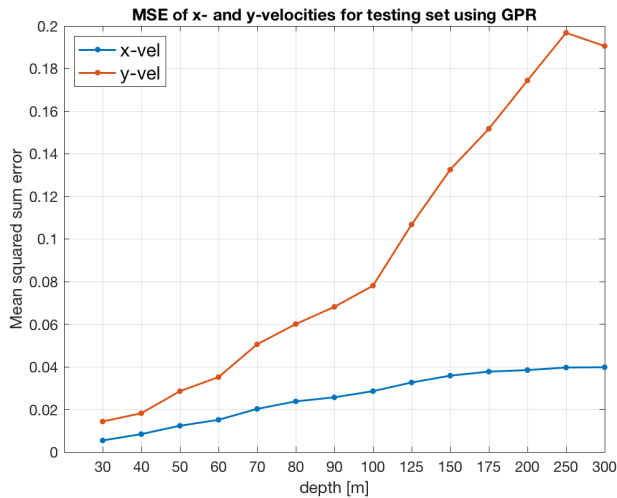
## Gaussian Process Regression

Gaussian Process Regression is another method applied to the data. This method is distinctly different from PCA and PLSR in many ways. It is not linear, but rather a parametric method that lets the data speak more for themselves. GPR is applied to exactly the same amount of data as the first PLSR, including 1000 points for training and the same amount for testing. This is done in order to compare the two methods. It is chosen to apply GPR to the estimation problem to try to obtain an improved regression. Moreover, by choosing estimation, computation time is limited as compared with forecasting as it includes less data in training the algorithm.

Figure 7.1 shows how well the estimation performs as compared with the real values, through the water column. Four depths distributed almost evenly through the water column are chosen, and the first 100 hours are shown in the timeseries. As seen, the method performs well for all depths in this case. Still, there is a noticeable progressively increasing difference between estimated and real current as the estimation is better for top currents than deeper currents. This can be further seen in the summed MSE for each variable in figure 7.2. All over, the error is fairly low. For the x-velocity component, there is a slight increase in error through the water column, but it is only moderately larger in the deeper part of the water column than in the top. The y-velocity component has a very different behavior, where the error increases much more and much more rapidly than for the x-velocity. At 250-300m, the error in y-velocity is almost four times larger than the error in x-velocity. Still, the error is generally low.



**Figure 7.1:** Timeseries the first 100 hours of x- and y-velocity for current at 30m depth, 80m, 150m and 300m.



**Figure 7.2:** MSE in each variable for x- and y-velocity

# Discussion

## 8.1 PCA

### First scenario

In the first scenario using PCA, a dataset of 12 variables through the water column down to 100m are used. This high resolution of measurement points are important because there are a lot of different phenomena causing the current strength in the ocean to vary, as explained in chapter 2. Thus, it is preferred to include as much of the dynamics as possible. At the same time there is high multi-colinearity in current strength data from the ocean, as seen in figure 5.1. This is because the water particles influence one another interchangeably. In larger depths, the measurements are sampled less frequently, because these waters are generally less subjected to mixing in the ocean, and thus the ocean current varies less. What PCA essentially does is to exploit the multi-colinearity of the data and re-arrange them so that there is no unnecessary redundancy in the variables. This is how the algorithm singles out the most dominant dynamics in the data. However, in oceanic applications, it is not certain that the most dominant dynamics are the most significant. This is the reason why the data are so well represented in most of the variables already at rank 1.

It is particularly interesting to see what PCs are used to describe the wind. A significant portion is explained in PC1, which indicates a certain correlation with the other variables. The largest portion of the wind variance is explained in PC2, indicating that the wind variance is the second largest, after the variance all variables have in common. Moreover, this PC is increasingly negatively correlated with the current through the water column. This indicates that there is a correlation between wind and deeper currents. However, as PCA is a time-independent method, it is not further indicated how they are correlated using this method.

The results from the first scenario shows a very well performing PCA, as the main intention with using PCA in this work is to explore efficient data reduction techniques. They show that these data are well suited for PCA, as they are highly correlated as seen in the biplot in figure 5.2 and the variance-explained plot in figure 5.3. Figure 5.2 also indicates which variable variance is explained in which PC, and how much, for the first 3 PCs. DoC is also included in this figure, and shows that DoC varies linearly with rank. Generally, PC1 includes some variance from all variables, deeper currents towards 100m and wind are explained in PC2 and surface- and top currents are largely explained in PC3. That indicates that the variance in bottom currents are generally larger than in surface currents. But this might be influenced by the wind variable, which has the largest variance and is largely explained in PC1 and PC2. What is interesting is that wind and current has variance in PC2, but the wind is negatively correlated with the bottom currents. This representation does not explain all relations between the variables, which is a complex composition of more than the first 3 PCs. Still, as seen in the reconstructed data, the first PCs are sufficient to reconstruct a large amount of the dynamics in the system at a low rank meaning that a high DoC is maintained.

For these particular data, rank 3 gives a well reconstructed current in all variables, while maintaining a high DoC, and is recommended as the minimal rank. This is because, as seen in figure 5.2, surface- and top currents are largely explained in PC3, and deeper currents explained in PC2. Higher order ranks are harder to explore, as the maximum dimensions for any plot is 3D. However, they can be relatively easily explored by doing the reconstructions after getting an indication of how many PCs are needed from the variance-explained plot, and by looking at error-plots as in figure 5.6. Although a recommendation of rank is made here based on the minimum amount of data included, it should be noted that the chosen rank needs to be adapted according to the specific application of the analyzed data.

### **Second scenario**

In the second scenario, PCA is performed in the horizontal 2D plane, looking at the correlation between wind- and current in the water column, and the interaction between the five grid point locations distributed as seen in figure 3.3. Two depths are chosen, where the dynamics are expected to have large differences. However, after analyzing them, the observed results in all the presented explanatory plots are very similar. This is further seen when doing the same analysis for corresponding directional data. Therefore, only one set of results are presented. Given that the main purpose of this second scenario analysis is to gather information on the correlation between the horizontal points, these results are considered satisfactory. The



horizontal measurements, which are the variables, does not maintain such a high measurement resolution as seen in scenario 1. Furthermore, a significantly higher number of observations are used, making a comparison between the two scenarios somewhat unfair. However, some observations of the differences between the two methods are made.

The results indicate that the data are not as strongly correlated as in the first scenario. In this case, the method really prove its power when it is applied to such a dataset, and still produces adequate reconstructions. The biplots in figure 5.7 and figure 5.8 show that the origin- and N10-points, and S10- and W10-points are highly correlated between them. The E10 point stands out from the other grid points, and it is reasonable to interpret that this variable shows the lowest correlation with the other variables, as it is close to completely described by PC2. Figure 5.9 shows that about 95% of the total variance can be explained at rank 3, which is a considerable amount, and gives an indication that the data in the grid points are correlated. The estimates of rank 2, rank 3 and rank 4 show that the reconstruction improves gradually corresponding with the higher rank. Still, there are larger variations in these data compared with the analysis done in the first scenario. It should also be noted that the 10km distances are interpolated values, which influences the relations between the origin and the 10km points in each direction.

For the two PCA scenarios, it is evident that PCA is better suited for the first scenario, where the variables are current measurements through the water column. The outcome from this analysis show that the data can be well reconstructed in a compressed model, and that the method is well suited for this type of data. That is useful to know before PLSR analysis, which is similar enough to PCA so that is reasonable to assume that the water column data is well suited for both methods. The second scenario, where measurement points in the horizontal plane are analyzed, shows that PCA is not as well suited for these data. Still, the method does work, and correlation between the points are explored.

Furthermore, PCA indicates that the wind stands out compared with the current through the water column. It has a strong correlation with surface currents, and a negative correlation with the deeper currents at 100m for these particular data. This might be related to time-dependency between the observations, which are not directly interpretable in PCA.

## 8.2 PLSR

### Estimation

For the first analysis done with PLSR, the same amount of data as in the first PCA is used, but for different variables. The reconstruction shows a very good estimation in the first 100m of the water column, before the error increases rapidly for larger depths. This might be related to the first 100m of the water column being highly correlated with the wind and top currents, which are used as predictor variables. As seen from chapter 2.5.1, the Ekman layer stretches through the surface typically down to between 100m to 150m as a boundary layer. This phenomena is wind-driven, and hence might be an explanation of the low error seen in figure 6.2. The error appears to be varying linearly down to a depth of 150m, and then the deepest currents increases exponentially. This also indicates that the correlation between the wind and deeper currents are less obvious, and PLSR might not be able to successfully estimate this part of the water column based on the given predictor variables. When using more data for training, some of the same are observed. The results of the estimation show that generally the top part of the water column is particularly well estimated, which is likely a result of a high correlation between the layers in the ocean.

However, the estimation show an increasing error through the water column, and in the deeper currents the error is generally high. From timeseries of full rank 32, anti-phase in x-velocity component and large offsets in y-velocity component are frequently observed between deeper current estimates and the real data. The amount of data included seems to have an impact on the analysis as seen in MSE plot in figure 6.10, where the error is higher for a smaller amount of data included both in x- and y-velocity components at rank 32. A reduced rank estimation gives a quite low error in the MSE plot in figure 6.11. Based on the corresponding time-series, this seems accurate in the upper parts of the water column, where there are little difference between the reconstruction of rank 4 and rank 32. However, in deeper currents, the amplitude of both velocity components are considerably lower than the real values, and almost resembling mean current for each component. Furthermore, the magnitude of the x-velocity is quite low compared with the y-velocity. Variance is maximized in the first components, which in many cases includes only the most dominant behavior, and not the most significant. As MSE shows the summed error in each variable, some information of how the error occurs in a matrix is lost. A selection of observed error compositions are shown in Appendix B.1.2. From the figures, it is seen that largest error can occur in any part of the water column, although it is most commonly observed in bottom currents.

Depending on the application of the estimated data and the allowable error, the estimation might be acceptable down to a particular depth. This might be useful for estimating current used in iceberg drift predictions, where data down to a depth of about 150m to 200m is necessary. This is given that wind measurements and current measurements to a depth of 20m can be obtained and used in the trained model. However, other methods should be applied to improve the estimation.

Using a different regression technique might be beneficial for ocean applications, particularly in estimation. As seen when both estimating current and predicting current, deeper current dynamics appear to be more complex than a linear method can reproduce in a satisfactory accurate manner. For estimation, oscillatory terms seems to be the main limitation of the method. Therefore, it might be preferred to use a method that specifically can handle that. Another alternative approach to PLSR estimation might be to do PLSR on mean currents, and then add tidal-current from a model post-analysis. However, this might not give a particularly more accurate result. Furthermore, estimating based on deeper surface currents of maybe only 10m further down also might improve the estimate. However, deeper estimates are harder to obtain. A different approach that might be interesting is to do separate analyzes the x- and y-velocity components. As the PLSR analysis is time-independent, this might not give different results. However, as the error in x- and y-velocity components give quite different results, PLSR might be satisfactory in one of the components, while a different method gives better results for the other.

## **Forecasting**

The variance explained plot shows that at rank 5, about 90% of the total variance is explained in the reconstructed data. Furthermore, the full rank 32 model results in over 99% of the total variance being explained. When increasing the rank to between about rank 15 to full rank 32, not much of the total variance appears to be added to the analysis from figure 6.12. That indicates that a full rank is not necessarily needed in forecasting, in order to explain a large amount of the variance. Already at rank 5, total variance explained is as high as 90% and at rank 7, the total variance explained is close to 95%. Either one of these ranks are potentially high enough for the main dynamics to be predicted. At higher ranks, the increase in total variance included stagnates. At around rank 18, increasing the rank adds very little variance to the dataset. However, as the method is based on largest variance, it is not given that the largest variance is the most significant variance. As seen when looking at difference between rank 32 and rank 18 in the timeseries plots for all amounts of training data, the difference is marginal. It is likely that the last components are mostly non-oscillatory components that gives

the timeseries various offsets. This behavior has been observed in the components when Singular Spectrum Analysis (SSA) was done in the project thesis during autumn (Tørresen, 2018). Therefore, the oscillatory components are considered more dominant by the PLSR algorithm, and lower ranks will get a offset compared with the actual data. This is also seen in timeseries of different ranks, where offsets in the timeseries are better reconstructed at higher ranks.

The results from PLSR forecasting in chapter 6.2, shows that rank 5 gives quite large deviations between forecasted current and real current. This is observed regardless of the included amounts of training data, and can also be seen in the corresponding MSE plot in figure 6.23. At full rank 32, the results show a very good forecasting, for all amounts of data included. Even deeper currents are generally very well predicted, even though the x-velocity component has some error in amplitude for the very deepest currents. However, based on observations from the timeseries, this is marginally better when using 10 years of training data than when using 2 and 5 years. The same is observed in the corresponding MSE plot in figure 6.22. Timeseries are also shown for rank 18, which show a very similar reconstruction as when using the full rank 32. A marginally larger error is seen in deeper currents in x-velocity component, but it is lower than when using rank 5. At rank 18, the data has a DoC of almost 44%. That means that by using data of a little more than half the size of the full model, an almost equivalent reconstruction can be obtained. Based on this, it is reasonable to determine a rank between 18-20 as an optimal rank. However, determining optimal rank is highly dependent on the application of the reduced data and a given allowable error. It should also be noted that forecasting is done for one hour ahead in time. As ocean currents do not change particularly much during one hour, a good prediction is expected. If a larger time horizon is included, the error is expected to increase correspondingly.

### **Comparing estimation and forecasting**

One particular problem with the PLSR estimation case is that in a full rank model, not all of the system variance is described as seen in figure 6.3, only about 90% to 95% at maximum rank. That means that even at full rank 32, there is an error that makes it inexact in parts of the reconstructed timeseries. In other analyzes, such as forecasting, this same amount of total variance explained is seen at considerably lower ranks close to the forecasting analysis minimum rank. This indicates that PLSR might not be the best analysis tool for estimation in this case and this type of data. In forecasting, the results using PLSR are much better, as the total variance explained reaches a high percentage quite fast. Here, it should be noted that the minimal rank in forecasting explains about the same amount of the total variance in the model as the minimum rank in forecasting. At ranks higher than about 18 when forecasting, very little of the total variance is added, and as seen in the

resulting timeseries, using data from rank 18 gives a very similar result as full rank 32. Using rank 18 includes a reduction of matrix size by approximately 43% of the original matrix. That means, by using a little more than half the data, a forecasting of almost the same quality as the full rank model can be obtained.

Although PLSR seems to be a better fit for forecasting than for estimation for these data, the differences in computation time are very different. Forecasting takes significantly more computation time than estimation in this case. The computer used for the analyses includes a 2,7GHz processor and a 8GB memory. With this, the PLSR algorithm when doing estimation using 10 years of data takes about 2 minutes, while the same analysis for forecasting takes about 30 minutes. However, arranging the matrices for estimation and selecting the data takes around 15 minutes in both cases additionally, depending on the amounts of data included. Generally, this computation time is not very high, which is an advantage of the PLSR method. However, the difference is considerable, and it is caused by the training data matrix being much larger in forecasting, when the entire water column measurements are used. A more powerful computer is highly recommended for further analyses.

## 8.3 GPR

Estimating currents using GPR also gives very good results. The estimation is among the best ones observed, particularly for deeper currents. However, there is still a significant error in the deepest part of the water column that can be observed in the timeseries. The problem is mainly seen in the oscillatory terms, where the full amplitude height of the real current is not reached. This might be improved by using a kernel function that includes an oscillatory component, as opposed to the kernel function that is used and specified in chapter 4.3. The error when using PLSR in figure 6.2 and the error when using GPR seen in figure 7.2 are quite similar in the first 100m through the water column. Moreover, the error in both figures increases when the water depth is larger than 100m. However, the value in deeper current MSE is significantly higher when using PLSR than GPR. The y-velocity component has the largest error for both analysis tools and when using GPR, the error in y-velocity is considerably higher than x-velocity error. This is expected to be largely because of the difference in magnitude between the two velocity components, where y-velocity is much higher than x-velocity.

A drawback of using GPR is the significant amount of computation time needed for such an analysis. While PLSR uses around one second to do this particular 1000 observations estimation, GPR uses 15 hours. This is largely because each variable is estimated individually in GPR, as opposed to linear weights in PCA/PLSR. To

perform the same analysis in estimation and forecasting on larger datasets will need more computational power than what has been available in this project.

An approach that has not been tried is using less data for GPR, which will push the running time down, and might still produce a better regression than PLSR. This has not been tested, as optimization of data amount related to running time has not been further explored. A different approach to be tried out is to do GPR on a compressed model by performing PCA to the data before the GPR-algorithm is applied, in order to reduce the running time.

In all regression analyses, a common denominator is that current prediction results have a slight phase shift between predicted and real data. In all results, this shift shows that the predicted data gives the result a little earlier than what is actually occurring. This is favored over it occurring later, and it is observed for all data.

## Conclusion and further work

Parameter reducing techniques applied to the hindcast data generally demonstrates that oceanic data are well suited for data compression. The most powerful compression tool used in this work is PCA. Analyzes done with PCA on highly correlated data through the water column are successful in efficiently reconstructing the dataset with a significantly reduced number of parameters. The graphical representations of correlation between the ocean current layers in the top 100m of the water column and the wind are shown in a biplot. It shows a very strong correlation in current layers. Moreover, the wind has a strong correlation with the upper currents, and a negative correlation with the deeper currents of this part of the water column. PCA is not as well suited for horizontal data, even though some information of the correlation between the grid points are obtained. However, a different method is recommended to extract information of how horizontal points relates.

Using PLSR for forecasting and estimation yield somewhat different results. When forecasting, a high percentage of total variance is obtained in a significantly reduced model. This allows a reconstruction at only rank 18, including a 44% reduction in amount of data (DoC), to be very similar to the full rank 32 model. Furthermore, PLSR gives good results in forecasting of current through the water column based on historic data one hour ahead of time, although the error in deeper parts is generally larger than in upper parts of the water column. Moreover, the required computation power are distinctly higher when using large amounts of data.

Using PLSR for estimation of current from 30m based on top currents and wind, produces fairly good estimates of currents close to the predictor variables. However, the error increases progressively down through the water column. In the lower 150m to 300m, quite large errors and frequent deviations from the real data are observed, both in full and reduced rank models. Increasing the amount of

training data slightly, improves the estimation without significant increase in computational time/power. Possible modifications to the data and/or the analysis tool might improve the quality of the estimation analysis. One such modification is using a different method. Therefore, GPR is applied to 6 weeks of data, trying to produce an improved estimation model. This method does result in an improved estimate in deeper currents, with a much lower MSE. However, GPR requires significant computation time which is considered too demanding to apply to the larger amounts of datasets used in this project.

PCA highlights the high correlation between currents in the upper 100m of the water column, which is further observed in the error plots of the prediction analyses. It is very clear that there is a strong correlation between the currents through the water column, and that wind is an important component. Particularly the surface layer has a very strong relation with the wind. However, the composition of all phenomena affecting the currents are complicated. This is further seen in the analysis tools not sufficiently being able to extract the most important dynamics for the predictions in the deeper layers. In the results, there are no clear indication that the seabed is affecting the analyses.

The significant required computation time might stop GPR from being well suited for time-sensitive applications, which short-term iceberg predictions can be. However, the low error of the results might make them useful in the design process of risers and umbilicals. Estimates of deeper currents based on a few measurements of the upper currents are very useful in this application, as this method is fast and fairly simple. Although the PLSR estimation results give quite inexact results, the estimates might be good enough down to a certain water depth. As icebergs float in the surface, with their keel going down to a certain depth, the bottom current accuracy might not be crucial for the application. Furthermore, they might not require data for the entire water column, as seen in the results.

## 9.1 Further work

There are many options on what to explore further from this work. For one, all analyses done in this thesis are performed on only one testing dataset, which only show one possible response. To ensure the robustness of the method, a variety of data should be applied to the trained models. This is particularly recommended in order to validate the results from forecasting using PLSR, which produce a regression model that is very similar to the real data. Moreover, analyzing the time-dimension, by for instance using Dynamic Mode Decomposition (DMD), is useful when choosing what data to use for training the algorithms.



Furthermore, doing estimation by using PLSR produced quite poor results compared to the other regressions done in this project. However, this should be further explored by using different testing data. Checking the robustness of the data, and if including prediction variables further down in the water column produce a better estimation. As forecasting is only done for one hour into the future, longer time should be explored. The predictions are expected to be poorer when doing this. However, finding how far into the future good predictions can be obtained are very useful. The amount of training data included might have a stronger impact on such analyses.

An approach that has not been tried is using less data for GPR, which will push the running time down, and might still produce a better regression than PLSR. This has not been tested, as optimization of data amount related to running time has not been further explored in this project. A different approach to be tried out is to do GPR on a compressed model by performing PCA to the data before the GPR-algorithm is applied, in order to reduce the running time.



# Bibliography

- British Oceanographic Data Centre (BODC) / General Bathymetric Chart of the Oceans (GEBCO), 2018. Gebco home page.  
URL <https://www.gebco.net/>
- Brown, J., Colling, A., Park, D., Phillips, J., Rothery, D., Wright, J., 1999. Waves, tides and shallow-water processes, 2nd Edition. Vol. vol. 4 of Oceanography series. Pergamon press, Oxford.
- Davis, R. A., 2001. Gaussian processes. Encyclopedia of Environmetrics.
- de Jong, S., 1993. Simpls: An alternative approach to partial least squares regression. Chemometrics and Intelligent Laboratory Systems, vol 18, issue 3, pp. 251-263.
- Ebden, M., 2008. Gaussian processes for regression: A quick introduction.
- Glen, W. G., III, W. J. D., Scott, D. R., 1989. Principal component analysis and partial least squares regression. Tetahedron Computer Methodology, Vol. 2, No. 6, pp 349 to 376.
- Hogben, L., Cline, A. K., Dhillon, I. S., 2007. Handbook of linear algebra. Chapman Hall, Ames, Iowa.
- Holmedal, L. E., 2002. Wave-current interactions in the vicinity of the sea bed. PhD Thesis, Department of Marine Hydrodynamics NTNU.
- Johnson, R. A., Wichern, D. W., 2007. Applied multivariate statistical analysis. Pearson, New Jersey.
- Jolliffe, I., 2002. Principal Component Analysis. Springer, University of Aberdeen, UK.

- 
- Løvås, G. G., 2013. Statistikk for universiteter og høyskoler. Universitetsforlaget AS.
- Maitra, S., Yan, J., 2008. Principle component analysis and partial least squares – two dimension reduction techniques for regression. In: Casualty Actuarial Society Discussion Paper Program Applying Multivariate Statistical Models. Québec, Canada.
- Martens, H., Martens, M., 2001. Multivariate Analysis of Quality. Wiley Sons Ltd.
- MathWorks, I., 2018. Documentation: Principal component analysis of raw data. URL <https://se.mathworks.com/help/stats/pca.html>
- MathWorks, I., 2019a. Documentation: Fit a gaussian process regression (gpr) model. URL <https://se.mathworks.com/help/stats/fitrgp.html>
- MathWorks, I., 2019b. Documentation: Partial least-squares regression. URL <https://se.mathworks.com/help/stats/plsregress.html>
- Morales, E., 2018. Coriolis effect. URL <https://www.yourdictionary.com/coriolis-effect>
- Mudge, S. M., 2015. Multivariate Statistical Methods and Source Identification in Environmental Forensics. Introduction to Environmental Forensics (Third Edition), Academic Press.
- Myrhaug, D., 2012. Oceanography: Current. Marinteknisk senter, Marinteknisk senter, Trondheim, Norway.
- Newland, D. E., 1993. Random vibrations, spectral and wavelet analysis, 3rd Edition. John Wiley Sons, New York.
- Ocean Motion, NASA, 2018. The ekman spiral. URL <http://oceanmotion.org/>
- Pearson, K., 1901. On lines and planes of closest fit to systems of points in space. Philosophical Magazine 2, 559–572.
- Pinet, P. R., 2016. Invitation to Oceanography. Jones Bartlett, Burlington, Mass.
- Quiñonero-Candela, J., Rasmussen, C. E., 2005. A unifying view of sparse approximate gaussian process regression. Journal of Machine Learning Research 6 (2005) pp. 1939-1959.

- 
- Rasmussen, C., Williams, C. K. I., 2006. Gaussian processes for machine learning. MIT Press(accessed at [www.gaussianprocess.org](http://www.gaussianprocess.org)).
- Røed, L. P., Lien, V., Melsom, A., Kristensen, N. M., Gusdal, Y., Ådlandsvik, B., Albretsen, J., 2015a. BaSIC Technical Report 4, Part I: Evaluation of the BaSIC4 long term hindcast results. MET report 2015 (no 5), 1–3.
- Røed, L. P., Lien, V., Melsom, A., Kristensen, N. M., Gusdal, Y., Ådlandsvik, B., Albretsen, J., 2015b. BaSIC Technical Report 4, Part II: Addendum. MET report 2015 (no 6), 1–3.
- Schlichting, H., Gersten, K., 2017. Boundary-Layer Theory, 9th Edition. Springer, Berlin, Heidelberg.
- Stewart, R. H., 2008. Introduction to Physical Oceanography. Prentice Hall, Texas.
- S.Wold, A.Ruthe, H.Wold, III, W. J. D., 1984. The colinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. SIAM J. Sci. Statist. Comput., 5.
- Tu, J. H., 2013. Dynamic mode decomposition: Theory and applications. PhD Thesis, Princeton.
- Tørresen, M., 2018. Big data analysis on ocean currents. NTNU(available by request).
- Vautard, R., Ghil, M., 1989. Singular Spectrum Analysis in Nonlinear Dynamics, with Application to Palaeoclimatic Time Series. Vol. 35 of Physica D.

---

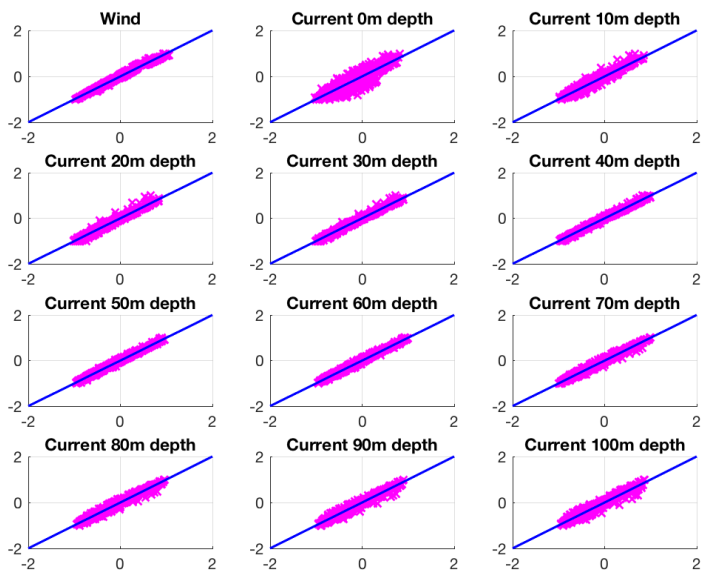
---

# Appendix

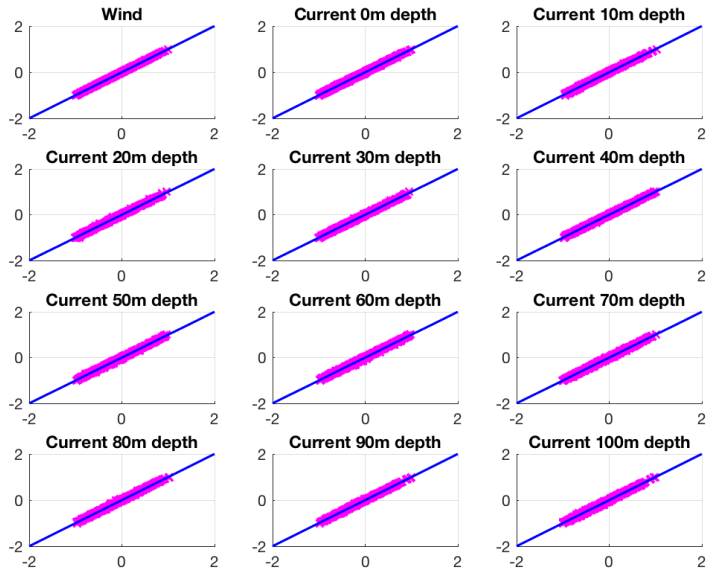
## Appendix A: PCA analysis

### A.1 First scenario: Twelve variables through the water column

Figure 9.1 and Figure 9.2 shows reconstructed and original dataset for rank 2 and rank 4, respectively, from chapter 5.1. They show the reconstructions getting gradually better represented when the rank is increased.



**Figure 9.1:** Similarity between the reconstructed data (x-axis) and the original timeseries data (y-axis) in each variable, for reconstruction of rank 2



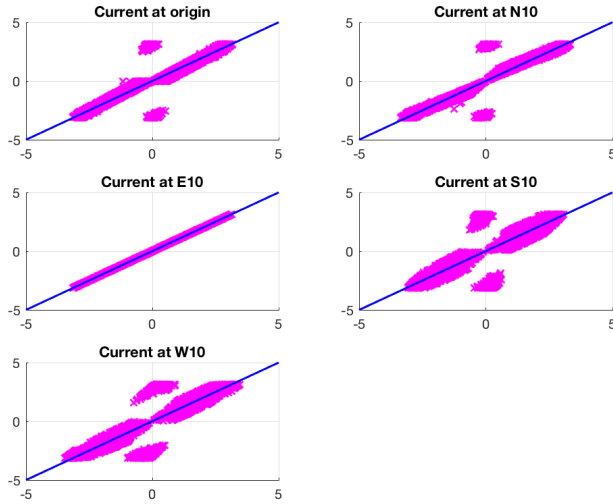
**Figure 9.2:** Similarity between the reconstructed data (x-axis) and the original timeseries data (y-axis) in each variable, for reconstruction of rank 4



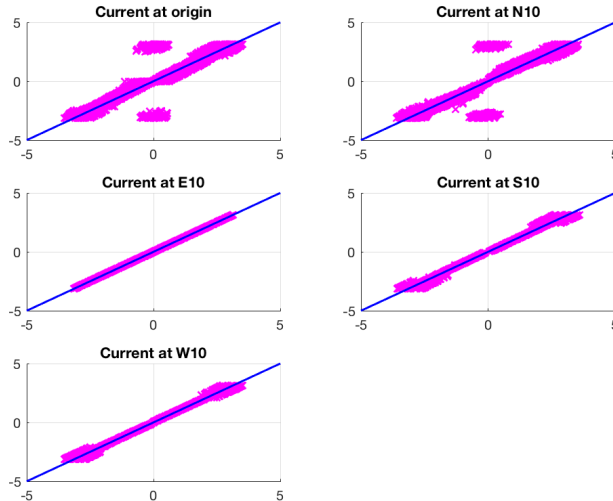
---

## A.2 Second scenario: Horizontal spatial dimension

The higher degree reconstructions are presented in figure 9.3 and figure 9.4. It is observed that the reconstruction improves as the rank increases.



**Figure 9.3:** Similarity between the reconstructed data (x-axis) and the original timeseries data (y-axis) in each variable, for reconstruction of rank 3



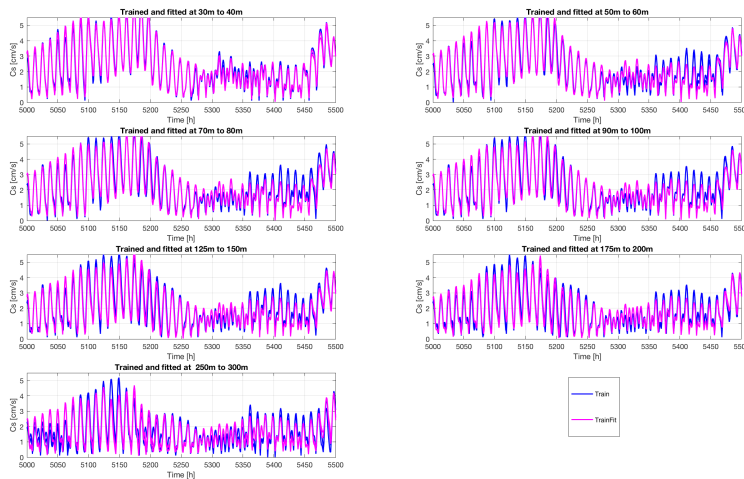
**Figure 9.4:** Similarity between the reconstructed data (x-axis) and the original timeseries data (y-axis) in each variable, for reconstruction of rank 4

---

# Appendix B: PLSR analysis

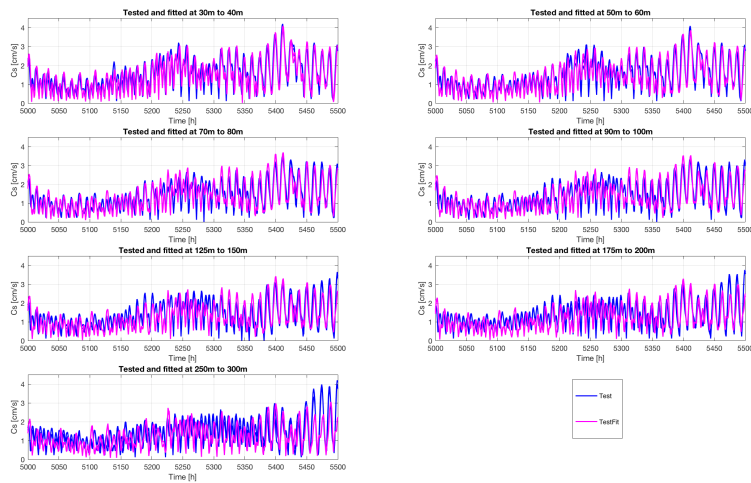
## B.1 Estimating deeper currents based on surface currents

### B.1.1 Trained model applied to training data and testing data

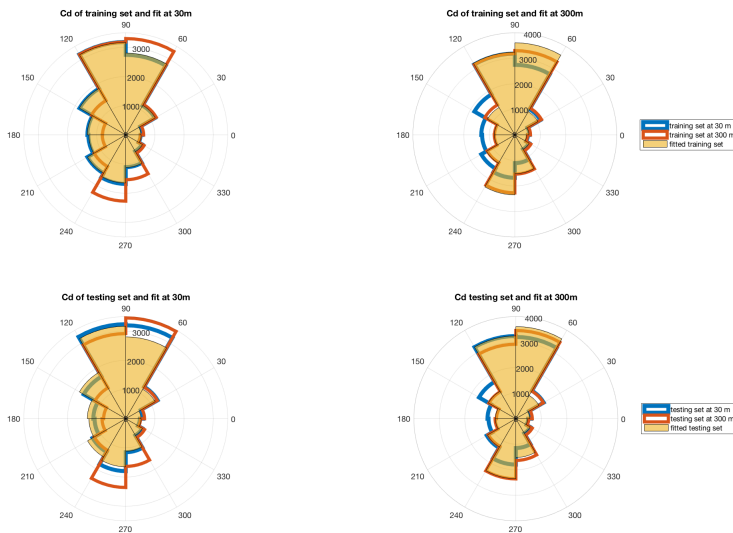


**Figure 9.5:** Timeseries of estimated current and real current from training set with 2 years of training data.  $ncomp = 32$

Figure 9.5 shows a small selection of the fitted data compared with the actual data when using 2 years of training data. Because the dataset is so large, even when using 2 years which is the smallest amount used in this PLSR analysis, only a small selection of the timeseries are shown. Because this is the actual data that the algorithm has been trained to estimate, the result should be pretty good. The corresponding fitted test and test data are shown below, in figure 9.6. There is a significant difference in the quality of the fit for the tested data, compared with the training dataset. This is an example of the difference between the train- and test data applied to the estimated model, and show some of the limitations in using a linear method, which will use the same weights for different data, obtained from its training.



**Figure 9.6:** Timeseries of estimated- and real current strength for test set with 2 years of training data.  $ncomp = 32$



**Figure 9.7:** Current directions observed for estimated- and real current strength for test set with 2 years of training data. The estimation of of full rank 32.

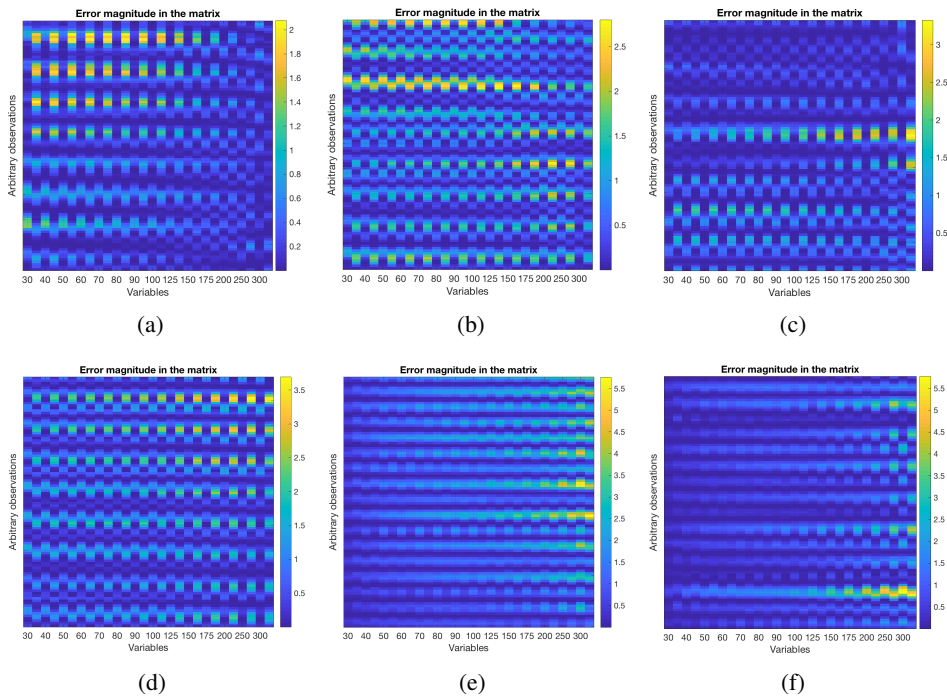
The current directions are shown in 9.7 for the 2 years train- and testing set and their respective fitted data.  $0^\circ$  is equivalent with true north. Only the boundaries

---

of 30m data and 300m data are shown in the compass plots, because these show the largest differences. The depth points in the rest of the water column shows a similar behavior. Training data is used in the first two compasses, therefore the result is very good. Still, there are some minor differences such as in the current direction of training set at 30m in the 60-90° sector, where the fit results in a larger number of observations within this sector than the training data does. Furthermore, for the testing set also shown in 9.7 it is seen that the regression yields good results for these data as well, yet there are distinct differences. At 30m depth it overestimates the number of observations in the three sectors between 120-240° and correspondingly underestimates the number of observations in the sector 60-90°. Looking at the entire water column, the current directions are very similar through the water column. It should also be noted that current directions are looked at in sectors, hence not in such a detailed manner as for current strength.

---

## B.1.2 Examples of error structures observed

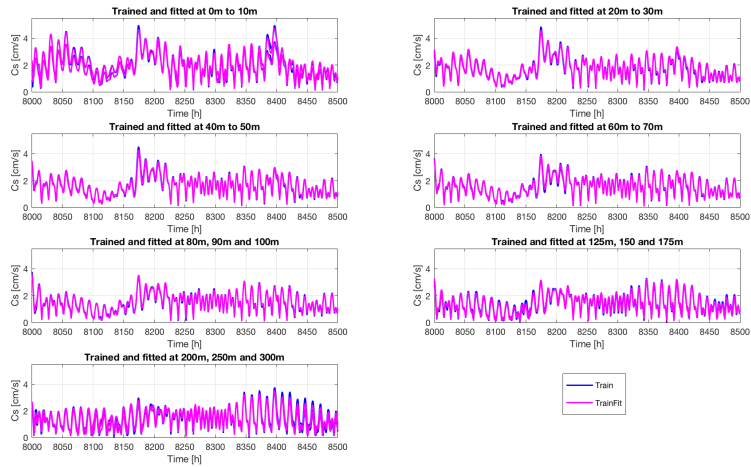


**Figure 9.8:** Examples of different structures of error magnitudes within the error matrix, for x- and y-velocities. They are shown as observed in the analyzed matrices of different ranks and chosen amounts of data: (a) Observed when using 10 years and rank 32, see that the error is smaller for the deeper currents; (b) Observed when using 10 years and rank 4, see that larger error occurs randomly; (c) Observed when using 5 years and rank 32, see a generally low error and some observations of higher error; (d) Observed when using 5 years and rank 32, see larger error occurring periodically; (e) Observed when using 2 years and rank 4, see higher error in deeper currents; and, (f) Observed when using 2 years and rank 32, see generally low error with larger errors in deeper currents.

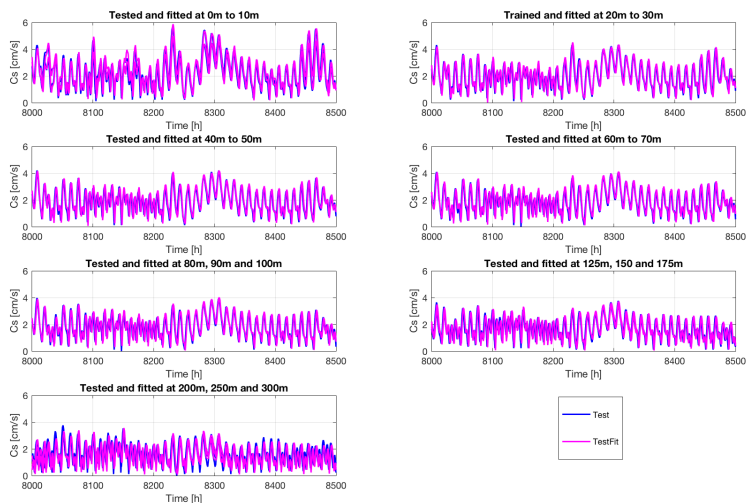
Figure 9.8 shows that a variety of compositions of error magnitude between the variables are observed in the analyzed data. That makes it quite hard to reach one final conclusion for the data.

---

## B.2 Forecasting currents based on historic water column measurements



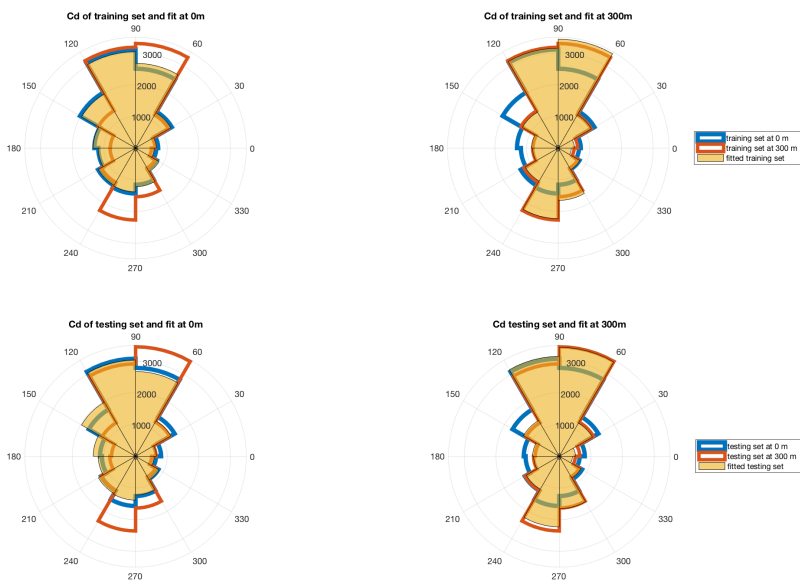
**Figure 9.9:** Timeseries of forecasted current and real current from training set with 2 years of training data.  $ncomp = 32$



**Figure 9.10:** Timeseries of forecasted- and real current strength for test set with 2 years of training data.  $ncomp = 32$

---

Figure 9.9 shows the 2 year timeseries of training data for a selection of points. The training data represents this particular amount of data very well. The testing set performs a bit poorer, as expected. The data are very well replicated. Correspondingly, figure 9.11 show an example of forecasted current direction when using training and testing set.



**Figure 9.11:** Current directions observed for forecasted- and real current strength for test set with 2 years of training data.  $n_{comp} = 32$

