Lappeenranta
University of Technology

Zubeda Mussa Seif

VARIATIONAL ENSEMBLE KALMAN FILTERING IN HYDROLOGY

Thesis for the degree of Doctor of Science (Technology) to be presented with due permission for public examination and criticism in Auditorium 1383 at Lappeenranta University of Technology, Lappeenranta, Finland on the 26th of August, 2015, at 12 pm.

Acta Universitatis
Lappeenrantaensis 653

Supervisor Professor, PhD Tuomo Kauranne
Faculty of Technology
Department of Mathematics and Physics
Lappeenranta University of Technology
Finland

Reviewers Professor Ionel Michel Navon
Department of Scientific Computing
Florida State University
Tallahassee, FL 32306–4120 (850) 644–6560
USA

Professor Heikki Järvinen
Department of Physics
University of Helsinki
Finland

Opponent Professor Ionel Michel Navon
Department of Scientific Computing
Florida State University
Tallahassee, FL 32306–4120 (850) 644–6560
USA

ISBN 978-952-265-832-6
ISBN 978-952-265-833-3 (PDF)
ISSN-L 1456-4491
ISSN 1456-4491

Lappeenrannan teknillinen yliopisto
Yliopistopaino 2015

Abstract

Zubeda Mussa Seif
VARIATIONAL ENSEMBLE KALMAN FILTERING IN HYDROLOGY
Lappeenranta, 2015
75 p.
Acta Universitatis Lappeenrantaensis 653
Diss. Lappeenranta University of Technology
ISBN 978-952-265-832-6, ISBN 978-952-265-833-3 (PDF), ISSN 1456-4491, ISSN-L 1456-4491

The current thesis manuscript studies the suitability of a recent data assimilation method, the Variational Ensemble Kalman Filter (VEnKF), to real-life fluid dynamic problems in hydrology. VEnKF combines a variational formulation of the data assimilation problem based on minimizing an energy functional with an Ensemble Kalman filter approximation to the Hessian matrix that also serves as an approximation to the inverse of the error covariance matrix. One of the significant features of VEnKF is the very frequent re-sampling of the ensemble: resampling is done at every observation step. This unusual feature is further exacerbated by observation interpolation that is seen beneficial for numerical stability. In this case the ensemble is resampled every time step of the numerical model. VEnKF is implemented in several configurations to data from a real laboratory-scale dam break problem modelled with the shallow water equations. It is also tried in a two-layer Quasi-Geostrophic atmospheric flow problem. In both cases VEnKF proves to be an efficient and accurate data assimilation method that renders the analysis more realistic than the numerical model alone. It also proves to be robust against filter instability by its adaptive nature.

Keywords: Data Assimilation, Variational Ensemble Assimilation, VEnKF, transport models.

Preface

The work presented in this thesis has been carried out at the Department of Computational Engineering, previously known as the Department of Mathematics and Physics at Lappeenranta University of Technology (LUT). First and the foremost, I thank the Almighty God for giving me the opportunity, strength and courage to study away from my family.

I would like to express my heartfelt gratitude to my supervisor, Professor Tuomo Kauranne for all the scientific guidance, constructive ideas, thoughtful suggestions and support throughout the entire research. He has been a source of knowledge, encouragement and inspiration and I have been very fortunate to have him as my supervisor. Thank you Tuomo for giving me opportunity to work with you. I would also like to thank Professor Heikki Haario for accepting me to pursue my PhD at LUT.

I would also like to thank the reviewers of this work, Professor Ionel Michel Navon from the Florida State University, USA and Professor Heikki Järvinen from the University of Helsinki, Finland for their valuable comments and thoughtful suggestions to improve this work.

My sincere thanks goes to my colleagues Idrissa Amour and Aleksandr Bibov for their technical support especially in programing with matlab and also being collaborators and co-authors of my research. I am very grateful to both of you.

For financial support, I would like to acknowledge the Science and Technology Higher Education Project (STHEP) of Dar es Salaam University College of Education (DUCE) in Tanzania, for the four years scholarship and Lappeenranta University of Technology for financing the remaining time.

During my studies, I could always count on the full support from my family. Warm thanks goes to my parents Mr. Mussa Seif Abdallah and Mariam J. Abdallah, my siblings Rehema, Fatuma, Ahmad, Zaina, Juma, Seif, Rajabu, Mustafa, Ally and Nassoro. Thank you my children Rahma and Akram for your love and patience during my absence. Special thanks go to my grandmother Zaina Ramadhani for taking care of my children during my absence. You have been like a mother to me and to Rahma and Akram, may the Almighty God bless you abundantly.

I would like to thank my friends in Lappeenranta, Mediatrice and her family, David Koloseni and Pendo Koloseni, Isambi Mbalawata and Jestina Mbalawata, Almasi Maguya, Frank Phillip, Gasper Mwanga, Daniel Osima, Felix John, Amani Metta, Seija Turunen and Deodata Madembwe Heiliö. I would like to give special thanks to Idrissa and his family for all the family love I was getting from your home. Thank you all for making my stay in Lappeenranta enjoyable academically, morally and socially. *"A friend in need is a friend indeed"*.

Special thanks to my co-workers and friends in particular, Dr. Emiliana Mwita and Dr. Fatma Bakari Hamad at Dar es Salaam University College of Education for your words of encouragement and support and always reminding me that "it will be done".

My dear husband Abdallah Mussa, there are no enough words to express my gratitude and love for you. Thanks for your patience, encouragement, love, and support which gave me a lot of strength throughout the entire period. You have made me feel very special even at difficulty times. You have been my strength through all my ups and downs. Thank you for being a good husband and a good father to our children Rahma and Akram. Thank you for being there for me and I am truly grateful.

I was not able to thank everyone who contributed to the success of this work, but I acknowledge and appreciate with thanks all your assistance and invaluable support. ASANTENI SANA

Lappeenranta, August 2015

Zubeda Mussa Seif

To my beloved
FAMILY

CONTENTS

LIST OF THE ORIGINAL ARTICLES AND THE AUTHOR'S CONTRIBUTION

This monograph thesis consists of an introductory part and two original refereed articles appeared or submitted in scientific journals. The articles and the author's contributions in them are summarized below.

I **Idrissa, A., Mussa, Z. S., A. Bibov and T. Kauranne**, Using ensemble data assimilation to forecast hydrological flumes, *Non Linear Process in Geophysics*, 20(6), 955-964, 2013.

II **Mussa, Z. S., Idrissa, A., A. Bibov and T. Kauranne**, Data assimilation of two-dimensional Geophysical flows with a Variational Ensemble Kalman Filter, *Non Linear Process in Geophysics Discussion* (NPGD)2014.

Zubeda Mussa is a co-author of Publication **I**, and a principal author of Publication **II**. In both papers, the author carried out experimentation and processed the results. In both articles, the author has participated in the substantially writing of the articles.

3D-Var	3 Dimension Variational Assimilation
4D-Var	4 Dimension Variation Assimilation
4D-EnVar	Four dimensional ensemble-variational data assimilation
CFD	Computational Fluid Dynamics
CFL	Courant–Friedrichs–Lewy
EKF	Extended Kalman Filter
EnKF	Ensemble Kalman Filter
EnSRF	Ensemble Square Root Filter
KF	Kalman Filter
LBFGS	Limited memory Broyden-Fletcher-Goldfarb-Shanno
NWP	Numerical Weather Prediction
LEnKF	Local Ensemble Kalman Filter
MLEF	Maximum Likelihood Ensemble Filter
QG	Quasi-Geostrophic model
RMSE	Root Mean Square Error
SLF	Statistical Linearization Filter
SWE	Shallow Water Equations
UKF	Unscented Kalman Filter
VKF	Variational Kalman Filter
VEnKF	Variational Ensemble Kalman Filter

OVERVIEW OF THE THESIS

CHAPTER I

Introduction

1.1 Background

In geophysics and atmospheric sciences, researchers have been using data assimilation to approximate the true state of a physical system. The analysis of these physical systems relies upon the forecast model, observation data available, and initial and boundary conditions. Daley (1991) describes this whole process in the case of meteorology. In order to predict the future state of the atmosphere, the present state of the atmosphere must be well characterized, and the governing equations (the model) which are used to predict the future state from the present state have to be well written. The analysis of the physical system at the current time is used as the initial state of the forecast to the next time point and this process, in which observations are combined with a dynamic model to produce the best estimate of the state of the system as accurately as possible, is called data assimilation (Talagrand, 1997; Wang et al., 2000; Navon, 2009).

Modern data assimilation methods, such as the Ensemble Kalman filter (EnKF) (Evensen, 2003) and Variational Kalman filtering (VKF) (Auvinen et al., 2010), have been developed for applications in computational fluid dynamics (CFD) and in operational weather forecasting. In these fields, the most critical task is to solve the corresponding equations of fluid dynamics, mostly shallow water equations (SWE) and the Navier-Stokes equations in different forms. Data assimilation in CFD therefore serves first and foremost the identification of the structure of the flow field. Yet in general it is difficult to observe the flow field directly. Instead, observations are made of quantities that flow along with the flow, such as tracers, or collective properties of the flow, such as pressure or temperature.

Data assimilation is of such central importance to the quality of weather forecasts, that it is worth a lot of development effort. A centerpiece of such efforts over the last thirty years has been the introduction of variational principles to data assimilation (Awaji et al., 2003; Bélanger and Vincent, 2004; Courtier and Talagrand, 1990; Le Dimet and Talagrand, 1986). Furthermore, hybrid methods that combine ensemble assimilation techniques and variational assimilation methods (Sasaki, 1970c,a,b) have been introduced. The goal of this research therefore is to apply a novel method for state estimation in data assimilation, the Variational Ensemble Kalman filter (VEnKF) developed to a large extent at the Department of Mathematics at Lappeenranta University of Technology by Solonen et al. (2012), to environmental problems presented by different types of hydrological models.

1.2 The Scope of the thesis

In this thesis we first introduce the benefit of data assimilation to hydrological modeling using wave meter data of a river model that was first introduced by Martin and Gorelick (2005). In the research work by Amour et al. (2013), we have shown how VEnKF is capable of producing better results than pure simulation when applied to the shallow water model. In this first application, the analysis is limited to a one dimensional set of observation whereby wave meter data of a measured laboratory dam break experiment by Bellos et al. (1991) has been used.

Further studies have been conducted to see whether VEnKF is able to capture cross flow synthetically. To achieve this, the dam break experiment by Bellos et al. (1991) has been modified to have a two dimensional setup of wave meters at the downstream end. VEnKF was then used to assimilate observations of a known flow pattern. VEnKF was later also used to assimilate observations of a two layer Quasi-Geostrophic (QG) model and its performance was compared with the classical extended Kalman filter.

1.3 Objectives

- The main objective of this thesis is to study a novel hybrid data assimilation method, the Variational Ensemble Kalman filter developed at Lappeenranta University of Technology, in real time applications to estimate the state of the dynamic system.
- To apply VEnKF to non-linear models described by the shallow water equations and the Quasi-Geostrophic model.
- To determine whether VEnKF can reproduce the turbulent behavior of the flow even when the pure simulation was not able to achieve this.

To achieve these objectives, VEnKF is applied to a large state estimation problem with highly non-linear model in hydrological modeling using a shallow water model and a QG model. The shallow water model was used to propagate the state and covariance in time and observations from a real dam break experiment were used to update the state. The main questions to be addressed are whether:

- VEnKF can capture the turbulent behavior of hydrological flows and be able to connect the simulated flow to the true flow.
- VEnKF will be able to learn and predict cross flow along a given flume topography, as required for a river flow.
- VEnKF will be able to conserve mass so that the total water storage remains the same for each ensemble member after each assimilation time step.

1.4 Outline

This thesis is organized as follows. After the introduction, Chapter II gives some background of data assimilation and its application to hydrological modeling. In Chapter III, a brief overview of both sequential and variational data assimilation techniques is presented. The hybrid variational

ensemble Kalman filter is also presented. The shallow water model, QG model, numerical solutions and the ability of VEnKF to represents these flows are presented in Chapter IV. Chapter V concludes the research work and suggestions for future research.

1.5 Author Contributions

The Author has done most of the writing and conducted almost all of the test runs of the experiments for shallow water equations (SWE). She has also programmed most of the modifications needed to the original SWE code taken from literature and to the VEnKF library written by one of the co-authors (A. Bibov).

CHAPTER II

Literature Review and Motivation

2.1 Data Assimilation

Data assimilation is the process of combining observations of the current and past state, and the dynamic system model (forecast) in order to produce the best estimate (analysis) of the current and future state of the system (Daley, 1991; Talagrand, 1997; Kalnay, 2003; Wu et al., 2008; Navon, 2009; Blum et al., 2009; van Leeuwen, 2011). Data assimilation has widely been used in numerical weather prediction (NWP) and other branches of geophysics. In weather forecasting, data assimilation is used to generate the initial conditions for an ensuing forecast, but also to continuously correct a forecast towards observations, whenever these observations are available in the course of the forecast (Daley, 1991; Ghil and Malanotte-Rizzoli, 1991; Kalnay, 2003; Fisher et al., 2009; Solonen and Järvinen, 2013). In oceanography, data assimilation has been used as a tool to describe ocean circulation (Stammer et al., 2002; Awaji et al., 2003; Bertino et al., 2003). In general data assimilation has been used for prediction of uncertainty (Moradkhani et al., 2005a), state estimation, parameter estimation (Navon, 1998) or both state and parameter estimation (Moradkhani et al., 2005b; Solonen, 2011; Järvinen et al., 2012; Laine et al., 2012; Mbalawata, 2014).

In data assimilation, the analysis and forecast can be described by means of a probability distributions whereby the analysis is the application of the Bayes theorem which states that, the posterior probability distribution $p(\mathbf{x}|\mathbf{y})$ of the true state $\mathbf{x}$ given observation $\mathbf{y}$, is given as

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{y})}, \tag{2.1}$$

where $p(\mathbf{y}|\mathbf{x})$ is the likelihood function, $p(\mathbf{x})$ is a prior probability which represents the prior knowledge of the state vector, and $p(\mathbf{y})$ is the normalization factor.

Definition 2.1.1 (Probabilistic state space model)**.** *A probabilistic state space model, which can be linear or non-linear, consists of a sequence of conditional probability distributions given as*

$$\begin{aligned} \mathbf{x}_k &\sim p(\mathbf{x_k}|\mathbf{x}_{k-1}), \\ \mathbf{y}_k &\sim p(\mathbf{y_k}|\mathbf{x}_k), \end{aligned} \tag{2.2}$$

for $k = 1, 2, \ldots$, where $\mathbf{x}_k \in \mathbb{R}^n$ is the state of the system at time step k assumed to be a Markov process whose initial distribution is $p(\mathbf{x}_0)$, $\mathbf{y}_k \in \mathbb{R}^m$ is the measurement at time step k, $p(\mathbf{x_k}|\mathbf{x}_{k-1})$

is the dynamic model which describes the stochastic dynamics of the system. The dynamic model can be a probability density, a counting measure or a combination of them depending on whether the state $\mathbf{x}_k$ is continuous, discrete or hybrid, $p(\mathbf{y_k}|\mathbf{x}_k)$ is the measurement model which describe the distribution of measurements given the state (Doucet et al., 2000; Särkkä, 2013).

Data assimilation finds the probability of the true state at time k conditioned on the measurements and the optimal filtering equation is thus given in two steps.

Prediction step: This step involves the computation of prediction distributions of $\mathbf{x}$ by Chapman-Kolmogorov equation given as,

$$p(\mathbf{x}_k|\mathbf{y}_{1:k-1}) = \int p(\mathbf{x}_k|\mathbf{x}_{k-1})p(\mathbf{x}_{k-1}|\mathbf{y}_{1:k-1})\mathrm{d}\mathbf{x}_{k-1} \tag{2.3}$$

Update step: Given the measurement $\mathbf{y}_k$, the posterior distribution is given by the Bayes' rule as,

$$p(\mathbf{x}_k|\mathbf{y}_{1:k}) = \frac{p(\mathbf{y}_k|\mathbf{x}_k)p(\mathbf{x}_k|\mathbf{y}_{1:k-1})}{\int p(\mathbf{y}_k|\mathbf{x}_k)p(\mathbf{x}_k|\mathbf{y}_{1:k-1})\mathrm{d}\mathbf{x}_k}. \tag{2.4}$$

Equations 2.3 and 2.4 can not be solved analytically for higher dimensional problems which are complex in real time applications. Several data assimilation techniques are being used to approximate Equations 2.3 and 2.4 (Todling, 1999). Examples of such techniques are Kalman filter (KF) (Kalman, 1960), extended Kalman filter (EKF), particle filtering techniques, Bayesian Optimal filter, statistical linearization filter (SLF), unscented Kalman filter (UKF) (Julier and Uhlmann, 2004; Chow et al., 2007; Kandepu et al., 2008), ensemble filtering techniques (Evensen, 1994; Houtekamer and Mitchell, 1998; Evensen, 2003), variational Kalman filter (VKF) (Auvinen et al., 2010), 3D and 4D variational assimilation techniques (Le Dimet and Talagrand, 1986; Courtier and Talagrand, 1990) and hybrid variational - ensemble data assimilation techniques (Hamill and Snyder, 2000; Zupanski, 2005; Zupanski et al., 2008; Liu et al., 2008; Gustafsson et al., 2014).

2.2 Data Assimilation in Geophysical and Atmospheric Sciences

In the past years, computational methods have been an essential tool in geophysical and atmospheric sciences. Modeling of geophysical problems is conducted using using computer simulation and solve the underlying partial differential equations using numerical schemes, such as the finite difference method (FDM), the finite element method (FEM) or the finite volume method (FVM) (Ciarlet et al., 2009; Durran, 2010; Lynch, 2008). In order to reduce uncertainties in numerical predictions, observations are combined with these numerical simulations to acquire more reliable predictions.

In the field of geophysical and atmospheric sciences, especially in numerical weather prediction (NWP), data assimilation has long been used to estimate the optimal state of a system by combining the system dynamics defined by the numerical model and real time measurements. The choice of the method to be used depends on the nature of the problem to be modeled and the available observations. However, variational assimilation methods such as 3D-Var and 4D-Var (Le Dimet and Talagrand, 1986; Fisher et al., 2009; Cacuci et al., 2013), have been commonly used in NWP although their use is limited by the need of a tangent linear and an adjoint model for the evaluation of

the gradient of the cost function which leads to a high computational cost (Le Dimet and Talagrand, 1986). The main idea in the use of these methods is to solve the underlying maximum a posterior optimization problem defined by a cost function proportional to the square of the distance between analysis and both background and observations (Bertino et al., 2003). Navon (2009) gives a review of these methods in application to NWP. See also the study by Courtier and Talagrand (1990).

Ensemble methods have been developed and used in geophysics application (Evensen, 1994). The ensemble Kalman filter (EnKF) that begins with (Evensen, 1994) and later by Houtekamer and Mitchell (1998); Doucet et al. (2000); Evensen (2003) uses a Monte Carlo approach such that the error covariance matrices are replaced by the corresponding sample covariance matrices calculated from the ensemble and the ensemble of states is propagated in time using the fully non-linear model (Evensen, 1994; Reichle et al., 2002a; Bertino et al., 2003; Hoteit et al., 2007; McMillan et al., 2013). Kalnay et al. (2007) and Gustafsson (2007) discuss the advantages and disadvantages of 4D-Var and EnKF in application to data assimilation. Although EnKF is suitable for high dimensional problem, it fails for highly non-linear model (Slivinski et al., 2015).

Several formulations of ensemble methods include the ensemble square root filter (EnSRF) (Whitaker and Hamill, 2002; Tippett et al., 2003), ensemble adjustiment Kalman Filter (EAKF) (Anderson, 2001), and the local ensemble Kalman filter (LEnKF) (Ott et al., 2004). Whitaker and Hamill (2002) pointed out that EnSRF is an example of an ensemble filter that does not require perturbed observations, it does not add sampling error as ENKF does and hence is more accurate. However, Lawson and Hansen (2004) have shown that a stochastic filter such as EnKF can handle non-linearity better than a deterministic filter such as EnSRF. On the other hand, LEnKF divides the state into local regions and the analysis is performed in each local region to obtain a local analysis mean and covariance and these are then used to construct the ensemble of the global field that is to be propagated to the next analysis time. Other Monte Carlo approaches include the use of a particle filter for non-linear problem (Doucet et al., 2000; Snyder et al., 2008; van Leeuwen, 2010, 2011).

In recent years, other techniques that combines ensemble methods and variational assimilation have been developed to form hybrid methods (Hamill and Snyder, 2000; Hunt et al., 2004; Liu et al., 2008; Buehner et al., 2013; Gustafsson et al., 2014). These methods have been found to produce comparable results with other assimilation techniques. In several studies different approaches have been used to present the prior error covariance. Hamill and Snyder (2000) showed that, the prior error covariance is obtained as a weighted sum of the sample covariance and the 3D-Var covariance by introducing a tuning parameter. The main drawback of the method is that it works under perfect model assumption. Liu et al. (2008) extend the ensemble 3D-Var to ensemble based 4D-Var (En4DVAR) and, using a shallow water model in a low dimension space, a test of its performance is made and found to produce similar result as that of 4D-Var with less computational cost. On the other hand, Buehner et al. (2013) made a comparison between 3D-Var, 4D-Var and a four dimensional ensemble-variational data assimilation (4D-EnVar) in deterministic weather prediction. They also used the same approach used by Hamill and Snyder (2000) to represent the prior error covariance. It has been found that the computational cost of the 4D-EnVar is lower than that of 4D-Var and 4D-EnVar analyses produce better forecasts than that of 3D-Var and similar or better forecasts when compared with 4D-Var in the troposphere of the tropics and in the winter extra-tropical region and similar or worse analyses in the summer extra-tropical region. In general the 4D-EnVar method proposed by Buehner et al. (2013) can be taken as the best alternative to 4D-Var in terms of simplicity and computational efficiency.

In Zupanski (2005), a maximum likelihood ensemble filter (MLEF) is proposed. MLEF uses

Bayesian theory and combine maximum likelihood and ensemble data assimilation. The state estimate is obtained as the state that maximizes the posterior probability density distribution (Zupanski, 2005). MLEF and other ensemble based variational algorithms (Hunt et al., 2004) use ensemble based prior error covariance. Unlike the variational ensemble Kalman Filter (VEnKF) by Solonen et al. (2012) which will be discussed in Chapter III, Section 3.1.6, MLEF does not include model error and it generates a single ensemble of forecasts at the beginning of the forecast and uses it for the whole assimilation process (Amour et al., 2013).

In hydrological and coastal models, data assimilation has not been applied very often. Liu et al. (2012) review some challenges in the application of data assimilation in hydrological forecasting. High non-linearity of the hydrological processes, high dimensionality of the state vector, the need to use large samples when using ensemble methods (Liu et al., 2012) and estimating the error covariance matrix for high dimensional state vectors (Kuznetsov et al., 2003; Blum et al., 2009) are described as the main challenges to be considered before the application of data assimilation techniques in hydrology. The main focus of hydrological modeling using data assimilation is to estimate the state and uncertainty of the dynamic system by combining observations (water level measurements, flow fields, soil moisture e.t.c) with the hydrological model, given the knowledge of the current state of the system. Hydrological modeling includes flood forecasting of river flows (Bélanger and Vincent, 2004; Madsen and Skotner, 2005) and soil moisture estimate (Reichle et al., 2002a). Bélanger and Vincent (2004) used the 4D-var assimilation technique to forecast floods using a simplified sediment model. In their study, 4D-var was found to outperform direct numerical simulation in producing an optimal analysis, however, it is computationally expensive in high dimensional problems and its application is hindered by the need of an adjoint model required in the evaluation of the gradient of the cost function. Furthermore, data assimilation was found useful in estimation of parameters of hydrological models (Moradkhani et al., 2005b; Lü et al., 2011).

Forecasting may be short-range, medium-range or long-range (Stensrud et al., 1999; Wood et al., 2002; Madsen and Skotner, 2005; Sene, 2010). In meteorology and hydrology forecasting is very important and has the advantage of (1) setting of action plan for disaster management, for example predicting flood and drought in advance, (2) Infrastructure development, (3) reducing damage and loss of life in case of disasters, and (4) disseminate information to the community. Thus, for hydrological modeling, the quality of forecast is of vital importance for decision making and immediate action plan. This can only be achieved when using data assimilation with a good and a reliable technique.

2.3 Motivation

The VEnKF method has been introduced and studied in Solonen et al. (2012) but only simple models have been used to validate the method. The study by Solonen et al. (2012) leaves open question whether VEnKF is a robust and valuable member in the family of approximate Kalman filters and whether if it can be applied to a real data assimilation problems. The main focus of this research therefore, is to study the behavior of VEnKF applied to a highly non-linear model where model error is also present. The emphasis will be on how easily VEnKF can be used and improving accuracy over other methods used in the past. VEnKF was applied to a real data assimilation problem using a shallow water model in one-dimensional and two-dimensional observation setting. VEnKF was further applied to a two-dimensional Quasi-Geostrophic model.

CHAPTER III

Data Assimilation Techniques

3.1 Filtering Techniques

Data assimilation techniques fall into two main categories namely, sequential assimilation methods and variational assimilation methods (Talagrand, 1997). Starting from a prior estimate for the initial state $\mathbf{x}_0$, the dynamic model is evolved to time k where the first observation is available. The predicted state of the system also known as the background state is denoted by $\mathbf{x}_k^p$. The difference between the predicted observation vector given by the background state and the vector of measured observations at this time is given by $\mathbf{K}\mathbf{x}_{k+1}^p - \mathbf{y}_{k+1}$. where, $\mathbf{K}$ is the observation operator. This difference is used to make a correction to the background state vector so as to get the improved state estimate $\mathbf{x}_k^{\text{est}}$ known as the analysis state. The model is then evolved forward again from the analysis state to the next time step where an observation is available and the process is repeated. This describes the sequential assimilation methods whereby the state is updated every time when observations become available (Nakamura et al., 2006). Examples of these methods include nudging (Zou et al., 1992b), particle filter (Moradkhani et al., 2005a; Snyder et al., 2008), the Kalman filter (Kalman, 1960) and its variants and the ensemble Kalman filter and its variants (Evensen, 2003, 2009; Houtekamer and Mitchell, 1998, 2001).

On the other hand, variational assimilation methods, which are computationally more expensive than the sequential assimilation methods, use a batch of data at a specific time interval. These methods solve the underlying maximum a posterior estimate (MAP) equivalent to minimizing the optimization problem that measures the model to data misfit (Bertino et al., 2003) defined by the cost function as presented in Section 3.1.4. However, their use is limited by the need of a tangent linear and adjoint code for the propagation of the covariance (Auvinen et al., 2010). Examples of these methods include optimal interpolation, three-dimensional and four-dimensional variational data assimilation. In this chapter only the Kalman filter (KF), the extended Kalman filter (EKF), the ensemble Kalman filter (EnKF), the variational Kalman filter (VKF) and the variational ensemble Kalman filter (VEnKF) are reviewed.

3.1.1 Kalman Filter

Kalman filter (Kalman, 1960) is an optimal recursive data processing algorithm for estimation of state of dynamic system from noisy measurements in linear Gaussian state space models (Grewal

and Andrews, 2001) subjected to additive Gaussian noises as given by Equations (3.1) and (3.2). KF operates by propagating mean and covariance of the state in time and the task is to estimate the state $\mathbf{x}_k \in \mathbb{R}^n$ of dimension $n \times 1$ governed by dynamic process

$$\mathbf{x}_k = \mathbf{M}_{k-1}\mathbf{x}_{k-1} + \mathbf{q}_{k-1}, \tag{3.1}$$

with a measurement $\mathbf{y}_k \in \mathbb{R}^m$ of dimension $m \times 1$ governed by the measurement model

$$\mathbf{y}_k = \mathbf{K}_k\mathbf{x}_k + \mathbf{r}_k, \tag{3.2}$$

where $\mathbf{M}_{k-1}$ is $n \times n$ transition matrix of the dynamic model, $\mathbf{K}_k$ is $m \times n$ linear, measurement model matrix, $\mathbf{q}_{k-1}$ and $\mathbf{r}_k$ are the model error and the observation error assumed to be normally distributed zero mean random variables with covariance matrices $\mathbf{Q}_{k-1}$ and $\mathbf{R}_k$ respectively. KF assumes that the model and measurement noises are independent. The main task is to estimate the state $\mathbf{x}_k$ and its error covariance $\mathbf{C}_k^{est}$ at time point k given the measurements $\mathbf{y}_k$.

KF algorithm consists of two main steps: (i) the prediction (forecast) step, where the state of the system is predicted based on the previous state and (ii) the update (analysis) step where the state is updated based on the available measurement at that time. The mathematical equations of the KF provide a recursive efficient computation of dynamic states from which the mean of the squared error is minimized and this can be described by Algorithm 3.1.

From Algorithm (3.1), $\mathbf{x}_k^p$ is a prior state estimate, $\mathbf{x}_k^{\text{est}}$ is a posterior state estimate, $\mathbf{C}_k^p$ is a prior estimate error covariance, and $\mathbf{C}_k^{\text{est}}$ is a posterior estimate error covariance. The posterior estimate is also Gaussian and therefore it can be estimated from its mean and covariance.

One of the disadvantages of KF is that it is limited to linear dynamic models and Gaussian noise. Furthermore, KF assumes that the state vector of the dynamic model has n unknowns and therefore the error covariance matrix has n^2 unknowns and thus, the propagation of the error covariance matrix leads to a cost of $2n$ model integrations. For non-linear models, other filtering techniques like the extended Kalman filter, unscented Kalman filter, ensemble Kalman filter and particle filters (Jazwinski, 1970; Julier and Uhlmann, 2004; Evensen, 2003; Doucet et al., 2000) are used instead. The extended Kalman filter in presented in the next section.

3.1.2 Extended Kalman Filter

The extended Kalman filter (EKF) is the extension of KF to non-linear optimal filtering problems by forming a Gaussian approximation to the distribution of states and measurements using a Taylor series expansion (Särkkä, 2013). Incorporating the Kalman filter with repeated linearizations of a non-linear dynamical system leads to the EKF that can be used for non-linear models. The dynamic process (Equation 3.1) and the measurement model (Equation 3.2) are now written, respectively, in the form of:

$$\mathbf{x}_k = \mathscr{M}(\mathbf{x}_{k-1}) + \mathbf{q}_{k-1}, \tag{3.3}$$

$$\mathbf{y}_k = \mathscr{K}(\mathbf{x}_k) + \mathbf{r}_k, \tag{3.4}$$

where $\mathscr{M}$ denotes the non-linear model and $\mathscr{K}$ is the non-linear observation operator.

The filter uses the full non-linear evolution model Equation 3.3 to produce a prior estimate: $\mathbf{x}_k^p = \mathscr{M}(\mathbf{x}_k^{\text{est}})$. Non-linear dynamical models require a linearization when deriving the error covariance

Algorithm 3.1 Kalman filter

The prediction and update step equations for KF are:

i) Initialization: Select initial guess $\mathbf{x}_0^{\text{est}}$ and covariance $\mathbf{C}_0^{\text{est}}$ and set k = 1.

ii) Prediction step:

(a) Move the state estimate and covariance in time

$$\mathbf{x}_k^p = \mathbf{M}_{k-1}\,\mathbf{x}_{k-1}^{\text{est}},$$
$$\mathbf{C}_k^p = \mathbf{M}_{k-1}\,\mathbf{C}_{k-1}^{\text{est}}\mathbf{M}_{k-1}^{\mathrm{T}} + \mathbf{Q}_{k-1}.$$

iii) Update step:

(a) Compute the Kalman gain

$$\mathbf{G}_k = \mathbf{C}_k^p\mathbf{K}_k^{\mathrm{T}}\left(\mathbf{K}_k\mathbf{C}_k^p\mathbf{K}_k^{\mathrm{T}} + \mathbf{R}_k\right)^{-1},$$

(b) Compute the state estimate

$$\mathbf{x}_k^{\text{est}} = \mathbf{x}_k^p + \mathbf{G}_k\left(\mathbf{y}_k - \mathbf{K}_k\mathbf{x}_k^p\right),$$

(c) Compute the covariance estimate

$$\mathbf{C}_k^{\text{est}} = \mathbf{C}_k^p - \mathbf{G}_k\mathbf{K}_k\mathbf{C}_k^p.$$

iii) Set $k \to k+1$ and go to step (ii).

evolution equation and thus the measurement model and the dynamic model functions need to be differentiable.

The covariance estimate is obtained by first linearizing the prediction model about $\mathbf{x}_{k-1}^{\text{est}}$:

$$\mathbf{M}_k = \frac{\partial \mathscr{M}(\mathbf{x}_{k-1}^{\text{est}})}{\partial x}, \tag{3.5}$$

so that the prior covariance estimate is given by

$$\mathbf{C}_k^p = \mathbf{M}_k \mathbf{C}_{k-1}^{est} \mathbf{M}_k + \mathbf{Q}_k. \tag{3.6}$$

The measurement model is then linearized about the prior estimate $\mathbf{x}_k^p$ using:

$$\mathbf{K}_k = \frac{\partial \mathscr{K}(\mathbf{x}_k^p)}{\partial x}. \tag{3.7}$$

The full non-linear observation operator is then used to update the state so as to get the current state estimate and the corresponding error covariance estimate:

$$\mathbf{x}_k^{\text{est}} = \mathbf{x}_k^p + \mathbf{G}_k \left(\mathbf{y}_k - \mathscr{K}(\mathbf{x}_k^p) \right), \tag{3.8}$$

$$\mathbf{C}_k^{\text{est}} = \mathbf{C}_k^p - \mathbf{G}_k \mathbf{K}_k \mathbf{C}_k^p. \tag{3.9}$$

The algorithmic formulation of the EKF is shown in Algorithm 3.2.

EKF is effective in many practical cases, easy to use and computationally efficient. However, the method fails to account for the fully non-linear dynamics in higher dimensional problems and hence fails to represent the error probability density because, if n is the dimension of the state vector, and if m is the size of the observation space then it requires storage and multiplication of $n \times n$ matrices and the inversion of $m \times m$ matrices and so, the error covariance matrix has n^2 unknowns and $2n$ model integrations, (Auvinen et al., 2010; Evensen, 2009). So for models with $n \sim O(10^7)$ for example in meteorology and oceanography, matrix storage and computation become prohibitively expensive. This makes the basic formulation of KF and EKF impossible to implement in higher dimension problems.

The linearization in Equation 3.5 and 3.7 requires the measurement and the dynamic model to be differentiable and can be obtained by using finite differences approach which is computationally expensive for models in higher dimension (Särkkä, 2013). The linearization also may lead to poor error covariance evolution which, in some models, lead to unstable error covariance growth, (Evensen, 2009; Blum et al., 2009).

EKF is restricted to Gaussian noise processes, thus models with discrete valued random variables can not use this filtering method (Särkkä, 2013). These factors leads to introduction of other filters to be discussed in the coming sections.

3.1.3 Ensemble Kalman Filter

As pointed out earlier that, the EKF is optimal in estimation of atmospheric states however, it is computational expensive (Houtekamer and Mitchell, 2005) and work well with model which does not have severe non-linearities (Särkkä, 2013). The ensemble Kalman filter (EnKF) which was first introduced by Evensen (1994) was proposed as a stochastic or Monte Carlo alternative to the

Algorithm 3.2 Extended Kalman Filter

The prediction and update step for EKF with additive noise are

i) Initialization: Select initial guess $\mathbf{x}_0^{\text{est}}$ and covariance $\mathbf{C}_0^{\text{est}}$ and set k = 1.

ii) Prediction step

(a) Compute prediction

$$\mathbf{x}_k^p = \mathscr{M}(\mathbf{x}_k^{\text{est}}),$$

(b) Propagate estimate covariance

$$\mathbf{C}_k^p = \mathbf{M}_k \mathbf{C}_{k-1}^{est} \mathbf{M}_k + \mathbf{Q}_k.$$

iii) Update step:

(a) Compute the Kalman gain

$$\mathbf{G}_k = \mathbf{C}_k^p \mathbf{K}_k^{\mathrm{T}} \left(\mathbf{K}_k \mathbf{C}_k^p \mathbf{K}_k^{\mathrm{T}} + \mathbf{Q}_k \right)^{-1},$$

(b) Compute the state estimate

$$\mathbf{x}_k^{\text{est}} = \mathbf{x}_k^p + \mathbf{G}_k \left(\mathbf{y}_k - \mathscr{K}(\mathbf{x}_k^p) \right),$$

(c) Compute the covariance estimate

$$\mathbf{C}_k^{\text{est}} = \mathbf{C}_k^p - \mathbf{G}_k \mathbf{K}_k \mathbf{C}_k^p.$$

iv) Set $k \to k+1$ and go to step (ii).

EKF. EnKF does not need the integration of the state error covariance matrix (Houtekamer and Mitchell, 1998; Evensen, 2003; Reichle et al., 2002b; Houtekamer and Mitchell, 2005), instead, the uncertainty in the state is represented as N samples and thus, it solves the problems of dimensionality and non-linearity suffered by EKF. Different from EKF, the EnKF use the non-linear model to propagate the ensemble of model trajectories. Like KF, there are two steps in EnKF: prediction step (forecast step) and update step (analysis step). In the prediction step, an ensemble of forecast states is computed, and used to compute the error covariances and the sample mean which is used to define the state estimate. The Kalman gain $\mathbf{G}_k$ is computed from these sample mean and error covariances and it is used to assimilate the measurements to produce the analysis of ensemble states. For a linear model, the EnKF converges exactly to the KF with the increase of ensemble size.

There are various versions of EnKF that differ in the computation of update ensemble. The EnKF can be a stochastic filter or a deterministic filter, depending on the added vectors (Kalnay et al., 2007). In the stochastic case, the EnKF uses Kalman gain together with random perturbations while in the deterministic case, the EnKF uses a non-random transformation on the forecast ensemble. The *perturbed observation filter* is the EnKF where the measurement ensemble is created by adding a random vector to the actual measurement (Whitaker and Hamill, 2002). EnKF scheme uses the Kalman filter update equations whereby in the update step, the intuition is to use the Kalman gain to combine the forecast ensembles, measurements and measurement noise.

Now, consider a bunch of N-dimensional random vectors $s_{k,i} \sim \mathcal{N}(\mathbf{x}_k^{est}, \mathbf{C}_k^{est})$ which are Gaussian distributed with mean $\mathbf{x}_k^{est}$ and covariance $\mathbf{C}_k^{est}$, where $k \in \mathbb{N}$, $i = 1, \ldots, N$, and N is the ensemble cardinality. Consider a matrix $\mathbf{X}_k$ depending on $s_{k,i}$, which is defined by the following:

$$\mathbf{X}_k = \left(\left(s_{k,1} - \bar{s}_k \right), \ldots, \left(s_{k,N} - \bar{s}_k \right) \right) / \sqrt{N-1}. \tag{3.10}$$

Here $\bar{\mathbf{s}}_k = \frac{1}{N} \sum_{i=1}^{N} s_{k,i}^p$ denotes the mean of ensemble $s_{k,i}$. A single EnKF data assimilation step defines a procedure of propagating $s_{k,i}$ to $s_{(k+1),i}$ and the algorithmic formulation of EnKF is summarized in Algorithm 3.3. The ensemble Kalman filter can be implemented directly on top of a non-linear model as it does not require either tangent linear or adjoint code and is therefore easy to program. EnKF can perform better when the ensembles are statistically representative, i.e. when ensemble sizes are large relative to the dimension of the state vector. However, dynamical models describing the state of geophysical systems especially in meteorology and oceanography are of very high dimension of $O(10^7)$ to $O(10^9)$ (Blum et al., 2009). Larger ensembles are therefore impossible to implement due to high computational cost. Due to the fact that the EnKF uses a much smaller number of ensemble members than the number of the state variables, it suffers from a number of disadvantages.

- As the number of ensemble members is always smaller than the dimension of the state, the prior error covariance matrix is always under-estimated (Hamill et al., 2001) and this can lead to unreliable prediction.
- Under-sampling leads to ensemble in-breeding (Houtekamer and Mitchell, 1998, 2001; Whitaker and Hamill, 2002) and thereby the analysis error covariance is always underestimated.
- The underestimated prior error covariance produces unrealistic long range spurious correlations between distant points (Houtekamer and Mitchell, 1998; Hamill et al., 2001; Anderson, 2001).

Algorithm 3.3 The ensemble Kalman filter

i) Select the initial guess $\mathbf{x}_0^{\text{est}}$ and covariance $\mathbf{C}_0^{\text{est}}$ and set $k = 1$.

ii) Prediction step

(a) Propagate each ensemble member forward using a stochastic model

$$\mathbf{s}_{k,i}^p = \mathscr{M}\left(s_{k-1,i}^{est}\right) + \mathbf{q}_{k,i}^p,\ i = 1,\ldots,N.$$

(b) Compute sample mean and sample covariance

$$\bar{\mathbf{s}}_k = \frac{1}{N}\sum_{i=1}^{N} \mathbf{s}_{k,i}^p$$
$$\mathbf{C}_k^p = \mathbf{X}_k\mathbf{X}_k^{\mathrm{T}},$$

iii) Update step

(a) Compute the Kalman gain

$$\mathbf{G}_k = \mathbf{C}_k^p\mathbf{K}_k^{\mathrm{T}}\left(\mathbf{K}_k\mathbf{C}_k^p\mathbf{K}_k^{\mathrm{T}} + \mathbf{R}_k\right)^{-1}.$$

(b) Update ensemble members

$$\mathbf{s}_{k,i}^{\text{est}} = \mathbf{s}_{k,i}^p + \mathbf{G}_k\left(\mathbf{y}_k - \mathbf{K}_k\mathbf{s}_{k,i}^p + \mathbf{r}_k\right)$$

(c) Calculate the next state estimate as the sample mean of the ensembles

$$\mathbf{x}_k^{\text{est}} = \bar{\mathbf{s}}_{(k),i}.$$

(iv) Set $k \to k+1$ and go to step (ii).

- Underestimation of the prior covariance can also lead to filter divergence (Houtekamer and Mitchell, 1998; Whitaker and Hamill, 2002).

Different techniques have been developed to overcome these problems. The problem of ensemble inbreeding and filter divergence have been dealt with using covariance inflation (Anderson and Anderson, 1999) while covariance localization has been used to remove unrealistic spurious correlations (Houtekamer and Mitchell, 1998; Hamill, 2001). Houtekamer and Mitchell (1998) suggested that, in order to avoid sampling error that causes spurious correlation, it is necessary to use a cut-off distance. This avoids the use of observations that are far away from the grid point analyzed. Other methods of covariance localization include the local ensemble Kalman filter (LEnKF) of Ott et al. (2004) and a systematic error correction algorithm of Anderson (2012). The LEnKF is an ensemble square root filter where the analysis is estimated using the all ensemble members and observations in the local regions and the local analyses are then used to form a global analysis.

3.1.4 Variational Kalman Filter

Variational data assimilation approaches are used to many numerical weather prediction problems (Le Dimet and Talagrand, 1986; Courtier and Talagrand, 1990; Schlatter, 2000) most of which are applied on the shallow water equations (SWE). A variational formulation of the Kalman filter (VKF) can be used as an alternative to KF and EKF when the computational cost increases and the classical Kalman filters are impractical to implement (Auvinen et al., 2009, 2010). Variational method like 4D-Var has proven to produce the same results as Kalman filter (Li and Navon, 2001) under some given conditions including linear model and observation operator, Gaussian random noise for the state and observations and the use of fixed background error covariance.

Recall that, the Bayesian estimate of the true state $\mathbf{x}$, given the measurement $\mathbf{y}$, is the value which maximizes the posterior probability given by Equation (2.1). Since the probability of measurement does not depend on the true state, the maximum of the posterior probability is attained when the product $p(\mathbf{y} \mid \mathbf{x})p(\mathbf{x})$ is maximized and this is given by the minimum of the cost function defined as:

$$l(\mathbf{x} \mid \mathbf{y}_k) = \frac{1}{2}(\mathbf{x}-\mathbf{x}_k^p)^T(\mathbf{C}_k^p)^{-1}(\mathbf{x}-\mathbf{x}_k^p) + \frac{1}{2}(\mathbf{y}_k - \mathscr{K}(\mathbf{x}))^T\mathbf{R}_k^{-1}(\mathbf{y}_k - \mathscr{K}(\mathbf{x})). \tag{3.11}$$

Here, $\mathbf{C}_k^p$ is the prior error covariance matrix and $\mathbf{R}_k^{-1}$ is the covariance matrix of the measurement noise $\mathbf{r}_k$ and $\mathscr{K}$ is the observation operator that maps the model state onto observation space.

VKF described here, was first introduced by Auvinen et al. (2010) and its main idea is that given a set of observations $\mathbf{y}_k$ and a prior state vector $\mathbf{x}_k^p$, the state estimate or the analysis is the value of $\mathbf{x}$ which minimizes the cost function given by Equation (3.11) and the covariance estimate is given by the low memory approximation of the covariance given by the inverse Hessian. The minimization is done using a limited memory BFGS algorithm (L-BFGS) (Nocedal and Wright, 1999), (see Appendix B) whereby the inverse of the prior covariance $\mathbf{C}_k^p$ is also approximated using LBFGS (Wang et al., 1995) given that,

$$(\mathbf{C}_k^p)^{-1} = (\mathbf{M}_k\mathbf{C}_{k-1}^{\text{est}}\mathbf{M}_k^T + \mathbf{Q}_k)^{-1}. \tag{3.12}$$

The linear VKF method is summarized in Algorithm 3.4.

For the non-linear VKF method, if the non-linear model $\mathscr{M}_k$ can be linearized to $\mathbf{M}_k$ then, the covariance information can be propagated from one observation time to the next. However, this is not practical for problems in large dimension and instead, the tangent linear $\mathbf{M}_k^{\text{TL}}$ and the corresponding

Algorithm 3.4 The variational Kalman filter

i) Select the initial guess $\mathbf{x}_0^{\text{est}}$ and covariance $\mathbf{C}_0^{\text{est}}$ and set $k = 1$.

ii) Move the state estimate and covariance in time:

(a) Compute $\mathbf{x}_k^p = \mathbf{M}_k \mathbf{x}_{k-1}^{\text{est}}$.

(b) Define $\mathbf{C}_k^p = \mathbf{M}_k \mathbf{C}_{k-1}^{\text{est}} \mathbf{M}_k^T + \mathbf{Q}_k$ and use LBFGS to approximate $(\mathbf{C}_k^p)^{-1} = (\mathbf{M}_k \mathbf{C}_{k-1}^{\text{est}} \mathbf{M}_k^T + \mathbf{Q}_k)^{-1}$.

iii) Combine the prior with observations:

(a) Minimize $l(\mathbf{x} \mid \mathbf{y}_k) = \frac{1}{2}(\mathbf{x} - \mathbf{x}_k^p)^T (\mathbf{C}_k^p)^{-1} (\mathbf{x} - \mathbf{x}_k^p) + \frac{1}{2}(\mathbf{y}_k - \mathbf{K}_k \mathbf{x})^T \mathbf{R}_k^{-1} (\mathbf{y}_k - \mathbf{K}_k \mathbf{x})$ using the LBFGS method.

(b) Store the results of the minimization as the state estimate $\mathbf{x}_k^{\text{est}}$ and the inverse Hessian approximation as the covariance estimate $\mathbf{C}_k^{\text{est}}$.

(iv) Set $k \to k+1$ and go to step (ii).

Algorithm 3.5 Non-linear variational Kalman filter

i) Select the initial guess $\mathbf{x}_0^{\text{est}}$ and covariance $\mathbf{C}_0^{\text{est}}$ and set $k = 1$.

ii) Move the state estimate and covariance in time:

(a) Compute $\mathbf{x}_k^p = \mathscr{M}_k(\mathbf{x}_{k-1}^{\text{est}})$.

(b) Use LBFGS to approximate $(\mathbf{C}_k^p)^{-1} = (\mathbf{M}_k^{\text{TL}} \mathbf{C}_{k-1}^{\text{est}} \mathbf{M}_k^* + \mathbf{Q}_k)^{-1}$ if the tangent linear $\mathbf{M}_k^{\text{TL}}$ and the corresponding adjoint code $\mathbf{M}_k^*$ are available for the evolution model $\mathscr{M}$.

iii) Combine the prior with observations:

(a) Minimize $l(\mathbf{x}|\mathbf{y}_k) = \frac{1}{2}(\mathbf{x} - \mathbf{x}_k^p)^T (\mathbf{C}_k^p)^{-1} (\mathbf{x} - \mathbf{x}_k^p) + \frac{1}{2}(\mathbf{y}_k - \mathbf{K}_k^{TL}(\mathbf{x}))^T \mathbf{R}_k^{-1} (\mathbf{y}_k - \mathbf{K}_k^{TL}(\mathbf{x}))$ using LBFGS method.

(b) Store the results of the minimization as the state estimate $\mathbf{x}_k^{\text{est}}$ and the Hessian approximation as $\mathbf{C}_k^{\text{est}}$.

(iv) Set $k \to$ and go to step (ii).

adjoint operator $\mathbf{M}_k^*$ for the dynamic model $\mathscr{M}$ are used if available (Auvinen et al., 2010). The non-linear variational Kalman filter is summarized in Algorithm 3.5.

Example 3.1.1. *(Van der Pol Oscillator) (Gillijns et al., 2006): A first order Euler discretization of the equations of motions of the Van der Pol oscillator yield*

$$\begin{aligned} \mathbf{x}_{k+1} &= f(\mathbf{x}_k) \\ f(\mathbf{x}_k) &= \begin{bmatrix} x_{1,k} + h x_{2,k} \\ x_{2,k} + h(\alpha(1 - x_{1,k}^2)x_{2,k} - x_{1,k}) \end{bmatrix}, \end{aligned} \tag{3.13}$$

where $\mathbf{x}_k = [x_{1,k} \;\; x_{2,k}]^T$ *and* h *is the step size. We assume that the Van der Pol oscillator is driven by* w_k*, that is,*

$$\mathbf{x}_{k+1} = f(x_k) + w_k, \tag{3.14}$$

where $w_k \in \mathbb{R}^2$ *is zero mean white Gaussian noise with covariance matrix* $\mathbf{Q} \in \mathbb{R}^{2\times 2}$*. Assume that for all* $k \geq 0$*, measurements are available so that*

$$\mathbf{y}_k = C x_k + v_k, \tag{3.15}$$

where $v_k \in \mathbb{R}$ *is zero mean white Gaussian noise with covariance matrix* $\mathbf{R} > 0$ *and* C *selects* $x_{1,k}$ *or* $x_{2,k}$*. We can compare the performance of EKF and EnKF by estimating the state* x_k^{est} *so that the discrete time system is stable given that* $\alpha = 1$*,* $h = 0.1$ *and the prior covariance estimate is* $\mathbf{C}_k^p = diag(6.3e-4, 2.2e-4)$*.*

Figure 3.1 shows the state estimates when using EKF and EnKF. It can be observed that the performance of EnKF improves with increase of ensemble size as can be observed in in state estimate of variable x_1 *of Figure 3.1.*

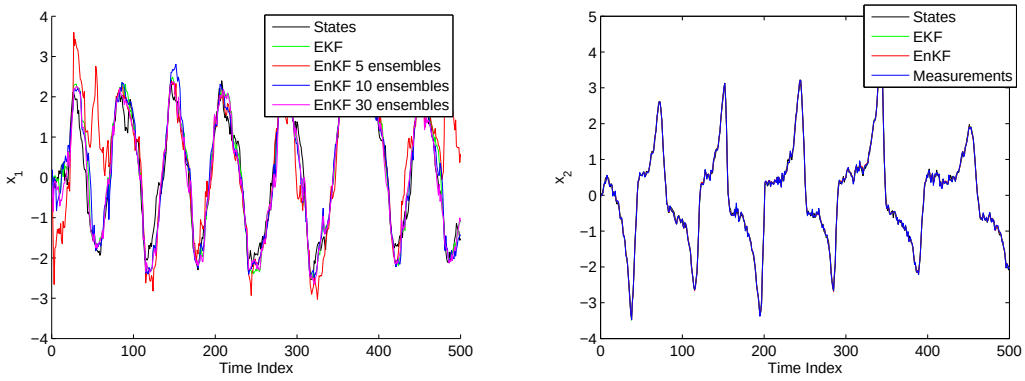


Figure 3.1: State estimate x_k^{est} of the Van der Pol oscillator

3.1.5 Hybrid data assimilation methods

In large scale state estimation in geosciences and in NWP, various ensemble based Kalman filter techniques and variational assimilation methods have been used. In recent years, hybrid data assimilation techniques that combine ensemble methods and variational assimilation have been developed

(Hamill and Snyder, 2000; Hunt et al., 2004; Liu et al., 2008; Buehner et al., 2013; Gustafsson et al., 2014; Lorenc et al., 2014). Slivinski et al. (2015) proposed another kind of hybrid method that combine a particle filter and an ensemble Kalman filter. Their theoretical formulation and test of their performances in Hamill and Snyder (2000); Liu et al. (2008); Gustafsson et al. (2014) and Hunt et al. (2004) have been presented.

Shen and Tang (2015) proposed two modified ensemble Kalman particle filter algorithms called nEnKPF and mEnKPF. The algorithms are obtained by modification of EnKF Kalman gain. Note that, each algorithm is a combination of EnKF and PF, whereby the combination is done by a continuous tuning parameter, which is automatically updated during the run. The proposed algorithms are modification of the EnKPF algorithm of Frei and Künsch (2013) to involve a nonlinear observation model.

The hybrid particle-ensemble Kalman filter (PEnKF) applied to a Lagrangian data assimilation takes the advantage of both particle filter and ensemble Kalman filter algorithms. The idea of a particle filter is to estimate the state of a highly nonlinear dynamic model. However, particle filter tend to suffer from the curse of dimensionality, whereas EnKF deals better with high dimension models.

Given a dynamic system defined as

$$\mathbf{x} = \mathbf{f}(\mathbf{x}, \theta),$$

the idea of particle ensemble Kalman filter is to split the model into two parts,

$$\dot{\mathbf{x}}^F = f_1(\mathbf{x}^F)$$

and,

$$\dot{\mathbf{x}}^D = f_2(\mathbf{x}^F, \mathbf{x}^D)$$

where, f_1 is a high dimensional part, f_2 a is highly non-linear part and $\mathbf{x} = [\mathbf{x}^F\, \mathbf{x}^D]^T$. It should be noted that the splitting technique of the dynamic model has also been done by Salman (2008) where the Fokker-Planck equation for $\mathbf{x}^D$ is solved by an advection diffusion equation.

Even though the PEnKF solves the problem of dimensionality and non-linearity, the issue of sampling error and ensemble in-breeding remain unsolved. It could be of interest if the EnKF in PEnKF is replaced by the variational ensemble Kalman filter.

Liu et al. (2008) proposed a hybrid method that combines EnKF and 4D-var (En4DVAR) whereby the background error is transformed to observation space as in EnKF. Together with this background error covariance matrix, it also uses a control vector that is preconditioned by a matrix calculated from the ensemble that makes the control vector to have dimension N as the ensemble cardinality and hence the cost function is in N-dimensional space. Therefore, apart from taking the benefits of EnKF and 4DVar, the proposed scheme does not require the tangent linear and adjoint codes.

Recently, Yang et al. (2015) proposed a hybrid method called 4DEnVar, where the 4DVar and EnKF are coupled. The proposed hybrid method introduces an empirical ensemble-based background covariance error that does not involve the tangent linear and adjoint codes. The authors proposed the 4DEnVar algorithms with localized covariance approach and local ensemble approach. The method is similar to the one proposed by Liu et al. (2008).

3.1.6 Variational Ensemble Kalman filter

We present another type of hybrid assimilation methods by Solonen et al. (2012) known as the variational ensemble Kalman filter (VEnKF), that use a cloud of points to represent both the error

covariance matrix and the state estimates and which does not require the use of tangent linear and adjoint code for the dynamic model. In VEnKF the state estimate (posterior estimate) is obtained by solving an optimization problem given by Equation (3.11) and the error covariance estimate is obtained as a limited memory approximation of the optimizer.

Thus, the formulation of the variational ensemble Kalman filter is based on the variational Kalman filter as introduced by Auvinen et al. (2010) and the ensemble Kalman filter as introduced by Evensen (2003). The state estimate in VEnKF is computed as a minimizer to the cost function (3.11) and the covariance estimate is the inverse Hessian of (3.11). The basic formulation of VEnKF can be found in details in Appendix A, however, here we present the main idea behind this method.

Consider a bundle of N-dimensional random vectors, $s_{k,i} \sim \mathcal{N}\left(\mathbf{x}_k^{est}, \mathbf{C}_k^{est}\right)$ (here we assume that model state vector as well as its covariance estimated at time instance $k-1$ are known). Therefore, the prediction step now can be formulated as follows:

$$\begin{aligned} \mathbf{x}_k^p &= \mathscr{M}\left(x_{k-1}^{est}\right), \\ \mathbf{s}_{k,i}^p &= \mathscr{M}\left(s_{k-1,i}\right), i = 1, \ldots, N. \end{aligned} \tag{3.16}$$

Define vector $\mathbf{X}_k$ as in section 3.1.3 but now instead of using the mean of the samples, we use the predicted state $\mathbf{x}_k^p$ evolved from the previous time as,

$$\mathbf{X}_k = \left(\left(s_{k,1} - \mathbf{x}_k^p\right), \ldots, \left(s_{k,N} - \mathbf{x}_k^p\right)\right) / \sqrt{N}, \tag{3.17}$$

where N as previously denotes the cardinality of ensemble $s_{k,i}$. Hence, the sampled approximation for the prior covariance can be defined by leveraging the prior ensemble $s_{k,i}^p$ computed on prediction step leading to the following,

$$\mathbf{C}_k^p = \mathbf{X}_k \mathbf{X}_k^T + \mathbf{Q}. \tag{3.18}$$

This sampled approximation allows to programmatically implement the prior covariance $\mathbf{C}_k^p$ as a low-memory subroutine since following (3.18), the computation of a matrix-vector product would only require storage of $\mathbf{X}_k$ (as before, it is assumed that $\mathbf{Q}$ is diagonal or implemented as a low-memory subroutine). Nevertheless, minimization of (3.11) makes use of $\left[\mathbf{C}_k^p\right]^{-1}$, which can be obtained by applying the Sherman Morrison-Woodbury (SMW) matrix identity defined as:

$$\left[\mathbf{C}_k^p\right]^{-1} = \mathbf{Q}^{-1} - \mathbf{Q}^{-1}\mathbf{X}_k \left(\mathbf{I} + \mathbf{X}_k^T \mathbf{Q}^{-1} \mathbf{X}_k\right)^{-1} \mathbf{X}_k^T \mathbf{Q}^{-1}. \tag{3.19}$$

Here, it is assumed that covariance $\mathbf{Q}$ is diagonal and therefore can be easily be inverted. Moreover, since $\mathbf{I} + \mathbf{X}_k^T \mathbf{Q}^{-1} \mathbf{X}_k$ is an N-by-N matrix and the ensemble size N is usually much smaller compared to the problem dimension, the inversions in (3.19) are considered feasible.

Minimization of (3.11) is done by the L-BFGS unconstrained optimizer described in Nocedal and Wright (1999) (see also Appendix B). The L-BFGS is a Quasi-Newton method, which uses the history of its iterations in order to approximate the inverse Hessian of the target cost function. Zou et al. (1993) pointed out that the LBFGS algorithm is one of the best optimization algorithm for large scale unconstrained optimization. Furthermore, the L-BFGS usually converges to the optimal point having a qualified inverse Hessian approximation in a number of iteration that is much smaller than the dimension of the problem. These characteristics of the method can be leveraged to minimize (3.11) as well as to compute its inverse Hessian, wherein both tasks are completed in a single pass.

The same idea may be used instead of SMW matrix identity to obtain $\left[C_k^p\right]^{-1}$ (see Solonen et al. (2012)). However, the L-BFGS only provides an approximation for the inverse Hessian of the target cost function (see Zou et al. (1993)), so formula (3.19) is suggested as the one preferable to use. Finally, putting together (3.16), (3.17), (3.18), (3.19) and the argumentation concerning the L-BFGS, the algorithmic formulation of VEnKF is as shown in Algorithm 3.6.

Algorithm 3.6 Variational Ensemble Kalman filter

i) Select the initial guess $\mathbf{x}_0^{\text{est}}$ and covariance $\mathbf{C}_0^{\text{est}}$ and set $k = 1$.

ii) Prediction step.

(a) Compute prior model state and move the ensemble forward as defined in (3.16).

(b) Define the approximative prior covariance operator $\mathbf{C}_k^p$ in accordance with (3.18).

(c) Apply SMW matrix identity or L-BFGS in order to define a low-memory operator representation of the inverse prior covariance $\left(\mathbf{C}_k^p\right)^{-1}$.

iii) Correction step.

(a) Apply L-BFGS to minimize (3.11) (Wang et al., 1995). Assign $\mathbf{x}_k^{est}$ to the minimizing point and $\mathbf{C}_k^{est}$ to the approximation of its inverse Hessian.

(b) Generate new ensemble $s_{k,i} \sim \mathcal{N}\left(\mathbf{x}_k^{est}, \mathbf{C}_k^{est}\right)$.

(iv) Set $k \to k+1$ and go to step (ii).

BFGS needs to set a set of convergence criterion, and the number of stored vectors is of order of 10 (Auvinen et al., 2010). The convergence criteria applied for this application is a combination of the following.

- Maximum number of iterations in the optimization process.
- Tolerance on the function value. If the difference between two consequent cost function values within the optimization process falls below this value, the optimization process stops and returns the last iteration's point as the optimal.
- Gradient stopping tolerance. If the norm of the gradient value obtained on an iteration step falls below this value, the iteration is being interrupted and the last iteration's point is returned as the optimal one.
- Maximum step size (e.g. norm of direction vector). If during optimization the norm of the direction vector exceeds this value, then the direction vector is normalized to maximum step value.
- Maximum store — number of BFGS vectors stored.

The attractive feature in the presented algorithm is that the operating ensemble is regenerated at every assimilation round, which allows us to avoid ensemble in-breeding inherent to EnKF. VEnKF was first tested using Lorenz 95 model and a large dimension heat equation and later VEnKF was

applied to a more realistic hydrological model as it has been shown in the study by Amour et al. (2013).

Different from all these hybrid filters discussed here, in the variational ensemble Kalman filter we estimate the posterior by solving an optimization problem using LBFGS and then using the search path of the optimizer we obtain the analysis error covariance. The biggest difference between VEnKF and other hybrid filters is the very frequent re-sampling of the ensemble. In this way, VEnKF genuinely propagates a distribution forward in time and not just a set of ensemble spread vectors.

3.1.7 Root Mean Square Error

Results obtained on the use of data assimilation methods have been used to compare theoretical and experimental test cases. The root mean square error (RMSE) in the state estimate is mostly used to show how well an assimilation scheme is performing. If $\mathbf{x}_k^t$ is the true solution and $\mathbf{x}_k^{est}$ is the filter estimate and N is the dimension of the state vector then the RMSE is defined as

$$RMSE = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(\mathbf{x}_k^{est} - \mathbf{x}_k^t)^2} = \sqrt{\frac{1}{N}\left\|\mathbf{x}_k^{est} - \mathbf{x}_k^t\right\|} \tag{3.20}$$

The RMSE can only show how the filter can estimate the mean of the state and not the quality of the uncertainty (Solonen et al., 2014). Table 3.1 shows the RMSE values obtained from example 1 when using EKF and EnKF. It can be observed that the values of RMSE of EnKF approache those of EKF when the number of ensemble members is increased.

Table 3.1: RMSE values

Case	Method	RMSE
1	EKF	0.3478
2	EnKF 5 members	0.7846
3	EnKF 10 members	0.3853
4	EnKF 30 members	0.3524
5	EnKF 40 members	0.3480

CHAPTER IV

VEnKF analysis of hydrological flows

4.1 The Models

4.1.1 The 2D Shallow Water Equations (SWE)

The Shallow Water Equations (SWE) (Martin and Gorelick, 2005; Sarveram et al., 2012; Casulli and Cheng, 1992) are a set of hyparbolic/parabolic Partial Differential Equations (PDE's) governing fluid flows in oceans, channels, river and estuaries. SWE are derived from the Navier-Stokes equations which are also derived from the law of conservation of mass and momentum. SWE are only valid for problems for which the vertical dimension is much smaller than the horizontal scale of the flow features (Tan, 1992), and they have long been used to model various natural and physical phenomenon such as tsunami waves, floods, tidal currents etc (Bellos et al., 1991; Bellos, 2004; Bélanger and Vincent, 2004; Chang et al., 2011). In data assimilation, SWE have also been used in numerical weather prediction (Kalnay, 2003) and in hydrological forecasting (Tossavainen et al., 2008; Chen and Navon, 2009)

The shallow water equations are governed by three equations namely the continuity equation, Equation (4.1), and the momentum equations, Equations (4.2) and (4.3). These equations result from depth avaraging of the Navier Stokes Equations and thus they are called the depth averaged shallow water equation.

$$\frac{\partial \eta}{\partial t} + \frac{\partial (HU)}{\partial x} + \frac{\partial (HV)}{\partial y} = 0, \tag{4.1}$$

$$\frac{\partial U}{\partial t} + U\frac{\partial U}{\partial x} + V\frac{\partial U}{\partial y} = -g\frac{\partial \eta}{\partial x} + \varepsilon\left(\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2}\right) + \gamma_T \frac{(U_\mathrm{a} - U)}{H} - S_\mathrm{fx} + fV, \tag{4.2}$$

$$\frac{\partial V}{\partial t} + U\frac{\partial V}{\partial x} + V\frac{\partial V}{\partial y} = -g\frac{\partial \eta}{\partial y} + \varepsilon\left(\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2}\right) + \gamma_T \frac{(V_\mathrm{a} - V)}{H} - S_\mathrm{fy} - fU, \tag{4.3}$$

where $U = (1/H)\int_{-h}^{\eta} u dz$ and $V = (1/H)\int_{-h}^{\eta} v dz$ are the depth averaged horizontal velocities in the x and y direction, respectively. Note that x and y here denote the Cartesian coordinates, η is the free surface elevation, g is the gravitational constant, t is time, ε is the horizontal eddy viscosity, f is the Coriolis parameter and $H = h + \eta$ is the total water depth, where h is the water depth measured from the undisturbed water surface, γ_T is the wind stress coefficient, U_a and V_a are wind

velocity components in the x and y direction respectively, S_{fx} and S_{fy} are the bottom friction terms in x and y direction, respectively. FU and FV represent a semi-Lagrangian advection operator. The relationship of H, h, and η are as shown in Figure 4.2. The shallow water model described here was used to simulate a physical laboratory experiment of a dam break by Bellos et al. (1991).

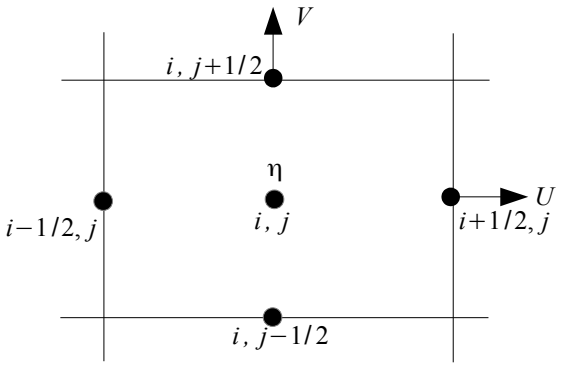


Figure 4.1: Variable location on a computational grid whereby U and V are defined at the face and η is defined at the volume center.

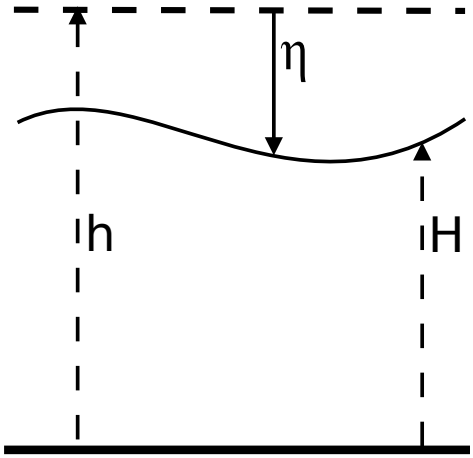


Figure 4.2: Variable definition on a computational grid whereby $H = h + \eta$.

The bottom friction terms are given as: $S_{\text{fx}} = gU\frac{\sqrt{U^2+V^2}}{Cz^2}$ and $S_{\text{fy}} = gV\frac{\sqrt{U^2+V^2}}{Cz^2}$ whereby the Chezy Cz coefficient is defined by the Manning's formula:

$$Cz = \frac{1}{Mn}H^{\frac{1}{6}}, \tag{4.4}$$

where Mn is the Manning's roughness coefficient.

4.1.2 Numerical Solution

To compute the numerical solution for the SWE, Equations (4.1), (4.2) and (4.3) are discretized using a semi-implicit and semi Lagrangian method combined with a finite volume discretization. These discretization methods have the advantage of providing a stable solution (Martin and Gorelick, 2005; Li et al., 1993; Sarveram et al., 2012). The basic idea in semi-implicit discretization is that some terms in a time dependent system are discretized implicitly, and explicit time stepping is used for the remaining terms (Fulton, 2004). In this study the free surface elevation in the momentum equations and the velocity in the free surface equations are discretized implicitly whereas other terms like the advective terms in the momentum equations, coriolis and horizontal viscosity are discretized explicitly (Sarveram et al., 2012; Martin and Gorelick, 2005; Casulli and Cheng, 1992; Li et al., 1993).

The discretization of Equations (4.1), (4.2) and (4.3) are respectively given as,

$$\begin{aligned}\eta_{i,j}^{N+1} =& \eta_{i,j}^{N} - \theta\frac{\Delta t}{\Delta x}\left(H_{i+1/2,j}^{N}U_{i+1/2,j}^{N+1} - H_{i-1/2,j}^{N}U_{i-1/2,j}^{N+1}\right)\\ &- \theta\frac{\Delta t}{\Delta x}\left(H_{i,j+1/2}^{N}V_{i,j+1/2}^{N+1} - H_{i,j-1/2}^{N}V_{i,j-1/2}^{N+1}\right)\\ &- (1-\theta)\frac{\Delta t}{\Delta x}\left(H_{i+1/2,j}^{N}U_{i+1/2,j}^{N} - H_{i-1/2,j}^{N}U_{i-1/2,j}^{N}\right)\\ &- (1-\theta)\frac{\Delta t}{\Delta x}\left(H_{i,j+1/2}^{N}V_{i,j+1/2}^{N} - H_{i,j-1/2}^{N}V_{i,j-1/2}^{N}\right)\end{aligned} \tag{4.5}$$

$$\begin{aligned}U_{i+1/2,j}^{N+1} =& FU_{i+1/2,j}^{N} - (1-\theta)\frac{g\Delta t}{\Delta x}\left(\eta_{i+1,j}^{N} - \eta_{i,j}^{N}\right) - \theta\frac{g\Delta t}{\Delta x}\left(\eta_{i+1,j}^{N+1} - \eta_{i,j}^{N+1}\right)\\ &- g\Delta t\frac{\sqrt{(U_{i+1/2,j}^{N})^2 + (V_{i+1/2,j}^{N})^2}}{Cz_{i+1/2,j}^{2}H_{i+1/2,j}^{N}}U_{i+1/2,j}^{N+1} + \Delta t\frac{\gamma_t(U_a - U_{i+1/2,j}^{N+1})}{H_{i+1/2,j}^{N}}\end{aligned} \tag{4.6}$$

$$\begin{aligned}V_{i,j+1/2}^{N+1} =& FV_{i,j+1/2}^{N} - (1-\theta)\frac{g\Delta t}{\Delta y}\left(\eta_{i,j+1}^{N} - \eta_{i,j}^{N}\right) - \theta\frac{g\Delta t}{\Delta y}\left(\eta_{i,j+1}^{N+1} - \eta_{i,j}^{N+1}\right)\\ &- g\Delta t\frac{\sqrt{(U_{i,j+1/2}^{N})^2 + (V_{i,j+1/2}^{N})^2}}{Cz_{i,j+1/2}^{2}H_{i,j+1/2}^{N}}V_{i,j+1/2}^{N+1} + \Delta t\frac{\gamma_t(V_a - V_{i,j+1/2}^{N+1})}{H_{i,j+1/2}^{N}}\end{aligned} \tag{4.7}$$

In the equations above, Δx is the computational volume length in the $x-$direction, Δy is the computational volume length in the $y-$direction and Δt is the computational time step (Martin and Gorelick, 2005). The parameter θ dictates the degree of implicitness of the solution, and its value ranges between 0.5 and 1, where $\theta = 0.5$ means that the approximation is centered in time and $\theta = 1.0$ means that the approximation is completely implicit (Casulli and Cheng, 1992; Li et al., 1993). For this case θ is set equal to 0.5.

4.1.3 Stability Criteria

For the semi-implicit, semi-Lagrangian used for the discretization of the SWE, the necessary condition for the convergence of the numerical approximations requires that the Courant-Fredrichs-Lewy

(CFL) criteria

$$C = |u\frac{\Delta t}{\Delta x}| \leq 1,$$

where u is the magnitude of the velocity component in the x-direction, Δt is the time step and Δx is the cell dimension.

4.1.4 Initial and Boundary conditions

Initially, we assume that in the domain the motion of fluid begins from an initial state of rest whereby $U = V = 0$ for $t <= 0$ and the initial total water depth $H = H_0$ is given.

Suitable boundary conditions must be applied to Equations (4.1), (4.2) and (4.3) in order to define the flow problem. Thus, different type of boundary conditions exist for these equations. For any simulation domain, the boundary condition may be a closed boundary which does not allow water to flow through the boundary, or an open boundary (inflow only, outflow only or both inflow and outflow) (Agoshkov et al., 1994).

For the closed boundary, the normal velocity component, tangential velocity component and the total water depth needs to be specified, For this case, on this boundary, the tangential and normal velocity component are both treated as zero i.e. $\mathbf{u} = (U,V)^T = 0$ however, there is no condition for the total water depth H (Agoshkov et al., 1994).

Furthermore, at the open boundary, no condition is imposed for the total water depth however, two types of radiation boundary conditions have been set (Martin and Gorelick, 2005):

(i) Projection of velocity normal to the domain

$$\frac{\partial U}{\partial t} + U_{upw}\frac{\partial U}{\partial n} = 0, \tag{4.8}$$

where U_{upw} is the upwinded normal direction velocity component, and n is the direction normal to the domain boundary (Martin and Gorelick, 2005).

(ii) To limit wave reflections at open boundaries (Givoli and Neta, 2003; Navon et al., 2004), the following condition is imposed

$$\frac{\partial \eta}{\partial t} + C_n\frac{\partial \eta}{\partial n} = 0, \tag{4.9}$$

where C_n is the propagation velocity from grid points around the boundary (Martin and Gorelick, 2005).

4.1.5 Dam Break Experiment

Dam break can be defined as uncontrolled release of water due to catastrophic failure of a dam resulting in a serious flooding at the down stream area (Biscarini et al., 2010; Chang et al., 2011). Studies on dam break flow are very significant for the aim of risk assessment to property damage and loss of lives, control of flood and emergence action plan. That is why for many years, dam break study has been a basic tool for researchers. Laboratory test experiments on dam break have been carried out to investigate the nature of flow especially in downstream areas (Bellos et al., 1991; Hu and Sueyoshi, 2010) and these experiments are used for validation of numerical models.

Numerical models based on shallow water equations have been developed in the past to represent the dam break flow given initial and boundary conditions (Morris, 2000; Chang et al., 2011; Biscarini et al., 2010). These models were developed under deterministic settings, and do not account for the uncertainty in the system and thus unreliable prediction. By the use of data assimilation, the observations are being incorporated with these numerical models with the advantage of improving prediction.

In this section, we present a dam break experiment of Bellos et al. (1991) and use data assimilation techniques to study the flow behavior after the dam break. The dam break experiment consists of a flume of length 21.1m and width 1.4m closed at the upstream end and open at the downstream end. It also has a curved constriction beginning at 5.0m and ending at 16.5m from the closed end. 8 sensors used to measure the depth of water were located at an approximate flume mid line as shown in Figure 4.4. A dam is located 8.5m from the closed end and this is the most narrow point of the flume. Initial water height behind the dam was 0.15m and the downstream end is initially dry. When the dam is broken instantly, flood waves sweep downstream and measurements from 7 out of 8 measurement locations were recorded and the total duration of the laboratory experiment is 70 seconds (Martin and Gorelick, 2005).

Figure 4.3 shows the flume geometry (Bellos, 2004) and Figure 4.4 shows the plan view of the geometric lay out of the experiment (Martin and Gorelick, 2005). Figure 4.5 is a snapshot showing the initial water height behind the dam at time $k = 0$. For the discretization of the domain, $\Delta x = 0.125m$ and $\Delta y = 0.05m$ are the grid spatial step and the computational time step is $\Delta t = 0.103$ with a total of 30×171 grid cells, while the Manning's roughness coefficient is 0.010 (Martin and Gorelick, 2005).

4.2 Faithfulness of VEnKF analysis against measurements

4.2.1 1D Set of observations

Prior to this study, VEnKF has been applied to a non-linear and chaotic synthetic model, the Lorenz 95 system and to a relative high dimensional heat equation and found to produce a better result than the standard ensemble Kalman filter (Solonen et al., 2012). In this section we present the application of VEnKF to a real data assimilation problem, by assimilating real data set published by Martin and Gorelick (2005) in the study namely MODfreeSurf2D using a SWE with 1D.

4.2.2 Interpolation of observation

One of the constraint encountered in the application of the VEnKF to acquire efficient minimization was the issue of sparse data set in-terms of space and time. Similar issue in the application of variational data assimilation methods in general has been addressed by Zou et al. (1992a). The data set published in Martin and Gorelick (2005) is very sparse both in time and in space. Data were recorded at an approximate average rate of 1 observation per 1.4 second. More precisely, it means that at a time instance only a small number of wave meters among those installed along the flume were producing actual measurements. These time instances had no alignment with the model integration time step. This sparsity hinders the application of data assimilation techniques since the amount of data obtained from the measurements is usually not enough to expose bias in the

Figure 4.3: Schematic picture of the dam break flume (Bellos, 2004).

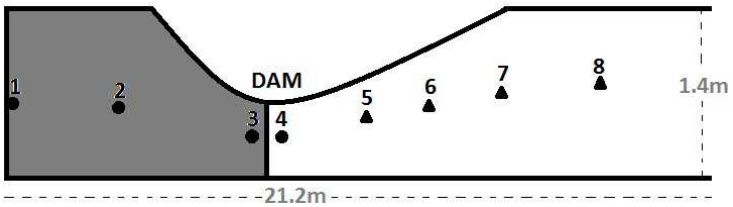


Figure 4.4: Geometrical layout of the dam break experiment (Plan view).

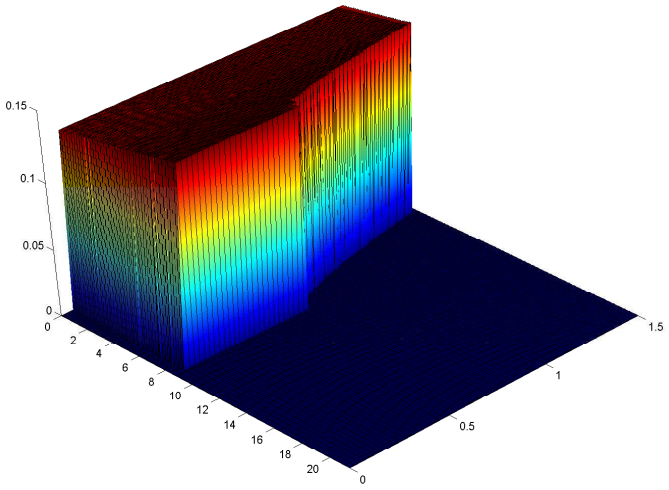


Figure 4.5: Initial water height behind the dam.

prediction model. Therefore simple interpolation technique in time and space has been applied in order to reduce the negative impact due to data sparsity.

The interpolation in time has been done using a spline function and it was organized as follows. The time axis was discretized with a discretization time step of $0.1s$. Thereafter, every time instance related to a measurement obtained from a wave-meter installed in the flume was aligned with the time discretization grid by rounding the time instances to the closest grid point. Since the time grid resolution is smaller than the rate of incoming measurements, some of the time grid points were left with no related observation. These gaps were filled by piecewise cubic interpolation defined by Hermite interpolating polynomials (Fritsch and Carlson, 1980). Figure 4.6 shows original measurements and time interpolated measurements from sensor number 2.

In terms of space, the data were given in only 7 spatial locations, whereas the model state consists of 5130 grid points. The data were much less for data assimilation method and therefore we use this known data set to determine the unknown data of neighboring data points. Thus, for each sensor the data obtained has to be extrapolated to a small neighborhoods of their spatial location. The interpolation has been done by introducing observation values to a 5×5 patch of the grid by sampling from the distribution $\mathcal{N}(y_*, \sigma^2)$, where y_* is the observation value at the sensor and $\sigma^2 = 0.001$. These neighborhoods were specified with the value at the center aligned to the spatial locations of the sensor. With these interpolations, the data are now observed at every time step and on total of 468 grid points. Figure 4.7 shows a spatial interpolated data computed for the 2nd sensor measurements.

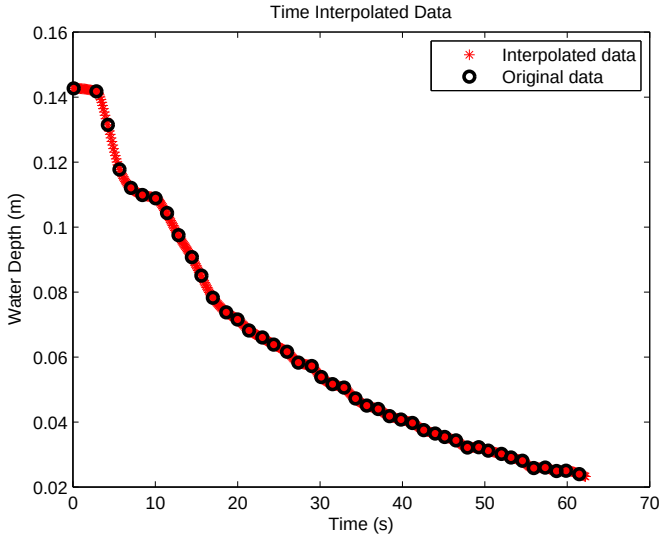


Figure 4.6: Time interpolated water depth at sensor number 2

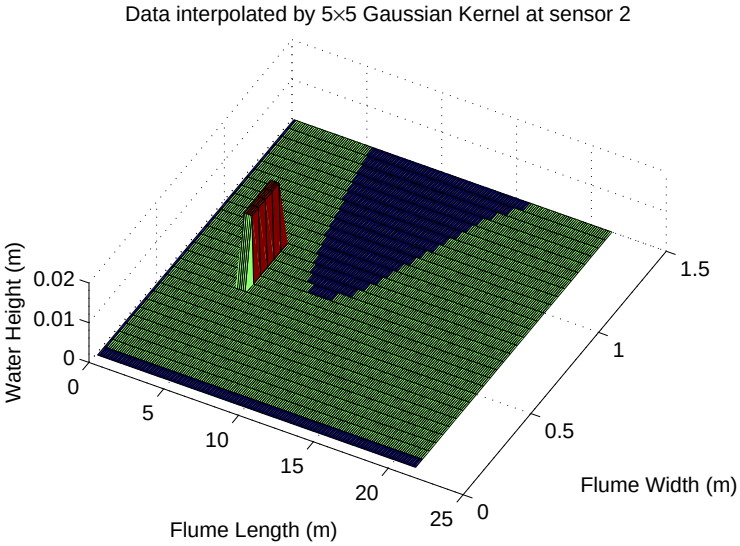


Figure 4.7: Space interpolated water depth at sensor number 2

4.2.3 Shore boundary definition and VEnKF parameters

The settings of the dam break experiment involves changing boundaries (a converging - diverging flume). In the application of VEnKF to the shallow water model, it was not possible to include information about the boundaries in the analysis. Since, in this study, we have a prior knowledge about the shoreline and the places where there is no water, we have used a strategy that allows us to account for the evolving boundaries. We include the information about the boundary in the model error covariance $\mathbf{Q}$. The model error covariance is defined in such a way that the grids which were located in the dry area (riverbank) have given a variance much smaller compared to the variance assigned to other grid points in the domain. This strategy allows the shore boundaries to be maintained by the VEnKF analysis.

The state vector for the dam break experiment is a vector of free surface elevation η, and horizontal velocities u and v in the x and y direction, respectively. We ran VEnKF on the shallow water model with 30×171 grid cells of the simulation domain, and thus the state vector has size approximately equal to 16000. The model error covariance used is $\mathbf{Q} = 0.0011^2\mathbf{I}$ and the observation error covariance is $\mathbf{R} = 0.001^2\mathbf{I}$. Initial state vector $\mathbf{x}_0^{est}$ equals to the initial water height in the flume and the initial covariance estimate $\mathbf{C}_0^{est}$ was set to identity matrix $\mathbf{I}$. The assimilation was conducted using 75 ensemble members and 25 stored vectors for the LBFGS with 25 iterations. With the interpolation done in Section 4.2.2, the number of data obtained is expected to give more reliable results with the VEnKF assimilation scheme.

4.2.4 VEnKF estimates with synthetic data of the dam break experiment

The ability of VEnKF was first examined using synthetic data obtained from the solution obtained by using direct model simulation by Martin and Gorelick (2005). To make the data more realistic, we add normally distributed noise with mean zero and variance 0.05. We compare the results from the 8 locations corresponding to the wave meter positions as given in Martin and Gorelick (2005). Using 50 ensembles, 25 iterations and 25 stored vectors, we compare VEnKF estimates with the data and the model simulation which we referred to here as the truth. VEnKF was used here as a *backtesting* and not for forecasting but the aim is to see whether VEnKF can handle disasters such as dam break especially in downstream locations. For this reason, the length of the forecast is just one computational time step. Figure 4.8 and 4.9 shows that the estimate follows the observations quite very well.

The root mean square error (RMSE) plot for this case is shown in Figure 4.10 for the entire simulation and it shows convergence of the VEnKF.

4.2.5 Experimental and assimilation results for a 1-D set of real observations

VEnKF was used to assimilate measurements of water depth for the dam break experiment published in Martin and Gorelick (2005).

Figure 4.11 shows the snapshots of the water profile of the experiment when using VEnKF at time steps $t = 33$, $t = 77$, $t = 127$ and $t = 302$.

Experimental data from 7 wave meters obtained on the dam break experiment by Bellos et al. (1991) and the simulation results published in Martin and Gorelick (2005) will be compared with the VEnKF assimilation results. Sensor number 7 had no measurements and therefore comparison

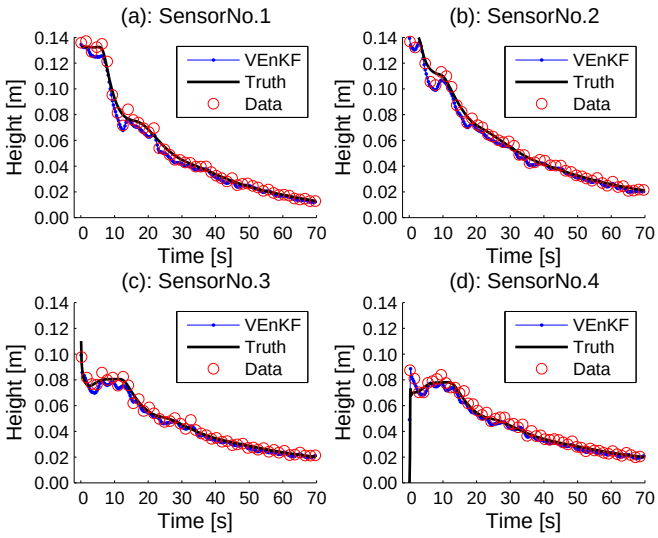


Figure 4.8: Comparison of VEnKF estimates, true water depth and the synthetic data of the dam break experiment for the first four sensors at the upstream end.

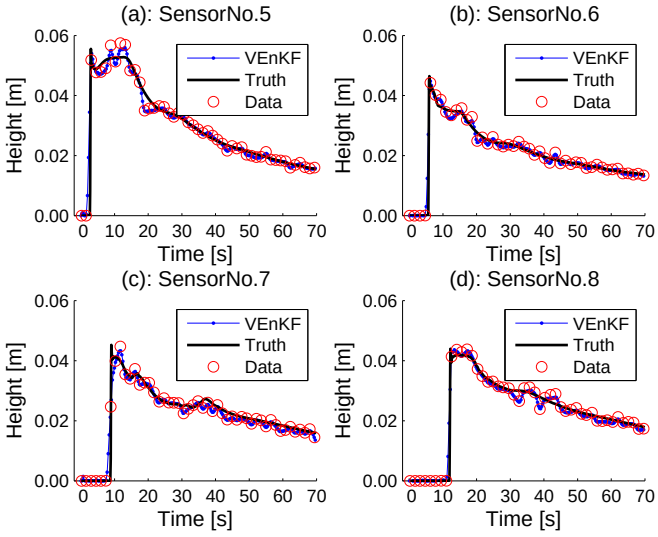


Figure 4.9: Comparison of VEnKF estimates, true water depth and the synthetic data of the dam break experiment for the last four sensors at the downstream end.

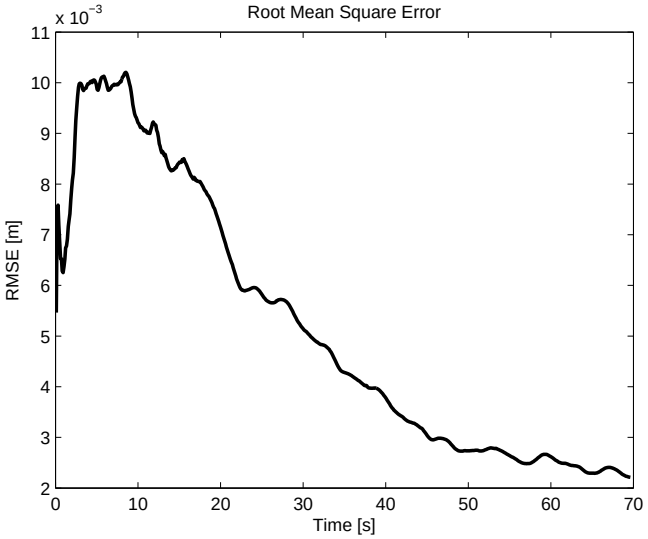


Figure 4.10: The RMSE plot for the entire time interval of assimilation

will be on simulated water depth and VEnKF results only. The initial condition applied is the given initial water depth at the upstream end and initial velocities $U = V = 0$ and the downstream end is dry as shown in Figure. 4.5. Boundary conditions are applied as explained in Section 4.1.4. The location of water/land boundaries is described accordingly by Equations (4.10) and (4.11) respectively (Martin and Gorelick, 2005).

$$H^{N+1}_{1+1/2,j} = max(0, h_{1+1/2}, j + \eta^{N+1}_{i,j}, h_{1+1/2}, j + \eta^{N+1}_{i+1,j}), \tag{4.10}$$

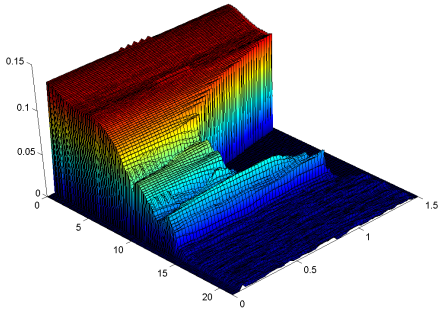
$$H^{N+1}_{i,j+1/2} = max(0, h_{i,j+1/2} + \eta^{N+1}_{i,j}, h_{i,j+1/2} + \eta^{N+1}_{i,j+1}). \tag{4.11}$$

Figures (4.12) and (4.13) show water depth of the 8 sensor locations for the comparison of VEnKF assimilation results with the experimental data and simulated water depth. In Figure 4.13 (c), comparison is between the simulated water height and that of VEnKF as no measurement was not given in this location.

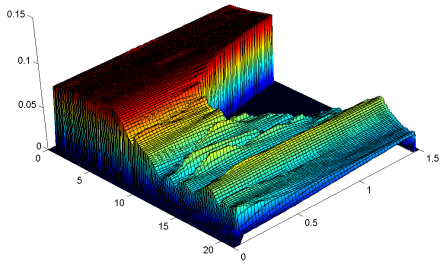
It can be observed that at the upstream end the simulated water depth by Martin and Gorelick (2005) matches well with the measured depth as it can be seen in Figure 4.12a-c. The turbulent behavior of water at the downstream end shown by the experimental data can not be observed on the graphs for the pure simulation. On the other hand, VEnKF results not only match with the measured depth but they also model the turbulent structure of the flow at the downstream end which is characterized by a super critical flow as it can be observed in Figure 4.13.

4.2.6 Spread of ensemble forecast

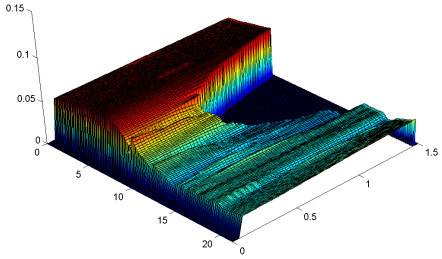
In several studies, the measure of forecast uncertainty has been done using ensemble spread of short range ensemble forecasts (Moradkhani et al., 2005a). Estimates of uncertainty aim at measuring the reliability of the model forecast at a given probability range. On the other hand, the ensemble spread is used as a measure of goodness of fit and can be used to represent the estimate of uncertainty (Xie



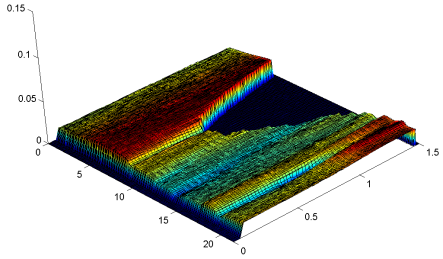
(a) $t = 33$



(b) $t = 77$



(c) $t = 127$



(d) $t = 302$

Figure 4.11: Water profile at different time steps of the assimilation.

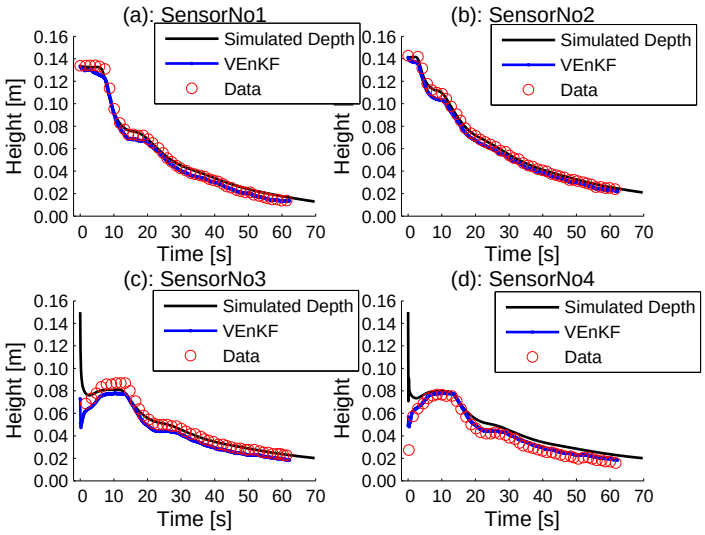


Figure 4.12: Water depth for the first four wave meters in the dam break flume at the upstream end.

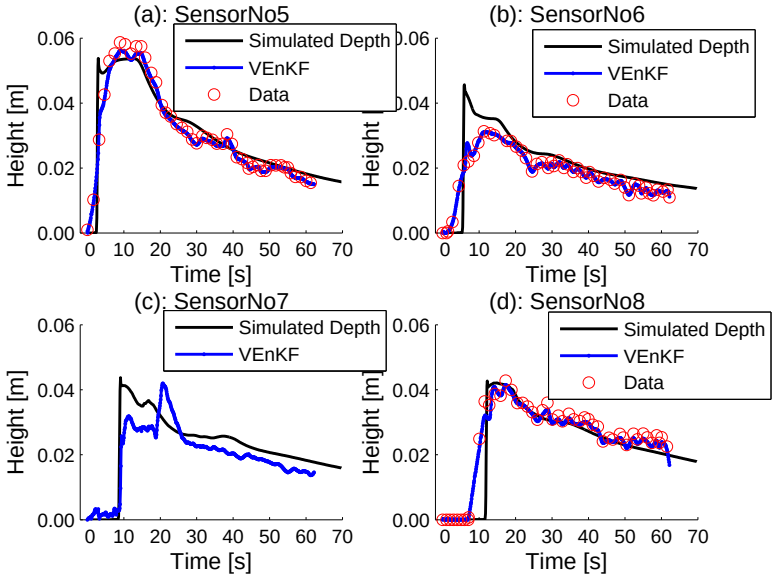


Figure 4.13: Water depth for the last four sensors in the dam break flume at the downstream end.

and Zhang, 2010). Ensemble spread can be increased by adjusting parameters of the model as it has been shown for example in hydrological data assimilation using a recursive ensemble Kalman filter by McMillan et al. (2013) that, increasing the water table parameters also increases the spread.

In this study, we also checked the performance of the VEnKF by considering the spread of the ensembles at the 95% confidence interval. As it can be observed in Figures 4.14 - 4.18, they illustrate ensemble spread at different sensor locations, the VEnKF ensemble occasionally has a tendency to diverge. In some locations and times, the ensemble divergence is seen as a spurious blow-up of ensemble spread. In other times and sensor locations, the entire ensemble appears to drift away from the trajectory that connects observations. The causes for this ensemble divergence are not very clear, but one possible candidate is the stochastic spatial extension of the observations that may cause local violations of the CFL condition in the area of observation extension. It is remarkable, however, that in no case does the VEnKF filter diverge. The analysis always stays close to the observations, even in cases when the entire ensemble diverts away from them.

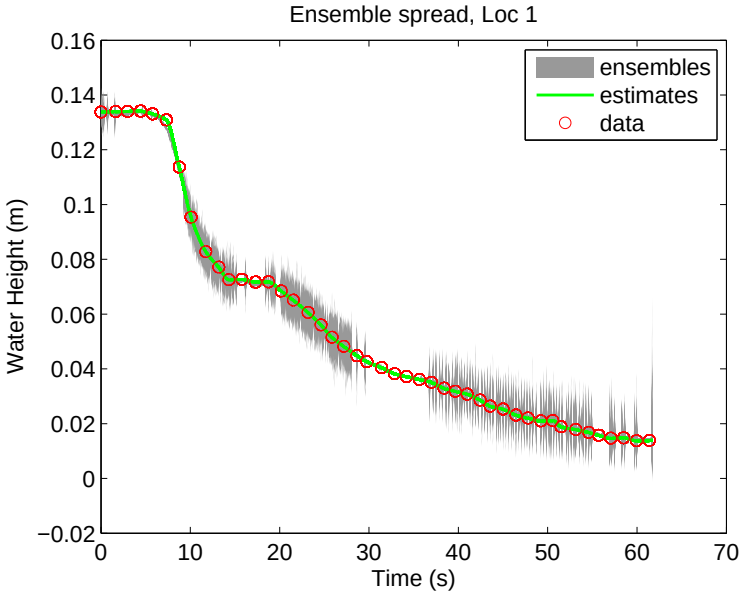


Figure 4.14: Ensemble spread at the 95% confidence interval of measurement location 1.

4.3 Ability of VEnKF analysis to represent two dimensional flow

4.3.1 2D observation settings

VEnKF was again tested with a 2D dam break problem. The same dam break experiment of Bellos et al. (1991) is tested here with new modifications. The observation locations at the downstream end were left unchanged as published in Martin and Gorelick (2005). We introduced parallel wave meters at the downstream end along the flume mid-line. As we know that a river flow comprises both cross flow and streamline flow, the aim of this setup is to examine whether VEnKF can be able to predict cross flow which is not identifiable with only a single line of sensors positioned along

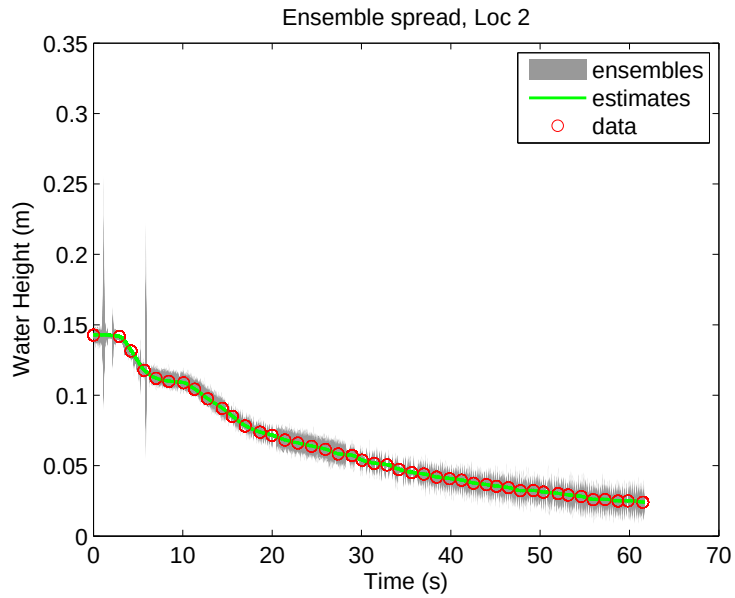


Figure 4.15: Ensemble spread at the 95% confidence interval of measurement location 2.

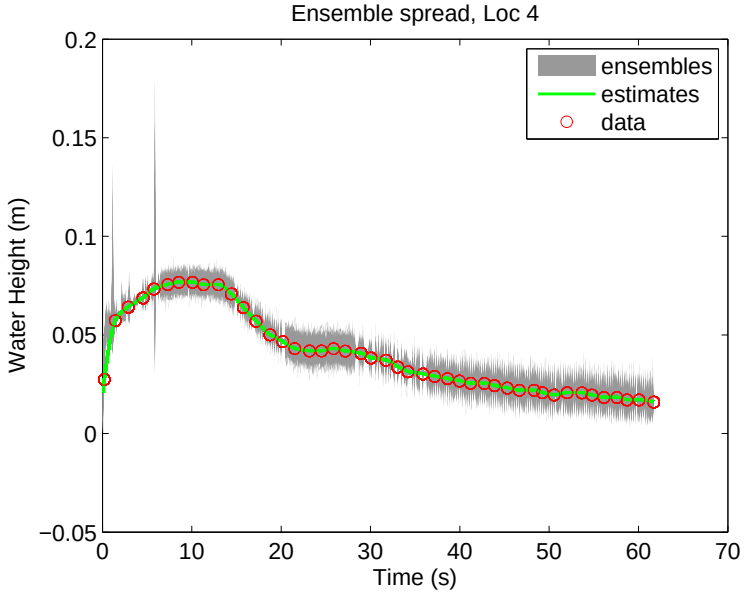


Figure 4.16: Ensemble spread at the 95% confidence interval of measurement location 4.

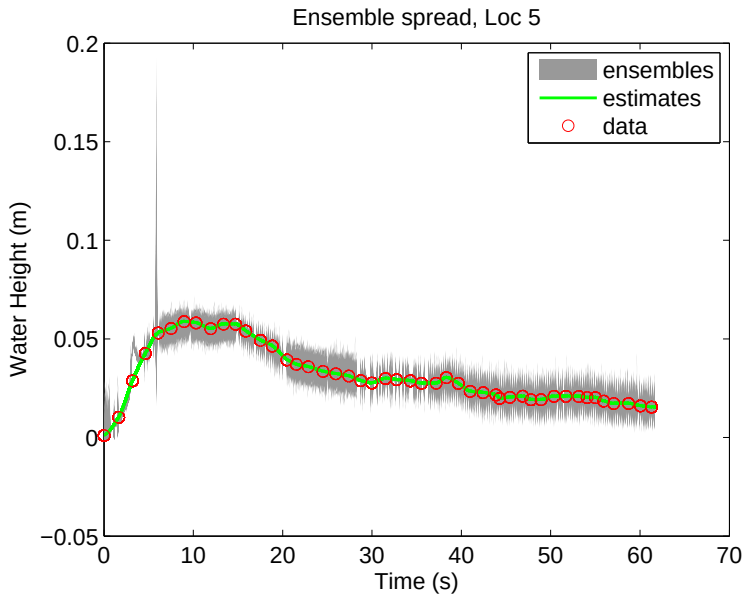


Figure 4.17: Ensemble spread at the 95% confidence interval of measurement location 5.

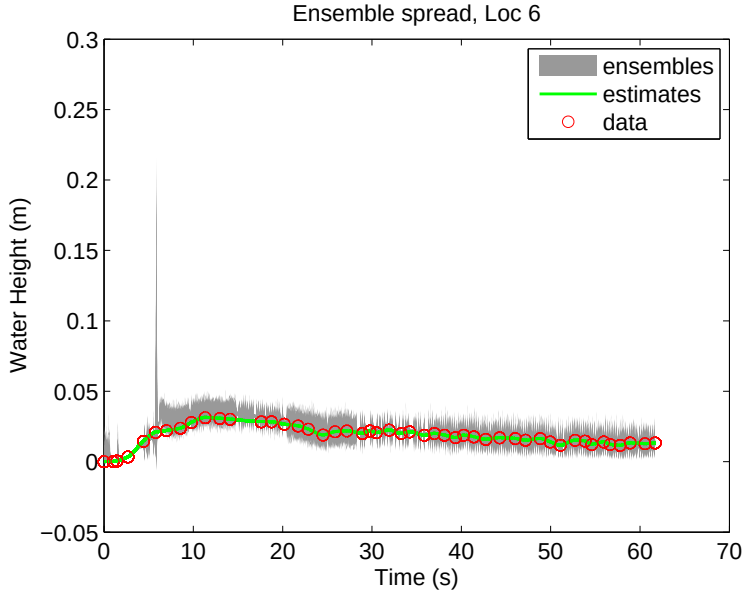


Figure 4.18: Ensemble spread at the 95% confidence interval of measurement location 6.

the flow mid-line. To accomplish this goal, the meters were placed in the same original position in the y-direction but pushed left and right from the flume mid-line by $4\Delta x$. This makes a total of 8 meters at the downstream end with the new position along the x-direction as $x' = x \pm 4\Delta x$ and $y' = y$ along the y-direction. From this new setting of the wave meters, we first assume that there is a cross flow along the flume and then we superimpose a sinusoidal wave across the flow on the true experimental observations. We have chosen the sine wave in such a way that the observations can not drop to zero during the time of assimilation. We also add to the new observations random noise which is normally distributed with mean 0 and standard deviation 0.001. Figure 4.19 shows this new setting of wave meters whereby other dimensions remain the same as in the one dimensional setting of observations.

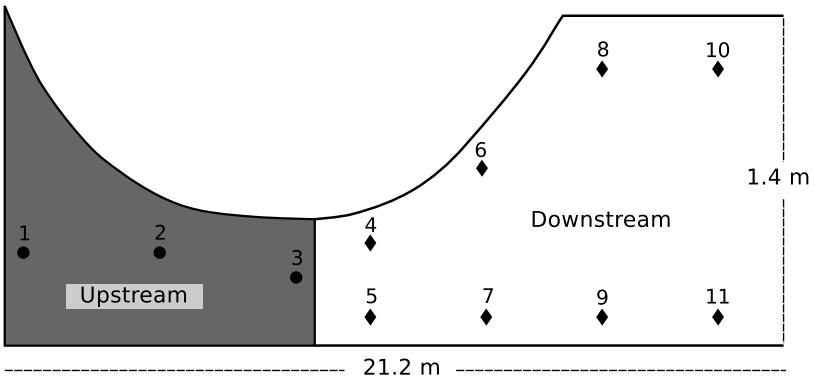


Figure 4.19: Parallel setup of wave meters at the downstream end.

The new data set obtained in this new setting of wave meters was again interpolated in time and in space as explained in section 4.2.2 using a square patch of 5×5.

4.3.2 Results with parallel setup of observations

Figures 4.20, 4.21, and 4.22 show the results of VEnKF when applied to the dam break problem with two rows of observations at the downstream end. It can easily be observed that there is no cross flow detected at the upstream end as shown by figure 4.20. Moreover, we can see a reasonable balance between measurements and the VEnKF analysis. VEnKF has been able to capture cross flow as can be observed by the presence of sinusoidal oscillations in the down stream end, as shown by Figures 4.21 and 4.22.

4.3.3 Impact of observation Interpolation with VEnKF

As mentioned in Section 4.2.2, given the sparse observations in 7 observation locations as published in Martin and Gorelick (2005), it was a challenge for data assimilation since the amount of data received at the time of assimilation was not enough to expose bias in the prediction model. Hence interpolation was necessary in terms of time and space. The aim here is to study the relationship between the time interpolation distance of observations and the ensemble variance.

When observations are interpolated so as to be captured at every time step or less frequently, we observe that the VEnKF algorithm always stays numerically stable, however, with long time intervals

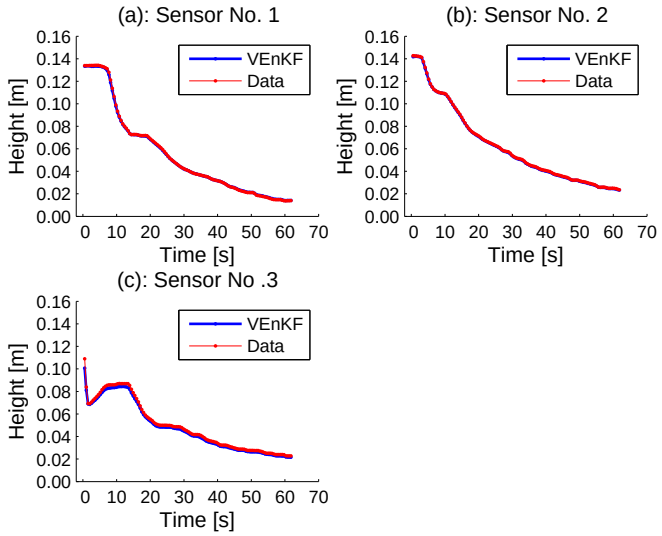


Figure 4.20: Upstream meters: no cross flows recorded by the VEnKF as was expected.

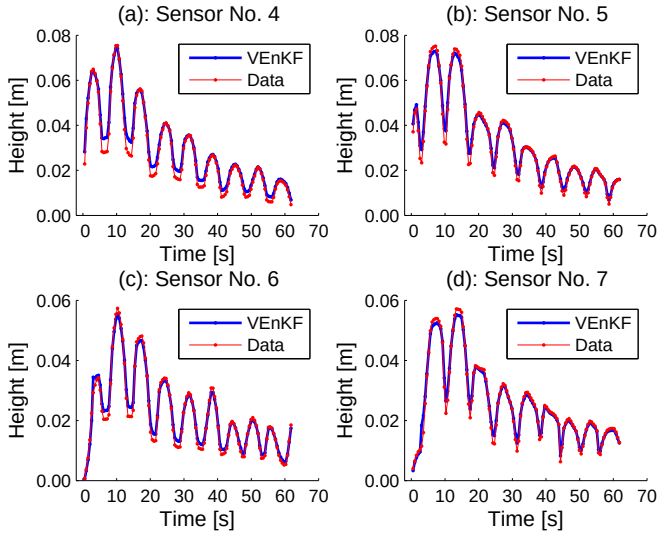


Figure 4.21: The VEnKF captures well the cross flows for the downstream locations.

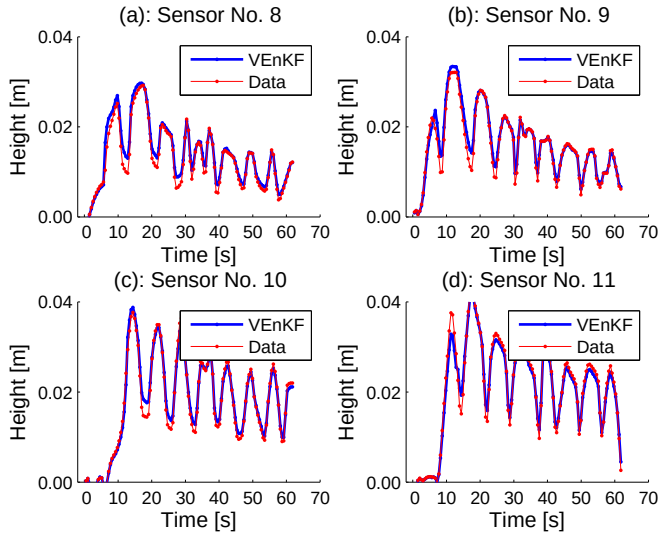


Figure 4.22: The VEnKF captures well the cross flows for the downstream locations.

between observations, the analysis fails to capture the waves present in the solution. Let us examine this behavior by considering the flow diagrams of sensor number 4 at different ensemble variances.

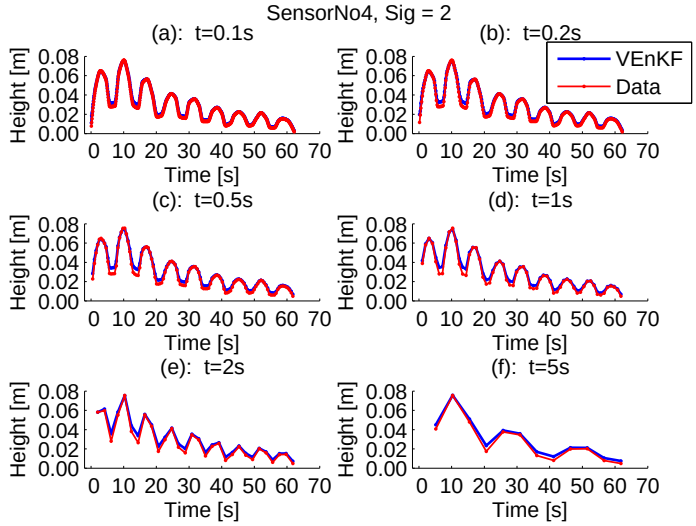


Figure 4.23: Results showing VEnKF converges to the true measurements with all observation intervals if ensemble variance is sufficient ($\sigma^2 = 4$). Note the aliasing of the sine wave to a lower frequency wave when the observation interval exceeds the wave frequency at time step 5s and the estimation problem violates the Nyquist limit. The filter then converges to the aliased solution.

From Figure 4.23 we can observe that the analysis converges to the observed measurements with ensemble variance $\sigma^2 = 4$. However, reducing the ensemble variance causes the filter to diverge slightly as in Figure 4.24 and when the ensemble variance is too small, the analysis diverges from the true solution as shown in Figure 4.25. We also found that there is a relationship between the

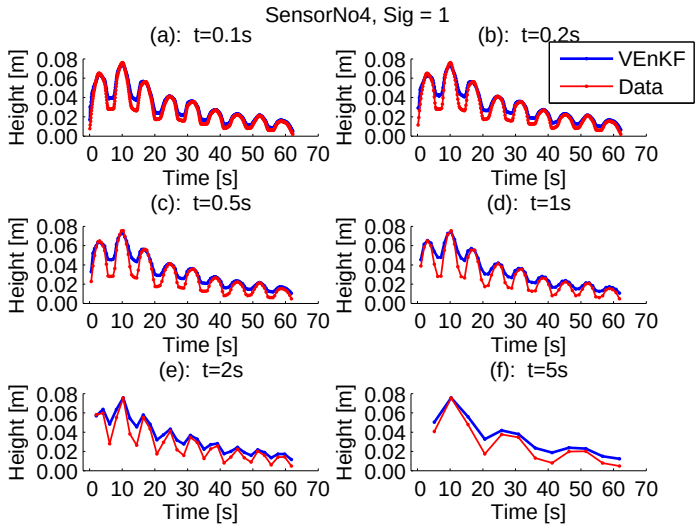


Figure 4.24: Border-line filter divergence with different observation intervals and border-line ensemble variance $\sigma^2 = 1$.

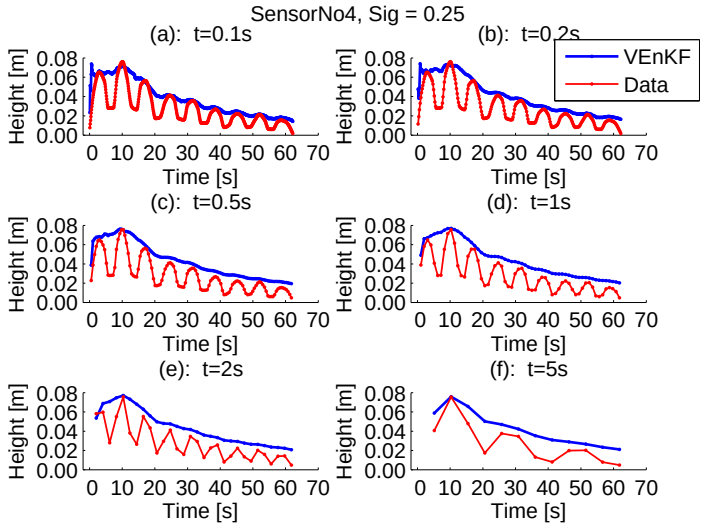


Figure 4.25: Results showing VEnKF divergence at all observation intervals with excessively small ensemble variance ($\sigma^2 = 0.625$). The solutions remain numerically stable in all cases.

time interpolation of observations and ensemble variances when studied at several range of values. We used the difference in Euclidean norm between the analysis and the true solution at different values of Δt and ensemble variance σ. Figure 4.26 shows level curves of this relationship when Δt and σ are plotted in logarithmic scale. It can easily be observed that the level curves are almost linear and if we study the slope of the level curves with respect to $\log(\sigma)$ and $\log(\Delta t)$ there is a linear relationship between the ensemble spread and time interpolation distance. This relationship is defined by a power law that guarantees the filter convergence of the form

$$\Delta t \approx \sigma^6,$$

or

$$\Delta t \approx var^3,$$

where σ is the standard deviation of the ensemble and $var = \sigma^2$ is the ensemble variance.

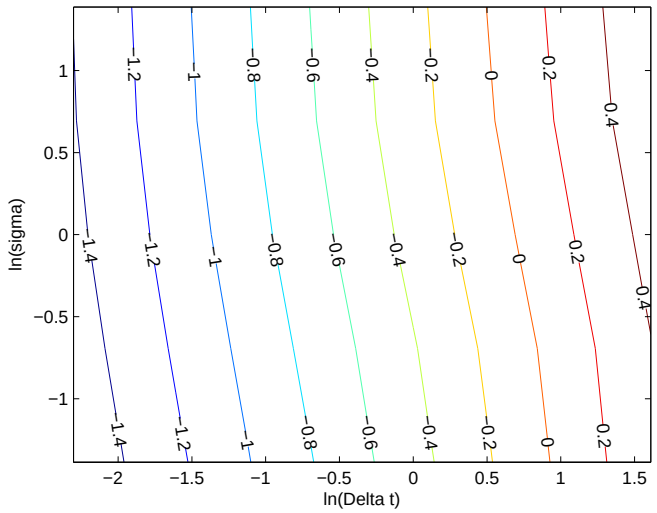


Figure 4.26: Empirical level curves for the difference in Euclidean norm between the analysis and true solution as a function of the logarithm of the observation interval Δt and the logarithm of ensemble standard deviation σ.

4.4 Mass conservation of VEnKF analyses

Different approaches have been used to solve the problem of mass conservation when using data assimilation techniques to estimate the state of the system. In a recent study to improve the spacial mean of a simulated soil moisture field by Li et al. (2012), the loss of water mass has been solved by the use of a mass conservation scheme, the conservative ensemble Kalman filter. The scheme use a correction term which guarantees that, the total water storage remains the same for each ensemble member after the ensemble update.

Data assimilation with VEnKF, where the ensemble is frequently re-sampled, suffers from lack of mass and entropy conservation. The primary numerical methods used in CFD are normally built to approximately conserve at least mass, but the least squares approximation implicit in data

assimilation is likely to reduce both mass and entropy. In Amour et al. (2013), the problem of mass conservation can easily be observed by considering Figure 4.27 which shows the fraction of the remaining mass after assimilation in comparison with that of pure simulation.

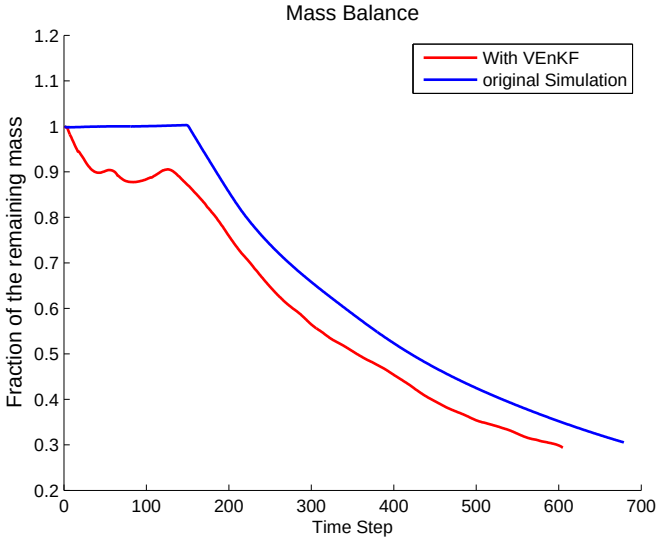


Figure 4.27: Fraction of the remaining mass for the dam break experiment with VEnKF and with the original simulation by Martin and Gorelick (2005).

When VEnKF was applied to the dam break experiment using Algorithm 3.6, the assimilated water height may not have the same mean as the model estimates. This is due to the fact that in some cases the updates generate negative water heights and if these values are replaced by zeros, they lead to a problem in mass conservation. Thus, the decrease of the posterior total mass relative to the prior total mass needs to be considered at the analysis stage. Table 4.1 shows the prior total mass relative to the posterior total mass for the first 10 time steps and it can easily be observed that there is a decrease of the prior total mass and hence underestimation of the posterior height fields.

Table 4.1: Prior total mass relative to the posterior total mass for the first 10 time steps

Time steps	Prior total mass	Posterior total mass
1	227.7522	224.0027
2	227.4255	218.7961
3	227.2985	210.5419
4	227.2594	204.6319
5	227.2625	201.0808
6	227.2753	198.5315
7	227.2577	196.7373
8	227.2705	195.3819
9	227.2931	194.4817
10	227.3066	194.2232

The modified VEnKF uses a correction term at the analysis step. We begin by calculating the

standard deviation σ of all the original observations as measured by the wave meters. Define the prior covariance matrix $\mathbf{C}_k^p$ as per Equation 3.18. Then forecast the prior and propagate the ensemble forward as per step (ii) of the VEnKF algorithm and then calculate the prior total mass m. In the analysis step, we apply LBFGS optimization to minimize the cost function (3.11) so as to get the posterior water height as the minimizer of the of the cost function (3.11) and the error covariance matrix $\mathbf{C}_k^{est}$ as the inverse of the Hessian of (3.11). We then calculate the posterior total mass m' by integrating total water height over the whole domain. The correction term to be added to the posterior water height is a value ε sampled from a normal distribution with mean d and standard deviation σ, where $d = (m - m')/number\ of\ gridpoints$. Thus, the posterior water height is now given by:

$$h^* = \mathbf{x}_k^{\text{est}} + \varepsilon \tag{4.12}$$

The modified version of the VEnKF for our dam break case is summarized in Algorithm 4.1.

Algorithm 4.1 The mass conservation VEnKF Algorithm

(i) Initialize the initial guess $\mathbf{x}_0^p$, $\mathbf{s}_0^p$ and set $k = 1$.

(ii) Calculate the standard deviation σ of all the original observations by the wave meters from the corresponding original model forecasts.

(iii) Forecast the prior with the numerical model one time step ahead:

  (a) Calculate prior total mass by integrating water height h over the whole domain.

(iv) Conduct the assimilation step minimization to arrive at posterior water height $\mathbf{x}_k^{\text{est}}$.

  (a) Apply LBFGS optimization to minimize Equation (3.11).

  (b) Calculate the posterior total mass m' by integrating water height equal to the minimizer of the cost function (3.11) and find mean mass difference of the posterior total mass from prior total mass, dividing it by the number of grid points, $d = (m - m')/number\ of\ gridpoints$.

  (c) Sample a correction term ε at every grid point from the normal distribution $\varepsilon \sim \mathcal{N}(d, \sigma)$ to arrive at the approximately mass and entropy conserving water height field h^*

  (d) Add this normally distributed correction term ε to the posterior height $h^* = \mathbf{x}_k^{\text{est}} + \varepsilon$

  (e) Update the ensemble $\mathbf{s}_{k,i}^{\text{est}}$ by sampling from $s_{k,i} \sim \mathcal{N}(h^*, \mathbf{C}_k^{est})$.

(v) Set $k \to k+1$ and go to step (iii).

With the new approach of using a correction term at the analysis step, the posterior water height field h^* have roughly the same mean and variance, respectively, as the prior height field and the true flow, hopefully approximately conserving both mass and entropy. Since the correction is done very frequently, the mean and standard deviation are likely to be small, but still positive, and not hopefully cause numerical instability. Table 4.2 shows the prior total mass relative to the posterior total mass for the first 10 time steps and it can easily be observed that there is a relative small difference of the prior total to the posterior total mass. With the use of mass conservative VEnKF for the dam break

Table 4.2: Prior total mass relative to the posterior total mass for the first 10 time steps using the mass conservative VEnKF

Time steps	Prior total mass	Posterior total mass
1	227.7522	227.6994
2	227.4255	227.5515
3	227.2985	227.3699
4	227.2594	227.3018
5	227.2625	227.2613
6	227.2753	227.3635
7	227.2577	227.3040
8	227.2705	227.3100
9	227.2931	227.2320
10	227.3066	227.2920

experiment we can show that mass is conserved by looking at Figure 4.28 which shows the fraction of the remaining mass after assimilation in comparison with that of pure simulation.

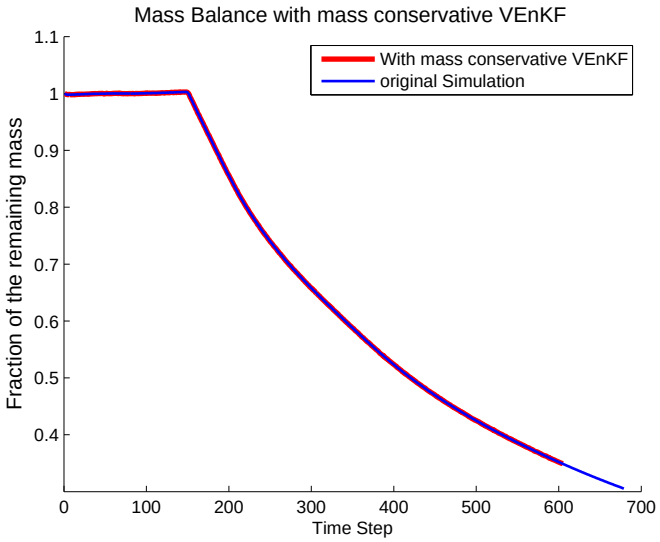


Figure 4.28: Fraction of the remaining mass for the dam break experiment with mass conservative VEnKF and with the original simulation by Martin and Gorelick (2005).

4.5 The two layer Quasi-Geostrophic model

The Quasi-Gestrophic (QG) model (Ikeda, 1981; Pedlosky, 1987) is an example of chaotic dynamics which can be run at a large scale setting with reasonable computational cost. Many research have been done on oceanic wind circulation using the QG model (Ikeda, 1981; Pedlosky, 1987; Medjo, 2000). This model has also been used to simulate flat double-layered geostrophic (slow) wind motion (Pedlosky, 1987).

The geometric layout of the 2-layer QG model is as shown in figure 4.29. The two atmospheric layers are lying one at the top and the other at the bottom of a cylindrical surface. $\bar{U}_1$ and $\bar{U}_2$

indicate the mean zonal wind speeds in the top and bottom layer respectively.

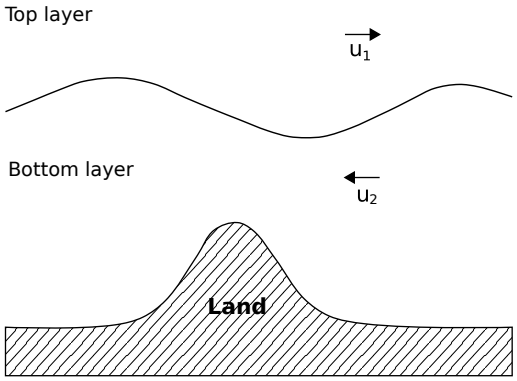


Figure 4.29: Geometric layout of the QG-model.

The formulation of the QG model is governed by the stream function and the potential vorticity. The relationship between these two components is given by the following equations:

$$q_1 = \nabla^2 \psi_1 - F_1 (\psi_1 - \psi_2) + \beta y, \tag{4.13}$$

$$q_2 = \nabla^2 \psi_2 - F_2 (\psi_2 - \psi_1) + \beta y + R_s, \tag{4.14}$$

where q_1 and q_2 are the top and bottom layer potential vorticities respectively and ψ_i are the stream functions.

The non-physical terms used to define the two layer QG model are described as follows:

Parameter	Description	Formula
F_i	layer interaction parameters	$F_i = \frac{f_0^2 L^2}{g \mathscr{D}_i}$
β	north ward gradient of the coriolis parameter	$\beta = \beta_0 \frac{L}{U}$
R_s	two-dimensional orography surface	$R_s = \frac{S(x,y)}{\eta \mathscr{D}_2}$
g	acceleration due to gravity	$g = \acute{g} \frac{\Delta\theta}{\bar{\theta}}$
$\mathscr{D}_i$	undisturbed depth of the corresponding model layer	
$\Delta\theta$	the temperature change across the layer interface	
$\bar{\theta}$	the mean potential temperature	
f_0	the coriolis parameter	
η	the Rossby number	$\eta = \frac{U}{f_0 L}$
$S(x,y)$	the orography term	
L and U	the main length and velocity scale respectively	

The stream function ψ is related to the geostrophic wind, the zonal wind u_i and meridional wind v_i by the following dependency:

$$\big(u_i, v_i\big) = \big(-\frac{\partial \psi_i}{\partial y}, \frac{\partial \psi_i}{\partial x}\big). \tag{4.15}$$

The QG model is assumed to obey the law of potential vorticity conservation,

$$\frac{D_1 q_1}{Dt} = \frac{D_2 q_2}{Dt} = 0 \tag{4.16}$$

where $\frac{D_i \cdot}{Dt} = \frac{\partial \cdot}{\partial t} + u_i \frac{\partial \cdot}{\partial x} + v_i \frac{\partial \cdot}{\partial y}$ denotes the material derivative. Equations (4.13), (4.14), (4.15) and (4.16) are the governing equations for the QG model.

4.5.1 Numerical approximation and VEnKF results

In our experiment, Equations (4.13)-(4.16) are integrated using a semi-Lagrangian approach (refer for example to Staniforth and Côté (1991)) using a finite-difference scheme. This numerical method is based on the core ideas of solving the QG-model equations explained by Fandry and Leslie (1984).

The test runs employ the VEnKF algorithm applied on top of the QG-model. More precisely, the model is instantiated twice in a twin experiment, where the first case that we call the truth run simulates the "nature" and is used to generate observations and the second case that we call the biased run runs with different layer depths and is leveraged as a prediction model. Both model instances were run at a dimension of 40-by-20 grid nodes in each layer thus having 1600 degrees of freedom. The layer depths used in the truth run were 6000m for the top layer and 4000m for the bottom layer with spatial discretization steps $\Delta x = \Delta y = 300km$ and time discretization of $6min$. Thus, data were collected at every 6min. In the biased run the layer depths were set to 5500m and 4500m, respectively. The rest of parameters were the same in both runs. The observations extracted from the truth run were perturbed by normally distributed zero-mean noise with standard deviation equal to 0.1. In addition, prior to starting the actual data assimilation, the truth and the biased runs were simulated for two weeks of the model time. This was done to establish divergence between the initial estimate of the VEnKF and the first bundle of observations.

Other parameters include the observation error covariance $\mathbf{R}$ which was set to $0.1\mathbf{I}$ whereby $\mathbf{I}$ is 800×800 identity matrix. The model error covariance is defined as

$$\mathbf{Q} = \begin{pmatrix} 0.2\mathbf{I} & 0.5\mathbf{I} \\ 0.5\mathbf{I} & 0.2\mathbf{I} \end{pmatrix},$$

The experiment was run by varying ensemble size and 50 iterations for the LBFGS optimization and the number of stored vectors was set to 50.

The dimension of the problem in the described numerical experiments was still small enough to allows the use of the EKF. Therefore, we compare the performance of the VEnKF algorithm with that of classical EKF in terms of the RMSE. Figure 4.30 shows the RMSE of EKF and that of VEnKF estimates at different ensemble sizes. It can easily be observed that the RMSE converge with 20 ensemble members, however, a larger ensemble size leads to more stable results.

Figure 4.31 contains the forecast skill curves for the VEnKF executed at different ensemble cardinalities as well as for the EKF. The forecast skill shows that when the VEnKF stabilizes (i.e. starting from 50 ensemble members) the effective forecast range stays at the same 7 days mark regardless of ensemble size growth. Expectedly, the EKF performs best, providing about 1 day longer range of effective forecast than VEnKF.

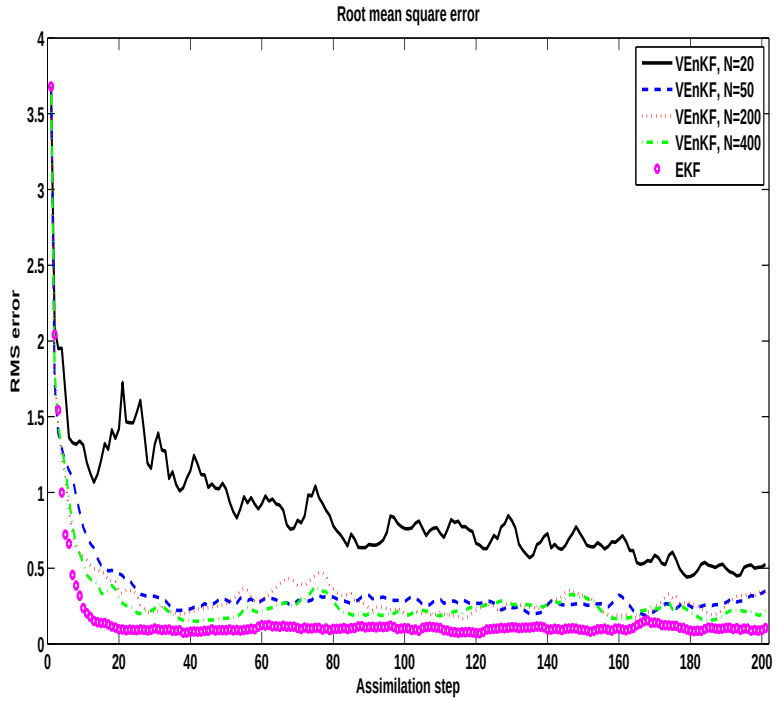


Figure 4.30: Root mean square error of the estimates in the QG-model when using EKF and VEnKF

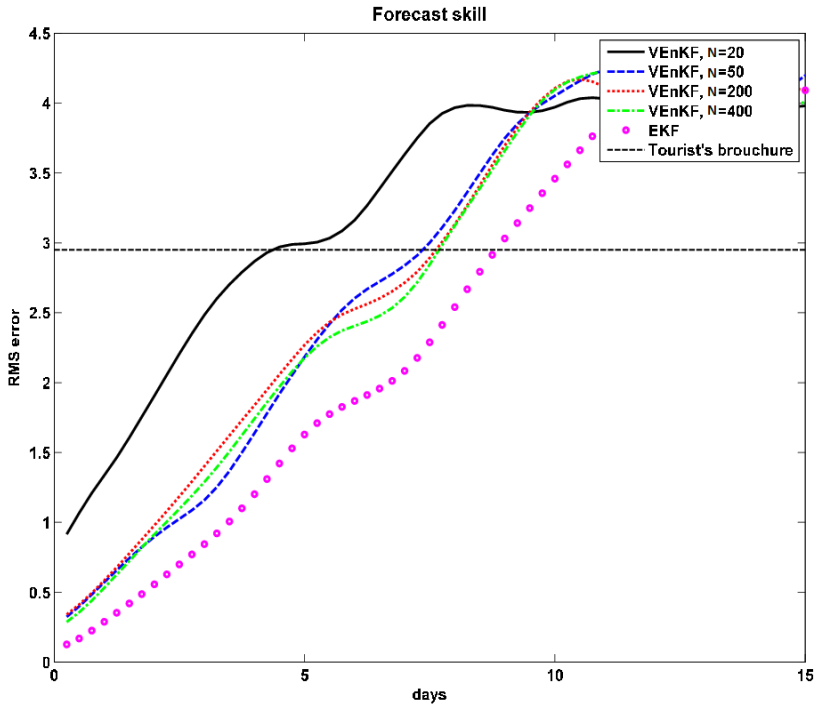


Figure 4.31: Forecast skill of VEnKF at different ensemble sizes and that of EKF on the QG-model

CHAPTER V

Discussion and Conclusions

The use of data assimilation in hydrological forecasting is becoming very popular due to the availability of various measurement devices that provide measurements of a given system and henceforth, hydrological models are combined with these real time measurements to estimate the state of a given system. The main goal of data assimilation is the quantification of uncertainty in the estimates, with low computational cost. Thus, efforts have been made by researchers to find different techniques in data assimilation that can be used in highly non-linear problems with high dimensional state. In various studies in the literature, it has been shown that the use of data assimilation techniques outperforms pure computer simulation using numerical schemes.

A new hybrid data assimilation method introduced by Solonen et al. (2012) called the variational ensemble Kalman filter was found useful for highly non-linear problems and in large dimensional applications. VEnKF combine an ensemble Kalman filter and variational data assimilation and it is easy to apply as it only uses the forward model. VEnKF does not require the construction and use of tangent linear and adjoint codes, as other variational methods like 3D-Var and 4D-Var do. In the present work, VEnKF was applied to a real data assimilation problem, the dam break problem in three different ways.

i. In the first case, VEnKF was applied to a dam break problem using a shallow water model consisting of measurements from a laboratory experiment. In this application, a one-dimensional set of observations which was sparse in terms of space and time was considered. This study provides an effective way of dealing with these sparse observations so as to allow the application of data assimilation. In terms of time, only a few sensors were able to produce measurements and these time instances had no alignment with the model integration time step. To address this, time interpolation was used. In terms of space, only a few data points out of the total set of grid points were available. It has been found that, in assimilating observations of the dam break experiment, VEnKF updates well the depth of water and it was able to reproduce the turbulent behavior of the flow, which could not be observed in pure simulation. The results obtained from this application were found to outperform the results from pure simulation.

ii. In the second application, VEnKF was applied to a modified laboratory dam break experiment by considering observations in a two-dimensional setting at the downstream end. In this second application VEnKF was able to learn and predict the cross flow as well as stream flow.

However, it was observed that cross flow was only achieved when the observations are in a two dimensional setting. In this application, also a study of convergence of VEnKF as a function of time interpolation distance and ensemble variance has been carried out.

iii. The application of VEnKF to the dam break problem and in its general application suffers from the problem of mass conservation due to frequent re-sampling of the ensemble and generation of random noise may sometimes be physically unrealistic. Thus, a new VEnKF algorithm referred to as the mass conserving VEnKF has been established. Using this new VEnKF algorithm, it was observed that, it is possible to account for the loss of total water mass.

iv. Lastly, VEnKF was applied in a two dimensional geophysical flow using a Quasi-Geostrophic (QG) model. In this experiment, synthetic measurements were used and the results obtained are comparable with the classical extended Kalman filter with increasing ensemble size.

The results obtained from all these applications indicate that VEnKF is a good candidate for data assimilation problems and can be applied in high dimensional non-linear models.

Agoshkov, V., Quarteroni, A., Saleri, F., 1994. Recent developments in the numerical simulation of shallow water equations I: boundary conditions. Applied Numerical Mathematics. 15 (2), 175–200.

Amour, I., Mussa, Z., Bibov, A., Kauranne, T., 2013. Using ensemble data assimilation to forecast hydrological flumes. Nonlinear Processes in Geophysics 20 (6), 955–964.

Anderson, J. L., 2001. An ensemble adjustment Kalman filter for data assimilation. Monthly weather review 129 (12), 2884–2903.

Anderson, J. L., 2012. Localization and sampling error correction in ensemble Kalman filter data assimilation. Monthly Weather Review 140 (7), 2359–2371.

Anderson, J. L., Anderson, S. L., 1999. A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. Monthly Weather Review 127 (12), 2741–2758.

Auvinen, H., Bardsley, J., Haario, H., Kauranne, T., 2010. The variational Kalman filter and an efficient implementation using limited memory BFGS. International Journal for Numerical Methods in Fluids 64 (3), 314–335.

Auvinen, H., Bardsley, J. M., Haario, H., Kauranne, T., 2009. Large-Scale Kalman Filter Using Limited Memory BFGS. Electronic Transaction on Numerical Analysis. 35, 217–233.

Awaji, T., Masuda, S., Ishikawa, Y., Sugiura, N., Toyoda, T., Nakamura, T., 2003. State estimation of the North Pacific Ocean by a four-dimensional variational data assimilation experiment. Journal of oceanography 59 (6), 931–943.

Bélanger, E., Vincent, A., 2004. Data assimilation (4D-VAR) to forecast flood in shallow-waters with sediment erosion. Journal of Hydrology 300, 114–125, doi:10.1016/j.jhydrol.2004.06.009.

Bellos, C., 2004. Experimental measurements of flood wave created by a dam break. European Water 7 (8), 3–15.

Bellos, C., Soulis, J., Sakkas, J., 1991. Computation of two-dimensional dam-break induced flows. Advances in Water Resources 14, 31–41.

Bertino, L., Evensen, G., Wackernagel, H., 2003. Sequential data assimilation techniques in oceanography. International Statistical Review 71 (2), 223–241.

Biscarini, C., Francesco, S. D., Manciola, P., 2010. CFD modelling approach for dam break flow studies. Hydrology and Earth System Sciences 14 (4), 705–718.

Blum, J., Le Dimet, F.-X., Navon, I. M., et al., 2009. Data assimilation for geophysical fluids. Handbook of Numerical Analysis: Computational Methods for the Atmosphere and the Oceans 14, 385–441.

Buehner, M., Morneau, J., Charette, C., 2013. Four-dimensional ensemble-variational data assimilation for global deterministic weather prediction. Nonlinear Processes in Geophysics 20 (5), 669–682.

Cacuci, D. G., Navon, I. M., Ionescu-Bujor, M., 2013. Computational methods for data evaluation and assimilation. CRC Press.

Casulli, V., Cheng, R., 1992. Semi-implicit finite difference methods for three-dimensional shallow water flow. International Journal for Numerical Methods in Fluid 15 (6), 629–648.

Chang, T.-J., Kao, H.-M., Chang, K.-H., Hsu, M.-H., 2011. Numerical simulation of shallow-water dam break flows in open channels using smoothed particle hydrodynamics. Journal of Hydrology 408 (1), 78–90.

Chen, X., Navon, I., 2009. Optimal control of a finite-element limited-area shallow-water equations model. Studies in Informatics and Control 18 (1), 41–62.

Chow, S.-M., Ferrer, E., Nesselroade, J. R., 2007. An unscented Kalman filter approach to the estimation of nonlinear dynamical systems models. Multivariate Behavioral Research 42 (2), 283–321.

Ciarlet, P. G., Temam, R., Tribbia, J., 2009. Computational methods for the atmosphere and the oceans: special volume. Vol. 14. Elsevier.

Courtier, P., Talagrand, O., 1990. Variational assimilation of meteorological observations with the direct and adjoint shallow-water equations. Tellus A 42 (5), 531–549.

Daley, R., 1991. Atmospheric data analysis. Cambridge university press.

Doucet, A., Godsill, S., Andrieu, C., 2000. On sequential Monte Carlo sampling methods for Bayesian filtering. Statistics and computing 10 (3), 197–208.

Durran, D. R., 2010. Numerical methods for fluid dynamics: With applications to geophysics. Vol. 32. Springer.

Evensen, G., 1994. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte-Carlo methods to forecast error statistics. Jounal of Geophysical Research 99 (C5), 10143–10162.

Evensen, G., 2003. The Ensemble Kalman Filter: theoretical formulation and practicle implementation. Ocean Dynamics 53, 343–367, dOI: 10.1007/s10236-003-0036-9.

Evensen, G., 2009. Data Assimilation: The Ensemble Kalman Filter. Jonh Wiley and Sons.

Fandry, C., Leslie, L., 1984. A Two-Layer Quasi-Geostrophic Model of Summer Trough Formation in the Australian Subtropical Easterlies. Journal of the Atmospheric Sciences 41, 807–818.

Fisher, M., Nocedal, J., Trémolet, Y., Wright, S. J., 2009. Data assimilation in weather forecasting: a case study in PDE-constrained optimization. Optimization and Engineering 10 (3), 409–426.

Frei, M., Künsch, H. R., 2013. Bridging the ensemble Kalman and particle filters. Biometrika 100 (4), 781–800.

Fritsch, F., Carlson, R., 27, March 1980. Monotone Piecewise Cubic Interpolation. SIAM Journal on Numerical Analysis 17 (2), 238–246.

Fulton, S. R., 2004. Semi-implicit time differencing. Tech. Rep. No. 2002-01, Department of Mathematics and Computer Science, Clarkson University.

Ghil, M., Malanotte-Rizzoli, P., 1991. Data assimilation in meteorology and oceanography. Advances in geophysics 33, 141–266.

Gillijns, S., Mendoza, O. B., Chandrasekar, J., De Moor, B., Bernstein, D., Ridley, A., 2006. What is the ensemble Kalman filter and how well does it work? In: American Control Conference, 2006. IEEE, pp. 6–pp.

Givoli, D., Neta, B., 2003. High-order nonreflecting boundary conditions for the dispersive shallow water equations. Journal of Computational and Applied Mathematics 158 (1), 49–60.

Grewal, M. S., Andrews, A. P., 2001. Kalman filtering, theory and practice using MATLAB, 2nd Edition. Wiley–IEEE Press.

Gustafsson, N., 2007. Discussion on '4D-Var or EnKF?'. Tellus A 59 (5), 774–777.

Gustafsson, N., Bojarova, J., Vignes, O., 2014. A hybrid variational ensemble data assimilation for the HIgh Resolution Limited Area Model (HIRLAM). Nonlinear Processes in Geophysics 21 (1), 303–323.

Hamill, T. M., 2001. Interpretation of rank histograms for verifying ensemble forecasts. Monthly Weather Review 129 (3), 550–560.

Hamill, T. M., Snyder, C., 2000. A hybrid ensemble Kalman filter-3D variational analysis scheme. Monthly Weather Review 128 (8), 2905–2919.

Hamill, T. M., Whitaker, J. S., Snyder, C., 2001. Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. Monthly Weather Review 129 (11), 2776–2790.

Hoteit, I., Triantafyllou, G., Korres, G., 2007. Using low-rank ensemble Kalman filters for data assimilation with high dimensional imperfect models. Journal of Numerical Analysis, Industrial and Applied Mathematics 2 (1-2), 67–78.

Houtekamer, P., Mitchell, H., 1998. Data assimilation using an ensemble Kalman filter technique. Monthly Weather Review 126, 796–811.

Houtekamer, P., Mitchell, H. L., 2005. Ensemble kalman filtering. Quarterly Journal of the Royal Meteorological Society 131 (613), 3269–3289.

Houtekamer, P. L., Mitchell, H. L., 2001. A sequential ensemble Kalman filter for atmospheric data assimilation. Monthly Weather Review 129 (1), 123–137.

Hu, C., Sueyoshi, M., 2010. Numerical simulation and experiment on dam break problem. Journal of Marine Science and Application 9 (2), 109–114.

Hunt, B., Kalnay, E., Kostelich, E., Ott, E., Patil, D., Sauer, T., Szunyogh, I., Yorke, J., Zimin, A., 2004. Four-dimensional ensemble Kalman filtering. Tellus A 56 (4), 273–277.

Ikeda, M., 1981. Meanders and detached eddies of a strong eastward-flowing jet using a two-layer quasi-geostrophic model. Journal of Physical Oceanography 11 (4), 526–540.

Järvinen, H., Laine, M., Solonen, A., Haario, H., 2012. Ensemble prediction and parameter estimation system: the concept. Quarterly Journal of the Royal Meteorological Society 138 (663), 281–288.

Jazwinski, A. H., 1970. Stochastic Processes and Filtering Theory. Academic Press.

Julier, S. J., Uhlmann, J. K., 2004. Unscented filtering and nonlinear estimation. Proceedings of the IEEE 92 (3), 401–422.

Kalman, R. E., 1960. A New Approach to Linear Filtering and Prediction Problems. Transactions of the ASME–Journal of Basic Engineering 82, 35–45.

Kalnay, E., 2003. Atmospheric modeling, data assimilation, and predictability. Cambridge university press.

Kalnay, E., Li, H., Miyoshi, T., Yang, S.-C., Ballabrera-Poy, J., 2007. 4D-Var or ensemble Kalman filter? Tellus A 59 (5), 758–773.

Kandepu, R., Foss, B., Imsland, L., 2008. Applying the unscented Kalman filter for nonlinear state estimation. Journal of Process Control 18 (7), 753–768.

Kuznetsov, L., Ide, K., Jones, C., 2003. A method for assimilation of Lagrangian data. Monthly Weather Review 131 (10), 2247–2260.

Laine, M., Solonen, A., Haario, H., Järvinen, H., 2012. Ensemble prediction and parameter estimation system: the method. Quarterly Journal of the Royal Meteorological Society 138 (663), 289–297.

Lawson, W. G., Hansen, J. A., 2004. Implications of stochastic and deterministic filters as ensemble-based data assimilation methods in varying regimes of error growth. Monthly Weather Review 132 (8), 1966–1981.

Le Dimet, F.-X., Talagrand, O., 1986. Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects. Tellus A 38 (2), 97–110.

Li, B., Toll, D., Zhan, X., Cosgrove, B., 2012. Improving estimated soil moisture fields through assimilation of AMSR-E soil moisture retrievals with an ensemble Kalman filter and a mass conservation constraint. Hydrology and Earth System Sciences 16 (1), 105–119.

Li, Y., Navon, I., Courtier, P., Gauthier, P., 1993. Variational data assimilation with a semi-Lagrangian semi-implicit global shallow-water equation model and its adjoint. Monthly Weather Review 121 (6), 1759–1769.

Li, Z., Navon, I., 2001. Optimality of variational data assimilation and its relationship with the Kalman filter and smoother. Quarterly Journal of the Royal Meteorological Society 127 (572), 661–683.

Liu, C., Xiao, Q., Wang, B., 2008. An ensemble-based four-dimensional variational data assimilation scheme. Part I: Technical formulation and preliminary test. Monthly Weather Review 136 (9), 3363–3373.

Liu, Y., Weerts, A. H., Clark, M., Hendricks Franssen, H.-J., Kumar, S., Moradkhani, H., Seo, D.-J., Schwanenberg, D., Smith, P., van Dijk, A. I. J. M., van Velzen, N., He, M., Lee, H., Noh, S. J., Rakovec, O., Restrepo, P., 2012. Advancing data assimilation in operational hydrologic forecasting: progresses, challenges, and emerging opportunities. Hydrology and Earth System Sciences 16 (10), 3863–3887.

Lorenc, A. C., Bowler, N. E., Clayton, A. M., Pring, S. R., Fairbairn, D., 2014. Comparison of hybrid-4DEnVar and hybrid-4DVar data assimilation methods for global NWP. Monthly Weather Review 143, 212–229.

Lü, H., Yu, Z., Zhu, Y., Drake, S., Hao, Z., Sudicky, E. A., 2011. Dual state-parameter estimation of root zone soil moisture by optimal parameter estimation and extended Kalman filter data assimilation. Advances in Water Resources 34 (3), 395–406.

Lynch, P., 2008. The origins of computer weather prediction and climate modeling. Journal of Computational Physics 227 (7), 3431–3444.

Madsen, H., Skotner, C., 2005. Adaptive state updating in real-time river flow forecasting - a combined filtering and error forecasting procedure. Journal of Hydrology 308 (1), 302–312.

Martin, N., Gorelick, S. M., 2005. MODFreeSurf2D: A MATLAB surface fluid flow model for rivers and streams. Computers & Geosciences 31 (7), 926–946.

Mbalawata, I. S., December 2014. Adaptive Markov chain Monte Carlo and Bayesian filtering for state space models. Ph.D. thesis, Lappeenranta University of Technology.

McMillan, H. K., Hreinsson, E. O., Clark, M. P., Singh, S. K., Zammit, C., Uddstrom, M. J., 2013. Operational hydrological data assimilation with the recursive ensemble Kalman filter. Hydrology and Earth System Sciences 17 (1), 21–38.

Medjo, T. T., 2000. Numerical simulations of a two-layer quasi-geostrophic equation of the ocean. SIAM journal on numerical analysis 37 (6), 2005–2022.

Moradkhani, H., Hsu, K.-L., Gupta, H., Sorooshian, S., 2005a. Uncertainty assessment of hydrologic model states and parameters: Sequential data assimilation using the particle filter. Water Resources Research 41 (5).

Moradkhani, H., Sorooshian, S., Gupta, H. V., Houser, P. R., 2005b. Dual state parameter estimation of hydrological models using ensemble Kalman filter. Advances in Water Resources 28 (2), 135–147.

Morris, M., 2000. Concerted Action on Dambreak Modelling-CADAM, Final Report. Tech. rep., HR Wallingford Limited.

Nakamura, K., Higuchi, T., Hirose, N., 2006. Sequential Data Assimilation: Information Fusion of a Numerical Simulation and Large Scale Observation Data. J. UCS 12 (6), 608–626.

Navon, I. M., 1998. Practical and theoretical aspects of adjoint parameter estimation and identifiability in meteorology and oceanography. Dynamics of Atmospheres and Oceans 27 (1), 55–79.

Navon, I. M., 2009. Data assimilation for numerical weather prediction: a review. In: Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications. Springer, pp. 21–65.

Navon, I. M., Neta, B., Hussaini, M. Y., 2004. A perfectly matched layer approach to the linearized shallow water equations models. Monthly Weather Review 132 (6), 1369–1378.

Nocedal, J., Wright, S., 1999. Numerical Optimization. Springer: Berlin.

Ott, E., Hunt, B. R., Szunyogh, I., Zimin, A. V., Kostelich, E. J., Corazza, M., Kalnay, E., Patil, D., Yorke, J. A., 2004. A local ensemble Kalman filter for atmospheric data assimilation. Tellus A 56 (5), 415–428.

Pedlosky, J., 1987. Geophysical Fluid Dynamics. Springer-Verlag, New York, Ch. Geostrophic Motion, pp. 22–57.

Reichle, R. H., McLaughlin, D. B., Entekhabi, D., 2002a. Hydrologic Data Assimilation with the Ensemble Kalman Filter. Monthly Weather Review 130, 103–114, doi: 10.1175/1520-0493.

Reichle, R. H., Walker, J. P., Koster, R. D., Houser, P. R., 2002b. Extended versus ensemble Kalman filtering for land data assimilation. Journal of hydrometeorology 3 (6), 728–740.

Salman, H., 2008. A hybrid grid/particle filter for Lagrangian data assimilation. I: Formulating the passive scalar approximation. Quarterly Journal of the Royal Meteorological Society 134 (635), 1539–1550.

Särkkä, S., 2013. Bayesian filtering and smoothing. Cambridge University Press.

Sarveram, H., Shamsai, A., Banihashemi, M. A., 2012. Two-dimensional simulation of flow pattern around a groyne using semi-implicit semi-Lagrangian method. International Journal of Physical Sciences 7 (20), 2775–2783.

Sasaki, Y., 1970a. Numerical variational analysis formulated under the constraints as determined by longwave equations and a low-pass filter. Mon. Wea. Rev 98 (12), 884–898.

Sasaki, Y., 1970b. Numerical variational analysis with weak constraint and application to surface analysis of severe storm gust. Mon. Wea. Rev 98, 899–910.

Sasaki, Y., 1970c. Some basic formalisms in numerical variational analysis. Monthly Weather Review 98 (12), 875–883.

Schlatter, T. W., 2000. Variational assimilation of meteorological observations in the lower atmosphere: a tutorial on how it works. Journal of atmospheric and solar-terrestrial physics 62 (12), 1057–1070.

Sene, K., 2010. Hydrological Forecasting. In: Hydrometeorology. Springer Netherlands, pp. 101–140.
URL http://dx.doi.org/10.1007/978-90-481-3403-8_4

Shen, Z., Tang, Y., 2015. A modified ensemble Kalman particle filter for non-Gaussian systems with nonlinear measurement functions. Journal of Advances in Modeling Earth Systems.

Slivinski, L., Spiller, E., Apte, A., Sandstede, B., 2015. A Hybrid Particle–Ensemble Kalman Filter for Lagrangian Data Assimilation. Monthly Weather Review 143 (1), 195–211.

Snyder, C., Bengtsson, T., Bickel, P., Anderson, J., 2008. Obstacles to high-dimensional particle filtering. Monthly Weather Review 136 (12), 4629–4640.

Solonen, A., November 2011. Bayesian methods for estimation, optimization and experimental design. Ph.D. thesis, Lappeenranta University of Technology.

Solonen, A., Bardsley, J. M., Bibov, A., Haario, H., 2014. Optimization-based Sampling in Ensemble Kalman Filtering. International Journal for Uncertainty Quantification.

Solonen, A., Hakkarainen, J., Auvinen, H., Amour, I., Haario, H., Kauranne, T., 2012. Variational Ensemble Kalman Filtering Using Limited Memory BFGS. Electronic Transaction on Numerical Analysis 39, 271–285, iSSN 1068-9613.

Solonen, A., Järvinen, H., 2013. An approach for tuning ensemble prediction systems. Tellus A 65.

Stammer, D., Wunsch, C., Fukumori, I., Marshall, J., 2002. State estimation improves prospects for ocean research. Eos, Transactions American Geophysical Union 83 (27), 289–295.

Staniforth, A., Côté, J., 1991. Semi-Lagrangian integration schemes for atmospheric models-a review. Monthly Weather Review 119 (9), 2206–2223.

Stensrud, D. J., Brooks, H. E., Du, J., Tracton, M. S., Rogers, E., 1999. Using ensembles for short-range forecasting. Monthly Weather Review 127 (4), 433–446.

Talagrand, O., 1997. Assimilation of observations, an introduction. Journal - Meteorological Society of Japan Series 2 75, 81–99.

Tan, W.-Y., 1992. Shallow water hydrodynamics: Mathematical Theory and Numerical Solution for a Two-dimensional System of Shallow-Water Equations. Elsevier.

Tippett, M. K., Anderson, J. L., Bishop, C. H., Hamill, T. M., Whitaker, J. S., 2003. Ensemble Square Root Filters. American Meteorological Society 131 (7), 1485–1490.

Todling, R., 1999. Estimation theory and foundations of atmospheric data assimilation. DAO Office Note 1.

Tossavainen, O.-P., Percelay, J., Tinka, A., Wu, Q., Bayen, A. M., 2008. Ensemble Kalman filter based state estimation in 2d shallow water equations using lagrangian sensing and state augmentation. In: Decision and Control, 2008. CDC 2008. 47th IEEE Conference on. IEEE, pp. 1783–1790.

van Leeuwen, P. J., 2010. Nonlinear data assimilation in geosciences: an extremely efficient particle filter. Quarterly Journal of the Royal Meteorological Society 136 (653), 1991–1999.

van Leeuwen, P. J., 2011. Efficient nonlinear data assimilation in geophysical fluid dynamics. Computer & Fluids 46, 52–58.

Wang, B., Zou, X., Zhu, J., 2000. Data assimilation and its applications. Proceedings of the National Academy of Sciences 97 (21), 11143–11144.

Wang, Z., Navon, I., Zou, X., Le Dimet, F., 1995. A truncated Newton optimization algorithm in meteorology applications with analytic Hessian/vector products. Computational Optimization and Applications 4 (3), 241–262.

Whitaker, J. S., Hamill, T. M., 2002. Ensemble data assimilation without perturbed observations. Monthly Weather Review 130 (7), 1913–1924.

Wood, A. W., Maurer, E. P., Kumar, A., Lettenmaier, D. P., 2002. Long-range experimental hydrologic forecasting for the eastern United States. Journal of Geophysical Research: Atmospheres 107 (D20), ACL 6–1–ACL 6–15.
URL http://dx.doi.org/10.1029/2001JD000659

Wu, L., Mallet, V., Bocquet, M., Sportisse, B., 2008. A comparison study of data assimilation algorithms for ozone forecasts. Journal of Geophysical Research: Atmospheres (1984–2012) 113 (D20).

Xie, X., Zhang, D., 2010. Data assimilation for distributed hydrological catchment modeling via ensemble Kalman filter. Advances in Water Resources 33 (6), 678–690.

Yang, Y., Robinson, C., Heitz, D., Mémin, E., 2015. Enhanced ensemble-based 4DVar scheme for data assimilation. Computers & Fluids.

Zou, X., Navon, I., Le Dimet, F., 1992a. Incomplete observations and control of gravity waves in variational data assimilation. Tellus A 44 (4), 273–296.

Zou, X., Navon, I., LeDimet, F., 1992b. An optimal nudging data assimilation scheme using parameter estimation. Quarterly Journal of the Royal Meteorological Society 118 (508), 1163–1186.

Zou, X., Navon, I. M., Berger, M., Phua, K. H., Schlick, T., Le Dimet, F.-X., 1993. Numerical experience with limited-memory quasi-Newton and truncated Newton methods. SIAM Journal on Optimization 3 (3), 582–608.

Zupanski, M., 2005. Maximum likelihood ensemble filter: Theoretical aspects. Monthly Weather Review 133 (6), 1710–1726.

Zupanski, M., Navon, I. M., Zupanski, D., 2008. The Maximum Likelihood Ensemble Filter as a non-differentiable minimization algorithm. Quarterly Journal of the Royal Meteorological Society 134 (633), 1039–1050.

APPENDIX I

Appendix

Appendix A: Variational Ensemble Kalman filter

This appendix gives the details of the variational ensemble Kalman filter as described by Solonen et al. (2012). Given the dynamic process and the measurement model,

$$\mathbf{x}_k = \mathscr{M}(\mathbf{x}_{k-1}) + \mathbf{q}_k, \tag{1}$$
$$\mathbf{y}_k = \mathscr{K}(\mathbf{x}_k) + \mathbf{r}_k, \tag{2}$$

The state estimation is obtained by minimizing the cost function defined in section 3.1.4 as

$$l(\mathbf{x} \mid \mathbf{y}_k) = \frac{1}{2}(\mathbf{x}-\mathbf{x}_k^p)^T(\mathbf{C}_k^p)^{-1}(\mathbf{x}-\mathbf{x}_k^p) + \frac{1}{2}(\mathbf{y}_k - \mathscr{K}(\mathbf{x}))^T\mathbf{R}_k^{-1}(\mathbf{y}_k - \mathscr{K}(\mathbf{x})). \tag{3}$$

Let $\{s_{k,i}\}_{i=1}^N$ be a bundle of N-dimensional ensembles sampled from a distribution of $\mathbf{X}_k^{est}$, $s_{k,i} \sim \mathscr{N}\left(\mathbf{x}_k^{est}, \mathbf{C}_k^{est}\right)$. Using the predicted state from the previous time point we define the a vector $\mathbf{X}_k$ as

$$\mathbf{X}_k = \left(\left(s_{k,1} - \mathbf{x}_k^p\right), \ldots, \left(s_{k,N} - \mathbf{x}_k^p\right)\right)/\sqrt{N}, \tag{4}$$

The sampled approximation of the prior covariance required in the cost function (3) is defined as

$$\mathbf{C}_k^p = Cov(\mathscr{M}(\mathbf{x}_{k-1}^{est}) + \mathbf{r}_k) = Cov(\mathscr{M}(\mathbf{x}_{k-1}^{est})) + Cov(\mathbf{q}_k) \approx \mathbf{X}_k\mathbf{X}_k^T + (Q). \tag{5}$$

The inverse of the prior covariance (5) can be approximated either by applying LBFGS optimization to the artificial optimization problem

$$argmin_u \mathbf{u}^T(\mathbf{X}_k\mathbf{X}_k^T + \mathbf{Q})\mathbf{u},$$

where $\mathbf{u}$ is a vector and $\mathbf{Q}$ is assumed to be diagonal. The LBFGS optimization gives a recursive algorithm that can allow the computation of matrix-vector product when the matrix is in LBFGS form, (see Appendix B for more details) .

Alternatively the inverse of the prio covariance can be calculated by Sherman Morrison-Woodbery (SMW) matrix identity defined as

$$\left[\mathbf{C}_k^p\right]^{-1} = \mathbf{Q}^{-1} - \mathbf{Q}^{-1}\mathbf{X}_k\left(\mathbf{I} + \mathbf{X}_k^T\mathbf{Q}^{-1}\mathbf{X}_k\right)^{-1}\mathbf{X}_k^T\mathbf{Q}^{-1}. \tag{6}$$

where the matrix $\mathbf{Q}$ is assumed diagonal therefore it can be inverted. The inverse of the matrix $\left(\mathbf{I}+\mathbf{X}_k^T\mathbf{Q}^{-1}\mathbf{X}_k\right)^{-1}$ is considered feasible since the ensemble size is assumed to have a lower dimension compared to the dimension of the state.

Minimization of the cost function (3) is done using LBFGS unconstrained optimization Nocedal and Wright (1999) (see Appendix B). The LBFGS optimization gives a low storage approximation for the covariance $\mathbf{C}_k^{est}$ as the inverse Hessian of (3).

Appendix B: LBFGS Optimization

To approximate the inverse of the prior error covariance matrix $\mathbf{C}_k^p$, the LBFGS is applied to an auxiliary optimization problem

$$argmin_u \mathbf{u}^T(\mathbf{X}_k\mathbf{X}_k^T+\mathbf{Q})\mathbf{u},$$

where $\mathbf{u}$ is a vector and $\mathbf{Q}$ is assumed to be diagonal.

The first term of the auxiliary optimization can be evaluate using LBFGS algorithm to the quadratic function $q(u)=\frac{1}{2}\mathbf{u}^T\mathbf{A}\mathbf{u}$ given an initial guess $\mathbf{u}_0$. The LBFGS algorithm for quadratic function problem reads as follows.

Algorithm .1 LBFGS algorithm for quadratic problem

i) Choose an initial guess $\mathbf{u}_0$ and an initial inverse Hessian $\mathbf{H}_k^0$.

ii) Compute the gradient $\mathbf{g}_k=\nabla q(\mathbf{u}_k)=\mathbf{A}\mathbf{u}_k$.

iii) Compute the search direction $\mathbf{p}_k=\mathbf{H}_k\mathbf{g}_k$ where $\mathbf{H}_k$ in the LBFGS approximation of the inverse Hessian outlined below.

iv) Compute step size $\alpha_k=\mathbf{g}_k^T\mathbf{p}_k/\mathbf{p}_k^T\mathbf{A}\mathbf{p}_k$.

v) Compute $\mathbf{u}_{k+1}=\mathbf{u}_k-\alpha_k\mathbf{p}_k$ and

vi) Set $k\rightarrow k+1$ and go to step 1

The inverse Hessian $\mathbf{H}_k$ is approximated using BFGS formula defined as

$$\mathbf{H}_{k+1}=\mathbf{V}_k^T\mathbf{H}_k\mathbf{V}_k+\rho_k\mathbf{s}_k\mathbf{s}_k^T$$

where

$$\rho_k=1/(y_k^T\mathbf{s}_k)$$
$$\mathbf{V}_k=\mathbf{I}-\rho_k y_k\mathbf{s}_k^T$$
$$\mathbf{s}_k=\mathbf{u}_{k+1}-\mathbf{u}_k$$
$$\mathbf{y}_k=\mathbf{g}_{k+1}-\mathbf{g}_k.$$

LBFGS need not the storage of full matrix $\mathbf{H}_k$ but only a certain number n of the most recent vector of the pair $\{\mathbf{s}_i,\mathbf{y}_i\}$. Hence the recursive formula of the Hessian matrix can be written be written in

the following way

$$\begin{aligned}\mathbf{H}_k = &(\mathbf{V}_{k-1}^T \dots \mathbf{V}_{k-n})\mathbf{H}_k^0(\mathbf{V}_{k-n} \dots \mathbf{V}_{k-1}) \\ &+ \rho_{k-n}(\mathbf{V}_{k-1}^T \dots \mathbf{V}_{k-n+1})\mathbf{s}_{k-n}\mathbf{s}_{k-n}^T(\mathbf{V}_{k-n+1} \dots \mathbf{V}_{k-1}) \\ &+ \rho_{k-n+1}(\mathbf{V}_{k-1}^T \dots \mathbf{V}_{k-n+2})\mathbf{s}_{k-n+1}\mathbf{s}_{k-n+1}^T(\mathbf{V}_{k-n+2} \dots \mathbf{V}_{k-1}) \\ &+ \dots \\ &+ \rho_{k-1}\mathbf{s}_{k-1}\mathbf{s}_{k-1}^T.\end{aligned}$$

In LBFGS, there is no unique way of choosing the initial inverse Hessian $\mathbf{H}_k^0$. One way is to choose an identity matrix or a fixed diagonal covariance $\mathbf{H}_k^0 = \gamma_k \mathbf{I}$ where the parameter $\gamma_k = (\mathbf{y}_k^T \mathbf{s}_k)/(\mathbf{y}_k^T \mathbf{y}_k)$ will estimate the size of the covariance along the last search direction (Nocedal and Wright 1999). This scaling factor is calculated after the first step has been computed but before the BFGS update is performed (Nocedal and Wright 1999).

To avoid storage of a full inverse Hessian, $\mathbf{H}_k$ is kept in vector form and there exist an iterative algorithm for computing the matrix-vector product with the inverse Hessian. Assumung that the initial inverse Hessian can be decomposed into $\mathbf{H}_k^0 = \mathbf{L}_0\mathbf{L}_0^T$, the LBFGS inverse Hessian formula can be written as

$$\mathbf{H}_k = \mathbf{B}_0\mathbf{B}_0^T + \sum_i^n \mathbf{b}_i\mathbf{b}_i^T,$$

where

$$\begin{aligned}\mathbf{B}_0 &= (\mathbf{V}_{k-1}^T \dots \mathbf{V}_{k-n}^T)\mathbf{L}_0 \\ \mathbf{b}_1 &= \sqrt{\rho_{k-1}}\mathbf{s}_{k-1} \\ \mathbf{b}_i &= \sqrt{\rho_{k-i}}(\mathbf{V}_{k-1}^T \dots \mathbf{V}_{k-i+1}^T)\mathbf{s}_{k-i}, \;\; i = 2, \dots, n.\end{aligned}$$

The step length is chosen such that $rho_i \geq 0$ for all i this makes the square root $\sqrt{\rho_i}$ always be calculated. Thus, we can sample zero mean random variables from the covariance $\mathbf{H}_k$ by calculating

$$\mathbf{r} = \mathbf{B}_0\mathbf{z} + \sum_{i=1}^n w_i\mathbf{b}_i,$$

where $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $w_i \sim \mathcal{N}(0, i)$. It can simply be verified that $Cov(\mathbf{r}) = \mathbf{H}_k$.

ACTA UNIVERSITATIS LAPPEENRANTAENSIS

616. METSO, SARI. A multimethod examination of contributors to successful on-the-job learning of vocational students. 2014. Diss.

617. SIITONEN, JANI. Advanced analysis and design methods for preparative chromatographic separation processes. 2014. Diss.

618. VIHAVAINEN, JUHANI. VVER-440 thermal hydraulics as computer code validation challenge. 2014. Diss.

619. AHONEN, PASI. Between memory and strategy: media discourse analysis of an industrial shutdown. 2014. Diss.

620. MWANGA, GASPER GODSON. Mathematical modeling and optimal control of malaria. 2014. Diss.

621. PELTOLA, PETTERI. Analysis and modelling of chemical looping combustion process with and without oxygen uncoupling. 2014. Diss.

622. NISKANEN, VILLE. Radio-frequency-based measurement methods for bearing current analysis in induction motors. 2014. Diss.

623. HYVÄRINEN, MARKO. Ultraviolet light protection and weathering properties of wood-polypropylene composites. 2014. Diss.

624. RANTANEN, NOORA. The family as a collective owner – identifying performance factors in listed companies. 2014. Diss.

625. VÄNSKÄ, MIKKO. Defining the keyhole modes – the effects on the molten pool behavior and the weld geometry in high power laser welding of stainless steels. 2014. Diss.

626. KORPELA, KARI. Value of information logistics integration in digital business ecosystem. 2014. Diss.

627. GRUDINSCHI, DANIELA. Strategic management of value networks: how to create value in cross-sector collaboration and partnerships. 2014. Diss.

628. SKLYAROVA, ANASTASIA. Hyperfine interactions in the new Fe-based superconducting structures and related magnetic phases. 2015. Diss.

629. SEMKEN, R. SCOTT. Lightweight, liquid-cooled, direct-drive generator for high–power wind turbines: motivation, concept, and performance. 2015. Diss.

630. LUOSTARINEN, LAURI. Novel virtual environment and real-time simulation based methods for improving life-cycle efficiency of non-road mobile machinery. 2015. Diss.

631. ERKKILÄ, ANNA-LEENA. Hygro-elasto-plastic behavior of planar orthotropic material. 2015. Diss.

632. KOLOSENI, DAVID. Differential evolution based classification with pool of distances and aggregation operators. 2015. Diss.

633. KARVONEN, VESA. Identification of characteristics for successful university-company partnership development. 2015. Diss.

634. KIVYIRO, PENDO. Foreign direct investment, clean development mechanism, and environmental management: a case of Sub-Saharan Africa. 2015. Diss.

635. SANKALA, ARTO. Modular double-cascade converter. 2015. Diss.

636. NIKOLAEVA, MARINA. Improving the fire retardancy of extruded/coextruded wood-plastic composites. 2015. Diss.

637. ABDEL WAHED, MAHMOUD. Geochemistry and water quality of Lake Qarun, Egypt. 2015. Diss.

638. PETROV, ILYA. Cost reduction of permanent magnet synchronous machines. 2015. Diss.

639. ZHANG, YUNFAN. Modification of photocatalyst with enhanced photocalytic activity for water treatment. 2015. Diss.

640. RATAVA, JUHO. Modelling cutting states in rough turning of 34CrNiMo6 steel. 2015. Diss.

641. MAYDANNIK, PHILIPP. Roll-to-roll atomic layer deposition process for flexible electronics applications. 2015. Diss.

642. SETH, FRANK. Empirical studies on software quality construction: Exploring human factors and organizational influences. 2015. Diss.

643. SMITH, AARON. New methods for controlling twin configurations and characterizing twin boundaries in 5M Ni-Mn-Ga for the development of applications. 2015. Diss.

644. NIKKU, MARKKU. Three-dimensional modeling of biomass fuel flow in a circulating fluidized bed furnace. 2015. Diss.

645. HENTTU, VILLE. Improving cost-efficiency and reducing environmental impacts of intermodal transportation with dry port concept – major rail transport corridor in Baltic Sea region. 2015. Diss.

646. HAN, BING. Influence of multi-phase phenomena on semibatch crystallization processes of aqueous solutions. 2015. Diss.

647. PTAK, PIOTR. Aircraft tracking and classification with VHF passive bistatic radar. 2015. Diss.

648. MAKKONEN, MARI. Cross-border transmission capacity development – Experiences from the Nordic electricity markets. 2015. Diss.

649. UUSITALO, ULLA-MAIJA. Show me your brain! Stories of interdisciplinary knowledge creation in practice. Experiences and observations from Aalto Design Factory, Finland. 2015. Diss.

650. ROOZBAHANI, HAMID. Novel control, haptic and calibration methods for teleoperated electrohydraulic servo systems. 2015. Diss.

651. SMIRNOVA, LIUDMILA. Electromagnetic and thermal design of a multilevel converter with high power density and reliability. 2015. Diss.

652. TALVITIE, JOONAS. Development of measurement systems in scientific research: Case study. 2015. Diss.