3-18-2011

4-D Var Data Assimilation and POD Model Reduction Applied to Geophysical Dynamics Models

Xiao Chen
Florida State University

THE FLORIDA STATE UNIVERSITY

COLLEGE OF ARTS AND SCIENCES

4-D VAR DATA ASSIMILATION AND POD MODEL REDUCTION APPLIED TO GEOPHYSICAL DYNAMICS MODELS

By

XIAO CHEN

A Dissertation submitted to the
Department of Mathematics
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Degree Awarded:
Spring Semester, 2011

The members of the Committee approve the Dissertation of Xiao Chen defended on March 18, 2011.

Ionel Michael Navon
Professor Directing Dissertation

Mark Sussman
Professor Co-Directing Dissertation

Robert Hart
University Representative

Xiaoming Wang
Committee Member

Erlebacher Gordon
Committee Member

Approved:

Philip L. Bowers, Chair
Department of Mathematics

Dr. Joseph Travis, Dean, College of Arts and Sciences

The Graduate School has verified and approved the above named committee members.

This thesis is dedicated to my beloved wife and my parents for their patient love, endless encouragement and unconditional support.

ACKNOWLEDGEMENTS

The realization of this work was only possible due to the several people's collaboration, to whom I desire to express my gratitude.

I would first like to acknowledge my gratitude to my Ph.D advisor, Professor I. M. Navon through out my past five years at Florida State University. His encouragement, guidance and support from the initial to the final level have enabled me to develop a comprehensive understanding of the research project. Prof. Navon has helped me to develop the capability to carry on a large variety of challenging research topics on my own independently. He also provided me the freedom to collaborate with many other researchers working on similar domains. Working with Prof. Navon is inspiring and I was contaminated by his enthusiasm. He is the one of the smartest and knowledgeable people I have ever known.

I am also very grateful to Prof. Mark Sussman for his unconditional help. He provided insightful discussions and suggestions about my research many times last year. He has to fly to FSU in order to attend my Ph.D defense out of his busy schedule. I would like to express my gratitude to Prof. Xiaoming Wang, Prof. Gordon Erlebacher, Prof. Robert Hart, Prof. Xiaolei Zhou, Prof. Gunzburger Max, Prof. Peterson Janet and Prof. Ming Ye for providing valuable inputs and advices. This dissertation would not have been possible without their patient help.

I would like to thank many of my dear friends including Mr. Zheng Guoxian, Dr. Zheng Weiwei, Dr. Willy, Mr. Kwok, Phil, Takahara, Mingxin Chen, Peilin Yu, Qiang Zhang, Xiangrong Xu, Jianyang Liu, Dr. Nan Liu, Dr. Xia Liao, Dr. Sandosh and Charley from my days at Florida State University.

I owe my deepest gratitude to Dr. Xiaoguang Li for his technical support. He was always available and willing to help whenever I ask. I am indebted to many of my colleagues for supporting me including Dr. Fang, Dr. Juan Du, Dr. Jeff Steward, Dr. Santha Akella, Prof. Bill Hu, Prof. Adrian Sandu, Dr. Mihai Alexe, Dr. Xinya Li, Dr. Jinshan Xie, Dr.

Surujan, Dr. Qinshan Chen, Dr. Jardak, Dr. Burkardt John, Dr. Hailin Deng, McLaughlin Ben, Dr. Chunhong Qi, Liang Li, Nathan Lay and Dr. Charles Tong. I offer my regards and blessings to all of those who supported me in any way during the completion of my Ph.D degree. The successful completion of this manuscript was made possible through the invaluable contribution of a number of people. To say "thank you" to all of you is clearly not even enough to express my gratitude.

It is an honor for me to thank the financial support for this research which is provided partially by the National Science Foundation (NSF) from 2007 to 2008. I would especially like to thank my parents for their continued moral support and my wife who was always there for me with her love and continued support. Finally, I would like to thank the Lord Father for his truly knowing my weakness, showing me the way like a bright morning star and giving me the strength to complete my Ph.D degree.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Standard spatial discretization schemes for dynamical system (DS), usually lead to large-scale, high-dimensional, and in general, nonlinear systems of ordinary differential equations.Due to limited computational and storage capabilities, Reduced Order Modeling (ROM) techniques from system and control theory provide an attractive approach to approximate the large-scale discretized state equations using low-dimensional models.

The objective of 4-D variational data assimilation (4-D Var) is to obtain the minimum of a cost functional estimating the discrepancy between the model solutions and distributed observations in time and space. A control reduction methodology based on Proper Orthogonal Decomposition (POD), referred to as POD 4-D Var, has been widely used for nonlinear systems with tractable computations.

However, the appropriate criteria for updating a POD ROM are not yet known in the application to optimal control. This is due to the limited validity of the POD ROM for inverse problems. Therefore, the classical Trust-Region (TR) approach combined with POD (TRPOD) was recently proposed as a way to alleviate the above difficulties. There is a global convergence result for TR, and benefiting from the trust-region philosophy, rigorous convergence results guarantee that the iterates produced by the TRPOD algorithm will converge to the solution of the original optimization problem.

In order to reduce the POD basis size and still achieve the global convergence, a method was proposed to incorporate information from the 4-D Var system into the ROM procedure by implementing a dual weighted POD (DWPOD) method.

The first new contribution in my dissertation consists in studying a new methodology combining the dual weighted snapshots selection and trust region POD adaptivity (DWTR-

POD). Another new contribution is to combine the incremental POD 4-D Var, balanced truncation techniques and method of snapshots methodology. In the linear DS, this is done by integrating the linear forward model many times using different initial conditions in order to construct an ensemble of snapshots so as to generate the forward POD modes. Then those forward POD modes will serve as the initial conditions for its corresponding adjoint system. We then integrate the adjoint system a large number of times based on different initial conditions generated by the forward POD modes to construct an ensemble of adjoint snapshots. From this ensemble of adjoint snapshots, we can generate an ensemble of so-called adjoint POD modes. Thus we can approximate the controllability Grammian of the adjoint system instead of solving the computationally expensive coupled Lyapunov equations. To sum up, in the incremental POD 4-D Var, we can approximate the controllability Grammian by integrating the TLM a number of times and approximate observability Grammian by integrating its adjoint also a number of times.

A new idea contributed in this dissertation is to extend the snapshots based POD methodology to the nonlinear system. Furthermore, we modify the classical algorithms in order to save the computations even more significantly. We proposed a novel idea to construct an ensemble of snapshots by integrating the tangent linear model (TLM) only once, based on which we can obtain its TLM POD modes. Then each TLM POD mode will be used as an initial condition to generate a small ensemble of adjoint snapshots and their adjoint POD modes. Finally, we can construct a large ensemble of adjoint POD modes by putting together each small ensemble of adjoint POD modes. To sum up, our idea in a forthcoming study is to test approximations of the controllability Grammian by integrating TLM once and observability Grammian by integrating adjoint model a reduced number of times.

Optimal control of a finite element limited-area shallow water equations model is explored with a view to apply variational data assimilation(VDA) by obtaining the minimum of a functional estimating the discrepancy between the model solutions and distributed observations. In our application, some simplified hypotheses are used, namely the error of the model is neglected, only the initial conditions are considered as the control variables, lateral boundary conditions are periodic and finally the observations are assumed to be distributed in space and time. Derivation of the optimality system including the adjoint state, permits

computing the gradient of the cost functional with respect to the initial conditions which are used as control variables in the optimization. Different numerical aspects related to the construction of the adjoint model and verification of its correctness are addressed. The data assimilation set-up is tested for various mesh resolutions scenarios and different time steps using a modular computer code. Finally, impact of large-scale unconstrained minimization solvers L-BFGS is assessed for various lengths of the time windows.

We then attempt to obtain a reduced-order model (ROM) of above inverse problem, based on proper orthogonal decomposition(POD), referred to as POD 4-D Var. Different approaches of POD implementation of the reduced inverse problem are compared, including a dual-weighed method for snapshot selection coupled with a trust-region POD approach. Numerical results obtained point to an improved accuracy in all metrics tested when dual-weighing choice of snapshots is combined with POD adaptivity of the trust-region type. Results of ad-hoc adaptivity of the POD 4-D Var turn out to yield less accurate results than trust-region POD when compared with high-fidelity model.

Finally, we study solutions of an inverse problem for a global shallow water model controlling its initial conditions specified from the 40-yr ECMWF Re-Analysis (ERA-40) datasets, in presence of full or incomplete observations being assimilated in a time interval (window of assimilation) presence of background error covariance terms. As an extension of this research, we attempt to obtain a reduced-order model of above inverse problem, based on proper orthogonal decomposition (POD), referred to as POD 4-D Var for a finite volume global shallow water equations model based on the Lin-Rood [89, 90, 91, 92, 96] flux-form semi-Lagrangian semi-implicit time integration scheme. Different approaches of POD implementation for the reduced inverse problem are compared, including a dual-weighted method for snapshot selection coupled with a trust-region POD adaptivity approach. Numerical results with various observational densities and background error covariance operator are also presented. The POD 4-D Var model results combined with the trust region adaptivity exhibit similarity in terms of various error metrics to the full 4-D Var results, but are obtained using a significantly lesser number of minimization iterations and require lesser CPU time. Based on our previous and current research work, we conclude that POD 4-D Var certainly warrants further studies, with promising potential for its extension to operational 3-D numerical weather prediction models.

CHAPTER 1

INTRODUCTION

Computational simulation, or more generally, computational science is now regarded as the third discipline, besides the classical disciplines of pure theory and real experiment in science and industry. It has now become a useful part of modeling many of systems in physics, chemistry, biology, economics and engineering allowing us to get deep insight into the operations of those systems.

The ever increasing demand for real-time simulation, control and prediction of complex systems, places a heavy burden on the shoulders of computational mathematicians, since appropriate mathematical modeling often leads to optimal control problems of dynamical system (DS), where the governing equations are partial differential equations (PDE).

The past several decades produced major advances in techniques for solving DS numerically, thanks to the increase in computational power and speed-up in numerical algorithms. Many essential problems that were impossible to solve a few decades before can be solved trivially and routinely nowadays.

However, there are still many of the tools in either simulation and control that have gone largely unused, because standard spatial discretization schemes for high resolution DS usually lead to very large-scale, high-dimensional and in general nonlinear systems of ordinary differential equations.

In order to perform robust simulation and active control of complex virtual models, model order reduction (MOR) or control order reduction (COR) as a branch from system and control theory provide an attractive approach to approximate large-scale discretized systems of state equations using low-dimensional Reduced-Order Modeling (ROM) (Figure 1.1).

In the former case of MOR, one seeks to find a low-dimensional approximation for a high-dimensional DS, based on a few dominant modes of the underlying DS, resulting in simulation

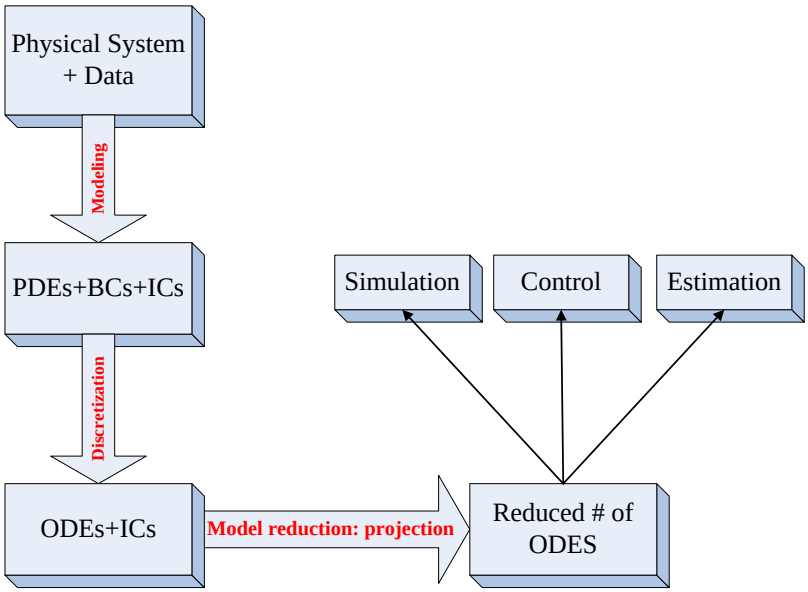


Figure 1.1: Process to a Reduced-Order Modeling

with reduced computational complexity. It is required that such a simplification preserve the essential input-output behaviors (Figure 1.2) of the high-fidelity, but with relatively small approximation errors. It is also required that the reduction procedure should be reliable, computationally efficient and restricted to a limited storage capacity.

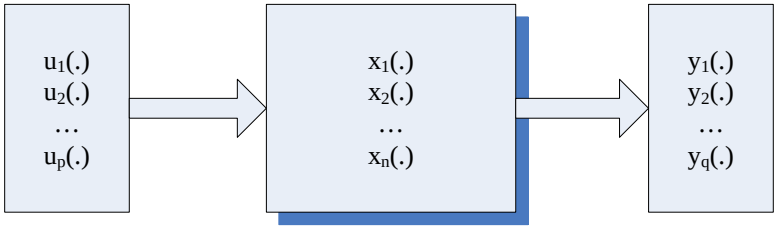


Figure 1.2: Input-output systems

The basic MOR methods for DS can be classified into three general categories as shown in Figure 1.3: Singular Value Decomposition (SVD) based methods, Krylov methods and combination into so-called SVD-Krylov methods.

Using the SVD methodology, one can obtain an optimal lower-rank approximation of a matrix, as measured by l_2 norm. One of the SVD-based approach designed for linear DS is

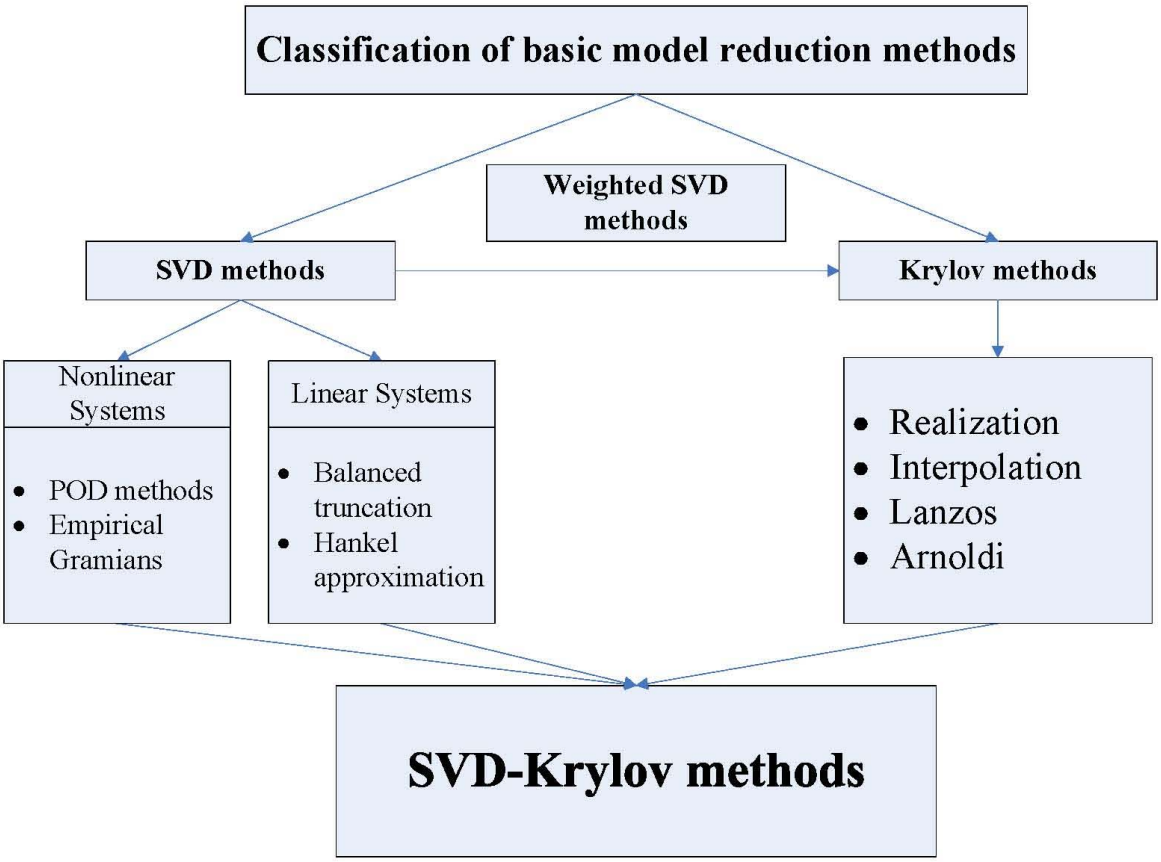


Figure 1.3: Classification of basic Reduced-Order Modeling methodologies

called Balanced Model Truncation (BMT). The central concept of BMT is to find a ROM of the high-fidelity linear DS such that the degree of reachability and degree of observability of each state are the same. This is achieved by simultaneously diagonalizing the reachability and the observability Grammian. BMT was first introduced by Mullis and Roberts [1] and later in the systems and control literature by Moore [2]. One of the BMT methods is called Lyapunov Balanced Reduction (LBR), which is implemented by solving reachability and the observability Lyapunov equations simultaneously. The stability of LBR was found by Pernebo and Silverman [3] and a bound on the approximation error of LBR was provided by Enns [4]. Besides the LBR, there are other types of balancing methods. The stochastic balancing method was first proposed by Desai and Pal [5] for balancing stochastic systems and later generalized by Green [6, 7]. The relative error bound for stochastic balancing is

due to Green [6]. A closely related balancing method is positive real balancing by Desai and Pal [5], which is applied for model reduction of positive real DS by solving two positive real Riccati equations. Another method which also requires solving two Riccati equations, is bounded real balancing which is applied to the bounded real systems. This method, together with the absolute error bound, was first introduced by Opdenacker and Jonckheere [8]. Also, LQG balancing referred to as the closed-loop balancing was introduced by Jonckheere and Silverman [9]. In the meantime, LBR was extended to the frequency weighted balanced reduction by Enns [4]. Stability of frequency weighted balancing methods was studied by Wang [10] and Zhou [11]. For a detailed survey on balancing related model reduction methods, please see Gugercin and Antoulas [12]. A closely related method based on BMT is called Optimal Hankel Norm Reduction [13], in which the truncated system of size can be calculated and specified given a so-called Hankel norm.

Another SVD-based methodologies designed for nonlinear DS is well known as Proper Orthogonal Decomposition (POD). The POD ROM method essentially identifies an orthonormal basis for representing the given data in a certain least squares optimal sense. Historically, POD goes by the names of Karhunen-Loeve decomposition (KLD) [14, 15] or Principal Components Analysis (PCA) and before them it was discovered by Kosambi [16]. The method originated in the work of Pearson [17] who invented the principal component analysis (PCA) which involves a mathematical procedure that transforms a number of possibly correlated variables into a smaller number of uncorrelated variables called principal components. It was also put forward in statistical framework by Hotelling [18]. POD was introduced in the context of analysis of turbulent flow by Lumley [19], Berkooz et al. [20]. Sirovich [21] introduced the idea of snapshots. See the book of Holmes [22] and the book of Michael Kirby [23]. Wiener [24] proposed Polynomial chaos decomposition as an extension to KLD. Dongbin Xiu [25] presented a new method for solving stochastic differential equations based on Galerkin projections and extensions of Wieners' Polynomial Chaos [24].

However, there are some drawbacks of POD ROM in a transient DS, where dominant flow structures tend to change significantly as the underlying DS traverses from unstable trajectories to reference attractors. The price of the low-dimensionality entails a lack of stability especially for transitional and turbulent flows (Couplet et al., 2005 [39]; Noack et al., 2010 [27]; Galletti et al., 2004 [37]; Gloerfelt, 2006 [40]). This either restricts reduced order models to a narrow range of parameters or to a short-time integration span. To improve

the accuracy of POD-Galerkin models, the effect of these unresolved modes must be included to provide insight into the turbulent energy.

A modification of POD ROM consists of calibration in flow problems and adding a shift-mode so that it includes an accurate representation of the unstable steady solution [26]. Noack et al proposed a systematic strategy to eliminate dynamic degrees of freedoms in Galerkin systems of incompressible fluid flows. The proposed system reduction strategy was derived from a Finite-Time Thermodynamics closure [27]. Also, the POD ROM method heavily relies on the input of DS and the time instances at which the snapshots are taken. Consequently, singular values and modes obtained by POD ROM are not invariants of DS. Also, Abury [28] studied POD ROM for Kuramoto-Sivashinsky equation and found that a model based on the leading six POD modes could not reproduce the right dynamics, even though those six POD modes represent 99.9995% of the variance. Similar problems with models based on POD modes were reported by Armbruster et al. (1992) [29] in a study of Kolmogorov flow in a regime of bursting behavior. Majda [30] studied Charney-DeVore model [34] and compared POD methods, optimal persistence patterns(OPPs) (introduced by DelSole [32])and principal interaction patterns(PIPs) (introduced by Hasselmann [33]) It is shown that the PIPs and OPP based ROM methods are superior to the POD based ROM methods.

Nevertheless, POD ROM method has been widely applied to high-complexity linear DS as well as nonlinear DS, due to its tractable computation using SVD eigenvalue solvers. Recently, the method of empirical Grammians has been proposed to remedy the issues arising in POD methods, at the expense of added computational complexity. Furthermore, snapshots-based balanced truncation was developed by Lall [45] and Rowley [46].

A different type of MOR method, mainly based on iterative solver, is Krylov-based approximation. These methods do not depend on the SVD. Instead, they relie on moment matching of the impulse response of the system. If the moment matches infinity, two widely used methods fall under this category, namely the Lanczos [48] and the Arnoldi [49] procedures. Otherwise, if the moment matches zero, the problem becomes Pade approximation. For general matching at an arbitrary point, the problem becomes the so-called rational Lanczos procedure. For an overview of relavant materials, please refer to Antoulas and Sorensen [50]. It turns out to be that SVD-based methodologies have a number of desirable features, namely, there exists a local/global error bound and the

stability is preserved. The drawback is that they can only be applied to DS with relatively low dimensions. On the other hand, the Krylov-based methodologies are iterative in nature and thus can be applied to DS with very high dimensions. The drawbacks are that the resulting ROM has no guaranteed error bound and stability is not necessarily preserved. These considerations lead to the so-called SVD-Krylov-based methodologies, which aims to combine the advantages and eliminate the disadvantages of two approximation methodologies described above. In this dissertation, we are focused on the SVD-based methodologies using ARPACK package. The ARPACK [51] package is designed by Lehoucq, Sorensen and Yang to compute a few eigenvalues and corresponding eigenvectors of large sparse or structured matrices, using the Implicitly Restarted Arnoldi Method (IRAM) or, in the case of symmetric matrices, the corresponding variant of the Lanczos algorithm.

In the latter case of COR, one seeks to find a way to drastically decrease the dimension of the control space without significantly compromising the quality of optimization but sizably reducing the cost in memory and CPU time. The difficulties encountered in COR, as opposed to just MOR, have been a significant target of systems and control theory during the last twenty years or so. A control reduction methodology based on POD, referred to as POD 4-D Var, has been widely used with tractable computations. One approach consists of an ad-hoc adaptive method, namely ad-hoc POD 4-D Var, in which new snapshots are regularly determined during the optimization process when the effectiveness of the existing POD ROM to represent the DS is considered to be insufficient. For recent work on ad-hoc POD 4-D Var, see Hinze and Kunish [52, 53, 54, 55, 56], Cao [58], Fang [62, 63, 64, 65, 66], Vermeulen [67], Luo [60, 61], Sachs [68] and Altaf [69]. However, the appropriate criteria for updating a POD ROM are not yet known in the application to optimal control. This is due to the limited validity of the POD ROM for inverse problems. Therefore, the classical Trust-Region (TR) approach [70, 71] combined with POD (TRPOD) was recently proposed as a way to overcome the above difficulties. The trust region method goes back to [72, 73, 74]. See also work of [75] followed by important work of [76, 77]. Finally the terminology of trust region and Cauchy point was put forward by [78] and systematized by [79]. The trust-region proper orthogonal decomposition (TRPOD) was recently proposed in [81, 82] as a way to overcome difficulties related to POD reduced order modeling (ROM) used for solving partial differential equations (PDE) constrained optimization problem. For a comprehensive survey on the techniques combining POD with the concept of trust-region (TRPOD) with general model functions,

see Toint [83, 85]. For an introduction to trust region methods, see Nocedal and Wright [70]. TRPOD presents a framework for decision as to when an update of the POD-ROM is necessary during the optimization process. Moreover, from a theoretical point of view, we have a global convergence result for TRPOD [81] proving that the iterates produced by the optimization algorithm, started at an arbitrary initial iterate, will converge to a local optimizer for the original model. In order to reduce the POD basis size and still achieve the global convergence, another novel method was proposed to incorporate information from the 4-D Var system into the ROM procedure by implementing a dual weighted POD (DWPOD) method by Daescu [102]. Yaremchuk [103] proposed a version of the reduced 4-D Var in a sequence of low-dimensional subspaces of the control space. This method does not require development of the tangent linear, adjoint and POD model for implementation. Vahid [104] presented a version of Equation-Free/Galerkin-Free Reduced-Order Modeling of the Shallow Water Equations Based on Proper Orthogonal Decomposition.

The contributions of this dissertation are as follows. The first novelty in the dissertation is to study a new methodology combining the dual weighted snapshots selection and trust region POD adaptivity (DWTRPOD). The second novelty in the dissertation is to combine the incremental POD 4-D Var, balanced truncation techniques and method of snapshots methodology. In the linear DS, this is done by integrating the linear forward model many times using different initial conditions in order to construct an ensemble of snapshots so as to generate the forward POD modes. Those forward POD modes then serve as the initial conditions for its corresponding adjoint system. We then integrate the adjoint system a large number of times based on different initial conditions generated by the forward POD modes to construct an ensemble of adjoint snapshots. From this ensemble of adjoint snapshots, we can generate an ensemble of so-called adjoint POD modes. Thus we can approximate the controllability Grammian of the adjoint system instead of solving the computationally expensive coupled Lyapunov equations. To sum up, in the incremental POD 4-D Var, we can approximate the controllability Grammian by integrating the TLM a large number of times and approximate observability Grammian by integrating its adjoint also a large number of times.

The third novelty is to extend the snapshots based POD methodology to the nonlinear system. Furthermore, we modify the classical algorithms in order to save the computations even more significantly. We proposed a novel idea to construct an ensemble of snapshots

by integrating the tangent linear model (TLM) only once, based on which we can obtain its TLM POD modes. Then each TLM POD mode will be used as an initial condition to generate a small ensemble of adjoint snapshots and their adjoint POD modes. Finally, we can construct a large ensemble of adjoint POD modes by putting together each small ensemble of adjoint POD modes. To sum up, we can approximate the controllability Grammian by integrating TLM only once and approximate observability Grammian by integrating adjoint model only a reduced number of times.

The plan of this dissertation is as follows. In chapter 2, we briefly review the time and frequency domain spaces and their corresponding norms. In chapter 3, we discuss the classical model reduction methodologies widely used for linear systems. In chapter 4, we focus on the theories and issues arising the Proper Orthogonal Decomposition techniques, followed by chapter 5 in which we discuss the POD reduced 4-D Var assimilation approaches. Then, in chapter 6, we explor the feasibility of carrying out a modular structured variational data assimilation (VDA) using a finite-element method of the nonlinear shallow water equations model on a limited area domain. In chapter 7, we address the POD model reduction along with inverse solution of a two-dimensional finite-element shallow-water equations model on a limited area domain. In chapter 8, we address a POD model reduction along with inverse solution of a two-dimensional global shallow water equations model. Our intention in this chapter is to generalize the efficient state-of-the-art POD implementation from our previous chapter on finite element SWE on the limited area to global FV-SWE model with realistic initial conditions. Finally the dissertation concludes with a summary and conclusions chapter 9, in which directions of future research are also outlined.

CHAPTER 2

SPACES AND NORMS

2.1 Norms of finite-dimensional vectors and matrices

Let $\mathbb{V} \to \mathbb{R}$ be a linear space over the field of reals $\mathbb{R}$ or complex numbers $\mathbb{C}$. A norm on $\mathbf{V}$ is a function

$$v: \ \mathbb{V} \to \mathbb{R}$$

such that the following three properties are satisfied:

- strict positiveness: $v(\mathbf{x}) \geq \mathbf{0} \ \forall \mathbf{x} \in \mathbb{V}$ with equality iff $x = 0$
- triangle inequality: $v(\mathbf{x}+\mathbf{y}) \leq v(\mathbf{x}) + v(\mathbf{y}) \ \ \forall \mathbf{x}, \mathbf{y} \in \mathbb{V}$
- positive homogeneity: $v(\alpha \mathbf{x}) = |\alpha| \, v(\mathbf{x}) \ \ \forall \alpha \in \mathbb{C}, \ \forall \mathbf{x} \in \mathbb{V}$

For any vector $\mathbf{x} = \begin{pmatrix} x_1 & \cdots & x_n \end{pmatrix}^{\mathbf{T}} \in \mathbb{C}^{\mathbf{n}}$, the Holder or $p-$norm is defined as follows:

$$\|\mathbf{x}\|_p = \begin{cases} (\sum_{i=1}^{n} |x_i|^p)^{\frac{1}{p}} & 1 \leq p < \infty \\ \\ \max\limits_{1 \leqslant i \leqslant n} |x_i| & p = \infty \end{cases}$$

An important class of matrix norms are those that are *induced* by the vector $p-$norm defined above. More precisely, for a given matrix $\mathbf{A} = (a_{ij}) \in \mathbb{C}^{n \times m}$, the induced $p-$norm is defined as follows:

$$\|\mathbf{A}\|_p = \sup_{x \neq 0} \frac{\|\mathbf{A}\mathbf{x}\|_p}{\|\mathbf{x}\|_p}$$

In particular, for $p = 1, 2, \infty$, the following expressions hold:

$$\|\mathbf{A}\|_p = \begin{cases} \max\limits_{1\leq j\leq m} \sum_{i=1}^{n} |a_{ij}| & p=1 \\ \max\limits_{1\leq i\leq n} \sum_{j=1}^{m} |a_{ij}| & p=2 \\ \left(\lambda_{max}\left(\mathbf{A}\mathbf{A}^*\right)\right)^{\frac{1}{2}} & p=\infty \end{cases}$$

where $\left(\lambda_{max}\left(\mathbf{A}\mathbf{A}^*\right)\right)^{\frac{1}{2}}$ denotes the square root of the largest eigenvalue of the positive-semidefinite matrix $\mathbf{A}\mathbf{A}^*$, $\mathbf{A}^*$ denotes the conjugate transpose of the matrix $\mathbf{A}$.

There exist other matrix norms besides the induced matrix norms. An important case is the Schatten $p-$norms. These non-induced norms are unitarily invariant. To define them, we introduce the singular value decomposition as follows:

Given a matrix $\mathbf{A} \in \mathbb{C}^{\mathrm{n}\times\mathrm{m}}$, there exist unitary matrices

$$\mathbf{U} = \begin{pmatrix} \mathbf{u_1} & \mathbf{u_2} & \dots & \mathbf{u_n} \end{pmatrix}, \ \mathbf{U}\mathbf{U}^* = \mathbf{I_n}$$

$$\mathbf{V} = \begin{pmatrix} \mathbf{v_1} & \mathbf{v_2} & \dots & \mathbf{v_m} \end{pmatrix}, \ \mathbf{V}\mathbf{V}^* = \mathbf{I_m}$$

where $\mathbf{I_n}$ and $\mathbf{I_m}$ denote $n \times n$ identity matrix and $m \times m$ identity matrix respectively, such that

$$\mathbf{A} = \mathbf{U}\sum\mathbf{V}^*$$

where

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_1 & 0 \\ 0 & 0 \end{pmatrix} \in \mathbb{C}^{\mathrm{n}\times\mathrm{m}}$$

$$\mathbf{\Sigma}_1 = \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{pmatrix} \in \mathbb{R}^{\mathrm{r}\times\mathrm{r}}$$

with $r = \mathrm{Rank}\left(\mathbf{A}\right)$ and the singular values

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$$

while the columns of $\mathbf{U}$ and $\mathbf{V}$ are called the left and right singular vectors of $\mathbf{A}$, respectively.

Hence, for a given matrix $\mathbf{A} = (a_{ij}) \in \mathbb{C}^{n\times m}$ for $m \leq n$, the non-induced Schatten $p-$norm is defined as follows:

$$\|\mathbf{A}\|_{s,p} = \begin{cases} \left(\sum_{i=1}^{m} \left(\sigma_i \left(\mathbf{A}\right)\right)^p\right)^{\frac{1}{p}} & 1 \leq p < \infty \\ \\ \sigma_{max}\left(\mathbf{A}\right) & p = \infty \end{cases}$$

where $\sigma_i \left(\mathbf{A}\right)$ can be viewed as the *singular values* of $\mathbf{A}$ or as the square root of the ith largest eigenvalues of $\mathbf{A}\mathbf{A}^*$

Similarly, in particular, for $p = 1, 2, \infty$, the following expressions hold:

$$\|\mathbf{A}\|_{s,p} = \begin{cases} \sum_{i=1}^{m} \sigma_i \left(\mathbf{A}\right) & p = 1 \quad \text{trace norm} \\ \\ \left(\sum_{i=1}^{m} \left(\sigma_i \left(\mathbf{A}\right)\right)^2\right)^{\frac{1}{2}} = \text{trace}\left(\mathbf{A}^*\mathbf{A}\right) & p = 2 \quad \text{Frobenius norm} \\ \\ \sigma_{max}\left(\mathbf{A}\right) = \left(\lambda_{max}\left(\mathbf{A}\mathbf{A}^*\right)\right)^{\frac{1}{2}} & p = \infty \quad \text{spectral} - \text{norm} \end{cases}$$

2.2 Time and frequency domain spaces and norms

Consider a linear space $\mathbb{V}$ defined over $\mathbb{R}$, not necessarily finite dimensional. Let a norm be defined on$\mathbb{V}$, satisfying the strict positiveness, triangle inequality and positive homogeneity, then $\mathbb{V}$ is called a *normed space*. In such spaces the concept of convergence can be defined as follows. We say that a sequence $\{x_k\}_{k=1}^{\infty} \in \mathbb{V}$ converges to $\mathbf{x}_*$ if the sequence of real numbers $v\left(\mathbf{x}_k - \mathbf{x}_*\right)$=$\|\mathbf{x}_k - \mathbf{x}_*\|$ goes to zero as k goes to infinity. A sequence $\{x_k\}_{k=1}^{\infty} \in \mathbb{V}$ is a *Cauchy sequence* if for all $\epsilon > 0$, there exists an integer N such that $\|\mathbf{x}_p - \mathbf{x}_q\| < \epsilon$ for all $p, q > N$. If every Cauchy sequence converges, then $\mathbb{V}$ is called *complete*.

2.2.1 Banach and Hilbert spaces

A Banach space is a normed linear space $\mathbb{V}$ that is complete defined over the field of scalars $\mathbb{F}$ that can either be the field of reals $\mathbb{R}$ or the field of complex numbers $\mathbb{C}$. Hilbert spaces have more structure than Banach spaces with additional structure resulting from the existence of an *inner product*. The inner product is a function from the Cartesian product $\mathbb{V} \times \mathbb{V} \to \mathbb{R}$:

$$\langle\ ,\ \rangle : \mathbb{V} \times \mathbb{V} \to \mathbb{R},\ (\mathbf{x}, \mathbf{y}) \mapsto \langle \mathbf{x}, \mathbf{y} \rangle \in \mathbb{R}$$

$$v: \ \mathbb{V} \to \mathbb{R}$$

such that the following four properties are satisfied:

- strict positiveness: $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0 \ \forall \mathbf{x} \in \mathbb{V}$ with equality iff $x = 0$
- linearity in the first argument: $\langle \alpha \mathbf{x} + \beta \mathbf{y}, \mathbf{z} \rangle = \alpha \langle \mathbf{x}, \mathbf{z} \rangle + \beta \langle \mathbf{y}, \mathbf{z} \rangle \ \ \forall \mathbf{x}, \mathbf{y} \in \mathbb{V}, \ \forall \alpha, \beta \in \mathbb{F}$
- conjugate symmetry:$\langle \mathbf{x}, \mathbf{y} \rangle^* = \langle \mathbf{y}, \mathbf{x} \rangle$

This inner product *induces* a norm on $\mathbb{V}$, namely,

$$x \longmapsto \|\mathbf{x}\| = \langle \mathbf{x}, \mathbf{x} \rangle^{\frac{1}{2}}$$

2.2.2 The time-domain Lebesgue spaces l_p and L_p

l_p spaces

Considering the discretized vector-valued sequences $\mathbf{f} : \mathbb{Z}_+ \to \mathbb{R}^n$, the Holder $p-$norms of these sequences can be defined as:

$$\|\mathbf{f}\|_{l_p} = \begin{cases} \left(\sum_{i \in \mathbb{Z}^+} \|\mathbf{f}(i)\|_p^p \right)^{\frac{1}{p}} & 1 \leq p < \infty \\ \sup_{i \in \mathbb{Z}^+} \|\mathbf{f}(i)\|_p & p = \infty \end{cases}$$

The corresponding l_p spaces are defined as

$$l_p^n(\mathbb{Z}_+) = \left\{ \mathbf{f} : \mathbb{Z}_+ \to \mathbb{R}^n, \ \|\mathbf{f}\|_{l_p} < \infty \right\}$$

L_p spaces

Similarly, considering the continuous-time vector-valued functions $\mathbf{f} : \mathbb{R}_+ \to \mathbb{R}^n$, the Holder $p-$norms of these functions can be defined as:

$$\|\mathbf{f}\|_{L_p} = \begin{cases} \left(\int_{t \in \mathbb{R}^+} \|\mathbf{f}(t)\|_p^p \right)^{\frac{1}{p}} & 1 \leq p < \infty \\ \sup_{t \in \mathbb{R}^+} \|\mathbf{f}(t)\|_p & p = \infty \end{cases}$$

$$L_p^n(\mathbb{R}_+) = \left\{ \mathbf{f} : \mathbb{R}_+ \to \mathbb{R}^n, \ \|\mathbf{f}\|_{L_p} < \infty \right\}$$

2.2.3 The frequency-domain Lebesgue spaces h_p and H_p

h_p spaces

considering the complex matrix-valued function $\mathbf{F} : \mathbb{C} \to \mathbb{C}^{n\times m}$, which is defined over the analytic closed unit disc $\overline{D}$, the Schatten $p-$norms of these functions can be defined as:

$$\|\mathbf{F}\|_{h_p} = \begin{cases} \left(\frac{1}{2\pi} \sup_{|r|<1} \int_0^{2\pi} \left\| \mathbf{F}\left(re^{i\theta}\right) \right\|_p^p d\theta \right)^{\frac{1}{p}} & 1 \le p < \infty \\ \sup_{z\in\overline{D}} \|\mathbf{F}(\mathbf{z})\|_p = \sup_{z\in\overline{D}} \sigma_{max}\left(\mathbf{F}(\mathbf{z})\right) & p = \infty \end{cases}$$

The corresponding h_p spaces are defined as

$$h_p^{n\times m}\left(\overline{D}\right) = \left\{ \mathbf{F} : \mathbb{C} \to \mathbb{C}^{n\times m},\ \|\mathbf{F}\|_{h_p} < \infty \right\}$$

H_p spaces

Let $\mathbb{C}_+ \subset \mathbb{C}$ denote the (open) right half of the complex plane:$s = x + iy \in \mathbb{C}$,$x > 0$. Considering the complex matrix-valued function $\mathbf{F} : \mathbb{C} \to \mathbb{C}^{n\times m}$, which is defined over the analytic closed unit disc $\mathbb{C}_+$, the Schatten $p-$norms of these functions can be defined as:

$$\|\mathbf{F}\|_{H_p} = \begin{cases} \left(\sup_{x>0} \int_{-\infty}^{\infty} \|\mathbf{F}(x+iy)\|_p^p d\theta \right)^{\frac{1}{p}} & 1 \le p < \infty \\ \sup_{z\in\mathbb{C}_+} \|\mathbf{F}(\mathbf{z})\|_p = \sup_{z\in\mathbb{C}_+} \sigma_{max}\left(\mathbf{F}(\mathbf{z})\right) & p = \infty \end{cases}$$

The corresponding H_p spaces are defined as

$$H_p^{n\times m}\left(\mathbb{C}_+\right) = \left\{ \mathbf{F} : \mathbb{C} \to \mathbb{C}^{n\times m},\ \|\mathbf{F}\|_{H_p} < \infty \right\}$$

The search for the $\|\mathbf{F}\|_{H_\infty}$ above can be simplified by making use of the *maximum modulus theorem*, which states that a function $\mathbf{F}$ continuous inside a domain $D \subset \mathbb{C}$ as well as on its boundary ∂D and analytic inside D attains its maximum on the *boundary* ∂D of D. Hence, we obtain

$$\|\mathbf{F}\|_{h_\infty} = \sup_{\theta\in[0,2\pi]} \sigma_{max}\left(\mathbf{F}\left(e^{i\theta}\right)\right)$$

$$\|\mathbf{F}\|_{H_\infty} = \sup_{y\in\mathbb{R}} \sigma_{max}\left(\mathbf{F}\left(iy\right)\right)$$

Sobolev spaces $H^1(\Omega)$ and $H_0^1(\Omega)$

Let Ω be a bounded domain in $\mathbb{R}^d$,

$$H^1 = H^1(\Omega) = \left\{ v(x),\ x \in \Omega :\ v \in L_2(\Omega)\ \frac{\partial v}{\partial x_j} \in L_2(\Omega),\ j = 1, \ldots, d \right\}$$

$$H_0^1 = H_0^1(\Omega) = \left\{ v(x),\ x \in \Omega : v \in H^1(\Omega),\ v = 0,\ x \in \partial\Omega \right\}$$

$$H_g^1 = H_g^1(\Omega) = \left\{ v(x),\ x \in \Omega : v \in H^1(\Omega),\ v = g,\ x \in \partial\Omega \right\}$$

CHAPTER 3

LINEAR SYSTEMS AND MODEL TRUNCATION

3.1 Linear state-space systems

We consider linear state-space systems

$$G: \begin{cases} \dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} \quad \mathbf{x}(0) = \mathbf{x}_0 \\ \\ \mathbf{y} = \mathbf{C}\mathbf{x} + \mathbf{D}\mathbf{u} \end{cases}$$

with state $\mathbf{x}(t) \in \mathbb{R}^n$, input $\mathbf{u}(t) \in \mathbb{R}^m$, and output $\mathbf{y}(t) \in \mathbb{R}^p$. Let $\phi(\mathbf{u}, \mathbf{x}, t)$ denote the solution of the state equations . In particular, for the continuous-time state equations the solution can be written as

$$\phi(\mathbf{u}, \mathbf{x}, t) = e^{\mathbf{A}t}\mathbf{x}_0 + \int_0^t e^{\mathbf{A}(t-\tau)}\mathbf{B}\mathbf{u}(t)\, d\tau,\ t \geq 0$$

$$\mathbf{y}(t) = \mathbf{C}\phi(\mathbf{u}, \mathbf{x}, t) + \mathbf{D}\mathbf{u}(t),\ t \geq 0$$

where

$$e^{\mathbf{A}t} = \mathbf{I}_n + \frac{t}{1!}\mathbf{A} + \frac{t^2}{2!}\mathbf{A}^2 + \cdots + \frac{t^k}{k!}\mathbf{A}^k + \cdots$$

Furthermore, the linear state-space systems G with m inputs and p outputs can be viewed as an operator mapping the input space to the output space, in particular, we will be concerned with systems which may be written by means of *convolution integral*

$$G:\ u \longrightarrow y,\ y(t) = \int_{-\infty}^{\infty} \mathbf{h}(t, \tau)\, \mathbf{u}(\tau) d\tau,\ t \in \mathbb{R}$$

where $\mathbf{h}(t, \tau)$ is a matrix-valued function called the *kernel* of system G

Let the input $\mathbf{u}(t) \in \mathbb{R}^m$ be the impulse function that may be represented mathematically by a Dirac delta function as follows:

$$\delta(t) =: \begin{cases} \infty & t = 0 \\ \\ 0 & t \neq 0 \end{cases}$$

Then in the time domain, $\mathbf{h}$ is called the impulse response and can be written as

$$\mathbf{h}(t) = \begin{cases} \mathbf{C}e^{\mathbf{A}t}\mathbf{B} + \delta(t) D & t \geq 0 \\ \\ 0 & t < 0 \end{cases}$$

It is easier to perform the analysis in the frequency domain for linear state-space systems. In order to convert to the frequency domain, we need apply the Laplace Transform to the impulse response $\mathbf{h}(t)$ to determine the transfer function $\mathbf{H}(s)$ of the system G.

$$\mathbf{H}(s) = \mathcal{L}(\mathbf{h}(t)) = \int_{-\infty}^{\infty} e^{-st}\mathbf{h}(t)dt = \mathbf{D} + \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} \in H_p^{n\times m}$$

where $s \in \mathbb{C}$ is the complex frequency.

Hence, the Laplace Transform converts linear differential equations into algebraic expressions which are easier to manipulate. The Laplace Transform also converts functions with a real dependent variable (such as time) into functions with a complex dependent variable(such as frequency).

As a measure of system size and to measure the distance between two different systems, we regularly use the $H_\infty-$norm

$$\|\mathbf{H}\|_{H_\infty} = \sup_{s\in\mathbb{C}_+} \sigma_{max}(\mathbf{H}(\mathbf{s}))$$

where $\mathbb{C}_+$ denotes the open complex right-half plane, $\sigma_{max}(\mathbf{H}(\mathbf{s}))$ denotes the largest singular value of the matrix $(\mathbf{H}(\mathbf{s}))$in the MIMO(Multi-input-Multi-Output) case. In the SISO(Single-Input-Single-Output) case, this is equal to the magnitude of the complex number $(\mathbf{H}(\mathbf{s}))$.

Furthermore, $\|\mathbf{H}\|_{H_\infty}$ is finite iff $\mathbf{H}(\mathbf{s})$is stable, i.e, $\mathbf{H}(s)$ has no poles in the closed right complex half planes $\mathbb{C}_+$ and we obtain

$$\|\mathbf{H}\|_{H_\infty} = \sup_{y\in\mathbb{R}} \sigma_{max}(\mathbf{H}(iy))$$

3.2 Approximation criteria and projection

A reduced-order systems or an approximation of G of its order n is a state-space system G_r

$$G_r : \begin{cases} \dot{\mathbf{z}} = \mathbf{A_r z} + \mathbf{B_r u} \quad \mathbf{z}(0) = \mathbf{z}_0 \\ \\ \mathbf{y_r} = \mathbf{C_r z} + \mathbf{D_r u} \end{cases}$$

such that $\mathbf{z}(t) \in \mathbb{R}^r$ where $r << n$.

The main approximation criterion we are interested in is to make $\|G - G_r\|_{H_\infty}$ small, with the motivation to measure worst-case error of approximation to the original system G by the reduced-order system G_r. Other criteria include the relative criterion $\|G^{-1}(G - G_r)\|_{H_\infty}$ and the frequency-weighted criterion $\|W_1(G - G_r)W_2\|_{H_\infty}$

The approximation G_r can often be obtained by means of *orthogonal truncation* as follows:

Change the coordinates $\mathbf{x}(t) = \mathbf{T}\bar{\mathbf{x}}(t)$, where we find a *suitable* invertible matrix $\mathbf{T} \in \mathbb{R}^{n\times n}$ and transform the state-space model into

$$\bar{\mathbf{A}} = \mathbf{T}^{-1}\mathbf{A}\mathbf{T} = \begin{pmatrix} \bar{\mathbf{A}}_{11} & \bar{\mathbf{A}}_{12} \\ \bar{\mathbf{A}}_{21} & \bar{\mathbf{A}}_{22} \end{pmatrix},\ \bar{\mathbf{A}}_{\mathbf{11}} \in \mathbb{R}^{r\times r}$$

$$\bar{\mathbf{B}} = \mathbf{T}^{-1}\mathbf{B} = \begin{pmatrix} \bar{\mathbf{B}}_1 \\ \bar{\mathbf{B}}_2 \end{pmatrix},\ \bar{\mathbf{B}}_{\mathbf{1}} \in \mathbb{R}^{r\times m}$$

$$\bar{\mathbf{C}} = \mathbf{C}\mathbf{T} = \begin{pmatrix} \bar{\mathbf{C}}_1 & \bar{\mathbf{C}}_2 \end{pmatrix},\ \bar{\mathbf{C}}_{\mathbf{1}} \in \mathbb{R}^{p\times r}$$

$$\bar{\mathbf{D}} = \mathbf{D} \in \mathbb{R}^{p\times m}$$

Hence G_r is obtained by

$$\begin{pmatrix} \mathbf{A}_r & \mathbf{B}_r \\ \mathbf{C}_r & \mathbf{D}_r \end{pmatrix} = \begin{pmatrix} \bar{\mathbf{A}}_{11} & \bar{\mathbf{B}}_1 \\ \bar{\mathbf{C}}_{\mathbf{1}} & \bar{\mathbf{D}} \end{pmatrix}$$

Such an orthogonal truncation can also be viewed as a *projection* from the original state-space G in $\mathbb{R}^n$ to the reduced state-space $\mathbb{R}^r$.

3.3 Petrov projection and Galerkin projection

Generally, for non-orthogonal projection, we have the transformations

$$\mathbf{W^T} = \begin{pmatrix} \mathbf{I}_r & \mathbf{0}_{r\times(n-r)} \end{pmatrix} \mathbf{T^{-1}},\ :\mathbb{R}^n \to \mathbb{R}^r$$

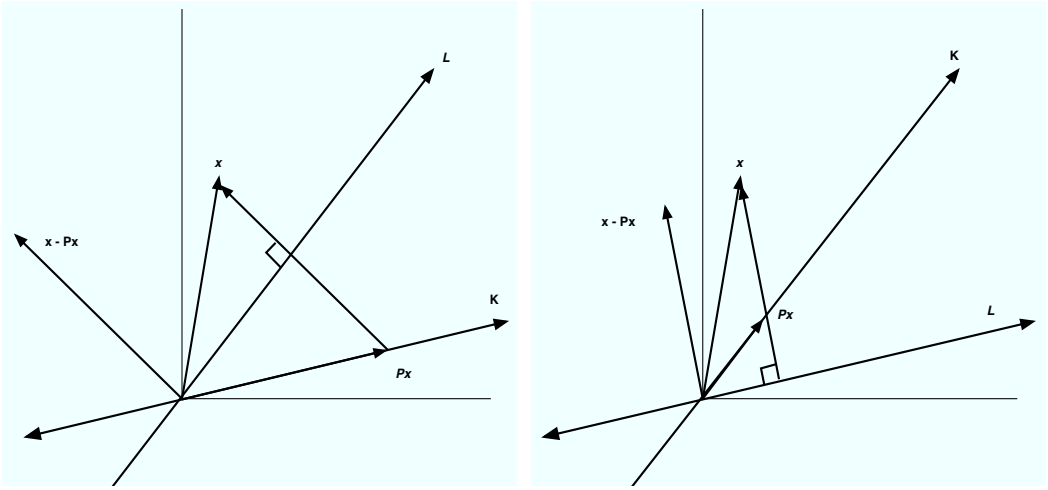
$$\mathbf{V} = \mathbf{T} \begin{pmatrix} \mathbf{I}_r \\ \mathbf{0^T}_{r\times(n-r)} \end{pmatrix},\ :\mathbb{R}^r \to \mathbb{R}^n$$

Notice that $\mathbf{W}$ and $\mathbf{V}$ satisfy

$$\mathbf{W^T V} = \mathbf{I_r},\ \mathbf{VW^T} = \left(\mathbf{VW^T}\right)\left(\mathbf{VW^T}\right)$$

Such a projection is called a *Petrov Galerkin projection.* If $\mathbf{W^T} = \mathbf{V^T}$, i.e., $\mathbf{T}$ is an orthogonal matrix, then the projection is called a Galerkin projection. For Petrov Galerkin projection, we have that

$$\begin{pmatrix} \mathbf{A}_r & \mathbf{B}_r \\ \mathbf{C}_r & \mathbf{D}_r \end{pmatrix} = \begin{pmatrix} \mathbf{W^T AV} & \mathbf{W^T B} \\ \mathbf{CV} & \mathbf{D} \end{pmatrix}$$



(a) Petrov Galerkin projection (b) Galerkin Projection

Figure 3.1: Petrov Projection and Galerkin Projection

to understand the Petrov Galerkin projection, we can consider the following analysis. Assuming we want to try to express the solution $\mathbf{x}(t) \in \mathbb{R}^n$ of the original model G only in r variables in the reduced-order model G_r, we can write

$$\mathbf{x}(t) = \mathbf{V}\mathbf{z}(t) \in \text{Range}\,(\mathbf{V})\,,\ \mathbf{x}(t) \in \mathbb{R}^n, \mathbf{z}(t) \in \mathbb{R}^r$$

so that the residual can be expressed as

$$\mathbf{R} = \mathbf{V}\dot{\mathbf{z}} - \mathbf{A}\mathbf{V}\mathbf{z} - \mathbf{B}\mathbf{u}$$

The Petrov Galerkin projection requires that the projection of residual $\mathbf{R}$ into Range $(\mathbf{W})$ is zero, which can be represented by

$$\mathbf{W}^\mathbf{T}\mathbf{R} = \mathbf{0}$$

Hence, we obtain the equation

$$\mathbf{W}^\mathbf{T}(\mathbf{V}\dot{\mathbf{z}} - \mathbf{A}\mathbf{V}\mathbf{z} - \mathbf{B}\mathbf{u}) = \dot{\mathbf{z}} - \mathbf{W}^\mathbf{T}\mathbf{A}\mathbf{V}\mathbf{z} - \mathbf{W}^\mathbf{T}\mathbf{B}\mathbf{u} = \mathbf{0}$$

Therefore,

$$\dot{\mathbf{z}} = \mathbf{W}^\mathbf{T}\mathbf{A}\mathbf{V}\mathbf{z} + \mathbf{W}^\mathbf{T}\mathbf{B}\mathbf{u}$$

3.4 Balanced Truncation

Let $\mathbf{D} = \mathbf{0}$ in linear state-space systems G, so we obtain the impulse response, which can be written as

$$\mathbf{h}\,(t) = \mathbf{C}e^{\mathbf{A}t}\mathbf{B}\ t \geq 0$$

Let $\mathbf{h_r}\,(t) = e^{\mathbf{A}t}\mathbf{B}$, $t \geq 0$ and $\mathbf{h_o}\,(t) = \mathbf{C}e^{\mathbf{A}t},\ t \geq 0$ be the input-to-state and the state-to-output response of the system G, respectively. The *reachability Grammian* is then defined as

$$\mathbf{P} = \int_0^\infty \mathbf{h_r}\,(t)\,\mathbf{h_r^*}\,(t)\,dt$$

while the *observability Gramian* is defined as

$$\mathbf{Q} = \int_0^\infty \mathbf{h_o}\,(t)\,\mathbf{h_o^*}\,(t)\,dt$$

In practice, assuming $\mathbf{A}$ is Hurwitz space (all eigenvalues of are in the open left complex half plane), reachability Grammian $\mathbf{P}$ and observability Grammian $\mathbf{Q}$ are obtained as solutions to the following differential Lyapunov equations

$$\dot{\mathbf{P}} = \mathbf{AP} + \mathbf{PA^T} + \mathbf{BB^T} = \mathbf{0},\ \mathbf{P}(\mathbf{0}) = \mathbf{0}$$

$$\dot{\mathbf{Q}} = \mathbf{A^T Q} + \mathbf{QA} + \mathbf{C^T C} = \mathbf{0},\ \mathbf{Q}(\mathbf{0}) = \mathbf{0}$$

Let $t \to \infty$and we conveniently obtain algebraic Lyapunov equations

$$\mathbf{AP} + \mathbf{PA^T} + \mathbf{BB^T} = \mathbf{0}$$

$$\mathbf{A^T Q} + \mathbf{QA} + \mathbf{C^T C} = \mathbf{0}$$

The eigenvalues of the product of the controllability and observability Grammians play an important role in system theory and control. Those eigenvalues are invariant under coordinate transformations. Hence, we define the *Hankel singular values* as

$$\sigma_i = \left(\lambda_i(\mathbf{PQ})\right)^{\frac{1}{2}}$$

With the above quantification of observability and controllability, one might be tempted to prescribe some algorithm like eliminating the least observable or least controllable dimensions in the state space to yield a lower-order approximate model. The procedure has four steps

1. Compute the reachability gramian $\mathbf{P}$ and and observability gramian $\mathbf{Q}$
2. Compute the Cholesky factor $\mathbf{R}$ of $\mathbf{P}$, that is $\mathbf{Q} = \mathbf{R^T R}$
3. Compute the singular value decomposition $\mathbf{RPR^T} = \mathbf{U\Sigma^2 U^T}$
4. Use the coordinate transformation $\mathbf{x}(t) = \mathbf{T}\bar{\mathbf{x}}(t)$, where $\mathbf{T} = \mathbf{R^{-1}U\Sigma}^{\frac{1}{2}}$
5. Compute the truncation G_r after the transformation of the state-space model G

$$\bar{\mathbf{A}} = \mathbf{T}^{-1}\mathbf{A}\mathbf{T} = \begin{pmatrix} \bar{\mathbf{A}}_{11} & \bar{\mathbf{A}}_{12} \\ \bar{\mathbf{A}}_{21} & \bar{\mathbf{A}}_{22} \end{pmatrix},\ \bar{\mathbf{A}}_{\mathbf{11}} \in \mathbb{R}^{r\times r}$$

$$\bar{\mathbf{B}} = \mathbf{T}^{-1}\mathbf{B} = \begin{pmatrix} \bar{\mathbf{B}}_1 \\ \bar{\mathbf{B}}_2 \end{pmatrix},\ \bar{\mathbf{B}}_{\mathbf{1}} \in \mathbb{R}^{r\times m}$$

$$\bar{\mathbf{C}} = \mathbf{C}\mathbf{T} = \begin{pmatrix} \bar{\mathbf{C}}_1 & \bar{\mathbf{C}}_2 \end{pmatrix},\ \bar{\mathbf{C}}_{\mathbf{1}} \in \mathbb{R}^{p\times r}$$

$$\bar{\mathbf{D}} = \mathbf{D} \in \mathbb{R}^{p\times m}$$

$$\begin{pmatrix} \mathbf{A}_r & \mathbf{B}_r \\ \mathbf{C}_r & \mathbf{D}_r \end{pmatrix} = \begin{pmatrix} \bar{\mathbf{A}}_{11} & \bar{\mathbf{B}}_1 \\ \bar{\mathbf{C}}_{\mathbf{1}} & \bar{\mathbf{D}} \end{pmatrix}$$

Under the new coordinate system, it is easy to verify that

$$\bar{\mathbf{P}} = \mathbf{T}^{-1}\mathbf{P}\mathbf{T}^{-\mathbf{T}},\ \bar{\mathbf{Q}} = \mathbf{T}^{\mathbf{T}}\mathbf{Q}\mathbf{T}$$

Such a similarity transformation is called *balancing transformation*, since

$$\bar{\mathbf{P}} = \left(\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{U}^{\mathbf{T}}\mathbf{R}\right)\mathbf{P}\left(\mathbf{R}^{\mathbf{T}}\mathbf{U}\mathbf{\Sigma}^{-\frac{1}{2}}\right)$$

$$= \left(\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{U}^{\mathbf{T}}\right)\mathbf{U}\mathbf{\Sigma}^{\mathbf{2}}\mathbf{U}^{\mathbf{T}}\left(\mathbf{U}\mathbf{\Sigma}^{-\frac{1}{2}}\right) = \left(\mathbf{\Sigma}^{-\frac{1}{2}}\right)\mathbf{\Sigma}^{\mathbf{2}}\left(\mathbf{\Sigma}^{-\frac{1}{2}}\right) = \Sigma$$

$$\bar{\mathbf{Q}} = \left(\mathbf{R}^{-\mathbf{1}}\mathbf{U}\mathbf{\Sigma}^{\frac{1}{2}}\right)^{\mathbf{T}}\left(\mathbf{R}^{\mathbf{T}}\mathbf{R}\right)\left(\mathbf{R}^{-\mathbf{1}}\mathbf{U}\mathbf{\Sigma}^{\frac{1}{2}}\right) = \left(\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{U}^{\mathbf{T}}\right)\left(\mathbf{U}\mathbf{\Sigma}^{\frac{1}{2}}\right) = \Sigma$$

Geometrically, balancing transformation obtain the observability and controllability ellipsoids so that they are identical and their principal axes corresponding to the left singular vectors of the Hankel singular value decomposition.

Approximation by balanced truncation preserves stability, and the H_∞- norm (the maximum of the frequency response) of the error system is bounded by twice the sum of neglected Hankel singular values

$$\|G - G_r\|_{H_\infty} \leq 2\left(\sigma_{r+1} + \ldots + \sigma_n\right)$$

CHAPTER 4

PROPER ORTHOGONAL DECOMPOSITION

4.1 Karhunen-Loeve Expansion

Let us consider an ensemble of snapshots written as:

$$\Phi = \{\phi_1, \phi_2, \ldots, \phi_n\} \tag{4.1}$$

It was shown that there exists an optimal representation

$$\Psi = \{\psi_1, \psi_2, \ldots, \psi_M\} \tag{4.2}$$

in the sense that the following average error is minimal.

$$\min_{\{\psi^1,\psi^2,\ldots,\psi^M\}} \frac{1}{n} \sum_{i=1}^{n} \left\| \phi_i - \sum_{j=1}^{M} \alpha_j \psi_j \right\|^2 \tag{4.3}$$

where$\|\cdot\|^2$ represents the usual L_2 norm.

The minimal average value is obtained if the basis elements satisfy the eigenfunctions problem

$$\int K(x,y)\,\psi_i(y)\,dy = \lambda_i \psi_i(x) \tag{4.4}$$

and

$$\langle \psi_i, \psi_j \rangle = \int \psi_i(x)\,\psi_j(x)\,dx = \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \tag{4.5}$$

where

$$K(x,y) = \frac{1}{n} \sum_{i=1}^{n} \psi_i(x)\,\psi_i(y) \tag{4.6}$$

This is the essence of the Karhunen-Loeve Expansion procedure, Proper Orthogonal Decomposition (POD) and Principal Component Analysis (PCA).

4.2 Essence of POD

For a complex temporal-spatial flow $y(x,t)$, an ensemble of snapshots is chosen in the analysis time interval $[0,T]$ written as

$$\left\{y^1, y^2, \ldots, y^n\right\} \tag{4.7}$$

where $y^i(x) = y(x,t_i)$, $i = 1,\ldots,n$, n is the number of snapshots.

Define the ensemble average of the snapshots as

$$\bar{y}(x) = \frac{1}{n}\sum_{i=1}^{n} y^i \tag{4.8}$$

Subtracting the mean from each snapshot, we obtain

$$\mathbf{Y} = \left[y^1 - \bar{y}, y^2 - \bar{y}, \ldots, y^n - \bar{y}\right] \tag{4.9}$$

We expand $y(t,x)$as

$$y^{POD}(x,t) = \bar{y}(x) + \sum_{i=1}^{M} \alpha_i(t)\,\psi_i(x) \tag{4.10}$$

Where the POD basis vector $\psi_i(x)$ and M are judiciously chosen to capture the dynamics of the flow as follows. First, define the spatial correlation matrix $\mathbf{K}^{n\times n}$ with entries as follows:

$$K_{ij} = \int_\Omega \left(y^i - \bar{y}\right)^{\mathbf{T}} \left(y^i - \bar{y}\right) d\Omega,\ 1 \le i,j \le n \tag{4.11}$$

Thus, the eigenvalue problem

$$\mathbf{K}\psi_i = \lambda_i \psi_i \tag{4.12}$$

is solved to obtain the eigenvalues $\lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_n \ge 0$ of $\mathbf{K}^{n\times n}$ with its corresponding orthonormal eigenvectors $\xi_1,\ldots,\xi_n$.

Hence, the corresponding POD modes are thus obtained by defining

$$\psi_i = \frac{1}{\sqrt{\lambda_i}}\mathbf{Y}\xi_i, \quad i = 1,\ldots,M \tag{4.13}$$

and

$$\langle \psi_i, \psi_j \rangle = \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \ne j \end{cases} \tag{4.14}$$

One can define a relative information content to choose a low-dimensional basis of size $M << n$ by neglecting modes corresponding to the small eigenvalues. We define

$$I(m) = \frac{\sum_{i=1}^{i=m} \lambda_i}{\sum_{i=1}^{i=n} \lambda_i} \tag{4.15}$$

and choose M such that $M = \arg\min \{I(m) : I(m) > \gamma\}$, where $0 \leq \gamma \leq 1$ is the percentage of total information retained in the reduced space and the tolerance γ must be chosen to be close unity in order to capture most of the energy of the snapshots basis.

4.3 Method of snapshots

An ensemble of snapshots is chosen in the analysis time interval $[0, T]$ written as $\{y^1, y^2, \ldots, y^n\}$ where $y^i \in \mathbb{R}^N$, $i = 1, \ldots, n$, n is the number of snapshots and N is the dimension of the discrete mesh.

Define ensemble average of the snapshots as

$$\bar{y}(x) = \frac{1}{n} \sum_{i=1}^{n} y^i \tag{4.16}$$

Subtracting the mean from each snapshot, we obtain the following$N \times n$ dimensional matrix

$$\mathbf{Y} = \left[y^1 - \bar{y}, y^2 - \bar{y}, \ldots, y^n - \bar{y}\right] \tag{4.17}$$

The POD modes

$$\Psi = \{\psi_1, \psi_2, \ldots, \psi_M\} \tag{4.18}$$

of order $M \leq n$ provide an optimal representation of the ensemble data in a $M-$dimensional state subspace by minimizing the averaged projection error

$$\min_{\{\psi^1, \psi^2, \ldots, \psi^M\}} \frac{1}{n} \sum_{i=1}^{n} \left\| \left(y^i - \bar{y}\right) - \Pi_{\Psi,M} \left(y^i - \bar{y}\right) \right\|^2$$
$$\text{s.t. } \left\langle \psi^i, \psi^j \right\rangle = \delta_{ij} \tag{4.19}$$

where $\Pi_{\Psi,M}$ is the projection operator onto the $M-$dimensional space

$$span \left\{\psi^1, \psi^2, \ldots, \psi^M\right\} \tag{4.20}$$

and

$$\Pi_{\Psi,M} = \sum_{i=1}^{M} \langle y, \psi_i \rangle \psi_i \tag{4.21}$$

Define the spatial correlation matrix

$$\mathbf{A} = \mathbf{Y}\mathbf{Y}^{\mathbf{T}} \tag{4.22}$$

To compute the POD modes $\psi_i \in \mathbb{R}^N$, one must solve an $N-$dimensional eigenvalue problem

$$\mathbf{A}\psi_i = \lambda_i \psi_i \tag{4.23}$$

Since in practice the number of snapshots is much less than the the state dimension, $n << N$, an efficient way to compute the reduced basis is to introduce a $n-$dimensional matrix as follows:

$$\mathbf{K}^{n\times n} = \mathbf{Y}^{\mathbf{T}}\mathbf{Y} \tag{4.24}$$

and compute the eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0$ of $\mathbf{K}^{n\times n}$ with its corresponding eigenvectors $\xi_1, \ldots, \xi_n$

Hence, the corresponding POD modes are thus obtained by defining

$$\psi_i = \frac{1}{\sqrt{\lambda_i}}\mathbf{Y}\xi_i, \quad i = 1, \ldots, M \tag{4.25}$$

where

$$\langle \psi_i, \psi_j \rangle = \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \tag{4.26}$$

One can define a relative information content to choose a low-dimensional basis of size $M << n$ by neglecting modes corresponding to the small eigenvalues. We define

$$I(m) = \frac{\sum_{i=1}^{i=m} \lambda_i}{\sum_{i=1}^{i=n} \lambda_i} \tag{4.27}$$

and choose M such that $M = \arg\min \{I(m) : I(m) > \gamma\}$, where $0 \leq \gamma \leq 1$ is the percentage of total information retained in the reduced space and the tolerance γ must be chosen to be close unity in order to capture most of the energy of the snapshots basis.

4.4 POD Galerkin Projection and Error estimation

For an atmospheric or oceanic temporal-spatial flow $y(x,t)$ defined in time interval $[0,T]$

$$\frac{dy}{dt} = F(y,t)$$

$$y(x,0) = y_0(x) \tag{4.28}$$

To obtain a reduced model, we can first solve the PDE to obtain an ensemble of snapshots, then use a Galerkin projection scheme of the model equations onto the space spanned by the POD basis elements. We obtain the system of ODE as follows:

$$\frac{d\alpha_i}{dt} = \left\langle F\left(\bar{y} + \sum_{i=1}^{M} \alpha_i \psi_i, t\right), \psi_i \right\rangle \tag{4.29}$$

along with the initial conditions:

$$\alpha_i(0) = \langle y(x,0) - \bar{y}, \psi_i \rangle = \langle y_0 - \bar{y}, \psi_i \rangle, \quad i = 1, \cdots, M \tag{4.30}$$

the error between POD solution obtained by Galerkin projection scheme and the true solution can be written as

$$\int_0^T \left\| y(t) - y^{POD}(t) \right\|_{L_2} dt \tag{4.31}$$

The difference between the true solution $y(t)$ and the continuous FEM solution $y^{POD}(t)$ obtained by Galerkin projection scheme can be decomposed into

$$y(t) - y^{POD}(t) = (y(t) - \Pi_M y(t)) + \left(\Pi_M y(t) - y^{POD}(t)\right) = \varrho(t) + \upsilon(t) \tag{4.32}$$

where $\varrho(t) \in \mathbf{\Psi}^{\perp}$ and $\upsilon(t) \in \Psi$. To estimate $\varrho(t)$, we have

$$\int_0^T \|\varrho(t)\|_{L_2}^2 dt = \int_0^T \|(y(t) - \Pi_M y(t))\|_{L_2}^2 dt = \sum_{i=M+1}^{i=\infty} \lambda_i \tag{4.33}$$

To estimate $\upsilon(t)$, we consider

$$\begin{aligned}
\dot{\upsilon}(t) &= \Pi_M \dot{y}(t) - \dot{y}^{POD,h}(t) = \left(\dot{y}(t) - \dot{y}^{POD,h}(t)\right) + \left(\Pi_M \dot{y}(t) - \dot{y}(t)\right) \\
&= \left(F(\dot{y}, t) - F\left(\dot{y}^{POD,h}, t\right)\right) + \left(\Pi_M \dot{y}(t) - \dot{y}(t)\right) \\
&= \left(\mathbf{F}(y, t) - \mathbf{F}\left(y^{POD}, t\right)\right) + o\left(\left\| y - y^{POD} \right\|_{L_2}\right) + \left(\Pi_M \dot{y}(t) - \dot{y}(t)\right) \\
&= \mathbf{F}\left(y - y^{POD}\right) + o\left(\left\| y - y^{POD} \right\|_{L_2}\right) + \left(\Pi_M \dot{y}(t) - \dot{y}(t)\right) \\
&= \mathbf{F}(\varrho(t) + \upsilon(t)) + o\left(\|\varrho(t) + \upsilon(t)\|_{L_2}\right) + \left(\Pi_M \dot{y}(t) - \dot{y}(t)\right)
\end{aligned} \tag{4.34}$$

where $\mathbf{F}$ is the linearization of the nonlinear model F.

Hence, we obtain that

$$\upsilon(t)^{\mathbf{T}} \dot{\upsilon}(t) = \upsilon(t)^{\mathbf{T}} \left(\mathbf{F}(\varrho(t) + \upsilon(t)) + o\left(\|\varrho(t) + \upsilon(t)\|_{L_2}\right) + \left(\Pi_M \dot{y}(t) - \dot{y}(t)\right)\right) \tag{4.35}$$

By the Lax-Milgram lemma and Young's inequality, we obtain that

$$\frac{d}{dt}\left\|\upsilon\left(t\right)\right\|_{L_2}^2 \leq C\left(\left\|\varrho\left(t\right)\right\|_{L_2}^2 + \left\|\upsilon\left(t\right)\right\|_{L_2}^2 + o\left(\left\|\varrho\left(t\right)\right\|_{L_2}^2 + \left\|\upsilon\left(t\right)\right\|_{L_2}^2\right) + \left\|\Pi_M\dot{y}\left(t\right) - \dot{y}\left(t\right)\right\|_{L_2}^2\right) \tag{4.36}$$

for any $t \in [0, T]$.

Hence we obtain

$$\frac{d}{dt}\left\|\upsilon\left(t\right)\right\|_{L_2}^2 \leq C\left(\left\|\varrho\left(t\right)\right\|_{L_2}^2 + \left\|\upsilon\left(t\right)\right\|_{L_2}^2 + \left\|\Pi_M\dot{y}\left(t\right) - \dot{y}\left(t\right)\right\|_{L_2}^2\right) \tag{4.37}$$

Integrating the ODE above using t he initial condition$\upsilon\left(0\right) = 0$ and apply the Gronwall lemma, we obtain

$$\begin{aligned}\left\|\upsilon\left(t\right)\right\|_{L_2}^2 &\leq C\left(\left\|\varrho\left(t\right)\right\|_{L_2}^2 + \left\|\Pi_M\dot{y}\left(t\right) - \dot{y}\left(t\right)\right\|_{L_2}^2\right) \\ &= C\left(\sum_{i=M+1}^{i=\infty}\lambda_i + \left\|\Pi_M\dot{y}\left(t\right) - \dot{y}\left(t\right)\right\|_{L_2}^2\right)\end{aligned} \tag{4.38}$$

Finally, we obtain

$$\int_0^T \left\|y\left(t\right) - y^{POD}\left(t\right)\right\|_{L_2} dt \leq C\left(\sum_{i=M+1}^{i=\infty}\lambda_i + \left\|\Pi_M\dot{y}\left(t\right) - \dot{y}\left(t\right)\right\|_{L_2}^2\right) \tag{4.39}$$

Hence, the error between POD solution obtained by a Galerkin projection scheme and the true solution is bounded by the decay of the eigenvalues of POD and the snapshots approximation quality for $\dot{y}\left(t\right)$.

4.5 Links between POD and balanced truncation

We consider linear state-space systems

$$\begin{aligned}\dot{\mathbf{x}} &= \mathbf{A}\mathbf{x} + \mathbf{u} \\ \mathbf{y} &= \mathbf{x}\end{aligned} \tag{4.40}$$

with state $\mathbf{x}(t) \in \mathbb{R}^n$, input $\mathbf{u}(t) \in \mathbb{R}^m$, and output $\mathbf{y}(t) \in \mathbb{R}^p$. Let $\phi\left(\mathbf{u}, \mathbf{x}, t\right)$ denote the solution of the state equations . In particular, for the continuous-time state equations the solution can be written as

$$\phi(\mathbf{u}, \mathbf{x}, t) = e^{\mathbf{A}t}\mathbf{x}_0 \tag{4.41}$$

$$\mathbf{y}(t) = \phi(\mathbf{u}, \mathbf{x}, t) = e^{\mathbf{A}t}\mathbf{x}_0,\ t \geq 0 \tag{4.42}$$

Construct an ensemble of solutions Y

$$Y = \{y_1(t), y_2(t), \ldots, y_m(t)\} \tag{4.43}$$

by using different unit impulses

$$x_0^i = e_i, \quad i = 1, \cdots m \tag{4.44}$$

and

$$u = 0 \tag{4.45}$$

Hence

$$y^i(t) = e^{\mathbf{A}t}\mathbf{x}_0 = e^{\mathbf{A}t}e_i, \quad i = 1, \cdots m \tag{4.46}$$

Thus, we obtain that

$$\begin{aligned}
\mathbf{P} &= \int_0^\infty \mathbf{h_r}(t)\,\mathbf{h_r^*}(t)\,dt \\
&= \int_0^\infty e^{\mathbf{A}t}e^{\mathbf{A}^*t}dt \\
&= \int_0^\infty e^{\mathbf{A}t}\,[e_1, \cdots, e_m]\begin{bmatrix} e_1 \\ \vdots \\ e_m \end{bmatrix} e^{\mathbf{A}^*t}dt \\
&= \int_0^\infty [y_1, \cdots, y_m]\begin{bmatrix} y_1^* \\ \vdots \\ y_m^* \end{bmatrix} dt \\
&= \int_0^\infty [y_1, \cdots, y_m]\begin{bmatrix} y_1^* \\ \vdots \\ y_m^* \end{bmatrix} dt \\
&= \int_0^\infty y_1y_1^* + y_2y_2^* + \cdots + y_my_m^* dt
\end{aligned} \tag{4.47}$$

Note that the similarity of the expression above to the method of snapshots using POD basis vectors. Indeed, if data from simulations is used to find the impulse responses, then it is usually given at discrete times $t_1, \cdots, t_m$, and the integral above becomes a quadrature sum. We may construct the ensemble of snapshots by

$$Y = \{y(t_1), y(t_2), \ldots, y(t_m)\} \tag{4.48}$$

Thus, we obtain *the controllability Grammian*

$$\mathbf{P} = \mathbf{Y}\mathbf{Y}^* \tag{4.49}$$

Let's consider the adjoint of linear state-space systems as follows:

$$\dot{\mathbf{z}} = \mathbf{A}^*\mathbf{z} + \mathbf{v} \tag{4.50}$$

If data from simulations is used to find the impulse responses of the adjoint system above, we may construct the ensemble of so-called adjoint snapshots by

$$\mathbf{Z} = \{z(t_1), z(t_2), \ldots, z(t_p)\} \tag{4.51}$$

Similarly, we obtain the *observability Grammian* derived as

$$\begin{aligned} \mathbf{Q} &= \int_0^\infty \mathbf{h_o}(t)\, \mathbf{h_o^*}(t)\, dt \\ &= \mathbf{Z}\mathbf{Z}^* \end{aligned} \tag{4.52}$$

One of the difficulties with the POD/Galerkin method is that the inner product used for the computing the POD modes and projecting the dynamics is arbitrarily chosen. Sometimes, an appropriate inner product is obvious, if POD is constructed in the framework of finite element space. Sometimes, a suitable inner product is not obvious, and different choices can give totally different results. From the discussion above, it is clear that the deepest connection between the POD methodologies and balanced truncated model reduction methods is that balanced truncation may be viewed as a special case of POD, using impulse responses from simulations and using the observability Grammian as an inner product defined as follows.

$$\langle a, b\rangle = a^* \mathbf{Q} b \tag{4.53}$$

Therefore, the POD modes with respect to this inner product are eigenvectors of

$$\mathbf{R} = \mathbf{PQ} \tag{4.54}$$

This eigenvectors are Hankel eigenvectors which will be orthogonal to the above mentioned inner product, instead of respect to the standard inner product.

4.6 Variants of POD

4.6.1 Centroidal Voronoi Tessellation

Given an ensemble of snapshots

$$Y = \left\{y^1, y^2, \ldots, y^n\right\} \tag{4.55}$$

where $y^i \in \mathbb{R}^N$, $i = 1, \ldots, n$, n is the number of snapshots and N is the dimension of discrete mesh, the following set

$$\left\{T^1, T^2, \ldots, T^n\right\}$$

is a tessellation of Y if for $i = 1, \cdots, r$

$$\begin{aligned} T^i &\subset Y \\ T^i &\cap T^j = \phi \ , i \neq j \\ \cup_{i=1}^r T^i &= Y \end{aligned} \tag{4.56}$$

Given an ensemble of vectors

$$Z = \left\{z^1, z^2, \ldots, z^r\right\} \tag{4.57}$$

where $z^i \in \mathbb{R}^N$ (but not necessarily to Y), $i = 1, \ldots, r$, the Voronoi region corresponding to the vector z^i is defined by

$$V^i = \left\{y \in Y : \ \left|y - z^i\right| \leq \left|y - z^j\right| \ \ for \ j = 1, \cdots, r \ j \neq i\right\} \tag{4.58}$$

The ensemble of vectors

$$V = \left\{V^1, V^2, \ldots, V^r\right\} \tag{4.59}$$

is called a Voronoi tessellation or Voronoi diagram of Y corresponding to Z and the vectors in Z are called the generators of the Voronoi diagram.

Given a density function $\rho(u) \geq 0$, defined for $u \in Y$, the mass centroid z^* of any subset $T \subset Y$ is defined by

$$\sum_{u \in T} \rho(u) |u - z^*|^2 = \inf_{z \in R^N} \sum_{u \in T} \rho(u) |u - z|^2$$

In general, $z_i^* \neq z^i$, $i = 1, \ldots, r$, i.e., the centers of mass of the Voronoi regions are not the same as the generators of those regions. However, if $z_i^* = z^i$, $i = 1, \ldots, r$ we refer to the Voronoi tessellation as being a Centroidal Voronoi tessellation or CVT for short.

CVT's are optimal in the following sense.

Given an ensemble of snapshots $Y = \{y^1, y^2, \ldots, y^n\}$ and an ensemble of vectors $Z = \{z^1, z^2, \ldots, z^r\}$, where $y^i, z^i \in \mathbb{R}^N$, $i = 1, \ldots, n$, n is the number of snapshots and N is the dimension of the discrete mesh, we define the error of a tessellation $V = \{V^1, V^2, \ldots, V^r\}$ of Y with respect to Z by

$$\mathcal{F}\left(\left\{z^i, V^i\right\}_{i=1}^r\right) = \sum_{i=1}^{r} \sum_{v \in V^i} \rho(v) \left|v - z^i\right|^2$$

then, it can be shown that a necessary condition for such kind of error measure $\mathcal{F}$ to be minimized is that the pair $\{z^i, V^i\}_{i=1}^r$ forms a CVT of Y.

CVT's of discrete sets are closely related to optimal k-means clusters so that Voronoi regions and centroids can be referred to as clusters and cluster centers, respectively. The error $\mathcal{F}$ also often referred to as the variance, cost distortion error, or mean square error. CVT's have been successfully used in data compression with one particular application to image reconstruction.

Fortunately, one does not have to choose between POD and CVT, but can combine the two methods in several different ways to define a hybrid CVT based POD method CVOD, such that CVOD offers the possibility of taking advantage of the best features of both POD and CVT. Also, CVOD is computationally less expensive than POD since it requires the solution of several smaller eigenvalue problems instead of one large one.

4.7 Limitations of POD ROM

SVD based methods, in particular POD methods, are used to investigate relationships between fields because POD modes capture most of the observed covariance with the fewest pairs of patterns (see DelSole [32]). Majda and Wang [31] showed how measuring predictability for NWP with a Gaussian distribution can be simplified through the special use of a linear change of coordinates based on empirical orthogonal functions (EOF) basis.

However, SVD based methods are not ideal in all cases. For instance in predictability studies, one would like to determine patterns that are the "most predictable" according to the measure of forecast skill. This is because that truncated POD modes represent only a tiny amount of variance that can be crucial in the generation of certain types of dynamics (see Majda [30]). In particular, systems that exhibit sudden transitions between different states (i.e., bursting behavior) will be susceptible to these type of problems when trying to model them using POD modes. Low-dimensional truncation of the POD basis inhibits transfers of energy between the large and small scales (unresolved) of the fluid flow. Therefore, the price of the low-dimensionality entails a lack of stability especially for transitional and turbulent flows characterized by high Reynolds numbers (Couplet et al., 2005 [39]; Noack et al., 2010 [27]; Galletti et al., 2004 [37]; Gloerfelt, 2006 [40]). This either restricts reduced order models to a narrow range of parameters or to a short-time integration span. To improve the accuracy of POD-Galerkin models, the effect of these unresolved modes, which are taken from the small scales of the fluid flow, for instance by including eddy viscosity terms, must be included to provide an insight into the turbulent energy.

A methods named calibration has been proposed in order to improve the accuracy of POD reduced-order modeling due to solutions of optimization problems. The idea of calibration is to use information from the temporal dynamics of the POD model known in advance to correct whole or part of the coefficients from the POD Galerkin projection. Various calibration methods have been developed to enhance the stability of POD-Galerkin models (Couplet et al., 2005 [39]; Gloerfelt, 2006 [40]; Galletti et al., 2005 [41]; Pastoor et al., 2008 [42]). These calibration terms are computed by minimizing a cost functional defined as either: the difference between the amplitude coefficients predicted by the calibrated POD and those from the POD; or a weak constraint functional, where the constraints are calibrated POD equations and are enforced by introducing Lagrange multipliers or adjoint variables.

A modification of POD ROM consists of calibration in flow problems and adding a shift-mode so that it includes an accurate representation of the unstable steady solution [26]. Noack et al 2010 pointed out that the challenge of computing high Reynolds number turbulent flows rests on the fact that the discretization of all dynamically relevant scales from large to Kolmogorov scales is not possible. This impossibility leads to necessity of turbulence models for modeling the effect of unresolved scales on the resolved flow. He proposed a systematic strategy to eliminate dynamic degrees of freedoms in Galerkin systems of incompressible fluid flows. The proposed system reduction strategy was derived from a Finite-Time Thermodynamics closure [27].

Also, the POD ROM method heavily relies on the input of DS and the time instances at which the snapshots are taken. Consequently, singular values and modes obtained by POD ROM are not invariants of DS. Aubry [28] studied POD ROM for Kuramoto-Sivashinsky equation and found that a model based on the leading six POD modes could not reproduce the right dynamics, even though those six POD modes represent 99.9995% of the variance. Similar problems with models based on POD modes were reported by Armbruster et al. (1992) [29] in a study of Kolmogorov flow in a regime of bursting behavior. Majda [30] studied Charney-DeVore model [34] and compared POD methods, optimal persistence patterns(OPPs) (introduced by DelSole [32])and principal interaction patterns(PIPs) (introduced by Hasselmann [33]) It is shown that the PIPs and OPP based ROM methods are superior to the POD based ROM methods.

4.7.1 Optimal Persistence Patterns

A technique named optimal persistence patterns (OPPs) is described for determining the set of patterns in a time-varying field whose corresponding time series remain correlated for the longest times. The basic idea is to obtain patterns that, when projected on a time-varying field, produce time series that optimize a measure of decorrelation time. The decorrelation time is measured by one of the integrals

$$T_1 = \int_0^\infty \rho_\tau d\tau \tag{4.60}$$

or

$$T_2 = 2\int_0^\infty \rho_\tau^2 d\tau \tag{4.61}$$

in which ρ_τ is the correlation function depending on time lag τ. The idea is, given a data set $g(t)$ in some phase space P, to find a vector $e_1 \in P$, which can be solved by eigenvector methods, such that the time series $v(t) = e_1^{\mathbf{T}}(t)\, g(t)$ attain maximum for T_1 or T_2, then a second vector e_2,orthogonal in some sense to e_1, that again maximizes T_1 or T_2 and so on. The ordering of the patterns based on their persistence or correlation time makes the OPPs an interesting type of optimal basis. If one aims to reproduce the long time-scale behavior of a system, a set of patterns with maximal correlation times is a natural candidate for the basis of a reduced model.

4.7.2 Principal Interaction Patterns

The Principal Interaction Patterns (PIPs) was introduced by Hasselmann [33] and improved by Kwasniok [35]. Consider a high-fidelity dynamic model represented by a system of coupled ODEs

$$\dot{x} = F(x) \tag{4.62}$$

with state vector functions $x(t) \in \mathbb{R}^n$. If we integrate the system above from $t = 0$ to $t = \tau$, we start from x^0 and we end up with x^τ.

Let x_p be the projections of x onto a number PIPs basis vectors based on projection P , which yields a reduced system of coupled ODEs

$$\dot{x_p} = F(x_p) \tag{4.63}$$

with state vector functions $x_p(t) \in \mathbb{R}^r$. Similarly, if we integrate the reduced system above from $t = 0$ to $t = \tau$, we start from x_p^0 and we end up with x_p^τ.

Hence, the difference at $t = \tau$ between the PIPs ROM and that of the original model is computed by

$$d^\tau = x_p^\tau - x^\tau \tag{4.64}$$

We can integrate the norm of the difference d^τ as follows (see Kwasniok 1996 [35]):

$$Q = \int_0^{\tau_{max}} \|d^\tau\|_{L_2}^2 \, d\tau \tag{4.65}$$

Finally, assuming we know the integration time τ_{max}, we compute the ensemble average of Q over all initial states x^0 on the attractor. Thus, we obtain an error function dependent on the projection P:

$$E(P) = \overline{Q(x^0, P)}$$

Therefore, the PIPs basis vectors can be obtained by minimizing the error functions $E(P)$ with respect to the projection P. It should be noted the PIPS ROM is very sensitive to the choice of time τ_{max} (see Kwasniok 2004 [36]).

4.7.3 Calibrated POD for flows with high Reynolds numbers

High Reynolds number ocean flows exhibit dynamics on a wide range of scales. They display a combination of organized or coherent structures associated with the phase-averaged/spatially phase-correlated components that exhibit the most evident structure and apparently disorganized or incoherent structures associated with the random components.

The energy transfer/interaction between the different coherent/inherent structure flows plays an important role in high Reynolds number flows. Low-order truncation of the POD basis, however, inhibits transfers between the large and small (unresolved) scales of the fluid flow. As a consequence there is a lack of dissipation in POD/ROM and the reduced order model may diverge. Therefore, at higher Reynolds numbers where more kinetic energy is constrained within the smaller scales, i.e., more POD snapshots as well as more bases should be retained for a realistic representation (Galletti et al., 2004 [37]). To improve the accuracy of POD-Galerkin models, the effect of these unresolved modes must be included to provide an insight into the turbulent energy.

It is shown by Ma et al., 2002 [38] that a low-dimensional Galerkin model used to simulate three dimensional high Reynolds number system didn't capture accurately both the limit circle and the transition of three-dimensionality. The flow they studied exhibits a small divergence eventually rendering the system unstable. If more modes are included for the flow they studied, the onset of divergence is delayed but the same picture emerges.

It is not surprising that the POD reduced model derived using the Galerkin approach is not sufficiently accurate in reproducing the dynamics of higher Reynolds number flows since the truncation applied in the POD subspace inhibits transfers between the different scales of the fluid flow. The neglected POD modes correspond to small scale structures and introduce dissipative errors in the model. As a consequence, the system may lose its long-term stability. The stabilization of a reduced order model can be achieved by introducing an artificial dissipation by using a Sobolev H_1 inner product norm.

4.7.4 Balanced POD

The POD/Galerkin method can yield unpredictable results, and is sensitive to details such as the empirical data used (Rathinam and Petzold, 2003 [43]), and the choice of inner product (Colonius and Freund, 2002 [44]). Balanced truncation was developed in the control theory community for stable, linear, input-output systems, and does not suffer the same limitations as the POD method. Most notably, balanced truncation has error bounds that are close to the lowest error possible from any reduced-order model.

However it becomes intractable as the number of variables exceeds 10,000, i.e. impractical for many discrete CFD systems. Balanced proper orthogonal decomposition is a concept introduced with the aim of combining POD and balanced truncation. The goal is to compute balanced truncations, or approximations to these, with computational cost similar to POD. Several previous methods have combined ideas from POD and balanced truncation, including the original work of Moore ,1981 [2]. The method presented here relies heavily on the work of Lall et al. 2002 [45], who used empirical Grammians to generalize balanced truncation to nonlinear systems.

It was shown by Ilak [47] that some important features from the control designers point of view for the 3-D system, such as impulse response, frequency response, capturing of actuation and performance at off-design Reynolds number, were captured very well by the balanced POD reduced-order models. It was also found by Ilak [47] that, while the leading POD modes capture very well the energy of the perturbation, the corresponding models do not capture the dynamics well: in particular, POD models fail to reproduce the energy growth of the perturbation. It was shown that the energy growth of the perturbation is captured only if modes with very low energy content are included in the POD models, while Balanced POD models that include only the leading balancing modes performed very well.

CHAPTER 5

POD 4-D VARIATIONAL DATA ASSIMILATION

5.1 4-D variational data assimilation problem

The model can be written as:

$$\frac{\partial \mathbf{X}(t)}{\partial t} = \mathcal{F}(\mathbf{X}(t)) \tag{5.1}$$

where $\mathcal{F}$ is the model operator and the discretized form of the numerical model can be written as:

$$\mathbf{X}(t_r) = M_{0\rightarrow r}\mathbf{X_0} \tag{5.2}$$

where initial condition $\mathbf{X_0}$ is the control variable for the given numerical model, $M_{0\rightarrow r}$ is the predefined discretized nonlinear model forecast operator, mapping the initial condition $\mathbf{X_0}$ into the model solution $\mathbf{X_r}$ at time t_r

In its general form, the *4D-Var data assimilation* (4-D Var), is defined as the minimization with respect to the initial condition $\mathbf{X_0}$ of the following discrete cost functional:

$$J(\mathbf{X_0}) = \frac{1}{2}(\mathbf{X} - \mathbf{X_b})\,\mathbf{B^{-1}}\,(\mathbf{X} - \mathbf{X_b}) + \frac{1}{2}\sum_{r=0}^{n}\left(H_r(\mathbf{X_r}) - \mathbf{Y_r}\right)^{\mathbf{T}}\mathbf{O_r^{-1}}\left(H_r(\mathbf{X_r}) - \mathbf{Y_r}\right) \tag{5.3}$$

subject to the model as a strong constraint, (i.e. assuming the model is perfect) so that the sequence of model states $\mathbf{X_r}$ at time t_r must be a solution for the given model equations, where $\mathbf{B}$ is the background covariance matrix, $\mathbf{X_r}$ is the model solution at time t_r, $\mathbf{O_r}$ is the observation error covariance matrix at time t_r, H_r is the observation operator at time t_r, representing projection of model variables into the observational variables. Since the $M_{0\rightarrow r}(\mathbf{X_0})$ is a nonlinear operator, the *4D-Var data assimilation method* becomes a

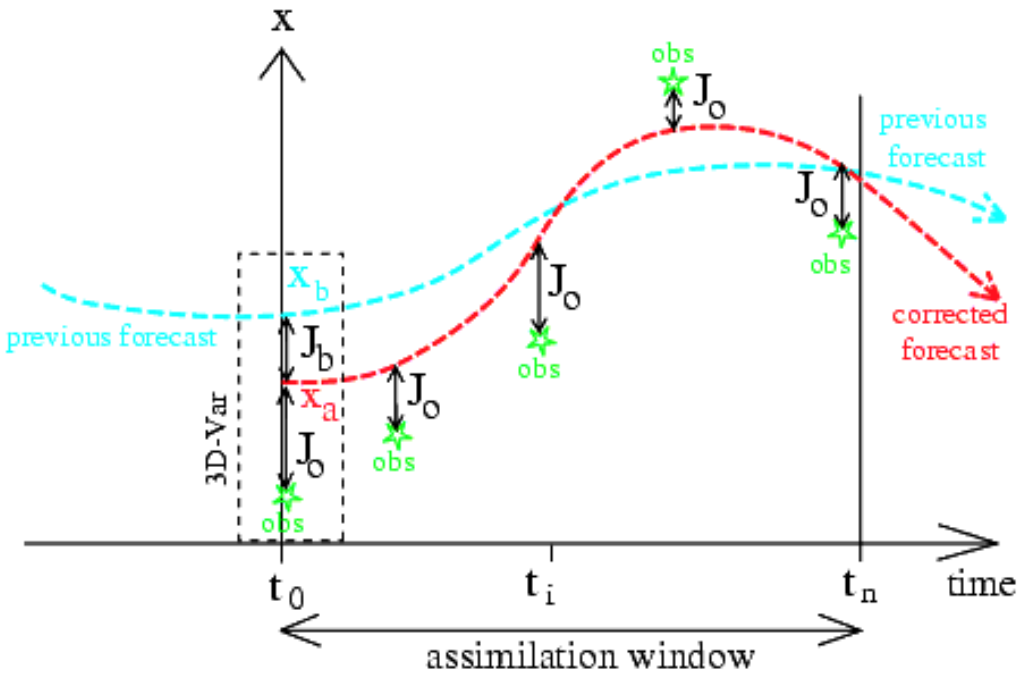


Figure 5.1: 4 D-Var in a numerical forecasting system

nonlinear constrained optimization problem, with respected to the control variable $\mathbf{X_0}$ and it is very difficult to solve. Fortunately, it can be greatly simplified with two hypotheses.

The first hypothesis is the causality, in which the forecast model can be expressed as the product of intermediate forecast steps, so that the nonlinear model forecast operator $M_{0\to r}$ can be factorized into $M_{0\to r} = M_r M_{r-1} \dots M_1$, where each operator M_r denotes the discretized nonlinear forecast operator step from time $r-1$ to r and we have $\mathbf{X_r} = M_r \mathbf{X_{r-1}}$. Hence, by recurrence, we have

$$\mathbf{X_r} = M_r M_{r-1} \dots M_1 \mathbf{X_0} \tag{5.4}$$

Another hypothesis is that, at each time step from both from $r-1$ to r, we obtain that the linearization of observation operator H_r can be written as $\mathbf{H_r}$, and that forecast operator M_r can also be linearized so that the predefined discretized nonlinear model forecast operator can be differentiated(perturbed) to obtain a so-called tangent linear model(TLM) :

$$\mathbf{X'(t_r)} = \mathbf{M}_r \mathbf{X_0'} \tag{5.5}$$

where $\mathbf{M}_r$ represents the linearization of the discretized nonlinear model forecast operator.

Under those hypotheses, the quadratic cost functional above can be written as a summation as:

$$J = J^b + J^o = J^b + \sum_{r=0}^{n} \left(J^o\right)_r \tag{5.6}$$

where J^b and J^o are the background and observation terms respectively.

In order to obtain the optimal initial conditions of shallow water equations model that minimizes J above, the gradient of J needs to be calculated with respect to the control variable $\mathbf{X_0}$ as:

$$\nabla J = \nabla J^b + \nabla J^o \tag{5.7}$$

where the first term ∇J^b can be easily obtained as:

$$\nabla J^b = \mathbf{B^{-1}} \left(\mathbf{X} - \mathbf{X_b}\right) \tag{5.8}$$

and the second term ∇J^o requires the adjoint model integration which shall be briefly derived as follows:

On the one hand, consider the change in the cost functional J resulting from a small perturbation $\mathbf{X_0^{'}}$ in the initial condition, which can be written as:

$$\left(J^o\left(\mathbf{X}_0\right)\right)^{'} = J^0\left(\mathbf{X}_0 + \mathbf{X_0^{'}}\right) - J^0(\mathbf{X}_0) = \sum_{r=0}^{n} \mathbf{H_r^T}\left(\mathbf{O_r^{-1}}\left(\mathbf{H_r}(\mathbf{X_r}) - \mathbf{y_r}\right)\right)^T \mathbf{X_r^{'}} \tag{5.9}$$

On the other hand, to first order we can write the Taylor expansion of J as:

$$\left(J^o\left(\mathbf{X}_0\right)\right)^{'} = \left(\nabla J^o\left(\mathbf{X}_0\right)\right)^T \mathbf{X_0^{'}} + o\left(\|\mathbf{X_0}\|_2\right) \tag{5.10}$$

Furthermore, we can find the gradient of the cost function by using the adjoint of the Tangent Linear Model of the given nonlinear time-dependent forward model and we obtain

$$\nabla J\left(\mathbf{X}_0\right) = \mathbf{B^{-1}}\left(\mathbf{X} - \mathbf{X_b}\right) + \sum_{r=0}^{n} \mathbf{M_r^T H_r^T O_r^{-1}}\left(\mathbf{H_r}(\mathbf{X_r}) - \mathbf{Y_r}\right) \tag{5.11}$$

where $\mathbf{M_r^T}$ represents the adjoint of model at the r^{th} time step while the weighted differences

$$\mathbf{H_r^T O_r^{-1}} \left(\mathbf{H_r}(\mathbf{X_r}) - \mathbf{Y_r}\right) \tag{5.12}$$

are forcing terms which are added to the r.h.s of the adjoint model whenever an observational time is reached.

5.2 Dual-weighted POD basis

An ensemble of snapshots is chosen in the analysis time interval $[0, T]$ written as

$$\left\{y^1, y^2, \ldots, y^n\right\}$$

where $y^i \in \mathbb{R}^N$, $i = 1, \ldots, n$, n is the number of snapshots and N is the dimension of discrete mesh.

Define the weighted ensemble average of the finite-element represented data as

$$\bar{y} = \sum_{i=1}^{i=n} w_i y^i \tag{5.13}$$

where the snapshots weights w_i are such that $0 < w_i < 1$ and $\sum_{i=1}^{n} w_i = 1$, and they are used to assign a degree of importance to each member of the ensemble. Time weighting is usually considered, and in the standard approach $w_i = \frac{1}{n}$.

introduce a general form of inner product

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{A}} = \mathbf{x^T A y} \tag{5.14}$$

where $\mathbf{A}$ is a symmetric positive definite matrix of the dimention N. For the standard Euclidean norm, $\mathbf{A}$ is just the identity matrix.

The POD basis of order $M \leq n$ provides an optimal representation of the ensemble data in M- dimensional state subspace by minimizing the averaged projection error

$$\min_{\{\psi^1, \psi^2, \ldots, \psi^M\}} \sum_{i=1}^{n} w_i \left\| \left(y^i - \bar{y}\right) - \Pi_{\Psi,M} \left(y^i - \bar{y}\right) \right\|_{\mathbf{A}}^2$$

$$\text{s.t. } \left\langle \psi^i, \psi^j \right\rangle_{\mathbf{A}} = \delta_{ij} \tag{5.15}$$

where $\Pi_{\Psi,M}$ is the projection operator onto the $M-$dimensional space

$$span\left\{\psi^1, \psi^2, \ldots, \psi^M\right\} \tag{5.16}$$

$$\Pi_{\Psi,M} = \sum_{i=1}^{M} \langle y, \psi_i \rangle_{\mathbf{A}} \psi_i$$

Build the weighted spatial correlation matrix

$$\mathbf{C} = \mathbf{YWY^T} \tag{5.17}$$

The POD modes $\psi^i \in \mathbb{R}^N$ are eigenvectors to the $N-$dimensional eigenvalue problem

$$\mathbf{CA}\psi_i = \lambda_i \psi_i$$

Since in practice the number of snapshots is much less than the state dimension, $n << N$, an efficient way to compute the reduced basis is to introduce a $n-$dimensional matrix as follows,

$$\mathbf{K}^{n\times n} = \mathbf{W}^{\frac{1}{2}}\mathbf{K}\mathbf{W}^{\frac{1}{2}} = \mathbf{W}^{\frac{1}{2}}\mathbf{Y^T}\mathbf{A}\mathbf{Y}\mathbf{W}^{\frac{1}{2}} \tag{5.18}$$

and compute the eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0$ of $\mathbf{K}^{n\times n}$ with its corresponding eigenvectors $\xi_1, \ldots, \xi_n$.

The POD basis vectors are obtained by defining

$$\psi_i = \frac{1}{\sqrt{\lambda_i}} \mathbf{Y}\mathbf{W}^{\frac{1}{2}}\xi_i, \quad i = 1, \ldots, M \tag{5.19}$$

where

$$\langle \psi_i, \psi_j \rangle_{\mathbf{A}} = \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \tag{5.20}$$

One can define a relative information content to choose a low-dimensional basis of size $M << n$ by neglecting modes corresponding to the small eigenvalues. We define

$$I(m) = \frac{\sum_{i=1}^{i=m} \lambda_i}{\sum_{i=1}^{i=n} \lambda_i} \tag{5.21}$$

and choose M such that

$$M = \arg\min \{ I(m) : \ I(m) > \gamma \} \tag{5.22}$$

where $0 \leq \gamma \leq 1$ is the percentage of total information by the reduced space and the tolerance γ must be chosen to be close to the value one in order to capture most of the energy of the snapshots basis.

The aim of 4-D VAR data assimilation is that of fusing observational data and model predictions to obtain an optimal representation of the state of the atmosphere. In the full

nonlinear 4-D Var [109], this process is implemented by minimizing the cost functional as follows.

$$J\left(y_0\right)=\frac{1}{2}\left(y_0-y^b\right)^{\mathbf{T}}\mathbf{B}^{-1}\left(y_0-y^b\right)+\frac{1}{2}\sum_{k=0}^{k=n}\left(H_k y_k-y_k^o\right)^{\mathbf{T}}\mathbf{R}_k^{-1}\left(H_k y_k-y_k^o\right) \tag{5.23}$$

where y^b is the background prior state estimation and $\mathbf{B}$ is the background error covariance covariance matrix, $\mathbf{R}$ is the observational error covariance matrix, H is the observation operator, y_0 is a vector containing control variables such as initial conditions, y_k is a vector containing the solution of variables from the model at the time level k, y_k^o is the observation at time level k, and n is the number of time levels.

The snapshots are essentially a set of instantaneous model solutions, obtained from experimental data or from a simulation. They are then used to compute the POD basis vectors to yield an optimal representation of the data so that for any given basis vector size, the L_2 norm of the error between the original and reconstructed snapshot is minimized.

The construction of POD basis vectors depends not only on the features of model dynamics itself, but it also requires to properly account for the features from the 4-D VAR data assimilation. Furthermore, these two features may be quite different from each other. A recent method to avoid this problem is referred to as optimality system proper orthogonal decomposition [56]. By implementing a dual-weighted proper orthogonal decomposition (DWPOD) method [102], we can incorporate the information from the 4-D VAR into the POD reduced-order modeling.

The specification of dual weights w_k associated with the snapshots may have a significant impact on which modes are selected as dominant and thus included into the POD basis. The dual-weighted approach makes use of the time-varying sensitivities of the 4-D Var cost functional with respect to perturbations in the state at each time level where the snapshots are taken.

Assuming the cost functional $J\left(y\left(t\right)\right)$ is defined explicitly in terms of each state $y\left(t\right)$ at time step t, then for any fixed time step $\tau<t$, the model can be written as

$$\forall\tau<t,\ y\left(t\right)=M_{\tau\rightarrow t}\left(y\left(\tau\right)\right)=M_{\tau,t}\left(y\left(\tau\right)\right) \tag{5.24}$$

such that implicitly, the cost functional J can be viewed as a function of the previous state $y\left(\tau\right)$, to first-order approximation, the impact of small errors/perturbations δy_i in the

state error at a snapshot time $t_i \leq t$ on J may be estimated using the tangent linear model $\mathbf{M}(t_i, t)$ and its adjoint model $\mathbf{M^T}(t, t_i)$:

$$\delta J \approx \left\langle \nabla J_{y(t)}\left(y\left(t\right)\right), \delta y\left(t\right)\right\rangle = \left\langle \nabla J_{y(t)}\left(y\left(t\right)\right) \mathbf{M}(t_i, t)\delta y\left(t_i\right)\right\rangle$$

$$= \left\langle \mathbf{M^T}(t, t_i)\nabla J_{y(t)}\left(y\left(t\right)\right), \delta y\left(t_i\right)\right\rangle = \left\langle y^*_{t_i}, \delta y\left(t_i\right)\right\rangle \tag{5.25}$$

where $y^*_{t_i} = \mathbf{M^T}(t, t_i)\nabla J_{y(t)}\left(y\left(t\right)\right)$ are the adjoint variables at time step t_i.

In particular, the model can be written as

$$\forall k,\ y_k = M_{k-1\rightarrow k}\left(y_{k-1}\right) = M_k\left(y_{k-1}\right) \tag{5.26}$$

where $M_{k-1\rightarrow k}$ is defined as the model forecast operator from time $k-1$ to k.

In order to derive the algorithm for the computation of dual weighs by using the adjoint model, we explicitly choose $\tau = t_i = k-1$ and $t = k$, to the first-order approximation, the impact of perturbations δy_{k-1} in state vectors on cost functional J_k may be estimated using tangent linear model $\mathbf{M}_k$ and its adjoint model $\mathbf{M_k^T}$:

$$\delta J_k \approx \langle \nabla J_k, \delta y_k\rangle = \langle \nabla J_k, \mathbf{M}_k\delta y_{k-1}\rangle = \left\langle \mathbf{M_k^T}\nabla J_k, \delta y_{k-1}\right\rangle = \left\langle y^*_{k-1}, \delta y_{k-1}\right\rangle \tag{5.27}$$

where $y^*_{k-1} = \mathbf{M_k^T}\nabla J_k$ are the adjoint variables at time step t_{k-1}.

Hence, it follows (see Equ (23)) that

$$|\delta J_k| \approx \left\langle y^*_{k-1}, \delta y_{k-1}\right\rangle = \left|\left\langle (\mathbf{A})^{-\mathbf{1}} y^*_{k-1}, \delta y_{k-1}\right\rangle_{\mathbf{A}}\right|$$

$$\leq \left\|(\mathbf{A})^{-\mathbf{1}} y^*_{k-1}\right\|_{\mathbf{A}} \left\|\delta y_{k-1}\right\|_{\mathbf{A}} \tag{5.28}$$

where $\mathbf{A}$ is a symmetric positive definite matrix of dimension N. For the standard Euclidean norm, $\mathbf{A}$ is just the identity matrix.

Hence, the dual weights w_k associated with the snapshots selection are defined as normalized values in the following:

$$c_k = \left\|(\mathbf{A})^{-1} y^*_k\right\|_{\mathbf{A}}$$

$$w_k = c_k / \sum_{j=1}^{j=n} c_j, \quad k = 1, \cdots, n \tag{5.29}$$

and provide a measure of the relative impact of the perturbations of state variables on the cost functional. A large value of weight w_k indicates that state errors at time step t_k plays an important role in the optimization. In other words, the dual weights are chosen in order that information from the data assimilation system is incorporated directly into the optimality criteria that determines the POD basis functions. Hence, the dual-weighted POD incorporates not only information from the dynamical system, but also information from the data assimilation system. The traditional POD basis aims at capturing the most energetic modes of the dynamical system, while the dual-weighted approach may also capture lower energy modes that can be significant for the successful implementation of 4-D Var.

From an implementation point of view, the evaluation of all dual weights requires only one adjoint model integration.

1. Initialize the adjoint variables y^* at final time to zero: $y_n^* = 0$
2. For each step $k-1$ the adjoint variables y_{k-1}^* are obtained by $y_{k-1}^* = \mathbf{M}_k^{\mathbf{T}} y_k^* + \mathbf{H}_k^{\mathbf{T}} \mathbf{R}_k^{-1} \left(\mathbf{H}_k y_k - y_k^o \right)$
3. We obtain $y_0^* = y_0^* + \mathbf{B}^{-1} \left(y_0 - y^b \right)$ where y^b is the background prior state estimation.
4. Compute $c_k = \left\| (\mathbf{A})^{-1} y_k^* \right\|_{\mathbf{A}}$ and $w_k = c_k / \sum_{j=1}^{j=n} c_j, \quad k = 1, \cdots, n$

where $\mathbf{M}_k$ is the tangent linear model and $\mathbf{H}_k$ is the linearized observation operator at time step k.

Hence, the evaluation of the dual weights only requires the integration of the adjoint model backward in time. Since the adjoint model is available during the implementation of 4-D VAR data assimilation, no additional cost is required for the development of DWPOD 4-D VAR over the classic POD 4-D VAR.

5.3 Reduced-order POD 4-D Var

In order to reduce the computational cost of 4-D Var data assimilation (Vermeulen and Heemink 2006 [67]), we consider minimization of the cost functional in a space whose

dimension is much smaller than that of the original one. A way to drastically decrease the dimension of the control space without significantly compromising the quality of the final solution but sizably decreasing the cost in memory and CPU time of 4-D Var motivates us to choose to project the control variable on a basis of characteristic vectors capturing most of the energy and the main directions of variability of the model, i.e. SVD. One would then attempt to control the vector of initial conditions in the reduced space model.

The reduced-order cost functional can be expressed as

$$J^{POD}\left(y_0^{POD}\right) = \frac{1}{2}\left(y_0^{POD} - y^b\right)^{\mathbf{T}} \mathbf{B}^{-1}\left(y_0^{POD} - y^b\right)$$

$$+ \frac{1}{2}\sum_{k=0}^{k=n}\left(H_k y_k^{POD} - y_k^o\right)^{\mathbf{T}} \mathbf{R}_k^{-1}\left(H_k y_k^{POD} - y_k^o\right) \tag{5.30}$$

where $\mathbf{B}$ is the background error covariance matrix, $\mathbf{R}_k$ is the observation error covariance matrix at time level k, H_k is the observation operator at time level k. y^b is the background prior state estimation. y_0^{POD} is a vector containing the control variables (here, initial conditions) represented by the POD basis. y_k^{POD} is a vector containing the solution of variables obtained from the reduced-order model at the time level k.

In a POD reduced-order model, the initial value y_0^{POD} and the reduced-order model solution y_k^{POD} can be expressed as

$$\begin{array}{l} y_0^{POD} = \bar{y} + \sum_{i=1}^{i=M} \alpha_i\left(0\right)\psi_i = \bar{y} + \Psi\alpha_0 \\ y_k^{POD} = \bar{y} + \sum_{i=1}^{i=M} \alpha_i\left(t^k\right)\psi_i = \bar{y} + \Psi\alpha_k \end{array} \tag{5.31}$$

where an ensemble of POD basis is

$$\Psi = \left\{\psi^1, \psi^2, \ldots, \psi^M\right\} \tag{5.32}$$

Hence, we can rewrite the reduced-order cost functional $J^{POD}\left(y_0^{POD}\right)$ dependent on y_0^{POD} as an explicit cost functional $J_\alpha^{POD}\left(\alpha_0\right)$ dependent on α_0 that is the coefficient in the POD basis vectors Ψ. Once we find the minimizer of $\alpha_0^{min} = \min_{\alpha_0} J_\alpha^{POD}\left(\alpha_0\right)$, we can express the retrieved initial condition $y_0^{POD} = \bar{y} + \Psi\alpha_0$ in the POD reduced-order model cost functional

$$J_\alpha^{POD}\left(\alpha_0\right) = \frac{1}{2}\left(\bar{y} + \Psi\alpha_0 - y^b\right)^{\mathbf{T}} \mathbf{B}^{-1}\left(\bar{y} + \Psi\alpha_0 - y^b\right)$$

$$+\frac{1}{2}\sum_{k=0}^{k=n}\left(H_k\left(\bar{y}+\Psi\alpha_k\right)-y_k^o\right)^{\mathbf{T}}\mathbf{R}_k^{-1}\left(H_k\left(\bar{y}+\Psi\alpha_k\right)-y_k^o\right) \tag{5.33}$$

The reduced model can be written as:

$$\forall k,\ \alpha_k = M_{0\to k}^{POD}\left(\alpha_0\right) \tag{5.34}$$

By denoting

$$\forall k,\ \alpha_k = M_{k-1\to k}^{POD}\left(\alpha_{k-1}\right) = M_k^{POD}\left(\alpha_{k-1}\right) \tag{5.35}$$

-

and by recurrence we obtain that

$$\alpha_k = M_k^{POD}\cdots M_1^{POD}\alpha_0 \tag{5.36}$$

The reduced-order cost functional $J_\alpha^{POD}\left(\alpha_0\right)$ that is dependent on α_0 can be divided into two components:

$$J_\alpha^{POD} = J_\alpha^{POD,b} + J_\alpha^{POD,o} \tag{5.37}$$

where the background cost functional that is dependent on α_0 is written as

$$J_\alpha^{POD,b} = \frac{1}{2}\left(\bar{y}+\Psi\alpha_0-y^b\right)^{\mathbf{T}}\mathbf{B}^{-1}\left(\bar{y}+\Psi\alpha_0-y^b\right) \tag{5.38}$$

and the observational cost functional that is dependent on α_0 is written as

$$J_\alpha^{POD,o} = \frac{1}{2}\sum_{k=0}^{k=n}\left(H_k\left(\bar{y}+\Psi\alpha_k\right)-y_k^o\right)^{\mathbf{T}}\mathbf{R}_k^{-1}\left(H_k\left(\bar{y}+\Psi\alpha_k\right)-y_k^o\right) \tag{5.39}$$

Denoting "normalized departures "

$$d_k = \mathbf{R}_k^{-1}\left(H_k\left(\bar{y}+\Psi\alpha_k\right)-y_k^o\right) \tag{5.40}$$

the contributions to the observational cost functional that is dependent on α_0 can be written as

$$J_{\alpha,k}^{POD,o} = \left(H_k\left(\bar{y}+\Psi\alpha_k\right)-y_k^o\right)^{\mathbf{T}} d_k \tag{5.41}$$

Hence the reduced-order cost functional that is dependent on α_0 can be rewritten as

$$J_\alpha^{POD} = J_\alpha^{POD,b} + \sum_{k=0}^{n} J_{\alpha,k}^{POD,o} \tag{5.42}$$

Therefore, the gradient of the reduced-order cost functional with respect to the α_0 can be derived as

$$\nabla_{\alpha_0} J_\alpha^{POD} = \nabla_{\alpha_0} J_\alpha^{POD,b} + \sum_{k=0}^{n} \nabla_{\alpha_0} J_{\alpha,k}^{POD,o} \tag{5.43}$$

$$\nabla_{\alpha_0} J_\alpha^{POD} = \mathbf{\Psi^T B}^{-1} \left(\bar{y} + \Psi\alpha_0 - y^b\right) + \sum_{k=0}^{n} \left(\mathbf{M_1^{POD}}\right)^{\mathbf{T}} \ldots \left(\mathbf{M_k^{POD}}\right)^{\mathbf{T}} \mathbf{\Psi^T H_k^T} d_k \tag{5.44}$$

where $\left(\mathbf{M_k^{POD}}\right)^{\mathbf{T}}$ is the POD reduced-order adjoint model at time step k.

From an implementation point of view (see Vermeulen and Heemink [67] and Kunisch [54, 55]), we can compute the gradient $\nabla_{\alpha_0} J_\alpha^{POD}$ in the following steps.

1. Initialize the *reduced-order adjoint variables* α^* at final time to zero: $\alpha_n^* = 0$

2. For each step $k-1$ the adjoint variables α_{k-1}^* is obtained by adding the *reduced-order adjoint forcing term* $\mathbf{\Psi^T H_k^T} d_k$ to α_k^* and by performing the *reduced-order adjoint integration* of reduced-order model by multiplying the result by $\left(\mathbf{M_k^{POD}}\right)^{\mathbf{T}}$, i.e. $\alpha_{k-1}^* = \left(\mathbf{M_k^{POD}}\right)^{\mathbf{T}} \left(\alpha_k^* + \mathbf{\Psi^T H_k^T} d_k\right)$

3. At the end of recurrence, the value of the adjoint variable $\alpha_0^* = J_{\alpha_0}^o$ yields the gradient of the observational cost functional

4. Compute $\nabla_{\alpha_0} J_\alpha^{POD,b} = \mathbf{\Psi^T B}^{-1} \left(\bar{y} + \Psi\alpha_0 - y^b\right)$ and we obtain $\nabla_{\alpha_0} J_\alpha^{POD} = \nabla_{\alpha_0} J_\alpha^{POD,b} + \nabla_{\alpha_0} J_\alpha^{POD,o}$

5.4 Trust-Region based optimal control approach

5.4.1 Classical trust-region method

Historically the trust region method goes back to [72, 73, 74]. See also work of [75] followed by important work of [76, 77]. Finally the terminology of trust region and Cauchy point was put forward by [78] and systematized by [79]

The classical trust-region method [80] aims to define a region around the current iterate within which it trusts the model to be an adequate representation of the objective function f, and then choose the step to be the approximate minimizer of the model in the trust region,

i.e/ choosing direction and length of the step simultaneously. The algorithm approximates only a certain region (the so-called trust region) of the objective function with a model function (often a quadratic). It is assumed that the first two terms of the quadratic model function m_k at each iterate x_k are identical the first two terms of the Taylor-series expansion of f around x_k in the following:

$$m_k^{quad}(u_k + s) = f_k + \nabla f_k^{\mathbf{T}} s + \frac{1}{2} s^{\mathbf{T}} \mathbf{Q}_k s \tag{5.45}$$

where $f_k = f(u_k)$ and $\nabla f_k = \nabla f(u_k)$ and $\mathbf{Q}_k$ is an approximation to the Hessian and more generally $\mathbf{Q}_k$ is some symmetric matrix.

To obtain each step, we seek a solution of the following sub-problem for which we only need an approximate solution to obtain convergence and good practical behavior [70]

$$\min m_k^{quad}(u_k + s) = f_k + \nabla f_k^{\mathbf{T}} s + \frac{1}{2} s^{\mathbf{T}} \mathbf{Q}_k s \tag{5.46}$$

$$\text{subject to } \|s\| \leq \delta_k \tag{5.47}$$

where $\delta_k > 0$ is the trust-region radius.

In the strategy for choosing the trust-region radius δ_k at each iteration, we define the actual reduction

$$ared_k(s_k) = f(u_k) - f(u_k + s_k) \tag{5.48}$$

and

$$pred_k(s_k) = m_k^{quad}(u_k) - m_k^{quad}(u_k + s) \tag{5.49}$$

Thus, we can define the ratio

$$\rho_k = \frac{ared_k(s_k)}{pred_k(s_k)} \tag{5.50}$$

We measure agreement between quadratic model function m_k^{quad} and the objective function $f(u_k)$ as a criterion for choosing trust-region radius $\delta_k > 0$. If the ratio ρ_k is negative, the new objective value is greater than the current value so that the step must be rejected. On the other hand, if ρ_k is close to 1, there is good agreement between the approximate quadratic model m_k^{quad} and the object function f_k over this step, so it is safe to expand the trust region radius for the next iteration. If ρ_k is positive but not close to 1, we do not alter the trust region radius, but if it is close to zero or negative, we shrink the trust region radius.

Outline of basic trust-region algorithm:

Let $0 < \eta_1 < \eta_2 < 1$, $0 < \gamma_1 < \gamma_2 < 1 \leq \gamma_3$, δ_0 be given, set $k = 0$

1. Compute the minimizer s^k of

$$\min m_k^{quad}(u_k + s)$$

$$\text{subject to } \|s\| \leq \delta_k$$

2. Compute the new $f(u_k + s_k)$and

$$\rho_k = \frac{ared_k(s_k)}{pred_k(s_k)} \tag{5.51}$$

3. Update the trust-region radius:

- If $\rho_k \geq \eta_2$: implement outer projection $u_{k+1} = u_k + s_k$ and increase trust-region radius $\delta_{k+1} = \gamma_3 \delta_k$ and GOTO 1
- If $\eta_1 < \rho_k < \eta_2$: implement outer iteration $u_{k+1} = u_k + s_k$ and decrease trust-region radius $\delta_{k+1} = \gamma_2 \delta_k$ and GOTO 1
- If $\rho_k \leq \eta_1$: set $y_0^{(k+1)} = y_0^{(k)}$ and decrease trust-region radius $u_{k+1} = u_k$ and GOTO 3

The predicted decrease based on a quadratic model function can be analyzed using the concept of Cauchy decrease. For this reason, we consider the quadratic model approximation for objective function f expanded in the steepest descent direction.

$$m_k^{quad}(u_k - \lambda \nabla f_k) = f_k - \lambda \|\nabla f_k\|^2 + \frac{1}{2}\lambda^2 \nabla f_k{}^{\mathbf{T}} \mathbf{Q}_k \nabla f_k \tag{5.52}$$

for $\nabla f_k \neq 0$ and $\lambda \geq 0$.

Hence, the minimization of 5.52 within the trust-region δ_k yields the so-called Cauchy step

$$s_k^c = -\lambda_k^c \nabla f_k \tag{5.53}$$

where λ_k^c can be computed efficiently [76, 77] as long as simple convexity arguments holds.

Therefore, we define

$$u_{k+1}^c = u_k^c + s_k^c \tag{5.54}$$

We can denote model decrease related to the Cauchy step by

$$pred_k\left(s_k^c\right) = m_k^{quad}\left(u_k\right) - m_k^{quad}\left(u_k + s_k^c\right) \tag{5.55}$$

Since the s_k and s_k^c are both within the trust-region δ_k, it is reasonable to compare the predicted decrease $pred_k\left(s_k\right)$ to the model decrease $pred_k\left(s_k^c\right)$ related to the Cauchy step.

It can be showed that $pred_k\left(s_k\right)$ satisfies a fraction of Cauchy decrease condition

$$pred_k\left(s_k\right) \geqslant c_{fcd}\left(pred_k\left(s_k^c\right)\right) \tag{5.56}$$

For the Cauchy decrease, it has been shown in [76, 77] that

$$pred_k\left(s_k^c\right) \geq \frac{1}{2}\left\|\nabla f_k\right\| \min\left\{\delta_k, \frac{\left\|\nabla f_k\right\|}{\left\|\mathbf{Q}_k\right\|}\right\} \tag{5.57}$$

Therefore, we can derive the sufficient decrease condition

$$ared_k\left(s_k\right) \geq \eta_1 pred_k \geq \eta_1 c_{fcd}\left(pred_k\left(s_k^c\right)\right) \geqslant \frac{\eta_1 c_{fcd}}{2}\left\|\nabla f_k\right\| \min\left\{\delta_k, \frac{\left\|\nabla f_k\right\|}{\left\|\mathbf{Q}_k\right\|}\right\} \tag{5.58}$$

Finally, we obtain that

$$f\left(u_k + s_k\right) \leqslant f\left(u_k\right) - \frac{\eta_1 c_{fcd}}{2}\left\|\nabla f_k\right\| \min\left\{\delta_k, \frac{\left\|\nabla f_k\right\|}{\left\|\mathbf{Q}_k\right\|}\right\} \tag{5.59}$$

Accordingly, the weak and global convergence theorems can be proved based on the fraction of Cauchy decrease condition under some assumptions [70].

In practice, the quadratic model functions is provided with inexact gradient information, since the ∇f_k can only be approximated numerically, for instance, by its corresponding adjoint model. Also, it is clear that $\mathbf{Q}_k$ denotes the approximation to true Hessian matrix $\nabla^2 f_k$. This leads to trust-region method with quadratic model functions as follows.

$$m_k^{iquad}\left(u_k + s\right) = f_k + g_k^{\mathbf{T}} s + \frac{1}{2} s^{\mathbf{T}} \mathbf{Q}_k s \tag{5.60}$$

where g_k denotes an approximation to the true gradient ∇f_k and $\mathbf{Q}_k$denotes an approximation to the true Hessian matrix, i.e., $m_k^{iquad}\left(u_k + s\right)$ is a quadratic model based on inexact gradient information.

Similar to the quadratic model based on exact gradient information, the modified sufficient decrease condition for the quadratic model based on inexact gradient information can be derived as follows:

$$f\left(u_k + s_k\right) \leqslant f\left(u_k\right) - \frac{\eta_1 c_{fcd}}{2} \left\|g_k\right\| \min\left\{\delta_k, \frac{\left\|g_k\right\|}{\left\|\mathbf{Q}_k\right\|}\right\} \tag{5.61}$$

Finally, it is noted that the convergence results for g_k can be carried over from ∇f_k (see Carter [84])

In this situation, the Cauchy step is simply replaced with

$$s_k^c = -\lambda_k^c g_k \tag{5.62}$$

Having introduced the basic ideas of trust-region methods with quadratic model functions, we now turn to the more general trust-region method dealing with nonlinearity of the model function as follows:

$$\min\ m_k^{nonlin}\left(u_k + s\right)$$

$$\text{subject to}\ \ \|s\| \leq \delta_k$$

As pointed out in the previous sections, we have dealt with the trust-region method applied to quadratic model functions with exact or inexact gradient information, The convergence behavior of those trust-region methods relied on the sufficient decrease conditions 5.59 and 5.61. Furthermore, those sufficient decrease conditions are based on the fraction of the Cauchy decrease condition 5.56 and 5.57, where Cauchy step can be computed efficiently. However, since we are dealing with nonlinear model functions we are dealing with now, the Cauchy step and Cauchy decrese condition are no longer available in terms of the classical definition.

Toint [83] has proposed a so-called step determination algorithm generalizing the Cauchy step s_k^{nonlin} and Cauchy decrease condition to the case of general nonlinear model functions. In that sense, one can expect to get the desired sufficient decrease condition for the nonlinear mode functions based on the fraction of the generalized Cauchy decrease conditions.

$$pred_k\left(s_k\right) \geqslant c_{nonlin}\left(pred_k\left(s_k^{nonlin}\right)\right) \tag{5.63}$$

Furthermore, in 5.61, $\|\mathbf{Q}_k\|$ denotes the induced norm (see Chapter 1) of the Hessian of the model function. In that sense, it represents a measure of the model's curvature at current iteration. In general, Toint [83] defined a concept to compute the curvature w_k for the nonlinear model functions.

$$pred_k\left(s_k^{nonlin}\right) \geq \frac{1}{2}\left\|\nabla f_k\right\| \min\left\{\delta_k, \frac{\left\|g_k\right\|}{1+w_k}\right\} \tag{5.64}$$

Consequently, we can obtain a similar sufficient decrease condition for nonlinear model functions in general as follows:

$$f\left(u_k+s_k\right) \leqslant f\left(u_k\right) - \frac{\eta_1 c_{nonlin}}{2}\left\|g_k\right\| \min\left\{\delta_k, \frac{\left\|g_k\right\|}{1+w_k}\right\} \tag{5.65}$$

5.4.2 Trust-region POD method

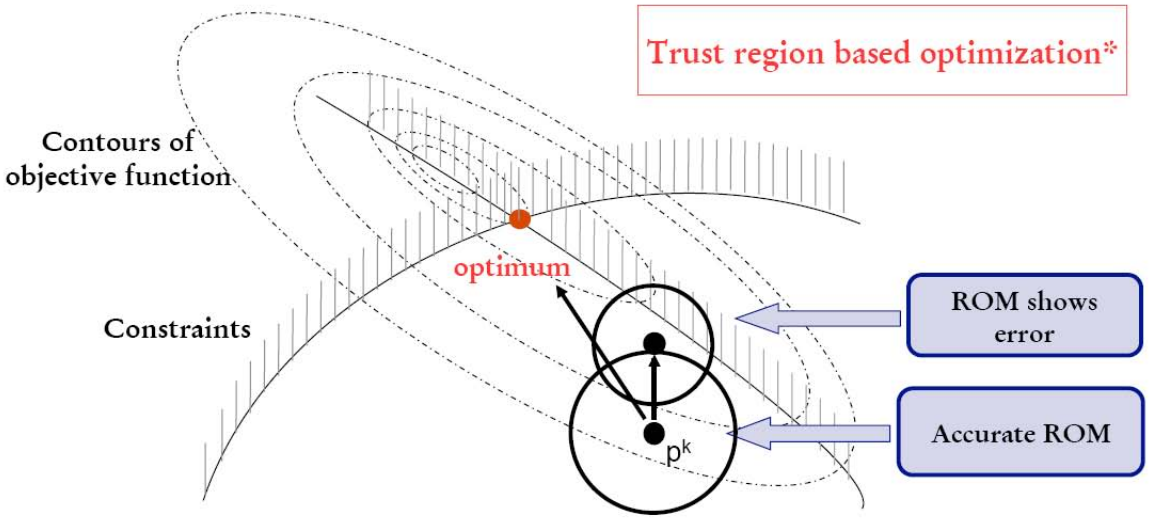


Figure 5.2: Trust-region based POD reduced-order optimization method

In the previous cases, we have introduced inportant results for the trust-region methodologies, which can be applied to the POD reduced order modeling as follows. In this work, the POD reduced order model is based on the solution of the original model for specified control variables (e.g. initial and boundary conditions). It is therefore necessary to reconstruct the POD reduced order model when the resulting control variables from the latest optimization iteration are significantly different from the ones upon which the POD model is based. Hence, it is natural to improve the POD reduced order control model successively by updating the

snapshots which are used to generate the POD basis in the process of reduced-order 4-D Var.

For the reduced-order cost functional [110, 111]

$$J^{POD}\left(y_0^{POD}\right) = \frac{1}{2}\left(y_0^{POD} - y^b\right)^{\mathbf{T}} \mathbf{B}^{-1}\left(y_0^{POD} - y^b\right) + \frac{1}{2}\sum_{k=0}^{k=n}\left(\mathbf{H}_k y_k^{POD} - y_k^o\right)^{\mathbf{T}} \mathbf{R}_k^{-1}\left(\mathbf{H}_k y_k^{POD} - y_k^o\right) \tag{5.66}$$

or its explicit version

$$J_\alpha^{POD}\left(\alpha_0\right) = \frac{1}{2}\left(\bar{y} + \Psi\alpha_0 - y^b\right)^{\mathbf{T}} \mathbf{B}^{-1}\left(\bar{y} + \Psi\alpha_0 - y^b\right) + \frac{1}{2}\sum_{k=0}^{k=n}\left(\mathbf{H}_k\left(\bar{y} + \Psi\alpha_k\right) - y_k^o\right)^{\mathbf{T}} \mathbf{R}_k^{-1}\left(\mathbf{H}_k\left(\bar{y} + \Psi\alpha_k\right) - y_k^o\right) \tag{5.67}$$

defined above, we first start with a random perturbation of given initial condition $y_0^{(0)}$ and compute the snapshots that correspond to the flow behavior forced by $y_0^{(0)}$. We then use these snapshots to compute the first POD basis $\mathbf{\Psi}^{(0)}$ and build up the corresponding POD based control model forced by applying inner projection $\alpha_0^{(0)} = \left\langle y_0^{(0)} - \bar{y}, \mathbf{\Psi}^{(0)} \right\rangle$. We now implement the inner minimization iteration based on $\mathbf{\Psi}^{(0)}$ to obtain the new control variable $\alpha_0^{(1)}$ in the reduced-order space. When we carry out an outer iteration, we obtain $y_0^{(1)} = \bar{y} + \mathbf{\Psi}^{(0)}\alpha_0^{(1)}$. If we use $y_0^{(1)}$ for the computation of new snapshots and a new POD basis $\mathbf{\Psi}^{(1)}$, we can improve the initial condition of the PDE and thus improve the POD based model. However, the outer projection computing new snapshots and corresponding new POD basis is computationally expensive and should only occur at rare instances controlled by appropriate criteria. One criterion for adaptivity consists of an ad-hoc rule that an outer projection should occur whenever the value of the objective function cannot be decreased beyond a prescribed tolerance between two consecutive inner minimization iterations . Also, this criterion will abort the outer iteration cycle when the value of the objective function is less than a given tolerance. The trust-region POD approach for adaptivity is both efficient and mathematically correct, being based on the trust-region globalization properties derived from optimization theory [83].

Therefore, to find a new step s^k, we minimize with respect to s

$$\min m_k\left(\alpha_0^{(k)}+s\right) := J_\alpha^{POD}\left(\alpha_0^{(k)}+s\right) \tag{5.68}$$

$$\text{subject to } \|s\| \leq \delta_k \tag{5.69}$$

Based on trust-region strategy from optimization [82, 111] , we can decide to increase or decrease the trust-region radius by comparing the actual(for the full order model)

$$J\left(\bar{y}+\mathbf{\Psi}^{(k-1)}\alpha_0^{(k)}\right) - J\left(\bar{y}+\mathbf{\Psi}^{(k-1)}\left(\alpha_0^{(k)}+s_k\right)\right) \tag{5.70}$$

with the predicted decrease(for the reduced-order model)

$$m_k\left(\alpha_0^{(k)}\right) - m_k\left(\alpha_0^{(k)}+s_k\right) \tag{5.71}$$

Outline of trust-region POD algorithm:

Let $0<\eta_1<\eta_2<1$, $0<\gamma_1<\gamma_2<1\leq\gamma_3$ and $y_0^{(0)}$, δ_0 be given, set $k=0$

1. Compute snapshot set $\mathcal{Y}_k^{SNAP}$ based on initial condition $y_0^{(k)}$
2. Compute the POD basis $\mathbf{\Psi}^{(k)}$ and build up the corresponding POD based control model based on the initial condition $\alpha_0^{(0)} = \left\langle y_0^{(0)}-\bar{y}, \mathbf{\Psi}^{(0)}\right\rangle$
3. Compute the minimizer s^k of

$$\min m_k\left(\alpha_0^{(k)}+s\right)$$

$$\text{subject to } \|s\| \leq \delta_k$$

4. Compute the new $J\left(\bar{y}+\mathbf{\Psi}^{(k-1)}\left(\alpha_0^{(k)}+s_k\right)\right)$ of the full model and

$$\rho_k = \frac{J\left(\bar{y}+\mathbf{\Psi}^{(k-1)}\alpha_0^{(k)}\right) - J\left(\bar{y}+\mathbf{\Psi}^{(k-1)}\left(\alpha_0^{(k)}+s_k\right)\right)}{m_k\left(\alpha_0^{(k)}\right) - m_k\left(\alpha_0^{(k)}+s_k\right)} \tag{5.72}$$

5. Update the trust-region radius:

- If $\rho_k \geq \eta_2$: implement outer projection $y_0^{(k+1)} = \bar{y}+\mathbf{\Psi}^{(k-1)}\left(\alpha_0^{(k)}+s_k\right)$ and increase trust-region radius $\delta_{k+1}=\gamma_3\delta_k$ and GOTO 1

- If $\eta_1 < \rho_k < \eta_2$: implement outer iteration $y_0^{(k+1)} = \bar{y} + \mathbf{\Psi}^{(k-1)} \left(\alpha_0^{(k)} + s_k \right)$ and decrease trust-region radius $\delta_{k+1} = \gamma_2 \delta_k$ and GOTO 1
- If $\rho_k \leq \eta_1$: set $y_0^{(k+1)} = y_0^{(k)}$ and decrease trust-region radius $\delta_{k+1} = \gamma_1 \delta_k$ and GOTO 3

In the trust-region POD optimal control algorithm above, the gradient of $m_k \left(\alpha_0^{(k)} + s \right)$ with respect to s plays an important role in the constrained minimization sub-problem

$$\min m_k \left(\alpha_0^{(k)} + s \right)$$
$$\text{subject to } \|s\| \leq \delta_k$$

On the one hand, if δ_k is large enough, the norm constraint is not active then s_k is just in the vicinity of the unconstrained minimum. On the other hand, if δ_k is small, then the higher order terms in s play a less important role than the linear term, i.e. for some constant β_k it holds $s_k \approx -\beta_k \nabla J_\alpha^{POD} \left(\alpha_0^{(k)} \right)$. As δ_k is increasing we obtain a continuous change from the direction of steepest descent to the direction of the minimum of $J_\alpha^{POD} \left(\alpha_0^{(k)} \right)$. Therefore good gradient information is required, which can be obtained by performing the reduced-order adjoint backward in time integration .

Following the trust-region philosophy, it is not necessary to determine the exact step solution of the constrained problem above. It is sufficient to compute a trial step s_k that achieves only a certain amount of decrease for the full model. We can use a backtracking approach to find the sufficient decrease. For recent work on stable Galerkin reduced order models see Barone [112].

5.4.3 Dual weighted TRPOD approach

A new methodology combining the dual weighted snapshots and trust region POD adaptivity is put forward, allowing us to enhance the benefits already derived from by using DWPOD. The combined algorithm proceeds as follows illustrated in the algorithm flowchart.

5.5 Incremental balanced truncated POD 4-D Var

The model can be written as:

$$\frac{\partial \mathbf{X}(t)}{\partial t} = \mathcal{F}(\mathbf{X}(t)) \tag{5.73}$$

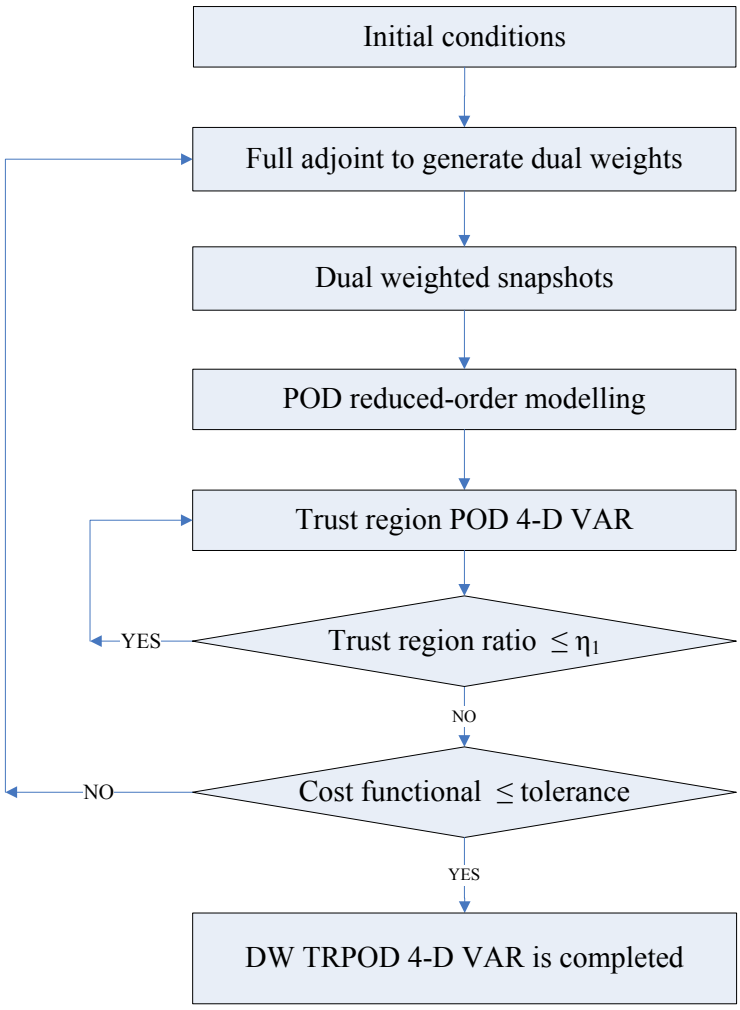


Figure 5.3: Dual weighted TRPOD approach flowchart

and the discretized form of the numerical model can be written as:

$$\mathbf{X}(t_i) = M_{0\to i}\mathbf{X_0} \tag{5.74}$$

where initial condition $\mathbf{X_0}$ is the control variable for the given numerical model, $M_{0\to i}$ is the predefined discretized nonlinear model forecast operator, mapping the initial condition $\mathbf{X_0}$ into the model solution $\mathbf{X_i}$ at time t_i

In its general form, the *4D-Var data assimilation* (4-D Var), is defined as the minimization with respect to the initial condition $\mathbf{X_0}$ of the following discrete cost functional:

$$J(\mathbf{X_0}) = \frac{1}{2}(\mathbf{X}-\mathbf{X_b})\,\mathbf{B^{-1}}\,(\mathbf{X}-\mathbf{X_b}) + \frac{1}{2}\sum_{i=0}^{n}(H_i(\mathbf{X_i})-\mathbf{Y_i})^{\mathbf{T}}\,\mathbf{O_i^{-1}}\,(H_i(\mathbf{X_i})-\mathbf{Y_i}) \quad (5.75)$$

subject to the strong nonlinear constraint

$$\mathbf{X_i} = M_i\mathbf{X_0} \quad (5.76)$$

Assuming that the model is perfect, so that the sequence of model states $\mathbf{X_i}$ at time t_i must be a solution for the given model equations, where $\mathbf{B}$ is the background covariance matrix, $\mathbf{X_i}$ is the model solution at time t_i, $\mathbf{O_i}$ is the observation error covariance matrix at time t_i, H_i is the observation operator at time t_i, representing projection of model variables into the observational variables.

For the *incremental 4-D Var*, in the outer loop we set the initial guess $\mathbf{X_0^{(0)}}$ to be equal to the background.

Therefore, for $k = 1, \cdots, K$ we compute

$$\mathbf{X_i^{(k)}} = M_i\mathbf{X_0^{(k)}} \quad (5.77)$$

In the inner loop, we solve the linear minimization problem

$$\begin{aligned} J(\delta\mathbf{X_0^{(k)}}) &= \frac{1}{2}\left(\delta\mathbf{X_0^{(k)}}-\delta\mathbf{X_b^{(k)}}\right)\mathbf{B^{-1}}\left(\delta\mathbf{X_0^{(k)}}-\delta\mathbf{X_b^{(k)}}\right) \\ &+ \frac{1}{2}\sum_{i=0}^{n}\left(\mathbf{H_i}(\delta\mathbf{X_i^{(k)}})-\mathbf{d_i^o}\right)^{\mathbf{T}}\mathbf{O_i^{-1}}\left(\mathbf{H_i}(\delta\mathbf{X_i^{(k)}})-\mathbf{d_i^o}\right) \end{aligned} \quad (5.78)$$

where

$$\mathbf{d_i^o} = \mathbf{H}_i(\mathbf{X_i^{(k)}}) - \mathbf{Y}_i$$

subject to the strong linear constraint

$$\delta\mathbf{X_{i+1}^{(k)}} = \mathbf{M_i}\delta\mathbf{X_i^{(k)}} \quad (5.79)$$

In the end of each outer loop, we update

$$\mathbf{X_0^{(k+1)}} = \mathbf{X_0^{(k)}} + \delta\mathbf{X_0^{(k)}} \quad (5.80)$$

Inside of each inner iteration, let's remove the upper index k and assume

$$H_i = \mathbf{I}_{n\times n} \tag{5.81}$$

and

$$\mathbf{M}_i = \mathbf{M} \tag{5.82}$$

Hence, we obtain that

$$\begin{aligned} \delta\mathbf{X_{i+1}} &= \mathbf{M}\delta\mathbf{X_i} \\ d_i &= \delta X_{\mathbf{i}} \end{aligned} \tag{5.83}$$

where

$$\delta\mathbf{X_i} \in R^n \tag{5.84}$$

is the a perturbation about the current state variable and

$$\mathbf{M} \in R^{n\times n}$$

is the the linearization of the nonlinear model operator about the current state variable.

In order to do the balanced truncation, for $i = -1, \cdots, n$ we can setup a input-output system as follows

$$\begin{aligned} \delta\mathbf{X_{i+1}} &= \mathbf{M}\delta\mathbf{X_i} + \mathbf{u_i} \\ d_i &= \delta X_{\mathbf{i}} \end{aligned} \tag{5.85}$$

where

$$\delta\mathbf{X_i},\ u_i\ d_i \in R^n \tag{5.86}$$

Let

$$\delta\mathbf{X}_{-1} = 0 \tag{5.87}$$

and

$$u_{-1} \sim N\left(0, D_0\right),\ u_i = 0,\ i = 1, \cdots n \tag{5.88}$$

Integrate this input-output system and we construct an ensemble of tangent linear model solutions as follows.

$$\delta\mathbf{X} = \{\delta\mathbf{X}_1, \delta\mathbf{X}_2, \ldots, \delta\mathbf{X}_n\} \tag{5.89}$$

And we construct the POD modes based on the snapshots of $\delta\mathbf{X}$

$$\Psi = \{\psi_1, \psi_2, \ldots, \psi_r\} \tag{5.90}$$

where

$$r \ll n$$

Afterwards, we can integrate the adjoint of this input-ouput system

$$\delta\mathbf{X}^*_{\mathbf{i+1}} = \mathbf{M}^*\delta\mathbf{X}^*_{\mathbf{i}} \tag{5.91}$$

with initial conditions

$$\delta X^*_{\mathbf{0}} = \psi_j,\ j = 1, \cdots r$$

Now, we can construct a series of so-called adjoint snapshots as follows for $j = 1, \cdots r$ repectively

$$\delta\mathbf{X}^*_{\mathbf{j}} = \left\{\delta\mathbf{X}^*_{j\,1}, \delta\mathbf{X}^*_{j\,2}, \ldots, \delta\mathbf{X}^*_{j\,n}\right\},\ j = 1, \cdots r \tag{5.92}$$

Put all them together, we have contructed a large ensemble of adjoint snaphots

$$\delta\mathbf{X}^* = \left\{\delta\mathbf{X}^*_1, \delta\mathbf{X}^j_2, \ldots, \delta\mathbf{X}^j_r\right\},\ j = 1, \cdots r \tag{5.93}$$

It is noted that

$$\delta\mathbf{X} \in R^{n\times n} \tag{5.94}$$

and

$$\delta\mathbf{X}^* \in R^{n\times rn} \tag{5.95}$$

Hence, we can define the controllability matrix as

$$\mathbf{P} = \delta\mathbf{X}\left(\delta\mathbf{X}\right)^{\mathbf{T}} \in R^{n\times n} \tag{5.96}$$

and the observability matrix as

$$\mathbf{Q} = \delta\mathbf{X}^* \left(\delta\mathbf{X}^*\right)^{\mathbf{T}} \in R^{n\times n} \tag{5.97}$$

Hence, we define the *Hankel singular values* as

$$\sigma_i = \left(\lambda_i \left(\mathbf{PQ}\right)\right)^{\frac{1}{2}} \tag{5.98}$$

and find the eigenvectors such that

$$\mathbf{T}^{-1}\mathbf{PQT} = \Sigma^{\mathbf{2}} \tag{5.99}$$

where

$$\Sigma = \mathrm{diag}\left(\sigma_1, \cdots, \sigma_n\right) \tag{5.100}$$

Those eigenvalues of the product of the controllability and observability Grammians are invariant under coordinate transformations. Approximation by balanced truncation preserves stability, and the H_∞- norm (the maximum of the frequency response) of the error system is bounded by twice the sum of neglected Hankel singular values $2\left(\sigma_{r+1} + \ldots + \sigma_n\right)$.

Define balanced optimal projections as

$$\mathbf{U^T} = \begin{pmatrix} \mathbf{I}_r & \mathbf{0}_{r\times(n-r)} \end{pmatrix} \mathbf{T^{-1}},\ :\mathbb{R}^n \to \mathbb{R}^r$$

$$\mathbf{V} = \mathbf{T}\begin{pmatrix} \mathbf{I}_r \\ \mathbf{0}_{r\times(n-r)} \end{pmatrix},\ :\mathbb{R}^r \to \mathbb{R}^n$$

Finally, we obtain the reduced order inner loop problem by balanced optimal projections as follows:

$$\begin{aligned} J(\delta\hat{X}_{\mathbf{0}}^{(k)}) &= \frac{1}{2}\left(\delta\hat{\mathbf{X}}_{\mathbf{0}}^{(\mathbf{k})} - \mathbf{U}^{\mathbf{T}}\delta\mathbf{X}_{\mathbf{b}}^{(\mathbf{k})}\right)\left(\mathbf{U}^{\mathbf{T}}\mathbf{B_0}\mathbf{U}\right)^{-\mathbf{1}}\left(\delta\hat{\mathbf{X}}_{\mathbf{0}}^{(\mathbf{k})} - \mathbf{U}^{\mathbf{T}}\delta\mathbf{X}_{\mathbf{b}}^{(\mathbf{k})}\right) \\ &+ \frac{1}{2}\sum_{i=0}^{n}\left(\mathbf{HV}(\delta\hat{\mathbf{X}}_{\mathbf{i}}^{(\mathbf{k})}) - \hat{\mathbf{d}}_{\mathbf{i}}^{\mathbf{o}}\right)^{\mathbf{T}}\mathbf{O}_{\mathbf{i}}^{-\mathbf{1}}\left(\mathbf{HV}(\delta\hat{\mathbf{X}}_{\mathbf{i}}^{(\mathbf{k})}) - \hat{\mathbf{d}}_{\mathbf{i}}^{\mathbf{o}}\right) \end{aligned} \tag{5.101}$$

where

$$\hat{\mathbf{d}}_{\mathbf{i}}^{\mathbf{o}} = \mathbf{HV}\mathbf{d}_{\mathbf{i}}^{\mathbf{o}}$$

subject to the strong linear constraint

$$\delta\hat{\mathbf{X}}_{\mathbf{i+1}}^{(\mathbf{k})} = \mathbf{U}^{\mathbf{T}}\mathbf{M}_{\mathbf{i}}\delta\hat{\mathbf{X}}_{\mathbf{i}}^{(\mathbf{k})} \tag{5.102}$$

and set

$$\delta\mathbf{X}_{\mathbf{0}}^{(\mathbf{k})} = \mathbf{V}\delta\hat{\mathbf{X}}_{\mathbf{0}}^{(\mathbf{k})} \tag{5.103}$$

CHAPTER 6

4-D VAR OF FINITE-ELEMENT LIMITED-AREA SHALLOW-WATER EQUATIONS MODEL

The shallow-water equations are frequently used to simulate the earth's atmosphere, which can be thought of as a thin (practically zero in height), semi-incompressible fluid that is flowing over the surface of a rotating globe (the earth). The shallow-water equations are the simplest form of the equations of motion that show how the fluid flow will evolve in response to rotational and gravitational accelerations of the earth, forming waves.

This chapter explores the feasibility of carrying out a modular structured variational data assimilation (VDA) using a finite-element method of the nonlinear shallow water equations model on a limited area domain, in which we improve the methodology (Courtier and Talagrand 1987; Zhu et al. 1994) and addresses issues in the development of the adjoint of a basic finite-element model. Specific numerical difficulties in the adjoint derivation, for example, the treatment of the adjoint of the iterative process required for solving the systems of linear algebraic equations resulting from the finite-element discretizations using Crank-Nicholson time differencing scheme (see Wang et al. 1972; Douglas and Dupont 1970) are explicitly addressed. The systems of algebraic linear equations resulting from the finite-element discretizations of the shallow-water equations model were solved by a Gauss-Seidel iterative method. To save computer memory, a compact storage scheme for the banded and sparse global matrices was used (see Hinsman, 1975). We emphasize the development of the tangent linear (TLM) and the adjoint models of the finite-element shallow-water equations model and illustrate its use on various retrieval cases when the initial conditions are as control variables.

The plan of this chapter is as follows. The finite-element Galerkin method for the shallow-water equations model on an f plane, the derivation of its tangent linear model and its adjoint

are briefly described in Section 6.1. The full finite element discretizations of the model of the nonlinear shallow-water equations model is described in Section 6.2. Section 6.3 introduces the optimal control methodology including the development of the tangent linear model and its adjoint as well as formulation of the cost functional aimed at allowing the derivation of optimal initial conditions reconciling model forecast and observations in a window of data assimilation by minimizing the cost functional measuring lack of fit between model forecast and observations. Particular attention is paid to the development of adjoint of iterative Gauss-Seidel solver. Verification of the correctness of the adjoint is carried out in a detailed manner for all stages of the calculations (i.e. TLM, adjoint and gradient test). A detailed description of the entire optimal control set-up code organization is provided and illustrated.

Set-up of numerical experiments and the experimental design are detailed in Section 6.4. Basic assimilation experiments using a random perturbation of the initial conditions as observations and their results are presented. Particular attention is paid to the effectiveness of limited memory Quasi-Newton method L-BFGS for minimizing the cost functional in retrieving optimal initial conditions. Various scenarios involving mesh resolution, different time steps as well as various lengths of the assimilation windows are tested and numerical conclusions are drawn (See Zhu, Navon and Zou 1994 [118]).

6.1 Shallow-Water equations model on an f plane

The shallow-water equations model is one of the simplest forms of the equations of motion for incompressible fluid for which the depth is relatively small compared to the horizontal dimensions, which can be applied to describe the horizontal structure of an atmosphere. They describe the evolution of an incompressible fluid in response to gravitational and rotational accelerations (See Tan 1992 and Vreugdenhil 1994 Galewsky 2004).

The shallow-water equations can be written as:

$$\frac{\partial \vec{v}}{\partial t} + \overrightarrow{v} \cdot \nabla \overrightarrow{v} + \nabla \phi + f \overrightarrow{k} \times \overrightarrow{v} = 0 \tag{6.1}$$

$$\frac{\partial \phi}{\partial t} + \nabla \cdot (\phi \overrightarrow{v}) = 0 \tag{6.2}$$

$$(x, y) \in [0, L] \times [0, D], \;\; t \geqslant 0$$

where L and D are the dimensions of a rectangular domain of integration, $\overrightarrow{v}$ is a vector function:

$$\overrightarrow{v} = (u(x, y, t), v(x, y, t)) \tag{6.3}$$

where u and v are the velocity components in the x and y axis respectively, $\phi = gh$ is the geopotential height, h is the depth of the fluid and g is the acceleration of gravity. The vector $\overrightarrow{k}$ is the vertical unit vector pointing away from the center of the planet. The scalar function f is the Coriolis parameter defined by the β-plane approximation:

$$f = \hat{f} + \beta \left(y - \frac{D}{2} \right) \tag{6.4}$$

The Coriolis parameter

$$\hat{f} = 2\Omega \sin \theta \tag{6.5}$$

is defined at a mean latitude θ_0, where Ω is the angular velocity of the earth's rotation and θ is latitude.

6.1.1 Initial and boundary conditions

The shallow-water equations require specifying appropriate initial and boundary conditions.

An initial condition is imposed as:

$$w(x, y, 0) = \varphi(x, y) \tag{6.6}$$

where state variables are $w = w(x, y, t) = (\vec{v}(x, y, t), \phi(x, y, t))$

with periodic boundary conditions in the x-direction:

$$w(0, L, t) = w(0, D, t) \tag{6.7}$$

and a solid wall boundary condition in the y-direction is:

$$\vec{v}(x, 0, t) = \vec{v}(x, D, t) = 0 \tag{6.8}$$

The geopotential $\varphi(x, y)$ will be specified later in the numerical experiments.

6.1.2 Linearization of the Shallow-Water equations model

The linearization of the shallow-water equations model (1) - (2) can be written as:

$$\frac{\partial \overrightarrow{v}'}{\partial t} + \overrightarrow{v}' \cdot \nabla \overrightarrow{v} + \overrightarrow{v} \cdot \nabla \overrightarrow{v}' + \nabla \phi' + f \overrightarrow{k} \times \overrightarrow{v}' = 0 \tag{6.9}$$

$$\frac{\partial \phi'}{\partial t} + \nabla \cdot (\phi' \overrightarrow{v}) + \nabla \cdot (\phi \overrightarrow{v}') = 0 \tag{6.10}$$

where the prime denotes a perturbation around the basic state variables.

The form above can also be written explicitly (Jacques Blum, Franois-Xavier Le Dimet, I. Michael Navon 2008) as continuous tangent linear model (TLM):

$$\frac{\partial u'}{\partial t} + u'\frac{\partial u}{\partial x} + v'\frac{\partial u}{\partial y} + \frac{\partial \phi'}{\partial x} + u\frac{\partial u'}{\partial x} + v\frac{\partial u'}{\partial y} - fv' = 0$$

$$\frac{\partial v'}{\partial t} + u'\frac{\partial v}{\partial x} + v'\frac{\partial v}{\partial y} + \frac{\partial \phi'}{\partial y} + u\frac{\partial v'}{\partial x} + v\frac{\partial v'}{\partial y} + fu' = 0$$

$$\frac{\partial \phi'}{\partial t} + \frac{\partial (\phi' u)}{\partial x} + \frac{\partial (\phi' v)}{\partial y} + \frac{\partial (\phi u')}{\partial x} + \frac{\partial (\phi v')}{\partial y} = 0$$

and its first order continuous adjoint model with weighting forcing terms may be written as:

$$-\frac{\partial u^\star}{\partial t} = -\left(-u\frac{\partial u^\star}{\partial x} - \frac{\partial (vu^\star)}{\partial y} + v^\star\frac{\partial v}{\partial x} + fv^\star - \phi\frac{\partial \phi^\star}{\partial x}\right) + W_u(u - u^o)$$

$$-\frac{\partial v^\star}{\partial t} = -\left(u^\star\frac{\partial u}{\partial y} - fu^\star - v\frac{\partial v^\star}{\partial y} - \frac{\partial (uv^\star)}{\partial x} - \phi\frac{\partial \phi^\star}{\partial y}\right) + W_v(v - v^o)$$

$$-\frac{\partial \phi^\star}{\partial t} = -\left(-\frac{\partial u^\star}{\partial x} - \frac{\partial v^\star}{\partial y} - u\frac{\partial \phi^\star}{\partial x} - v\frac{\partial \phi^\star}{\partial y}\right) + W_\phi(\phi - \phi^o)$$

with final conditions equal to zeros:

$$u(T) = v(T) = \phi(T) = 0$$

By integrating the first order continuous adjoint model reversely in time, the gradient of a given cost functional J is obtained by the adjoint model solutions as follows:

$$\nabla J(w_0) = \nabla J(u_0, v_0, \phi_0) = w^\star(0) = \begin{pmatrix} u^\star(0) \\ v^\star(0) \\ \phi^\star(0) \end{pmatrix}$$

where $w^\star = (u^\star, v^\star, \phi^\star)$ is the first order adjoint variable vector, W_u, W_v, W_ϕ are weighting factors which are chosen to be the inverse of estimates of the statistical root-mean-square observational errors on geopotential and wind components respectively. In our test problem, values of $W_\phi = 10^{-4}m^{-4}s^4$ and $W_u = W_v = 10^{-2}m^{-2}s^2$ are used.

The operator form of the discretized (9) - (10) can be written as(see Navon et al. 1992)

$$w'(x, y, t) = \mathbf{P}\left(w(x, y, t)\right) w'(x, y, 0) \tag{6.11}$$

where the control variable $w'(x, y, 0)$ is the random perturbation variable of the initial state variable $w(x, y, 0)$, while $\mathbf{P}\left(w(x, y, t)\right)$ represents the tangent linear operator, so that we can obtain the control variable $w'(x, y, t)$ that contains the values of wind fields and geopotential field at the final time step.

Generally speaking, there are two approaches which could be employed for calculating the gradient of the cost functional with respect to the initial conditions of shallow water equations. The first approach is called continuous adjoint, in which we need to differentiate the nonlinear shallow water equations model with respect to its initial conditions first and then discretize its adjoint PDE to compute the approximate gradient of the given cost functional. Another approach is called discrete approach, in which we need to approximate the nonlinear PDE by a discretized nonlinear system of equations first and then differentiate the discretized nonlinear system with respect to the parameters. The discrete adjoint approach is easy to implement with the help of automatic differentiation tools, such as ADIFOR and TAMC. In the following sections, we demonstrate the methodology of discrete adjoint to carry on the VDA.

6.2 Discretization of the SWE model

6.2.1 Formulation of Galerkin Finite-Element model

We employ linear piecewise polynomials on triangular elements in the formulation of Galerkin Finite-Element model (1) - (2) for the sake of of simplicity. Over each given element, a variable ξ can be written as(see Zienkiewicz 2005)

$$\xi_{el} = \sum_{j=1}^{3} \xi_j(t) V_j(x, y)$$

where $\xi_j(t)$ represents the scalar node value of variable ξ at the node of the triangular element, and V_j represents a basis function(interpolation function) defined by the coordinates of the nodes.

The advection terms in the continuity equation (2) are usually integrated by parts using Green's theorem to shift the derivative from the variable to the basis function, which yields:

$$\left\langle \frac{d\phi}{dt}, V_i \right\rangle + \langle \nabla \cdot (\phi \overrightarrow{v}), V_i \rangle = 0 \tag{6.12}$$

$$\Rightarrow \left\langle \frac{d\phi}{dt}, V_i \right\rangle + \int \nabla \cdot (\phi V_i \overrightarrow{v}) - \langle \phi \overrightarrow{v}, \nabla V_i \rangle = 0 \tag{6.13}$$

where the notation:

$$\left\langle \overrightarrow{f}, V_i \right\rangle = \sum_{elements}^{M} \iint \overrightarrow{f}(x, y) \cdot V_i dx dy \tag{6.14}$$

defines the inner product when a function is multiplied by the trial function V_i. where $\cdot$ represents the inner product between two real vectors. In Galerkin FEM method, we choose the trial function to coincide with the test function. Taking into account the boundary conditions (see Navon 1979), the second term of equation (13) vanishes so that we obtain the final expression for the continuity equation:

$$\left\langle \frac{d\phi}{dt}, V_i \right\rangle - \langle \phi \overrightarrow{v}, \nabla V_i \rangle = 0 \tag{6.15}$$

Following the Galerkin FEM, the momentum equation (1) becomes:

$$\left\langle \frac{d\vec{v}}{dt}, V_i \right\rangle + \langle \overrightarrow{v} \cdot \nabla \overrightarrow{v}, V_i \rangle + \langle \nabla \phi, V_i \rangle + \left\langle f \overrightarrow{k} \times \overrightarrow{v}, V_i \right\rangle = 0 \tag{6.16}$$

Over each element, we denote wind fields and geopotential fields

$$\vec{v} = \sum_{j=1}^{3} \vec{v}_j(t) V_j(x, y), \quad \phi = \sum_{j=1}^{3} \phi_j(t) V_j(x, y) \tag{6.17}$$

where $\vec{v}_j(t)$ and $\phi_j(t)$ are the time-dependent nodal values of wind fields and geopotential fields respectively.

Upon substituting (17) into (15) - (16), one obtains:

$$\left\langle \frac{d\phi_j}{dt} V_j, V_i \right\rangle - \langle \phi_j \vec{v}_k V_j V_k, \nabla V_i \rangle = 0 \tag{6.18}$$

$$\left\langle \frac{d\vec{v}}{dt} V_j, V_i \right\rangle + \langle \vec{v}_k \cdot \nabla \vec{v}_k, V_i \rangle + \langle \nabla \phi_k, V_i \rangle + \left\langle f \overrightarrow{k} \times \vec{v}_k, V_i \right\rangle = 0 \tag{6.19}$$

According to the definition (14), we may write (18) explicitly as:

$$\left\langle \frac{\partial \phi_j}{\partial t} V_j, V_i \right\rangle - \left\langle \phi_j u_k V_j V_k, \frac{\partial V_i}{\partial x} \right\rangle - \left\langle \phi_j v_k V_j V_k, \frac{\partial V_i}{\partial y} \right\rangle = 0 \tag{6.20}$$

We may also write (19) explicitly as:

$$\begin{aligned} &\left\langle \begin{pmatrix} \frac{\partial u_j}{\partial t} V_j \\ \frac{\partial v_j}{\partial t} V_j \end{pmatrix}, V_i \right\rangle + \left\langle V_k(u_k, v_k) \begin{pmatrix} u_j \frac{\partial V_j}{\partial x} & u_j \frac{\partial V_j}{\partial y} \\ v \frac{\partial V_j}{\partial x} & v_j \frac{\partial V_j}{\partial y} \end{pmatrix}, V_i \right\rangle \\ &\quad + \left\langle \begin{pmatrix} \phi_k \frac{\partial V_i}{\partial x} \\ \phi_k \frac{\partial V_i}{\partial y} \end{pmatrix}, V_i \right\rangle + \left\langle \begin{pmatrix} -f v_k V_k \\ f u_k V_k \end{pmatrix}, V_i \right\rangle = 0 \end{aligned}$$

$$\Rightarrow \left\langle \frac{\partial u_j}{\partial t} V_j, V_i \right\rangle + \left\langle u_k V_k u_j \frac{\partial V_j}{\partial x}, V_i \right\rangle$$

$$+ \left\langle v_k V_k u_j \frac{\partial V_j}{\partial y}, V_i \right\rangle + \left\langle \phi_k \frac{\partial V_i}{\partial x}, V_i \right\rangle + \langle -f v_k V_k, V_i \rangle = 0 \tag{6.21}$$

and

$$\Rightarrow \left\langle \frac{\partial v_j}{\partial t} V_j, V_i \right\rangle + \left\langle u_k V_k v_j \frac{\partial V_j}{\partial x}, V_i \right\rangle$$

$$+ \left\langle v_k V_k v_j \frac{\partial V_j}{\partial y}, V_i \right\rangle + \left\langle \phi_k \frac{\partial V_i}{\partial y}, V_i \right\rangle + \langle f u_k V_k, V_i \rangle = 0 \tag{6.22}$$

6.2.2 Time integration

A time-extrapolated Crank-Nicholson time differencing scheme was applied for integrating in time the system of ordinary differential equations resulting from the application of the Galerkin FEM (see Navon 1979,1987). The shallow-water equations system were then coupled at every time step so that the equations become quasi-linearized(see Wang et al.

1972; Douglas and Dupont 1970), since an average is taken at time level $n-1$ and time level n of expressions, while the nonlinear advective terms are linearized by estimating them at time level $n+\frac{1}{2}$ using the following second-order approximation in time:

$$w^{\star} = \tfrac{3}{2}w^{n} - \tfrac{1}{2}w^{n-1} + o(\Delta t^{2}) \tag{6.23}$$

where the state variables $w = w(x, y, t) = (\vec{v}(x, y, t), \phi(x, y, t))$.

At each time step the shallow-water equations system was coupled, i.e. the solution of each equation after one iteration at a given time step was used to solve the other two equations for the same iteration for the same time step.

Upon introducing a finite difference discretization in time into the continuity equation (20), which is the first to be solved at a given time step, one obtains

$$\mathbf{M}\left(\phi_j^{n+1} - \phi_j^{n}\right) - \frac{\triangle t}{2}\mathbf{K}_1\left(\phi_j^{n+1} + \phi_j^{n}\right) = 0 \tag{6.24}$$

where

$$\mathbf{M} = \iint\limits_{ele} V_i V_j dA \tag{6.25}$$

and

$$\mathbf{K}_1 = \iint\limits_{ele} V_j V_k u_k^{\star}\frac{\partial V_i}{\partial x}dA + \iint\limits_{ele} V_j V_k v_k^{\star}\frac{\partial V_i}{\partial y}dA \tag{6.26}$$

In this continuity equation, we need to use Crank-Nicholson to extrapolate $u^{\star}$ and $v^{\star}$ at the current time step so that we can proceed to solve ϕ^{n+1} at the next time step from $(u^{\star}, v^{\star}, \phi^{n})$.

By introducing the same finite difference scheme into the u-momentum equations (21), one obtains:

$$\mathbf{M}\left(u_j^{n+1} - u_j^{n}\right) + \frac{\triangle t}{2}\mathbf{K}_2\left(u_j^{n+1} + u_j^{n}\right) + \frac{\triangle t}{2}\left(\mathbf{K}_{21}^{n+1} + \mathbf{K}_{21}^{n}\right) + \triangle t\mathbf{P_2} = 0 \tag{6.27}$$

where

$$\mathbf{K}_2 = \iint\limits_{ele} u_k^{n} V_i V_k \frac{\partial V_j}{\partial x}dA + \iint\limits_{ele} v_k^{*} V_i V_k \frac{\partial V_j}{\partial y}dA \tag{6.28}$$

$$\mathbf{K}_{21} = \iint_{ele} \phi_k^{n+1} V_i \frac{\partial V_k}{\partial x} dA \tag{6.29}$$

$$\mathbf{P}_2 = -\iint_{ele} f v_k^{\star} V_k V_i dA \tag{6.30}$$

In this u-momentum equation, since we already know the most recent solution ϕ^{n+1} from solving the continuity equation above, we only need to extrapolate $v^{\star}$ at the current time step so that we can proceed to solve u^{n+1} at the next time step from $(u^n, v^{\star}, \phi^{n+1})$.

Finally, from the v-momentum equation (22), one obtains:

$$\mathbf{M}\left(v_j^{n+1} - v_j^n\right) + \frac{\triangle t}{2}\mathbf{K}_3\left(v_j^{n+1} + v_j^n\right) + \frac{\triangle t}{2}\left(\mathbf{K}_{31}^{n+1} + \mathbf{K}_{31}^n\right) + \triangle t \mathbf{P}_3 = 0 \tag{6.31}$$

where

$$\mathbf{K}_3 = \iint_{ele} u_k^{n+1} V_i V_k \frac{\partial V_j}{\partial x} dA + \iint_{ele} v_k^n V_i V_k \frac{\partial V_j}{\partial y} dA \tag{6.32}$$

$$\mathbf{K}_{31} = \iint_{ele} \phi_k^{n+1} V_i \frac{\partial V_k}{\partial y} dA \tag{6.33}$$

$$\mathbf{P}_3 = \iint_{ele} f u_k^{n+1} V_k V_i dA \tag{6.34}$$

In this v-momentum equation, since we already know the most recent solution for both ϕ^{n+1} and u^{n+1} at the current time step , we don't need any extrapolations at the current time step and we can proceed to solve v^{n+1} at the next time step from $(u^{n+1}, v^n, \phi^{n+1})$.

6.2.3 Gauss-Seidel iterative method for the compact matrix of the Galerkin finite-element model.

In this Galerkin finite-element model, a compact matrix form was adopted due to the local support property over the triangle mesh. In particular, the $N \times N$ global matrix, assembled from each small element matrix, has at most seven nonzero elements in each row of the matrix. Hence, we can store the global matrix into a compact matrix of size $N \times 7$. (see Zhu, Navon and Zou 1994).

In order to implement boundary conditions in the Galerkin finite-element model, we have adopted the approach suggested by Payne and Irons(see Payne 1963) and mentioned by Huebner(see Huebner1975). This approach consists in modifying the diagonal terms of the global matrix associated with the nodal variables by multiplying them by a large number, say 10^{16} (chosen with a view to the significant number of digits possible with the given computer and the size of the field variables), while the corresponding term in the right-hand vector is replaced by the specified boundary nodal variable multiplied by the same large factor times the corresponding diagonal term. This procedure is repeated until all prescribed boundary nodal variables have been treated(see Navon 1979).

6.3 Optimal Control of FE-SWE Model

6.3.1 Code organization

The nonlinear Galerkin FEM Model, TLM test, transpose test(Input/Output test), Gradient Test, and L-BFGS optimization were all written by a modularized FORTRAN90 language. In the graphs as follows, we only show the modularized Galerkin FEM code as well as the modularized L-BFGS optimization code flowchart.

In nonlinear Galerkin FEM model (Figure 6.1), four different modules are written as *Mesh*, *Assemble Matrix*, *Nonlinear Forward Model*, and *solver*. For example, in *Module Mesh*, we encapsulated a large amount of information such as the mesh size, the local and global element, compact local support, the area of each element, the coordinate and derivative of each node, and special geometries of the boundary structure.

In the graph of modularized L-BFGS optimization flowchart (Figure 6.2), we encapsulated the nonlinear Galerkin FEM model as well as its corresponding adjoint model. In the calls graph of L-BFGS implementation (Figure 6.3), we briefly list the function calls and subroutine calls to each other within each of the relevant modules.

6.3.2 Techniques in coding the adjoint of FE-SWE model

- Reset some temporary variables to zeros when using them in different statements ;
- Saving and loading the state variables calculated in the forward model ;
- Identifying the reused adjoint control variables in all the subroutines;

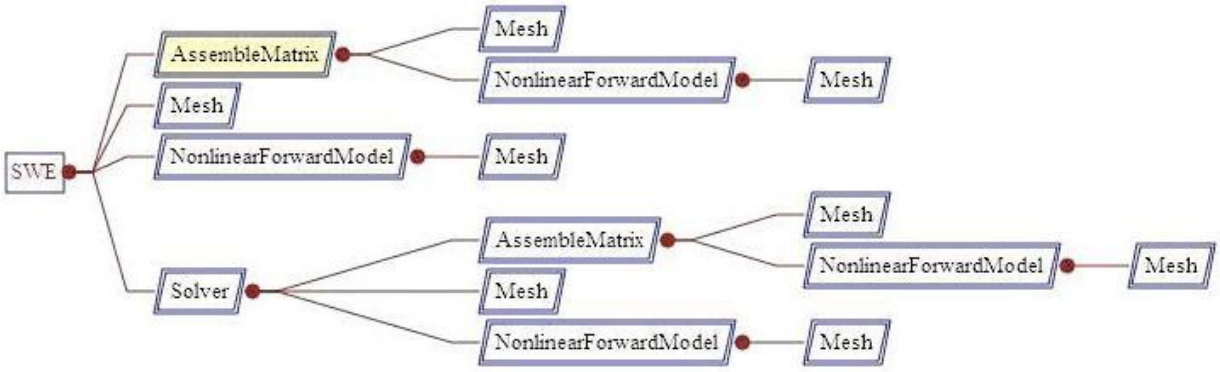


Figure 6.1: Modularized Galerkin FEM code organization

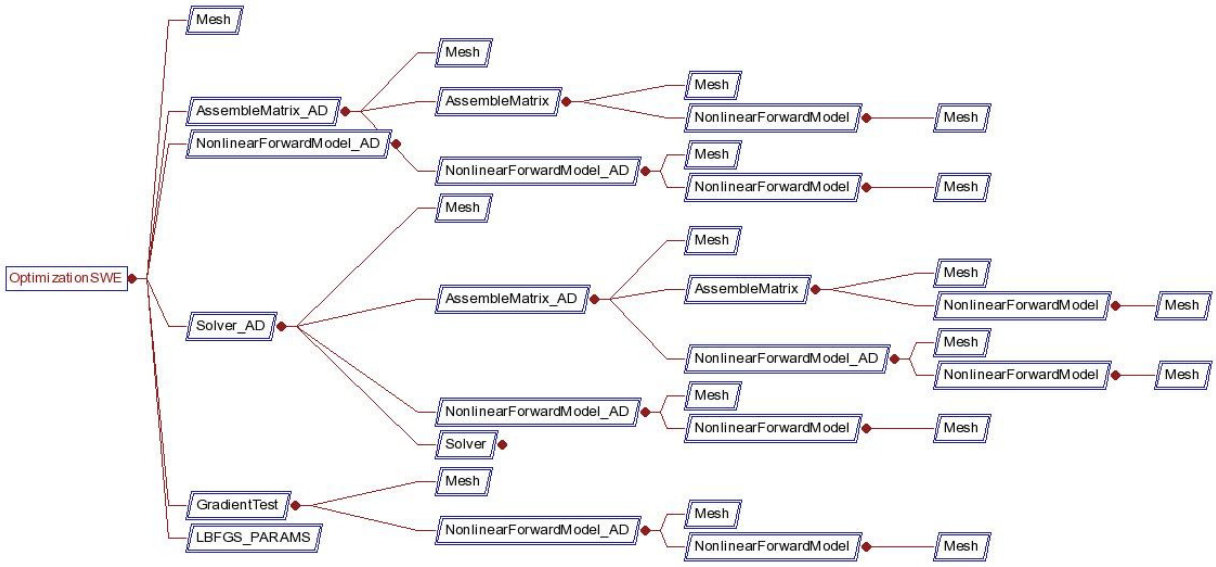


Figure 6.2: Modularized L-BFGS code organization

  - Reset the accumulations of reused adjoint variables to zeros when one period of accumulation is finished;
  - Finish the accumulations of reused adjoint variables only when calculating backwards into its first use ;

- Handle the adjoint of iterative solver such as Gauss-Seidel ;
- Handle the adjoint of boundary conditions ;

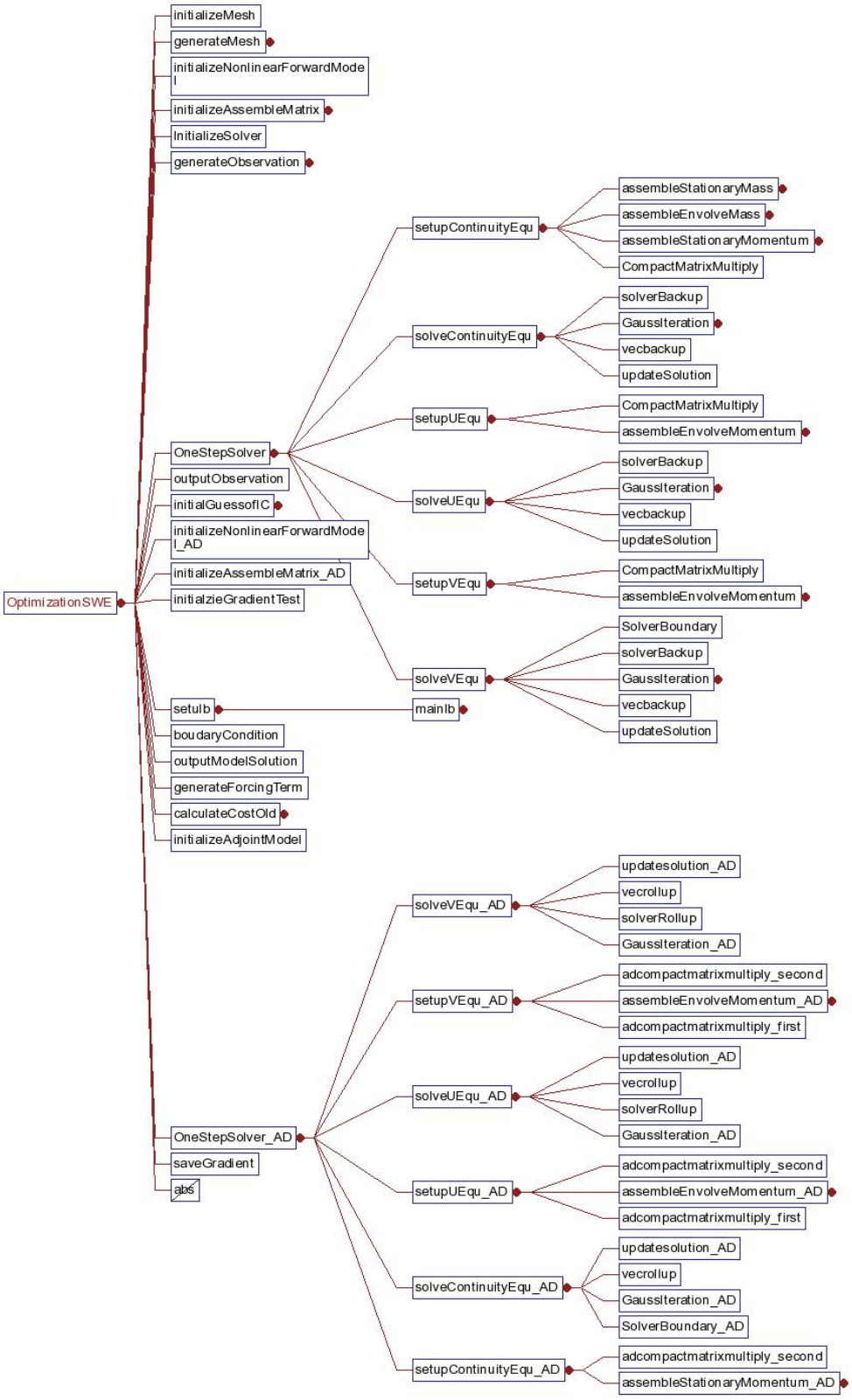


Figure 6.3: Calls graph of L-BFGS implementation

- Identifying the inputs and outputs of each subroutine and the whole program ;
- Make adjoint subroutines and parameters generic so that they can be reused for different adjoint variables without rewriting them over and over again.

6.3.3 Adjoint of iterative solver

The challenging part in the development of adjoint for nonlinear time-dependent discrete Galerkin Finite-Element model consists in the treatment of the Gauss-Seidel iterative procedure to solve the continuity equation systems and u-momentum equation systems as well as v-momentum linear systems, because some of the control variables to be solved at the current iteration level are reused while some are not(see Zhu, Navon and Zou 1994).

The key issues related to developing the adjoint of Gauss-Seidel iterative procedure are as follows:

We need to record the maximum number of the iterations when we integrate the nonlinear model forward in time, then, in order to obtain the adjoint of the Gauss-Seidel iterative procedure, the relationship of being reused among all the control variables must be analyzed. Finally, since the piecewise linear triangular Galerkin Finite-Element model has a local support of at most six nodes, while the minimum number of nodes is four when the node is on the boundary. Hence, the variable value at any given node inner or boundary is related to no more than six neighboring nodes surrounding it, and sometimes they are input variables and sometimes they are output variables. We are only concerned with the input variables when we speak about the reused variables, in other words, some of input variables in the iterative procedure are reused while other input variables are not, depending on the position in the grid as well as level of the iterations itself.

In addition, some control variables are firstly used in the setup of the *continuity system* and it will be used later twice in the setup of the *u-momentum system*. When dealing with situation to reuse adjoint variables in the adjoint code, we need to save the accumulated reused adjoint variables when calculating backwards into its first use. In other words, when we write the adjoint code, we need restore all the following accumulations into its first use when we finish the accumulation of reused adjoint variables.

6.3.4 TLM test

Prior to checking the correctness of the adjoint model, we need to check the correctness of the discrete TLM (Figure 6.4). One idea is to consider a state vector $\mathbf{X}$ and a perturbation $\mathbf{X}'$ so that we can use Taylor expansion to verify the correlation between nonlinear Galerkin FEM and its corresponding TLM:

$$\psi(\alpha) = \frac{\mathbf{G}\left(\mathbf{X} + \alpha\mathbf{X}'\right) - \mathbf{G}(\mathbf{X})}{\alpha\mathbf{P}(\mathbf{X}')} = 1 + O(\alpha) \tag{6.35}$$

where $\mathbf{G}$ denotes the nonlinear Galerkin FEM and $\mathbf{P}$ represents its TLM operator, αdefines the perturbation factor . Both the nonlinear Galerkin FEM and its TLM are integrated for a 5-hours period with various α values decreasing, and the results show that the correlationship between Nonlinear Galerkin FEM model and its TLM is almost equal to one as α tends to zero (Figure 6.5).

Therefore, if the TLM test can be correct, we only need to code the adjoint model directly from the discrete TLM by rewriting the code of TLM statement by statement in the opposite direction. This simplifies not only the complexity of constructing the adjoint model but also avoids the inconsistency generally arising from the derivation of the adjoint equations in analytic form followed by the discrete approximation(due to non-commutativity of discretization and adjoint operators)

In addition, we also use an alternative idea to test the TLM (and thus the adjoint). It's called the *complex-step derivative approximation*. It is reasonably straightforward to implement, and it requires only slight modifications in the forward model code. The feature of this method is that it can avoid some cancellations in the finite difference calculation that will result in the loss of digit accuracy(see Martins 2003).

6.3.5 Transpose test

The correctness of the adjoint model checked by following the algebraic expression :

$$(\mathbf{PX})^{\mathbf{T}}(\mathbf{PX}) = \mathbf{X}^{\mathbf{T}}\left(\mathbf{P}^{\mathbf{T}}(\mathbf{PX})\right) \tag{6.36}$$

where $\mathbf{X}$ represents the perturbation of input of the Galerkin FEM model, while the TLM denoted by $\mathbf{P}$ represents either a single *DO* loop or a *subroutine*. Each of them has its adjoint image *DO* loop or a *subroutine*, respectively. The left hand side involves only

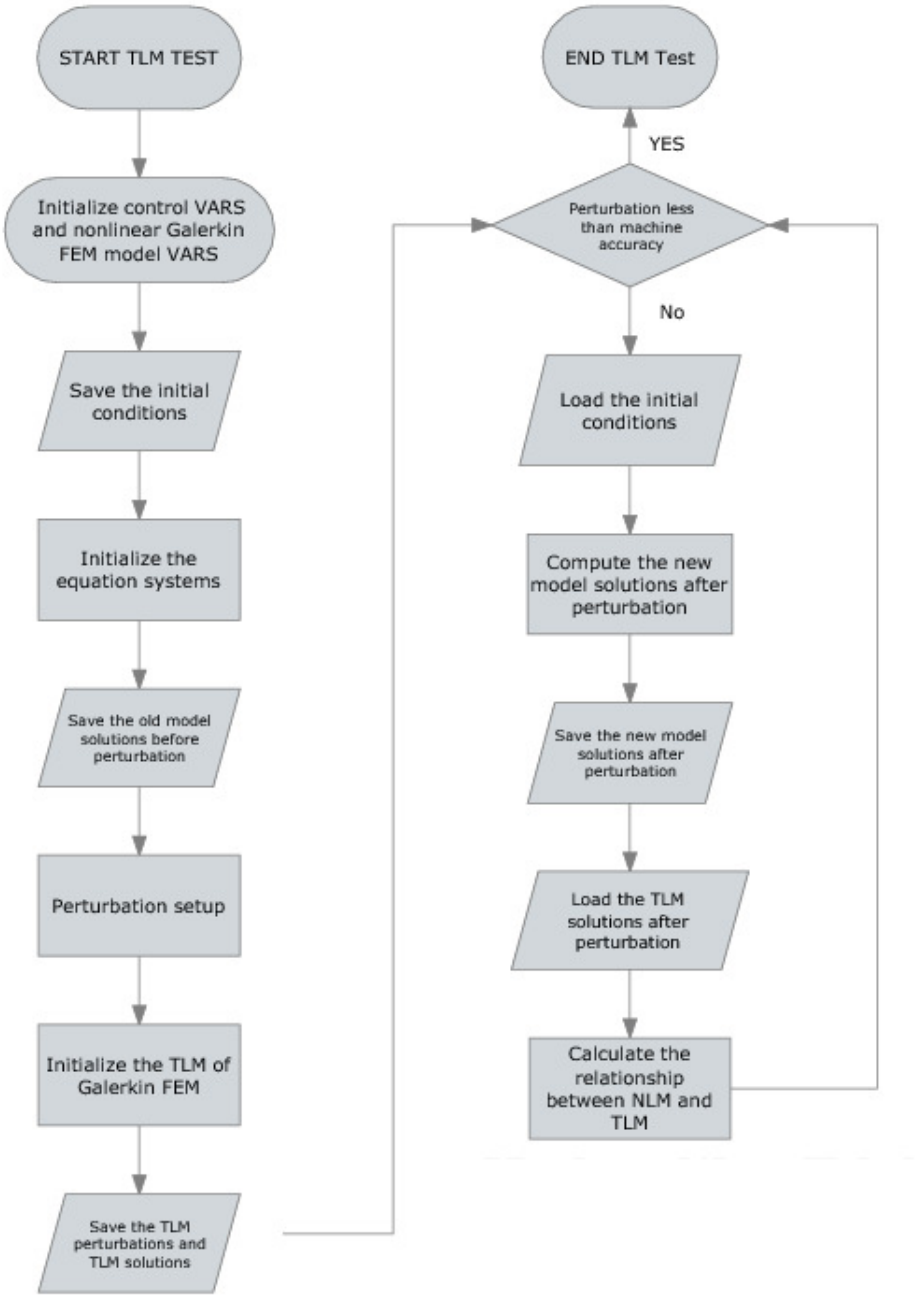


Figure 6.4: Flowchart of the Test of Tangent Linear Galerkin Finite-Element Model

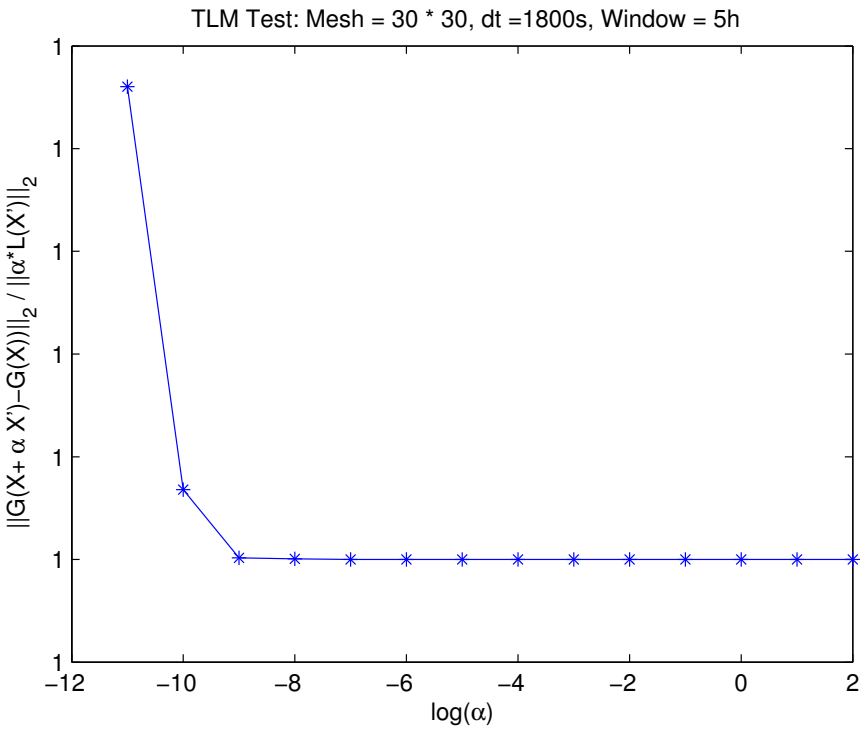


Figure 6.5: Correlation between Nonlinear Galerkin FEM model and its TLM, where α defines the perturbation factor.

the tangent linear code, while the right hand side involves also the adjoint code. When we implement it, we first run the TLM code and use the output vector as the input vector of the adjoint calculation. There are some issues where we need to be careful, when running the test. First, we need to make sure all the state variables have been saved when we integrate TLM forward and restored or loaded when we integrate its adjoint backward. Second, we may need to run the different inputs to make sure we go thorough a rigorous check of the adjoint code into each single part of it. Finally, the results obtained illustrated that a 13 digits accuracy can be achieved in the input/output tests by using DOUBLE PRECISION.

6.3.6 Gradient test

We also tested the accuracy of the gradient of the cost function by using the so-called α test as follows (Figure 6.6 and Figure 6.7):

$$F(\alpha) = \frac{J(\mathbf{X} + \alpha\mathbf{X}') - J(\mathbf{X})}{\alpha \left(\nabla J\right)^T \left(\nabla J\right)} = 1 + O(\alpha) \tag{6.37}$$

and the results show that the vector we obtained from the adjoint model is almost equal to the gradient as α decreasingly tends to zero, if α is not too close to the machine accuracy(see Navon 1992).

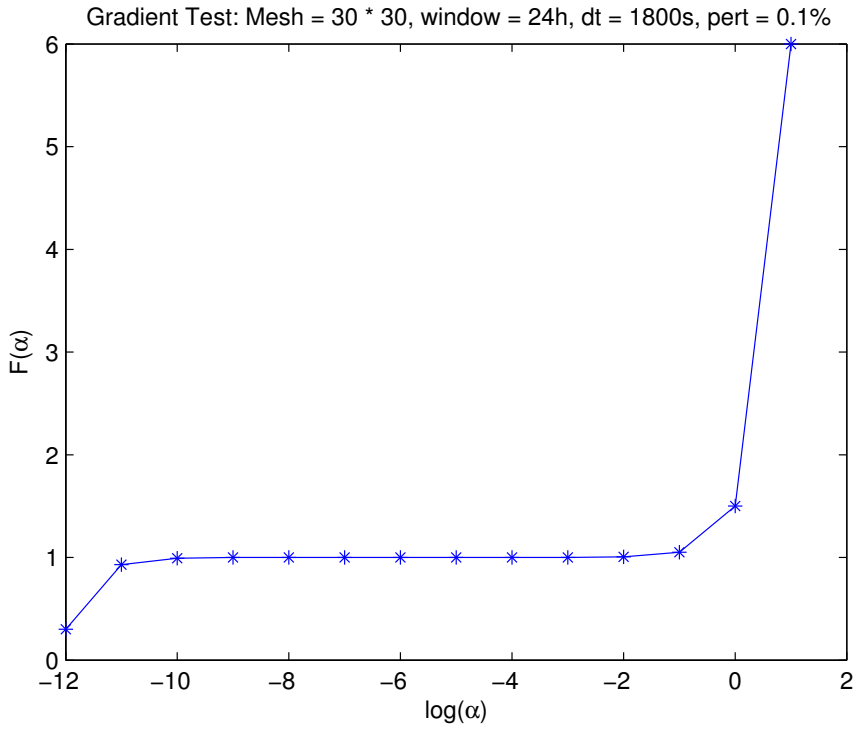


Figure 6.6: Gradient Test:Variation of $F(\alpha)$ with respect to $\log \alpha$.

6.4 Numerical Experiments

6.4.1 Description of Problem

The test problem used here adopts the initial conditions (Figure 6.8 and Figure 6.9) from the initial height field condition No.1 of Grammeltvedt (see Grammeltvedt 1969):

$$h(x,y) = H_0 + H_1 \tanh\left(\frac{9(D/2-y)}{2D}\right) + H_2\left(1/\cosh^2\left(\frac{9(D/2-y)}{D}\right)\right)\sin\left(\frac{2\pi x}{L}\right) \tag{6.38}$$

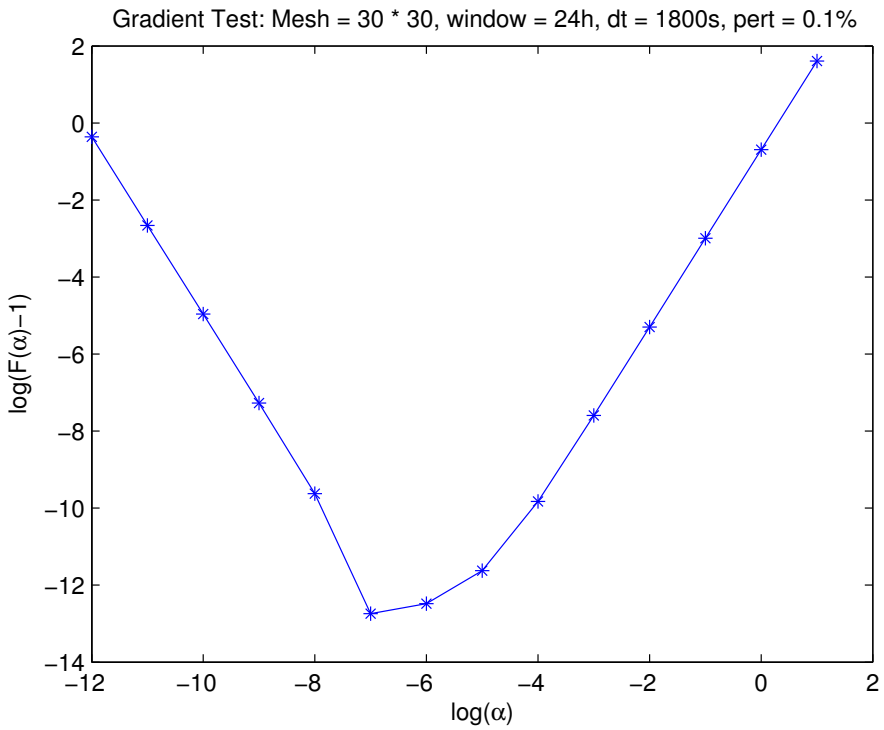


Figure 6.7: Gradient Test:Variation of $\log(F(\alpha) - 1)$ with respect to $\log \alpha$, where α defines the perturbation factor.

where this initial condition has energy in wave number one in the x-direction.

The initial velocity fields were derived from the initial height field using the geostrophic relationship:

$$u = -\left(\frac{g}{f}\right)\frac{\partial h}{\partial y} \quad v = \left(\frac{g}{f}\right)\frac{\partial h}{\partial x} \tag{6.39}$$

The dimensional constants used here are:

$$\begin{gathered} L = 4400km,\ D = 6000km,\ \bar{f} = 10^{-4}s^{-1},\ \beta = 1.5 \times 10^{-11}s^{-1}m^{-1}, \\ g = 10ms^{-1},\ H_0 = 2000m,\ H_1 = 220m,\ H_2 = 133m. \end{gathered} \tag{6.40}$$

and the space increments used here are

$$\Delta x = \Delta y = 400km \tag{6.41}$$

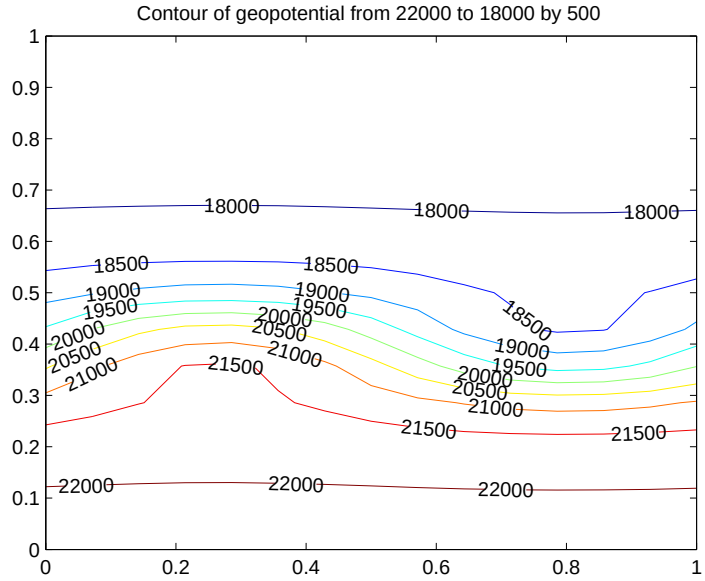


Figure 6.8: Initial geopotential

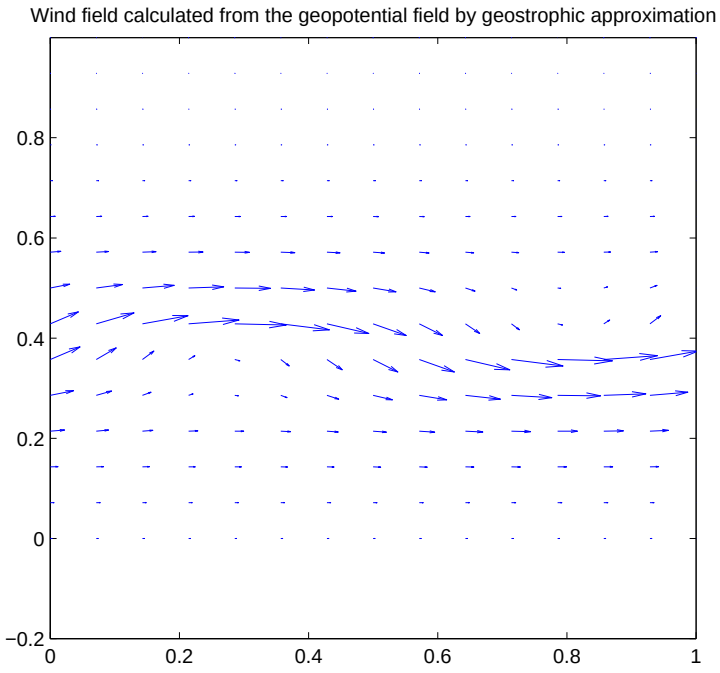


Figure 6.9: Initial wind fields

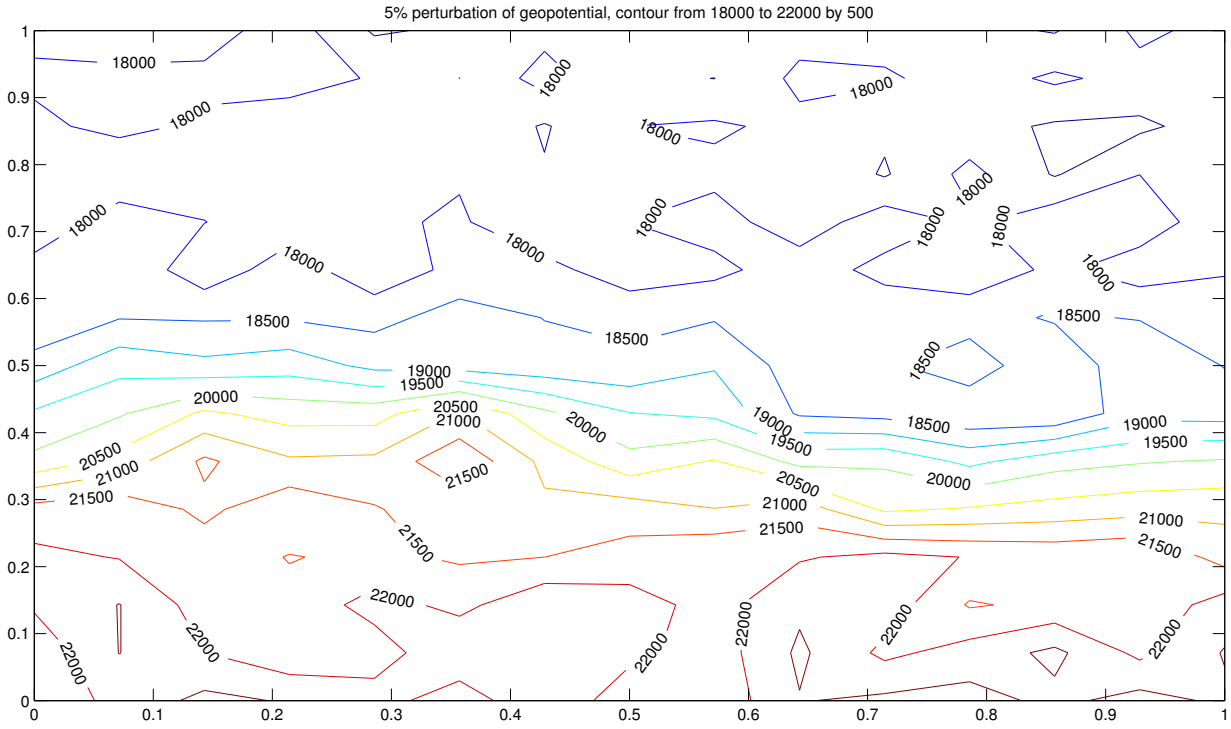


Figure 6.10: 5% random perturbation of the initial geopotential

6.4.2 Perturbation of initial conditions

We applied a 5% uniform random perturbations (Figure 6.10 and Figure 6.11) on the initial conditions in order to provide twin-experiment "observations" and we also computed the errors between the retrieved initial conditions related to the perturbed data and the reference state variables.

6.4.3 Retrieving the optimal initial conditions by applying L-BFGS

The accuracy of a short-range numerical weather prediction greatly depends on the initial and boundary conditions. The following experiments illustrate the technology to retrieve the optimal initial condition from a noisy initial conditions. First, we randomly perturb the initial conditions to generate the so-called observations at each time step. Second, we generate another rondom perturbations of the initial conditions to obtain a initial guess of the initial conditions in the optimization. In this paper, we tried limited quasi-Newton method of

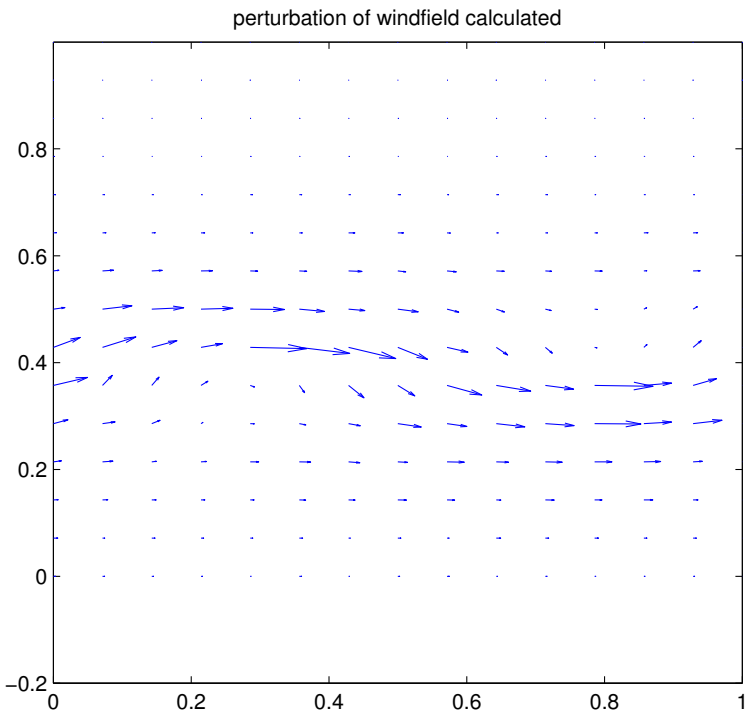


Figure 6.11: 5% random perturbation of the initial wind-field

Liu and Nocedal(1980,1989) and Richard and Nocedal(1995) to minimize the misfit between model solutions and artificial observations. The code is written in FORTRAN90 modularized with the control variables allocatable, so that any different mesh size can be tested in this code with a high accuracy. We also tested the different time steps as well as different data assimilation windows. The control variables are all defined as DOUBLE PRECISION so that a very high accuracy of approximation of the gradient of the cost functional with respect to the initial conditions can be achieved. In L-BFGS, we setup the number seven as the number of corrections($M = 7$)(See Liu and Nocedal 1989).

Testing different observations

The first experiment (Figure 6.12, Figure 6.13 and Figure 6.14) is performed on a short assimilation window for 12 hours with a small mesh size consisting of 15×15 grid points and we use a unconstrained minimization algorithm L-BFGS to minimize the cost functional. The adjoint model is integrated backward in time, with a forcing term being added, consisting

Table 6.1: L-BFGS: Data assimilation window = 12h, $\Delta x = \Delta y = 400km$, $\triangle t = 1800$s, and minimization convergence tolerance $\epsilon = 10^{-11}$

Random perturbations	Iterations	Function evaluations
5%	31	108
1%	28	99

of the difference between forecast and observation, interpolated at the same time and space location every time when an observation is encountered. We found out (Table 6.1) if we use 5% perturbation for both observations and initial guess, the L-BFGS converges in 31 iterations with 108 function evaluations to converge to prescribed tolerance $\epsilon = 10^{-11}$ (Figure 6.15 and Figure 6.16), but if we use 1% random perturbations, it will only take 28 iterations with 99 function evaluations to converge, which means both good observations and good initial guess will reduce the assimilation time required.

Furthermore, if we extend the assimilation window from 12 hours to 48 hours, the L-BFGS minimization fails to achieve the prescribed tolerance no matter how accurate the observations and initial guess we choose for the optimization algorithms. If the mesh size is too coarse, say 5×5 grid points, even if we use 12 hours assimilation window, we will still fail to converge by using L-BFGS, which means either a too large assimilation window or a too small mesh size will affect the ability of the L-BFGS algorithm to converge to achieve the prescribed tolerance.

Testing different mesh resolutions

By increasing the mesh resolution from 15×15 to 30×30 (Figure 6.17 and Figure 6.18) and still using L-BFGS, we found out that we can achieve a stricter tolerance $\epsilon = 10^{-16}$, although it requires more iterations and function evaluations to converge (Table 6.2). Hence, it can be observed that the rate of the convergence of the cost functional associated with the coarse mesh is faster than the rate of convergence corresponding to the fine-resolution models, however, the value of the cost functional associated with the fine mesh can be reduced to achieve a higher level of tolerance that is by five orders of magnitude better than minimization of the cost functional achieved for the coarse mesh.

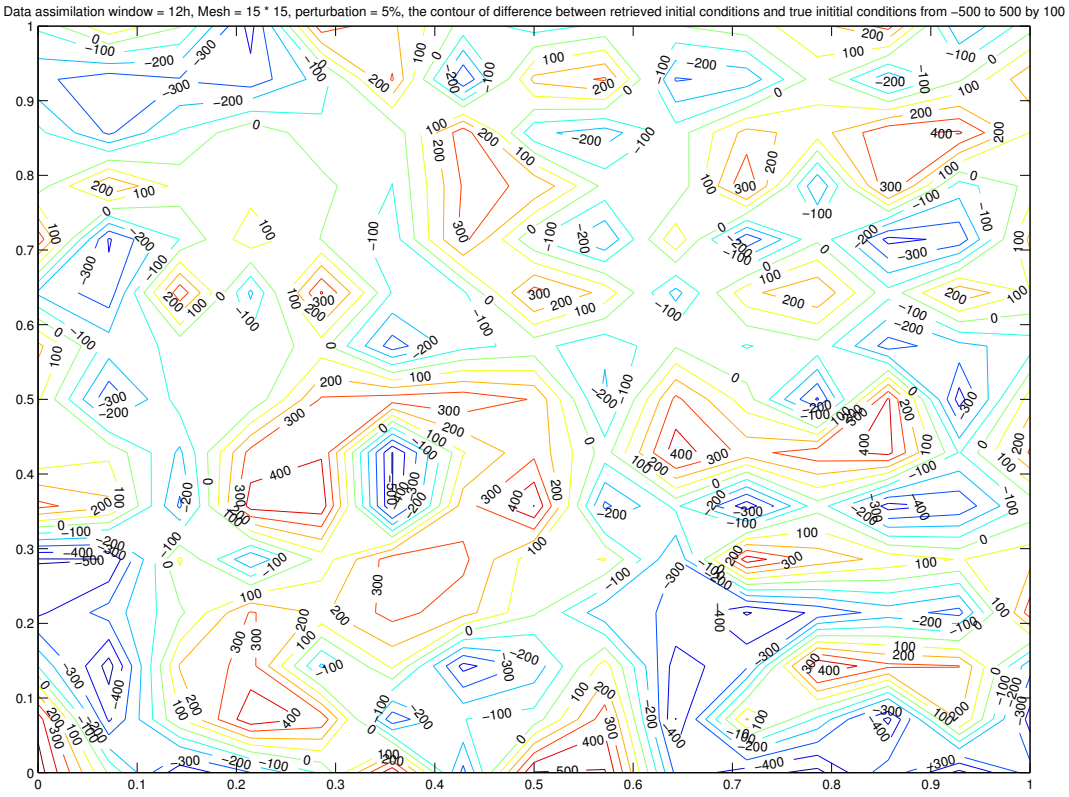


Figure 6.12: Data assimilation window = 12h, $\Delta x = \Delta y = 400km$, random perturbation = 5%. The contours of difference between retrieved initial geopotential and true initial geopotential are plotted.

Testing different time steps

By decreasing the time length from 1800s to 900s while keeping an identical data assimilation window of 12 hours, which requires more time steps, we can achieve a convergence of minimization with tolerance $\epsilon = 10^{-15}$ by using a coarse mesh size= 15×15, which is beneficial especially when there are not enough observations of a fine mesh in space available everywhere but we could have the ability to measure them for every short time step length, we may still retrieve a very high accuracy of optimal initial conditions by shrinking each time step length and expanding number of data assimilation steps (Table 6.3).

This can also be explained by noting that the results from the fine mesh integrated contain

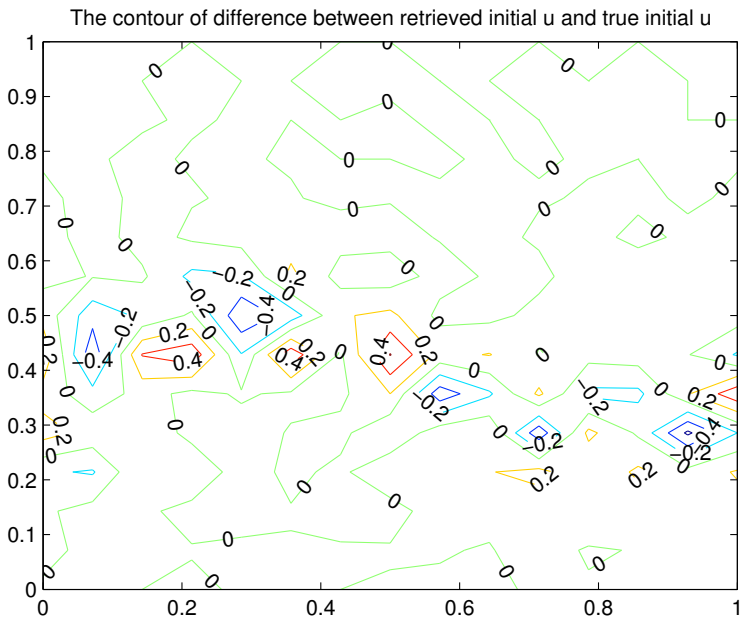


Figure 6.13: Data assimilation window = 12h, $\Delta x = \Delta y = 400km$, random perturbation = 5%. The contours of difference between retrieved initial u-momentum and true initial u-momentum from -0.5 to 0.5 by 0.2 are displayed.

Table 6.2: Results of using L-BFGS: data assimilation window = 12h, $\Delta x = \Delta y = 200km$, mesh resolution= 30×30, $\triangle t$= 1800s, and minimization convergence tolerance $\epsilon = 10^{-16}$

Random perturbations	Iterations	Function evaluations
5%	42	162
1%	38	149

more small-scale features than the corresponding ones from the coarse mesh integrated, and the dimension of the control variables also impacts upon the convergence rate so that the retrieval with fine-mesh model data becomes more difficult. The presence of small-scale results in an increase in the condition number of the Hessian of the cost function of the fine-mesh resolution model due to the introduction of small eigenvalues in the spectrum of the Hessian(see Axellson and Barker 1984). This situation becomes more apparent when the data assimilation is carried after a long time window of assimilation allowing reflections from limited boundaries thus causing short wave number noisy contaminations.

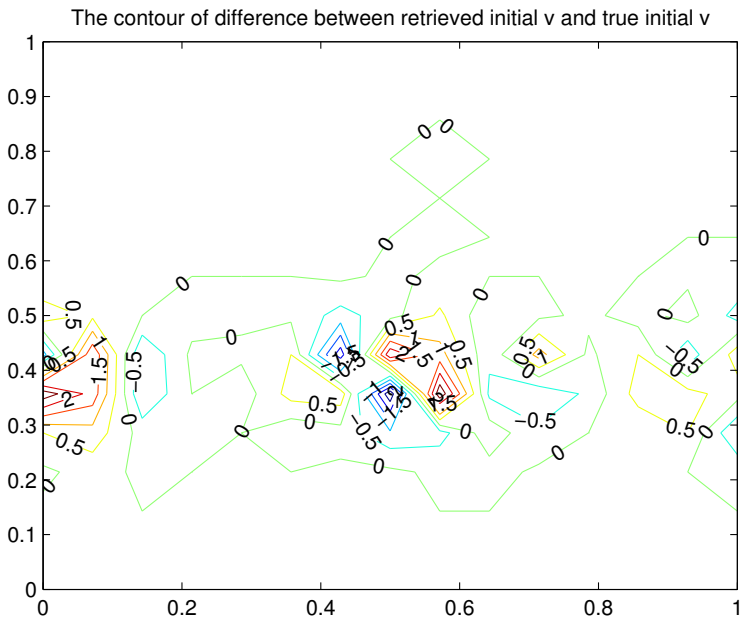


Figure 6.14: Data assimilation window = 12h, $\Delta x = \Delta y = 400km$, random perturbation = 5%. The contours of difference between retrieved initial v-momentum and true initial v-momentum from -0.3 to 0.3 by 0.05 are also displayed.

Table 6.3: Results of using L-BFGS: data assimilation window = 12h, $\Delta x = \Delta y = 400km$, random perturbations = 5%, $\triangle t$= 900s, and tolerance of convergence of minimization is $\epsilon = 10^{-15}$

mesh size	Iterations	Function evaluations
15×15	28	97
30×30	35	140

In this Chapter, we developed a modularized code written in FORTRAN90 to present a VDA scheme using Galerkin FEM and its adjoint to generate minimization algorithms used to minimize cost functional so as to yield optimal initial conditions using model forecast with observations. The challenging part in this paper is how to handle the reused variables especially in constructing the adjoint of Gauss-Seidel iterative procedure for the Finite-Element Shallow-Water equations model over a limited area domain.

The large-scale unconstrained minimization limited-memory quasi-Newton method writ-

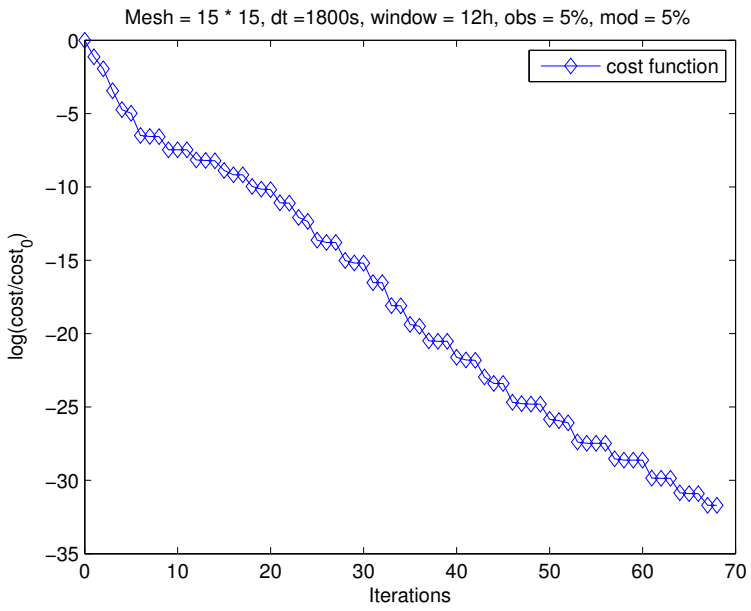


Figure 6.15: L-BFGS minimization: Normalized cost function scaled by initial cost function versus the number of minimization iterations

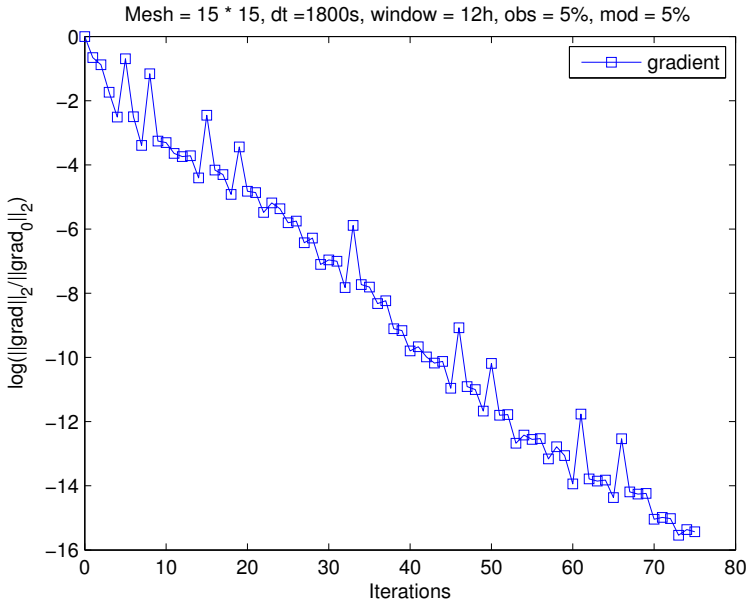


Figure 6.16: L-BFGS minimization: The norm of gradient scaled by initial norm of the gradient versus the number of minimization iterations

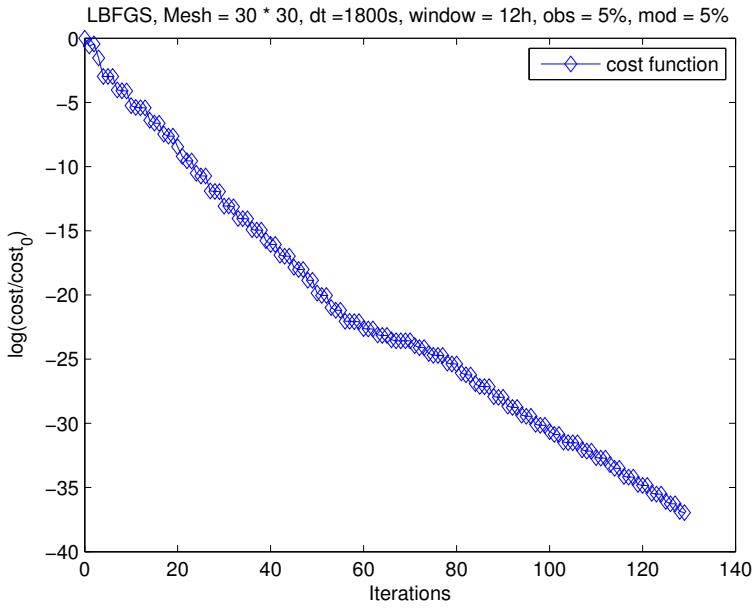


Figure 6.17: L-BFGS minimization: Normalized cost function scaled by initial cost function versus the number of minimization iterations

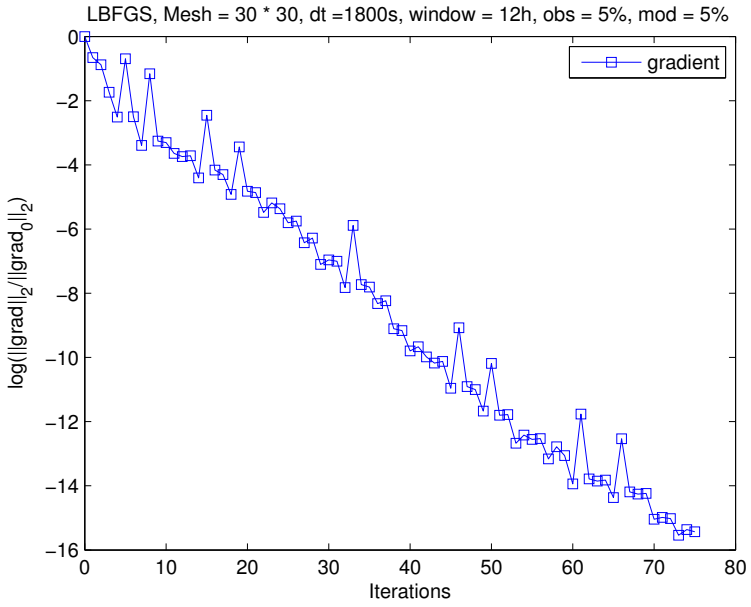


Figure 6.18: L-BFGS minimization: The norm of gradient scaled by initial norm of the gradient versus the number of minimization iterations

ten by Liu and Nocedal(1989) was used to minimize the cost functional consisting of difference between model solutions and observations over the large assimilation window. We used the full random perturbation of the No.1 of Grammeltvedt initial conditions(1969) to generate the observations and initial guess of the true initial conditions. We then carried the VDA numerical experiments using the adjoint model to assimilate the noisy observations.

The minimization of the cost functional was able to retrieve the true initial conditions when a coarse mesh size was employed. We also found out that the more accurate the observations as well as the initial guess of the initial conditions, the faster the rate of convergence of the minimization of the cost functional and the more accurate was the retrieval of the true initial conditions.

However, when carrying the L-BFGS to implement the VDA, it took a very long time to converge when applied to a very fine mesh and it failed to converge when a coarse mesh was employed. When we employed a coarse mesh in the model while using L-BFGS minimization and when observations were inserted frequently while shorter time steps were employed, we obtained similar accuracy results as in the case of fine mesh retrieval of the optimal initial conditions.

As we extended the length of the time window of the data assimilation of the forecast model, we impacted on the validity of the TLM model assumption and it became more and more difficult to employ the VDA scheme, since both effects of nonlinearity as well as limited area boundary conditions reflections impacted on the data assimilation procedure. To retrieve a high accuracy of optimal initial conditions, a fine mesh size is therefore required.

CHAPTER 7

ADAPTIVE POD 4-D VAR APPLIED TO FE-SWE MODEL

In this chapter, we address the POD model reduction along with inverse solution of a two-dimensional finite-element shallow-water equations model on a limited area domain. While there is a body of experience using POD model reduction for the shallow-water equations as well as for POD applied to 4-D VAR data assimilation of the shallow-water equations our intention is to draw on state of the art methodologies for efficient POD implementation, i.e. combining efficient snapshot selection in the presence of data assimilation system namely merging dual weighting of snapshots with trust region POD techniques.

The trust-region proper orthogonal decomposition (TRPOD) was recently proposed in [81, 82] as a way to overcome difficulties related POD ROM use in solving the Partial Differential Equation (PDE) constrained optimization problem. Combining POD technique with the concept of trust-region with general model functions (see Toint and Conn [71, 83] for a comprehensive survey or Nocedal and Wright [70] for an introduction to trust region methods) presents a framework for decision as to when an update of the POD ROM is necessary during the optimization process. Moreover, from a theoretical point of view for TRPOD, we have a global convergence result [81] proving that the iterates produced by the optimization algorithm, started at an arbitrary initial iterate, will converge to a local optimizer for the original mode.

The novelty of this contribution consists in assessing the combined effect of use of TRPOD in conjunction with dual weighting Data Assimilation System (DAS) snapshot selection in the framework of a relatively affordable, yet relevant model. One expects a beneficial cumulative effect from the combination of these two techniques. Comparisons to ad-hoc update adaptivity of the POD 4-D VAR and full 4-D VAR (high fidelity model) are carried

out for a variety of metrics to validate theoretical results in the light of numerical experiments. Indeed the combination of TRPOD and dual-weighted snapshots yields the best results in all metrics (see [100, 102]). For recent work on POD 4-D VAR, see [57, 58, 62, 63, 64, 65, 69].

The plan of the chapter is as follows: Section 7.1 provides the description of the generation of POD using a finite-element formulation. Section 7.2 details the POD Galerkin projection of FE-SWE model. Section 7.3 provides the framework of POD for reduced-order 4-D Var data assimilation of FE-SWE model. Section 7.4 details the numerical experiments carried out in order to validate accuracy of the POD reduced order model and the POD 4-D VAR approach for the various numerical methods enumerated above. For recent work on POD 4-D VAR, see [57, 58, 62, 63, 64, 65, 69]. In particular we compare ad-hoc adaptivity for POD 4-D VAR with trust-region adaptivity in combination with dual weighted snapshots. Finally, we provide error analysis of dual weighted trust-region POD 4-D Var compared to the high fidelity model. A discussion of numerical results thus obtained ensues. Finally we conclude with a conclusion section.

7.1 Generation of POD using Finite-Element formulation

The proper orthogonal decomposition identifies basis functions or modes which optimally capture the average energy content from numerical or experimental data. POD was introduced in the context of analysis of turbulent flow by Lumley [19], Berkooz et al. [20]. Sirovich[21] introduced the idea of snapshots. See also the book of Holmes [22].

Let Ω be a bounded domain in $\mathbb{R}^n$, the $L_2(\Omega)$ is defined as

$$L_2(\Omega) = \left\{ f(x),\ x \in \Omega : \int_\Omega f^2 d\Omega < \infty \right\} \tag{7.1}$$

with inner product

$$\langle f, g \rangle = \int_\Omega f g d\Omega \quad \forall f, g \in L_2(\Omega) \tag{7.2}$$

Given a set of sampled data

$$\mathbf{Y}^h = \left\{ y^{h,1}, y^{h,2}, \ldots, y^{h,n} \right\} \tag{7.3}$$

where $y^{h,i} \in L_2(\Omega)$ and $V = span\left(\mathbf{Y^h}\right) \subseteq \mathbb{R}^n$.

Let $\mathbf{K}$ be the correlation matrix of the data defined by

$$\mathbf{K} = \mathbf{Y^h}\left(\mathbf{Y^h}\right)^{\mathbf{T}} \tag{7.4}$$

where $\mathbf{K} = (k_{ij})_{n\times n}$, $k_{ij} = \left\langle y^{h,i}, y^{h,j}\right\rangle, i, j = 1, \ldots n$.

Then from all the subspaces $V_M \subset V$ with a fixed dimension M =dim(V_M) <dim(V),

$$\min_{V_M} \|\mathbf{Y} - \Pi_M \mathbf{Y}\| = \sum_{i=M+1}^{n} \lambda_i \tag{7.5}$$

where $\{\lambda_i\}_{i=1}^n$ are the non-negative ordered eigenvalues of symmetric matrix $\mathbf{K}$ and $\mathbf{\Psi^h} = \{\psi_i^h\}_{i=1}^n$ are the corresponding eigenvectors.

such that

$$\left\langle \psi^{h,i}, \psi^{h,j}\right\rangle = \delta_{ij} = \begin{cases} 1 & i = j \\ \\ 0 & i \neq j \end{cases} \tag{7.6}$$

Thus, the optimal subspace is given by

$$V_M = span\left\{ \ \psi_1^h, \ \ \psi_2^h, \ \ \ldots \ \ , \psi_M^h \ \right\} \tag{7.7}$$

in the sense that such type of POD bases describes more energy on average of the ensemble than any other linear basis of the same dimension,

with optimal orthogonal projection $\Pi_M : V \to V_M$, where $\Pi_M^2 = \Pi_M$ given by

$$\Pi_M = \sum_{i=1}^{M} \psi_i^h \left(\psi_i^h\right)^{\mathbf{T}} \tag{7.8}$$

Historically, in other disciplines, the same procedure goes by the names of Karhunen-Loeve decomposition (KLD)(see [14], [15]) or principal components analysis (PCA) and before them it was discovered by Kosambi [16].

The POD bases are applied with the Galerkin weak-form finite element method to create a reduced-order numerical model with reduced computational cost. It is well known that under some circumstances, Galerkin projections can produce unstable equilibrium points and limit cycles where the full system possesses stable equilibrium points and limit cycles. If energy-based inner product is used, then Galerkin projection preserves the stability of an equilibrium point at the origin [112, 113]. Snapshots bases consist of the flow solution for several flow solutions corresponding to different sets of parameter values evaluated at

different time instants of the model evolution. This involves solving the fully discretized model and saving states at various time instants in the time interval under consideration [22].

An ensemble of nodal-value represented snapshots chosen in the analysis time interval $[0, T]$ can be written as

$$\left\{y^1, y^2, \ldots, y^n\right\} \tag{7.9}$$

where $y^i \in \mathbb{R}^N$, $i = 1, \ldots, n$, n is the number of snapshots and N is the number of nodes in the mesh.

Define the weighted ensemble average of the finite-element represented data as

$$\bar{y} = \sum_{i=1}^{i=n} w_i y^i \tag{7.10}$$

where the snapshots weights w_i are such that $0 < w_i < 1$ and $\sum_{i=1}^{n} w_i = 1$, and they are used to assign a degree of importance to each member of the ensemble. Time weighting is usually considered, and in the standard approach $w_i = \frac{1}{n}$.

Hence,the finite-element represented POD solution can be expressed as

$$y^{POD} = \bar{y} + \sum_{i=1}^{i=M} \alpha_i(t)\, \psi_i \tag{7.11}$$

where

$$\Psi = \left\{\psi^1, \psi^2, \ldots, \psi^M\right\} \tag{7.12}$$

The nodal-value represented POD bases vectors $\mathbf{\Psi}$ and number of POD basis M are judiciously chosen to capture the dynamics of the flow as follows in the procedure described below:

1. The first step in creating a POD basis is to obtain a set of possible solution fields over the domain of the given problem. These fields will be generated through Finite Element (FE) analysis as described above, and are referred to as snapshots. The snapshot selection is crucial to the generalization capabilities of the POD basis, and a strategy to create the set of snapshots is vital.

2. Compute mean value of snapshots

$$\bar{y} = \sum_{i=1}^{i=n} w_i y^i \tag{7.13}$$

3. Subtract the mean from each snapshot and we obtain

$$\mathbf{Y} = \left\{y^1 - \bar{y}, y^2 - \bar{y}, \ldots, y^n - \bar{y}\right\} \tag{7.14}$$

4. Denote the finite-element basis [114] by

$$[\mathbf{V}] = [V_1, \ldots, V_n] \tag{7.15}$$

Compute the symmetric positive definite matrix

$$\mathbf{A} = \mathbf{V^T V} \tag{7.16}$$

and introduce a general form of inner product

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{A}} = \mathbf{x^T A y} \tag{7.17}$$

The POD basis of order $M \leq n$ provides an optimal representation of the ensemble data in M- dimensional state subspace by minimizing the averaged projection error

$$\min_{\{\psi^1, \psi^2, \ldots, \psi^M\}} \sum_{i=1}^{n} w_i \left\| \left(y^i - \bar{y}\right) - \Pi_{\Psi,M} \left(y^i - \bar{y}\right) \right\|_{\mathbf{A}}^2$$

$$\text{s.t. } \left\langle \psi^i, \psi^j \right\rangle_{\mathbf{A}} = \delta_{ij} \tag{7.18}$$

where $\Pi_{\Psi,M}$ is the projection operator onto the $M-$dimensional space

$$span\left\{\psi^1, \psi^2, \ldots, \psi^M\right\} \tag{7.19}$$

and

$$\Pi_{\Psi,M} = \sum_{i=1}^{M} \langle y, \psi_i \rangle_{\mathbf{A}} \psi_i \tag{7.20}$$

5. Build the weighted spatial correlation matrix

$$\mathbf{C} = \mathbf{YWY^T} \tag{7.21}$$

The POD modes $\psi^i \in \mathbb{R}^N$ are eigenvectors to the $N-$dimensional eigenvalue problem

$$\mathbf{CA}\psi_i = \lambda_i \psi_i$$

Since in practice the number of snapshots is much less than the the state dimension, $n << N$, an efficient way to compute the reduced basis is to introduce a $n-$dimensional matrix as follows:

$$\mathbf{K}^{n\times n} = \mathbf{W}^{\frac{1}{2}}\mathbf{K}\mathbf{W}^{\frac{1}{2}} = \mathbf{W}^{\frac{1}{2}}\mathbf{Y}^{\mathbf{T}}\mathbf{A}\mathbf{Y}\mathbf{W}^{\frac{1}{2}} \tag{7.22}$$

and compute the eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0$ of $\mathbf{K}^{n\times n}$ with its corresponding eigenvectors $\xi_1, \ldots, \xi_n$

6. The nodal-value represented POD basis vectors are obtained by defining

$$\psi_i = \frac{1}{\sqrt{\lambda_i}}\mathbf{Y}\mathbf{W}^{\frac{1}{2}}\xi_i, \quad i = 1, \ldots, M \tag{7.23}$$

and corresponding finite-element represented continuous POD basis can be expressed as

$$\left\{\psi^{h,1}, \psi^{h,2}, \ldots, \psi^{h,M}\right\} = \left\{\mathbf{V}\psi^1, \mathbf{V}\psi^2, \ldots, \mathbf{V}\psi^M\right\} \tag{7.24}$$

where

$$\left\langle \psi^{h,i}, \psi^{h,j} \right\rangle = \left\langle \psi^i, \psi^j \right\rangle_{\mathbf{A}} = \delta_{ij} = \begin{cases} 1 & i = j \\ \\ 0 & i \neq j \end{cases} \tag{7.25}$$

One can define a relative information content to choose a low-dimensional basis of size $M << n$ by neglecting modes corresponding to the small eigenvalues.

7.2 POD Galerkin Projection of FE-SWE model

For an atmospheric or oceanic flow defined in time interval $[0, T]$

$$\frac{dy}{dt} = F\left(y, t\right)$$

$$y\left(x, 0\right) = y_0\left(x\right) \tag{7.26}$$

To obtain a reduced model, we can first solve the PDE to obtain an ensemble of snapshots, then use a Galerkin projection scheme of the model equations onto the space spanned by the POD basis elements. We obtain the system of ODE as follows:

$$\frac{d\alpha_i}{dt} = \left\langle F\left(\bar{y}^h + \sum_{i=1}^{i=M} \alpha_i \psi_i^h, t\right), \psi_i^h \right\rangle \tag{7.27}$$

along with the initial conditions:

$$\alpha_i(0) = \left\langle y^h(x,0) - \bar{y}^h, \psi_i^h \right\rangle = \left\langle y_0 - \bar{y}, \psi_i \right\rangle_{\mathbf{A}}, \quad i = 1, \cdots, M \tag{7.28}$$

To obtain a POD reduced FE-SWE model, we reconstruct FE-SWE solutions based on POD basis as follows:

$$u = \bar{u} + \sum_{i=1}^{i=n_u} \alpha_i^u \psi_i^u$$

$$v = \bar{v} + \sum_{i=1}^{i=n_v} \alpha_i^v \psi_i^v$$

$$\phi = \bar{\phi} + \sum_{i=1}^{i=n_\phi} \alpha_i^\phi \psi_i^\phi$$

In the derivations below, we use the notaion $\langle\ ,\ \rangle$ to represent inner product, and $(f \centerdot g)$ to represent the pointwise product.

The system of ODE for $\dot{\alpha}_k^u$ after galerkin projection is as follows:

$$\dot{\alpha}_k^u = \left\langle \left(f \centerdot \bar{v} - \bar{u} \centerdot (\bar{u})_x - \bar{v} \centerdot (\bar{u})_y - \left(\bar{\phi}\right)_x \right), \psi_k^u \right\rangle$$

$$- \sum_{i=1}^{i=n_u} \alpha_i^u \left\langle \left(\psi_i^u \centerdot (\bar{u})_x + \bar{u} \centerdot (\psi_i^u)_x + \bar{v} \centerdot (\psi_i^u)_y \right), \psi_k^u \right\rangle$$

$$+ \sum_{i=1}^{i=n_v} \alpha_i^v \left\langle \left(f \centerdot \psi_i^v - \psi_i^v \centerdot (\bar{u})_y \right), \psi_k^u \right\rangle - \sum_{i=1}^{i=n_\phi} \alpha_i^\phi \left\langle \left(\psi_i^\phi\right)_x, \psi_k^u \right\rangle$$

$$- \sum_{i=1}^{i=n_u} \sum_{j=1}^{j=n_u} \alpha_i^u \alpha_j^u \left\langle \left(\psi_i^u \centerdot \left(\psi_j^u\right)_x \right), \psi_k^u \right\rangle - \sum_{i=1}^{i=n_u} \sum_{j=1}^{j=n_v} \alpha_i^u \alpha_j^v \left\langle \left((\psi_i^u)_y \centerdot \psi_j^v \right), \psi_k^u \right\rangle \tag{7.29}$$

By introducing the following notations,

$$b_k^1 = \left\langle \left(f \centerdot \bar{v} - \bar{u} \centerdot (\bar{u})_x - \bar{v} \centerdot (\bar{u})_y - \left(\bar{\phi}\right)_x \right), \psi_k^u \right\rangle$$

$$l_{ki}^{11} = - \left\langle \left(\psi_i^u \centerdot (\bar{u})_x + \bar{u} \centerdot (\psi_i^u)_x + \bar{v} \centerdot (\psi_i^u)_y \right), \psi_k^u \right\rangle$$

$$l_{ki}^{12} = \left\langle \left(f \cdot \psi_i^v - \psi_i^v \cdot (\bar{u})_y \right), \psi_k^u \right\rangle$$

$$l_{ki}^{13} = -\left\langle \left(\psi_i^\phi \right)_x, \psi_k^u \right\rangle$$

$$Q_{ijk}^{11} = -\left\langle \left(\psi_i^u \cdot \left(\psi_j^u \right)_x \right), \psi_k^u \right\rangle$$

$$Q_{ijk}^{12} = -\left\langle \left((\psi_i^u)_y \cdot \psi_j^v \right), \psi_k^u \right\rangle$$

we rewrite the system of ODE for $\dot{\alpha}_k^u$ as follows:

$$\begin{aligned} \dot{\alpha}_k^u &= b_k^1 + \sum_{i=1}^{i=n_u} l_{ki}^{11} \alpha_i^u + \sum_{i=1}^{i=n_v} l_{ki}^{12} \alpha_i^v + \sum_{i=1}^{i=n_\phi} l_{ki}^{13} \alpha_i^\phi \\ &+ \sum_{i=1}^{i=n_u} \sum_{j=1}^{j=n_u} Q_{ijk}^{11} \alpha_i^u \alpha_j^u + \sum_{i=1}^{i=n_u} \sum_{j=1}^{j=n_v} Q_{ijk}^{12} \alpha_i^u \alpha_j^v \quad k = 1, \cdots, n_u \end{aligned} \tag{7.30}$$

Using the matrix notation as follows:

$$\overrightarrow{\alpha^u} = (\alpha_1^u(t), \alpha_2^u(t), \ldots, \alpha_{n_u}^u(t))^{\mathbf{T}}$$

$$\overrightarrow{\alpha^v} = (\alpha_1^v(t), \alpha_2^v(t), \ldots, \alpha_{n_v}^v(t))^{\mathbf{T}}$$

$$\overrightarrow{\alpha^\phi} = (\alpha_1^\phi(t), \alpha_2^\phi(t), \ldots, \alpha_{n_\phi}^\phi(t))^{\mathbf{T}}$$

$$\mathbf{b^1} = \left(b_k^1 \right)_{n_u \times 1}$$

$$\mathbf{L^{11}} = \left(L_{ki}^{11} \right)_{n_u \times n_u}$$

$$\mathbf{L^{12}} = \left(L_{ki}^{12} \right)_{n_u \times n_v}$$

$$\mathbf{L^{13}} = \left(L^{13}_{ki}\right)_{n_u \times n_\phi}$$

$$\mathbf{Q^{11}} = \left(Q^{11}_{ijk}\right)_{n_u \times n_u \times n_u}$$

$$\mathbf{Q^{12}} = \left(Q^{12}_{ijk}\right)_{n_u \times n_u \times n_v}$$

Therefore, we could write the system of ODE for $\dot{\alpha}^u_k$ into a matrix as follows:

$$\mathbf{A_u}\dot{\overrightarrow{\alpha^u}} = \mathbf{b^1} + \mathbf{L^{11}}\overrightarrow{\alpha^u} + \mathbf{L^{12}}\overrightarrow{\alpha^v} + \mathbf{L^{13}}\overrightarrow{\alpha^\phi} + \left(\overrightarrow{\alpha^u}\right)^{\mathbf{T}} \mathbf{Q^{11}} \left(\overrightarrow{\alpha^u}\right) + \left(\overrightarrow{\alpha^u}\right)^{\mathbf{T}} \mathbf{Q^{12}} \left(\overrightarrow{\alpha^v}\right) \tag{7.31}$$

Similarly, the system of ODE for $\dot{\alpha}^v_k$ after galerkin projection is as follows:

$$\begin{aligned}
\dot{\alpha}^v_k = &- \left\langle f \cdot \bar{u} + \bar{u} \cdot (\bar{v})_x + \bar{v} \cdot (\bar{v})_y + \left(\bar{\phi}\right)_y, \psi^v_k \right\rangle \\
&- \sum_{i=1}^{i=n_u} \alpha^u_i \left\langle f \cdot \psi^u_i + \left(\psi^u_i \cdot (\bar{v})_x\right), \psi^v_k \right\rangle \\
&- \sum_{i=1}^{i=n_v} \alpha^v_i \left\langle \left(\psi^v_i \cdot (\bar{v})_y + \bar{u} \cdot (\psi^v_i)_x + \bar{v} \cdot (\psi^v_i)_y\right), \psi^v_k \right\rangle - \sum_{i=1}^{i=n_\phi} \alpha^\phi_i \left\langle \left(\psi^\phi_i\right)_y, \psi^v_k \right\rangle \\
&- \sum_{i=1}^{i=n_v} \alpha^v_i \left\langle \left(\psi^v_i \cdot (\bar{v})_y + \bar{u} \cdot (\psi^v_i)_x + \bar{v} \cdot (\psi^v_i)_y\right), \psi^v_k \right\rangle - \sum_{i=1}^{i=n_\phi} \alpha^\phi_i \left\langle \left(\psi^\phi_i\right)_y, \psi^v_k \right\rangle \\
&- \sum_{i=1}^{i=n_u} \sum_{j=1}^{j=n_v} \alpha^u_i \alpha^v_j \left\langle \left(\psi^u_i \cdot \left(\psi^v_j\right)_x\right), \psi^v_k \right\rangle - \sum_{i=1}^{i=n_v} \sum_{j=1}^{j=n_v} \alpha^v_i \alpha^v_j \left\langle \left(\psi^v_i \cdot \left(\psi^v_j\right)_y\right), \psi^v_k \right\rangle
\end{aligned} \tag{7.32}$$

Using the notations as follows:

$$b^2_k = - \left\langle f \cdot \bar{u} + \bar{u} \cdot (\bar{v})_x + \bar{v} \cdot (\bar{v})_y + \left(\bar{\phi}\right)_y, \psi^v_k \right\rangle$$

$$l_{ki}^{21} = -\left\langle f \centerdot \psi_i^u + \left(\psi_i^u \centerdot (\bar{v})_x\right), \psi_k^v \right\rangle$$

$$l_{ki}^{22} = -\left\langle \left(\psi_i^v \centerdot (\bar{v})_y + \bar{u} \centerdot (\psi_i^v)_x + \bar{v} \centerdot (\psi_i^v)_y\right), \psi_k^v \right\rangle$$

$$l_{ki}^{23} = -\left\langle \left(\psi_i^\phi\right)_y, \psi_k^v \right\rangle$$

$$Q_{ijk}^{21} = -\left\langle \left(\psi_i^u \centerdot \left(\psi_j^v\right)_x\right), \psi_k^v \right\rangle$$
$$Q_{ijk}^{22} = -\left\langle \left(\psi_i^v \centerdot \left(\psi_j^v\right)_y\right), \psi_k^v \right\rangle$$

Hence, we could rewrite the system of ODE for $\dot{\alpha}_k^v$ as follows:

$$\begin{aligned} \dot{\alpha}_k^v &= b_k^2 + \sum_{i=1}^{i=n_u} l_{ki}^{21}\alpha_i^u + \sum_{i=1}^{i=n_v} l_{ki}^{22}\alpha_i^v + \sum_{i=1}^{i=n_\phi} l_{ki}^{23}\alpha_i^\phi \\ &\quad + \sum_{i=1}^{i=n_u}\sum_{j=1}^{j=n_v} Q_{ijk}^{21}\alpha_i^u\alpha_j^v + \sum_{i=1}^{i=n_v}\sum_{j=1}^{j=n_v} Q_{ijk}^{22}\alpha_i^v\alpha_j^v \quad k = 1, \cdots, n_v \end{aligned} \tag{7.33}$$

Using the matrix notation as follows:

$$\mathbf{b^2} = \left(b_k^2\right)_{n_v \times 1}$$

$$\mathbf{L^{21}} = \left(L_{ki}^{21}\right)_{n_v \times n_u}$$

$$\mathbf{L^{22}} = \left(L_{ki}^{22}\right)_{n_v \times n_v}$$

$$\mathbf{L^{23}} = \left(L_{ki}^{23}\right)_{n_v \times n_\phi}$$

$$\mathbf{Q^{21}} = \left(Q_{ijk}^{22}\right)_{n_u \times n_v \times n_v}$$

$$\mathbf{Q^{22}} = \left(Q_{ijk}^{21}\right)_{n_v \times n_v \times n_v}$$

Therefore, we could write the system of ODE for $\dot{\alpha}_k^v$ into a matrix as follows:

$$\mathbf{A_v}\dot{\overrightarrow{\alpha^v}} = \mathbf{b^2} + \mathbf{L^{21}}\overrightarrow{\alpha^u} + \mathbf{L^{22}}\overrightarrow{\alpha^v} + \mathbf{L^{23}}\overrightarrow{\alpha^\phi} + \left(\overrightarrow{\alpha^u}\right)^{\mathbf{T}} \mathbf{Q^{21}} \left(\overrightarrow{\alpha^v}\right) + \left(\overrightarrow{\alpha^v}\right)^{\mathbf{T}} \mathbf{Q^{22}} \left(\overrightarrow{\alpha^v}\right) \tag{7.34}$$

Finally, the system of ODE for $\dot{\alpha}_k^\phi$ after galerkin projection is as follows:

$$\begin{aligned}
\dot{\alpha}_k^\phi = &- \left\langle \bar{u} \cdot \left(\bar{\phi}\right)_x + \bar{v} \cdot \left(\bar{\phi}\right)_y + \bar{\phi} \cdot (\bar{u})_x + \bar{\phi} \cdot (\bar{v})_y , \psi_k^\phi \right\rangle \\
&- \sum_{i=1}^{i=n_u} \alpha_i^u \left\langle \left(\left(\bar{\phi}\right)_x \cdot \psi_i^u + \bar{\phi} \cdot \left(\psi_i^u\right)_x \right), \psi_k^\phi \right\rangle \\
&- \sum_{i=1}^{i=n_v} \alpha_i^v \left\langle \left(\psi_i^v \cdot \left(\bar{\phi}\right)_y + \bar{\phi} \cdot \left(\psi_i^v\right)_y \right), \psi_k^\phi \right\rangle \\
&- \sum_{i=1}^{i=n_\phi} \alpha_i^\phi \left\langle \left(\psi_i^\phi \cdot (\bar{u})_x + \psi_i^\phi \cdot (\bar{v})_y + \bar{u} \cdot \left(\psi_i^\phi\right)_x + \bar{v} \cdot \left(\psi_i^\phi\right)_y \right), \psi_k^\phi \right\rangle \\
&- \sum_{i=1}^{i=n_u} \sum_{j=1}^{j=n_\phi} \alpha_i^u \alpha_j^\phi \left\langle \left(\psi_i^u \cdot \left(\psi_j^\phi\right)_x + \left(\psi_i^u\right)_x \cdot \psi_j^\phi \right), \psi_k^\phi \right\rangle \\
&- \sum_{i=1}^{i=n_v} \sum_{j=1}^{j=n_\phi} \alpha_i^v \alpha_j^\phi \left\langle \left(\psi_i^v \cdot \left(\psi_j^\phi\right)_y + \left(\psi_i^v\right)_y \cdot \psi_j^\phi \right), \psi_k^\phi \right\rangle
\end{aligned} \tag{7.35}$$

Using the notations as follows:

$$b_k^3 = - \left\langle \bar{u} \cdot \left(\bar{\phi}\right)_x + \bar{v} \cdot \left(\bar{\phi}\right)_y + \bar{\phi} \cdot (\bar{u})_x + \bar{\phi} \cdot (\bar{v})_y , \psi_k^\phi \right\rangle$$

$$l_{ki}^{31} = - \left\langle \left(\left(\bar{\phi}\right)_x \cdot \psi_i^u + \bar{\phi} \cdot \left(\psi_i^u\right)_x \right), \psi_k^\phi \right\rangle$$

$$l_{ki}^{32} = - \left\langle \left(\psi_i^v \cdot \left(\bar{\phi}\right)_y + \bar{\phi} \cdot \left(\psi_i^v\right)_y \right), \psi_k^\phi \right\rangle$$

$$l_{ki}^{33} = - \left\langle \left(\psi_i^\phi \cdot (\bar{u})_x + \psi_i^\phi \cdot (\bar{v})_y + \bar{u} \cdot \left(\psi_i^\phi\right)_x + \bar{v} \cdot \left(\psi_i^\phi\right)_y \right), \psi_k^\phi \right\rangle$$

$$Q_{ijk}^{31} = - \left\langle \left(\psi_i^u \cdot \left(\psi_j^\phi\right)_x + \left(\psi_i^u\right)_x \cdot \psi_j^\phi \right), \psi_k^\phi \right\rangle$$

$$Q_{ijk}^{32} = -\left\langle \left(\psi_i^v \cdot \left(\psi_j^\phi \right)_y + \left(\psi_i^v \right)_y \cdot \psi_j^\phi \right), \psi_k^\phi \right\rangle$$

Using the matrix notation as follows:

$$\mathbf{b^3} = \left(b_k^3 \right)_{n_\phi \times 1}$$

$$\mathbf{L^{31}} = \left(L_{ki}^{31} \right)_{n_\phi \times n_u}$$

$$\mathbf{L^{32}} = \left(L_{ki}^{32} \right)_{n_\phi \times n_v}$$

$$\mathbf{L^{33}} = \left(L_{ki}^{33} \right)_{n_\phi \times n_\phi}$$

$$\mathbf{Q^{31}} = \left(Q_{ijk}^{31} \right)_{n_u \times n_\phi \times n_\phi}$$

$$\mathbf{Q^{32}} = \left(Q_{ijk}^{32} \right)_{n_v \times n_\phi \times n_\phi}$$

Therefore, we could write the system of ODE for $\dot{\alpha}_k^\phi$ into a matrix as follows:

$$\mathbf{A}_\phi \dot{\overrightarrow{\alpha^\phi}} = \mathbf{b^3} + \mathbf{L^{31}} \overrightarrow{\alpha^u} + \mathbf{L^{32}} \overrightarrow{\alpha^v} + \mathbf{L^{33}} \overrightarrow{\alpha^\phi} + \left(\overrightarrow{\alpha^u} \right)^{\mathbf{T}} \mathbf{Q^{31}} \left(\overrightarrow{\alpha^\phi} \right) + \left(\overrightarrow{\alpha^v} \right)^{\mathbf{T}} \mathbf{Q^{32}} \left(\overrightarrow{\alpha^\phi} \right) \tag{7.36}$$

To sum up, we have the system of ODE as follows:

$$\begin{cases} \mathbf{A_u} \dot{\overrightarrow{\alpha^u}} = \mathbf{b^1} + \mathbf{L^{11}} \overrightarrow{\alpha^u} + \mathbf{L^{12}} \overrightarrow{\alpha^v} + \mathbf{L^{13}} \overrightarrow{\alpha^\phi} + \left(\overrightarrow{\alpha^u} \right)^{\mathbf{T}} \mathbf{Q^{11}} \left(\overrightarrow{\alpha^u} \right) + \left(\overrightarrow{\alpha^u} \right)^{\mathbf{T}} \mathbf{Q^{12}} \left(\overrightarrow{\alpha^v} \right) \\ \mathbf{A_v} \dot{\overrightarrow{\alpha^v}} = \mathbf{b^2} + \mathbf{L^{21}} \overrightarrow{\alpha^u} + \mathbf{L^{22}} \overrightarrow{\alpha^v} + \mathbf{L^{23}} \overrightarrow{\alpha^\phi} + \left(\overrightarrow{\alpha^u} \right)^{\mathbf{T}} \mathbf{Q^{21}} \left(\overrightarrow{\alpha^v} \right) + \left(\overrightarrow{\alpha^v} \right)^{\mathbf{T}} \mathbf{Q^{22}} \left(\overrightarrow{\alpha^v} \right) \\ \mathbf{A}_\phi \dot{\overrightarrow{\alpha^\phi}} = \mathbf{b^3} + \mathbf{L^{31}} \overrightarrow{\alpha^u} + \mathbf{L^{32}} \overrightarrow{\alpha^v} + \mathbf{L^{33}} \overrightarrow{\alpha^\phi} + \left(\overrightarrow{\alpha^u} \right)^{\mathbf{T}} \mathbf{Q^{31}} \left(\overrightarrow{\alpha^\phi} \right) + \left(\overrightarrow{\alpha^v} \right)^{\mathbf{T}} \mathbf{Q^{32}} \left(\overrightarrow{\alpha^\phi} \right) \end{cases} \tag{7.37}$$

along with the initial conditions:

$$\begin{cases} \alpha_k^u(0) = \left(u(\vec{x}, 0) - \bar{u}, \psi_k^u \right), & k = 1, \cdots, n_u \\ \alpha_k^v(0) = \left(v(\vec{x}, 0) - \bar{v}, \psi_k^v \right), & k = 1, \cdots, n_v \\ \alpha_k^\phi(0) = \left(\phi(\vec{x}, 0) - \bar{\phi}, \psi_k^\phi \right), & k = 1, \cdots, n_\phi \end{cases} \tag{7.38}$$

where

$$\overrightarrow{\alpha^u} = (\alpha_1^u(t), \alpha_2^u(t), \ldots, \alpha_{n_u}^u(t))^{\mathbf{T}}$$

$$\overrightarrow{\alpha^v} = (\alpha_1^v(t), \alpha_2^v(t), \ldots, \alpha_{n_v}^v(t))^{\mathbf{T}}$$

$$\overrightarrow{\alpha^\phi} = (\alpha_1^\phi(t), \alpha_2^\phi(t), \ldots, \alpha_{n_\phi}^\phi(t))^{\mathbf{T}}$$

In the following sections, we consider a total energy norm defined as

$$\|\mathbf{y}\|^2_{\mathbf{A^{FEM}}} = \frac{1}{2}\left(\|u\|^2_{L_2(\Omega)} + \|v\|^2_{L_2(\Omega)} + \frac{g}{\overline{h}}\|u\|^2_{L_2(\Omega)}\right)$$

$$= \frac{1}{2}\left(u^{\mathbf{T}}\mathbf{A}u + v^{\mathbf{T}}\mathbf{A}v + \frac{g}{\overline{h}}h^{\mathbf{T}}\mathbf{A}h\right) = \mathbf{y}^{\mathbf{T}}\mathbf{A}^{\mathbf{FEM}}\mathbf{y} \tag{7.39}$$

where $\mathbf{A} = \mathbf{V}^{\mathbf{T}}\mathbf{V}$ is a symmetric positive definite matrix and $[\mathbf{V}] = [V_1, \ldots, V_N]$ is the finite-element basis, $\overline{h}$ is the mean height of the reference data at the initial time. Hence $\mathbf{A^{FEM}}$ can be viewed as a symmetric positive definite block-wise diagonal matrix:

$$\mathbf{A^{FEM}} = \text{diag}\left(\begin{array}{ccc} \frac{1}{2}\mathbf{A} & \frac{1}{2}\mathbf{A} & \frac{g}{2\overline{h}}\mathbf{A} \end{array}\right) \tag{7.40}$$

7.3 Optimal Control of POD reduced FE-SWE model

Let us define

$$\overrightarrow{\alpha} = \left(\overrightarrow{\alpha^u}, \overrightarrow{\alpha^v}, \overrightarrow{\alpha^\phi}\right)^{\mathbf{T}} \tag{7.41}$$

$$F = \begin{pmatrix} F_1 \\ F_2 \\ F_3 \end{pmatrix} =$$

$$\begin{pmatrix} \mathbf{b^1} + \mathbf{L^{11}}\overrightarrow{\alpha^u} + \mathbf{L^{12}}\overrightarrow{\alpha^v} + \mathbf{L^{13}}\overrightarrow{\alpha^\phi} + \left(\overrightarrow{\alpha^u}\right)^{\mathbf{T}}\mathbf{Q^{11}}\left(\overrightarrow{\alpha^u}\right) + \left(\overrightarrow{\alpha^u}\right)^{\mathbf{T}}\mathbf{Q^{12}}\left(\overrightarrow{\alpha^v}\right) \\ \mathbf{b^2} + \mathbf{L^{21}}\overrightarrow{\alpha^u} + \mathbf{L^{22}}\overrightarrow{\alpha^v} + \mathbf{L^{23}}\overrightarrow{\alpha^\phi} + \left(\overrightarrow{\alpha^u}\right)^{\mathbf{T}}\mathbf{Q^{21}}\left(\overrightarrow{\alpha^v}\right) + \left(\overrightarrow{\alpha^v}\right)^{\mathbf{T}}\mathbf{Q^{22}}\left(\overrightarrow{\alpha^v}\right) \\ \mathbf{b^3} + \mathbf{L^{31}}\overrightarrow{\alpha^u} + \mathbf{L^{32}}\overrightarrow{\alpha^v} + \mathbf{L^{33}}\overrightarrow{\alpha^\phi} + \left(\overrightarrow{\alpha^u}\right)^{\mathbf{T}}\mathbf{Q^{31}}\left(\overrightarrow{\alpha^\phi}\right) + \left(\overrightarrow{\alpha^v}\right)^{\mathbf{T}}\mathbf{Q^{32}}\left(\overrightarrow{\alpha^\phi}\right) \end{pmatrix} \tag{7.42}$$

It is easy to verify that

$$\frac{\partial F_1}{\partial \overrightarrow{\alpha^u}} = \mathbf{L^{11}}\left(\delta\overrightarrow{\alpha^u}\right) + \tag{7.43}$$

$$\begin{pmatrix} \left(\overrightarrow{\alpha^u}\right)^{\mathbf{T}}\left(\mathbf{Q_1^{11}} + \left(\mathbf{Q_1^{11}}\right)^{\mathbf{T}}\right)\delta\overrightarrow{\alpha^u} \\ \vdots \\ \left(\overrightarrow{\alpha^u}\right)^{\mathbf{T}}\left(\mathbf{Q_{n_u}^{11}} + \left(\mathbf{Q_{n_u}^{11}}\right)^{\mathbf{T}}\right)\delta\overrightarrow{\alpha^u} \end{pmatrix} + \begin{pmatrix} \left(\overrightarrow{\alpha^v}\right)^{\mathbf{T}}\left(\mathbf{Q_1^{12}}\right)^{\mathbf{T}}\delta\overrightarrow{\alpha^u} \\ \vdots \\ \left(\overrightarrow{\alpha^v}\right)^{\mathbf{T}}\left(\mathbf{Q_{n_u}^{12}}\right)^{\mathbf{T}}\delta\overrightarrow{\alpha^u} \end{pmatrix} \tag{7.44}$$

$$\frac{\partial F_1}{\partial \overrightarrow{\alpha^v}} = \mathbf{L^{12}}\left(\delta\overrightarrow{\alpha^v}\right) + \begin{pmatrix} \left(\overrightarrow{\alpha^u}\right)^{\mathbf{T}}\mathbf{Q}_1^{12}\delta\overrightarrow{\alpha^v} \\ \vdots \\ \left(\overrightarrow{\alpha^u}\right)^{\mathbf{T}}\mathbf{Q}_{n_u}^{12}\delta\overrightarrow{\alpha^v} \end{pmatrix} \tag{7.45}$$

$$\frac{\partial F_1}{\partial \overrightarrow{\alpha^\phi}} = \mathbf{L^{13}}\left(\delta\overrightarrow{\alpha^\phi}\right) \tag{7.46}$$

$$\frac{\partial F_2}{\partial \overrightarrow{\alpha^u}} = \mathbf{L^{21}}\left(\delta\overrightarrow{\alpha^u}\right) + \begin{pmatrix} \left(\overrightarrow{\alpha^v}\right)^{\mathbf{T}}\left(\mathbf{Q_1^{21}}\right)^{\mathbf{T}}\delta\overrightarrow{\alpha^u} \\ \vdots \\ \left(\overrightarrow{\alpha^v}\right)^{\mathbf{T}}\left(\mathbf{Q_{n_v}^{21}}\right)^{\mathbf{T}}\delta\overrightarrow{\alpha^u} \end{pmatrix} \tag{7.47}$$

$$\frac{\partial F_2}{\partial \overrightarrow{\alpha^v}} = \mathbf{L^{22}}\left(\delta\overrightarrow{\alpha^v}\right) + \tag{7.48}$$

$$\begin{pmatrix} \left(\overrightarrow{\alpha^u}\right)^{\mathbf{T}}\mathbf{Q_1^{21}}\delta\overrightarrow{\alpha^v} \\ \vdots \\ \left(\overrightarrow{\alpha^u}\right)^{\mathbf{T}}\mathbf{Q_{n_v}^{21}}\delta\overrightarrow{\alpha^v} \end{pmatrix} + \begin{pmatrix} \left(\overrightarrow{\alpha^v}\right)^{\mathbf{T}}\left(\mathbf{Q_1^{22}} + \left(\mathbf{Q_1^{22}}\right)^{\mathbf{T}}\right)\delta\overrightarrow{\alpha^v} \\ \vdots \\ \left(\overrightarrow{\alpha^v}\right)^{\mathbf{T}}\left(\mathbf{Q_{n_v}^{22}} + \left(\mathbf{Q_{n_v}^{22}}\right)^{\mathbf{T}}\right)\delta\overrightarrow{\alpha^v} \end{pmatrix} \tag{7.49}$$

$$\frac{\partial F_2}{\partial \overrightarrow{\alpha^\phi}} = \mathbf{L^{23}}\left(\delta\overrightarrow{\alpha^\phi}\right) \tag{7.50}$$

$$\frac{\partial F_3}{\partial \overrightarrow{\alpha^u}} = \mathbf{L^{31}}\left(\delta\overrightarrow{\alpha^u}\right) + \begin{pmatrix} \left(\overrightarrow{\alpha^\phi}\right)^{\mathbf{T}}\left(\mathbf{Q_1^{31}}\right)^{\mathbf{T}}\delta\overrightarrow{\alpha^u} \\ \vdots \\ \left(\overrightarrow{\alpha^\phi}\right)^{\mathbf{T}}\left(\mathbf{Q_{n_\phi}^{31}}\right)^{\mathbf{T}}\delta\overrightarrow{\alpha^u} \end{pmatrix} \tag{7.51}$$

$$\frac{\partial F_3}{\partial \overrightarrow{\alpha^v}} = \mathbf{L^{32}}\left(\delta\overrightarrow{\alpha^v}\right) + \begin{pmatrix} \left(\overrightarrow{\alpha^\phi}\right)^{\mathbf{T}}\left(\mathbf{Q_1^{32}}\right)^{\mathbf{T}}\delta\overrightarrow{\alpha^v} \\ \vdots \\ \left(\overrightarrow{\alpha^\phi}\right)^{\mathbf{T}}\left(\mathbf{Q}_{n_\phi}^{32}\right)^{\mathbf{T}}\delta\overrightarrow{\alpha^v} \end{pmatrix} \tag{7.52}$$

$$\frac{\partial F_3}{\partial \overrightarrow{\alpha^{\phi}}} = \mathbf{L^{33}}\left(\delta\overrightarrow{\alpha^{\phi}}\right) + \begin{pmatrix} \left(\overrightarrow{\alpha^{u}}\right)^{\mathbf{T}} \mathbf{Q_1^{31}} \delta\overrightarrow{\alpha^{\phi}} \\ \vdots \\ \left(\overrightarrow{\alpha^{\mathbf{u}}}\right)^{\mathbf{T}} \mathbf{Q}_{n_\phi}^{31} \delta\overrightarrow{\alpha^{\phi}} \end{pmatrix} + \begin{pmatrix} \left(\overrightarrow{\alpha^{\mathbf{v}}}\right)^{\mathbf{T}} \mathbf{Q_1^{32}} \delta\overrightarrow{\alpha^{\phi}} \\ \vdots \\ \left(\overrightarrow{\alpha^{\mathbf{v}}}\right)^{\mathbf{T}} \mathbf{Q}_{n_\phi}^{32} \delta\overrightarrow{\alpha^{\phi}} \end{pmatrix} \tag{7.53}$$

Let us define

$$\mathcal{Q}_{11}(u) = \sum_{i=1}^{n_u} e_i \left(\overrightarrow{\alpha^{u}}\right)^{\mathbf{T}} \left(\mathbf{Q_i^{11}} + \left(\mathbf{Q_i^{11}}\right)^{\mathbf{T}}\right)$$

$$\mathcal{Q}_{12}^{u}(u) = \sum_{i=1}^{n_u} e_i \left(\overrightarrow{\alpha^{u}}\right)^{\mathbf{T}} \left(\mathbf{Q_i^{12}}\right)$$

$$\mathcal{Q}_{12}^{v}(v) = \sum_{i=1}^{n_u} e_i \left(\overrightarrow{\alpha^{v}}\right)^{\mathbf{T}} \left(\mathbf{Q_i^{12}}\right)^{\mathbf{T}}$$

$$\mathcal{Q}_{21}^{u}(u) = \sum_{i=1}^{n_v} e_i \left(\overrightarrow{\alpha^{u}}\right)^{\mathbf{T}} \left(\mathbf{Q_i^{21}}\right)$$

$$\mathcal{Q}_{21}^{v}(v) = \sum_{i=1}^{n_v} e_i \left(\overrightarrow{\alpha^{v}}\right)^{\mathbf{T}} \left(\mathbf{Q_i^{21}}\right)^{\mathbf{T}}$$

$$\mathcal{Q}_{22}(v) = \sum_{i=1}^{n_v} e_i \left(\overrightarrow{\alpha^{v}}\right)^{\mathbf{T}} \left(\mathbf{Q_i^{22}} + \left(\mathbf{Q_i^{22}}\right)^{\mathbf{T}}\right)$$

$$\mathcal{Q}_{31}^{u}(u) = \sum_{i=1}^{n_v} e_i \left(\overrightarrow{\alpha^{u}}\right)^{\mathbf{T}} \left(\mathbf{Q_i^{31}}\right)$$

$$\mathcal{Q}_{31}^{\phi}(\phi) = \sum_{i=1}^{n_\phi} e_i \left(\overrightarrow{\alpha^{\phi}}\right)^{\mathbf{T}} \left(\mathbf{Q_i^{31}}\right)^{\mathbf{T}}$$

$$\mathcal{Q}_{32}^{v}(v) = \sum_{i=1}^{n_\phi} e_i \left(\overrightarrow{\alpha^{v}}\right)^{\mathbf{T}} \left(\mathbf{Q}_i^{32}\right)^{\mathbf{T}}$$

$$\mathcal{Q}_{32}^{\phi}(\phi) = \sum_{i=1}^{n_\phi} e_i \left(\overrightarrow{\alpha^{\phi}}\right)^{\mathbf{T}} \mathbf{Q}_i^{32}$$

Therefore,

$$\mathbf{A_u}\frac{\left(\delta\overrightarrow{\alpha^u_{n+1}}-\delta\overrightarrow{\alpha^u_n}\right)}{\mathbf{dt}}=\left(L^{11}+\mathcal{Q}_{11}(u)+\mathcal{Q}^v_{12}(v)\right)\left(\delta\overrightarrow{\alpha^u_n}\right)+\left(L^{12}+\mathcal{Q}^u_{12}(u)\right)\left(\delta\overrightarrow{\alpha^v_n}\right)+L^{13}\left(\delta\overrightarrow{\alpha^\phi_n}\right) \tag{7.54}$$

$$\mathbf{A_v}\frac{\left(\delta\overrightarrow{\alpha^v_{n+1}}-\delta\overrightarrow{\alpha^v_n}\right)}{\mathbf{dt}}=\left(L^{21}+Q^v_{21}(v)\right)\left(\delta\overrightarrow{\alpha^u_n}\right)+\left(L^{22}+\mathcal{Q}^u_{21}(u)+\mathcal{Q}_{22}(v)\right)\left(\delta\overrightarrow{\alpha^v_n}\right)+L^{23}\left(\delta\overrightarrow{\alpha^\phi_n}\right) \tag{7.55}$$

$$\begin{aligned}\mathbf{A}_\phi\frac{\left(\delta\overrightarrow{\alpha^\phi_{n+1}}-\delta\overrightarrow{\alpha^\phi_n}\right)}{\mathbf{dt}} &= \left(L^{31}+Q^\phi_{31}(\phi)\right)\left(\delta\overrightarrow{\alpha^u_n}\right)+\left(L^{32}+Q^\phi_{32}(\phi)\right)\left(\delta\overrightarrow{\alpha^v_n}\right)+\\ &\quad \left(L^{33}+Q^u_{31}(u)+Q^v_{32}(v)\right)\left(\delta\overrightarrow{\alpha^\phi_n}\right)\end{aligned} \tag{7.56}$$

Let's define

$$F_{11}(u,v)=\left(L^{11}+\mathcal{Q}_{11}(u)+\mathcal{Q}^v_{12}(v)\right)$$

$$F_{12}(u)=\left(L^{12}+\mathcal{Q}^u_{12}(u)\right)$$

$$F_{21}(v)=\left(L^{21}+Q^v_{21}(v)\right)$$

$$F_{22}(u,v)=\left(L^{22}+\mathcal{Q}^u_{21}(u)+\mathcal{Q}_{22}(v)\right)$$

$$F_{31}(\phi)=\left(L^{31}+Q^\phi_{31}(\phi)\right)$$

$$F_{32}(\phi)=\left(L^{32}+Q^\phi_{32}(\phi)\right)$$

$$F_{33}(u,v)=\left(L^{33}+Q^u_{31}(u)+Q^v_{32}(v)\right)$$

Hence, we obtain

$$\mathbf{A_u}\frac{\left(\delta\overrightarrow{\alpha^u_{n+1}}-\delta\overrightarrow{\alpha^u_n}\right)}{\mathbf{dt}}=F_{11}\left(u,v\right)\left(\delta\overrightarrow{\alpha^u_n}\right)+F_{12}\left(u\right)\left(\delta\overrightarrow{\alpha^v_n}\right)+L^{13}\left(\delta\overrightarrow{\alpha^\phi_n}\right) \tag{7.57}$$

$$\mathbf{A_v}\frac{\left(\delta\overrightarrow{\alpha^v_{n+1}}-\delta\overrightarrow{\alpha^v_n}\right)}{\mathbf{dt}}=F_{21}\left(v\right)\left(\delta\overrightarrow{\alpha^u_n}\right)+F_{22}\left(u,v\right)\left(\delta\overrightarrow{\alpha^v_n}\right)+L^{23}\left(\delta\overrightarrow{\alpha^\phi_n}\right) \tag{7.58}$$

$$\mathbf{A}_\phi\frac{\left(\delta\overrightarrow{\alpha^\phi_{n+1}}-\delta\overrightarrow{\alpha^\phi_n}\right)}{\mathbf{dt}}=F_{31}\left(\phi\right)\left(\delta\overrightarrow{\alpha^u_n}\right)+F_{32}\left(\phi\right)\left(\delta\overrightarrow{\alpha^v_n}\right)+F_{33}\left(u,v\right)\left(\delta\overrightarrow{\alpha^\phi_n}\right) \tag{7.59}$$

Thus, we obtain

$$\mathbf{A_u}\delta\overrightarrow{\alpha^u_{n+1}}=\left(\mathbf{A_u}+dtF_{11}\left(u,v\right)\right)\left(\delta\overrightarrow{\alpha^u_n}\right)+dtF_{12}\left(u\right)\left(\delta\overrightarrow{\alpha^v_n}\right)+dtL^{13}\left(\delta\overrightarrow{\alpha^\phi_n}\right) \tag{7.60}$$

$$\mathbf{A_v}\delta\overrightarrow{\alpha^v_{n+1}}=dtF_{21}\left(v\right)\left(\delta\overrightarrow{\alpha^u_n}\right)+\left(\mathbf{A_v}+dtF_{22}\left(u,v\right)\right)\left(\delta\overrightarrow{\alpha^v_n}\right)+dtL^{23}\left(\delta\overrightarrow{\alpha^\phi_n}\right) \tag{7.61}$$

$$\mathbf{A}_\phi\delta\overrightarrow{\alpha^\phi_{n+1}}=dtF_{31}\left(\phi\right)\left(\delta\overrightarrow{\alpha^u_n}\right)+dtF_{32}\left(\phi\right)\left(\delta\overrightarrow{\alpha^v_n}\right)+\left(\mathbf{A}_\phi+dtF_{33}\left(u,v\right)\right)\left(\delta\overrightarrow{\alpha^\phi_n}\right) \tag{7.62}$$

We can rewrite them into matrix and we obtain the TLM as follows:

$$\begin{pmatrix} A_u\delta\overrightarrow{\alpha^u_{n+1}} \\ A_v\delta\overrightarrow{\alpha^v_{n+1}} \\ A_\phi\delta\overrightarrow{\alpha^\phi_{n+1}} \end{pmatrix}=$$

$$\begin{pmatrix} \left(\mathbf{A_u}+dtF_{11}\left(u,v\right)\right) & dtF_{12}\left(u\right) & dtL^{13} \\ dtF_{21}\left(v\right)\left(\delta\overrightarrow{\alpha^u_n}\right) & \left(\mathbf{A_v}+dtF_{22}\left(u,v\right)\right) & dtL^{23} \\ dtF_{31}\left(\phi\right) & dtF_{32}\left(\phi\right) & \left(\mathbf{A}_\phi+dtF_{33}\left(u,v\right)\right) \end{pmatrix}\begin{pmatrix} \delta\overrightarrow{\alpha^u_n} \\ \delta\overrightarrow{\alpha^v_n} \\ \delta\overrightarrow{\alpha^\phi_n} \end{pmatrix} \tag{7.63}$$

Hence the the adjoint model can be written as

$$\begin{pmatrix} A_u\left(\overrightarrow{\alpha^u_n}\right)^* \\ A_v\left(\overrightarrow{\alpha^v_n}\right)^* \\ A_\phi\left(\overrightarrow{\alpha^\phi_n}\right)^* \end{pmatrix}=$$

$$\begin{pmatrix} \left(\mathbf{A_u}+dtF_{11}\left(u,v\right)\right)^{\mathbf{T}} & \left(dtF_{21}\left(v\right)\left(\delta\overrightarrow{\alpha_n^u}\right)\right)^{\mathbf{T}} & \left(dtF_{31}\left(\phi\right)\right)^{\mathbf{T}} \\ \left(dtF_{12}\left(u\right)\right)^{\mathbf{T}} & \left(\mathbf{A_v}+dtF_{22}\left(u,v\right)\right)^{\mathbf{T}} & \left(dtF_{32}\left(\phi\right)\right)^{\mathbf{T}} \\ \left(dtL^{13}\right)^{\mathbf{T}} & \left(dtL^{23}\right)^{\mathbf{T}} & \left(\mathbf{A}_{\phi}+dtF_{33}\left(u,v\right)\right)^{\mathbf{T}} \end{pmatrix} \begin{pmatrix} \left(\overrightarrow{\alpha_{n+1}^u}\right)^* \\ \left(\overrightarrow{\alpha_{n+1}^v}\right)^* \\ \left(\overrightarrow{\alpha_{n+1}^{\phi}}\right)^* \end{pmatrix} \tag{7.64}$$

The reduced model can be written as:

$$M^{POD}:\ R^d \rightarrow R^m,\ \alpha \mapsto X = M^{POD}(\alpha) \tag{7.65}$$

Denoting the POD basis by ϕ^{POD} and the observations by D, the reduced cost functional can be expressed as:

$$J:\ R^d \rightarrow R,\ \alpha \mapsto \frac{1}{2}\left((\phi M^{POD}(\alpha)+\bar{X}-D, \phi^{POD}M^{POD}(\alpha)+\bar{X}-D\right) \tag{7.66}$$

The TLM of reduced model is denoted as $\mathbf{M^{POD}}$,then the differentiation of cost fucntional w.r.t the control variable α in the reduced space can derived as follows:

$$\delta J(\alpha) = \left(\phi^{POD}M^{POD}(\alpha)+\bar{X}-D, \phi^{POD}\mathbf{M^{POD}}\delta\alpha\right)$$

$$= \left(\left(\mathbf{M^{POD}}\right)^*\left(\phi^{POD}\right)^{\mathbf{T}}\left(\phi^{POD}M^{POD}(\alpha)+\bar{X}-D\right), \delta\alpha\right) \tag{7.67}$$

One the other hand,

$$\delta J(\alpha) = \left(\nabla J\left(\alpha\right), \delta\alpha\right)$$

Hence,

$$\nabla J\left(\alpha\right) = \left(\mathbf{M^{POD}}\right)^*\left(\phi^{POD}\right)^{\mathbf{T}}\left(\phi^{POD}M^{POD}(\alpha)+\bar{X}-D\right) \tag{7.68}$$

in which the gradient of the reduced cost functional w.r.t the control variables in the reduced space is the adjoint of the projection of forcing term from the full spacel., therefore the first order adjoint model with the forcing terms may be written as

$$\begin{pmatrix} A_u\left(\overrightarrow{\alpha_n^u}\right)^* \\ A_v\left(\overrightarrow{\alpha_n^v}\right)^* \\ A_{\phi}\left(\overrightarrow{\alpha_n^{\phi}}\right)^* \end{pmatrix} =$$

$$\begin{pmatrix} \left(\mathbf{A_u} + dtF_{11}\left(u,v\right)\right)^{\mathbf{T}} & \left(dtF_{21}\left(v\right)\left(\delta\overrightarrow{\alpha_n^u}\right)\right)^{\mathbf{T}} & \left(dtF_{31}\left(\phi\right)\right)^{\mathbf{T}} \\ \left(dtF_{12}\left(u\right)\right)^{\mathbf{T}} & \left(\mathbf{A_v} + dtF_{22}\left(u,v\right)\right)^{\mathbf{T}} & \left(dtF_{32}\left(\phi\right)\right)^{\mathbf{T}} \\ \left(dtL^{13}\right)^{\mathbf{T}} & \left(dtL^{23}\right)^{\mathbf{T}} & \left(\mathbf{A}_\phi + dtF_{33}\left(u,v\right)\right)^{\mathbf{T}} \end{pmatrix} \begin{pmatrix} \left(\overrightarrow{\alpha_{n+1}^u}\right)^* \\ \left(\overrightarrow{\alpha_{n+1}^v}\right)^* \\ \left(\overrightarrow{\alpha_{n+1}^\phi}\right)^* \end{pmatrix}$$

with the forcing terms in each time step integrating backward whenever encountered

$$\mathbf{W_u}\left(\psi^{\mathbf{u}}\right)^{\mathbf{T}}\left(\psi^{\mathbf{u}}M^{POD}(\overrightarrow{\alpha_n^u}) + \bar{u} - u^{obs}\right) \tag{7.69}$$

$$\mathbf{W_v}\left(\psi^{\mathbf{v}}\right)^{\mathbf{T}}\left(\psi^{\mathbf{v}}M^{POD}(\overrightarrow{\alpha_n^v}) + \bar{v} - v^{obs}\right) \tag{7.70}$$

$$\mathbf{W}_\phi\left(\psi^{\phi}\right)^{\mathbf{T}}\left(\psi^{\phi}M^{POD}(\overrightarrow{\alpha_n^\phi}) + \bar{\phi} - \phi^{obs}\right) \tag{7.71}$$

with final conditions

$$\left(\overrightarrow{\alpha^u}\right)^*_{t=t_f} = 0$$

$$\left(\overrightarrow{\alpha^v}\right)^*_{t=t_f} = 0$$

$$\left(\overrightarrow{\alpha^\phi}\right)^*_{t=t_f} = 0$$

By integrating the first order continuous adjoint model reversely in time, the the gradient of the reduced cost functional w.r.t the control variables in the reduced space is thus obtained as

$$\nabla J(\overrightarrow{\alpha_0}) = (\overrightarrow{\alpha})^*(0) = \begin{pmatrix} \left(\overrightarrow{\alpha^u}\right)^*(0) \\ \left(\overrightarrow{\alpha^v}\right)^*(0) \\ \left(\overrightarrow{\alpha^\phi}\right)^*(0) \end{pmatrix} \tag{7.72}$$

where $(\overrightarrow{\alpha})^* = (\left(\overrightarrow{\alpha^u}\right)^*, \left(\overrightarrow{\alpha^v}\right)^*, \left(\overrightarrow{\alpha^\phi}\right)^*)$ is the first order adjoint variable vector, W_u, W_v, W_ϕ are weighting factors which are chosen to be the inverse of estimates of the statistical root-mean-square observational errors on geopotential and wind components respectively. In our test problem, values of $W_\phi = 10^{-4}m^{-4}s^4$ and $W_u = W_v = 10^{-2}m^{-2}s^2$ are used.

7.4 Discussion of numerical results obtained by trust-region POD 4-D Var combined with dual-weighted snapshots selection

The model test problem used here adopts the following initial conditions (Figure 6.8 and Figure 6.9) from the initial height field condition No.1 of Grammeltvedt [115]:

$$h(x,y) = H_0 + H_1 \tanh\left(\frac{9(D/2-y)}{2D}\right) + H_2\left(1/\cosh^2\left(\frac{9(D/2-y)}{D}\right)\right)\sin\left(\frac{2\pi x}{L}\right) \quad (7.73)$$

where this initial condition has energy in wave number one in the x-direction.

The initial velocity fields were derived from the initial height field using the geostrophic relationship:

$$u = -\left(\frac{g}{f}\right)\frac{\partial h}{\partial y} \quad v = \left(\frac{g}{f}\right)\frac{\partial h}{\partial x} \quad (7.74)$$

The dimensional constants used here are:

$$\begin{gathered} L = 4400km,\ D = 6000km,\ \bar{f} = 10^{-4}s^{-1},\ \beta = 1.5\times 10^{-11}s^{-1}m^{-1}, \\ g = 10ms^{-1},\ H_0 = 2000m,\ H_1 = 220m,\ H_2 = 133m. \end{gathered} \quad (7.75)$$

and the space increments used here are

$$\Delta x = \Delta y = 200km, \quad \Delta t = 1800s \quad (7.76)$$

We employed linear piecewise polynomials on triangular elements in the formulation of Galerkin finite-element shallow-water equations model [116], in which the global matrix was stored into a compact matrix (see [118]). A time-extrapolated Crank-Nicholson time differencing scheme was applied for integrating in time the system of ordinary differential equations resulting from the application of the Galerkin finite-element method and the Courant-Friedrichs-Levy (CFL) criterion was $\sqrt{gH_0}\left(\frac{\Delta t}{\Delta x}\right) < \frac{\sqrt{2}}{2}$ (see [121, 122]), based on which the shallow-water equations system was then coupled at every time step so that the equations are quasi-linearized (see [117]).

In order to implement boundary conditions in the Galerkin finite-element model, we have adopted the approach suggested by Payne and Irons [119] and mentioned by Huebner [120]. This approach consists in modifying the diagonal terms of the global matrix associated with

the nodal variables by multiplying them by a large number, say 10^{16}, while the corresponding term in the right-hand vector is replaced by the specified boundary nodal variable multiplied by the same large factor times the corresponding diagonal term. This procedure is repeated until all prescribed boundary nodal variables have been treated (see [121]).

In the numerical experiment, we applied a 1% uniform random perturbations on the initial conditions in order to provide twin-experiment "observations". We also computed the errors between the retrieved initial conditions related to 5% uniform random perturbations of the true initial conditions as the initial guess of the reduced-order 4-D Var (Figure 6.10 and Figure 6.11). The data assimilation was carried on a 48 hours window using the $\Delta t = 1800$s in time and a mesh of 30×24 grid points in space, thus we generated 96 snapshots by integrating the full finite-element shallow-water equations model forward in time, from which we choose 10 POD bases for each of the $(u(x, y), v(x, y), \phi(x, y))$ to capture over 99.9% of the energy. The dimension of control variables vector for the reduced-order 4-D Var thus is $10 \times 3 = 30$.

In the process of POD 4-D Var, the resulting control variables from the latest optimization iteration are projected to the full model to generate new POD bases. The new POD bases then replace the previous ones resulting in a new POD reduced-order model. We found that both the root mean square error and correlation error metrics between the full model solutions and reduced-order solutions were improved after each outer projection was carried out.

The Polak Ribiere nonlinear conjugate gradient (CG) technique [123] was employed for high-fidelity full model 4-D VAR and all variants of ad-hoc POD 4-D Var, while the steepest-descent strategy was employed for the trust-region POD 4-D Var within the trust-region radius and provides a sufficient reduction of the high-fidelity model quantified in terms of the Cauchy point [70]. In the ad-hoc POD 4-D Var, the POD bases are re-calculated when the value of the cost function cannot be decreased by more than 10^{-1} for ad-hoc POD 4-D Var and 10^{-2} for ad-hoc DWPOD 4-D Var between the consecutive minimization iterations. In the trust-region 4-D Var, the POD bases are re-calculated when the ratio ρ_k is larger than the trust-region parameter η_1 in the process of updating the trust-region radius.

The unweighted ad-hoc POD 4-D Var as a reduced order approach required a smaller computation cost but could not achieve the same cost functional reduction as the high-fidelity model 4-D Var. The dual weighted ad-hoc POD 4-D Var (Figure 5.3 and Figure 7.1) achieves a better reduction of the cost functional. However, neither of the above-mentioned methods

can attain the minimum of the high fidelity 4-D VAR model cost functional. Furthermore, the unweighted snapshots trust-region POD 4-D Var yield an additional cost functional reduction compared to the ad-hoc approach, albeit at a higher computational cost. Finally, the dual weighted trust-region POD 4-D Var achieves almost exactly the same cost functional reduction as the full high fidelity 4-D Var model, resulting in an additional decrease of four orders of magnitude compared to the minimization of the cost functional obtained by applying the unweighted ad-hoc POD 4-D Var (see Table 7.1), showing that the combination of the dual-weighted approach and trust-region method to model reduction is significantly beneficial in the achievement of a local minimum of optimization almost identical to one obtained by the high fidelity full 4-D VAR.

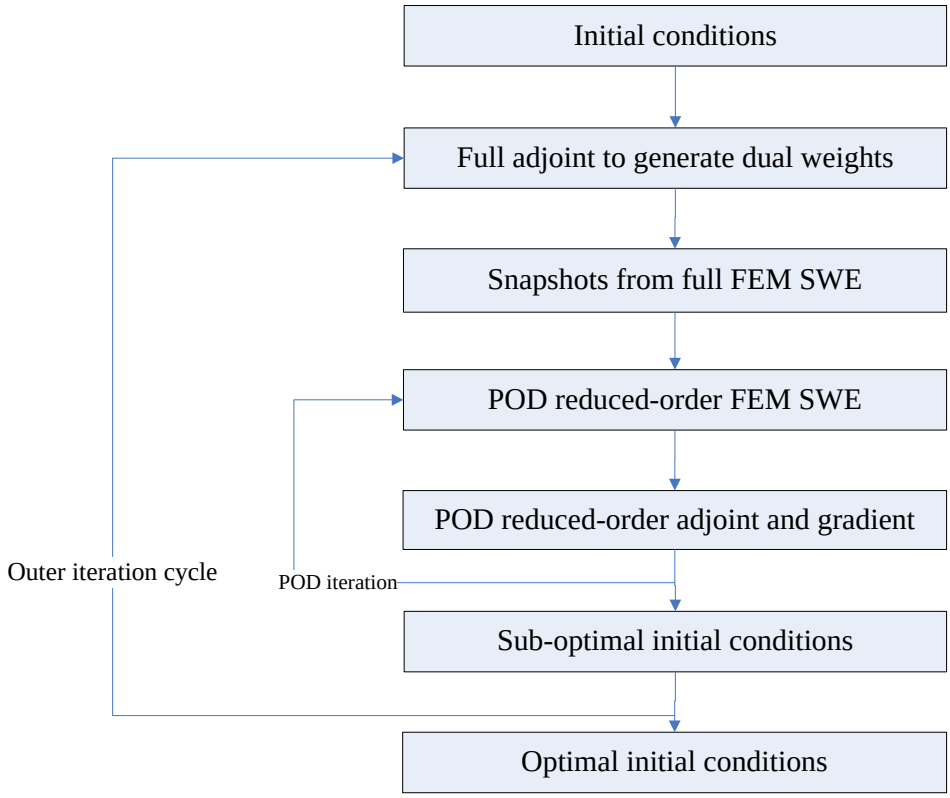


Figure 7.1: Flowchart of the methodology using adaptive POD reduced-order model for dual-weighted snapshots of the full model

Table 7.1: Comparison of iterations, outer projections, error and CPU time for ad-hoc POD 4-D Var, ad-hoc dual weighed POD 4-D Var, trust-region POD 4-D Var, trust-region dual weighed POD 4-D Var and the full model 4-D Var.

POD 4-D Var	ADPOD	DWAHPOD	TRPOD	DWTRPOD	Full
Iterations	22	42	46	57	80
Outer projections	2	6	10	12	N/A
Error	10^{-1}	10^{-2}	10^{-5}	10^{-8}	10^{-10}
CPU time (s)	15.2	38.7	121.2	142.8	222.6

In Figure 7.2, it is noticed that the dual-weighed 4-D Var absorbs the information from the full 4-D Var model and mimics the behavior of the full model 4-D Var thus being able to achieve better reduction of the cost functional. It is also noticed that in the dual weighted approach, the reduced basis is adjusted to according to the norm of the full adjoint variable. The dual weights are decreasing in time for the snapshots without sharp transients (Figure 7.3) due to the fact that observations are available in each time step in our experiments. Furthermore, the dual weights on the snapshot data are distinct from one outer projection to the next. The importance of snapshots for longer windows of assimilation may assume a preponderant importance after each outer iteration. However, it should be emphasized that the benefit obtained for POD 4-D Var using the dual-weighted procedure diminishes as the dimension of the reduced space increases.

Once the retrieved initial condition is obtained by implementing the dual weighted trust-region 4-D Var, we can compare the results from the POD reduced-model with those from the full model. To quantify the performance the dual weighted trust-region 4-D Var, we use two metrics namely the root mean square error (RMSE) and correlation of the difference between the POD reduced-order simulation and high-fidelity model.

In particular, the RMSE (Figure 7.4a and Figure 7.4b) between variants of the POD reduced-model solution and the true one at the time level i is used to estimate the error of the POD model.

$$\mathrm{RMSE}^{\mathrm{i}} = \sqrt{\frac{\sum_{j=1}^{j=N} \left(U_{i,j} - U_{i,j}^{POD}\right)^2}{n}}, \quad i = 1, \ldots, n \tag{7.77}$$

where $U_{i,j}$ and $U_{i,j}^{POD}$ are the state variables obtained by the full model and ones obtained by optimal POD reduced-order model of time level i at node j, respectively, and N is the total

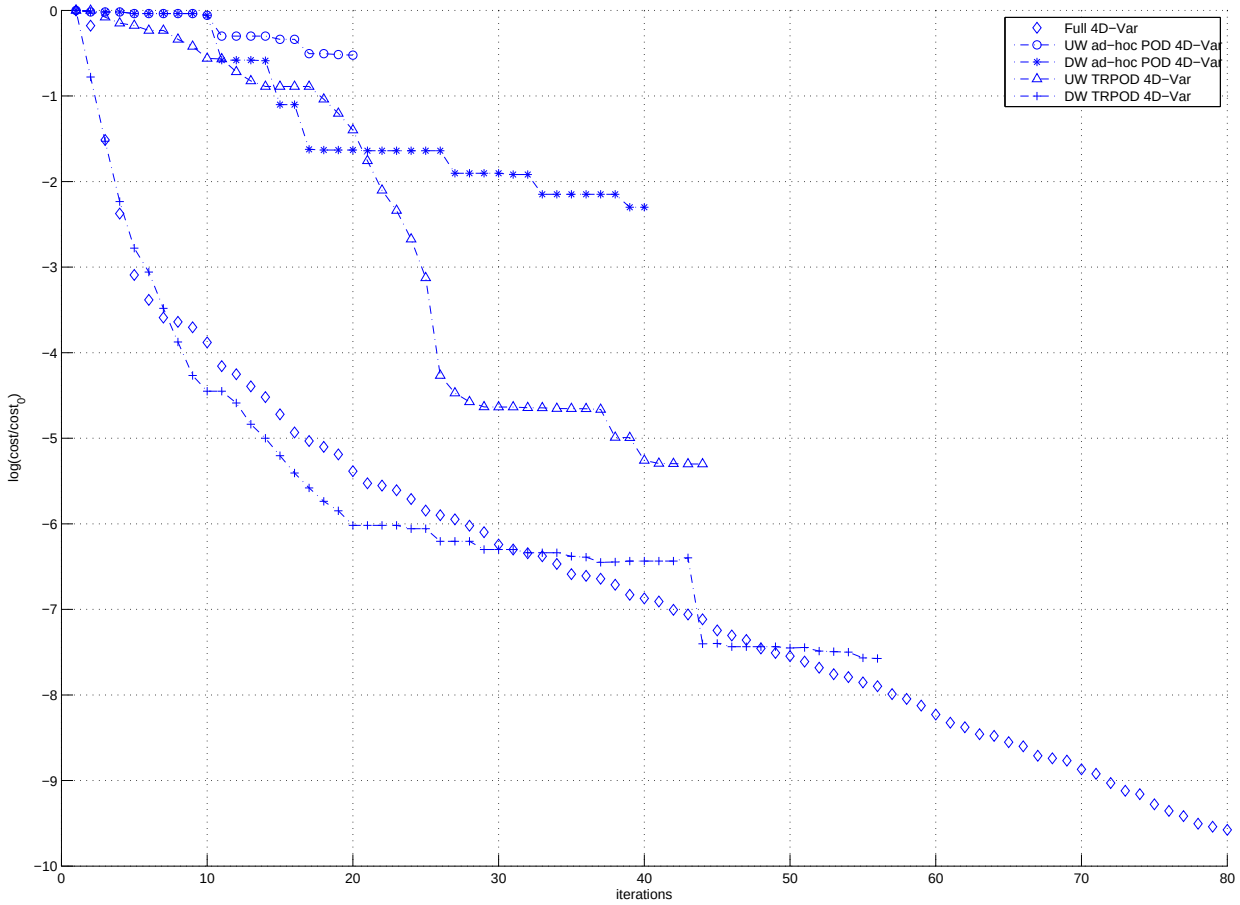


Figure 7.2: Comparison of the performance of minimization of cost functional in terms of number of iterations for ad-hoc POD 4-D Var, ad-hoc dual weighed POD 4-D Var, trust-region POD 4-D Var, trust-region dual weighed POD 4-D Var and the full model 4-D Var.

number of nodes over the domain. U and U^{POD} are used to either denote the geopotential or the velocity of the full model and the POD reduced-order model, respectively.

In (Figure 7.5a and Figure 7.5b), the correlation r defined below is used as an additional metric to evaluate quality of the inversion simulation

$$r_i = \frac{cov_{12}^i}{\sigma_1^i \sigma_2^i} \tag{7.78}$$

where

$$\sigma_1^i = \sum_{j=1}^{j=N} \left(U_{i,j} - \overline{U}_j\right)^2, \quad \sigma_2 = \sum_{j=1}^{j=N} \left(U_{i,j}^{POD} - \overline{U^{POD}}_j\right)^2,, \quad i = 1, \ldots, n \tag{7.79}$$

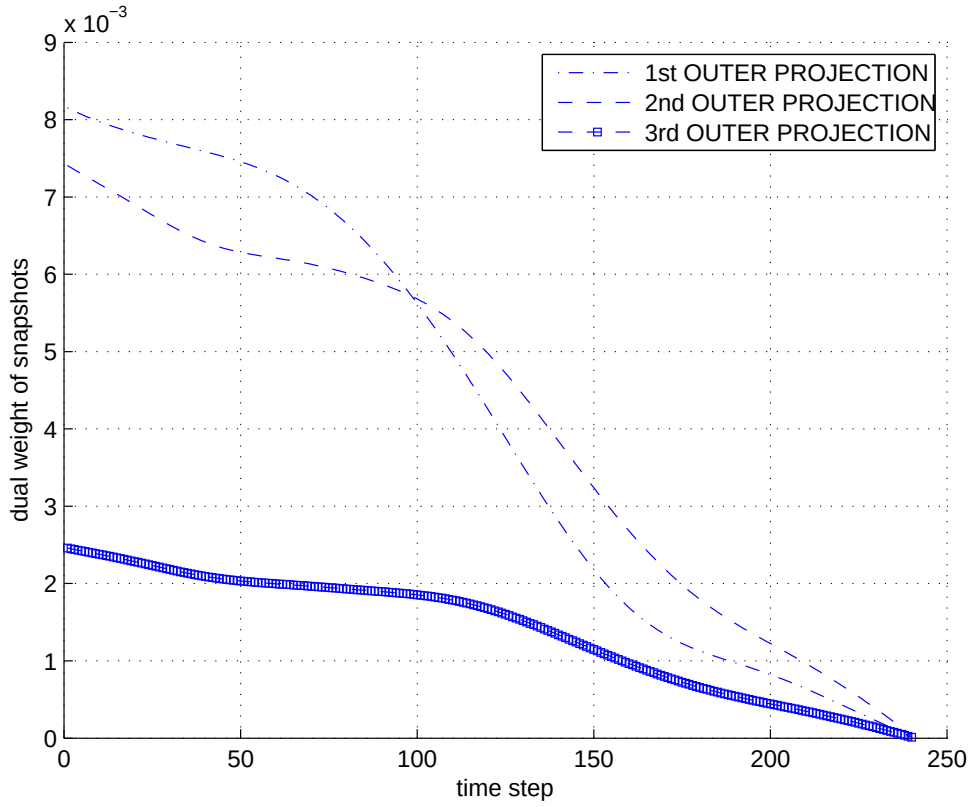


Figure 7.3: The dual weights of the snapshots data determined by the full adjoint variable for the trust-region POD 4-D Var

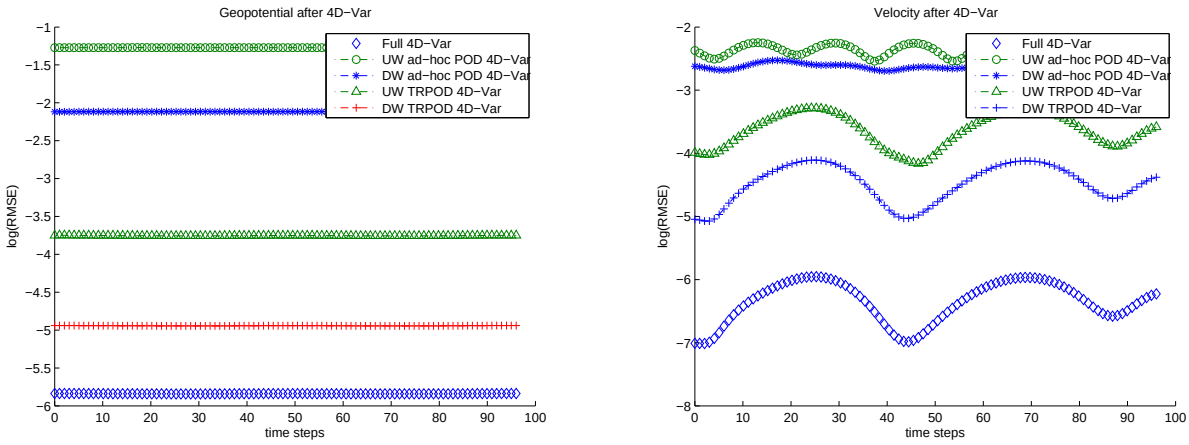
4

$$cov_{12} = \sum_{j=1}^{j=N} \left(U_{i,j} - \overline{U}_j\right) \left(U_{i,j}^{POD} - \overline{U^{POD}}_j\right), , \quad i = 1, \ldots, n \tag{7.80}$$

where $\overline{U}_j$ and $\overline{U^{POD}}_j$ are the means over the simulation period $[0, T]$ obtained by the full model and ones obtained by optimal POD reduced-order model at node j, respectively.

Even though it turned out to be advantageous to combine the dual-weighed approach with the trust-region POD 4-D Var, it should be emphasized that this advantage diminishes when we increase the number of POD bases for each component of the $(u(x, y), v(x, y), \phi(x, y))$ from 10 to 20 by applying both metrics mentioned above. However, increasing the dimension of the POD reduced-order space from 30 to 60 can increase the computational cost of POD reduced-order 4-D Var .This agrees with results obtained in [102] that for practical applications, the

dual-weighted procedure may be of particular benefit for use only with small dimensional bases in the context of adaptive order reduction as the minimization approaches the optimal solution. For other beneficial effects of POD 4-D Var related to its use in the framework of second order adjoint of a global shallow water equations model see Daescu and Navon (2007) [101]

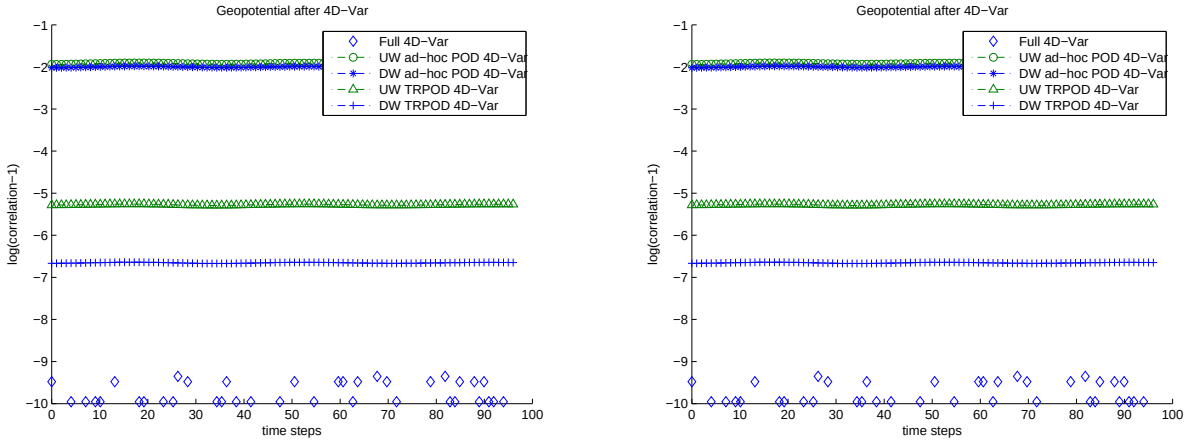


(a) Geopotential (b) Wind field

Figure 7.4: Comparison of the RMSE of between ad-hoc POD 4-D Var, ad-hoc dual weighed POD 4-D Var, trust-region POD 4-D Var, trust-region dual weighed POD 4-D Var and the full model 4-D Var.

In this Chapter, we solved an inverse problem for the POD reduced-order shallow-water equations model using a finite-element formulation, controlling its initial conditions in presence of observations being assimilated in a time window. In this POD 4-D Var, we developed the full adjoint of the finite-element shallow-water equations model and the reduced-order adjoint for POD reduced-order model. We integrated the full adjoint model backward in time to compute the time-varying sensitivities of the full 4-D Var cost functional with respect to time-varying model states, from which we derived the dual weights of the ensemble of snapshots. Also, we integrated the reduced-order adjoint model backward in time to compute gradient of reduced-order cost functional.

In the numerical experiments, we compared several variants of POD 4-D Var, namely unweighted ad-hoc POD 4-D Var, dual-weighed ad-hoc POD 4-D Var, unweighted trust-region POD 4-D Var and dual-weighed trust-region POD 4-D Var, respectively. We found



(a) Geopotential

(b) Wind field

Figure 7.5: Comparison of correlation between ad-hoc POD 4-D Var, ad-hoc dual weighed POD 4-D Var, trust-region POD 4-D Var, trust-region dual weighed POD 4-D Var and the full model 4-D Var.

that the ad-hoc POD 4-D Var version yielded the least reduction of the cost functional compared with the trust-region 4-D VAR . We assume that this result may be attributed to lack of feedbacks from the high-fidelity model . On the other hand, the trust-region POD 4-D Var version yielded a sizably better reduction of the cost functional, due to inherent properties of TRPOD allowing local minimizer of the full problem to be attained by minimizing the TRPOD sub-problem. Thus trust-region 4-D Var resulted in global convergence to the high fidelity local minimum starting from any initial iterates.

The dual-weighted proper orthogonal decomposition selection of snapshots allows propagation of information from the data assimilation system onto the reduced order model, possibly capturing lower energy modes that may play significant role in successful implementation of 4-D Var data assimilation. Combining the dual-weighted approach with the trust-region POD approach to model reduction results in a significant enhanced benefit achieving a local minimum of reduced cost function optimization almost identical to the one obtained by the high fidelity full 4-D VAR model. Hence we achieve a double benefit while running a reduced-order inversion at an acceptable computational cost, at least for the shallow-water equations model in a two-dimensional spatial domain.

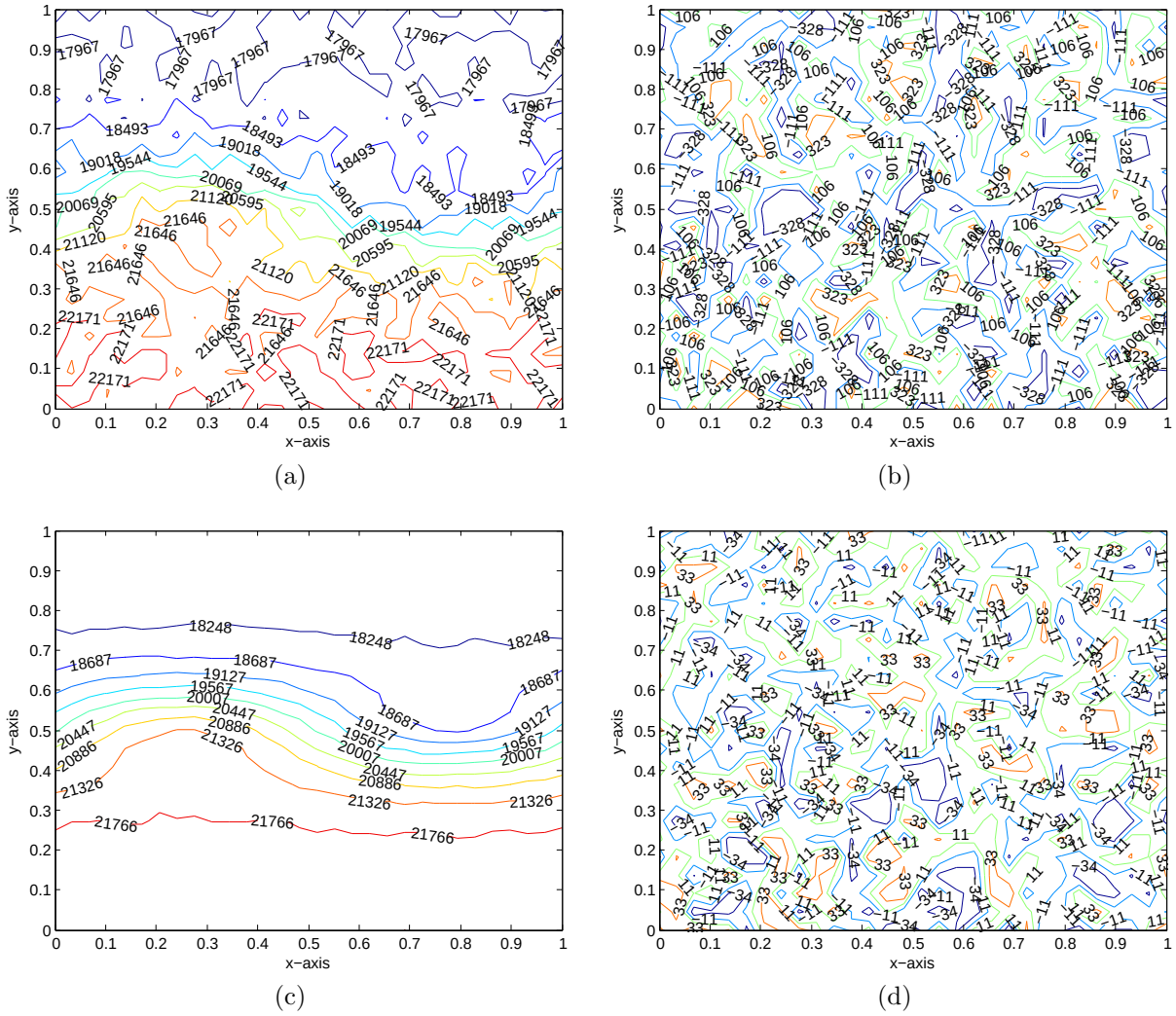


Figure 7.6: Errors between the retrieved initial geopotential and true initial geopotential applying dual weighted trust-region POD 4-D Var to the 5% uniform random perturbations of the true initial conditions taken as initial guess. (a) shows the contour of 5% perturbation of true initial geopotential; (b) shows the contour of difference between 5% perturbation of true initial geopotential; (c) shows the contour of retrieved initial geopotential after 2days with dt = 1800s; (d) shows the contour of difference between retrieved initial geopotential and true initial geopotentials.

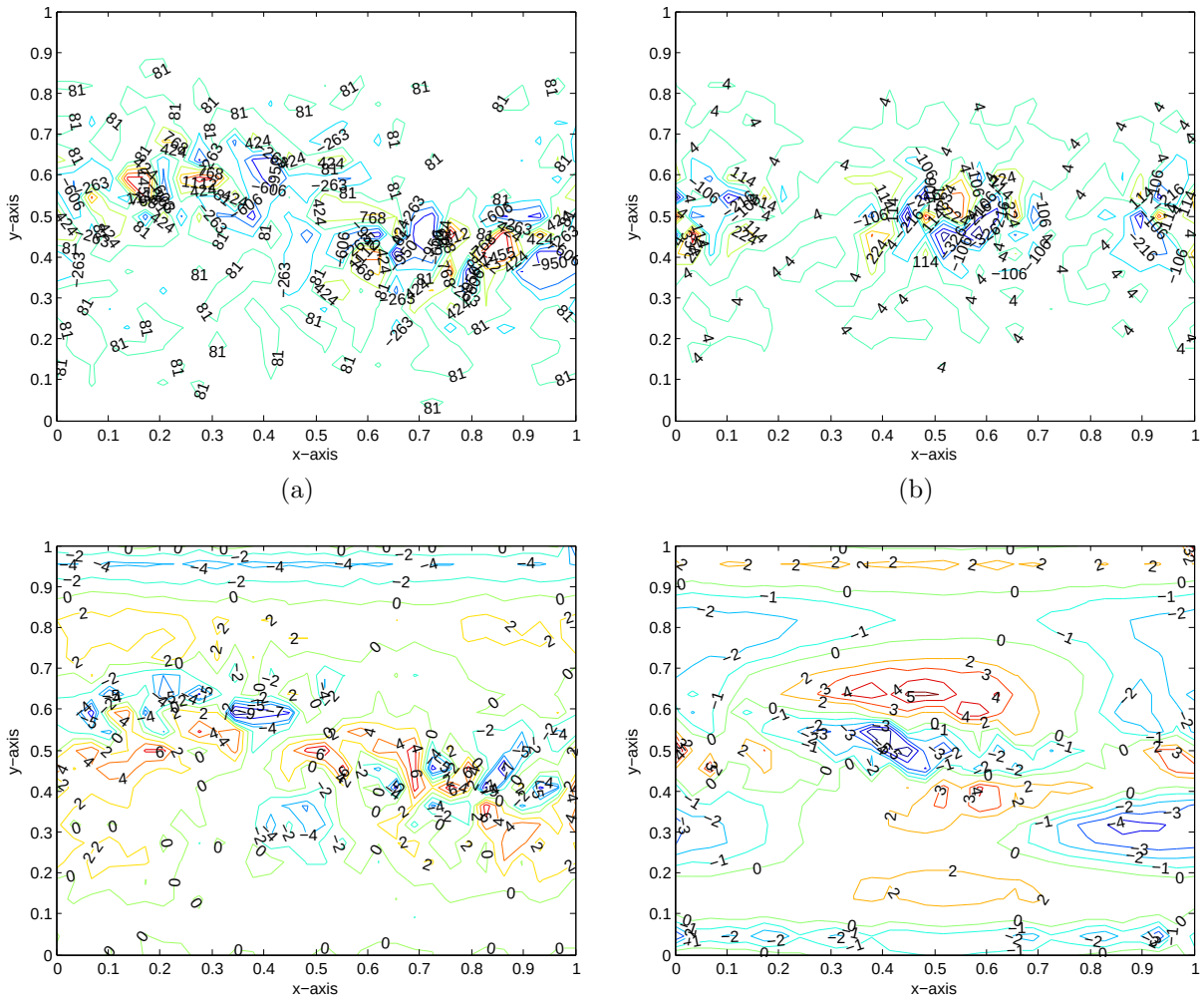
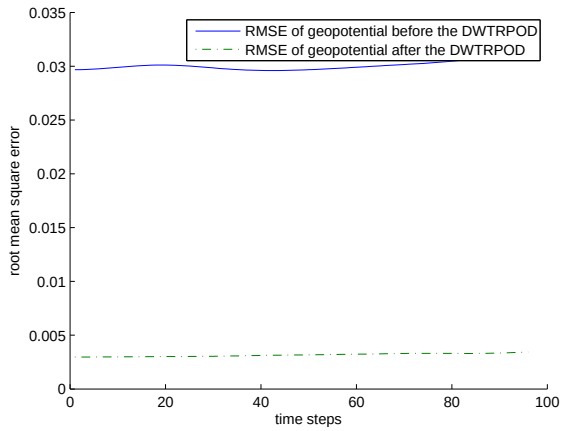
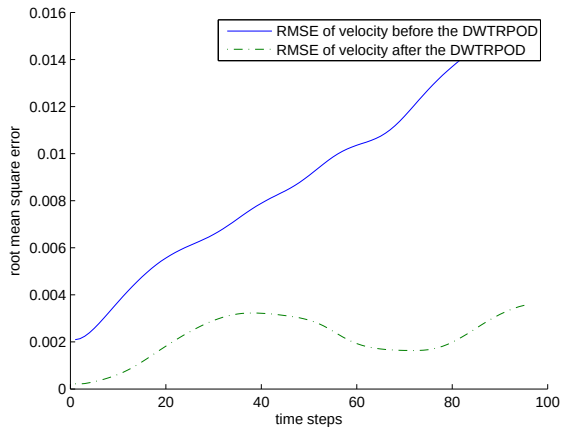


Figure 7.7: Errors scaled by 100 between the retrieved initial wind field and true initial wind field applying dual weighted trust-region POD 4-D Var to the 5% uniform random perturbations of the true initial conditions taken as the initial guess. (a) shows the contour of difference between true initial u-velocity and perturbed initial u-velocity; (b) shows the contour of difference between true initial v-velocity and perturbed initial v-velocity; (c) shows the contour of difference between retrieved initial u-velocity and true initial u-velocity; (d) shows the contour of difference between retrieved initial v-velocity and true initial v-velocity.

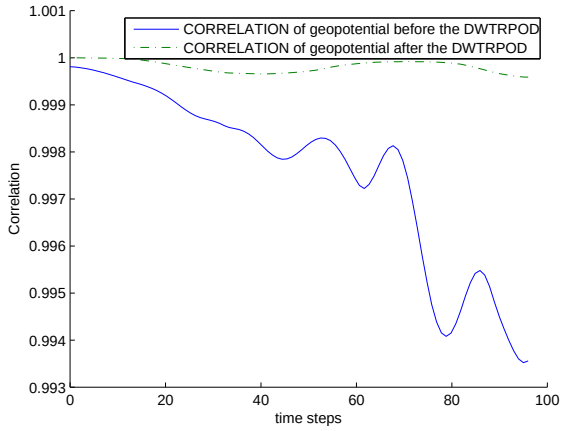


(a) Geopotential

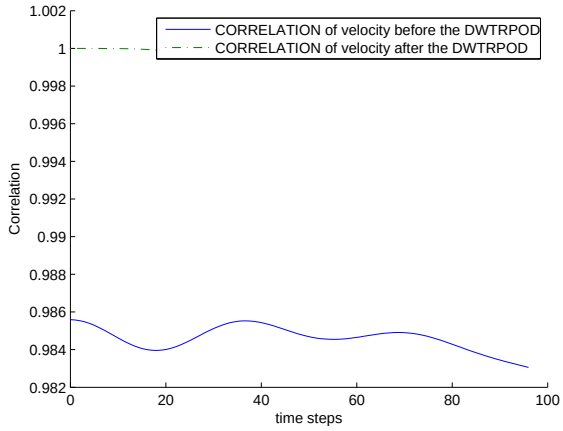


(b) Wind field

Figure 7.8: Comparison of the RMSE between the full model and the ROM before and after the data assimilation applying dual weighted trust-region POD 4-D Var to the 5% uniform random perturbations of the true initial conditions taken as the initial guess.



(a) Geopotential



(b) Wind field

Figure 7.9: Comparison of the correlation between the full model and the ROM before and after data assimilation applying dual weighted trust-region POD 4-D Var to the 5% uniform random perturbations of the true initial conditions serving as initial guess.

CHAPTER 8

GENERALIZATION TO A REAL-LIFE MODEL IN TWO SPACE DIMENSIONS PLUS TIME

In this chapter, we address a POD model reduction along with inverse solution of a two-dimensional global shallow water equations model. Solutions of SWE([86, 87, 88]) exhibit some of the important properties of large scale atmospheric flow and the equations have certain important features (such as horizontal dynamical aspects) in common with more complicated NWP models. Therefore, derivation and testing of various algorithms for solving SWE has often been a first step towards developing new atmosphere and ocean general circulation models. The explicit flux-form semi-Lagrangian finite volume scheme has been used to solve the SWE henceforth referred to as FV-SWE[89, 90, 91, 92, 96] in the forward model integration. This is used for NASA GEOS-5 and also at NCAR in the dynamics core

Our intention here is to generalize the efficient state-of-the-art POD implementation from our previous work on finite element SWE on the limited area [97, 98](FE-SWE) to global FV-SWE model with realistic initial conditions, i.e., combining efficient snapshot selection in the presence of data assimilation system by merging dual weighting of snapshots with trust region POD techniques [99].

In previous chapter (Chen et al.[98]), we studied the effect of combining TRPOD in conjunction with dual weighting Data Assimilation System (DAS) snapshot selection in the framework of Galerkin-projection based POD-ROM for FE-SWE on the limited area without a background error covariance term, in which the observations were available at all the time steps and distributed at all the grid points during the entire window of assimilation. As in the previous paper, one of the goals of this article is to confirm that dual weighted TRPOD 4-D Var can also be applied to the global FV-SWE model even when a Galerkin projection scheme is unavailable from full space to POD reduced-order space in the case of

complete observations distributed in space and time and with an unbalanced background error covariance term being provided as well. Furthermore, we study the performance of TRPOD 4-D Var in the case of incomplete observations in both time and space with or without a balanced background error covariance term being included in the cost functional. In the framework of TRPOD 4-D Var with background error covariance term inclusion, an ideal preconditioning of the POD 4-D Var is derived so that the Hessian matrix of the POD reduced-order background error covariance matrix becomes the identity matrix. In this paper we show that TRPOD 4-D Var performs satisfactorily in presence of incomplete observations, just as in the case of full 4-D Var, if a geostrophically balanced background error covariance matrix is available during the implementation of data assimilation.

In the numerical experiments, we compare the ad-hoc update adaptivity of the POD 4-D VAR, the trust-region update adaptivity with or without dual weighting and full 4-D Var (high fidelity model) in the case of full observations with or without the background error covariance term being included in the cost functional. We confirm that the combination of TRPOD and dual-weighted snapshots yields the best results in all error metrics (see [98, 100, 101, 102]). The advantage of TRPOD adaptivity over ad-hoc POD adaptivity is due to the fact that TRPOD can appropriately determine a trust region within which the step stays and step size is not too small, so that it is guaranteed to compute a sufficient decrease for the cost functional of the full model by projecting the Quasi-Newton direction of POD reduced order cost functional into the trust region box as a substitute for the Cauchy point in the standard trust region methods using quadratic approximation. Hence, TRPOD by comparing the actual reduction and predicted reduction[83, 85] can successively refresh the POD basis, based on updated control values, allowing it to keep the full 4-D Var and the POD reduced-order 4-D synchronized so as to ensure that the POD reduced order 4-D Var takes updated information from the full model to evolve the POD reduced model to the local optimizer of the high fidelity model.

Also, we notice that there are almost twice as many outer projections (refreshing the snapshots) related to TRPOD adaptivity compared to the number of projections in ad-hoc adaptivity in the framework of Galerkin projection in our previous work ([98]). In this work there are almost thrice as many outer projections related to TRPOD adaptivity as compared to ad-hoc adaptivity without a Galerkin projection scheme in presence of incomplete observations in time and space. The CPU time required by TRPOD 4-D Var is

still a fraction of the CPU required by the full 4-D Var, due to the fact that most of the functional evaluations are carried out in the lower dimensional POD 4-D Var while the full model will be evaluated only when a appropriate descent direction in the TRPOD reduced-order space is obtained. Therefore, TRPOD avoids unnecessary full model evaluations and also reduces the cost of minimization inside the inner TRPOD loop.

8.1 Global finite-volume shallow-water equations model

In spherical coordinates the vorticity divergence form of the SWE can be written as the mass conservation law for a shallow layer of water

$$\frac{\partial h}{\partial t} + \nabla \cdot (\mathbf{V}h) = 0 \tag{8.1}$$

and the vector invariant form of momentum equations

$$\frac{\partial u}{\partial t} = \Omega v - \frac{1}{A\cos\theta}\frac{\partial\left(\kappa + \Phi\right)}{\partial \lambda} \tag{8.2}$$

$$\frac{\partial v}{\partial t} = -\Omega u - \frac{1}{A}\frac{\partial\left(\kappa + \Phi\right)}{\partial \theta} \tag{8.3}$$

$$(\lambda, \theta) \in [-\frac{\pi}{2}, \frac{\pi}{2}] \times [-\pi, \pi], \;\; t \geqslant 0$$

where h represents the fluid height (above the surface height h_s), $\mathbf{V} = (u, v)$, u and v represent the zonal and meridional wind velocity components respectively, θ and λ are the latitudinal and longitudinal directions respectively, ω is the angular speed of rotation of the earth, a is radius of the earth. The free surface geopotential is given by $\Phi = \Phi_s + gh$, where $\phi_s = gh_s$, $\kappa = \frac{1}{2}\mathbf{V} \cdot \mathbf{V}$ is the kinetic energy, and $\Omega = 2\omega sin\theta + \nabla \times \mathbf{V}$ is the absolute vorticity.

In this paper we have used a descretized (finite volume, semi-Lagrangian) version of the above SWE model, which serves as the dynamical core in the community atmosphere model (CAM), version 3.0, and its operational version implemented at NCAR and NASA is known as finite volume-general circulation model (FV-GCM). In brief, a two grid combination based on C-grid and D-grids is used for advancing from time step t_n to $t_n + \Delta t$. In the first half

of the time step, advective winds (time centered winds on the C-grid: (u^*, v^*)) are updated on the C-grid, and in the other half of the time step, the prognostic variables (h, u, v) are updated on the D-grid.

Using the finite volume method, within each cell of the discrete grid, if we consider a piecewise linear approximation to the solution, whose slope is *limited* in a certain way depending on the values of the solution at the neighboring grid cells, one can consistently derive a family of van Leer schemes. We will follow the suggestion in [92] and always use unconstrained van Leer[93, 94, 95] scheme to advect winds on the C-grid. The same advection scheme will be used on D-grid as well. This strategy provides solutions whose accuracy is comparable to those obtained by using more CPU demanding advection schemes, for e.g., constrained van Leer schemes.

8.2 Generation of dual weighted POD reduced model applied to FV-SWE

An ensemble of snapshots is chosen in the analysis time interval $[0, T]$ written as $\{y^1, y^2, \ldots, y^n\}$ where $y^i = (h^i, u^i, v^i)^{\mathbf{T}} \in \mathbb{R}^N$, $i = 1, \ldots, n$, n is the number of snapshots and $N = 3N_xN_y$ is triple the dimension of discrete mesh, N_x and N_y are the mesh points of the latitudinal and longitudinal directions respectively. Our choice of snapshots number was to take a snapshot at every time step ($\Delta t = 450$ sec) of the window of assimilation whose length was taken in our case to be 15 hours. We could have chosen another snapshot distribution, however, we elected to implement this choice as the most intuitive one (15 hours = 120 time steps of 450 sec, each). Define the dual weighted ensemble average of the snapshots as $\bar{y} = \sum_{i=1}^{i=n} w_i y^i$ where the snapshots weights w_i are such that $0 < w_i < 1$ and $\sum_{i=1}^{n} w_i = 1$, and they are used to assign a degree of importance to each member of the ensemble. Time weighting is usually considered, and in the standard approach $w_i = \frac{1}{n}$. Subtracting the mean from each snapshot, we obtain the following $N \times n$ dimensional matrix $\mathbf{Y} = [y^1 - \bar{y}, y^2 - \bar{y}, \ldots, y^n - \bar{y}]$.

The POD modes $\Psi = \left\{\psi^1, \psi^2, \ldots, \psi^M\right\}$ of order $M \leq n$ provide an optimal representation of the ensemble data in a $M-$dimensional state subspace by minimizing the averaged projection error

$$\min_{\{\psi^1, \psi^2, \ldots, \psi^M\}} \sum_{i=1}^{n} w_i \left\| \left(y^i - \bar{y}\right) - \Pi_{\Psi, M} \left(y^i - \bar{y}\right) \right\|^2$$

$$\text{s.t. } \left\langle \psi^i, \psi^j \right\rangle_{l_2} = \delta_{ij} \tag{8.4}$$

where $\Pi_{\Psi,M}$ is the projection operator onto the $M-$dimensional space $\text{Span}\left\{\psi^1, \psi^2, \ldots, \psi^M\right\}$

$$\Pi_{\Psi,M} = \sum_{i=1}^{M} \langle y, \psi_i \rangle_{l_2} \psi_i$$

We define the dual weighted spatial correlation matrix, $\mathbf{A} = \mathbf{YWY^T}$, where $\mathbf{W} = \text{diag}\left\{w_1, w_2, \ldots, w_n\right\}$ is the diagonal matrix of weights.

To compute the dual weighted POD modes $\psi^i \in \mathbb{R}^N$, one must solve an $N-$dimensional eigenvalue problem, $\mathbf{A}\psi_i = \lambda_i \psi_i$.

In practice the number of snapshots is much less than the the state dimension, $n << N$, an efficient way to compute the reduced basis is to introduce an $n-$dimensional matrix as follows:

$$\mathbf{K}^{n\times n} = \mathbf{W}^{\frac{1}{2}}\mathbf{Y^T}\mathbf{Y}\mathbf{W}^{\frac{1}{2}} \tag{8.5}$$

and compute the eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0$ of $\mathbf{K}^{n\times n}$ with its corresponding eigenvectors $\xi_1, \ldots, \xi_n$

Hence, the corresponding POD modes are thus obtained by defining

$$\psi_i = \frac{1}{\sqrt{\lambda_i}}\mathbf{Y}\mathbf{W}^{\frac{1}{2}}\xi_i, \quad i = 1, \ldots, M \tag{8.6}$$

where

$$\left\langle \psi^i, \psi^j \right\rangle_{l_2} = \delta_{ij} = \begin{cases} 1 & i = j \\ \\ 0 & i \neq j \end{cases} \tag{8.7}$$

Define the following vectors

$$\begin{aligned} \mathbf{h}^k &= \begin{pmatrix} h_1^k & h_2^k & \cdots & h_N^k \end{pmatrix}^{\mathbf{T}} \ \mathbf{u}^k = \begin{pmatrix} u_1^k & u_2^k & \cdots & u_N^k \end{pmatrix}^{\mathbf{T}} \ \mathbf{v}^k = \begin{pmatrix} v_1^k & v_2^k & \cdots & v_N^k \end{pmatrix}^{\mathbf{T}} \\ \mathbf{h}^* &= \begin{pmatrix} h_1^* & h_2^* & \cdots & h_N^* \end{pmatrix}^{\mathbf{T}} \ \mathbf{u}^* = \begin{pmatrix} u_1^* & u_2^* & \cdots & u_N^* \end{pmatrix}^{\mathbf{T}} \ \mathbf{v}^* = \begin{pmatrix} v_1^* & v_2^* & \cdots & v_N^* \end{pmatrix}^{\mathbf{T}} \end{aligned}$$

thus $\mathbf{h}^*$, $\mathbf{u}^*$and $\mathbf{v}^*$ are obtained on C-grid [91, 92], in the following way,

$$\begin{aligned} \mathbf{h}^* =& \mathbf{h}^k + \mathbf{F}_h^c\left(\mathbf{h}^k, \mathbf{u}^k, \mathbf{v}^k\right) \\ \mathbf{u}^* =& \mathbf{u}^k + \frac{\Delta t}{2}\mathbf{F}_u^c\left(\mathbf{h}^k, \mathbf{u}^k, \mathbf{v}^k\right) \\ \mathbf{v}^* =& \mathbf{v}^k + \frac{\Delta t}{2}\mathbf{F}_v^c\left(\mathbf{h}^k, \mathbf{u}^k, \mathbf{v}^k\right) \end{aligned} \tag{8.8}$$

Define the following vectors

$$\Psi \;=\; \left(\begin{array}{ccc} \Psi_h & \Psi_u & \Psi_v \end{array}\right)^{\mathbf{T}},\; \bar{y} = \left(\begin{array}{ccc} \overline{\mathbf{h}} & \overline{\mathbf{u}} & \overline{\mathbf{v}} \end{array}\right)^{\mathbf{T}} \tag{8.9}$$

and we obtain the POD reduced-order model on C-grid by projection as follows, where the α coefficients are the modal coefficients of the flow field with respect to the POD basis;

$$\begin{aligned} \alpha_h^* =& \alpha_h^k + \Psi_h^{\mathbf{T}} \left(\mathbf{F}_h^c \left(\Psi_h \alpha_h^k + \overline{\mathbf{h}}, \Psi_u \alpha_u^k + \overline{\mathbf{u}}, \Psi_v \alpha_v^k + \overline{\mathbf{v}}\right) - \overline{\mathbf{h}}\right) \\ \alpha_u^* =& \alpha_u^k + \frac{\Delta t}{2} \Psi_u^{\mathbf{T}} \left(\mathbf{F}_u^c \left(\Psi_h \alpha_h^k + \overline{\mathbf{h}}, \Psi_u \alpha_u^k + \overline{\mathbf{u}}, \Psi_v \alpha_v^k + \overline{\mathbf{v}}\right) - \overline{\mathbf{u}}\right) \\ \alpha_v^* =& \alpha_v^k + \frac{\Delta t}{2} \Psi_v^{\mathbf{T}} \left(\mathbf{F}_v^c \left(\Psi_h \alpha_h^k + \overline{\mathbf{h}}, \Psi_u \alpha_u^k + \overline{\mathbf{u}}, \Psi_v \alpha_v^k + \overline{\mathbf{v}}\right) - \overline{\mathbf{v}}\right). \end{aligned} \tag{8.10}$$

Similarly, we can rewrite the D-grid [91, 92] time integration as the following vector formulation,

$$\begin{aligned} \mathbf{h}^{k+1} =& \mathbf{h}^k + \mathbf{F}_h^d \left(\mathbf{h}^*, \mathbf{u}^*, \mathbf{v}^*\right) \\ \mathbf{u}^{k+1} =& \mathbf{u}^k + \frac{\Delta t}{2} \mathbf{F}_u^d \left(\mathbf{h}^*, \mathbf{u}^*, \mathbf{v}^*\right) \\ \mathbf{v}^{k+1} =& \mathbf{v}^k + \frac{\Delta t}{2} \mathbf{F}_v^d \left(\mathbf{h}^*, \mathbf{u}^*, \mathbf{v}^*\right) \end{aligned} \tag{8.11}$$

and the POD reduced-order model on D-grid by projection as below,

$$\begin{aligned} \alpha_h^{k+1} =& \alpha_h^k + \Psi_h^{\mathbf{T}} \left(\mathbf{F}_h^d \left(\Psi_h \alpha_h^* + \overline{\mathbf{h}}, \Psi_u \alpha_u^* + \overline{\mathbf{u}}, \Psi_v \alpha_v^* + \overline{\mathbf{v}}\right) - \overline{\mathbf{h}}\right) \\ \alpha_u^{k+1} =& \alpha_u^k + \frac{\Delta t}{2} \Psi_u^{\mathbf{T}} \left(\mathbf{F}_u^d \left(\Psi_h \alpha_h^* + \overline{\mathbf{h}}, \Psi_u \alpha_u^* + \overline{\mathbf{u}}, \Psi_v \alpha_v^* + \overline{\mathbf{v}}\right) - \overline{\mathbf{u}}\right) \\ \alpha_v^{k+1} =& \alpha_v^k + \frac{\Delta t}{2} \Psi_v^{\mathbf{T}} \left(\mathbf{F}_v^d \left(\Psi_h \alpha_h^* + \overline{\mathbf{h}}, \Psi_u \alpha_u^* + \overline{\mathbf{u}}, \Psi_v \alpha_v^* + \overline{\mathbf{v}}\right) - \overline{\mathbf{v}}\right), \end{aligned} \tag{8.12}$$

where $\alpha_h^k \in R^{M_h}$, $\alpha_u^k \in R^{M_u}$ and $\alpha_v^k \in R^{M_v}$, $k = 0, 1, 2, \cdots, n$ and initial values are

$$\alpha_h^0 \;=\; \Psi_h^{\mathbf{T}} \left(\mathbf{h}^0 - \overline{\mathbf{h}}\right)\;\; \alpha_u^0 = \Psi_u^{\mathbf{T}} \left(\mathbf{u}^0 - \overline{\mathbf{u}}\right)\;\; \alpha_v^0 = \Psi_v^{\mathbf{T}} \left(\mathbf{v}^0 - \overline{\mathbf{v}}\right) \tag{8.13}$$

Formulas (8.10) and (8.12) are the POD reduced-order model for the FV-SWE model (8.1) (8.2) and (8.3), and it only includes $(M_h + M_u + M_v) \times n$ degrees of freedom, where M_h, M_u, $M_v << N$ compared to the numerical FV-SWE model which contains $3N \times n$ degrees of freedom (see Akella and Navon (2006) [124]).

8.3 Preconditioning of the POD 4-D Var applied to FV-SWE

The 4-D Var cost functional

$$J(y_0) = \underbrace{\frac{1}{2}\left(y_0 - y^b\right)^{\mathbf{T}} \mathbf{B}^{-1}\left(y_0 - y^b\right)}_{J_b} + \underbrace{\frac{1}{2}\sum_{k=0}^{k=n}\left(\mathbf{H}_k y_k - y_k^o\right)^{\mathbf{T}} \mathbf{R}_k^{-1}\left(\mathbf{H}_k y_k - y_k^o\right)}_{J_o} \tag{8.14}$$

can be separated into $J = J_b + J_o$, where $J_b = \frac{1}{2}\left(y_0 - y^b\right)^{\mathbf{T}} \mathbf{B}^{-1}\left(y_0 - y^b\right)$ is the background cost functional and

$$J_o = \frac{1}{2}\sum_{k=0}^{k=n}\left(\mathbf{H}_k y_k - y_k^o\right)^{\mathbf{T}} \mathbf{R}_k^{-1}\left(\mathbf{H}_k y_k - y_k^o\right) \tag{8.15}$$

is the observational cost functional. Let $\delta y = y_0 - y^b$, so that the background cost functional can be rewritten as, $J_b = \frac{1}{2}(\delta y)^{\mathbf{T}} \mathbf{B}^{-1}(\delta y)$.

Define an approximation to the control variable, $y_0 \approx \Psi\alpha_0 + \bar{y}$, where the POD modes are given by $\Psi = \left\{\psi^1, \psi^2, \ldots, \psi^M\right\}$ and the dual weighted ensemble average of the snapshots is given as before, in which α_0 is the corresponding control variable in the $M-$dimensional POD reduced-order space. Define the coefficient, $\alpha_b = \Psi^{\mathbf{T}}(y_b - \bar{y})$, and we obtain the background term y_b in terms of POD modes, $y_b = \Psi\alpha_b + \bar{y}$. From the above equations we obtain, $\delta y = y_0 - y^b \approx (\Psi\alpha_0 + \bar{y}) - (\Psi\alpha_b + \bar{y}) = \Psi(\alpha_0 - \alpha_b)$. Let $\delta\alpha = \alpha_0 - \alpha_b$ so that $\delta y = \Psi\delta\alpha$.

Hence, the 4-D Var cost functional in (8.14) can be approximated by

$$J(y_0) \approx \hat{J}(\delta\alpha) = \hat{J}_b(\delta\alpha) + \hat{J}_o(\delta\alpha) \tag{8.16}$$

where

$$\hat{J}_b(\delta\alpha) = \frac{1}{2}(\delta\alpha)^{\mathbf{T}}\left(\Psi^{\mathbf{T}}\mathbf{B}^{-1}\Psi\right)(\delta\alpha) \tag{8.17}$$

$$\hat{J}_o(\delta\alpha) = \frac{1}{2}\sum_{k=0}^{k=n}\left(\mathbf{H}_k\mathbf{M}_k(y_b + \Psi\delta\alpha) - y_k^o\right)^{\mathbf{T}} \mathbf{R}_k^{-1}\left(\mathbf{H}_k\mathbf{M}_k(y_b + \Psi\delta\alpha) - y_k^o\right) \tag{8.18}$$

Since the inverse of the background error covariance matrix $\mathbf{B}^{-1}$ is a symmetric positive definite matrix (S.P.D), it is easy to verify that $\Psi^{\mathbf{T}}\mathbf{B}^{-1}\Psi$ is S.P.D from the fact that $\Psi^{\mathbf{T}}\Psi = \mathbf{I}$.

Define

$$\hat{\mathbf{B}}^{-1} = \Psi^{\mathbf{T}}\mathbf{B}^{-1}\Psi. \tag{8.19}$$

Therefore $\hat{\mathbf{B}}^{-1}$ is S.P.D and (8.17) can be written as,

$$\hat{J}_b(\delta\alpha) = \frac{1}{2}(\delta\alpha)^{\mathbf{T}}\hat{\mathbf{B}}^{-1}(\delta\alpha) \tag{8.20}$$

Since $\hat{\mathbf{B}}^{-1}$ is S.P.D, we can find the square-root matrix

$$\hat{\mathbf{B}} = \hat{\mathbf{B}}^{\frac{1}{2}}\hat{\mathbf{B}}^{\frac{\mathbf{T}}{2}} \tag{8.21}$$

using the inverse Cholesky decomposition methodology without finding $\hat{\mathbf{B}}$ itself. Define a transformation $\delta\alpha = \hat{\mathbf{B}}^{\frac{1}{2}}v^{\alpha}$. Hence, we obtain that

$$\begin{aligned}
\tilde{J}_b(v^{\alpha}) &= \hat{J}_b(\delta\alpha) = \hat{J}_b\left(\hat{\mathbf{B}}^{\frac{1}{2}}v^{\alpha}\right) = \frac{1}{2}(\delta\alpha)^{\mathbf{T}}\hat{\mathbf{B}}^{-1}(\delta\alpha) = \frac{1}{2}\left(\hat{\mathbf{B}}^{\frac{1}{2}}v^{\alpha}\right)^{\mathbf{T}}\hat{\mathbf{B}}^{-1}\left(\hat{\mathbf{B}}^{\frac{1}{2}}v^{\alpha}\right) \\
&= \frac{1}{2}\left(\hat{\mathbf{B}}^{\frac{1}{2}}v^{\alpha}\right)^{\mathbf{T}}\left(\hat{\mathbf{B}}^{\frac{1}{2}}\hat{\mathbf{B}}^{\frac{\mathbf{T}}{2}}\right)^{-1}\left(\hat{\mathbf{B}}^{\frac{1}{2}}v^{\alpha}\right) = \frac{1}{2}(v^{\alpha})^{\mathbf{T}}\hat{\mathbf{B}}^{\frac{\mathbf{T}}{2}}\hat{\mathbf{B}}^{-\frac{\mathbf{T}}{2}}\hat{\mathbf{B}}^{\frac{1}{2}}\hat{\mathbf{B}}^{\frac{1}{2}}v^{\alpha} \\
&= \frac{1}{2}(v^{\alpha})^{\mathbf{T}}v^{\alpha}.
\end{aligned} \tag{8.22}$$

The methodology of construction of $\mathbf{B}^{1/2}$ and $\mathbf{B}^{T/2}$ using univariate correlation and multivariate geostrophic balancing operators is detailed as follows. (see also Akella, 2006[124]).

The model variables $(\mathbf{h}, \mathbf{u}, \mathbf{v})$ are partitioned into balanced and unbalanced components. The so-called balancing operator, $\mathbf{K}_b$ acts on the unbalanced components of the model variables and in-turn, $\mathbf{K}_b = \mathbf{K'}_b + I$. Following [129], $\mathbf{K'}_b$ is formulated using the linear balance equations, based on geostrophic balance (written in spherical coordinates) and hydrostatic hypothesis.

Geostrophic balance:

$$\begin{aligned}
u &= -\frac{1}{\rho f}\left[\frac{1}{a}\frac{\partial p}{\partial\theta}\right], \\
v &= \frac{1}{\rho f}\left[\frac{1}{a\cos\theta}\frac{\partial p}{\partial\lambda}\right].
\end{aligned}$$

Hydrostatic hypothesis: $p = \rho g h$.

Which implies,

$$\begin{aligned}
u &= -\frac{g}{f}\left[\frac{1}{a}\frac{\partial h}{\partial\theta}\right], \\
v &= \frac{g}{f}\left[\frac{1}{a\cos\theta}\frac{\partial h}{\partial\lambda}\right].
\end{aligned}$$

Therefore

$$\mathbf{K}_b = \mathbf{K'}_b + I = \begin{pmatrix} I & 0 & 0 \\ -\frac{g}{a\,f}\frac{\partial}{\partial\theta} & I & 0 \\ \frac{g}{a\,f\,\cos\theta}\frac{\partial}{\partial\lambda} & 0 & I \end{pmatrix}$$

which is a lower triangular matrix, since our control vector is of the form $(\mathbf{h},\, \mathbf{u},\, \mathbf{v})^T$.

Remark: At the North and South poles, one sided differences have been used for computing the above derivative with respect to the latitude and at the equator, where $\theta = \pi/2$, we have used the average values of the derivative (with respect to the longitude) from the two neighboring latitude circles, above and below the equator.

Using the balance operator, we can write $\mathbf{B} = \mathbf{K}_b\,\mathbf{B}_u\,\mathbf{K}_b^T$, where $\mathbf{B}_u$ is a block diagonal error covariance matrix for the unbalanced component of the variables (see [130]), which implies that the cross-covariances between the unbalanced variables is taken to be negligible. Thus $\mathbf{B}_u = \mathbf{\Sigma_b}\,\mathbf{C}\,\mathbf{\Sigma_b}$, where $\mathbf{\Sigma_b}$ is a block-diagonal matrix of the background-error variances in the grid point space, such that the diagonal entries represent error variances at every grid point (in this work, we prescribed $\Sigma_b = [2000\,I, 100\,I, 100\,I]$).

$\mathbf{C}$ is a symmetric matrix of background-error correlations for the unbalanced component of the variables. Assuming that $\mathbf{C}$ is block-diagonal, which is a valid assumption, since $\mathbf{B}_u$ has already been assumed to be block-diagonal, we obtain the square-root factorization $\mathbf{C} = \mathbf{C}^{1/2}\,\mathbf{C}^{T/2}$.

Thus the square-root factorization of the background error covariance can be written as,

$$\begin{aligned} \mathbf{B} &= \mathbf{K}_b\,\mathbf{B}_u\,\mathbf{K}_b^T = \mathbf{K}_b\,(\mathbf{\Sigma_b}\,\mathbf{C}\,\mathbf{\Sigma_b})\,\mathbf{K_b^T} = \mathbf{K_b}\,(\mathbf{\Sigma_b}\,\mathbf{C^{1/2}}\,\mathbf{C^{T/2}}\mathbf{\Sigma_b})\,\mathbf{K_b^T} \\ &= (\mathbf{K}_b\,\mathbf{\Sigma_b}\,\mathbf{C^{1/2}})\,(\mathbf{C^{T/2}}\mathbf{\Sigma_b}\,\mathbf{K_b^T}) \\ &= \mathbf{B}^{1/2}\,\mathbf{B}^{T/2}. \end{aligned} \tag{8.23}$$

Notice that the above formulation ensures that $\mathbf{B}$ is symmetric and positive definite, both of these properties are usually required to be satisfied by any preconditioning matrix. The analysis increment is given by $\delta\mathbf{x} = \mathbf{B}^{1/2}\mathbf{v} = \mathbf{K}_b\,\mathbf{\Sigma_b}\,\mathbf{C^{1/2}}\mathbf{v}$. Since $\mathbf{C}$ is block-diagonal, the operation $\mathbf{C}^{1/2}\,\mathbf{v}$ can be split into individual operators $\mathbf{C}_\alpha^{1/2}\,\mathbf{v}_\alpha$, that act independently on different components of the variable $\mathbf{v}$, such as $\mathbf{v}_\alpha$. For each variable, the univariate operator can be factorized into $\mathbf{C}_\alpha = \mathbf{C}_\alpha^{1/2}\,\mathbf{C}_\alpha^{T/2}$. The procedure suggested by [130] has been implemented to model the univariate correlation operator, has been implemented to model the univariate correlation operator, $\mathbf{C}_\alpha$ as an isotropic diffusion operator, assuming Gaussianity with a decorrelation length equal to $500\,km$.

We considered height field which was comprised of a single Dirac delta pulse located at equator and longitude 180^o, and prescribed no wind field, the action of $\mathbf{B}$ on such a field is shown in Figure 8.1. We see the effect of the correlation operator on the Dirac pulse and also on the wind field obtained under geostrophic balance assumption Figure 8.2, which is parallel to the isobars of the pressure. Since there is a *high pressure* at the center, the direction of the wind is clockwise in the Northern hemisphere and anti-clockwise in the Southern hemisphere; at the equator due to the balancing of the pressure gradient and Coriolis forces, the wind blows straight.

Therefore the gradient of the background cost functional, $\tilde{J}_b\left(v^\alpha\right)$ with respect to v^α is given by,

$$\nabla_{v^\alpha}\tilde{J}_b = v^\alpha \tag{8.24}$$

and the Hessian of the background cost functional, $\tilde{J}_b\left(v^\alpha\right)$, with respect to v^α is given by,

$$\nabla^2_{v^\alpha}\tilde{J}_b = \mathbf{I}_M. \tag{8.25}$$

To summarize, we obtain that the cost functional can be approximated by,

$$J\left(y_0\right) \approx \tilde{J}\left(v^\alpha\right) = \tilde{J}_b\left(v^\alpha\right) + \tilde{J}_o\left(v^\alpha\right) = \frac{1}{2}\left(v^\alpha\right)^{\mathbf{T}} v^\alpha + \tilde{J}_o\left(v^\alpha\right), \tag{8.26}$$

and the gradient of the cost functional with respect to v^α is given by chain rule,

$$\begin{aligned} \nabla_{v^\alpha}\tilde{J} &= v^\alpha + \left(\nabla_{v^\alpha}\alpha^0\right)^{\mathbf{T}}\left(\left(\nabla_{\alpha^0}y^0\right)^{\mathbf{T}}\nabla_{y^0}J_o\right) \\ &= v^\alpha + \hat{\mathbf{B}}^{\frac{\mathbf{T}}{2}}\Psi^{\mathbf{T}}\nabla_{y^0}J_o. \end{aligned} \tag{8.27}$$

8.4 POD 4-D Var using full ERA-40 observations

8.4.1 ERA-40 observations

Reanalyzed data on a $2.5^o \times 2.5^o$ grid (500 hPa pressure level- geopotential height and velocity fields) from the ERA-40, 40-year reanalysis system (http://www.ecmwf.int/research/era/), valid at 0000 UTC 2 February 2001 was used to specify the initial conditions for forward model integration. These initial conditions were unchanged in all the following test cases. As for boundary conditions, since the domain being considered is spherical, it is obvious that the boundary conditions remain unchanged. The *unconstrained van Leer scheme* with a $2.5^o \times 2.5^o$ (144×72 cells) grid resolution and time step of $\Delta t = 450$ seconds, has been

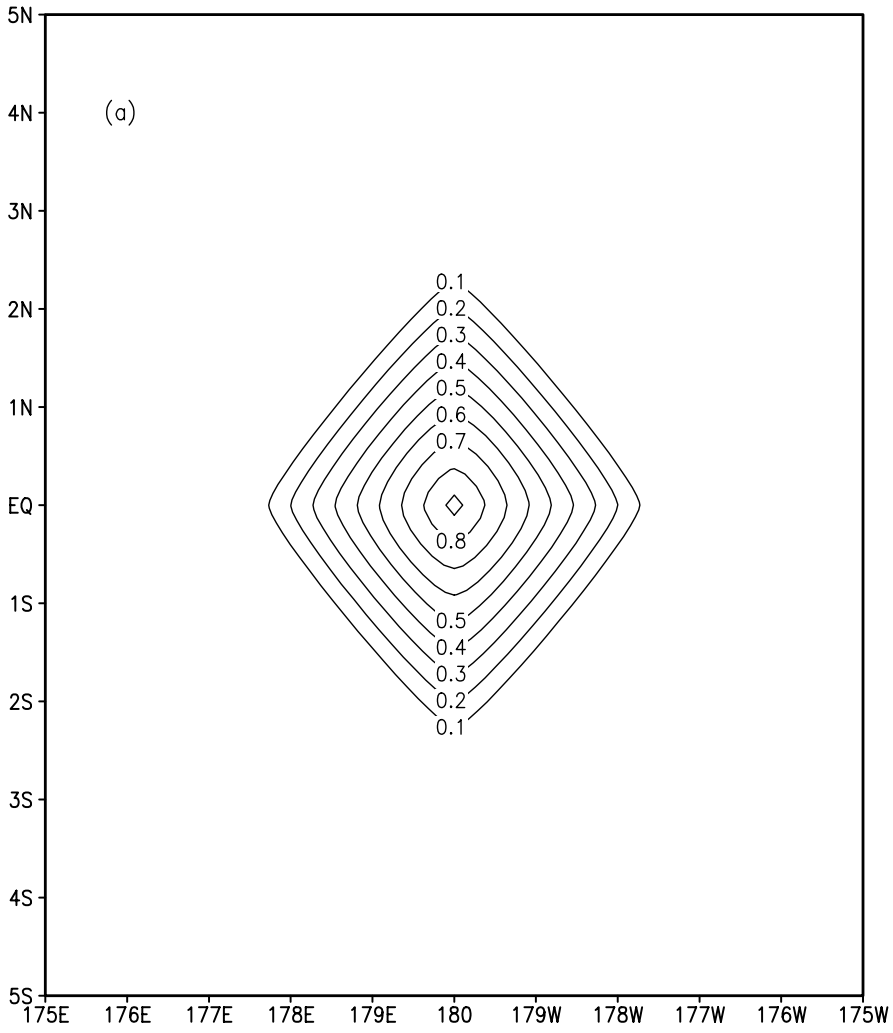


Figure 8.1: Result obtained by operating with **B** on a single Dirac delta pulse in the height field: isolines of the height field

used in this article, to generate a *reference trajectory*. Synthetic observations are obtained by randomly perturbing the reference trajectory, in which the observational error covariance matrix has been taken to be a block diagonal matrix $R = \begin{bmatrix} 10^4 I & 10^2 I & 10^2 I \end{bmatrix}$, where I is a identity matrix. For the entries in R, the values of the variances are specified based on typical values of the variables. The zonal and meridional winds vary on a scale of $10m/sec - 100m/sec$. Hence, a value of 100 was specified for their variances. For the geopotential height field, $\Phi = gh$ varies on a scale of $10^4 m^2/sec^2$.

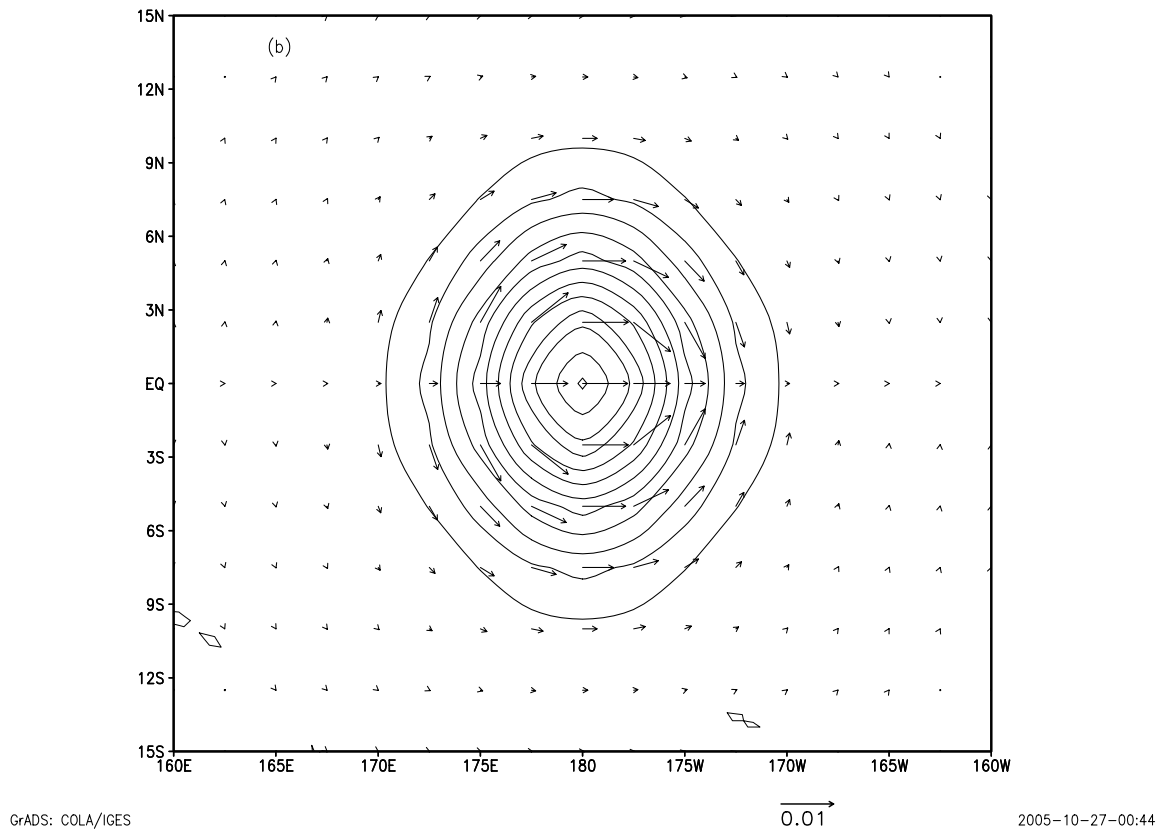


Figure 8.2: Result obtained by operating with **B** on a single Dirac delta pulse in the height field: geostrophic wind plotted along with the isolines of the height field

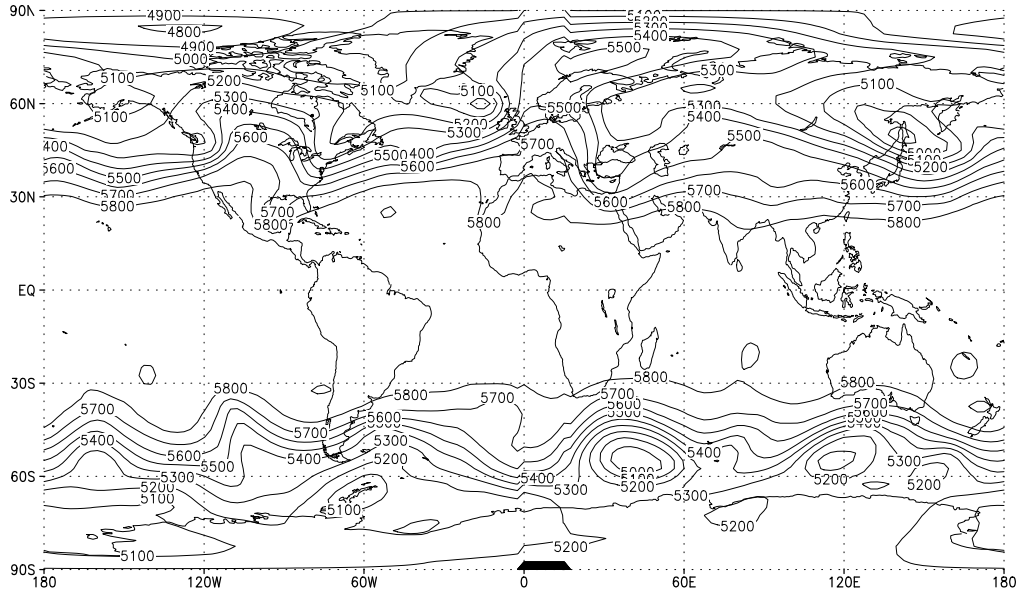
In the numerical experiment, we carried out a 1% normally distributed random perturbation on the true initial conditions over the entire vector $X = [u, v, h]$ field in Figure(8.3a) specified from ERA-40 in order to provide twin-experiment "observations". Also, the 18-h forecast of the FV-SWE model was taken to be forecast verification time displayed in Figure(8.3b). The 4-D Var optimization loop was stopped when the l_2 norm of the gradient was less than a tolerance of 10^{-3}. Since we didn't change the tolerance, the results we obtained are not affected. It is obvious that if we were to make the tolerance more stringent, the optimization would have required more iterations. The reduction of the cost functional is measured by the value of the current cost functional normalized by the initial one with or without the logarithmic scale. We computed the errors between the true initial conditions and the retrieved initial conditions related to a 1% normally distributed random

perturbations of the true initial conditions as the initial guess of the reduced-order 4-D Var. The data assimilation was carried on a 15 hours window using the $\Delta t = 450$s in time and a mesh of 144×72 grid points in space and the observations are available every 3 hours in time including the initial time. Thus we have $144 \times 72 \times 3 \times 6$ observations distributed in time and space.

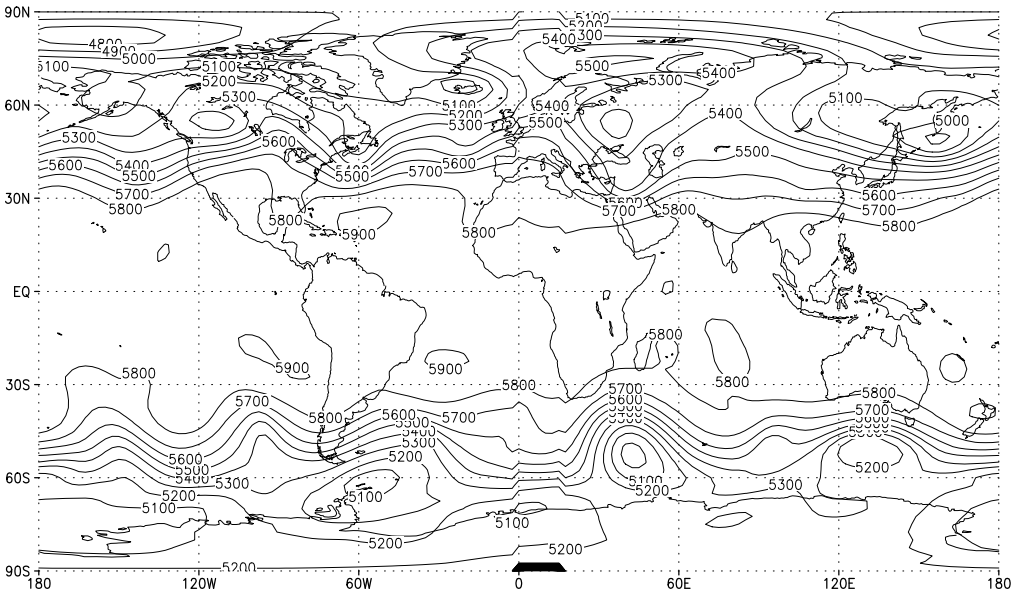
Now we generated 120 snapshots by integrating the full FV-SWE model forward in time, from which we choose 15 POD modes or 15 DWPOD modes for each of the $(u(x,y), v(x,y), \phi(x,y))$ to capture over 99.9% of the energy. The singular value decomposition for both POD modes and DWPOD modes from the snapshots is displayed in Figure(8.4a). The energy captured by the leading POD modes or DWPOD modes from the snapshots as a function of the dimension of the POD reduced space is displayed in Figure(8.4b). Also, the isopleths of the POD modes of dimension 1, 5 and 10 are displayed in Figure(8.5). The other POD modes, though not plotted show a gradual shift in where most energy is localized; that is, the leading POD modes display most energy uniformly distributed almost on the entire globe, whereas the latter POD modes show a shift towards the north and south poles, we attribute this observation to our particular FV-SWE model. Similar observation was made by Akella and Navon (2006) [124] in terms of where the largest errors in the retrieved initial conditions were obtained in their 4-D Var twin experiments using the FV-SWE model. Furthermore, the dimension of control variables vector for the POD reduced-order 4-D Var thus is $15 \times 3 = 45$ compared to $144 \times 72 \times 3 = 31104$ for the full 4-D Var

8.4.2 POD reduced-order 4-D Var experiments

Two POD reduced-order 4-D Var experiments are set up, in which the first experiment, hereafter refereed as DAS-I, had no background term included in the POD reduced-order cost functional and the second one, hereafter referred as DAS-II, had the background error covariance term included in the POD reduced-order cost functional. The background state was generated using a 1% normal random perturbations on the initial conditions, in which the background error covariance matrix has been taken to be a block diagonal matrix $B = \begin{bmatrix} 2 \times 10^4 I & 10^2 I & 10^2 I \end{bmatrix}$. In practice, by applying random number generator using CPU clock cycle, we made sure that the seeds used to generate pseudo normal random perturbations for twin-experiment "observations" are nearly uncorrelated with the seeds

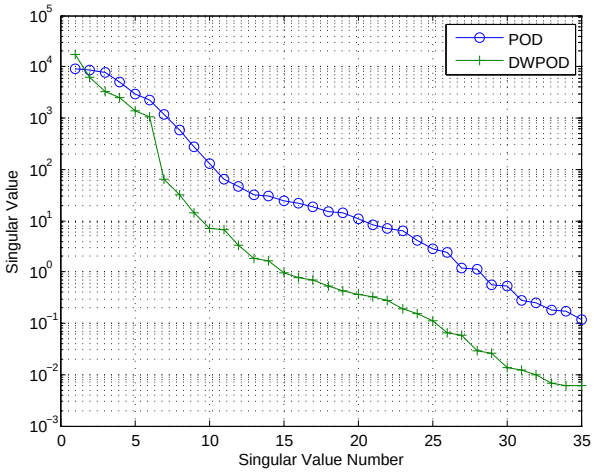


(a) The configuration at the initial time specified from ERA-40 data sets

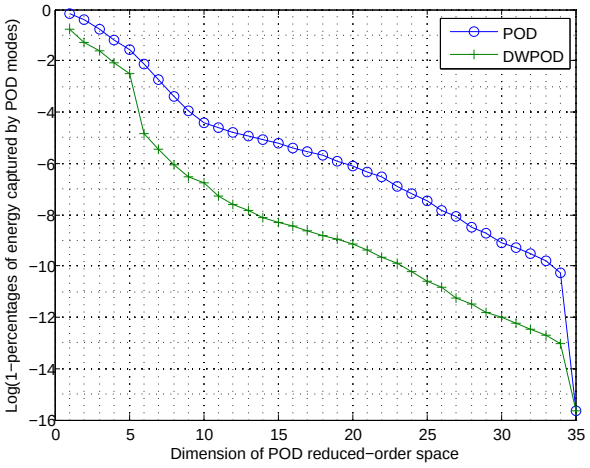


(b) The 18-h forecast of the FV-SWE model using unconstrained Van-Leer advection scheme.

Figure 8.3: Isopleths of the geopotential height for the reference trajectory



(a) Unweighted SVD and dual weighted SVD



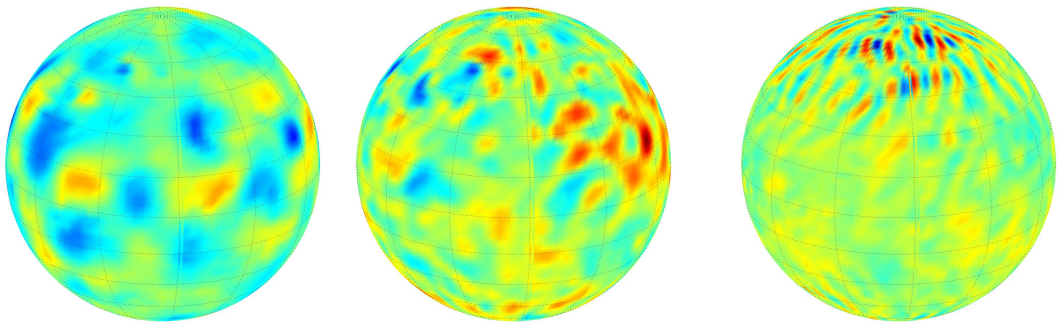
(b) The percentage of energy captured by POD

Figure 8.4: Singular value decomposition

used to generate normal random perturbations for background terms in the reduced-order cost functional.

In the process of POD 4-D Var, the resulting control variables from the latest optimization iteration are projected to the full model to generate new POD bases. The new POD bases then replace the previous ones resulting in a new POD reduced-order model. We found that the root mean square error metrics between the full model solutions and reduced-order solutions were consistently improved after each outer projection was carried out.

The limited memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS) update algorithm for quasi-Newton minimization ([125]) was employed for high-fidelity full model 4-D Var and all variants of ad-hoc POD 4-D Var, while a variant of the LBFGS, called LBFGS-B[126, 127] which can handle box-constraints on the variables was employed for the trust-region POD 4-D Var within the trust-region radius and provides a sufficient reduction of the high-fidelity model quantified in terms of the Cauchy point [70]. In the ad-hoc POD 4-D Var[58, 65], the POD bases are re-calculated when the value of the cost function cannot be decreased by more than a factor of 0.5 for ad-hoc POD 4-D Var and 0.1 for ad-hoc DWPOD 4-D Var between consecutive minimization iterations. The reason for the particular choice of these values is based on numerical experience and relative rate of convergence of the ad-hoc and



(a) 1st POD mode (b) 5th POD mode (c) 10th POD mode

Figure 8.5: Isopleths of the POD modes of dimension 1, 5 and 10 respectively

dual weighted POD methods, respectively. In the trust-region 4-D Var, the POD bases are re-calculated when the ratio ρ_k is larger than the trust-region parameter η_1 in the process of updating the trust-region radius.

The unweighted ad-hoc POD 4-D Var as a reduced order approach required a smaller computational cost but could not achieve the same cost functional reduction as the high-fidelity model 4-D Var. The dual weighted ad-hoc POD 4-D Var achieves a better reduction of the cost functional. However, neither of the above mentioned methods can attain the minimum of the high fidelity 4-D Var model cost functional. Furthermore, the unweighted snapshots trust-region POD 4-D Var yields an additional cost functional reduction compared to the ad-hoc approach, albeit at a higher computational cost. Finally, the dual weighted trust-region POD 4-D Var achieves almost exactly the same cost functional reduction as the full high fidelity 4-D Var model, resulting in an additional decrease of four orders of magnitude compared to the minimization of the cost functional obtained by applying the unweighted ad-hoc POD 4-D Var (see Figure 8.6a and Figure 8.6b), showing that the combination of the dual-weighted approach and trust-region method to model reduction is significantly beneficial in attaining a local minimum of the cost functional almost identical to one obtained by the high fidelity full 4-D Var, while the computation of effort for dual weighted trust-region POD 4-D Var is much less than the one required for full 4-D Var(see Table 8.1a and Table 8.1b).

Table 8.1: Comparison of iterations, outer projections, error, and CPU time for ad-hoc POD 4-D Var, trust-region POD 4-D Var, trust-region dual-weighted POD 4-D Var and full 4-D Var

(a) DAS-I

DAS I	UWAHPOD	DWAHPOD	UWTRPOD	DWTRPOD	Full
Iterations	23	24	16	23	42
Outer projections	2	2	14	14	NA
$\log\left(\frac{J_f}{J_0}\right)$	$10^{-0.37}$	$10^{-0.69}$	$10^{-1.78}$	$10^{-2.32}$	$10^{-2.50}$
CPU time (s)	117.1	149.2	143.2	181.7	601.7

(b) DAS-II

DAS II	UWAHPOD	DWAHPOD	UWTRPOD	DWTRPOD	Full
Iterations	14	59	50	62	100
Outer projections	2	2	15	16	NA
$\frac{J_f}{J_0}$	0.72	0.54	0.17	0.13	0.10
CPU time (s)	100.3	207.7	280.1	352.5	966.7

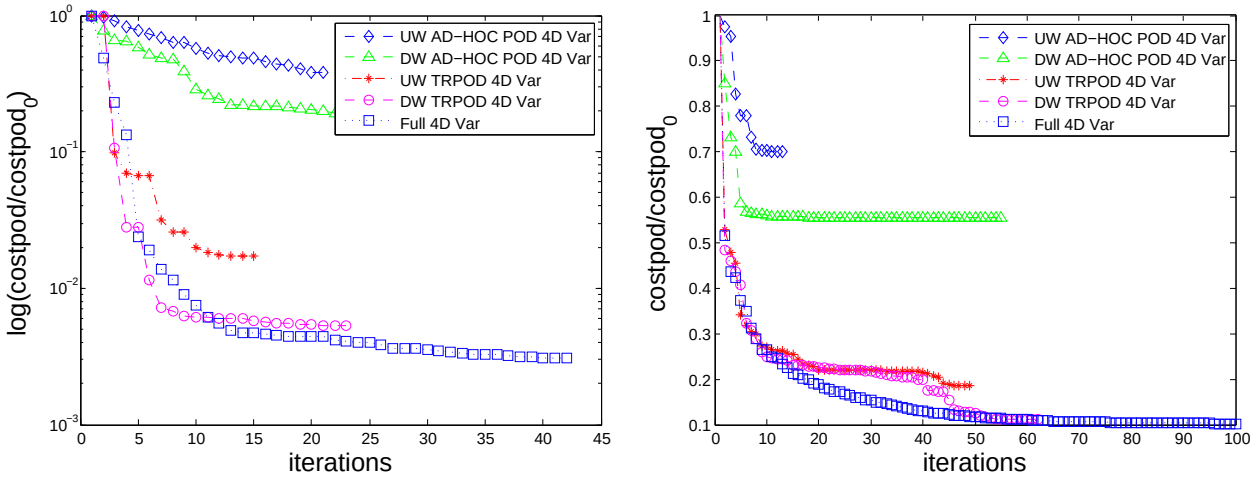
In Figure(8.7a) and Figure(8.7b), we found that the minimization of the cost functional using full 4-D Var will be terminated if the scaled norm of the gradient of the cost functional can decrease by 2 orders of magnitude, while the one using DWTRPOD 4-D Var will be terminated if the corresponding scaled norm of the gradient can decrease by 3 orders of magnitude, which can be explained by the fact that the POD reduced-order space is dimensionally lower than the full space.

Once the retrieved initial condition is obtained by implementing the dual weighted trust-region 4-D Var, we can compare the results from the POD reduced-model with those from the full model. To quantify the performance of the dual weighted trust-region 4-D Var, we used the metric namely the root mean square error (RMSE) of the difference between the POD reduced-order simulation and high-fidelity model.

In particular, the RMSE between variants of the POD reduced-model solution and the true one at the time level i is used to estimate the error of the POD model.

$$\mathrm{RMSE}^{\mathrm{i}} = \sqrt{\frac{\sum_{j=1}^{j=N}\left(U_{i,j} - U_{i,j}^{POD}\right)^2}{N}}, \quad i = 1, \ldots, n \tag{8.28}$$

where $U_{i,j}$ and $U_{i,j}^{POD}$ are the state variables obtained by the full model and ones obtained



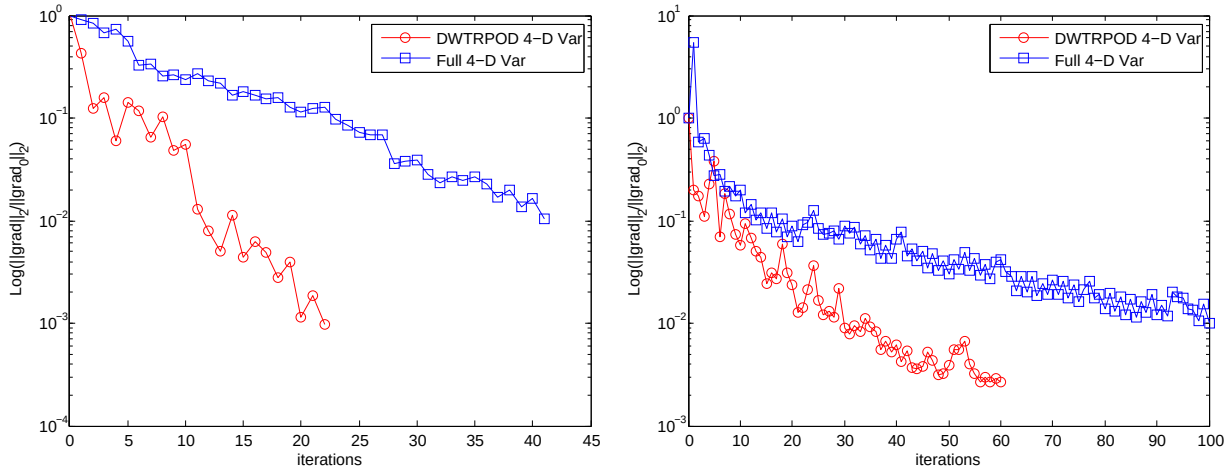
(a) DAS-I

(b) DAS-II

Figure 8.6: Comparison of the performance of the iterative minimization process of the scaled cost functional for unweighted ad-hoc POD 4-D Var, dual weighted ad-hoc POD 4-D Var, unweighted trust-region POD 4-D Var, dual weighted trust-region 4-D Var, and full model 4-D Var respectively.

by optimal POD reduced-order model of time level i at node j, respectively, and N is the total number of nodes over the domain. U and U^{POD} are used to either denote the geopotential or the velocity of the full model and those corresponding to the POD reduced-order model, respectively.

Even though it turned out to be advantageous to combine the dual-weighed approach with the trust-region POD 4-D Var, it should be emphasized that this advantage diminishes when we increase the number of POD bases for each component of the $(u(x,y), v(x,y), \phi(x,y))$ from 15 to 25. This remark is based on RMSE and also the difference between the 18h-forecast using true initial conditions and the one using retrieved initial condition after data assimilation. However, increasing the dimension of the POD reduced-order space from 45 to 75 can increase the computational cost of POD reduced-order 4-D Var. This agrees with results obtained in [102] that for practical applications, the dual-weighted procedure may be of particular benefit for use only with small dimensional bases in the context of adaptive order reduction as the minimization approaches the optimal solution. For other beneficial effects of POD 4-D Var related to its use in the framework of second order adjoint of a global



(a) DAS-I (b) DAS-II

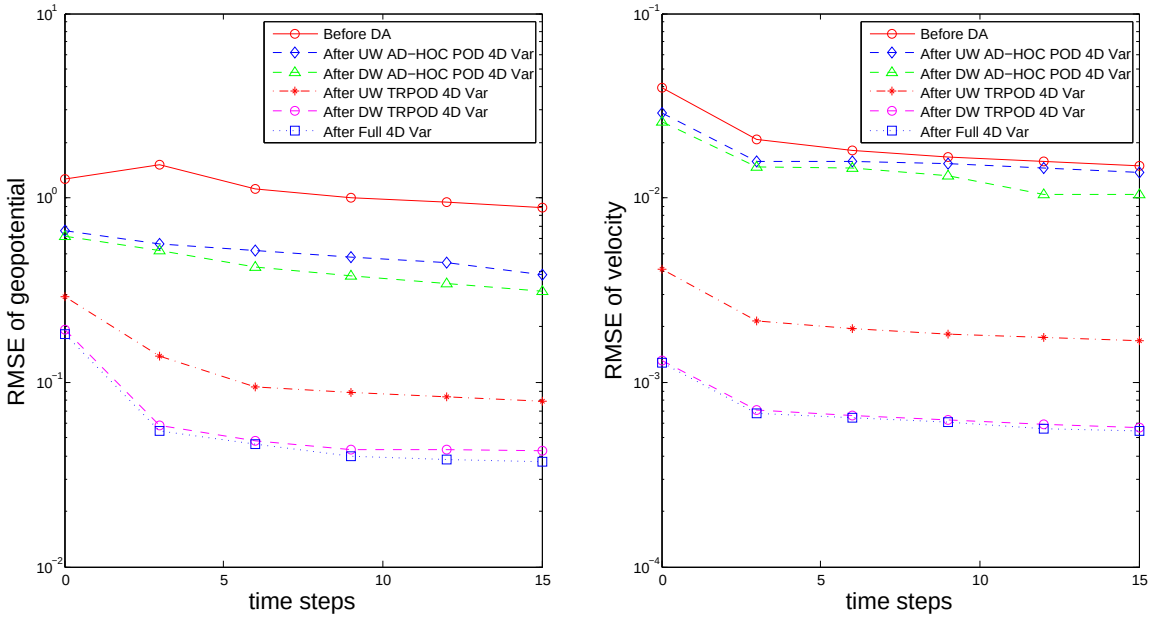
Figure 8.7: Comparison of the performance of the iterative minimization process of the scaled norm of the gradient of the cost functional for dual weighted trust-region 4-D Var and full model 4-D Var.

shallow water equations model see Daescu and Navon (2007) [101].

Finally in Figure(8.9a) and Figure(8.9b) we compared the errors in retrieved initial conditions without and with background error covariance terms (i.e., DAS-I and DAS-II experiments). Notice that in both cases the largest errors occur in the polar regions (see note in Section 5.1). With the background term, we obtained an improved estimation of the true initial condition in DAS-II, compared to DAS-I, as evident through the RMSE plots (Figure(8.8a) and Figure(8.8b)) as well. Such advantages of the background term in 'full' 4 -D Var are well documented in [105].

8.4.3 Nonlinearity in the projection

Due to the complexity of the Lin-Rood finite volume code, the numerical fluxes had to be computed at the element boundaries. This required us to go back to the full model in order to evaluate the numerical fluxes, in order to deal with the nonlinearity in the projection. The numerical problem of reducing the complexity of evaluating the nonlinear terms of the POD reduced model in the context of finite volume (FV) requires for this quadratic nonlinearity a pre-computing of a special POD-Galerkin projection. However, the pre-computing technique



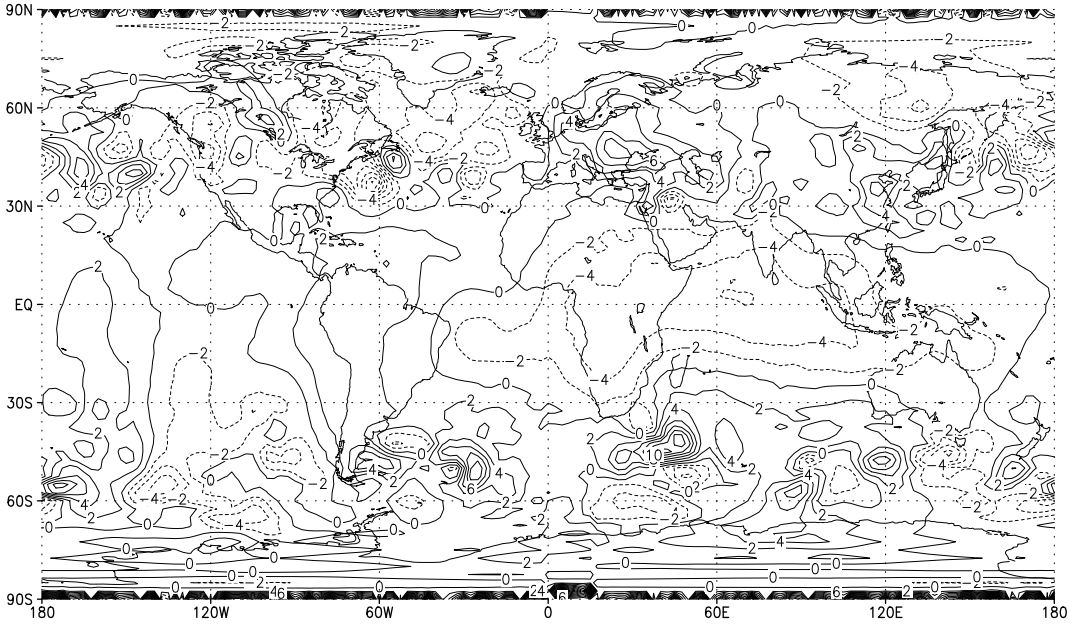
(a) RMSE of Geopotential in DAS-II

(b) RMSE of Wind velocity in DAS-II

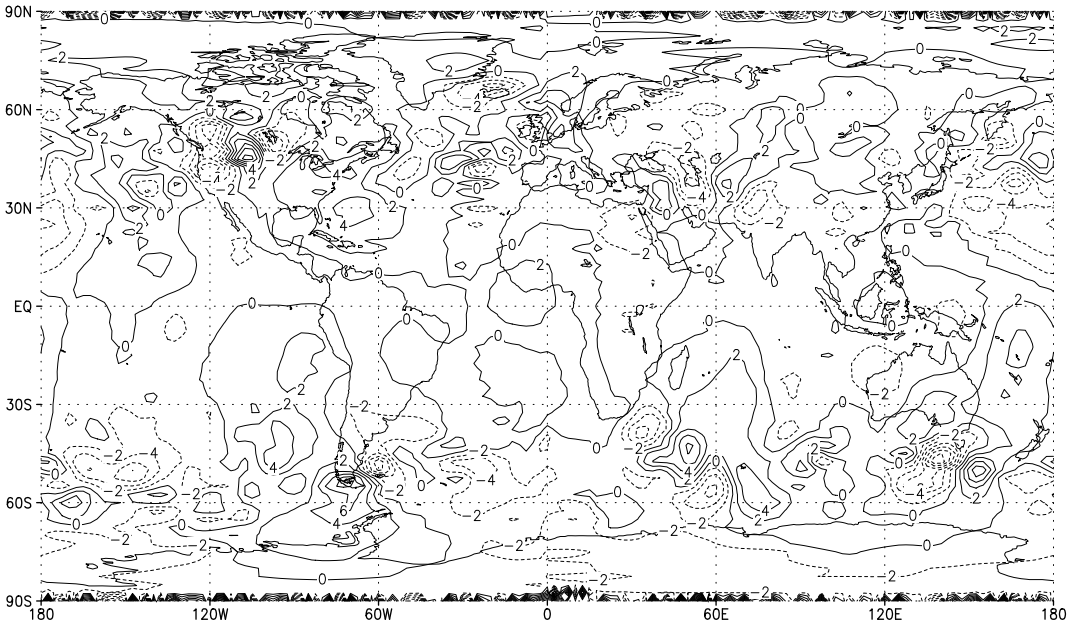
Figure 8.8: Comparison of the RMSE in DAS-II experiments among unweighted ad-hoc POD 4-D Var, dual weighted ad-hoc POD 4-D Var, unweighted trust-region POD 4-D Var, dual weighted trust-region 4-D Var, and full model 4-D Var respectively.

proved to be very difficult to implement due to the algorithmic features of the Lin Rood FV scheme. This explains why we obtained only a speed up of a factor of order 3 as shown in Table 8.1a and Table 8.1b.

An elegant solution to this problem was put forward by Chaturantabut (2008) [106], Chaturantabut and Sorensen (2010a, b) [107, 108] where they proposed a method referred as a Discrete Empirical Interpolation Method (DEIM). DEIM achieves a complexity reduction of the nonlinearities which is proportional to the number of reduced variables while POD retains a complexity proportional to the original number of variables. The DEIM approach approximates a nonlinear function by combining projection with interpolation. DEIM constructs specially selected interpolation indices that specify an interpolation-based projection so as to provide a nearly l_2 optimal subspace approximation to the nonlinear term, without the expense of orthogonal projection.



(a) DAS-I



(b) DAS-II

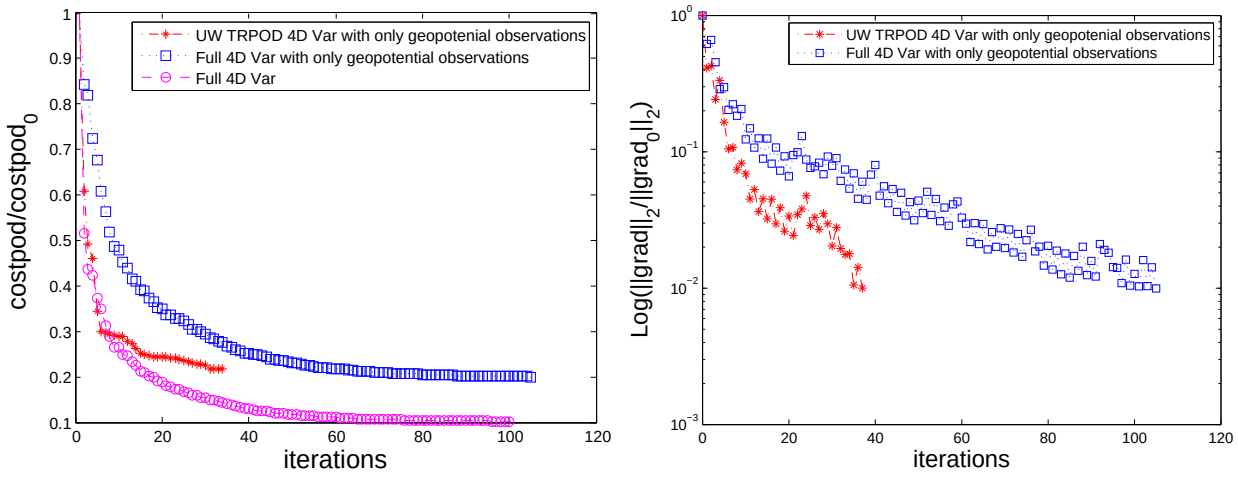
Figure 8.9: Isopleths(scaled by multiplying 1000) of the geopotential height for the difference between the 18h-forecast using true initial conditions and the one using retrieved initial condition after DWTRPOD 4-D Var.

8.5 Results with incomplete observations

8.5.1 The observations of height field only

In DAS-II, meteorological observations are temporally available every 3 hours but spatially distributed at all the grid points. So the question arises as to what will happen if we decrease the number of observations in space [128], i.e, observational operator in the cost functional becomes a sparse matrix.

Suppose that only the geopotential field is observed but the observations for the wind field are unavailable (i.e., the number of observations is decreased from $144 \times 72 \times 3 \times 6$ to $144 \times 72 \times 6$). We refer to this case by DAS-III(a), in which the initial perturbed field is the same as the one used to start DAS-I. In DAS-III(a), the numerical results in Figure(8.10a) show that it takes more iterations for the the cost functional of full 4-D Var with only incomplete observations to converge than the one with full observations. Furthermore, the POD reduced cost functional in DAS-III(a) using the UWTRPOD 4-D Var can be reduced to almost the same degree of magnitude as full 4-D Var in DAS-III(a) displayed in Figure(8.10a). Also, in DAS-III(a) the norm of the gradient of POD reduced cost functional using UWTRPOD 4-D Var and the cost functional using full 4-D Var both decrease by only 2 orders of magnitude, displayed in Figure(8.10b). In Figure(8.11a), an additional experiment was carried out comparing results for UWTRPOD 4-D Var, DWTRPOD 4-D Var as well as full 4-D Var in the case of observations being available only for the geopotential field. It was also found out that the results for DWTRPOD 4-D Var produced similar results as those obtained in the case of DAS III (b) (c) (d) (not shown) experiments with incomplete observations. In Figure(8.11b), corresponding results were displayed for the scaled norm of the gradient for DWTRPOD 4-D Var and full 4-D Var. Again, the other experiments (not shown) exhibited similar results of incomplete observations. In Figure(8.12), we obtained the errors in retrieved initial conditions using UWTRPOD 4-D Var with incomplete observations.(i.e., only the geopotential observations are available) Notice that in this case the largest errors are still dominant in the polar regions, while the overall RMSE becomes larger than the results obtained in DAS-II.



(a) Scaled cost functional

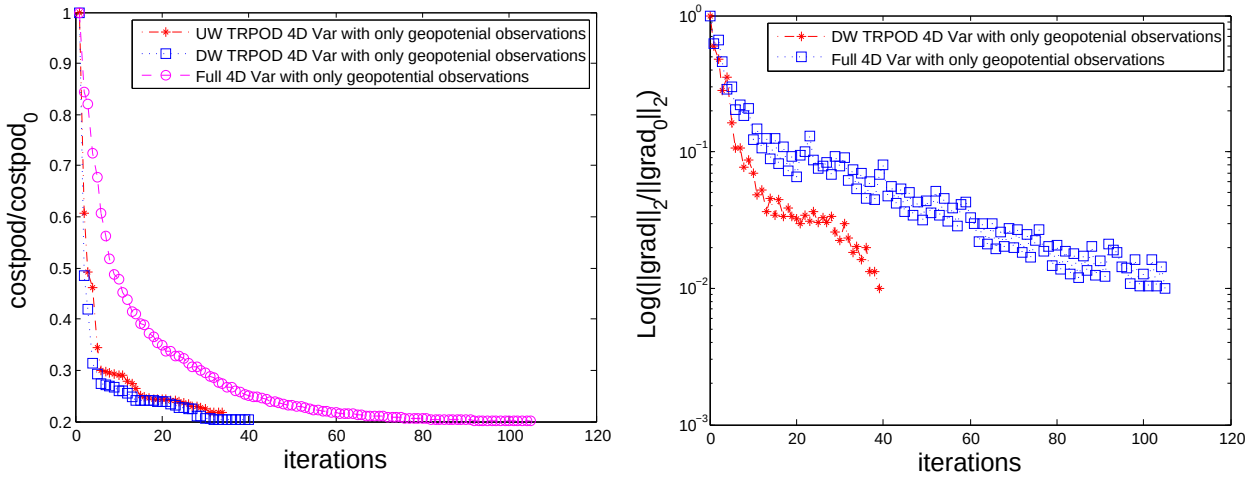
(b) Scaled norm of the gradient

Figure 8.10: DAS-III(a)(Observations of height field only): Comparison of the performance of the iterative minimization process of the scaled cost functional and the scaled norm of the gradient of the cost functional for unweighted trust-region POD 4-D Var and full 4-D Var.

8.5.2 Incomplete observations in space

Next we consider fewer observations along the longitudinal direction. From the earlier number of 144 observations, we specified only 72. Hence the observational resolution is 72×72. But we have observations for [h,u,v] at every three hours as in DAS-II. The reduction in cost functional and scaled gradient norm are plotted in Figure(8.13a) and Figure(8.13b) respectively. Notice that the performance of both the full 4-D Var and UWTRPOD is affected due the alternating observations in one direction.

We follow on the above approach and test what happens when instead of having fewer observations along the longitudinal direction , we have lesser observations along the latitudinal direction, i.e., instead of 72, have only 36 observations, which implies an observational resolution of 144×36. Notice that the performance is not as severely impacted (see Figure(8.14a) and Figure(8.14b)) as in earlier results with 5×2.5 observational resolution. Based on the above two experiments, with observations at 5×2.5 and 2.5×5 grid resolutions, though the cost functional and gradient norm could minimized, as remarked for e.g., Zou et al.[128], such alternating sparsity of the observations affects the condition



(a) Scaled cost functional

(b) Scaled norm of the gradient

Figure 8.11: DAS-III(a)(Observations of height field only): Comparison of the performance of the iterative minimization process of the scaled cost functional and the scaled norm of the gradient of the cost functional for unweighted trust-region POD 4-D Var, dual weighted trust-region POD 4-D Var and full 4-D Var.

number of the Hessian of cost functional, resulting in a poorly conditioned minimization problem. Based on our results, we remark that the POD 4-D Var also suffers from the ill-conditioning as the full 4-D Var for such an observational grid resolution.

In addition, we conducted another experiment where we retained observations of height field at all grid points, whereas the wind components, u and v were observed as following. The observations for the winds fields were not available from 20 degrees North/ South to the North/ South poles, that is, we masked the observations for u and v fields near the poles. The decrease in scaled cost and gradient norm are plotted in Figure 8.15a and Figure 8.15b, respectively. We note a comparable performance of the TRPOD 4D-Var and the full 4D-Var. This example illustrates that the background error covariance, which was implemented using geostrophic balance assumptions is beneficial in POD 4-D Var case, just like it is for the full 4-D Var.

In this chapter, we solved an inverse problem for the POD reduced-order global shallow water equations model using a finite-volume formulation, controlling its initial conditions in presence of observations being assimilated in a time window. In this POD 4-D Var, we

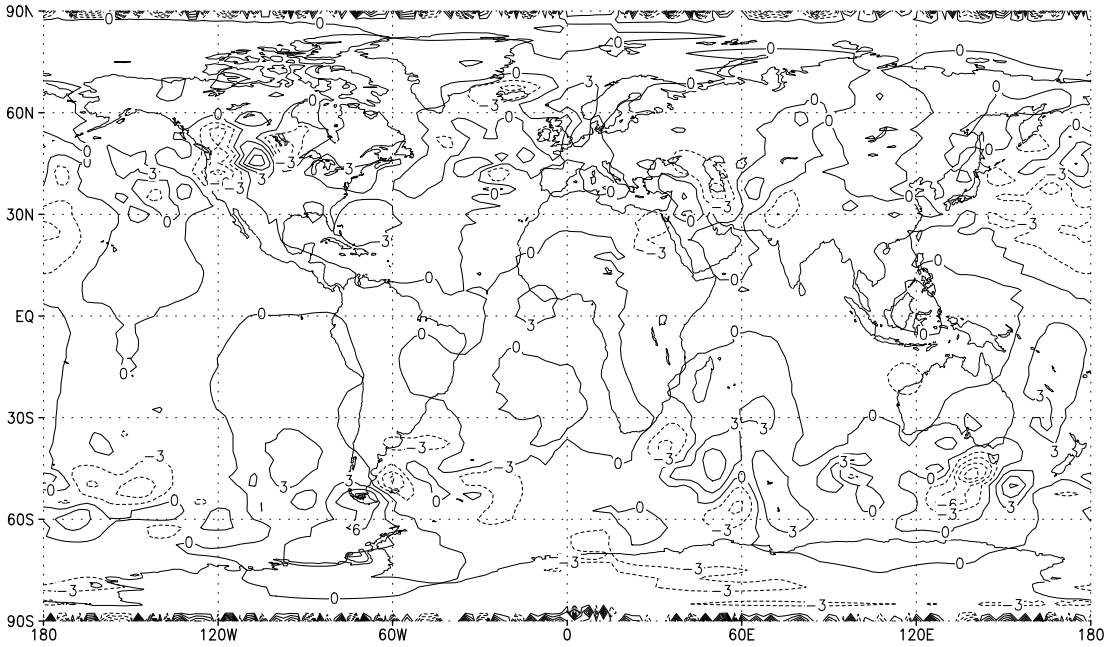
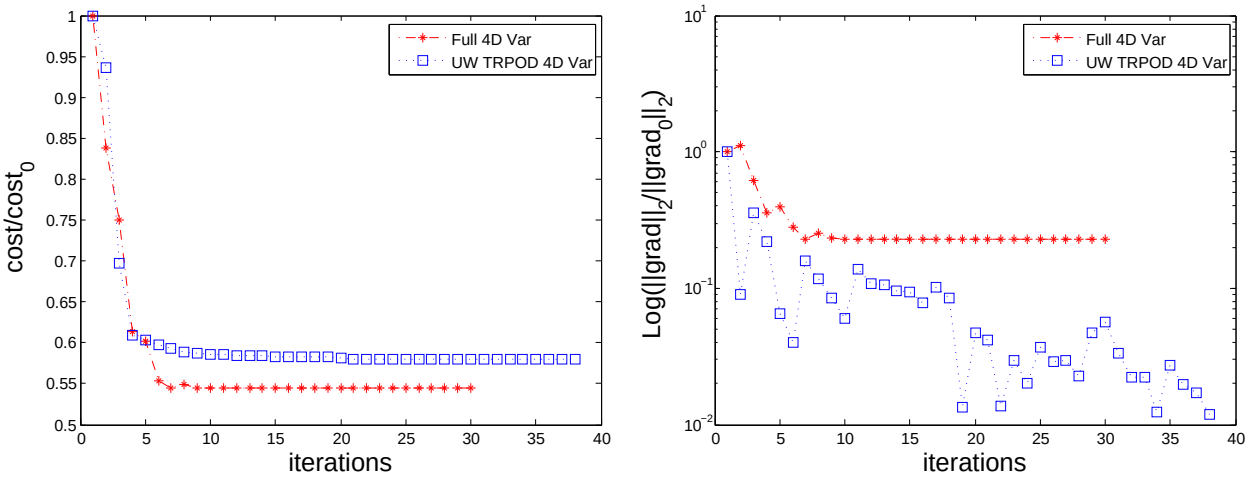


Figure 8.12: DAS-III: Isopleths(scaled by multiplying 1000) of the geopotential height for the difference between the 18h-forecast using true initial conditions and the one using retrieved initial condition after UWTRPOD 4-D Var.

developed the full adjoint of the FV-SWE and by projection we obtained the reduced-order adjoint for POD reduced-order model. We integrated the full adjoint model backward in time to compute the time-varying sensitivities of the full 4-D Var cost functional with respect to time-varying model states, from which we derived the dual weights of the ensemble of snapshots. Also, we projected the gradient of the full cost functional onto the gradient of the POD reduced-order cost functional. Furthermore, after the projection of full background error covariance matrix to low dimensional reduced space, an ideal preconditioning of the POD 4-D Var was obtained so that the Hessian matrix of the POD reduced-order background error covariance matrix became the identity matrix.

In the numerical experiments, we setup two types of 4-D Var experiments, namely, DAS-I without background terms and DAS-II with background term. For both DAS-I and DAS-II, we compared several variants of POD 4-D Var, namely unweighted ad-hoc POD 4-D Var, dual-weighted ad-hoc POD 4-D Var, unweighted trust-region POD 4-D Var and dual-weighted trust-region POD 4-D Var, respectively. We found that the ad-hoc POD 4-D



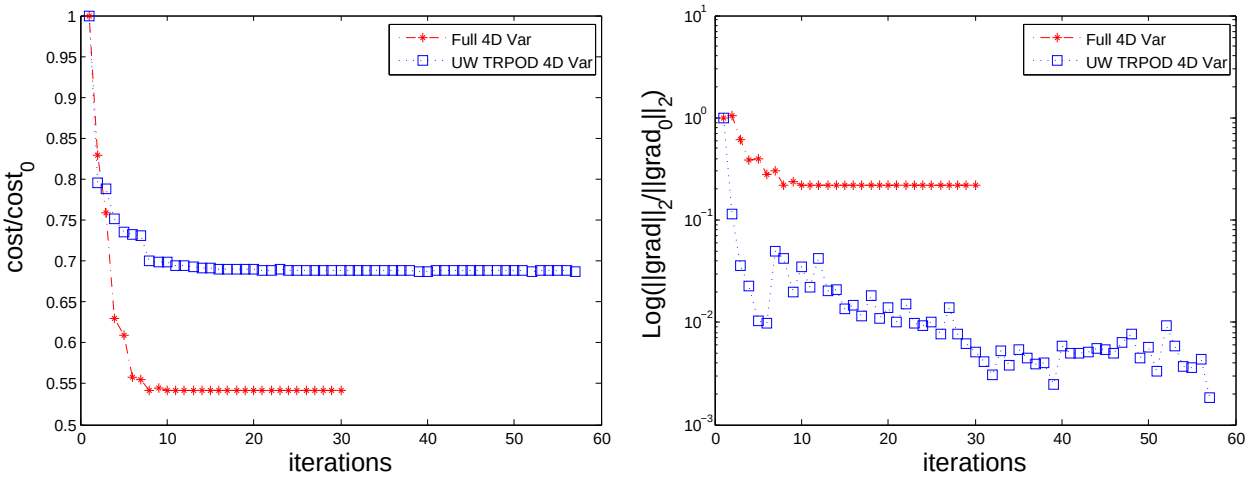
(a) Scaled cost functional

(b) scaled norm of the gradient

Figure 8.13: DAS-III(b)(5×2.5 Resolution): Comparison of the performance of the iterative minimization process of the scaled cost functional and the scaled norm of the gradient of the cost functional for unweighted trust-region POD 4-D Var and full 4-D Var.

Var version yielded the least reduction of the cost functional compared with the trust-region 4-D Var. We assume that this result may be attributed to lack of feedbacks from the high-fidelity model. On the other hand, the trust-region POD 4-D Var version yielded a sizably better reduction of the cost functional, due to inherent properties of TRPOD allowing local minimizer of the full problem to be attained by minimizing the TRPOD sub-problem. Thus trust-region 4-D Var resulted in global convergence to the high fidelity local minimum starting from any initial iterates. The experiments carried out in DAS-III with incomplete observational data indicate that in the case of insufficient data, the minimization is slower. Nevertheless many experiments with incomplete observations show satisfactory performance of the POD reduced 4-D Var, indicating its robustness to lack of observations.

The TRPOD approach for the optimal flow control problem can be viewed as a modification of classical trust region method with a non-quadratic POD model function. In our context, TRPOD was thus implemented for FV-SWE model in order to obtain the robust global convergence based on only a small number of POD basis function. The dual-weighted proper orthogonal decomposition selection of snapshots allows propagation of information from the data assimilation system onto the reduced order model, possibly



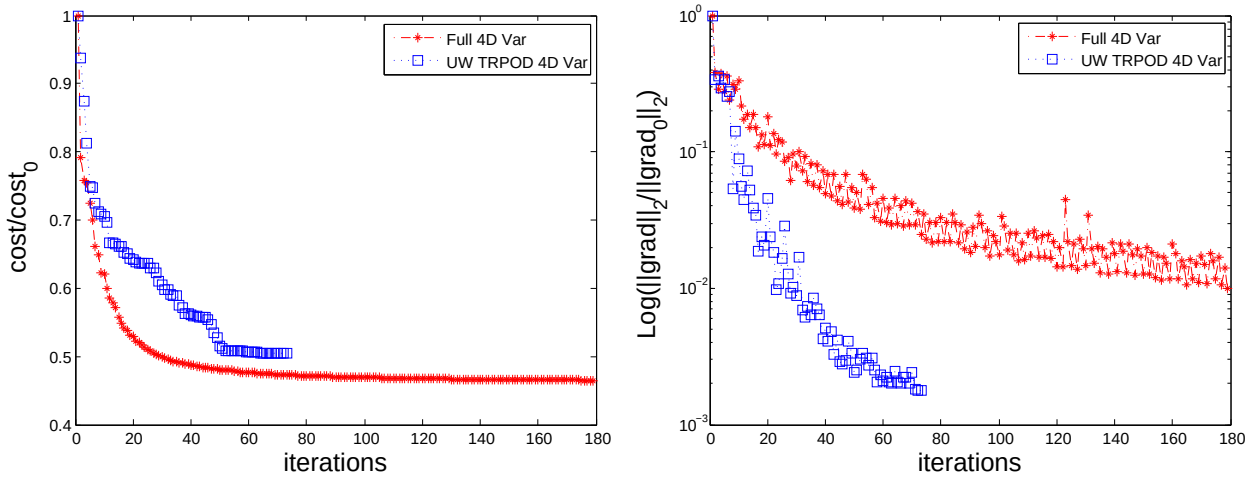
(a) Scaled cost functional

(b) Scaled norm of the gradient

Figure 8.14: DAS-III(c)(2.5×5 Resolution): Comparison of the performance of the iterative minimization process of the scaled cost functional and the scaled norm of the gradient of the cost functional for unweighted trust-region POD 4-D Var and full 4-D Var.

capturing lower energy modes that may play significant role in successful implementation of 4-D Var data assimilation. Combining the dual-weighted approach with the trust-region POD approach to model reduction results in a significant enhanced benefit achieving a local minimum of reduced cost function optimization almost identical to the one obtained by the high fidelity full 4-D Var model. Hence we achieve a double benefit while running a reduced-order inversion at an acceptable computational cost, at least for the shallow-water equations model in a two-dimensional spatial domain. Therefore, the advantage of the dual weighted TRPOD can be viewed as either the economization of the full 4-D Var without sacrificing the global convergence or the feasibility of implementation of optimal control of a large dynamical models based on a relatively lower dimensional POD control space.

In particular we observed that a similar reduction in cost functional and RMSE could be obtained using POD 4-D Var method, such as the dual weighted TRPOD compared to the full 4-D Var, but at a significantly less computational effort and reduced storage requirements (about 1/3 CPU-time less compared to full 4-D Var). These results indicate a potential for huge benefits within operational 4-D Var data assimilation systems with state of the art numerical weather prediction models. In order to obtain a drastic speed up of CPU-time by



(a) Scaled cost functional

(b) Scaled norm of the gradient

Figure 8.15: DAS-III(d): 2.5×2.5 Resolution with incomplete observations for u and v wind fields from 20^o north to north pole and 20^o south to south pole and complete observations for geopotential field, over entire globe. Comparison of the performance of the iterative minimization process of the scaled cost functional and the scaled norm of the gradient of the cost functional for unweighted trust-region POD 4-D Var and full 4-D Var.

at least an order of magnitude, we plan to explore implementation of DEIM to exploit the full potential of the POD reduced order model in the framework of dual weighted TRPOD in our future research work.

CHAPTER 9

SUMMARY AND FUTURE WORK

The ad-hoc POD 4-D Var lacks the ability to incorporate the information from the data assimilation system except for small intervals of time surrounding the initial window where the original snapshots were taken. Frequent projections from the reduced order POD model to the high fidelity model are required in order to retain the validity of the control. This is the motivation for applying the dual weighted POD basis to absorb the information from the optimality system of the 4-D Var. We also exploit the property of the global convergence of state-of-the-art trust-region POD adaptivity based on the ratio between predicted reduction of the POD reduced-order model and actual reduction of the high-fidelity model. Finally, we proposed a new methodology combining the dual weighted snapshots and trust region POD adaptivity, allowing us to enhance the benefits already provided by using dual weighted POD 4-D Var.

All of the above POD bases are generated by taking into account the controllability of DS and ignoring the observability of DS, however, some state variables that require little energy to control may require more energy to observe. On the other hand, the method of Balanced truncation (BMT) truncates the least controllable/observable states that have little effect on the input/output behavior. The central concept of the BMT is to find a ROM of high-fidelity linear DS such that the degree of reachability and the degree of observability of each state are the same.

In order to extend the classical linear BMT to the nonlinear DS, one can either compute the empirical observability grammian of the original system or the controllability grammian of the adjoint system. We have proposed a way to apply the method of snapshots to adjoint system and generate so-called adjoint POD modes, with which we can approximate the controllability grammian of the adjoint system instead of solving the computationally

expensive coupled Lyapunov equations explicitly.

We also discussed the possibility to extend the snapshots based POD methodology to the nonlinear system. Furthermore, we modify the classical algorithms in order to save the CPU time even more significantly. We proposed a novel idea to construct an ensemble of snapshots by integrating the tangent linear model (TLM) only once, based on which we can obtain its TLM POD modes. Then each TLM POD mode will be used as an initial condition to generate a small ensemble of adjoint snapshots and their adjoint POD modes. Finally, we can construct a large ensemble of adjoint POD modes by putting together each small ensemble of adjoint POD modes. In the incremental POD 4-D Var, we can approximate the controllability Grammian by integrating TLM only once and approximate observability Gramian by integrating adjoint model only a reduced number of times.

There have been several trends of the development in research recently on ROM by coupling POD with balanced truncated ROM, Krylov ROM, statistical inverse problems, and some control oriented method.

A dimension reduction method called Discrete Empirical Interpolation (DEIM) [106, 107, 108] was proposed and shown to dramatically reduce the computational complexity of the popular POD method for constructing reduced-order models for unsteady and/or parametrized nonlinear partial differential equations (PDEs). Michael Hinze and Martin Kunkel 2010 [132] investigated POD-based model order reduction for semiconductors in electrical networks using DEIM to treat the reduction of nonlinear components. In the context of structural problems involving plasticity or damage, strong topological changes in the structure might occur, and the initial snapshots might be too poor to represent accurately the solution of the damaged structure. John R. Singler and Belinda A. Batten 2010 [131] presented a Balanced POD algorithm for robust control design for linear distributed parameter systems to compute the nonstandard features of this robust control law.

Carlberg and Farhat 2009 [133] presented an adaptive POD-Krylov reduced-order modeling technique. They computed approximations to the structural state and sensitivities that are contained in the sum of POD and Krylov subspaces. Kerfriden et al., 2010 [134] developed a bridge between POD-based model order reduction techniques and the classical Newton-Krylov solvers to derive the POD-Krylov projection strategy and obtain an efficient solution procedure for highly nonlinear problems undergoing strong topological changes.

Imran Akhtar and Jeff Borggaard et al. 2010 [135] presented a approach called "control-

then-reduce" to control the von Karman vortex street. This approach ensures the feedback functional gains are well represented in the reduced basis functions.

Lieberman, Willcox and Ghattas [136] proposed a greedy algorithm for the construction of a reduced model with reduction in both parameter and state is developed for efficient solution of costly large-scale statistical inverse problems governed by partial differential equations with distributed parameters.

In future research work, I would like to consider the combination of the balanced truncation technique with the dual weighted trust-region POD reduced-oder 4-D Var and its generalization to nonlinear DS with real observations. Another research direction consists in exploring Discrete Empirical Interpolation Method (DEIM) proposed by [106, 107, 108] and application of DEIM to our FVSW model with realistic initial conditions. DEIM achieves a complexity reduction of the nonlinearities which is proportional to the number of reduced variables while POD retains a complexity proportional to the original number of variables. The DEIM approach approximates a nonlinear function by combining projection with interpolation. DEIM constructs specially selected interpolation indices that specify an interpolation-based projection so as to provide a nearly l_2 optimal subspace approximation to the nonlinear term, without the expense of orthogonal projection. Finally, it is also possible to expand the coefficient in the generalized polynomial chaos expansion (GPCE) in terms of POD basis based on GPCE POD stochastic Galerkin projection.

REFERENCES

[1] Mullis CT, Roberts RA. Synthesis of minimum roundoff noise fixed point digital filters. *IEEE Transactions on Circuits and Systems*, 1976; **CAS**(23):551-562. 1

[2] Moore BC. Principal component analysis in linear system: controllability, observability and model reduction. *IEEE Transactions on Automatic Control*, 1981; **AC**(26):551-562. 1, 4.7.4

[3] Pernebo L, Silverman LM. Model reduction with balanced realizations: An error bound and a frequency weighted generalization. *Proceedings of the 23rd IEEE Conference on Decision and Control*, 1982; **AC**(27):382-382. 1

[4] Enns D. Model reduction via balanced state space representation. *IEEE Transactions on Automatic Control*, 1984; Las Vegas, Nevada, USA. 1

[5] Desai IB, Pal D. A transformation approach to stochastic model reduction. *IEEE Transactions on Automatic Control*, 1984; **AC**(29). 1

[6] Green M. A relative error bound for balanced stochastic truncation. *IEEE Transactions on Automatic Control*, 1988; **AC**(33):961-965. 1

[7] Green M. Balanced stochastic realizations. *Journal of Linear Algebra and its Applications*, 1988; **98**:211-247. 1

[8] Opdenacker PC, Jonckheere EA. A contraction mapping preserving balanced reduction scheme and its infinity norm error bounds. *IEEE Transactions on Circuits and Systems*, 1988; **CAS**(35):184-189. 1

[9] Jonckheere EA, Silverman LM. A new of invariants for linear system-application to reduced order compensator design *IEEE Transactions on Automatic Control*, 1983; **AC**(28):953-964. 1

[10] Wang G, Sreeram V, Liu WQ. A new frequency weighted balanced truncation method and an error bound. *IEEE Transactions on Automatic Control*, 1999; **E**(44):1734-1737. 1

[11] Wang G, Sreeram V, Liu WQ. Frequency-weighted L1 norm and optimal Hankel norm model reduction. *IEEE Transactions on Automatic Control*, 1995; **T**(40):1687-1699. 1

[12] Gugercin S, Antoulas A. A survey of model reduction by balanced truncation and some new results. *International Journal of Control*, 2004; **77**:748-766. 1

[13] k. Glover. Optimal Hankel-norm approximations of linear multivariable systems and their error bounds. *International Journal of Control*, 1984; **39**(6):115-1936. 1

[14] Karhunen K. *Zur Spektraltheorie stochastischer Prozesse.* Ann. Acad. Sci: Fenicae,1946 1, 7.1

[15] Loeve M. *Fonctions aleatoires de second ordre.* C. R. Acad. Sci: Paris,1945 1, 7.1

[16] Kosambi DD. Statistics in function space. *J. Indian Math. Soc.*, 1943; **7**(1):76-88. 1, 7.1

[17] Pearson K. On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine*, 1901; **2**(1):559-572. 1

[18] Hotelling H. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 1933; **24**(1):417-441,498-520. 1

[19] Lumley JL. The structure of inhomogeneous turbulent flows. *Atmos. Turbu. Radio Wave Propagat.*, 1967; Yaglom AM and Tatarski VI eds. Moscow. Nauka : 166-178. 1, 7.1

[20] Berkooz G, Holmes PJ Lumley JL. The proper orthogonal decomposition in the analysis of turbulent flows. *Annual Review of Fluid Mechanics*,1993; **25**(1) :539–575 1, 7.1

[21] Sirovich L, Lumley JL, Berkooz G. Turbulence and the dynamics of coherent structures, part III: dynamics and scaling. *Quarterly of Applied Mathematics*, 1987; **45**(3):583-590. 1, 7.1

[22] Holmes P, Lumley JL, Berkooz G. *Turbulence, Coherent Structures, Dynamical Systems and Symmetry.* Cambridge Monographs on Mechanics: London, 1996. 1, 7.1, 7.1

[23] Michael K. *Geometrical Data Analysis: An empirical approach to dimensionality Reduction and the study of patterns.*(1st edn). Wiley-Interscience, 2000. 1

[24] Wiener N The Homogeneous Chaos *American Journal of Mathematics*, 1938; **60**(4):897-936. 1

[25] Xiu DB and Karniadakis GE The Wiener-Askey polynomial chaos for stochastic differential equations. *SIAM J. Sci Comput*, 2002; **24**(2):619-644. 1

[26] Noack BR, Afanasiev, Morzynski M, Tadmor G, Thiele F. A hierarchy of low-dimensional models for the transient and posttransient cylinder wake. *J. Fluid Mech*, 2003; **497**(d):335-363. 1, 4.7

[27] Noack BR, Schlegel M, Morzynski M, Tadmor G. System reduction strategy for Galerkin models of fluid flows. *Int. J. Numer. Meth. Fluids*, 2010; **63**(d):231-248. 1, 4.7

[28] Aubry N, Lian WY, Titi ES, Preserving symmetries in the proper orthogonal decomposition. *SIAM Journal on Scientific Computing*,1993; **14**(2):483-505. 1, 4.7

[29] Armbruster D, Heiland R, Kostelich EJ, and Nicolaenko B, Phase-space analysis of bursting behavior in Kolmogorov flow. *Physica D, 1992*; **58**:392-401. 1, 4.7

[30] Crommelin DT, Majda AJ. Strategies for Model Reduction: Comparing Different Optimal Bases. *J. Atmos. Sci*, 2004; **61**:2206-2217. 1, 4.7

[31] Majda AJ, Wang XM. Nonlinear Dynamics and Statistical Theories for Basic Geophysical Flows. Cambridge University Press, 2006; 471-472. 4.7

[32] DelSole T. Optimally Persistent Patterns in Time-Varying Fields. *J. Atmos. Sci*, 2004; **58**:1341-1356. 1, 4.7

[33] Hasselmann K. PIPs and POPs: The reduction of complex dynamical systems using principal interaction and oscillation patterns. *J. Geophys. Res.*, 1988 **93**:11015-11021. 1, 4.7, 4.7.2

[34] Charney JG, DeVore JG. Multiple flow equilibria in the atmosphere and blocking. *J. Atmos. Sci*, 1979; **36**:1205-1216. 1, 4.7

[35] Kwasniok, F. The reduction of complex dynamical systems using principal interaction patterns. *Physica D*, 1996; **92**:28-60. 4.7.2, 4.7.2

[36] Kwasniok, F. Empirical low-order models of barotropic flow. *J. Atmos. Sci*, 1996; **61**:235245. 4.7.2

[37] Galletti, B., Bruneau, C.H., Zannetti, L., Iollo, A. Low-order modelling of laminar flow regimes past a confined square cylinder. *Journal of Fluid Mechanics*, 2004; **503**:161170. 1, 4.7, 4.7.3

[38] Xia M, George EK. A low-dimensional model for simulating three-dimensional cylinder flow. textitJ. Fluid Mech, 2002; **358**:181190. 4.7.3

[39] Couplet, M., Basdevant, C., Sagaut, P. Calibrated reduced-order POD-Galerkin system for fluid flow modelling. *Journal of Computational Physics* , 2005; **207**:192220. 1, 4.7

[40] Gloerfelt, X., 2006. Compressible POD/Galerkin reduced-order model of self- sustained oscillations in a cavity. In: 12th AIAA/CEAS Aeroacoustics Conference (27th AIAA Aeroacoustics Conference), Cambridge, Massachusetts. 1, 4.7

[41] Galletti, B., Bruneau, C.H., Zannetti, L., Iollo, A., 2005. Accurate model reduction of transient flows. Technical Report, 5676, INRIA. 4.7

[42] Pastoor, M., Hennning, M., Noack, B.R., King, R., Tadmor, G., 2008. Feedback shear layer control for bluff body drag reduction. Journal of Fluid Mechanics. doi:10.1017/S0022112008002073. 4.7

[43] Rathinam, M, Petzold, LR. A new look at proper orthogonal decomposition. *SIAM J. Numer. Anal*, 2003; **41**(5):18931925. 4.7.4

[44] Colonius, T, Freund, JB. POD analysis of sound generation by a turbulent jet. *AIAA Paper*, 2002; **41**(5):18931925. 4.7.4

[45] Lall S. Marsden JE, Glavaski S. A subspace approach to balanced truncation for model reduction of nonlinear control systems. *International Journal of Robust and Nonlinear Control*, 2002; **12**:519-535. 1, 4.7.4

[46] Rowley CW. Model reduction for fluids using balanced proper orthogonal decomposition. *International Journal of Bifurcation and Chaos*, 2005; **15**(3):997-1013 1

[47] Ilak M. Model reduction and feedback control of transitional channel flow, 2009; Ph.D. dissertation advised by Clarence W. Rowley 4.7.4

[48] C. Lanczos. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *J. Res. Nat. Bureau Stan.*, 1950; **45**:255-282. 1

[49] WE. Arnoldi. The principle of minimized iterations in the solution of the matrix eigenvalue problem. *Quart. Appl. Math.*, 1951; **9**:17-29. 1

[50] Antoulas A., Sorensen C. Approximation of large-scale dynamical systems: an overview. *Int. J. Apppl. Math. Comput.*, 2001; **111**(5):1093-1121. 1

[51] Lehoucq, RB, Sorensen DC, Yang, C. ARPACK Users Guide: Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods. Philadelphia: *SIAM*. ISBN 978-0898714074, 1998. 1

[52] Afanasiev K and Hinze M. Adaptive control of a wake flow using proper orthogonal decomposition. Preprint, 1999. 1

[53] Hinze M and K. Kunish. Three Control Methods for Time-Dependent Fluid Flow. *Flow, Turbulence and Combustion*, 2000; **65**: 273-298. 1

[54] Kunisch K, Volkwein S. Galerkin proper orthogonal decomposition methods for parabolic problems. *Numerische Mathematik*, 2001; **90**(1):117-148. 1, 5.3

[55] Kunisch K, Volkwein, S. Galerkin Proper Orthogonal Decomposition Methods for a General Equation in Fluid Dynamics. *SIAM Journal on Numerical Analysis*, 2002; **40**(2):492-515. 1, 5.3

[56] Kunish K, Volkwein S. Proper orthogonal decomposition for optimality systems. *Mathematical modelling and numerical analysis*,2008; **42**(1) :1-23. 1, 5.2

[57] Cao Y, Zhu J, Luo Z, Navon IM. Reduced order modeling of the Upper Tropical Pacific ocean model using proper orthogonal decomposition. *Computers and Mathematics with Applications*, 2006 **52**(8-9):1373-1386. 7

[58] Cao Y, Zhu J, Navon IM, Luo Z. A reduced-order approach to four-dimensional variational data assimilation using proper orthogonal decomposition. *International Journal for Numerical Methods in Fluids*, 2007; **53**(10):1571-1583. 1, 7, 8.4.2

[59] Luo Z, Zhu J.,Wang R.,Navon IM. Proper Orthogonal Decomposition Approach and Error estimation of Mixed finite Element Methods for the tropical Pacific Ocean Reduced Gravity Model. *Computer Methods in Applied Mechanics and Engineering*, 2007; **196**(41-44):4184-4195.

[60] Luo Z, Zhou Y, Yang XZ. A reduced finite element formulation based on proper orthogonal decomposition for Burgers equation. *Applied Numerical Mathematics*, 2009; **59**(8):1933-1946. 1

[61] Luo Z, Zhou Y, Yang XZ. Du J, Zhu J, Luo Z. Navon I. M. An optimizing finite difference scheme based on proper orthogonal decomposition for CVD equations. *Int. J. Numer. Meth. Biomed. Engng*, 2011; **27**:78-94. 1

[62] Fang F, Pain CC, Piggott MD, Gorman GJ, Goddard, AJH. An adaptive mesh adjoint data assimilation method applied to free surface flows. *International Journal for Numerical Methods in Fluids*, 2005; **47**(8-9):995-1001. 1, 7

[63] Fang F, Piggott MD, Pain CC, Gorman GJ, Goddard, AJH. Adjoint data assimilation into a 3D unstructured mesh coastal finite element model. *Ocean Modeling*, 2006; **15**(1-2):3-38. 1, 7

[64] Fang F, Pain CC, Navon IM, Piggott MD, Gorman GJ, Allison P, Goddard, AJH. Reduced order modelling of an adaptive mesh ocean model. *International Journal for Numerical Methods in Fluids*, 2008; **59**(8):827-851. 1, 7

[65] Fang F, Pain CC, Navon IM, Piggott MD, Gorman GJ, Allison P, Goddard, AJH. A POD reduced order 4D-Var adaptive mesh ocean modeling approach. *International Journal for Numerical Methods in Fluids*, 2009; **60**(7): 709-732. 1, 7, 8.4.2

[66] Pain CC, Fang F, Navon IM, Cacuci DG. **Chen X**. The Independent Set Perturbation Adjoint method applied to a finite-element shallow-water equation model. Submitted for publication to *Computer Methods in Applied Mechanics and Engineering*, 2011. 1

[67] Vermeulen PTM, Heemink AW. Model-Reduced Variational Data Assimilation. *Monthly Weather Review*, 2006; **134**(10):2888-2899. 1, 5.3, 5.3

[68] Sachs EW, Schu M. Reduced order models(POD) for calibration problems in finance. *Numerical Mathematics and Advanced Applications*, ENUMATH, 2007; 735-742. 1

[69] Altaf MU, Heemink AW, Verlaan M. Inverse shallow water flow modeling using model reduction. *International Journal for Multi-scale Computational Engineering* ,2009; Submitted. 1, 7

[70] Nocedal J, Wright SJ. *Numerical Optimization*(2nd edn). Springer Series in Operations Research and Financial Engineering, 2006. 1, 5.4.1, 5.4.1, 7, 7.4, 8.4.2

[71] Conn AR, Gould NIM, Toint PL. *Trust-region methods*. SIAM: Philadelphia, 2000. 1, 7

[72] Levenberg K A Method For The Solution Of Certain Problems In Least Squares *Quarterly Journal on Applied Mathematics*,1944; **2** :164-168 1, 5.4.1

[73] Morrison DD *Proceedings of the Seminar on Tracking Programs and Orbit Determination.* Jet Propulsion Laboratory: USA, 1960 1, 5.4.1

[74] Marquardt D An Algorithm For Least-Squares Estimation Of Nonlinear Parameters *SIAM Journal on Applied Mathematics*,1963; **11** :431-441 1, 5.4.1

[75] Winfield D *Function and functional optimization by interpolation in data tables* Cambridge: USA, 1969. 1, 5.4.1

[76] Powell MJD A Fortran subroutine for solving systems of nonlinear algebraic equations, 1970d: 115-161 1, 5.4.1, 5.4.1, 5.4.1

[77] Powell MJD A Fortran Subroutine for Unconstrained Minimization Requiring First Derivatives of The Objective Function, 1970b: R-6469 1, 5.4.1, 5.4.1, 5.4.1

[78] Dennis JE *Numerical Analysis.* Proceedings of Symposia in Applied Mathematics series: 1978. 1, 5.4.1

[79] More JJ, Sorensen DC Computing A Trust Region Step *SISSC*,1983; **4**(3) :553-572 1, 5.4.1

[80] Celis M, Dennis JE, Tapia RA. A trust region strategy for nonlinear equality constrained optimization *Numerical Optimization 1984 (P. Boggs, R. Byrd and R. Schnabel, eds)* SIAM :Philadelphia, 1985, pp. 71-82. 5.4.1

[81] Fahl M. Trust-region methods for flow control based on Reduced Order Modeling. *Ph.D. thesis, Trier university*, 2000; 1, 7

[82] Arian E, Fahl M, Sachs EW. Trust-region proper orthogonal decomposition for flow control. *Institute for Computer Applications in Science and Engineering*,Technical Report: TR-2000-25, 2000; 1, 5.4.2, 7

[83] Toint PL. Global convergence of a class of trust-region methods for nonconvex minimization in Hilbert space. *IMA Journal of Numerical Analysis*, 1988; **8**(2):231-252. 1, 5.4.1, 5.4.1, 5.4.2, 7, 8

[84] Carter RG. On the global convergence of trust region algorithms using inexact gradient information. *SIAM J. Numer. Anal.*, 1988; **28**(1):251-265. 5.4.1

[85] Gratton S, Sarttenaer, Toint PL. Recursive trust-region methods for multi-scale nonlinear optimization. *SIAM J. on Optimization*, 2008; **19**(1):414-444. 1, 8

[86] Tan WY *Shallow-water hydrodynamics: mathematical theory and numerical solution for a two-dimensional system of shallow-water equations.* Elsevier oceanography series: Beijing, 1992. 8

[87] Vreugdenhil CB *Numerical methods for shallow-water flow.* Kluwer Academic Publishers: Boston, 1994. 8

[88] Galewsky J Cott RK, Polvani LM. An initial-value problem for testing numerical models of the global shallow-water equations. *Tellus*, 2004; **56**(5):429-440. 8

[89] Lin SJ, Chao WC, Sud YC, Walker GK. A class of the van Leer transport schemes and its applications to the moisture transport in a general circulation model. *Monthly Weather Review*, 1994; **122**: 1575-1593. (document), 8

[90] Lin SJ, Rood RB. Multidimensional flux-form semi-Lagrangian transport schemes. *Monthly Weather Review*, 1996; **124**: 2046-2070. (document), 8

[91] Lin SJ, Rood RB. An explicit flux-form semi-Lagrangian shallow-water model on the sphere, 1997; *Q. J. R. Meteorol. Soc*, 1997; **123**: 2477-2498. (document), 8, 8.2, 8.2

[92] Lin SJ, Rood RB. A finite-Volume integration method for computing pressure gradient force in general vertical coordinates, *Q. J. R. Meteorological Society*, 1997; **123**:1749-1762. (document), 8, 8.1, 8.2, 8.2

[93] Van Leer B. The quest for monotonicity, *Springer Lecture Notes in Physics*, 1973; 163-198. 8.1

[94] Van Leer B. Towards the ultimate conservative difference scheme: A new approach to numerical convection, *Journal of Computational Physics*, 1977; **23**:276-299. 8.1

[95] Van Leer B. Towards the ultimate conservative difference scheme: A second order sequel to Godunov's method, *Journal of Computational Physics*, 1977; **32**:101-136. 8.1

[96] Lin SJ. A "vertically Lagrangian" finite-volume dynamical core for global models, 2004; *Monthly Weather Review*, 2004; **132**: 2293-2307. (document), 8

[97] Chen X, Navon IM. Optimal Control of a Finite-Element Limited-Area Shallow-Water Equations Model. *Studies in Informatics and Control*, 2011; **18**(1):41-62. 8

[98] Chen X, Navon IM, Fang F. A dual-weighted trust-region adaptive POD 4D-Var applied to a finite-element shallow-water equations model, *Int. J. Numer. Meth. Fluids*, 2011; **65**(5):520-541. 8

[99] Chen X, Akella S, Navon IM. A dual weighted trust-region adaptive POD 4D-Var applied to a Finite-Volume Shallow-Water Equations Model on the sphere. In Early View in *Int. J. Numer. Meth. Fluids*(DOI: 10.1002/fld.2523), 2011. 8

[100] Bui-Thanh T, Willcox K, Ghattas O, Van Bloemen Waander B. Goal-oriented model constrained for reduction of large-scale systems. *Journal of Computational Physics*, 2007 **224**(2):880-896. 7, 8

[101] Daescu, DN, Navon, IM. Efficiency of a POD-based reduced second-order adjoint model in 4D-Var data assimilation. *International Journal for Numerical Methods in Fluids*, 2007; **53**(10):985-1004. 7.4, 8, 8.4.2

[102] Daescu, DN, Navon, IM. A dual weighted approach to order reduction in 4DVAR data assimilation. *Monthly Weather Review*, 2008; **136**(3):1026-1041. 1, 5.2, 7, 7.4, 8, 8.4.2

[103] Yaremchuk M, Nechaev D, Panteleev A Method of Successive Corrections of the Control Subspace in the Reduced-Order Variational Data Assimilation. *Monthly Weather Review*, 2009; **137**):2966-2978. 1

[104] Vahid E. Khosro A. Equation-Free/Galerkin-Free Reduced-Order Modeling of the Shallow Water Equations Based on Proper Orthogonal Decomposition. *Journal of fluids engineering*, 2009; **131**:071401-1-071401-13. 1

[105] Bannister R.N., A review of forecast error covariance statistics in atmospheric variational data assimilation. I: Characteristics and measurements of forecast error covariances., Quarterly Journal of the Royal Meteorological Society, 2008; **134**, 1951-1970. 8.4.2

[106] Chaturantabut, S. Dimension Reduction for Unsteady Nonlinear Partial Differential Equations via Empirical Interpolation Methods. *Master of Arts Thesis*, 2008; Rice University. 8.4.3, 9

[107] Chaturantabut S, Sorensen DC, Steven JC. Nonlinear model reduction via discrete empirical interpolation. *SIAM J. Sci. Comput*, 2010; **32**(5):2737-2764. 8.4.3, 9

[108] Anthony RK, Chaturantabut S, Sorensen DC. Morphologically accurate reduced order modeling of spiking neurons. *J Comput. Neurosci*, 2010; **28**:477-494. 8.4.3, 9

[109] Navon IM, Zou X, Derber J, Sela J. Variational data assimilation with an adiabatic version of the NMC spectral mode. *Monthly Weather Review*, 1992; **120**(7):1435-1443 5.2

[110] Bergmann M, Cordier L. Optimal control of the cylinder wake in the laminar regime by trust-region methods and pod reduced-order models. *Journal of Computational Physics*, 2008; **227**(16):7813-7840 5.4.2

[111] Bergmann M, Bruneau C, Iollo A. Enablers for robust pod models. *Journal of Computational Physics*, 2009; **228**(2):516-538 5.4.2, 5.4.2

[112] Barone MF, Kalashnikova I, Segalman DJ, Thornquist HK. Stable Galerkin reduced order models for linearized compressible flow *Journal of Computational Physics*, 2009; **228**(6):1932-1946 5.4.2, 7.1

[113] Rowley CW, Colonius T, Murray RM. Model reduction for compressible flows using POD and Galerkin projection. *Physica D*, 2004; **189**(1-2):115-129 7.1

[114] Aquino W Brigham JC, Earls CJ, Sukumak N. Generalized Finite Element Method using Proper Orthogonal Decomposition. *International Journal for Numerical Methods in Engineering*, 2009; In press. 4

[115] Grammeltvedt A. A survey of finite-difference schemes for the primitive equations for a barotropic fluid *Monthly Weather Review*, 1969; **97**(5):384-404 7.4

[116] Zienkiewicz OC, Taylor RL, Zhu JZ, Nithiarasu P. *The Finite Element Method for a barotropic fluid* Butterworth-Heinemann: 2005 7.4

[117] Wang HH, Halpern P, Douglas J, Dupont I. Numerical solutions of the one-dimensional primitive equations using Galerkin approximation with localized basic functions *Monthly Weather Review*, 1972; **100**(10):738-746 7.4

[118] Zhu K, Navon IM, Zou X. Variational data assimilation with a variable resolution finite-element shallow-water equations model *Monthly Weather Review*, 1994; **122**(5):946-965 6, 7.4

[119] Payne NA, Irons BM. Private communication to O. Zienkiewicz,1963 7.4

[120] Huebner KH, Dewhirst DL, Smith DE, Byrom TG. *The finite-element method for engineers* John Wiley & Sons: 2001 7.4

[121] Navon IM. Finite-element simulation of the shallow-water equations model on a limited area domain. *Appl. Math. Model*, 1979; **3**(1) :337-348 7.4

[122] Navon IM. FEUDX: a two-stage, high-accuracy, finite-element FORTRAN program for solving shallow-water equations. *Computers and Geosciences*, 1987; **13**(3) :255-285 7.4

[123] Polak E, Ribiere, G. Note sur la convergence de directions conjugues. *Rev. Francaise Informat. Recherche Operationelle*, 1969; **3**(16):35-43 7.4

[124] Akella S. Deterministic and Stochastic Aspects of Data Assimilation, PhD Thesis, 2006, URL: http://etd.lib.fsu.edu/theses/available/etd-04052006-192927/ 8.2, 8.3, 8.4.1

[125] Liu DC, Nocedal, A Limited Memory Algorithm for Bound Constrained Optimization, *Math Prog*, 1989; **45**:503-528. 8.4.2

[126] Byrd RH, Lu P, Nocedal J. On the limited memory BFGS method for large scale minimization, *SIAM Journal on Scientific and Statistical Computing*, 1995; **16**(5):1190-1208. 8.4.2

[127] Zhu C, Byrd RH, Nocedal J. L-BFGS-B: Algorithm 778: L-BFGS-B, FORTRAN routines for large scale bound constrained optimization, *ACM Transactions on Mathematical Software*, 1997; **23**(4): 550-560. 8.4.2

[128] Zou X, Navon, IM, Le Dimet FX. Incomplete observations and control of gravity waves in variational data assimilation. *Tellus*, 1992; **44**(A):273-296. 8.5.1, 8.5.2

[129] Derber, JC and Bouttier, F. A reformulation of the background error covariance in the ECMWF global data assimilation. Tellus, 1999; **51A**: 195– 221. 8.3

[130] Weaver A and Courtier P. Correlation modeling on the sphere using a generalized diffusion equation. Q. J. R. Met Soc., 2001; **121**: 1815-1846. 8.3

[131] John R. Singler and Belinda A. Batten *Balanced POD Algorithm for Robust Control Design for Linear Distributed Parameter Systems*. **Proceedings of American Control Conference**, 2010, 4881-4886 9

[132] M. Hinze, M. Kunkel. Discrete empirical interpolation in POD Model Order Reduction of Drift-Diffusion Equations in Electrical Networks. *SCEE Proceedings*, 2010, Toulouse. 9

[133] Kevin Carlberg, Charbel Farhat. An Adaptive POD-Krylov Reduced-Order Model for Structural Optimization. *8th World Congress on Structural and Multidisciplinary Optimization*, 2009. 9

[134] P. Kerfriden, P. Gosselet, S. Adhikari, S. Bordas, J.C. Passieux. POD-based model order reduction for the simulation of strong nonlinear evolutions in structures: application to damage propagation. *Materials Science and Engineering*, 2010, doi:10.1088/1757-899X/10/1/012165. 9

[135] Imran Akhtar, Jeff Borggaard. On Commutation of Reduction and Control: Linear Feedback Control of Von Karman Street. *5th Flow Conrol Conference*, 2010, Chicago, Illinois AIAA. 9

[136] Chad Lieberman, Karen Willcox, and Omar Ghattas. Parameter and state model reduction for large-scale statitical inverse problems, 2010. *SIAM Journal on Scientific Computing*,32(5) 2523-2542. 9

BIOGRAPHICAL SKETCH

Xiao Chen

I was born at Changsha, Hunan province in 1982. I have completed my M.S. degree in Applied Mathematics at Zhejiang University, Hangzhou, China in June, 2006. Currently, I am enrolled as a doctoral student in the department of mathematics at Florida State University, USA.

As a Ph.D. candidate, I have published three papers in some of the top journals of my field, and more papers are in progress for publication. I have also presented my research at many international conferences and attended many workshops. Recently, I have been invited to attend the Joint Mathematics meeting at New Orleans in January, 2011.

My Ph.D. dissertation, advised by Prof. I. M. Navon, is focused on Variational 4-D Data-Assimilation and model reduction for dynamical system. I have a strong background in mathematical modeling as evident from the fact that I received first prize in an undergraduate mathematical modeling contest in China. In addition, I have excellent programming skills to write well-organized codes for complex fluid dynamics models as well as their adjoint models.

I have been teaching a variety of math classes at Florida State University. My experience as an instructor has contributed to render me a better researcher since good collaboration requires effective communication.

I expect to satisfy the requirements of my doctoral degree in Applied and Computational Mathematics in April, 2011.