THE FLORIDA STATE UNIVERSITY

COLLEGE OF ARTS AND SCIENCES

# PRACTICAL OPTIMIZATION ALGORITHMS IN THE DATA ASSIMILATION OF LARGE-SCALE SYSTEMS WITH NON-LINEAR AND NON-SMOOTH OBSERVATION OPERATORS

By

JEFFREY L. STEWARD

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

Degree Awarded:
Spring Semester, 2012

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

Jeffrey L. Steward defended this dissertation on November 21, 2011.

The members of the supervisory committee were:

Ionel Michael Navon
Professor Directing Thesis

Guosheng Liu
University Representative

Max Gunzburger
Committee Member

Gordon Erlebacher
Committee Member

Milijia Zupanski
Committee Member

Napsu Karmitsa
Committee Member

The Graduate School has verified and approved the above-named committee members, and certifies that the dissertation has been approved in accordance with the university requirements.

# ACKNOWLEDGMENTS

First and foremost, I would like to acknowledge my wonderful and brilliant advisor, Professor Ionel Michael Navon, for his invaluable guidance and support throughout my graduate program. From the beginning of my graduate studies until the end, Professor Navon was indispensably helpful at each stage, giving me precisely the advice, guidance and support that I needed at each juncture. It is truly rare to find an advisor who legitimately cares for his students' well-being as much as Professor Navon, and I feel very blessed to have had the good fortune of working with this talented scientist.

Second of all, the Florida State University has a new and wonderful program in Computational Science, and I am incredibly grateful to Professor Navon, my committee, and everyone else in the DSC department including the excellent professors, administrative staff, and fellow students. My education throughout this interdisciplinary program was highly enjoyable and smooth thanks to the talents of teachers including Prof. Navon, Prof. Janet Peterson, Prof. Gordon Erlebacher, Prof. Max Gunzburger, Prof. David Kopriva, Prof. Kyle Gallivan, and Prof. Guosheng Liu. My interactions with Prof. Milija Zupanski, Dr. Emil Constantinescu, and Prof. Napsu Karmitsa have also been extremely valuable and rewarding.

Third, I would like to thank the National Science Foundation under the grant ATM-0931198 for their financial support in my studies. As a research assistant I was able to focus whole-heartedly on my studies without much of the rigmarole that graduate students are sometimes put through. I hope that this investment has paid dividends and will continue to do so. In my opinion, the greatness of a country and a society comes primarily through education, and thus I will do my best to repay this financial support that I have been given back to society.

Last but not least, I would like to thank my family and friends for their help and support. I definitely would not have been able to complete this program without the support and encouragement of my parents, friends, and loved one. Words cannot express my gratitude.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# ABSTRACT

This dissertation compares and contrasts three large-scale optimization algorithms in the use of variational and sequential data assimilation on two novel problems chosen to emphasize the challenges in non-linear and non-smooth data assimilation. The first problem explores the impact of a highly non-linear observation operator and highlights the importance of background information on the data assimilation problem. Using the RTTOV version 10 multiple-scattering radiative transfer model, the problem of the one-dimensional (column) data assimilation of all-sky infrared radiances is investigated. The presence of clouds creates a sharp transition between clear and cloudy regimes is a difficult case for data assimilation. Using six test cases of clear and cloudy skies with different size perturbations, a reasonable compromise method is found for initializing the background that allows both clear and cloudy skies to be assimilated. The second problem tackles large-scale data assimilation with a non-smooth observation operator using the shallow water equations. This problem is appropriate for handling a "kink" in the observation operator, i.e. a discontinuity in the first derivative or higher. Together, these two cases show both the importance of choosing an appropriate data assimilation method and, when a variational or variationally-inspired method is chosen, the importance of choosing the right optimization algorithm for the problem at hand.

# CHAPTER 1

# INTRODUCTION

While data assimilation – i.e. the integration of available observations of a physical system to improve the initial conditions, boundary conditions, or other model parameters – has been investigated right from the very advent of digital computers ([103]), there are still many open issues and challenges in the field. This dissertation research examines the overlap of data assimilation and optimization – i.e. the field of applied math that investigates algorithms to find extremal values of a cost or merit function. Using two medium-large size problems designed to highlight challenges in variational data assimilation, three different optimization algorithms appropriate for large-scale problems are compared and contrasted in order to investigate the suitability of the algorithms under study in both of these situations.

Other works have investigated non-smooth algorithms appropriate for very large problems for convex problems (e.g. [166]). The fully non-convex non-smooth variational data assimilation problem has also been investigated in [144], [90], [235], [101] and [12] on highly simplified problems. However, the existing optimization algorithms at the time those studies were conducted were not suitable for large-scale non-convex optimization [77]. The situation has now changed, however, due to both theoretical and algorithmic advances. An excellent comparison of non-smooth optimization algorithms for large-scale minimization is given in [106] showing positive results for problems with as many as 4000 variables. This thesis, then, is studying the applicability of these methods for highly non-linear and non-smooth problems with over 1000 variables. However, these methods are expected to be scalable to operational Numerical Weather Prediction (NWP) data assimilation scales, i.e. $10^6$ – $10^7$ variables.

The first application we consider is that of assimilating satellite radiant flux in the infrared spectrum within the atmosphere, which is a problem with a highly non-linear observation operator. Such observations are known as “all-sky” infrared radiances since these observations may or may not include the presence of clouds. An observation operator is a model – in this case the RTTOV version 10 radiative transfer model – of the mapping from the state of the model – in this case, temperature, pressure, water vapor, and cloud microphysical parameters – to the observations – in this case, radiant flux observed at the top of the atmosphere. In the infrared spectrum, there is a very steep transition between cloudy and clear regimes, as nearly all infrared energy is absorbed by clouds. The addition of these highly non-linear regime changes is challenging for the data assimilation problem and is explored here in the form of the assimilation of a single column of the atmosphere.

The final application is that of a theoretically even more difficult case – an observation operator with non-smooth “kinks.” Such kinks may arise in practice when different parameterization schemes are used or shifts between different regimes leave discontinuities in the first derivative of the observation operator. Using the shallow water equations – equations appropriate as a first-order approximation for the Earth's atmosphere – we study artificial non-smooth observation operators with increasingly sharp edges. In order to approach such problems, specialized tools from non-smooth optimization that are appropriate for large-scale problems are required. The differences and similarities between specially designed non-smooth optimization algorithms and traditional smooth optimization algorithms enlisted for non-smooth optimization will be highlighted.

This dissertation is organized as follows. Chapter 2 covers all of the optimization algorithms that will be used in this study. Chapter 3 introduces the data assimilation problem and the various approaches used here to solve it. Chapter 4 covers the highly non-linear cloudy IR assimilation case, and chapter 5 introduces the non-smooth observation operators problem with the shallow water equation model. Finally, chapter 6 relates the two test cases and makes conclusions regarding data assimilation in the presence of non-linear and/or non-smooth observation operators.

# CHAPTER 2

# OPTIMIZATION ALGORITHMS

In this dissertation, three optimization algorithms and their implementations are tested. These three methods are: the limited-memory BFGS (L-BFGS) implementation in the Harwell library ([136]), the non-linear conjugate gradient method CG-Descent of Hager and Zhang ([83]), and the limited-memory bundle method (LMBM) aimed at non-smooth optimization designed by Karmitsa et al. ([77], [78]).

The smooth optimization L-BFGS method, originally proposed by Nocedal in [161], has long been used in data assimilation (e.g. [242], [91]), and has recently been found to possess properties of a non-smooth optimization algorithm in [130], [129], and [185]. This method may offer promise for large-scale non-smooth optimal control problems, and these properties will be tested in this dissertation.

The non-linear conjugate gradient method CG-Descent was found in [5] and [7] to be among the most effective of conjugate gradient algorithms, and is thus used as a representative of the conjugate gradient family of optimization algorithms. While there is no theoretical or numerical basis to assume that this method will be suitable for non-smooth problems, it is included here nonetheless as a control of sorts.

The LMBM algorithm is a globally convergent non-smooth optimization algorithm specifically designed for large-scale, possibly non-convex minimization ([105], [78]). It is well-suited both numerically and theoretically for handling non-smooth problems, although its utility on highly non-linear realistic problems is, at present, unknown.

## 2.1 Non-Smooth Optimization

Since each of these methods will be used on a non-smooth problem, we develop each of the methods in the framework of non-smooth optimization. Before further discussion we introduce a few common definitions. In what follows we use an Euclidean norm, i.e. $||x|| = \left(\sum_{i=1}^{n} x_i^2\right)^{1/2}$.

**Definition 2.1.1.** A function $J : \mathbb{R}^n \to \mathbb{R}$ is *locally Lipschitz continuous* at $x \in \mathbb{R}^n$ with a constant $L > 0$ if there exists a positive number $\epsilon$ such that

$$|J(y) - J(z)| \leq L||y - z||$$

for all $y, z$ such that $||x - y|| \leq \epsilon$, $||x - z|| \leq \epsilon$.

Intuitively, $L$ is an upper limit of how fast the function changes at $x$ within the sphere of radius $\epsilon$. Note that the function $J$ itself cannot have a discontinuity, but the higher order derivatives may.

For a locally Lipschitz continuous function the classical directional derivative need not exist. However, we can generalize the concept of differentiability by defining a generalized directional derivative as follows.

**Definition 2.1.2.** Let $J : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz continuous function at a point $x \in \mathbb{R}^n$. The generalized directional derivative of $J$ at $x$ in the direction $p \in \mathbb{R}^n$ is defined by

$$J^o(x;p) = \limsup_{y \to x \text{ as } t \downarrow 0} \frac{J(y+tp) - J(y)}{t}$$

where $y \in \mathbb{R}^n$ and $t \in \mathbb{R}$.

Note that the only difference between this definition and the definition of a traditional directional derivative is the $\sup$, meaning that the largest directional derivative along any direction $y$ is taken. At a differentiable point, these limits will be the same along any direction; however, at a non-differentiable point, these values may be different, and choosing the largest is a choice of convenience that will be exploited later.

Unlike a traditional gradient, which is unique, at non-smooth points in general infinite sub-gradients exist as part of the *sub-differential* set, defined as follows:

**Definition 2.1.3.** Let $J : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz continuous function at a point $x \in \mathbb{R}^n$. Then the *sub-differential* of $J$ at $x$ is the set $\partial J(x)$ of vectors $\xi \in \mathbb{R}^n$ such that

$$\partial J(x) = \left\{ \xi \in \mathbb{R}^n | J^o(x;p) \geq \xi^{\mathrm{T}} p \text{ for all } p \in \mathbb{R}^n \right\}$$

Each vector $\xi \in \partial J(x)$ is called a *sub-gradient* of $J$ at $x$.

In analogy to the traditional interpretation of the gradient as a tangent hyperplane to the function $J$, intuitively, one can consider a sub-gradient at $x$ to be a normal vector of any tangent hyperplane that remains on or below the generalized directional derivatives $J^o(x;p)$ in all directions $p$.

When $J$ is differentiable at $x$, there is only one element of the sub-differential $\partial J(x)$, and it is the standard gradient. In the material that follows, the gradient of a variational cost function will be replaced with a sub-gradient in order to allow for, in general, non-smooth optimization. Because the gradient and the sub-gradient are identical at differentiable points, this change is primarily transparent to the model and adjoint development, although care must be taken to ensure that a value for the adjoint is chosen such that definition (2.1.3) holds at the discontinuities. In addition, the optimization algorithms themselves must take special care, as we will see. More details on the theory of non-smooth optimization can be found in the book of Makela and Neittaanmaki [144] and Bonnans et al. [31].

### 2.1.1 Survey

There are two main classes of non-smooth optimization methods: sub-gradient methods, and bundle methods ([144]). A sub-gradient method replaces the gradient from a traditional optimization algorithm with an arbitrary sub-gradient, while a bundle method replaces the function with a cutting-plane based on a bundle of sub-gradients.

The history of sub-gradient methods begins in the 1960s with scientists from the Soviet Union including N. Shor. The classic overview of these methods by Shor et al. can be found in [183]. Despite their simplicity and widespread adoption, one of the main handicaps of this family of methods is that no general convergence criteria can be established when the optimal point occurs at a non-smooth point. Smooth methods terminate based on the norm of the gradient becoming small, which is a necessary (but not sufficient) criteria for optimality (see e.g. [162]). However, non-smooth methods, which allow any arbitrary sub-gradient $\xi \in \partial J$, cannot rely upon the sub-gradient chosen at the optimal point going to zero; consider the classic example of $f(x) = |x|$, of which at the optimal point of $x = 0$ the sub-gradient of $\xi = 1$ is perfectly valid. In practice, while one can choose to terminate the optimization algorithm when changes in successive minimization algorithms are below some threshold, this practice does not have a mathematical basis behind it.

The approach which seems most promising for nonsmooth optimization is the bundle method, a survey of which can be found in [144], [143] and [77]. The basic approach of this method is to aggregate information from previous iterations into a *bundle* which gives information about the function. This method was pioneered by Lemarechal in [127] and Wolfe in [226]. Kiwiel reimagined bundle methods in terms of the cutting plane method in his book [110] by using the sub-gradients to form a piecewise linear approximation to the objective function. Both Lemarechal and Kiwiel's methods suffered from practical drawbacks (see e.g. [144]), and work on overcoming these limitations includes, among others [180] and [111]. The work of Vlcek and Luksan [212] for globally convergent line-search and the extension to large-scale problems by Karmitsa et al., e.g. [77], [78], [106] express the state-of-the-art in the field for the purposes of this dissertation.

In this work, we use two sub-gradient methods (the L-BFGS and CG-Descent methods) with no or only minor modifications to the line search. We also use LMBM, a bundle method as the name implies.

## 2.2 Optimization algorithms

In general, an iterative optimization algorithm can be formulated as

$$x_{k+1} = x_k + \alpha_k p_k \tag{2.1}$$

where $k$ is the iteration number, $p_k$ is the search direction and $\alpha_k$ is the step length. This procedure continues until some convergence criteria has been met.

If $J(x_{k+1}) \leq J(x_k)$ for all $k$, then an iterative method is called a *descent method* and the direction $p_k$ is called a *descent direction*. For smooth (continuously differentiable) objective functions, a descent direction may be generated by exploiting the fact that the direction opposite to the gradient is locally the direction of steepest descent. The step size $\alpha_k$ can then be determined, for example, by using a line search technique (see e.g. [162]). Furthermore, a necessary condition for local optimality is that the gradient goes to zero and by continuity becomes small on approach of an optimal point. This fact provides a useful stopping criterion for smooth iterative methods.

However, the direct application of smooth gradient-based methods to non-smooth problems may lead to a failure in convergence, in optimality conditions, or in gradient approximation. The usage of sub-gradients allows us to generalize well-developed gradient-based methods for non-smooth problems. In this section, we detail the optimization algorithms used in this study. An in-general non-unique sub-gradient $\nabla J(x_k) \in \partial J(x_k)$ at each iteration step $k$, used below, is defined to be $\nabla J_k$.

### 2.2.1 L-BFGS

In this work, we test an implementation of the limited-memory BFGS (Broyden-Fletcher-Goldfarb-Shanno) algorithm version VA15 of [136] in the Harwell library. The version of L-BFGS with sub-gradients is detailed in [252].

The L-BFGS method ([161]) is an adaptation of the BFGS method to large problems, achieved by changing the generalized Hessian update of the latter. The generalized Hessian is a matrix of second-order sub-derivatives generalized in analogy with the sub-gradient. The L-BFGS method uses an approximation $H_k$ to the inverse generalized Hessian which is updated at each time step (the generalized Hessian is a matrix of the second-order subderivatives). The search direction is found, in analogy to the Newton method, by

$$p_k = -H_k \nabla J_k \tag{2.2}$$

An inverse generalized Hessian approximation is updated at each iteration by

$$H_{k+1} = V_k^{\mathrm{T}} H_k V_k + \rho_k s_k s_k^{\mathrm{T}} \tag{2.3}$$

where $s_k = x_k - x_{k-1}$,

$$V_k = I - \rho_k y_k s_k^{\mathrm{T}} \tag{2.4}$$

$y_k = \nabla J_k - \nabla J_{k-1}$, and $\rho_k = 1/(y_k^{\mathrm{T}} s_k)$.

In the L-BFGS method, instead of forming the matrices $H_k$ explicitly (which would require a prohibitively large allocation of memory for even a medium-size problem), one only stores the vectors $s_k$ and $y_k$ obtained in the last $m$ iterations which define $H_k$ implicitly; a cyclical procedure is used to retain the latest vectors and discard the oldest ones. Thus, after the first $k$ iterations, equation (2.3) becomes

$$\begin{aligned} H_{k+1} &= \Pi_{k-\hat{m}}^{\mathrm{T}} H_0 \Pi_{k-\hat{m}} \\ &+ \rho_{k-\hat{m}} \Pi_{k-\hat{m}+1}^{\mathrm{T}} S_{k-\hat{m}} \Pi_{k-\hat{m}+1} \\ &+ \rho_{k-\hat{m}+1} \Pi_{k-\hat{m}+2}^{\mathrm{T}} S_{k-\hat{m}+1} \Pi_{k-\hat{m}+2} \\ &\quad \vdots \\ &+ \rho_k S_k \end{aligned} \tag{2.5}$$

where $\Pi_j = V_j V_{j+1} \cdots V_k$, $S_i = s_i s_i^{\mathrm{T}}$, $\hat{m} = \min(k, m-1)$, and the initial approximation $H_0$ is taken to be $I$.

Thus, only at most $2m$ correction pairs $s_i$ and $y_i$ for $i = 1, \ldots, m$ are needed, and no full matrix is ever stored in memory as equation (2.2) is solved. The only modification to the L-BFGS algorithm for this work is that we change the line search from strong to weak Wolfe conditions, as suggested in [130].

### 2.2.2 The nonlinear conjugate gradient algorithm CG-Descent

A generic conjugate gradient algorithm uses a step along the current negative gradient vector in the first iteration; successive directions are constructed so that they form a set of mutually conjugate vectors with respect to the Hessian. At each step, the new iterate is calculated from equation (2.1), and the search directions are expressed recursively as

$$p_k = -\nabla J_k + \beta_k p_{k-1} \tag{2.6}$$

with $p_0 = -\nabla J_0$.

Each variant of the conjugate gradient method chooses $\beta_k$ in a different manner. The method chosen here, the CG-Descent method by Hager and Zhang ([83]), choses $\beta_k$ by the relation:

$$\beta_k = \max\left(\bar{\beta}_k, \eta_k\right) \tag{2.7}$$

where

$$\bar{\beta}_k = \frac{1}{p_k^{\mathrm{T}} y_k} \left( y_k - 2p_k \frac{||y_k||^2}{p_k^{\mathrm{T}} y_k} \right)^{\mathrm{T}} \nabla J_k \tag{2.8}$$

($y_k$ is defined as in subsection 2.2.1), and

$$\eta_k = \frac{-1}{||p_k|| \min\left(\eta, ||\nabla J_k||\right)} \tag{2.9}$$

where $\eta$ is a configurable constant, set to 0.01 by default.

If $J$ is a quadratic and $\alpha_k$ is chosen to achieve the exact minimum of $J$ in the direction $p_k$, then $p_k^T \nabla J_{k+1} = 0$, and the formula for $\beta_k$ reduces to the traditional Hestenes-Stiefel scheme (see e.g. [162]).

The advantages of the new conjugate-gradient scheme are described in [83]. One such advantage is that this method has guaranteed descent with a smooth function and an inexact line search; however, no such guarantee exists for non-smooth functions.

### 2.2.3 LMBM

Both L-BFGS and CG-Descent were originally created for smooth optimization, and both are generalized for non-smooth methods by replacing the gradient with an arbitrary sub-gradient. However, there are some potentially serious drawbacks to this approach. First, it is theoretically possible for a non-descent search direction to occur as the direction opposite an arbitrary sub-gradient does not guarantee descent. Thus, a smooth line-search algorithm, used for step-size ($\alpha_k$) selection, may fail. Secondly, due to the fact that the norm of an arbitrary sub-gradient does not necessarily become small in the neighborhood of an optimal point, a convergence criterion based on this assumption, valid for smooth gradients, will also fail when the optimal point occurs at a discontinuity. Moreover, in general the convergence speed of sub-gradient methods can be poor ([144]).

In this subsection we describe the limited memory bundle method LMBM ([77], [78]) where the above-mentioned drawbacks are avoided by using the so-called bundling technique and null steps. The idea of bundling is that instead of using just one arbitrary sub-gradient, we approximate the whole sub-differential (see definition (2.1.3)) of the objective function by gathering the sub-gradients from previous iterations into a bundle, and null steps are used when the search direction is not a descent direction. In this way, we obtain more information about the local behavior of the function than what an individual arbitrary sub-gradient alone can yield.

LMBM was specifically developed for solving large-scale non-smooth optimization problems. It is characterized by the usage of null steps together with a simple aggregation of sub-gradients.

Moreover, as in L-BFGS, the limited memory approach is utilized in the calculation of the search direction

$$p_k = -H_k \nabla \tilde{J}_k \tag{2.10}$$

where $\nabla \tilde{J}_k \in \mathbb{R}^n$ is an aggregate sub-gradient and $H_k$ is not formed explicitly but calculated by the L-BFGS update (see equation (2.5)) after a serious step and by the L-SR1 update (see e.g. [34]) after a null step. The usage of null steps gives further information about the non-smooth objective in the case that the search direction is not "good enough." That is, a null step is taken when the descent criterion

$$J(x_k + t_R^k p_k) \leq J(x_k) + \varepsilon_R^k, \tag{2.11}$$

is not satisfied. Here $t_R^k$ is the step size and $\varepsilon_R^k > 0$ is the desired descent of $J$ at $x_k$. In the case of a null step, we set $x_{k+1} = x_k$ but information of the objective function is increased by storing the auxiliary point $\tilde{x}_{k+1} = x_k + t_R^k p_k$ and the corresponding auxiliary sub-gradient $\nabla J_{k+1} \in \partial J(\tilde{x}_{k+1})$. These values are used in the computation of the new aggregate sub-gradient that is used in the next iteration. A simple aggregation of sub-gradients guarantees the global convergence of the method (for more details see [78]) and make it possible to evaluate a sub-gradient based stopping criteria.

The pseudo-code of the LMBM algorithm is shown in table 2.1.

Table 2.1: LMBM Pseudo-code

```
PROGRAM LMBM
  INITIALIZE $x_1 \in \mathbb{R}^n$, $\nabla J_1 \in \partial J(x_1)$, and $\varepsilon_s > 0$;
  Set $k = 1$ and $p_1 = -\nabla J_1$;
  WHILE the termination condition
      $w_k \leq \varepsilon_s$ is not met
    Find step sizes $t_L^k$ and $t_R^k$;
    Set $x_{k+1} = x_k + t_L^k p_k$
    Evaluate $J(x_{k+1})$ and $\nabla J_{k+1} \in \partial J(x_k + t_R^k p_k)$;
    IF $t_L^k > 0$ THEN
      Compute the search direction $p_{k+1}$
        using $\nabla J_{k+1}$ and the L-BFGS update;
    ELSE
      Compute the aggregate
        sub-gradient $\nabla \tilde{J}_{k+1}$;
      Compute the search direction $p_{k+1}$
        using $\nabla \tilde{J}_{k+1}$ and the L-SR1 update;
    END IF
    Set $k = k + 1$;
  END WHILE
  RETURN final solution $x_k$;
END LMBM
```

# CHAPTER 3

# DATA ASSIMILATION

## 3.1 Introduction

Data assimilation (see e.g. [103]) aims to utilize observations of a system in combination with an estimate of the initial state of the system known as the *background*. Both the background and observations – along with their respective uncertainties – must be considered together in order to produce an "optimal" initial condition.

An observation operator is a function mapping between model space and the observation space. For example, an observation operator for satellite radiances may be a radiative transfer model that can utilize the numerically-modeled state of the atmosphere to produce synthetic observations that can be compared to actual satellite radiance observations. Non-linear observation operators are operators which have a non-linearity within them, and non-smooth observation operators are those that have a discontinuity in the derivative of order zero (function value) or higher. The lower the order of the discontinuity, the more difficult the issue is to deal with numerically and mathematically ([144]). Note that non-smooth observations are also non-linear. In this work, we will focus on highly non-linear observation operators and non-smooth operators with discontinuities in the first derivative.

In a general setting with standard assumptions, the data assimilation problem is shown in Table 3.1.

The background state $x_b$ is assumed to be of the form

$$x_b = x_{\text{true}}^{(0)} + \eta_b \tag{3.1}$$

where $x_{\text{true}}^{(0)}$ is the (unknown) true state of the system at the initial time and $\eta_b$ is a random variable with mean 0 and covariance matrix $B$. Note that $\eta_b$ need not be Gaussian; only the first two moments are required.

The observations $y^{(i)}$ are assumed to be of the form

$$y^{(i)} = \mathcal{H}\left(x_{\text{true}}^{(i)}\right) + \eta_{\text{obs}} \tag{3.2}$$

where $\eta_{\text{obs}}$ is a random variable with mean 0 and covariance matrix $R$. The observation errors are usually considered independent, so $R$ is a diagonal matrix.

Table 3.1: Data assimilation problem

Find an "optimal" trajectory $x^{(i)} \in \mathbb{R}^{N_{\text{state}}}$ at each time step $\{i = 1, \ldots, NT\}$ given:

- A "background" state of the model $x_b$ that approximates the true $x^{(0)}$ (from either a first guess or a previous prediction)
- Complete or partial noisy observations of the system $y^{(j)} \in \mathbb{R}^{N_{\text{obs}}^{(j)}}$ for some or all values $0 \leq j \leq NT$
- Background error covariance matrix $B \in \mathbb{R}^{N_{\text{state}} \times N_{\text{state}}}$ quantifying the covariance of the error between the background $x_b$ and the unknown true state
- Observation error covariance matrix $R_j \in \mathbb{R}^{N_{\text{obs}}^{(j)} \times N_{\text{obs}}^{(j)}}$ quantifying the covariance of the error between the observations $y^{(j)}$ and the unknown true observations
- A model $\mathcal{M} : \mathbb{R}^{N_{\text{state}}} \to \mathbb{R}^{N_{\text{state}}}$ that maps $x^{(i)}$ to $x^{(i+1)}$ (assumed here to be perfect)
- An observation operator $\mathcal{H}_j : \mathbb{R}^{N_{\text{state}}} \to \mathbb{R}^{N_{\text{obs}}^{(j)}}$ (also assumed to be perfect) that models the mapping between $x^{(j)}$ to the observations $y^{(j)}$

### 3.1.1 Bayesian framework

One can view data assimilation as a Bayesian problem ([140], [210]); that is, at time $k$, one can view the background probability density function $p(x^{(k-1)}|y^{(1:k-1)})$ as the belief in the distribution of $x^{(k-1)}$ given the previous observations $y^{(1:k-1)}$. At time 0, the first two moments of this distribution are given by $x_b$ and $B$. A common assumption is that the initial background is Gaussian.

From the background distribution and observations $y^{(k)}$, one could then find the exact predicted distribution $p(x^{(k)}|y^{(1:k-1)})$ based on the Chapman-Kolmogorov equation

$$p(x^{(k)}|y^{(1:k-1)}) = \int p(x^{(k)}|x^{(k-1)})p(x^{(k-1)}|y^{(1:k-1)})dx^{(k-1)} \tag{3.3}$$

where $p(x^{(k)}|x^{(k-1)})$ can be found based on the behavior of the model $\mathcal{M}$.

One could then use Bayes rule to find the *posterior* distribution $p(x^{(k)}|y^{(1:k)})$ by

$$p(x^{(k)}|y^{(1:k)}) \propto p(y^{(k)}|x^{(k)})p(x^{(k)}|y^{(1:k-1)}) \tag{3.4}$$

where $p(y^{(k)}|x^{(k)})$ is the *likelihood* of the observation $y^{(k)}$ given the state $x^{(k)}$ and can be inferred from the observation operator $\mathcal{H}$. The value $1/\int p(y^{(k)}|x^{(k)})p(x^{(k)}|y^{(1:k-1)})dx^{(k)}$ is the normalizing constant referred to as the *evidence*.

### 3.1.2 Methodology categorization

In practice, actually solving the integrals (3.3) and (3.4) for systems with a large dimension is far too expensive, and thus approximations to these solutions are required. Many different algorithms have been employed to solve the data assimilation problem in table 3.1. These approaches can be divided into three main categories:

- Variational: treat DA as an optimal control problem

  - 1D-Var – a single vertical column of the model state with no time dependence
  - 3D-Var – three dimensions with no explicit time dependence
  - 4D-Var – three dimensions with explicit time dependence

- Probabilistic: treat DA as an inference problem with ensembles
  - Ensemble Kalman filter (EnKF)
  - Other flavors of EnKF, e.g. the Local Ensemble Transform Kalman Filter (LETKF)
  - Particle filters

- Hybrid: try to combine the best of both approaches
  - 3D-Var plus EnKF type approaches, represented by the Maximum Likelihood Ensemble Filter (MLEF)

In this dissertation, we investigate a penalized 1D-Var problem for satellite radiances, and then compare and contrast 3D-Var, 4D-Var, EnKF, LETKF and MLEF for non-smooth data assimilation using the shallow water equations. A survey of data assimilation methods can be found in [27] and [159]. A survey detailing the pros and cons of variational versus probabilistic methods can be found in [104]. Reviews of Bayesian approaches include [196], [28], [30], and [29].

Because we intend to investigate non-linear and non-differentiable observation operators such as all-sky radiances, we expect EnKF-based filters (including LETKF) to fail due to the lack of theoretical support for non-linear operators as well as in view of studies such as [99]. In particular, EnKF-based filters rely heavily upon the tangent-linear hypothesis for the observation operator. The less accurate this is (i.e. the more non-linear the observation operator is), the worse these methods will perform. In light of [252], we expect 1D-Var, MLEF and 4D-Var, modified to handle non-smooth observations, to perform well. The reasoning behind this is that methods based on variational approaches use an iterative approach at each time step, and therefore the algorithms have multiple opportunities to re-apply the tangent linear hypothesis and thus use better approximations to the non-linear observation operator.

In the following sections, we will detail these methods in brief.

## 3.2 Variational methods

Variational methods ([123], [124]) arise from treating data assimilation as an optimal control problem. A cost function models the cost of a particular state, and this cost function is then minimized over the control variables to produce a state that is locally optimal.

### 3.2.1 1D-Var

The so called “1D-Var” method (e.g. [175]) is an approach that minimizes a cost function with a state of only a single column of the atmosphere, although in general this method can be applied to any three-dimensional model. Assuming the probability density functions for the background and model noise are Gaussian (although not necessarily assuming linearity of the cost function), one can construct a maximum likelihood or minimum variance estimate from the Bayesian framework presented in section 3.1.1 (see e.g. section 3.3 of [45]). In iterative form ([140]), one minimizes a cost function given as

$$J(x) = \frac{1}{2}\delta_b(x)^{\mathrm{T}} B^{-1}\delta_b(x) + \frac{1}{2}\delta_y(x)^{\mathrm{T}} R^{-1}\delta_y(x) \tag{3.5}$$

where $x \in \mathbb{R}^{n_{\text{state}}}$ is both the control and state variable and contains the state variables; $\delta_b(x) = x - x_b \in \mathbb{R}^{n_{\text{state}}}$ is the deviation from the background $x_b \in \mathbb{R}^{n_{\text{state}}}$, and $\delta_y(x) = y - \mathcal{H}(x) \in \mathbb{R}^{n_{\text{obs}}}$ is the deviation from the observations $y \in \mathbb{R}^{n_{\text{obs}}}$ when the observation operator $\mathcal{H} : \mathbb{R}^{n_{\text{state}}} \to \mathbb{R}^{n_{\text{obs}}}$ is applied to $x$. Here, again, $B \in \mathbb{R}^{n_{\text{state}}} \times \mathbb{R}^{n_{\text{state}}}$ is the background error covariance, and $R \in \mathbb{R}^{n_{\text{obs}}} \times \mathbb{R}^{n_{\text{obs}}}$ is the observation error covariance.

In order to solve (3.5), a subgradient of the function $J$ is needed. An arbitrary subgradient for a Lipschitz continuous operator $\mathcal{H}$ is given by

$$\nabla_x J(x) = B^{-1}\delta_b(x) - \mathbf{H}^{\mathrm{T}} R^{-1}\delta_y(x) \tag{3.6}$$

where $\mathbf{H}^{\mathrm{T}} = \frac{\partial \mathcal{H}}{\partial x}^{\mathrm{T}}$ is called the *adjoint* of $\mathcal{H}$, i.e. the transpose of the generalized Jacobian with respect to $x$ of the observation operator .

Using 1D-Var to assimilate a single column, and then passing these columns to either 3D-Var or 4D-Var is a common practice in data assimilation (see e.g. [96], [218]). This approach significantly reduces the computational burden versus a full three- or four-dimensional approach.

### 3.2.2 3D-Var

Like 1D-Var, 3D-Var (e.g. [140], [201], [46]) seeks construct to a maximum likelihood or minimum variance estimate from the Bayesian framework presented in section 3.1.1. The cost function, very similar to 1D-Var, is given by

$$J(x) = \frac{1}{2}\,\delta_b(x)^{\mathrm{T}} B^{-1}\delta_b(x) + \frac{1}{2}\,\delta_{y_k}(x)^{\mathrm{T}} R^{-1}\delta_{y_k}(x) \tag{3.7}$$

where $x$ is the control variable containing the three-dimensional control variables at time $k$, $\delta_{y_k} = y^{(k)} - \mathcal{H}(x)$, where $y^{(k)}$ are the observations at time $k$, and the other variables are the same as in the previous section. A subgradient is given by

$$\nabla_x J(x) = B^{-1}\delta_b(x) - \mathbf{H}^{\mathrm{T}} R^{-1}\delta_{y_k}(x) \tag{3.8}$$

It is clear that, assuming the minimum of the cost function $J(x)$ can be found, the 3D-Var solution is only “optimal” in the sense that it finds the model state that best fits the observations during each time step of a model. However, it is not necessarily the optimal trajectory over all the observations available. 4D-Var is the extension of this method to assimilating the observations at all available times.

### 3.2.3 4D-Var

Much like 1D-Var and 3D-Var, the 4D-Var approach to data assimilation is to minimize a cost function with the lack-of-fit between background and observations scaled by their respective uncertainties. The difference is now that observations at multiple times are assimilated. This translates to seeking the value $x^*$ that minimizes

$$J(x) = \tfrac{1}{2}\, \delta_b(x)^{\mathrm{T}}\, B^{-1}\delta_b(x) + \tfrac{1}{2} \textstyle\sum_{k=0}^{NT} \delta_{y_k}(x)^{\mathrm{T}}\, R^{-1}\delta_{y_k}(x) \tag{3.9}$$

subject to the constraint

$$x^{(k)}(x) = \mathcal{M}\left(x^{(k-1)}(x)\right), x^{(0)}(x) = x \tag{3.10}$$

and the other variables are the same as in 3D-Var, with the exception of

$$\delta_{y_k}(x) = y^{(k)} - \mathcal{H}(x^{(k)}(x)) \tag{3.11}$$

which is the difference between the observation at time $k$, $y^{(k)}$, and the model $\mathcal{M}$ at time $k$, $x^{(k)}(x)$, acted on by the observation operator $\mathcal{H}$.

In this work, the model $\mathcal{M}$ is taken as a strong constraint; in other words, the model constraint equation (3.10) is satisfied at each time-step and the model is taken to be perfect.

Once the optimal initial conditions $x^*$ have been found, the optimal trajectory $x^{(k)}(x^*)$ at each time step $k$ can be found by evolving the model forward in time using $\mathcal{M}$.

As with the previous methods, the subgradient of the cost function (3.9) is required. For a Lipschitz continuous model $\mathcal{M}$ and operator $\mathcal{H}$, this is given by

$$\nabla_x J(x) = B^{-1}\delta_b(x) + \sum_{k=0}^{NT} \left(\frac{\partial x^{(k)}}{\partial x}\right)^{\mathrm{T}} \mathbf{H}^{\mathrm{T}}(x) R^{-1}\delta_{y_k}(x) \tag{3.12}$$

where $\frac{\partial x^{(k)}}{\partial x} \in \mathbb{R}^{N_{\text{state}} \times N_{\text{state}}}$ is the Jacobian of $x^{(k)}$ with respect to $x$, evaluated at $x$. From equation (3.10),

$$\left(\frac{\partial x^{(k)}}{\partial x}\right)^{\mathrm{T}} = \left(\frac{\partial x^{(k-1)}}{\partial x}\right)^{\mathrm{T}} \mathbf{M}\left(x^{(k-1)}(x)\right)^{\mathrm{T}} \tag{3.13}$$

for $k = 1, \ldots, NT$ where $\mathbf{M}$ is the Jacobian of the model operator.

In order to compute this gradient, the adjoint of the tangent linear model is required. Either a discretize-then-differentiate or differentiate-then-discretize approach can be used [74]. In this work, we use adjoints calculated by the discretize-then-differentiate approach, i.e. a subderivative is found by applying the chain rule to the discrete operator $\mathcal{M}$ and reversing the order of the code.

By replacing the adjoints and gradients with their generalized counterparts, (3.6), (3.8) and (3.12) allow for, in general, non-smooth optimization.

## 3.3 Probabilistic filters

Probabilistic filters treat data assimilation as if it were a statistical inference problem. Using an ensemble of perturbed members and their statistics, a near-optimal inference can be gleaned under certain restrictive situations. In this section, we will briefly survey two popular approaches.

### 3.3.1 Ensemble Kalman Filter

The Kalman filter, first developed in [102], is the optimal predictor of state for the data assimilation problem if: 1) the noise random variables $\eta_b$ and $\eta_{\text{obs}}$ are Gaussian and 2) the models $\mathcal{M}$ and $\mathcal{H}$ are linear. These assumptions are rarely valid in the context of NWP. The Extended Kalman Filter (EKF), first developed by [100], uses the Jacobian (i.e. tangent linear model) of $\mathcal{M}$ and $\mathcal{H}$, although EKF is no longer an optimal predictor of state.

Because the dimension of the model state, $N_{\text{state}}$, may be on the order of $10^6$ to $10^7$ (or even larger), forming the covariance matrices directly can be a large challenge since the resulting covariance matrix will require storage on the order of $10^{12}$ or $10^{14}$, i.e. on the order of 1–100 terabytes for single precision data. The Ensemble Kalman Filter (EnKF), first introduced in [58], solves this problem by creating an ensemble of members and estimating the population covariance matrix through the ensemble sample covariance matrix.

We first define an *ensemble* of states that are perturbed replicates of the state.

$$X^{(k-1)} = \left[x_1^{(k-1)}, x_2^{(k-1)}, \dots, x_{N_{\text{ens}}}^{(k-1)}\right] \tag{3.14}$$

where $x_i^{(k-1)} \in \mathbb{R}^{N_{\text{state}}}$ is the $i^{th}$ ensemble at time $k$ and $N_{\text{ens}}$ is the number of ensembles. At the initial time, the ensemble members are given by $x_i^{(0)} = x_{\text{true}}^{(0)} + \eta_b$ where $\eta_b \sim N(0, B)$.

The EnKF algorithm can be broken into two steps: forecast and analysis. In the forecast step, the ensemble is moved forward in time by the model, and the sample covariance of the ensemble is calculated. In the analysis step, the forecast values are used to predict a "near-optimal" update given the ensemble statistics, state, observations, and corresponding models.

The $i^{th}$ member of the forecast ensemble $X_f^{(k)}$ at time $k > 0$ is given by

$$\left(X_f^{(k)}\right)_i = \mathcal{M}\left(x_i^{(k-1)}\right) \tag{3.15}$$

From this forecast ensemble, the sample forecast covariance that approximates the population covariance can be found:

$$P_f^{(k)} = \frac{1}{N_{\text{ens}} - 1} \sum_{i=1}^{N_{\text{ens}}} \left(\left(X_f^{(k)}\right)_i - \bar{x}_f^{(k)}\right)\left(\left(X_f^{(k)}\right)_i - \bar{x}_f^{(k)}\right)^{\mathrm{T}} \tag{3.16}$$

where $\bar{x}^{(k)} \in \mathbb{R}^{N_{\text{state}}}$ is the ensemble average of $X_f^{(k)}$. Note that this matrix is not actually stored in memory, but can be created through the outer-product formulation when needed.

EnKF utilizes perturbed observations, which are given by

$$Y^{(k)} = \left[y_1^{(k)}, y_2^{(k)}, \dots, y_{N_{\text{ens}}}^{(k)}\right] \tag{3.17}$$

where each $y_i^{(k)}$ has independently sampled noise as

$$y_i^{(k)} = y^{(k)} + \eta_{\text{obs}} \tag{3.18}$$

with $\eta_{\text{obs}} \sim N(0, R)$. The reason for these perturbed observations. as first formulated in [33], is that each of the observation replicates have the mean $y_{\text{true}}^{(k)}$ and covariance $R$.

While each of the observation replicates have covariance $R$ if $y^{(k)}$ is considered a constant, in a realistic best-case scenario where the uncertainty $R$ would be accurately quantified, $y^{(k)}$ would already include noise from $\eta_{\text{obs}}$, that is:

$$y^{(k)} = \mathcal{H}\left(x_{\text{true}}^{(k)}\right) + \eta_{\text{obs}_1} \tag{3.19}$$

In other words,

$$y_i^{(k)} = \mathcal{H}\left(x_{\text{true}}^{(k)}\right) + \eta_{\text{obs}_1} + \eta_{\text{obs}_2} \tag{3.20}$$

i.e. the noise is doubled. While the mean is still $y_{\text{true}}^{(k)} = \mathcal{H}\left(x_{\text{true}}^{(k)}\right)$, this approach in essence adds additional observation uncertainty and the use of $R$ for the covariance is inaccurate. This issue is overlooked or neglected by authors ranging from Burgers et al. ([33]) to Evensen ([59]).

If the noise covariance was additive (as is the case for Gaussian noise), and the matrix $R$ was used, this method would in fact underestimate the uncertainty by a factor of two, i.e.

$$\text{Cov}(y_i^{(k)}) = \text{Cov}\left(\eta_{\text{obs}_1} + \eta_{\text{obs}_2}\right) = 2R \tag{3.21}$$

This error could lead to underutilization of the observations.

Whitaker and Hamill ([223]) present a formulation of EnKF that can be used without perturbing the observations; methods based on this approach are called “deterministic” flavors of EnKF, and unlike perturbed observation flavors, can be considered realistic methods that accurately quantify the observation covariance.

Returning to the traditional formulation of EnKF, the new ensemble at time $k$ is given by

$$X^{(k)} = X_f^{(k)} + K^{(k)}\left(Y^{(k)} - \mathbf{H}_f X_f^{(k)}\right) \tag{3.22}$$

where

$$K^{(k)} = P_f^{(k)}\mathbf{H}^{\mathrm{T}}\left(\mathbf{H}P_f^{(k)}\mathbf{H}^{\mathrm{T}} + R\right)^{-1} \tag{3.23}$$

is the Kalman gain matrix arising from the Kalman filter solution (see e.g. [100]).

The new covariance matrix is given by

$$P^{(k)} = \left(I - K^{(k)}\right)P_f^{(k)} \tag{3.24}$$

where $I$ is the identity matrix. This is theoretically identical to the sample covariance of the updated ensemble $X^{(k+1)}$ from equation (3.16) and thus either equation can be used.

By taking $\nabla J = 0$ in the 3D-Var gradient equation (3.8), and using $P^{(k)} = B$, it can be shown that the EnKF update is theoretically identical to the 3D-Var analysis. However, since the 3D-Var is an iterative method, it may able to handle non-linear $\mathcal{H}$ better than EnKF-based approaches. On the other-hand, because EnKF accurately evolves the uncertainty $P^{(k)}$ in the situation forward in time, while 3D-Var in its most common formulation uses the fixed background error covariance model $B$, EnKF has the ability to scale the background term in a more realistic way, and for near-linear operators will find a superior solution when compared to 3D-Var. Thus, both 3D-Var and EnKF have strengths and weaknesses.

The matrix $P^{(k)}$ is a $N_{\text{ens}}$ rank estimate of the $N_{\text{state}} \times N_{\text{state}}$ matrix, where $N_{\text{ens}} << N_{\text{state}}$. Thus, $P^{(k)}$ is highly rank deficient and displays spurious correlations. These issues can lead to filter divergence where the EnKF error grows rapidly. One method for correcting these issues is covariance localization, introduced by [94]. This method replaces the Kalman gain $K^{(k)}$ in equation (3.22) with the modified Kalman gain $\hat{K}^{(k)}$,

$$\hat{K}^{(k)} = \rho \circ P_f^{(k)} \mathbf{H}^{\mathrm{T}} \left( \mathbf{H} \rho \circ P_f^{(k)} \mathbf{H}^{\mathrm{T}} + R \right)^{-1} \tag{3.25}$$

where $\rho \ \circ \ P^{(k)}$ is the Schur product (i.e. element-wise multiplication) between $\rho$ and $P^{(k)}$, where

$$\rho_{i,j}(D_{i,j}) = \exp\left[ - \left( \frac{D_{i,j}}{\ell} \right)^2 \right] \tag{3.26}$$

is the correlation matrix, $D_{i,j}$ is the distance between two grid points $(i, j)$, and $\ell$ is the correlation length.

Furthermore, additive covariance inflation, first suggested by [6], is also often used to prevent undesirable collapse of the covariance matrix $P^{(k)}$. This amounts to modifying each member $i$ of the ensemble by

$$\left( x_i^{(k)} \right)' = \lambda \left( x_i^{(k)} - \bar{x}^{(k)} \right) + \bar{x}^{(k)} \tag{3.27}$$

where $\lambda$ is the inflation parameter. For $\lambda = 1$, this does nothing, but for values of $\lambda > 1$, the amounts to an artificial adjustment of the covariance that gives more weight to the observations. This may be even more necessary considering the observation covariance error underestimation in the standard EnKF formulation.

EnKF relies heavily on the tangent linear observation operator $\mathbf{H}$ and is best suited for Gaussian noise. In the case of a highly non-linear or non-smooth observation operator, the tangent linear hypothesis will be invalid near the areas of strong non-linearity or discontinuities in the derivative and the ensemble noise would no longer be Gaussian. Therefore, one would expect EnKF to have difficulties with such operators. In chapter 5 we will quantify these difficulties on a non-smooth data assimilation problem.

### 3.3.2 Local Ensemble Transform Kalman Filter

The Ensemble Transform Kalman Filter (ETKF), first introduced by Bishop et al. in [26], is based on EnKF but avoids perturbing the observations. It is thus a *deterministic* filter (although this moniker is somewhat misleading as the ETKF is still a filter based upon statistical inference). It is a type of square-root Kalman filter.

A closely related implementation with covariance localization, the Local Ensemble Transform Kalman Filter (LETKF), is a popular choice. LETKF was first developed in [95] in 2007. While there are many different flavors of EnKF, this version is one of the most recent developments and seems to be one of the most promising methods in the EnKF family. It should therefore be very useful to the scientific community to include this in a comparison with the variational and hybrid methods, especially in the presence of non-linear and non-smooth observation operators.

In the development below, for convenience of notation, the time dependence (i.e. $(k)$) will be dropped temporarily.

LETKF distinguishes between global and local matrices. The global matrices are those defined as above, while local matrices are selected copies of data that have been reduced in size through covariance localization.

As with EnKF, the LETKF algorithm uses a global matrix $X_f^{(k)}$ of forecast states

$$X_f^{(k)} = \mathcal{M}\left(X^{(k-1)}\right) \tag{3.28}$$

A global matrix $Y_p \in \mathbb{R}^{N_{\text{obs}} \times N_{\text{ens}}}$ of the simulated observation perturbations for each ensemble is also created, where

$$(Y_p)_i = (Y_f)_i - \bar{y}_f \tag{3.29}$$

where $Y_f \in \mathbb{R}^{N_{\text{obs}} \times N_{\text{ens}}}$ is

$$(Y_f)_i = \mathcal{H}\left(\left(X_f^{(k)}\right)_i\right) \tag{3.30}$$

and $\bar{y}_f \in \mathbb{R}^{N_{\text{obs}}}$ is the global average of observations.

$$\bar{y}_f = \frac{1}{N_{\text{ens}}} \sum_i^{N_{\text{ens}}} (Y_f)_i \tag{3.31}$$

The global perturbation matrix $X_p \in \mathbb{R}^{N_{\text{state}} \times N_{\text{ens}}}$ is also used,

$$(X_p)_i \;=\; \left(X_f^{(k)}\right)_i - \bar{x}_f \tag{3.32}$$

where $\bar{x}_f \in \mathbb{R}^{N_{\text{state}}}$ is

$$\bar{x}_f = \frac{1}{N_{\text{ens}}} \sum_i^{N_{\text{ens}}} \left(X_f^{(k)}\right)_i \tag{3.33}$$

In LETKF, covariance localization is achieved by only including values that are “local” to a grid point in the analysis. Locality can be determined by several potential methods, but here a point will be considered local if its distance is less than the characteristic length, i.e. points $i$ and $j$ will be considered co-local if $D_{i,j} \leq \ell$. Only those values local to a grid point will be used in the following steps.

For each grid point $n$, copies of various variables are created, again only storing the local values. The dependence on $n$ will be temporarily neglected again for the sake of notation. These local copies are given by

$$X_p^{[\ell]} \;\in\; \mathbb{R}^{N_{\text{state}}^{[\ell]} \times N_{\text{ens}}} \tag{3.34}$$

$$\bar{x}_f^{[\ell]} \;\in\; \mathbb{R}^{N_{\text{state}}^{[\ell]}} \tag{3.35}$$

$$Y_p^{[\ell]} \;\in\; \mathbb{R}^{N_{\text{obs}}^{[\ell]} \times N_{\text{ens}}} \tag{3.36}$$

$$\bar{y}_f^{[\ell]} \;=\; \mathbb{R}^{N_{\text{obs}}^{[\ell]}} \tag{3.37}$$

$$R^{[\ell]} \;\in\; \mathbb{R}^{N_{\text{obs}}^{[\ell]} \times N_{\text{obs}}^{[\ell]}} \tag{3.38}$$

where $N_{\text{state}}^{[\ell]}$ and $N_{\text{obs}}^{[\ell]}$ are the number of local points for the grid point in question in the state and observation space, respectively.

The intermediate local matrix $C \in \mathbb{R}^{N_{\text{ens}} \times N_{\text{obs}}^{[\ell]}}$ is then computed,

$$C = \left(Y_p^{[\ell]}\right)^{\mathrm{T}} \left(R^{[\ell]}\right)^{-1} \tag{3.39}$$

and the local analysis covariance matrix $\tilde{P}_a \in \mathbb{R}^{N_{\text{ens}} \times N_{\text{ens}}}$ is given by

$$\tilde{P}_a = \left[\frac{(N_{\text{ens}})}{\lambda} I + C Y_p^{[\ell]}\right]^{-1} \tag{3.40}$$

Here, as in EnKF, $\lambda > 1$ is a covariance inflation factor.

The matrix $W_a \in \mathbb{R}^{N_{\text{ens}} \times N_{\text{ens}}}$ is calculated by

$$W_a = \left[(N_{\text{ens}} - 1)\, \tilde{P}_a\right]^{1/2} \tag{3.41}$$

where the square root is understood to be any symmetric square root decomposition such as the generalized Cholesky decomposition.

The vector $\bar{w}_a \in \mathbb{R}^{N_{\text{ens}}}$ is given by

$$\bar{w}_a = \tilde{P}_a C \left(y^{(k)} - \bar{y}_f^{[\ell]}\right) \tag{3.42}$$

This vector is added to each column of $W_a$, i.e.

$$(W_a)_i = (W_a)_i + \bar{w}_a \tag{3.43}$$

Finally, the new local analysis $X_a^{[\ell]} \in \mathbb{R}^{N_{\text{state}}^{[\ell]} \times N_{\text{ens}}}$ of the $i^{th}$ ensemble at the grid point $n$ is given by

$$\left(X_a^{[\ell]}\right)_i^n = X_p^{[\ell]} (W_a)_i + \bar{x}_f^{[\ell]} \tag{3.44}$$

In the simplest implementation of LETKF, this process is repeated for each grid point $n$. For each grid point $n$ and ensemble $i$, the updated ensemble analysis is given by

$$\left(X^{(k)}\right)_{n,i} = \left(X_a^{[\ell]}\right)_{L(n),i}^n \tag{3.45}$$

where $L(n)$ is a function mapping global indexing into local indexing. Note that the subscript $(n, i)$ here indicates the $n^{th}$ row and $i^{th}$ column of the matrix.

It should be emphasized that LETKF avoids directly using the tangent linear model $\mathbf{H}$ or its adjoint. However, it implicitly uses the assumption that the tangent linear hypothesis is valid in the development of the method ([95]). Because, unlike variational methods, only one iteration is used, if the tangent linear hypothesis is weak or even invalid, the method has no chance to correct for this. It is thus hypothesized that LETKF will suffer from the same issues as EnKF when faced with highly non-linear or non-smooth observation operators, although perhaps to a lesser extent since neither the tangent linear model $\mathbf{H}$ nor its adjoint is explicitly used. Finally, since unlike some flavors of EnKF, LETKF does not use perturbed observations, it will not underestimate the observation error covariance.

## 3.4 Hybrid filters

Hybrid filters take the best features from both the variational and probabilistic approaches. The only hybrid filter we will investigate in my dissertation is the Maximum Likelihood Ensemble Filter, although this is considered representative of all potential hybrid methods.

### 3.4.1 Maximum Likelihood Ensemble Filter

The Maximum Likelihood Ensemble Filter (MLEF) was first proposed by Zupanski in 2005 ([250]). The MLEF approach to data assimilation shares similarities with both a variational approach such as 3D-Var and an ensemble approach such as the Ensemble Kalman Filter, and especially the Ensemble Square Root Filter [202] and Ensemble Kalman Transform Filter. The derivation presented below is similar to [65] and [252] in which additional details can be found.

Like the Kalman filter family, the MLEF algorithm proceeds in two-stages: forecast and analysis. Suppose that a square-root analysis covariance matrix $\left(P^{(k-1)}\right)^{1/2}$ is available at time $k-1$ such that

$$P^{(k-1)} \approx \left(P^{(k-1)}\right)^{1/2} \left(P^{(k-1)}\right)^{T/2} \tag{3.46}$$

This is equal to the background $B$ at time 0, i.e. $P^{(0)} = B$. In MLEF, the columns of the forecast error covariance $\left(P_f^{(k)}\right)^{1/2}$ at time $k$ are given by

$$\begin{aligned} \left(P_f^{(k)}\right)^{1/2} &= \left[p_1^f, p_2^f, \ldots, p_{N_{\text{ens}}}^f\right] \\ p_i^f &= \mathcal{M}\left(x_a^{(k-1)} + p_i\right) - \mathcal{M}\left(x_a^{(k-1)}\right) \end{aligned} \tag{3.47}$$

where $N_{\text{ens}}$ are the number of ensembles, $x_a^{(k-1)}$ is the previous analysis value ($x_a^{(0)} = x_b$), and $p_i$ is the $i^{\text{th}}$ column of $\left(P^{(k-1)}\right)^{1/2}$.

Once the forecast covariance has been obtained, the analysis step can proceed. The analysis step of MLEF takes inspiration from variational methods, in particular 3D-Var. MLEF seeks (conceptually) to find the analysis $x_a^{(k)}$ that minimizes the cost function

$$J(x) = \frac{1}{2}\,\delta_f(x)^{\mathrm{T}}\,P_f^{-1}\delta_f(x) + \frac{1}{2}\,\delta_{y_k}(x)^{\mathrm{T}}\,R^{-1}\delta_{y_k}(x) \tag{3.48}$$

where $\delta_f = x - x_f^{(k)}$, $x_f^{(k)} = \mathcal{M}\left(x_a^{(k-1)}\right)$ and the other variables are the same as in section 3.2.3.

In order to avoid inverting the matrix $P_f$ – which will be rank deficient as it is approximated by matrix of rank at most $N_{\text{ens}}$ – a change of variables is introduced, namely:

$$\delta_f(x) = P_f^{1/2}\,(I + C(x))^{-\mathrm{T}/2}\,\zeta \tag{3.49}$$

where $\zeta$ are the new control variables, defined in the $N_{\text{ens}} \times 1$ space of ensembles, and $C \in \mathbb{R}^{N_{\text{ens}} \times N_{\text{ens}}}$ is a preconditioning matrix of the quadratic cost function (3.48). $C$ is formed as follows.

$$C(x) = Z^{\mathrm{T}} Z \tag{3.50}$$

where

$$Z_i(x) = R^{-1/2}\left[\mathcal{H}(x + p_i^f) - \mathcal{H}(x)\right] \tag{3.51}$$

Because of the mutual dependency between the preconditioner and $x$, $C$ is fixed to $C(x = x_f^{(k)})$, i.e. $C$ does not change during the minimization process.

The inversion of the symmetric matrix $(I + C)$, required by (3.49), is accomplished using a spectral decomposition of the form

$$I + C = V\Lambda V^{\mathrm{T}} \tag{3.52}$$

where $V$ is an orthogonal matrix of eigenvectors and $\Lambda$ is a diagonal matrix of eigenvalues. Once this decomposition is found, the required square-root can be found by

$$(I + C)^{-1/2} = V\Lambda^{-1/2}V^{\mathrm{T}} \tag{3.53}$$

Finally, the square root analysis covariance matrix at time $(k)$ is found by

$$P_a^{1/2} = P_f^{1/2}\left[I + C(x_a^{(k)})\right]^{-\mathrm{T}/2} \tag{3.54}$$

with the notation $C(x_a^{(k)})$ denoting that $C$ is recomputed at the solution $x_a^{(k)}$ of (3.48).

This update to the covariance matrix is similar to that of [26]; however, the main difference is that the observation operator is not restricted to be linear, and the Jacobian is not required.

Like the other ensemble-based methods listed here, MLEF also uses covariance localization. The covariance localization method is similar to LETKF approaches such as ([156] and [232]). However, unlike many LETKF approaches, MLEF only employs covariance localization in the horizontal directions, i.e. the covariance in the vertical direction is not localized. The reasoning behind this is that vertical localization can adversely affect vertical model balance, especially in the presence of clouds.

In summary, MLEF is an ensemble method that directly maximizes the posterior probability density function at each time step. It does not require the Jacobian of either the model or the observation operator. In addition, as shown by Jardak et al. ([99]), MLEF can perform well when the Gaussianity assumption for the model is no longer valid and/or when the observation operator is non-linear.

## 3.5 Summary

Table 3.2 shows the various pros and cons of the different methods that will be investigated. $\mathbf{M}^{\mathrm{T}}$ means that the adjoint of the model with respect to the state variables is required, while $\mathbf{H}$ means that the tangent linear model (or potentially its adjoint) of the observation operator is required. “Batches” refers to whether multiple observations (i.e. $y^{(j)}$ at multiple times $j$) values can be assimilated. “Uncertainty” means how the uncertainty is quantified using these methods. Further explanation is warranted here. For 3D-Var, at the optimal point, the Hessian singular values of the cost function at each time step can be found, which gives an estimate of the inverse uncertainty matrix in the solution at that time ([167]). Uncertainty quantification in 4D-Var functions in a similar fashion, although the uncertainty is now with respect to all of the assimilated observations at the initial time. EnKF

Table 3.2: Pros and cons of the various DA methods

| **Method name** | $\mathbf{M}^{\mathrm{T}}$ | **H** | **Batches** | **Uncertainty** | **Iterative** | **NL** $\mathcal{M}$ | **NL** $\mathcal{H}$ |
|---|---|---|---|---|---|---|---|
| 3D-Var | No | Yes | Single | Hessian | Yes | Yes | Yes |
| 4D-Var | Yes | Yes | Multiple | Hessian | Yes | Yes | Yes |
| EnKF | No | Yes | Single | Sample | No | Yes | No |
| LETKF | No | No | Single | Deterministic | No | Yes | Yes |
| MLEF | No | No | Single | Deterministic | Yes | Yes | Yes |

uses a sample covariance to estimate the population covariance, while LETKF uses a deterministic approach that modifies the covariance matrix columns directly. MLEF also modifies this matrix through a deterministic approach in a somewhat similar manner. “Iterative” means whether more than one iteration is used per assimilation cycle – a hallmark of variational methods. “NL $\mathcal{M}$” and “NL $\mathcal{H}$” denotes that the non-linear model or non-linear observation operator are used, respectively.

# CHAPTER 4

# NON-LINEAR OBSERVATION OPERATORS: ALL-SKY INFRARED RADIANCES

## 4.1 Introduction

In this chapter, we test an open data assimilation problem that has a highly non-linear observation operator and may potentially benefit from non-smooth optimization. The problem under consideration is the reconstruction of a single column of the atmosphere that may or may not contain clouds from satellite observations in the infrared spectrum. This problem exhibits a sharp transition from clear to cloudy scenes. The problem is also ill-posed in the sense that multiple states of the atmosphere can produce the same observations at the top of the atmosphere. However, the problem is regularized through the use of a background solution, which also provides a first-guess and prevents the solution from straying "too far" from this first guess.

The objective of this chapter is to evaluate the performance of our chosen optimization operators on the problem of data assimilation with highly non-linear observation operators. We also aim to highlight the importance of choosing a good background and background error covariance model.

For this problem, we consider synthetic data from the Atmospheric Infrared Sounder (AIRS) satellite with the Weather Research and Forecasting numerical weather prediction model and the ECWMF RTTOV fast radiative transfer model. We test the performance of a penalized 1D-Var algorithm with L-BFGS, CG-Descent, and the non-smooth optimization method LMBM with this observation operator.

As we will survey below, the problem of clear-sky data assimilation using infrared satellites is well understood, and while several studies have addressed cloud-fraction data assimilation, the full all-sky infrared data assimilation problem is currently a topic of intense research, especially at cloud-resolving scales.

At such scales, the presence of clouds introduces strong non-linearities in the observation (forward) operator with respect to the cloud micro-physical control variables between cloudy and clear-sky radiances due to the sharp transitions from clear skies to clouds within the atmosphere. This high degree of non-linearity ([122]) in the cost function may become an issue for any optimization algorithm being employed in a variational assimilation system. A highly non-linear or discontinuous cost function may lead to a poor solution or even a divergence of the minimization algorithm ([73]). Different methods are used to try to work around these non-linear and discontinuous issues including smoothing and regularization (e.g. [17], [97]); however, these resolutions are often ad-hoc and it is not clear that these remedies can be applied in more general settings. In addition, it is not known how these workarounds impact the final data assimilation solution. Many issues in this area still remain unresolved. For example, while in one recent study which contained cloud information in the initial condition, the data assimilation converged and improved the cloud information based on all-sky infrared radiances ([214]), in another study, where a cloud was not present in the initial condition, data assimilation with a traditional 4D-Var method for the infrared all-sky problem fails to reconstruct the desired cloud ([182]). In this chapter we will survey these problems and offer a new solution that treats the non-linearities and non-discontinuities directly through the use of a penalized 1D-Var with non-smooth optimization.

In Bauer et al. ([17]) and Geer et al. ([68]), some interesting statistics are given. While satellite observations provide 90-95% of the data that is assimilated, over 75% of these observations must be discarded due to cloud contamination and unknown surface emissivities. While all-sky microwave data is currently used around the world in operational centers, all-sky infrared data is still being investigated. While many cloudy-infrared products are already available ([122]), these products focus primarily on single-layer retrieval products from the infrared data such as cloud-top pressure and temperature, which introduce a source of error ([182]). This work examines assimilating all-sky infrared radiances directly through the use of a parameterized multiple-scattering radiative transfer model. The potential benefits of this approach are many, and include improving initial conditions of the optical depth and hydrometeor species and concentration of the cloudy column under investigation. When used in conjunction with microwave information, in particular, this method may greatly enhance cloud information in the initial conditions ([16]).

The material in this chapter is the subject of an upcoming paper prepared for publication [195].

### 4.1.1 Survey

The problem of clear-sky data assimilation using infrared satellites is well understood (e.g. [62], [43]). However, the full all-sky infrared data assimilation problem is currently a topic of active and intense research, especially at cloud-resolving scales ([193]).

Because of the large amount of useful information within cloud-affected radiances, the assimilation of such observations has long been a goal of numerical weather prediction. Thus the history of assimilating infrared satellite data is long and storied. We begin with early attempts to use this data in a meaningful way. We then switch our focus to the modern data assimilation of cloudy radar data, which in many ways informs techniques that can be used for the assimilation of cloudy infrared data. We then summarize modern attempts at assimilating microwave data, which is even closer in nature to infrared data, and conclude with the modern early results of assimilating all-sky infrared data.

**Early utilization of satellite data to initialize clouds.** The satellite era began on October 4th, 1957 when the Soviet Union successfully launched Sputnik I, the first artificial satellite. The United States followed with Explorer-I in 1958, and on April 1st 1960, NASA launched TIROS-I, which began providing observations of the earth from space for the purpose of improving meteorological predictions. Unfortunately TIROS-I only provided 78 days worth of data before failing, but proved that satellites could indeed provide useful information for NWP purposes. The TIROS series continued in the 1960s, with TIROS-7 in 1963 reaching an operational period of 1809 days before it was deactivated. With the subsequent introduction of the NIMBUS series of satellites, launched successively from 1964 until 1978, space-based remote-sensing of the atmosphere became an indispensable tool for meteorological applications ([192], [1], [190]). A multitude of satellites that have provided and/or continue to provide potentially useful data has since literally filled the sky. This survey focuses on the utilization of cloud data in numerical weather prediction models from satellite sources.

While many early studies investigated the potential for deriving properties of clouds from space-based instruments (e.g. Basharinov 1969 [13], Akvilonova 1973, [3]), this survey covers solely the use of this information to initialize NWP models.

One of the earliest practical methods for cloudy moisture analysis using satellites involved the "$CO_2$ slicing" technique. The method uses differences between multiple (two to four) channels of infrared data near the 15 $\mu$m $CO_2$ absorption band to detect low, medium, and high clouds with the assumption that clouds were blackbodies, a fairly reasonable assumption for the channels investigated. In addition to cloud-top heights, the method could also estimate effective cloud amount and even cloud motion winds based on tendencies (Chahine 1974, [38], 1977, [39], 1977 et al. [40]). Authors using this method include Smith et al. (1974, [189])), Smith and Platt (1978, [188]), Menzel et al. 1983, [153]), Schreiner et al. (1993, [181]), and Bayler et al. (2000, [21]). $CO_2$ slicing still remains in use as a first guess for modern implementations, e.g. Heilliette and Garand (2007, [87]).

Another early method for the usage of satellite data for initializing clouds was a technique known as "nephanalysis" where visible and/or infrared satellite data was used to manually prepare a synoptic scale chart with different types of clouds that was then used to assign vertical humidity profiles. Perkey (1976, [170]) was the first to use this technique to initialize a hydrostatic primitive equations model with predicted horizontal winds, virtual temperature, specific humidity, cloud or suspended water, and precipitation water. Perkey manually (and subjectively) inserted "bogus moisture soundings" based on the nephanalysis chart and found that precipitation was better predicted than by using radiosonde data alone. He concluded with the importance of investigating more realistic initialization techniques for this data.

Mills (1983, [154]) described the use of nephanalysis with an operational limited area model of 250 km horizontal resolution and 6 vertical levels to predict precipitation from a tropical cyclone over Australia. Starting from geostationary infrared data, cloud type and possible cloud-top heights were assigned. Then, using surface observations and radiosonde data, they assigned cloud-base heights and created vertical dew-point profiles. They found that their precipitation and low pressure center were greatly improved using this technique, and concluded by discussing possible techniques to make this system more objective and effective. One suggestion was that an "experienced operator could, ... using an interactive graphics terminal, overlay the objective moisture analysis and satellite imagery, and modify the moisture analysis in the cloudy areas as desired."

Other authors to use nephanalysis to initialize their model include Wolcott and Warner (1981, [225]), Cadet (1983, [35]), Mills and Davidson (1987, [155]), Norquist (1988, [163]), Mailhot et al. (1989, [142]), Bell and Hammon (1989, [22]), and Hamill et al. (1992, [84]). A closely related technique was used by Zou and Xiao (2000, [243]).

Another approach for initializing NWP methods from satellite observations was based on statistical methods. DesBois et al. (1982, [48]) used a clustering algorithm with one channel of visible and two channels of infrared data. An automatic classification scheme was used to identify clusters in the three-dimensional space of these three channels. Based on the pixels of data in a particular area, a histogram was constructed of the dark vs. bright and warm vs. cold pixels, from which different classes of clouds and surface albedo could be identified. This study allowed for partially transparent clouds. The benefits of this method were that many different classes of cloud and surface type could be quickly identified based on empirically trained models. Other authors to use this method for cloud classification include Koch et al. (1997, [112]).

Yet another early approach for utilizing model cloudy state based on satellite information is known as physical or diabatic initialization. In this approach, inverse model parameterizations, similarity theory, and relaxation were used to make the model be consistent with outgoing observations. In essence the model was initialized with a background state, and the values were aimed at the observations and "nudged" towards the observation target. Authors employing this approach include Krishnamurti et al. (1984, [117], 1988, [114], 1991, [116], 1993, [115]), Donner (1988, [51]), Puri and Miller ([172]), Heckley et al. (1990, [85]), Turpeinen et al. (1990, [208]), Davidson and Puri (1992, [47]), Manobianco et al. (1994, [145]), Raymond et al. (1995, [174]), and Kasahara et al. (1996, [107]).

The final early approach we survey was the method of incorporating many different sources to determine a three-dimensional cloud analysis. Macpherson et al. (1996, [141]) described the Moisture Observations Preprocessing System (MOPS) used operationally in the UK Meteorological Office, which converted a 31-level three-dimensional cloud analysis from a wide variety of data sources (satellite, radar and surface observations) into a set of relative humidity profiles to be used with a nudging scheme. Cloud-cover was determined by Meteosat infrared satellite brightness temperatures, with cloud-cover indicated if the brightness temperature was more than 5K less than the surface temperature. This is related to the so-called minimum residual method of Eyre et al. (1989, [60], [61]). In this method, the cloud-top is moved through each level of the model to find the brightness temperature that best matches the observations. While this approach is acceptable for actual brightness temperature calculations at a single layer, in Macpherson et al. cloud-top pressure was found by selecting the background temperature matched the infrared brightness temperature, a risky approach because the measured brightness temperature depends on many factors including temperature at all atmospheric levels, gas profiles, wavelength of the channel considered, zenith angle, etc. ([192]). In this study alone, the same column of atmosphere gives brightness temperatures of between 200 K and 300 K for various AIRS channels (see below); while some of these channels clearly peak at higher points in the atmosphere than the channels investigated in Macpherson et al., the point is that the brightness temperature does not necessarily correspond in a direct way to physical temperature, and this introduces a large potential source of error. Similar approaches include Albers et al. (1996, [4]), Zhang et al. (1998, [234], 1999 [233]), and Ducrocq et al. (2000, [53]). These authors used a variety of sources (with some authors even including pilot reports); however, they all employed the crude approximation that the highest cloud layer temperature corresponded to the infrared brightness temperature.

**Early variational approaches.** In 1992, Wu and Smith ([228]) took steps towards the variational data assimilation of relative humidity as a proxy for clouds with infrared observations. They minimized a cost function that measured the discrepancy between the model computed and observed outgoing radiation. The 15-level cost function they considered was based on outgoing long-wave radiation at the model initial time with no background or uncertainty term (thus making the problem ill-posed), and the optimization algorithm considered was based on Newton's method. The work of Zupanski (1993, [245]), Zou et al. (1993, [244]), Zupanski and Messenger (1995, [246]), Tsuyuki (1996, [206]), Zou and Kuo (1996, [241]), Kuo et al. (1997, [121]), Zhu and Navon (1999, [240]), and Guo et al. (2000, [75]) helped put this type of approach on more stable theoretical and numerical footings, albeit with idealized observation operators (i.e. direct observations of model variables rather than relating them to satellite observations) or with highly simplified observation operators and synthetic data.

One issue which hampered the development of effective variational methods was the presence of on/off switches in the forward parameterizations – a very common occurence for modeling clouds – which in turn lead to discontinuities in the model adjoint. This in turn could negatively affect the minimization, which, with few exceptions, are based on algorithms originally designed for smooth optimization. A variety of authors used smoothing methods and simplifications to address this problem including Zupanski (1993, [245]), Zou et al. (1993, [244]), Verline and Cotton (1993, [211]), Tsuyuki (1996, [206]), Xu (1996, [231]), and Janiskova (1999, [98]). This dissertation treats the problem directly by using non-smooth optimization methods.

### 4.1.2 Modern cloudy radar assimilation

In this dissertation, we arbitrarily demarcate the use of cloud microphysical variables with advanced data assimilation techniques as the entry into the "modern" era of cloudy assimilation. Assimilation of cloudy radar data was the first foray into this field.

The rate of precipitation and radar reflectivities have a near-linear relationship (see e.g. [135]). This allows radar data to be assimilated with near-linear observation operators, which makes the data assimilation problem highly tractable. Because it relies so heavily on scattering, radar gives detailed information about clouds and very little information about other areas. In some ways this is therefore complementary to infrared satellite data, which gives more complete vertical profiles away from clouds. Some of the earliest modern attempts at cloudy assimilation therefore were done using radar data.

The first pre-modern study of initializing NWP models with radar was done by Lin et al. (1993, [135]) using radar reflectivities and temperature retrievals to modify microphysical variables. They used a heuristic approach to initialize most variables, then used an empirical relationship between rain and radar reflectivity to determine initial model variables. No consideration for observational errors in the observation was considered, and many variables were simply initialized to zero. Bielli and Roux (1999, [25]) did a somewhat similar study using airborne Doppler radar.

Haddad et al. (1996, [80], [79]) describe both a full Bayesian and an extended Kalman filter approach to deriving rain profiles from radar observations based on several variables in an idealized observation operator using synthetic and actual data. They also included a "drift" parameter which is similar to a model error term. They reported encouraging results, although the number of model variables considered was small (less than 10). Similar studies include L'Ecuyer and Stephens (2002, [125]).

Sun and Crook (1997, [198], 1998 [199]) used a 4D-Var system with simulated Doppler radar data for three-dimensional wind, temperature and warm (no ice) microphysics. They included a penalty term in their cost function for "spatial and temporal smoothness." They used a six-equation model based on the anelastic approximation with explicit warm rainwater and an idealized, smooth and nearly linear observation operator (radial velocity and reflectivity) along with a simple, uncorrelated covariance model. Their model, like most modern models, had "on-off" parameterization switches for processes such as evaporation and autoconversion. They used the L-BFGS minimization method to solve their variational problem, and pointed out the problems associated with the non-smooth switches and non-linear parameterizations. They found that the non-smooth problems associated with their parameterizations were largely mitigated with no work needed, potentially due non-smooth-optimization-like properties of L-BFGS, although the highly non-linear aspects of their parameterization caused the minimization to fail. They mitigated these problems by making changes to the forward and adjoint models to smooth out the gradients. Their results showed that they were able to obtain very decent results but required further improvement in error covariance modeling and microphysics. This chapter looks at the non-linearity issue in the observation operator, and the non-smooth switches in observation operator in the next chapter. A future work will address the issues in the non-smooth model parameterizations of 4D-Var implementations.

Wu et al. (2000, [227]) extended these results using a 4D-Var system assimilating dual polarization radar data with the inclusion of hail for the same model and optimization algorithm. They reported results that were orders of magnitude worse than Sun and Crook, ostensibly due to the inclusion of ice microphysics. They also reported that the problem may be with 4D-Var itself "which inherently depends on linear approximations, (may be) unable to adequately assimilate the observations." This statement is slightly inaccurate in that at each step of the 4D-Var optimization, the algorithm relies upon the tangent linear model to come up with a search direction; for the smooth case, if the search direction is consistent with the forward model and is a descent direction, with a proper line-search 4D-Var is guaranteed to converge to a local minimum. No such guarantee exists with the non-smooth case with L-BFGS, which they investigated. Their cost history graph shows some irregularities with a non-uniform descent, which descent algorithms such as L-BFGS do not usually allow, implying that there is either an issue with the gradient of their cost function or their implementation of L-BFGS. Montmerle et al. (2001, [157]), Austin and Stephens (2001, [11]), Weygandt et al. (2002, [221], [222]), and Benedetti et al. (2003, [23] and [24]) perform similar studies but with better results.

Snyder and Zhang (2003, [191]) investigated using the Ensemble Kalman Filter (EnKF) without perturbed observations. They assimilated synthetic radar data with warm rain microphysical variables using the same model as Sun and Crook (1997, [198]). They also used an idealized, nearly linear observation operator that simulates radar reflectivities. They found that even though their observation operator was slightly non-linear, EnKF was still able to handle these observations and reconstruct the hydrometeor profiles. By using EnKF, they were able to avoid the need for a model adjoint or an observation operator adjoint and were also able to gather realistic sample error covariances.

Tong and Xue (2005, [204]) extended the work of Synder and Zhang by using the more advanced ARPS model with five species of microphysical variables including ice crystals which is quite similar to [92]. They used 100 ensemble members with observations assimilated every 5 minutes and obtained excellent results, even to the point of being able to distinguish different phases of the cloud. This work unfortunately used multiple observations (100) with the mean as the correct true observation and the correct observation error covariance were taken. As shown in the chapter on data assimilation, in the best case realistic scenario only one observation would be available with the noise already included; using multiple observations with the correct covariance model for each ensemble to improve the ensemble statistics is in fact "cheating." This is a flaw of all perturbed-observation EnKF studies in general, and thus only studies such as Synder and Zhang where unperturbed observations are used (or the covariance is doubled in the perturbed observation) should be relied upon. Caya et al. (2005, [37]) performed a similar study to compare 4D-Var and EnKF but did not use perturbed observations. Unfortunately they unrealistically only gave 4D-Var access to 3 observations of information (10 minutes) at a time, thus hampering it almost to the point of a "3.5 Var" method. Since 4D-Var theoretically (if not numerically) deals best with assimilating as long of an observation window as possible (especially since the background error covariance $B$

is only valid for 4D-Var at the initial time), this stacked the deck against 4D-Var somewhat. Even so, they found that 4D-Var and EnKF were very competitive.

Koizumi et al. (2005, [113]) describes an implementation of the Japanese Meteorological Agency (JMA) 4D-Var system with radar assimilation. They found that the assimilation of this data improves both precipitation and improves cloud location.

In Lopez and Bauer (2007, [139]) the operational 1D-Var+4D-Var assimilation scheme was tested on radar and hourly precipitation gage data. In this approach, rain rates were first assimilated through 1D-Var, and these were then used as "psuedo-observations" in the four-dimensional scheme. Lopez (2011, [138]) describes the full operational 4D-Var assimilation of National Centers for Environmental Prediction (NCEP) stage IV radar and rain gauge precipitation. A main conclusion of the paper is that assimilating six-hour precipitation accumulation rather than hourly data gives significantly better performance. This phenomenon may be attributable to the smoothing feature of the averaging operation, and thus the highly non-linear and non-differentiable "on/off" switches associated with precipitation can be appeased to some degree. Another major result of this paper is that only areas where both the background and model observations are precipitating are rainy can be assimilated. This dissertation will touch on these issues to some extent.

A plethora of studies of various different methods of assimilation of radar data have recently been conducted. These include Xiao et al. (2007, [230]) with 3D-Var; Kawabata et al. (2007, [109]) with 4D-Var with no microphysical variables; Sugimoto et al. (2009, [197]) with 3D-Var; Aksoy et al. (2009, [2]) with EnKF; Caumont et al. (2010, [36]) with 1D-Var+3D-Var which detailed microphysics – the benefit of the 1D-Var approach they use is that the model adjoint is not needed, while the benefit of the 3D-Var method is that the model adjoint is not needed; Lee et al. (2010, [126]), Routray et al. (2010, [178]), Li and Mecikalski (2010, [134]), Ha et al. (2011, [76]), all with 3D-Var; Kawabata et al. (2011, [108]) with 4D-Var with full microphysical variables; Dowell et al. (2011, [52]) with EnKF; and Zhao et al. (2011, [236]) with 3D-Var.

**Modern microwave data assimilation.** Except under heavy precipitation, the atmosphere is semi-transparent in the microwave spectrum, meaning that the nature of the radiation is not drastically changed in the presence of hydrometeors. The observation is also not highly non-linear ([17]). The situation is quite different for infrared radiation, however, where clouds are almost – but not quite – a blackbody that perfectly absorb/emit infrared radiation. The assimilation of microwave radiation has particular challenges, however, including the fact that the surface emissivity and polarization over the ocean are related to the wind speed ([56]). Despite these differences, like radar, many of the methods that have been used for microwave data assimilation can inform our assimilation of infrared brightness temperature.

Smith et al. (1992, [187]) and Mugnai et al. (1993, [158]) set the theoretical framework for data assimilation using a cloud-resolving model with five cold microphysics and an observation operator based on a full single-scattering radiative transfer model. They describe the nature of the microwave brightness temperature problem in relation to microphysical variables in great detail, and set forth a theoretical possibility of using an optimization method with a weighted combination of training input profiles that they would then use to minimize the difference between observed and simulated data in a "least-squares sense." However, they appeared not to take several important practical issues into account such as observation error, adjoints, well-posedness, uncertainty quantification or optimization algorithms, but none-the-less put forth an impressive effort especially from the observation operator standpoint. Giglio (1994, [69]) describes a somewhat similar but more practical approach, with emphasis on selecting appropriate cloud models (e.g. over land or ocean) and nudging the results towards better fit with the observations using regression. Kummerow and Giglio (1994, [118]) present this algorithm applied to actual SSM/I data and showed fair correlation with radar, especially over the ocean.

Kummerow et al. (1996, [120], 2001, [119], used a microwave radiative transfer model without/with scattering, respectively, along with a Bayesian approach that is essentially 1D-Var without a background (making the problem ill-posed when there are more model variables than degrees of freedom, see e.g. Stuart (2010, [196]) in order to derive three-dimensional cloud estimates and develop a cloud database for use in data assimilation. They conclude that their method has significant biases. Similar studies include Olson et al. (1996, [165]), Haddad et al. (1997, [81]), Marzano et al. (1999, [147]), Bauer (2001, [14], [15]), Di Michele et al. (2003, [49]), Grecu et al. (2004, [71]), Masunaga and Kummerow (2005, [148]), Di Michele et al. (2005, [50]), and Chiu and Petty (2006, [44]). Zhou et al. (2007, [239]) compared that cloud-databases based on these methods were not very accurate; a cloud-ensemble model run over 30 days without data assimilation (i.e. with no observations used) provided better statistical estimates of precipitation than those based directly on the observations. Ebert et al. (2007, [54]) reported similar results. This may be indicative of: the importance of including a background in the inversion problem; the need for improved models, microphysical schemes or prior distributions in the retrievals; the issue of the so-called "spin-up" problem whereby the model is not initialized in a dynamically consistent way; the fact that synoptic-scale dynamics have more impact on cloudy data than individual columns; or most likely some combination of these factors.

Fillion and Errico (1997, [63]) first analyzed the theoretical 1D-Var assimilation of derived rain-rate products for SSM/I microwave data. Using temperature, pressure, velocity, humidity and pressure on five levels as their 20 control variables, they used the M1QN3 L-BFGS implementation to minimize the 1D-Var cost function. Because they worked with derived products (which already contain major inaccuracies), their observation operator was simply the model parameterization schemes. They were able to make significant adjustments of model precipitation based on synthetic observations, and presented important theoretical contributions to the 1D-Var of rainy microwave data including a variable transform based on the eigenvalue decomposition of the background error covariance. Fillion and Mahfouf (2000, [64]) and Marecal and Mahfouf (2000, [146]) re-examined the 1D-Var technique on the Tropical Rainfall Measuring Mission (TRMM) microwave satellite using derived instantaneous surface rainfall rate products with near linear observation operators. They found that in areas where there was no precipitation in the background, 1D-Var was not able to trigger precipitation; a related issue will be addressed in this dissertation.

Aonashi and Liu (1999, [8]) were the first to investigate the direct assimilation of all-sky SSM/I polarized microwave radiance data using 1D-Var and a full radiative transfer model. Using several simplifying assumptions (including that the temperature was constant) and parameterizations, they assimilated both a derived rain flag product as well as the microwave brightness temperature using the control variables of divergence and precipitable water in rainy areas and total water in precipitation-free areas. Their four-stream microwave radiative transfer model allowed for cloud water, rain, snow and ice. They used the Gauss-Newton optimization algorithm, which is notoriously unstable, but were still able to obtain improved precipitation forecasts but did not see an improvement in humidity forecasts. Other authors such as Xiao et al. (2000, [229]) and Wiedner et al. (2004, [224]) also assimilated this data in a similar fashion.

Treadon et al. (2002, [205]) describes how NCEP has introduced rain retrievals from rain-affected microwave data into the operational global data assimilation system, while Tsuyuki et al. (2002, [207]) describes similar advances by the JMA.

Greenwald et al. (2002, [72]) describe a full all-weather radiative transfer observation operator and adjoint appropriate for visible and infrared data assimilation, while Weng and Liu (2003, [219]) discuss a microwave radiative transfer model and adjoint appropriate for data assimilation. Heidinger et al. (2006, [86]) and O'Dell et al. (2006, [164]) describe an observation operator and adjoint appropriate from the microwave to the infrared spectrum.

Bauer et al. (2005, [20]) present results on the accuracy of retrieving hydrometeor profiles using 1D-Var with synthetic microwave data and a full multiple-scattering radiative transfer model. They use the M1QN3 L-BFGS algorithm, and consider rain, snow and cloud, and found that they were able to achieve accurate retrievals. Bauer et al. (2006, [18], [19]) give an operational implementation of 1D-Var+4D-Var all-sky microwave assimilation at ECMWF, which has since been replaced by a full 4D-Var implementation (2010, Bauer et al. [17], 2010, Geer et al. [68]). In this implementation they use a full multiple-scattering radiative transfer model, RTTOV, which is used in this dissertation for infrared data instead. Again, unlike infrared data, this observation operator for microwave data is nearly linear. They showed that the 1D-Var+4D-Var and 4D-Var systems showed similar performance, with the 1D-Var actually achieving better results for precipitation data. They once again conclude that creating a cloud is much more difficult than removing it. Weng et al. (2007, [220]) describe a similar 1D-Var+4D-Var variational scheme for the assimilation of rain-affected microwave data, but they use the Community Radiative Transfer Model (CRTM), another multiple scattering model appropriate for the data assimilation of cloudy data. Boukabara et al. (2007, [32]) describes an all-weather 1D-Var microwave implementation.

Meirold-Mautner et al. (2007, [152]) compared the output from a radiative transfer model to that of observed satellite and infrared data and found good agreement. They concluded that the microwave observations were very sensitive to microphysical parameters such as snow dipole parameters, thus providing a good basis to evaluate (and in turn assimilate) these parameters within models.

Hou and Zhang (2007, [93]) generated a 4-year analysis by using a 1D-Var "variational continuous assimilation" that optimized the model error rather than the initial condition. The model was related to the observations through a simplified moist physics, and model error alone was used to account for the differences. The benefit of this approach is that a perfect model is no longer assumed, which for precipitation and cloud models is a quite poor assumption. However, by not considering any error in the initial conditions, however, the model dynamics are neglected entirely. While Hou and Zhang recognized the benefit of considering both types of errors, they nonetheless showed improvement in precipitation data by only considering model error.

Zupanski et al. (2011, [247]) discussed a prototype ensemble system based on WRF for the assimilation of downscaled microwave precipitation data using MLEF. They used horizontal wind components, temperature, pressure, water vapor, and five classes of hydrometeors. They included data from the AMSR-E and TMI satellites and used a delta-Eddington two-stream radiative transfer model with slant view. They tested a tropical storm case and found that the assimilation of precipitating radiances caused excessive surface rainfall, highlighting the need for improved microphysical schemes. Other studies using microwave data to initialize WRF have been conducted include Singh et al. (2011, [184]) with 3D-Var and AMSU data.

**Modern infrared data assimilation.** We now come to modern infrared data assimilation and its particular challenges of high-degrees of non-linearity.

A near-modern attempt at assimilating infrared data was Bayler et al. (2000, [21]). A successive correction algorithm was used for cloud initialization using infrared GOES sounder data with an 80 km resolution, Kessler microphysics (cloud, rain, and water vapor species) and a Kuo cumulus parameterization scheme. However, rather than using actual microphysical variables, Bayler et. al. used relative humidity above saturation to represent clouds. In addition, the Cressman scheme of successive correction was used, meaning that the observations were taken to be perfect and these observations were weighted by the distance from the analysis point. They used the $CO_2$ slicing and split window techniques to derive cloud-top pressure and effective cloud amount. Li et al. (2001, [133]) published a similar study.

Janiskova et al. (2002, [96]) made preliminary investigations of the 1D-Var data assimilation of temperature, humidity, and pressure as proxies for the development of stratiform clouds. Most importantly for this work, careful development of the linearization of the observation operator and cloud parameterization was described. This work used M1QN3, an implementation of the limited-memory BFGS algorithm. The use of cloud microphysical variables as control variables was discussed, but it was decided that directly including these variables would be too difficult due to the need for statistics of background errors and performance considerations. Finally, the authors noted a difficulty in triggering new clouds when the atmospheric state was far from the conditions of cloud formation.

McNally (2002, [150]) measured the cloud-modified, adjoint-derived sensitivity of the ECMWF model error to clouds, and found that there was a high degree of correlation between areas of baroclinic instability (which contributes to model error) and clouds. He also concluded that the use of hyper-spectral cloudy IR radiances could have a large impact on improving model error, although a great deal of effort would be needed to address various issues. The issues he noted include specifying background error statistics, optimally selecting which channels to use, and the large sensitivity of the observation to errors in cloud cover fraction.

Szyndel et al. (2004, [200]) examined 1D-Var assimilation of cloud data using a simplistic single-layer cloud model with an ensemble of related synthetic satellite observations. Their control variables consisted of an atmospheric profile without clouds, effective cloud amount and cloud-top pressure of their single layer cloud using 4–6 infrared channels. They used the minimum-residual method as a background first guess for their 1D-Var approach. The effective cloud amount and cloud-top pressure were "clamped" to between 0.01 and 1.0, respectively. As detailed below, this clamping can cause issues for the convergence of the algorithm, which indeed Szyndel et al. acknowledged. Regarding optimization algorithms, they used a Gauss-Newton method and the Levenberg-Marquardt method for least-squares fitting, both of which require storage on the order of $N^2$, where $N$ is the number of control variables. Since the only control variables used in their study were cloud-top pressure and effective cloud amount, this is not an issue. In this dissertation, where the cloud variables are considered at each cloud-level, such optimization methods become computationally infeasible, and the limited-memory approaches detailed above become vital to success.

Chevallier et al. (2004, [42]) investigated the issue of 4D-Var of cloud-affected AIRS and MVIRI infrared radiances using control variables of temperature, humidity, ozone, surface temperature and pressure. Like Janiskova et al. (2002, [96]), the decision was made not to include cloud variables as control variables due to the difficulty of specifying background errors and concerns about the non-linearity of the observation operator. The observation operator was kept simple "so that thresholds and strong nonlinearities do not make the 4D-Var minimization stop before reaching the absolute minimum of the cost function." This statement alludes to the fact that discontinuities and highly non-linear observation operators can cause the algorithms to terminate in their line searches, especially if they are not set in the context of non-smooth optimization (however, 4D-Var in general only converges to local minimums, not absolute ones). The model thus did not take scattering into account, but was rather based on a simple cloud-top pressure model. 35 "near-linear" channels of AIRS were chosen based on several criteria, and the recommendation is to assimilate only 3 regions of AIRS. In a similar study, Tompkins and Janiskova (2004, [203]) describe a variational model that determines cloud cover from infrared observations based on parameterizations and statistical properties.

Vukicevic (2004, [213]) investigated assimilating visible and infrared measurements for mesoscale cloud-state estimation using a 4D-Var algorithm with the RAMS model with explicit microphysics and an explicit visible and infrared radiative transfer observation operator. Vukicevic included model error, and found that she was able to achieve some improvement in cloud cover when compared to the background state; however, not much improvement was seen when the background state was clear. Greenwald (2004, [73]) studied the adjoint sensitivity of three infrared channels and found that these channels are sensitive to microphysical parameters. Wei et al. (2004, [216]) studied the AIRS channels and found the same. Vukicevic (2006, [214]) found that increasing the number of channels and frequency of observation had a clear impact on improving the assimilation results, that infrared could not trigger clouds, and that a simple linear model error approach was insufficient for controlling the error in boundary conditions.

Li et al. (2005, [131]) treat the 1D-Var problem with full microphysical particles including particle size and radiative transfer for AIRS data. Similar studies on the benefits of assimilating hyperspectral infrared channels were conducted by Smith et al. (2005, [190]) and Zhou et al. (2005, [237], 2007, [238]). [193].

Errico et al. (2007, [57]) reports on the outcome of a 2005 international workshop where the issues in assimilating cloudy satellite transfer was discussed. A variety of issues were identified that needed to be improved including: issues with the observations (including improving utilization of millimeter data); issues regarding models (including improved microphysical schemes, especially with ice); issues with radiative transfer (including quantifying satellite biases and standard deviations, using improved microphysical schemes); and issues with the data assimilation itself (including how to deal with highly non-linear and non-smooth processes, moving beyond perfect model assumptions, and improving uncertainty quantification). Errico et al. emphasizes the truly interdisciplinary nature of this problem.

Lopez (2007, [137]) reviews the issues facing the variational assimilation of cloud and precipitation radiance data from the perspective of the models. Based on time-scale arguments, he concludes that it is sufficient to assimilate humidity values as a proxy for cloud microphysics when the clouds are precipitating or convective, but for cirrus clouds the microphysical variables are needed. Lopez stresses the difficulties in dealing with a highly nonlinear observation operator. Lopez says that "improved linearity is usually achieved either by using smooth functions to describe each physical process or by artificially reducing or neglecting the perturbations of problematic quantities involved in the parameterization." This dissertation treats the non-linearities directly.

Weisz et al. (2007, [217]) demonstrate a model for determining cloud-top pressure for AIRS data based on training of a global data model. They use an eigenvalue regression to determine regression coefficients for either water or ice clouds. Li and Liu (2009, [132]) used this model and showed improved tracking of a hurricane using AIRS infrared data assimilation with the WRF/-DART testbed, which uses an Ensemble Adjustment Kalman Filter.

Heilliette and Garand (2007, [87]) describe a method for assimilating infrared radiances using a highly simplified cloud model with four parameters: single layer cloud height, 15 $\mu$m effective emissivity, and effective particle size of water and ice. Scattering was not considered as they used the RTTOV-8 model. They used a 1D-Var formulation for full columns of water vapor and temperature in addition to the four cloud parameters mentioned above. They considered that these four cloud parameters were independent and thus had a diagonal term in the background error covariance matrix $B$, and assimilated synthetic radiances corresponding to 100 channels of AIRS. Starting from a first guess for cloud parameters based on the $\mathrm{CO}_2$ slicing method, they found that they were able to reduce the variance of these variables and have a significant impact on the retrieved values. Pavelin et al. (2008, [169]) performed a similar study and found that while they achieved positive results for temperature and humidity, their cloudy parameters showed less accuracy when multiple channels were used, attributed in part to the lack of multiple scattering.

McNally (2009, [151]) describes a near-operational 4D-Var assimilation of infrared brightness temperature using an infrared radiative model based on simple brightness temperatures (i.e. a single-layer blackbody cloud with no multiple scattering). The background for this study was chosen using the minimum-residual method ([60]) using two channels. Only completely clear or overcast values were considered. i.e. no cloud fraction was chosen. By using the minimum residual method, no error covariances on the cloud-top were required, and only the cloud-top pressure was adjusted. This procedure showed slight positive impact on the RMS temperature and humidity. Pangaud et al. (2009, [168]) publish a similar study and show positive impact on geopotential up to 72 hours.

Seaman et al. (2010, [182]) use 4D-Var system to assimilate infrared radiances, and find that if no cloud is present in the initial condition, the analysis proceeds as if no cloud was present. In this dissertation we explain why this occurred and an easy remedy.

Zupanski et al. (2011, [249]) describe a system for assimilating synthetic GOES-R infrared data in cloudy conditions with the WRF model. This was a "non-identical twin" experiment as they used the RAMS model for creating their observations. Using a single channel of infrared data with a two-stream delta-Eddington radiative transfer model with a cloud optical property model, they examined the assimilation of a hurricane with potential temperature, specific humidity, and five hydrometeor classes as control variables, and experimented with leaving different hydrometeor classes out. They concluded that it is essential to use as many classes of hydrometeor as possible in order to obtain the maximum benefit of assimilating cloudy radiances, with the possible exception of rain or graupel. Polkinghorne and Vukicevic (2011, [171]) describe a 4D-Var system for the assimilation of GOES-8 infrared cloudy radiances. A cloud-mask is first used so that only the same cloud type is considered. They performed a wide variety of experiments including varying the assimilation window. They conclude that "while increasing the length of the assimilation window does not lead to a greater decrease in cost function, it does lead to a smoother dynamical response to the assimilation and a better forecast." That increasing the assimilation window does not decrease the cost function is somewhat apparent since the cost function is a non-decreasing function of assimilation window, i.e. the more non-exact observations considered, the greater the value of the cost function will be. They conclude that their main hindrance to achieving better errors was the cloud location problem, which may be fundamentally related to their cloud mask approach. It is probably better to solve the cloud location problem through an all-sky observation assimilation approach rather than determining an inaccurate cloud mask a priori; this is the approach is taken in this dissertation.

Finally, Bauer (2011, [16]) concludes with a survey of the all-sky and microwave and infrared schemes used operationally across the world. The focus is currently still on single-layer clouds for infrared data across the world, although research is being conducted into full gray cloud, multiple scattering infrared radiances. Most operational centers only assimilate fully clear or fully cloudy-skies, and effective cloud amount is not considered. Lavanant et al. (2011, [122]) discuss the cloud-derived products that are available for the Infrared Atmospheric Sounding Interferometer (IASI) satellite.

Thus concludes what is hoped to be a near comprehensive survey of the state-of-the-art in cloudy-sky assimilation.

### 4.1.3 Atmospheric Infrared Sounder

The Atmospheric Infrared Sounder (AIRS) satellite was launched on May 4th, 2002 by NASA with the intent of providing high-resolution data for improving climate and weather prediction. It has 2378 channels that range from 3.75 $\mu$m to 15.4 $\mu$m in wavelength ([10]). Each of these channels has a different *weighting function* or sensitivity that peaks at a different location in the atmosphere – figure 4.1 shows the weighting function for AIRS channel 300 (wavenumber 735.690 cm$^{-1}$ or a wavelength of 13.593 $\mu$m), and figure 4.2 shows the location in the atmosphere where the weighting function for each channel peaks. Both of these profiles were generated for a maritime tropical clear-sky profile (see section 4.3 below) with the default CO2 profile from RTTOV (see section 4.1.5). Both of these graphs were calculated by a model based on the HITRAN 2008 line-by-line database ([177]) using only the primary isotopologues of $CO_2$ and $H_2O$. In these graphs, the somewhat important effects of other gases such as $O_3$, $N_2O$, $CH_4$, and CO have been neglected, and thus should be taken as only an illustration of the fact that the weighting functions peak at different levels in the atmosphere. In this dissertation outside of this section, we do not use a line-by-line model, and instead use RTTOV; we assume the default RTTOV profile for all of the atmospheric constituents enumerated above except for water.

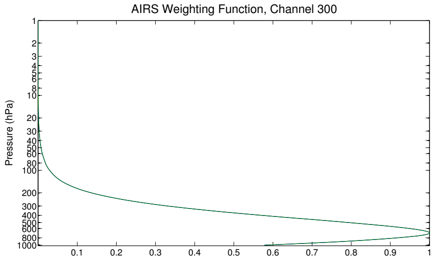


Figure 4.1: Weighting function for AIRS channel 300

Because each channel has a different weighting function with different sensitivity to the atmosphere, in aggregate these channels can be used to reconstruct an atmospheric profile. These 2378 wavelengths will be used in this experiment to provide synthetic observations that are generated from RTTOV v10, i.e. only infrared satellite observations will be used.

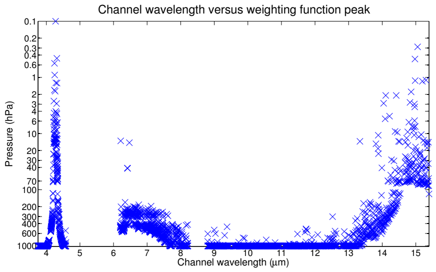


Figure 4.2: Channel wavelength versus channel peak pressure ($CO_2$ and $H_20$ only, clear sky profile)

### 4.1.4 Weather Research and Forecasting model

For these experiments, we use the Weather Research and Forecasting (WRF) model ([186]) to generate sample profiles of the atmosphere. WRF is a limited-area NWP model that has been used in many applications. In this experiment we use the WSM 6-class microphysics scheme ([92]) which has six species of hydrometeors: vapor, liquid cloud, rain, snow cloud, ice cloud, and graupel (hail). Since this experiment focuses primarily on RTTOV and the 1D-Var problem, and WRF is merely used to generate the initial profiles as exact values, we refer to ([186]) for more information about WRF.

### 4.1.5 RTTOV Version 10

In this experiment we use RTTOV version 10 as our radiative transfer model. Longwave radiation coming from the earth makes its way up to the top of the atmosphere, being absorbed along the way. In addition, the atmosphere itself emits radiation which is also partially absorbed by layers above it. This emission and absorption depends on pressure, temperature, wavelength, and the properties of gases such as water vapor, carbon dioxide, clouds and aerosols. A satellite pointed towards the Earth's atmosphere measures the upwelling radiation over a spectrum of frequencies. RTTOV uses atmospheric variables at various levels of the atmosphere to provide synthetic measurements of radiance that can be compared against the actual measurements from a satellite, and these measurements can then be used for the purpose of reconstructing the atmospheric profile, i.e. data assimilation can be performed.

RTTOV is the fast radiative transfer model originally written for the Tiros Operational Vertical Sounder (TOVS) in the early 1990s. The model was created to allow rapid calculations of radiative transfer for infrared and microwave satellites. It allows specification of temperature, gas concentrations (including water vapor, clouds and aerosols), as well as cloud and surface properties. RTTOV can either interpolate these values to user-defined pressure levels or use its own fixed pressure levels. The native pressure levels extend well into the mesosphere (with a minimum pressure of 0.005 hPa) ([89]).

The RTTOV model is trained using more time-consuming line-by-line and multiple-scattering models in order to derive a regression model. This means that RTTOV itself does not perform line-by-line or multiple-scattering calculations. Instead, the prediction scheme is calculated as a deviation from a reference profile, and is given by

$$d_{i,j} = d_{i,j-1} + \sum_{k=1}^{K} a_{i,j,k} X_{k,j} \tag{4.1}$$

where $d_{i,j}$ is the optical depth from level $j$ to space for the satellite channel $i$, $K$ is the number of predictors, and $X_{k,j}$ is the predictor $k$ and level $j$. Examples of potential predictors include the temperature, water vapor amount, cloud amount, carbon dioxide, etc. at the given level, although not all variables are required as defaults can be used. Finally, $a_{i,j,k}$ are the regression coefficients ([179]).

For clear-sky IR radiation, scattering can be neglected ([192]). However, for accurate simulation of cloudy-sky radiances – the target of this research – scattering must be taken into account, especially for optically thin clouds. In RTTOV version 10, the latest version at the time of publication, multiple scattering of clouds is included for the IR spectrum, which allows radiances to be calculated for multiple layer clouds. RTTOV is parameterized for six types of clouds: maritime stratus, continental stratus, maritime cumulus, continental clean cumulus, continental polluted cumulus, and cirrus (ice) clouds. Note that these are different designations than the clouds used in WRF, as WRF deals with cloud microphysics, while RTTOV has been parameterized according to large-scale cloud designations. Absolute cloud concentration (in g/m$^3$) and cloud fraction (0 - 1) at each layer can be specified for a maximum of two types of clouds per layer ([89]). The adjoint and tangent linear model of RTTOV is also available, and is detailed in ([89]). This makes RTTOV suitable at least in theory for cloudy-sky variational assimilation. This chapter will put that theory to the test in a realistic way.

## 4.2 Experimental setup

The details of our experimental setup are detailed here, while the numerical results are presented in the next section.

### 4.2.1 Choice of control variables

In this subsection we consider which atmospheric variables should be used as control variables. Under clear skies, the most important variables for infrared radiation are temperature and water vapor. It is apparent these two variables should be included as control variables. Pressure is another important variable, but since in this experiment we use the predefined RTTOV pressure levels to specify our remaining variables, we can treat pressure as a constant and not include it within the control variable set. Carbon dioxide and ozone also have an important impact on the satellite observations, especially near the 4.3 and 15 $\mu$m carbon dioxide band and the Wulf ozone bands (e.g. [192]), but in this study we choose to use the default RTTOV profiles.

Because RTTOV allows only two species of cloud to be specified at each layer, unfortunately not all of the five cloud hydrometeor species available from WRF can be integrated with RTTOV in the case of a mixed-phase cloud. Rather than trying to keep track of five variables in one case, and only two variables in the case of a mixed-phase cloud, in this study we simplify matters by combining QVapor (typically used for cumulus clouds) with QSnow (typically stratus clouds) from the WRF simulation. We assign this combined concentration to either maritime cumulus or continental clean cumulus, using the WRF LANDMASK variable to determine sea or land. As shown in section

4.3.1, the brightness temperature from RTTOV responds in a similar way for cumulus and stratus parameterizations, so we are somewhat justified in combining these two species together. We use QIce for cirrus clouds. This leaves only QRain and QGraup. In this study, we choose a cloudy profile that does not include rain or graupel in order to avoid the lack of a corresponding parameterization in RTTOV. These issues would need to be addressed before attempting full-scale assimilation; since rain and graupel would usually not be near the cloud top, and would be most likely to appear with cumulus clouds ([176]), one strategy would be to simply ignore them, and another would be to include their concentrations with the cumulus concentrations.

Cloud fraction at each level is also an important control variable. For RTTOV clouds, a maximum-random cloud overlap is used ([149]), and the cloud fraction refers to the fraction of clear versus cloudy radiance to use for all types of cloud at that particular layer. In other words,

$$d_j^{\text{total}} = C_j d_j^{\text{cloudy}} + (1 - C_j) d_j^{\text{clear}} \tag{4.2}$$

where $d_j^{\text{total}}$ is the total optical depth from level $j$ to space, $C_j$ is the cloud fraction at level $j$, and $d_j^{\text{cloudy}}$ and $d_j^{\text{clear}}$ are the cloudy and clear optical depths from level $j$ to space.

While cloud fraction is included in the control set, the presentation below will be simplified by showing effective clouds – that is, the cloud mixing ratio (kg/kg) for either QCloud or QIce times the cloud fraction at each level, i.e.

$$\text{QCloud}_{\text{eff}} = C_j \text{QCloud} \tag{4.3}$$

$$\text{QIce}_{\text{eff}} = C_j \text{QIce} \tag{4.4}$$

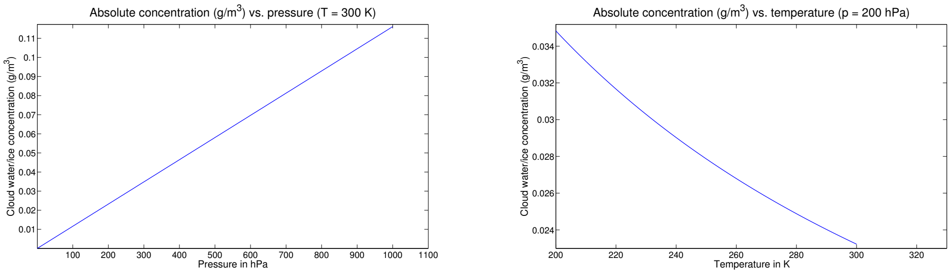
One final issue of clarification is that while WRF uses the reasonable unit of mixing ratio (kg/kg), which is the ratio of the mass of cloud species per kilogram of dry air, unfortunately (from a user's perspective) RTTOV chose to parameterize in terms of absolute concentration (g/m$^3$), which is the mass of the species per cubic meter of air. Because air at different pressures and temperatures has different volumes, unlike mixing ratio, the absolute concentration depends on both pressure and temperature, with the same amount of cloud species having a larger absolute concentration at higher pressures and a smaller absolute concentration at higher temperatures (see figure 4.3). In the graphs and results below, the cloud results will be given in mixing ratio (kg/kg) so as to avoid this bias. However, during the optimization process the control variables themselves use the native units (g/m$^3$) of absolute concentration and are only converted back to mixing ratio (kg/kg) in post-processing.

The procedure for converting from mixing ratio (q) to absolute concentration (ac) is to calculate the partial vapor pressure ($p_v$) in Pa, which is the contribution to pressure due to the cloud species:

$$p_v = \frac{\text{MR} * p}{\text{MR} + \text{M}_{\text{wv}}/\text{M}_{\text{air}}} \tag{4.5}$$

where $p$ is the overall pressure in Pa, $\text{M}_{\text{wv}} = 18.0152833$ g/mol is the molecular weight of water (which is the same for vapor, cloud, snow, or ice), and $\text{M}_{\text{air}} = 28.9644$ g/mol is the molecular weight of dry air.

The absolute concentration ac can then be found as



(a) Abs. concentration vs. p

(b) Abs. concentration vs. T

Figure 4.3: Absolute concentration of 1.0e-4 kg/kg of cloud species versus pressure and temperature. The pressure graph is fixed at 300 K, while the temperature graph is fixed at 200 hPa.

$$\mathrm{ac} = \frac{p_v * \mathrm{M_{wv}}}{\mathrm{T} * \mathrm{R}} \tag{4.6}$$

where T is the temperature in K and R $= 8.31447215$ m$^3$ Pa K$^{-1}$ mol$^{-1}$ is the universal gas constant.

The conversion back from absolute concentration to mixing ratio simply reverses this process and is straight-forward.

In summary, we choose the following five control variables, each specified at each of the $n_{\mathrm{lev}}$ native pressure levels of RTTOV: temperature (T), water vapor (QVapor), cloud fraction, water cloud (QCloud), and ice cloud (QIce). Cloud fraction is rolled into the water and ice cloud concentrations and presented as effective water cloud and effective ice cloud.

### 4.2.2 Twin experiment

In this chapter, we use a so-called "twin experiment" (see e.g. [58], [66]); that is, we start with our exact solution, extract observations of the exact system through our observation operator (in this case RTTOV), and perturb both the initial condition and the observations. We then use the perturbed state and observations in our data assimilation algorithm and compare the solution to our exact solution. This methodology allows us to assess and compare the success of our various methods.

We assume an exact background error and observation error covariance model, meaning that we perturb our true solution and observation so that the perturbations are Gaussian and have the same error covariance as the background error covariance and observation error covariance matrix, respectively.

### 4.2.3 Penalized 1D-Var Cost Function

Because RTTOV is based on a regression model, there are limits to which the regression was applied. RTTOV will take any input data that falls outside of these limits and will, depending on the option set, either generate an error or move it back within the limits. Because optimization algorithms will use a line search to explore values in the control space along a certain direction that may take it outside of the feasible region, it is important that the model being used does not generate errors in such a case, and therefore the only choice is to allow the values to be moved within the regression limits. This can confuse the optimization algorithm, which would otherwise have no knowledge of these regression limits, and instead see that the cost function appears to take on a constant value with respect to the out-of-bounds variable. What's worse, a constant value means a zero-gradient – the termination criteria for gradient-based optimization algorithms – and so in the worst case the algorithm may suspect that it has found a local minimum and terminate!

To avoid this highly undesirable situation, it is clear that the tools of constrained minimization are necessary. There are many methods for achieving some form of constrained minimization that range in complexity and effectiveness (see e.g. [162]). Because the true minimum would conceivably always fall well inside the regression limits, and the case when the minimum lies on the boundaries is by far the most difficult case, the "big guns" of constrained optimization such as the sequential quadratic programming or augmented Lagrangian methods may be overkill in this case. Thus, the relatively simple quadratic penalty method (again e.g. [162]) is chosen, and its formulation is:

$$J(x) = \frac{1}{2}\delta_b(x)^{\mathrm{T}} B^{-1}\delta_b(x) + \frac{1}{2}\delta_y(x)^{\mathrm{T}} R^{-1}\delta_y(x) + \frac{\mu}{2}\delta_p(x)^{\mathrm{T}} L^{-1}\delta_p(x) \tag{4.7}$$

where the variables are the same as in chapter 3, although here $x \in \mathbb{R}^{n_{\text{state}}}$ is both the control and state variable and contains each of T (temperature in K), QVapor (water vapor in ppmv), QCloud (cloud water vapor absolute concentration in g/m$^3$), and QIce (ice absolute concentration in g/m$^3$) at each of the $n_{\text{lev}}$ RTTOV pressure levels; and $\delta_p(x) \in \mathbb{R}^{n_{\text{state}}}$ is the deviation from the feasible area, and is given by $\delta_p(x) = x - P(x;\ell;u)$, where $P(x;\ell;u)$ is the projection of $x$ into the hypercube defined by $(\ell, u)$. In addition, because of the clamping to the feasible region, $\delta_y(x) = y - H(P(x;\ell;u))$. Here again $B \in \mathbb{R}^{n_{\text{state}}} \times \mathbb{R}^{n_{\text{state}}}$ is the background error covariance, $R \in \mathbb{R}^{n_{\text{obs}}} \times \mathbb{R}^{n_{\text{obs}}}$ is the observation error covariance, and $L \in \mathbb{R}^{n_{\text{state}}} \times \mathbb{R}^{n_{\text{state}}}$ can be considered a "penalty error covariance," although it is probably more convenient to consider it a scaling so that variables of drastically different sizes are penalized in the same way. $R$ and $L$ are diagonal matrices, with

$$R_{i,i} = \sigma_{obs}^2 \tag{4.8}$$

and

$$L = \begin{pmatrix} L_{\text{T}} & 0 & 0 & 0 \\ 0 & L_{\text{QVapor}} & 0 & 0 \\ 0 & 0 & L_{\text{QCloud}} & 0 \\ 0 & 0 & 0 & L_{\text{QIce}} \end{pmatrix} \tag{4.9}$$

where $L_*$ is a diagonal matrix and $L_{*_{i,i}}$ is a representative value for state level $i$ for the variable $*$ (T, QVapor, etc.) designed to bring the penalties to approximately equal values.

$\ell$ and $u$ are the lower and upper bound for the state variables. For T and QVapor, the lower and upper bounds are found in the RTTOV user guide ([89]). For QCloud and QIce, the lower bound is 0 and the upper bound is infinity. $\mu$ is a constant that scales the penalty term to a configurable level. The ability to change $\mu$ allows the impact of the infeasible penalty on the cost function to be set to a configurable level; the problem could also be solved repeatedly with the value of this penalty steadily increased, although this is seen as overkill due to the nature of this problem.

### 4.2.4 Background Error Covariance Model

The non-diagonal background error covariance model is an important part of the formulation of the data assimilation problem. In this work we use an exact background error covariance matrix to perturb the background $x_b$. We assume that the perturbation to the background vector is given by

$$x_b = x_{\text{true}} + \eta_x \tag{4.10}$$

where $x_{\text{true}}$ is the true state, $\eta_x \sim N(0, B)$, i.e. $\eta_x$ is a multi-variate normally distributed variable with mean 0 and covariance matrix $B$.

$B$ is the block matrix

$$B = \begin{pmatrix} B_{\text{T}} & 0 & 0 & 0 \\ 0 & B_{\text{QVapor}} & 0 & 0 \\ 0 & 0 & B_{\text{QCloud}} & 0 \\ 0 & 0 & 0 & B_{\text{QIce}} \end{pmatrix} \tag{4.11}$$

and

$$B^{-1} = \begin{pmatrix} B_{\text{T}}^{-1} & 0 & 0 & 0 \\ 0 & B_{\text{QVapor}}^{-1} & 0 & 0 \\ 0 & 0 & B_{\text{QCloud}}^{-1} & 0 \\ 0 & 0 & 0 & B_{\text{QIce}}^{-1} \end{pmatrix} \tag{4.12}$$

Each of the different $B_*$ matrices are created by the exponential squared kernel of

$$B_* = \Sigma_*^{1/2} \Sigma_*^{\text{T}/2} \tag{4.13}$$

and

$$\Sigma_{*_{i,j}}^{1/2} = \Sigma_{*_{j,i}}^{1/2} = \frac{\sigma_{*_i} + \sigma_{*_j}}{2} \exp(-r_{i,j}^2 / L_*^2) \tag{4.14}$$

Here, $\sigma_{*_i}$ is the standard deviation of the variable $*$ at level $i$, and $r$ represents the vertical distance in meters between two points, i.e.

$$r_{i,j}^2 = |H_i - H_j| \tag{4.15}$$

where $H_i$ and $H_j$ are the height, in meters, of levels $i$ and $j$ respectively. $L_*$ represents the correlation length, in meters, required for the correlation between two points of variable $*$ to reach $1/e \approx 0.36788$. The correlation lengths in this study are chosen based on heuristic, first-guess values to be (3000 m, 3000 m, 1000 m, 1000 m) for (T, QVapor, QCloud, QIce). Additional studies such as ([251]) could provide more accurate and realistic correlation lengths.

The normal perturbation $\eta_*$ that is a component of $\eta_x$ is created with the transformation $\eta_* = \Sigma_*^{1/2} Z_*$, where $Z_* \in R^{n_{\text{lev}}}$ and the components of $Z_{*_i} \sim N(0, 1)$ are independent identically distributed standard normal variables.

The size of $\sigma_{\text{T}_i}$ and $\sigma_{\text{QVapor}_i}$ are taken to be a percentage of the true profile values from section 4.3 with the percentage given in each test results. The size of $\sigma_{\text{QCloud}}$ and $\sigma_{\text{QIce}}$ are taken as a percentage of $10^{-4}$ kg/kg, which are then converted to absolute concentration for RTTOV. These same values (without the percentage) are also used as the penalty scaling for $L$.

### 4.2.5 Transformed cost function

While the matrix $B$ has, in theory, full rank, like many other background error covariances used in practice, this background error covariance matrix is numerically rank deficient due to the effect of having a much larger correlation length than the grid spacing. Thus, the inverse background error covariance cannot be computed numerically. This issue is addressed in our 1D-Var implementation by applying the change of variables transformation $\delta_b(x) = x - x_b = B^{1/2}z$ and using $z$ as the control variable such as in [63] and other studies. Since $B^{1/2}B^{\mathrm{T}/2} = B$, this removes the necessity of obtaining the inverse from the cost function, so that (4.7) becomes

$$J(z) = \frac{1}{2}z^{\mathrm{T}}z + \frac{1}{2}\,\delta'_y(z)^{\mathrm{T}}R^{-1}\delta'_y(z) + \frac{1}{2}\,\delta'_p(z)^{\mathrm{T}}L^{-1}\delta'_p(z) \tag{4.16}$$

where $\delta'_y(z) = \delta_y(B^{1/2}z+x_b) = y-\mathcal{H}(P(B^{1/2}z+x_b;\ell;u))$ and $\delta'_p(z) = B^{1/2}z+x_b-P(B^{1/2}z+x_b;\ell;u)$. The gradient of (4.16) is

$$\nabla_z J(z) = z - B^{1/2}\frac{\partial P}{\partial z'}\left(\frac{\partial \mathcal{H}}{\partial P}\right)^{\mathrm{T}} R^{-1}\delta'_y(z'(z)) + \mu B^{1/2}\left(I - \frac{\partial P}{\partial z'}\right) L^{-1}\delta_p(z'(z)) \tag{4.17}$$

where $z'(z) = B^{1/2}z + x_b$, but since

$$\left(\frac{\partial P}{\partial z'}\right)_i = \begin{cases} 1 & \text{if } z' \in (\ell_i, u_i) \\ 0 & \text{else} \end{cases} \tag{4.18}$$

and

$$\delta_p(z')_i = \begin{cases} 0 & \text{if } z' \in (\ell_i, u_i) \\ z' - P(z';\ell;u) & \text{else} \end{cases} \tag{4.19}$$

we have

$$\nabla_z J(z) = z - B^{1/2}\frac{\partial P}{\partial z'}\left(\frac{\partial \mathcal{H}}{\partial P}\right)^{\mathrm{T}} R^{-1}\delta'_y(z'(z)) + \mu B^{1/2}L^{-1}\delta_p(z') \tag{4.20}$$

### 4.2.6 Optimization settings

In this section we present the optimization settings used for each of our three algorithms. The optimization settings used for LMBM are shown in table 4.1. Details for these parameters can be found in the LMBM user's manual ([105]). The optimization settings for L-BFGS are shown in table 4.2, while the CG-Descent settings are shown in table 4.3. The L-BFGS parameters are described in [136], while an explanation of the CG-Descent parameters can be found in [82].

## 4.3 Test profiles

We test our data assimilation algorithm against two different profiles: one, a clear-sky profile taken over the Pacific ocean near the Baja peninsula, and a cloudy-sky profile of a cirrus cloud over the Gulf of Mexico, both from a WRF prediction of August 29 2011 at 00:00 UTC started from August 28 2011 at 00:00 UTC. Two maritime profiles were chosen at approximately the same

Table 4.1: LMBM optimization settings

| Parameter | Value | Parameter | Value |
|---|---|---|---|
| NA | 4 | MCU | 9 |
| MC | 9 | NW | Default |
| RPAR(1) | $10^{-3}$ | IPAR(1) | 2 |
| RPAR(2) | $10^{-8}$ | IPAR(2) | 300 |
| RPAR(3-5) | $10^{-2}$ | IPAR(3) | 300 |
| RPAR(6) | 1 | IPAR(4) | 1 |
| RPAR(7) | $10^{-2}$ | IPAR(6) | 0 |
| RPAR(8) | 1 | IPAR(7) | 1 |

Table 4.2: L-BFGS optimization settings

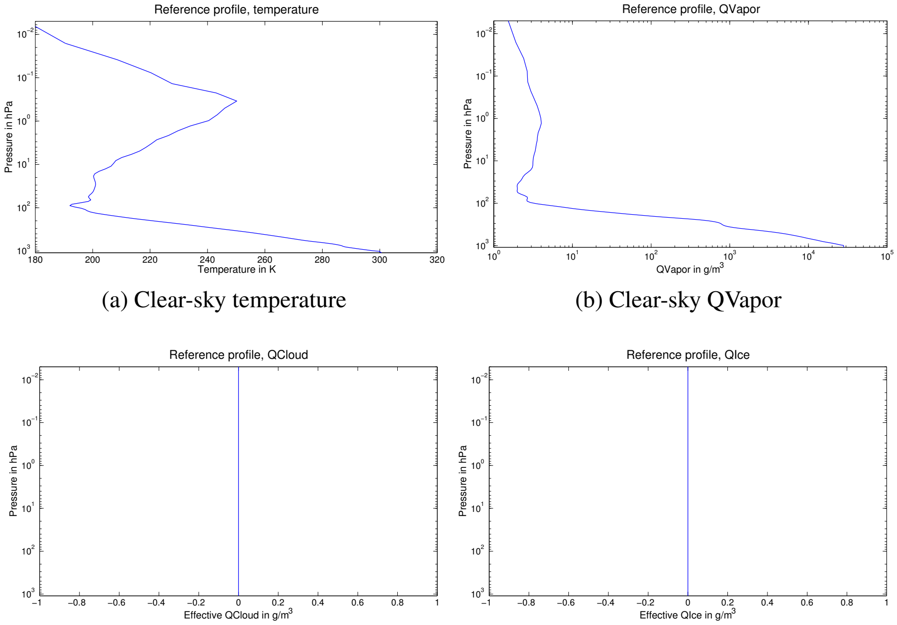
| Parameter | Value | Parameter | Value |
|---|---|---|---|
| MSave | 9 | $\epsilon$ | $10^{-1}$ |
| Max iter | 300 | diagco | false |

Table 4.3: CG-Descent optimization settings

| Parameter | Value | Parameter | Value |
|---|---|---|---|
| grad_tol | $10^{-3}$ | $\delta$ | $10^{-1}$ |
| $\sigma$ | 0.9 | eps | $10^{-3}$ |
| $\gamma$ | 0.66 | $\rho$ | 2.0 |
| $\eta$ | $10^{-2}$ | psi0 | $10^{-2}$ |
| psi1 | $10^{-1}$ | psi2 | 2.0 |
| QuadCutOff | $10^{-5}$ | StopFact | 0.0 |
| AWolfeFac | $10^{-3}$ | restart_fac | 1.0 |
| maxit_fac | 0.125 | feps | $10^{-3}$ |
| Qdecay | 0.7 | nexpand | 50 |
| nsecant | 50 | PertRule | true |
| StopRule | true | AWolfe | true |
| Step | false | debug | true |

latitude so that comparable pressure levels and temperatures could be compared, although of course the two profiles are from drastically different weather situations.

In order to simulate radiation coming from the earth, RTTOV uses pressure levels up to 0.005 hPa, which extends well into the thermosphere. Limited-area numerical weather prediction systems have historically focused primarily the troposphere (from the surface to approximately 100 hPa) while global weather systems extend into the stratosphere ($\approx$ 100 – 1 hPa); however, few NWP models have extended up into the mesosphere ($\approx$ 1 hPa to 0.01) or thermosphere, although this situation is changing for the very necessity of assimilating satellite radiance data ([55]). In this study, we only focus on the assimilation of a single column and use the 101 native RTTOV pressure levels for AIRS. The tropospheric and low stratospheric data are taken from a WRF simulation, while the remaining data is interpolated to be the same percentage as the top of the WRF data between RTTOV standard limits ([89]) at each pressure level. As discussed above, only temperature, water vapor, cloud concentration, ice concentration and cloud fraction are taken as control variables that are set at each level.



(a) Clear-sky temperature

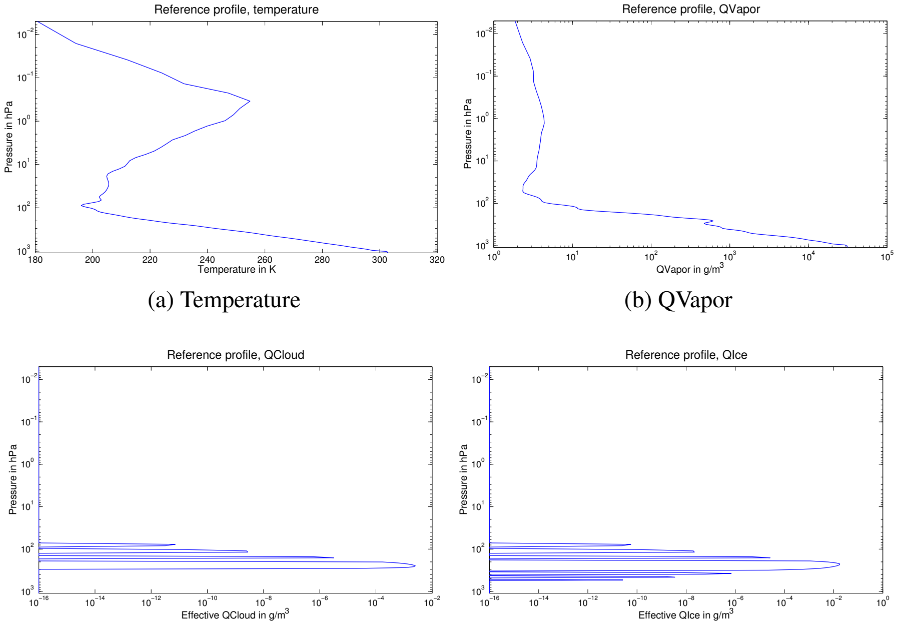
(b) Clear-sky QVapor

(c) Clear-sky QCloud

(d) Clear-sky QIce

Figure 4.4: Clear-sky profile for temperature, water vapor, effective cloud, and effective ice

Figure 4.4 shows the clear-sky profile, while figure 4.5 shows the cloudy profile used in this study. As these figures show, there is no ice or cloud in the clear profile, while the cloudy-profile has one large mixed-phase cirrus cloud that extends from around 300 hPa (approximately 9 km) to 200 hPa (approximately 11 km above the surface) and is thus approximately 2 km in depth. Several other smaller cirrus clouds are present above and below this height. The main cloud is optically thick and will prevent most if not all of the IR radiation from below this height to reach the top of atmosphere. Thus we would expect anything below this cloud will not be visible to the top of atmosphere.



(a) Temperature

(b) QVapor

(c) QCloud

(d) QIce

Figure 4.5: Cloudy-sky profile for temperature, water vapor, effective cloud, and effective ice

In all graphs below, in the interest of readability, we will display only up to the stratosphere (i.e. from the surface to 1 hPa).

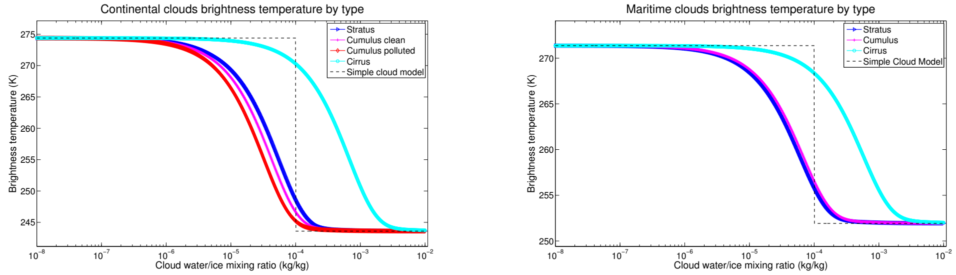
### 4.3.1 Impact of a single-layer cloud on brightness temperature

Before discussing cloudy-sky assimilation, it is worthwhile to examine what happens to the brightness temperatures reaching the top of the atmosphere as a cloud forms. As previously mentioned, RTTOV is parameterized using six different cloud types over land and sea. Figure 4.6 shows the response of increasing cloud mixing ratio from a clear profile to a cloudy profile at a layer approximately 6 km above the ground and 300 m thick for both continental and maritime clouds with a logarithmic scale in $x$. The brightness temperatures are taken for AIRS channel 300 (wavenumber 735.690 or a wavelength of 13.593 $\mu$m).

The results show a sharp drop of about $20 - 25K$ in brightness temperature as the mixing ratio goes from $10^{-6}$ to approximately $10^{-4}$ for stratus and cumulus clouds, while the drop occurs at about $10^{-2}$ for cirrus clouds.

In addition to the full-scattering model, RTTOV version 10 also includes a simple cloud model, where scattering is not considered, but only a cloud-top pressure is used for calculations. This simple model, here with a threshold of $10^{-4}$ kg/kg, is shown in figure 4.6 as the discontinuous jump from clear to overcast. Since even the most modern methods of non-smooth optimization cannot currently handle a jump in the function value, the simple model is considered unsuitable for direct cloudy-sky data assimilation.

Figure 4.7 shows this same information, but now in terms of mixing ratio versus the percentage

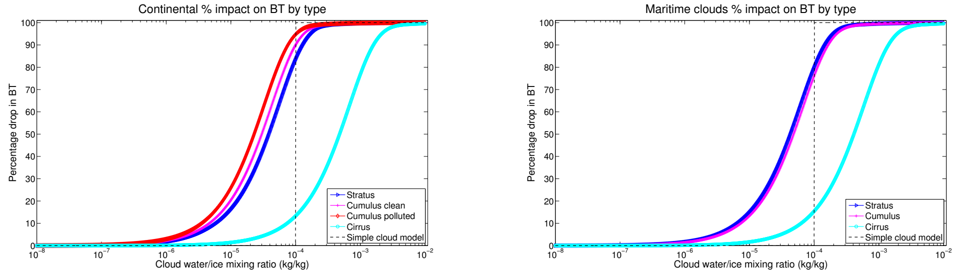


(a) Continental clouds

(b) Maritime clouds

Figure 4.6: Continental and maritime brightness temperature response to increasing clouds on a single 300 m thick layer

drop from the surface brightness temperature to the cloud-top brightness temperature for this case. We will make use of this graph throughout our experiments in judging what impact clouds will have on our profiles. While these results are specific to a particular channel, they give a rough baseline for the impact of clouds on the rest of the AIRS channels.



(a) Continental clouds

(b) Maritime clouds

Figure 4.7: Cloud concentration versus percentage drop in continental and maritime brightness temperature response to increasing clouds on a single 300 m thick layer

## 4.4 Clear-sky assimilation

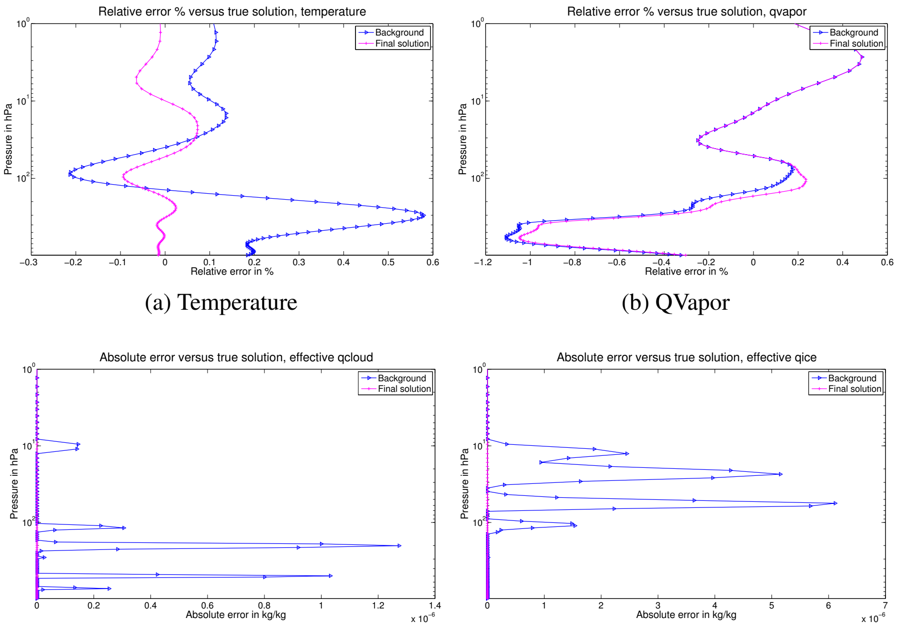
Clear-sky assimilation has taken place operationally for quite some time. This test case is thus for a solved problem; however, we also include the presence of clouds in the background, and see if the 1D-Var algorithm is able to remove the clouds with each of our three main optimization algorithms.

The experiment details are shown in table 4.4, and the reconstructed profiles are shown in figures 4.8 to 4.10 for LMBM, L-BFGS, and CG-Descent.

As we can see from the profiles reconstructed from a cloudy background, all three optimization methods were able to completely remove the background clouds. Thus we can consider this test to be a success for all three methods.

Table 4.4: Clear-sky Experimental setup

| T $\sigma$ % | QVapor $\sigma$ % | QCloud $\sigma$ % | QIce $\sigma$ % | Cloud fraction % | $\sigma_{\text{obs}}$ | $\mu$ |
|---|---|---|---|---|---|---|
| 0.1 | 0.1 | 10 | 10 | 10 | 0.1 | 1 |



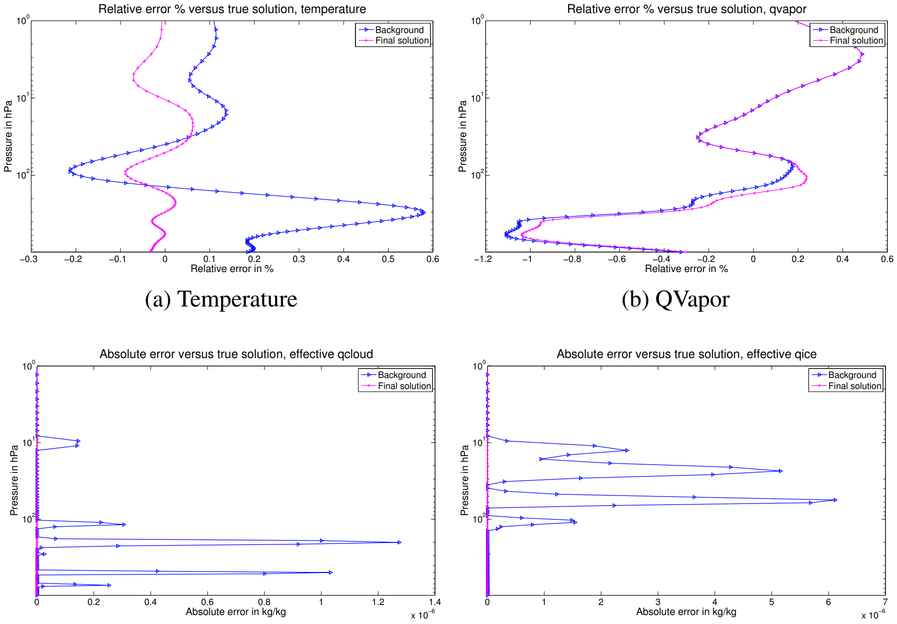
(a) Temperature

(b) QVapor

(c) QCloud

(d) QIce

Figure 4.8: LMBM reconstructed clear-sky profile



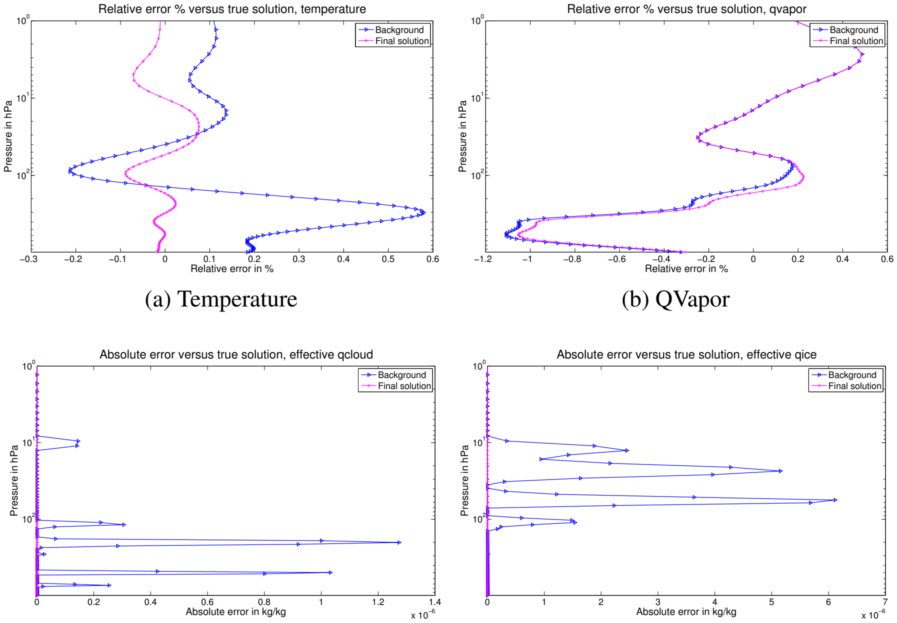
(a) Temperature

(b) QVapor

(c) Effective QCloud

(d) Effective QIce

Figure 4.9: L-BFGS reconstructed clear-sky profile



(a) Temperature

(b) QVapor

(c) QCloud

(d) QIce

Figure 4.10: CG-Descent reconstructed clear-sky profile

As shown in figure 4.11, all three methods converge to approximately the same value of the cost function. L-BFGS and LMBM reach this value the most quickly with approximately 20 function and gradient evaluations, while CG-Descent is the slowest with approximately 50 function and gradient evaluations needed. However, here all three methods can be considered competitive and efficient, as all three successfully remove the background clouds to converge to a cloud-free state.

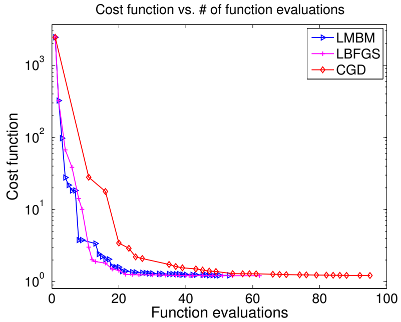


Figure 4.11: Optimization history for clear-sky profile

## 4.5 Cloudy-sky assimilation

The full cloudy-sky assimilation problem is that of trying to reconstruct an atmospheric profile that contains one or more clouds. As mentioned above, in RTTOV v10 the cloud concentration and cloud fraction at each layer can be specified for a maximum of two types of clouds per layer. In addition, in the current version of RTTOV (v10.1), only half of the layers can contain clouds ([89]). Finally, due to current bugs in RTTOV v10.1, currently neither fully overcast layers (cloud fraction = 100%) nor adjacent layers with identical cloud fractions are allowed. The workaround for the first issue is to have a cloud fraction near 100% (such as 99.999999%), while the workaround for the second is to offset identical layers by a small amount. These issues make coding the cost function and an accurate gradient more difficult to handle, and thus careful attention was paid to the so-called "alpha-test" of gradient consistency first derived in [160]. This is also an effective test for the accuracy of the RTTOV model adjoint.

### 4.5.1 Alpha test of gradient correctness

Starting from a Taylor series expansion of $J$ around $x_0$ of $x = x_0 + \alpha h^{\mathrm{T}}$, where $\alpha$ is a small number and $h$ is a unit vector (such as $\nabla J(x_0)/||\nabla J(x_0)||$), we have:

$$J(x) = J(x_0) + (x - x_0)^{\mathrm{T}} \nabla J(x_0) + O(\alpha^2) \tag{4.21}$$

Rearranging,

$$J(x) - J(x_0) = \alpha h^{\mathrm{T}} \nabla J(x_0) + O(\alpha^2) \tag{4.22}$$

or

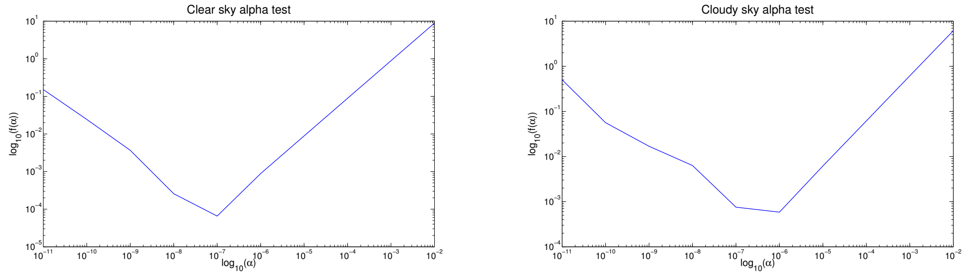
$$\frac{J(x) - J(x_0)}{\alpha h^{\mathrm{T}} \nabla J(x_0)} = 1 + O(\alpha) \tag{4.23}$$

If we define

$$\phi(\alpha) = 1 - \frac{J(x) - J(x_0)}{\alpha h^{\mathrm{T}} \nabla J(x_0)} \tag{4.24}$$

$\phi(\alpha)$ should be $O(\alpha)$ up to numerical precision.

The results for both clear and cloudy-skies in figure 4.12. We see in both cases the classic “V” shape, and although the cloudy-sky alpha test has some small bumps below $10^{-6}$, these can be attributed to loss of numerical precision. These results show that once the RTTOV issues mentioned above have been taken into account, the gradient of the cost-function is indeed consistent.



(a) Clear sky

(b) Cloudy sky

Figure 4.12: Alpha test of gradient consistency for clear and cloudy sky

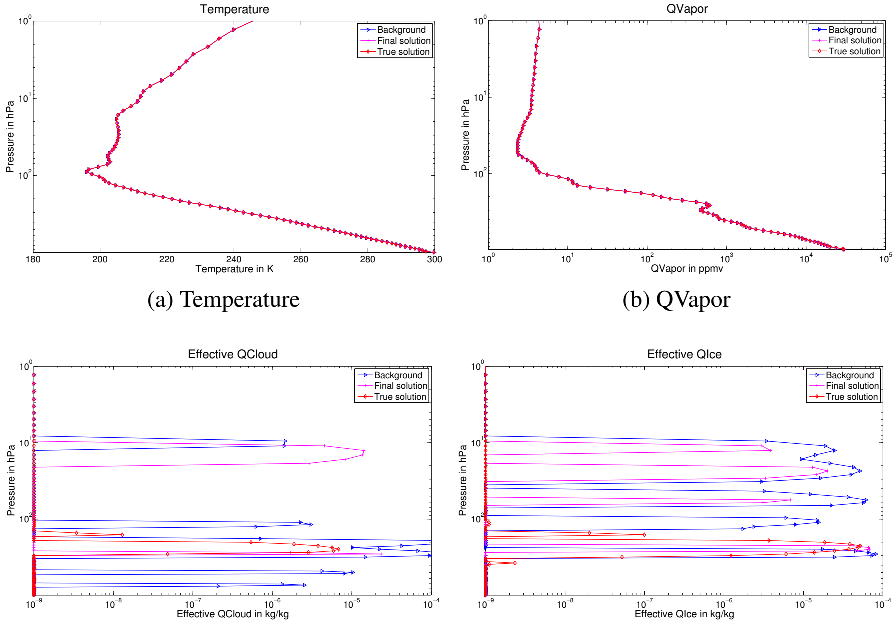
### 4.5.2 First test case

The first test case for cloudy-sky assimilation involves small QVapor and temperature perturbation but a large cloudy perturbation. Since QVapor and temperature both have such small perturbations, the background error covariance model will constrain the solution to stay near these profiles. The cost function will thus place more emphasis on reducing the clouds. The details of this experiment are shown in table 4.5, and the results of the assimilation for this case are shown in figures 4.13 to 4.18.

Table 4.5: Cloudy-Sky Test Case 1 Experimental setup

| T $\sigma$ % | QVapor $\sigma$ % | QCloud $\sigma$ % | QIce $\sigma$ % | Cloud fraction % | $\sigma_{\text{obs}}$ | $\mu$ |
|---|---|---|---|---|---|---|
| 0.01 | 0.01 | 100 | 100 | 100 | 10 | 1 |

We can see that the cloud background, especially for QIce, was particularly misleading, starting with cirrus clouds scattered above the main cloud. This situation proved difficult for the three test cases; LMBM was able to reduce most of the clouds slightly, but left spurious ice clouds on the order of 1e-5 kg/kg in the stratosphere. LMBM was also able to slightly improve upon the background temperature error, but left the water vapor background untouched. The other methods did not achieve as much reduction for these cirrus clouds and virtually no reduction for the temperature. Most importantly, the RTTOV adjoint from all three methods guided the background water vapor cloud from approximately 200 hPa up to 10 hPa, ostensibly to be near the large spurious cirrus cloud. In short, this test was a failure.



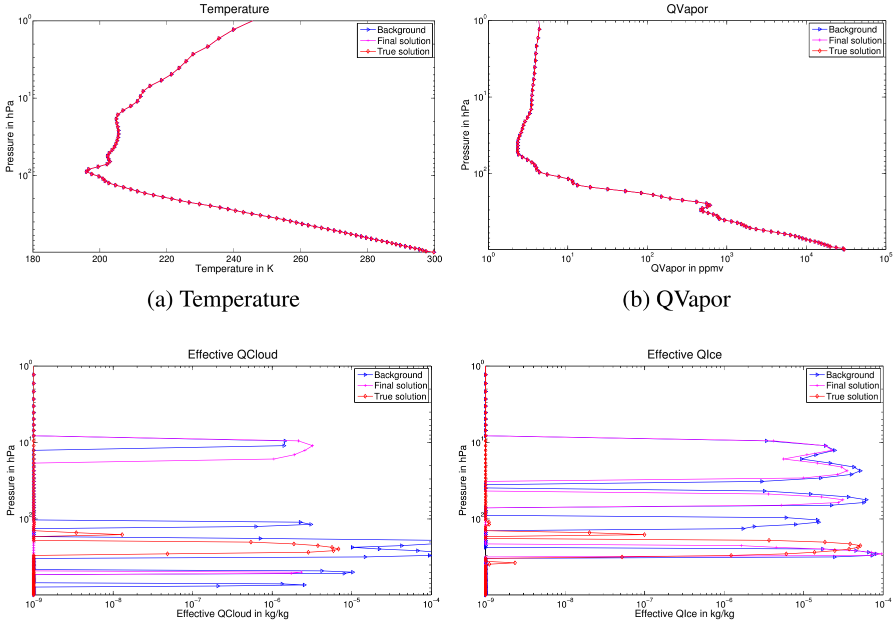
(a) Temperature

(b) QVapor

(c) QCloud

(d) QIce

Figure 4.13: LMBM reconstructed cloudy-sky profile for test case 1



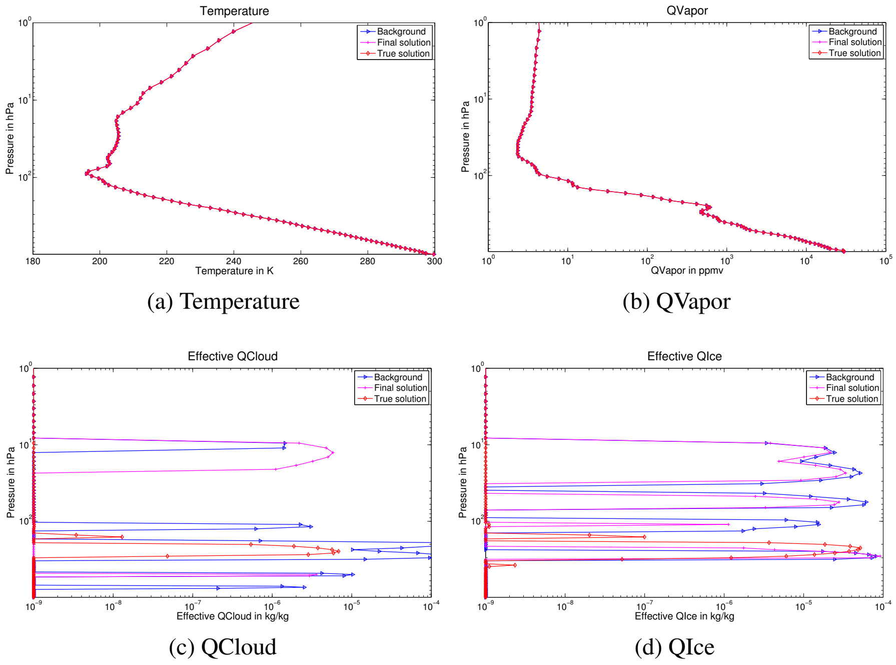
(a) Temperature

(b) QVapor

(c) Effective QCloud

(d) Effective QIce

Figure 4.14: L-BFGS reconstructed cloudy-sky profile for test case 1



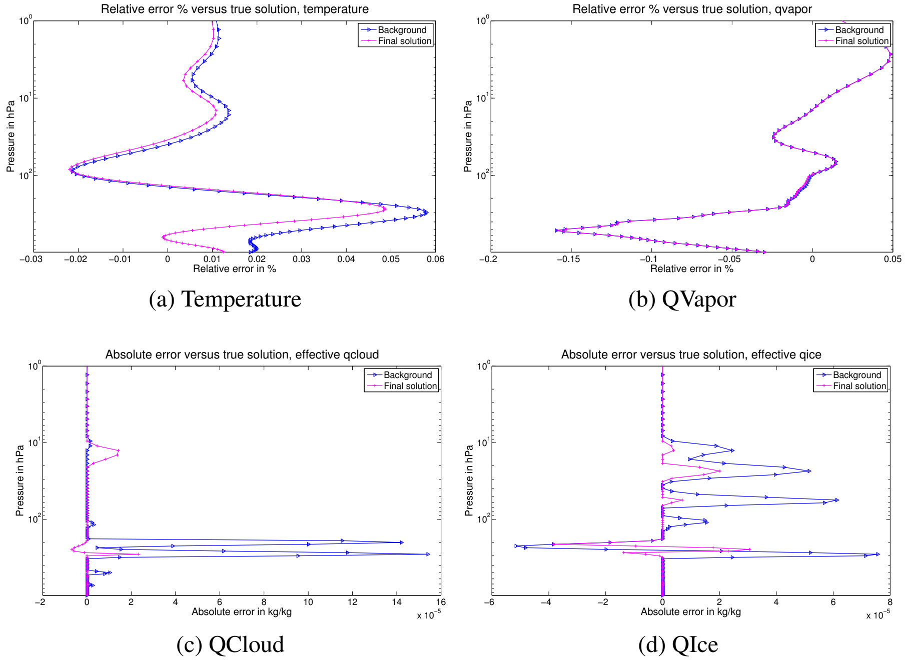
(a) Temperature

(b) QVapor

(c) QCloud

(d) QIce

Figure 4.15: CG-Descent reconstructed cloudy-sky profile for test case 1



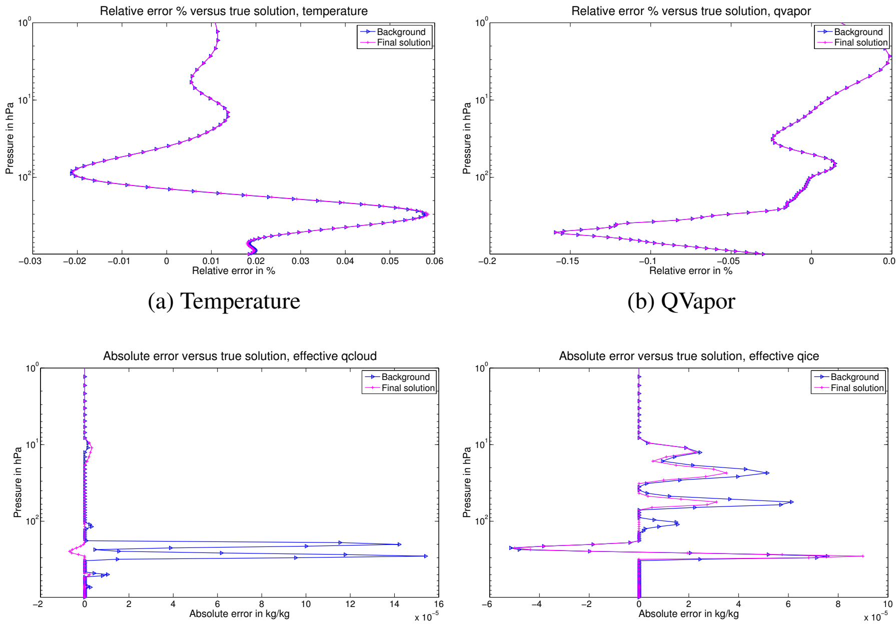
(a) Temperature

(b) QVapor

(c) QCloud

(d) QIce

Figure 4.16: LMBM reconstructed cloudy-sky profile error for test case 1



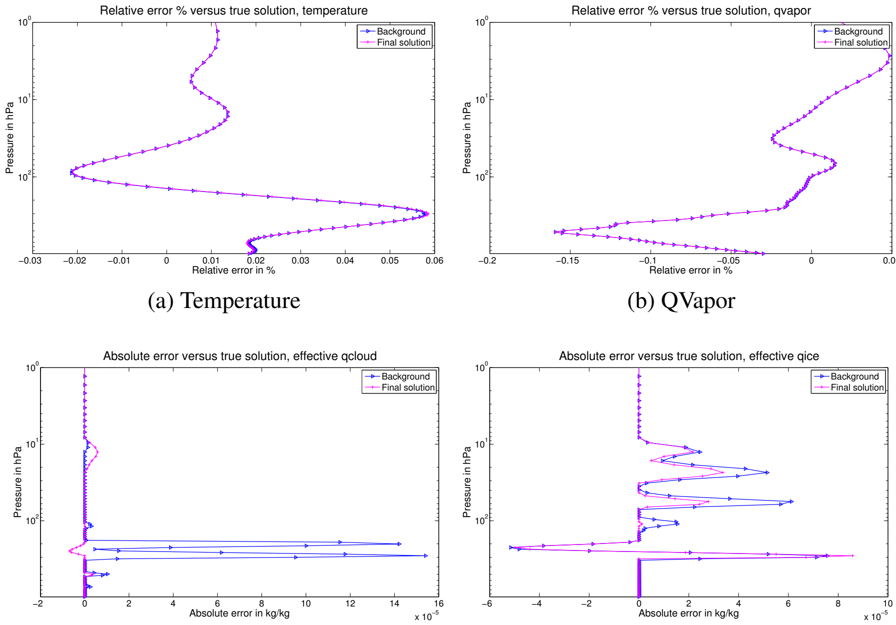
(a) Temperature

(b) QVapor

(c) Effective QCloud

(d) Effective QIce

Figure 4.17: L-BFGS reconstructed cloudy-sky profile error for test case 1



(a) Temperature

(b) QVapor

(c) QCloud

(d) QIce

Figure 4.18: CG-Descent reconstructed cloudy-sky profile error for test case 1

The history of the minimization in terms of function evaluations is shown in figure 4.19, and it becomes readily apparent why LMBM was able to achieve slightly better results. While the other minimization algorithms terminated with an error in the line-search, LMBM was able to continue to minimize the function for more than one additional order of magnitude before terminating.

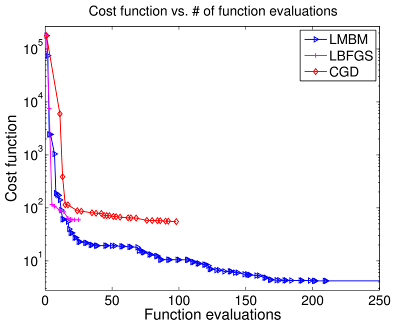


Figure 4.19: Optimization history for cloudy-sky test case 1

From this experiment, we see that with very large perturbations in background clouds, the optimization methods can completely fail to converge to the true state. Starting with large spurious ice clouds in the stratosphere, the methods converged to a false solution of a single stratus-like cloud. Since this provided good agreement with the observations (the cost function for LMBM converged to a solution on the order of 2e0), it is clear that this solution was able to explain the observations to some degree. As the cloudy-sky assimilation problem is known to be ill-posed (see e.g. [182], where water vapor and temperature were adjusted by the minimization algorithm rather than the desired cloud), it is clear that a large background error, even with a corresponding large covariance, can cause the final solution to converge to an erroneous cloudy state.

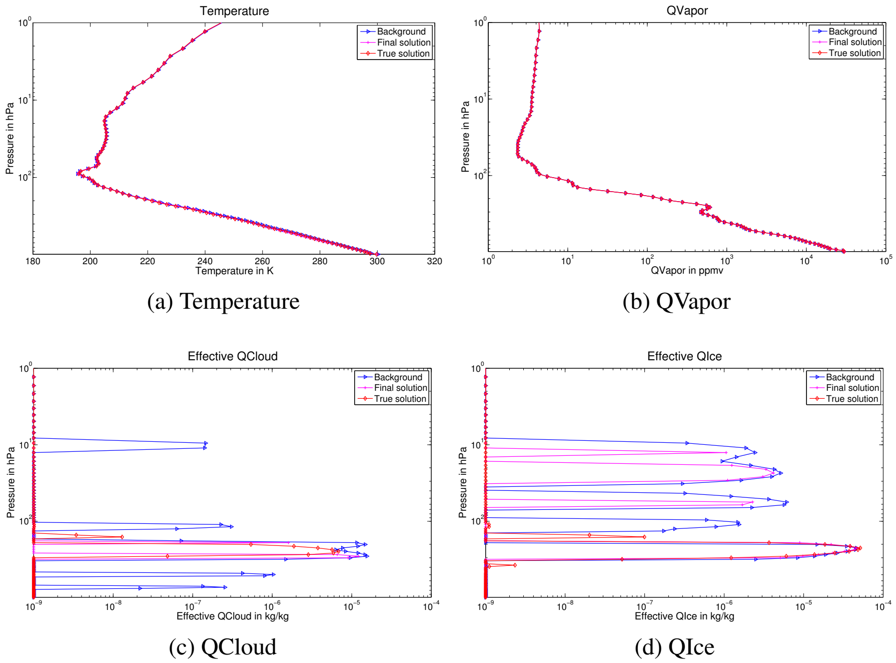
### 4.5.3 Second test case

In this test case, we test a medium standard deviation in both temperature and QVapor as well as water and ice cloud. The gives the cost function more weight on the temperature and water vapor, although it is still not as large in percentage as the cloud properties. The experiment details are given in table 4.6 and the error results are shown in figures 4.20 to 4.25.

Table 4.6: Cloudy-Sky Test Case 2 Experimental setup

| T $\sigma$ % | QVapor $\sigma$ % | QCloud $\sigma$ % | QIce $\sigma$ % | Cloud fraction % | $\sigma_{\text{obs}}$ | $\mu$ |
|---|---|---|---|---|---|---|
| 0.1 | 0.1 | 10 | 10 | 10 | 10 | 1 |

In this test case, as in the previous one, spurious cirrus clouds are present. However, since the perturbation is only on the order of 10%, rather than 100%, these spurious clouds are much smaller than the true cloud. All three methods are able to reduce the minor background clouds to a level around 1e-6, which has little impact on the final solution, and also keep the cloud-top pressure near



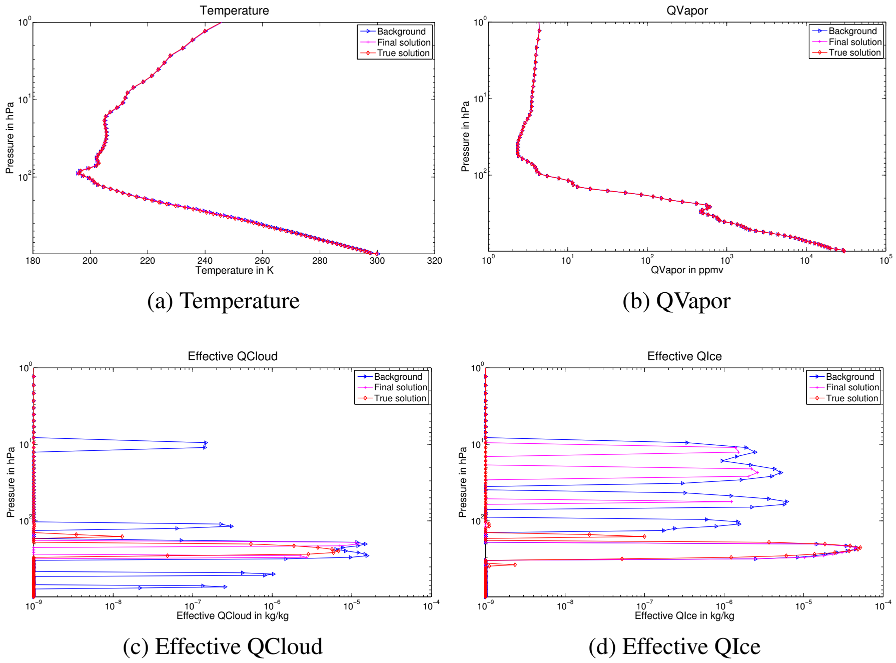
(a) Temperature

(b) QVapor

(c) QCloud

(d) QIce

Figure 4.20: LMBM reconstructed cloudy-sky profile for test case 2



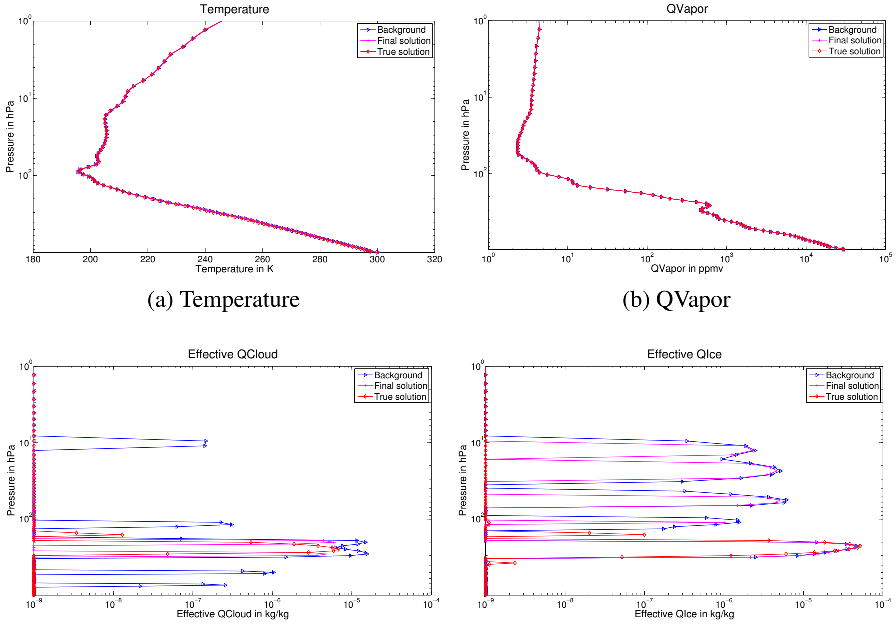
(a) Temperature

(b) QVapor

(c) Effective QCloud

(d) Effective QIce

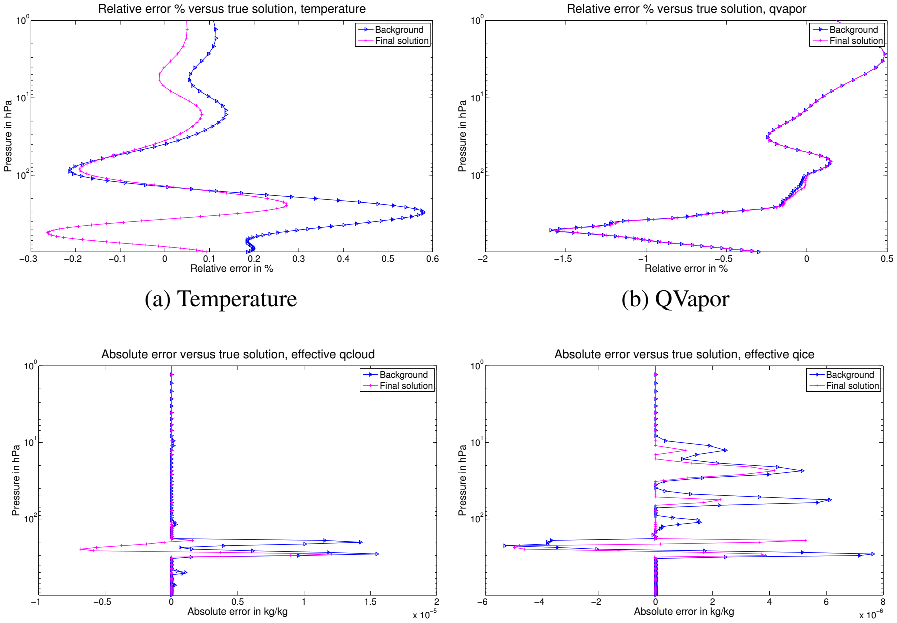
Figure 4.21: L-BFGS reconstructed cloudy-sky profile error for test case 2



(a) Temperature (b) QVapor

(c) QCloud (d) QIce

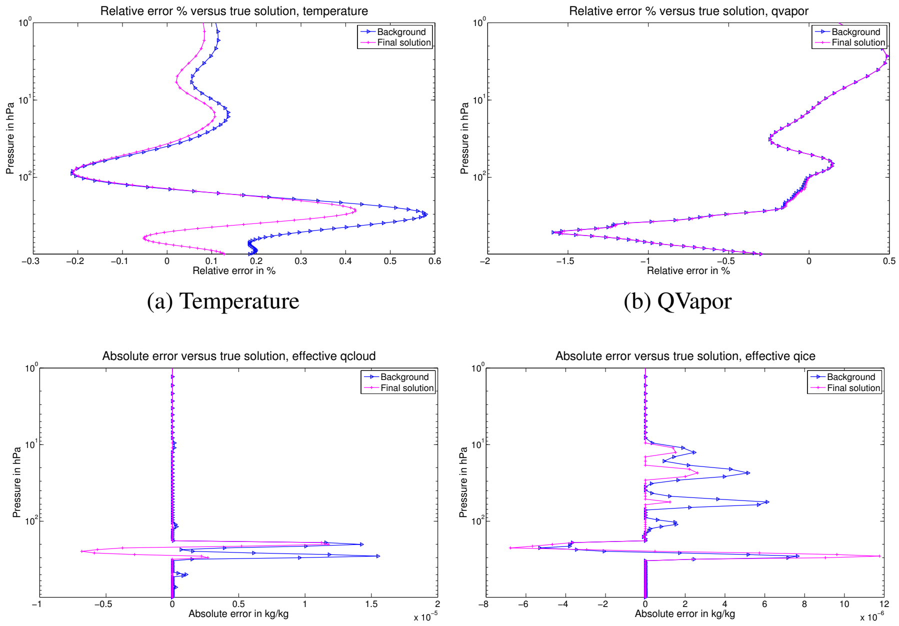
Figure 4.22: CG-Descent reconstructed cloudy-sky profile error for test case 2



(a) Temperature (b) QVapor

(c) QCloud (d) QIce

Figure 4.23: LMBM reconstructed cloudy-sky profile error for test case 2



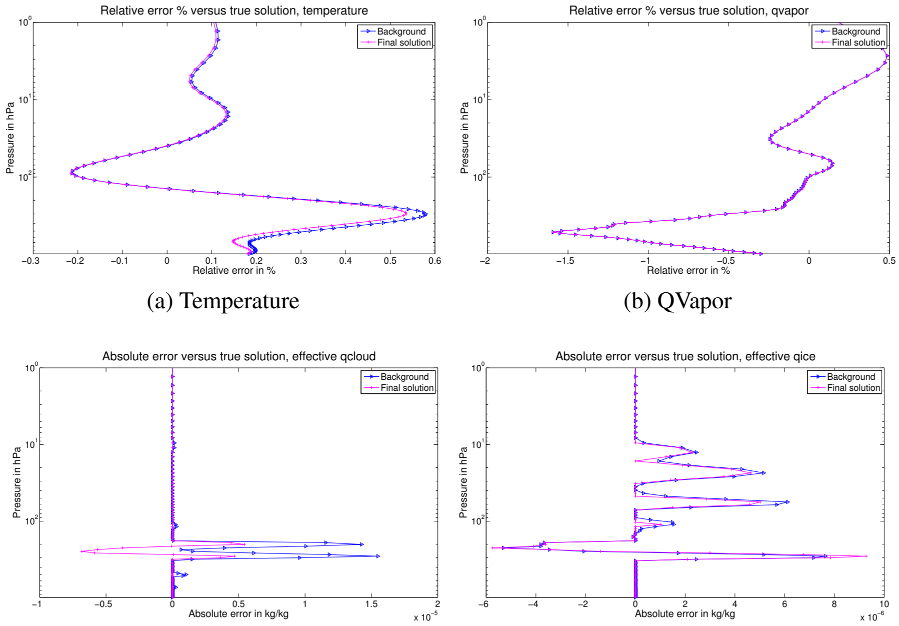
(a) Temperature

(b) QVapor

(c) Effective QCloud

(d) Effective QIce

Figure 4.24: L-BFGS reconstructed cloudy-sky profile error for test case 2



(a) Temperature

(b) QVapor

(c) QCloud

(d) QIce

Figure 4.25: CG-Descent reconstructed cloudy-sky profile error for test case 2

the true cloud. Both methods are able to reduce the temperature somewhat, although the water vapor background is again relatively untouched. This experiment can be considered a success.

Again, the history of the minimization in terms of function evaluations is shown in figure 4.26. Here both L-BFGS and LMBM are able to achieve a similar solution that is about an order of magnitude lower in cost, and unlike the previous experiment, all three methods terminate in less than 100 function evaluations.

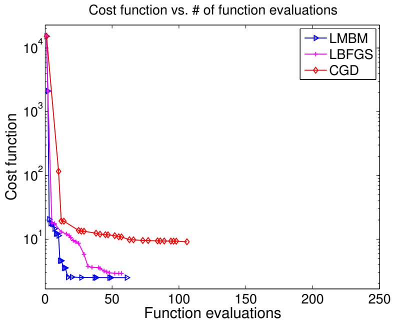


Figure 4.26: Optimization history for cloudy-sky test case 2

We can conclude from these results that when the cloud perturbation is not too large, cloudy assimilation assimilation is much more successful. All three methods did quite well on this test case, possibly because the true cloud was the largest cloud present in the background.

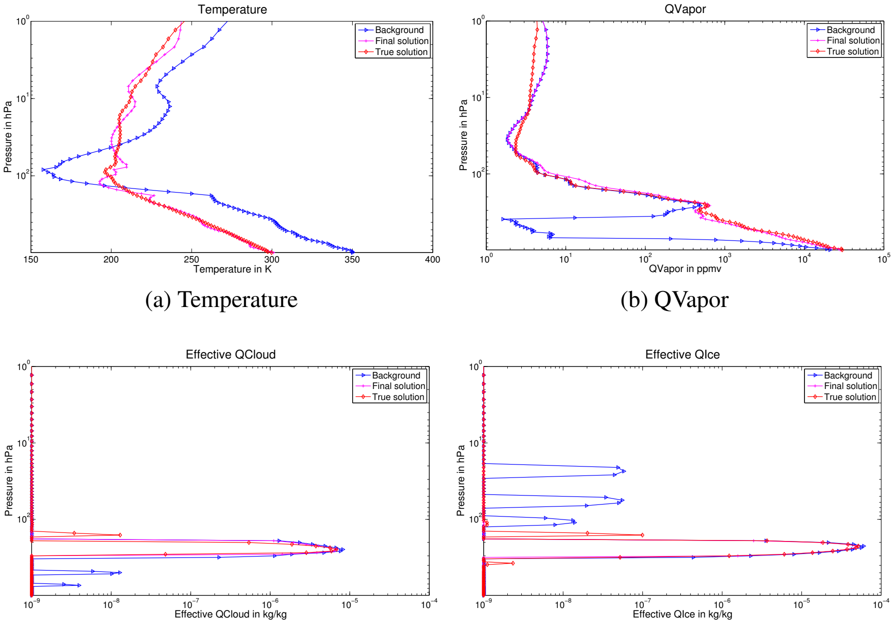
### 4.5.4 Third test case

In this final case for cloudy-sky assimilation, we test a very large standard deviation in temperature and QVapor and a small standard deviation in ice and cloud. The experiment details are given in table 4.7 and the results are shown in figures 4.27 to 4.32.

Table 4.7: Cloudy-Sky Test Case 3 Experimental setup

| T $\sigma$ % | QVapor $\sigma$ % | QCloud $\sigma$ % | QIce $\sigma$ % | Cloud fraction % | $\sigma_{\text{obs}}$ | $\mu$ |
|---|---|---|---|---|---|---|
| 10 | 10 | 1 | 1 | 1 | 1 | 1 |

These figures show that all three methods were able to achieve a very large reduction in error with respect to temperature, while the error for QVapor exhibited strange behavior. The large erroneous dry lower atmosphere in the background was largely fixed; however, an approximately 80% error in QVapor occurred above the cloud-top. This also corresponded with a large shift from too large of a cirrus cloud in the background to an underpredicted cirrus cloud. We can conclude from this that because the covariance for the water vapor was so large, the 1D-Var algorithm found it easier to change water vapor than the cirrus cloud. These results were also seen in [182], where by water vapor was changed rather than adding a cloud. We can conclude that water vapor and clouds are somewhat in competition since they absorb infrared radiation in similar ways, and an overly



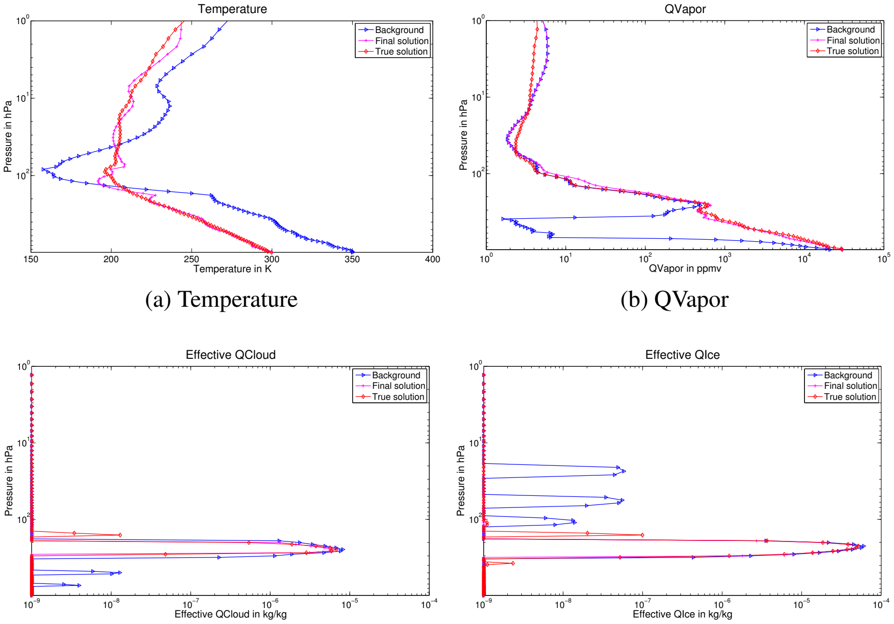
(a) Temperature

(b) QVapor

(c) QCloud

(d) QIce

Figure 4.27: LMBM reconstructed cloudy-sky profile for test case 3



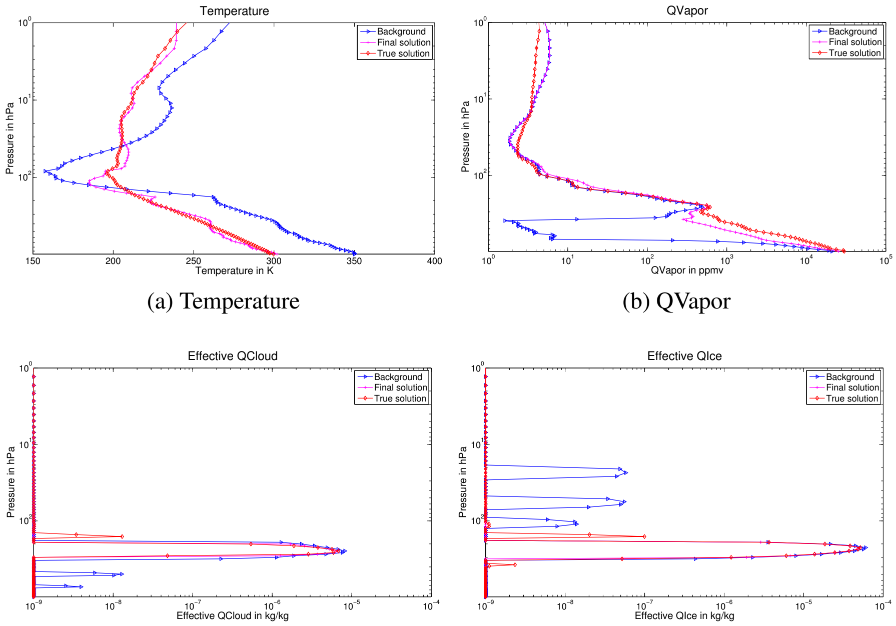
(a) Temperature

(b) QVapor

(c) Effective QCloud

(d) Effective QIce

Figure 4.28: L-BFGS reconstructed cloudy-sky profile for test case 3



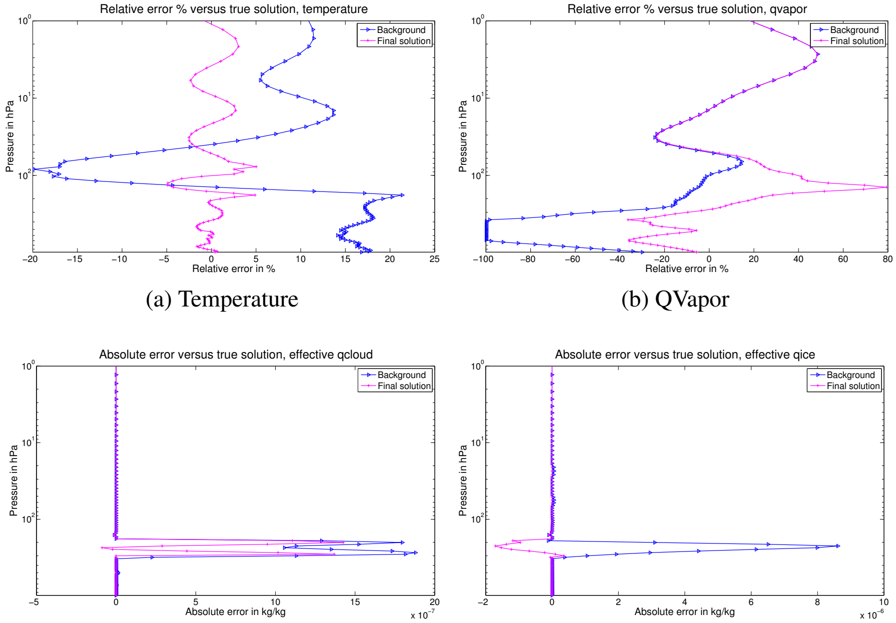
(a) Temperature

(b) QVapor

(c) QCloud

(d) QIce

Figure 4.29: CG-Descent reconstructed cloudy-sky profile for test case 3



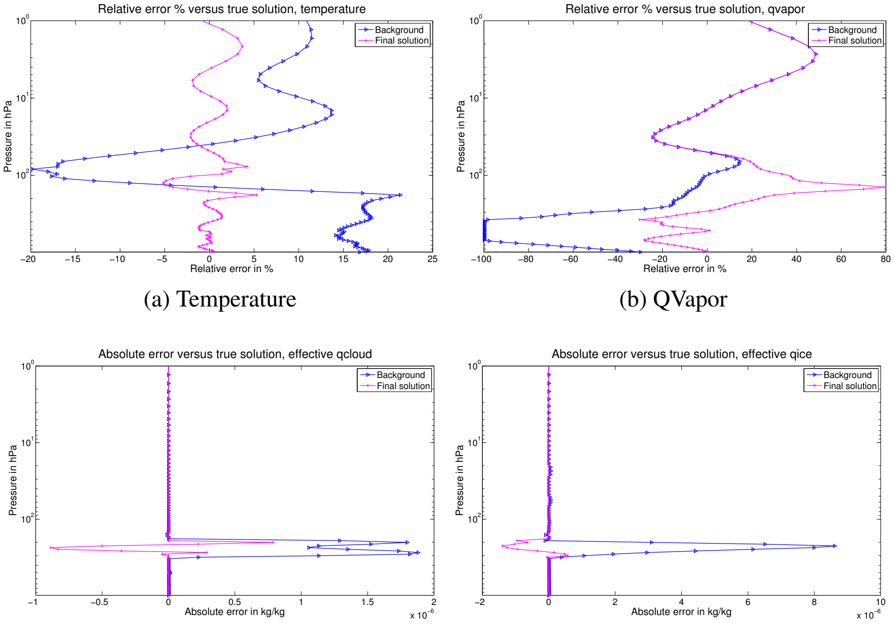
(a) Temperature

(b) QVapor

(c) QCloud

(d) QIce

Figure 4.30: LMBM reconstructed cloudy-sky profile error for test case 3



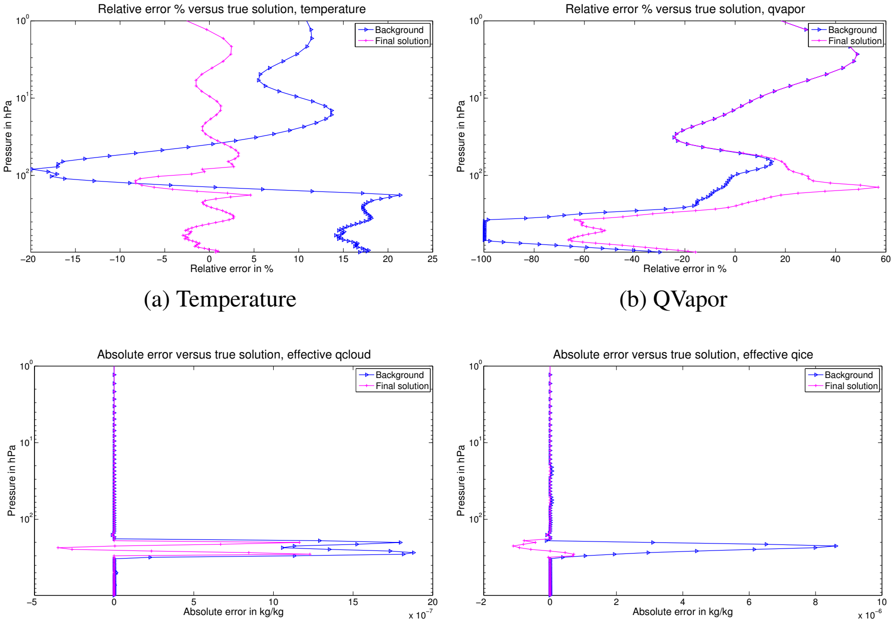
(a) Temperature

(b) QVapor

(c) Effective QCloud

(d) Effective QIce

Figure 4.31: L-BFGS reconstructed cloudy-sky profile error for test case 3



(a) Temperature

(b) QVapor

(c) QCloud

(d) QIce

Figure 4.32: CG-Descent reconstructed cloudy-sky profile error for test case 3

large background for water vapor can lead to a solution of an overly moist atmosphere rather than clouds.

These figures also demonstrates why cloudy assimilation is so difficult when compared to clear sky assimilation – the transition from clear to cloudy depends on a variable that scales from 0 to $10^{-2}$ with sharp transitions, while temperature and QVapor scale much more smoothly over a much broader range.

Again, the history of the minimization in terms of function evaluations is shown in figure 4.26. LMBM and L-BFGS are once again competitive, while CGD achieves an order of magnitude higher cost.

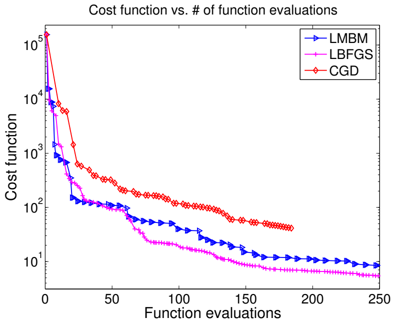


Figure 4.33: Optimization history for cloudy-sky test case 3

## 4.6 All-sky assimilation

All-sky data assimilation is the more realistic case where it is not known a priori whether the profile under consideration is clear or cloudy. While a background state may be available, in practice the particular locations and types of clouds predicted by a numerical weather prediction model are not considered to be particularly reliable at this time. The reasons for this include poorly understood and modeled cloud dynamics – consider that there are at least ten different microphysical parameterizations schemes included in WRF, and all of them give different representations of clouds ([173]). Clearly not all of these representations can be correct, and which scheme(s) to use in what situations is still an open question.

Furthermore, as we have seen above, it can be very difficult to clear an unwanted cloud that is in the background through the data assimilation. Indeed, the effect of including a background term is to constrain the solution so that the cost function is increased by moving away from it, and only when the data fits the observations better will the solution move away from the background. While this serves to regularize the problem, it can also be a hindrance; for a poor choice of background, the background term is like a chain that prevents the solution from moving in the right direction.

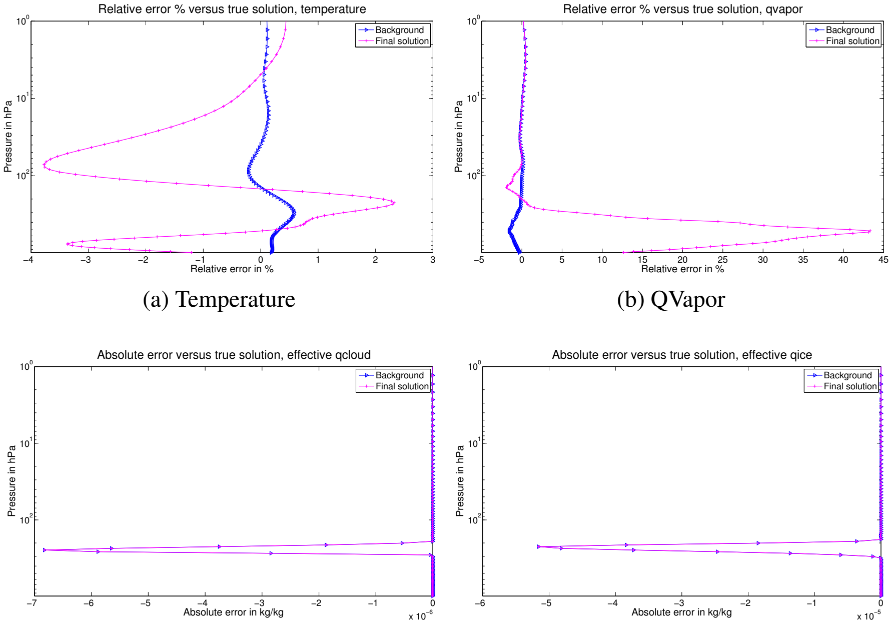
For these reasons, it is clear that an alternative strategy without the need for a well-specified cloud background is worth pursuing.

### 4.6.1 Starting from a clear sky

As we saw in our test of a cloudy background with clear observations, our 1D-Var data assimilation with RTTOV can clear clouds out quite effectively. It is natural to ask if the opposite might be true – by starting from a clear sky, can RTTOV push the gradient of the cost function in the direction of clouds?

Table 4.8: Cloudy-sky starting clear experimental setup

| T $\sigma$ % | QVapor $\sigma$ % | QCloud $\sigma$ % | QIce $\sigma$ % | Cloud fraction % | $\sigma_{\text{obs}}$ | $\mu$ |
|---|---|---|---|---|---|---|
| .1 | .1 | 10 | 10 | 10 | 10 | 1 |



(a) Temperature

(b) QVapor

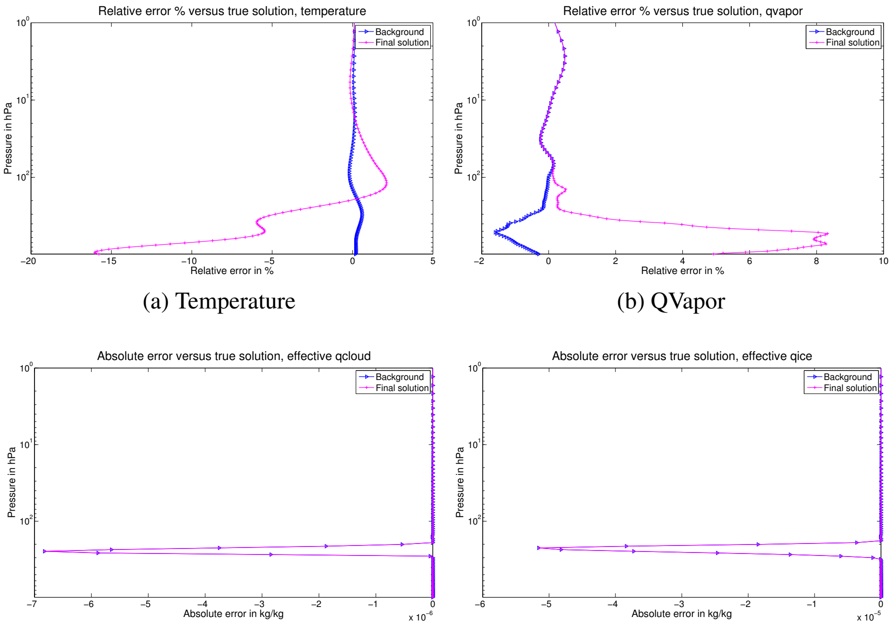
(c) QCloud

(d) QIce

Figure 4.34: LMBM reconstructed cloudy-sky profile error from starting clear

As we can see, no cloud was reconstructed by this test case – instead the data assimilation for all three algorithms added additional water vapor and reduced the temperature near the location of the cloud, even though these values were constrained by a very small covariance. In short, this test is a complete failure.

Therefore, as evidenced by this test, unfortunately the answer to the question above is "no" – one cannot go from a completely clear-sky (with QCloud and QIce set to zero) and hope to reconstruct clouds with RTTOV. Because RTTOV does not use scattering for clear skies, when the profile is perfectly clear, the RTTOV adjoint does not hit the code for calculating cloudy gradients, and thus the gradient remains zero. This may explain, at least in part, the results seen in [182].



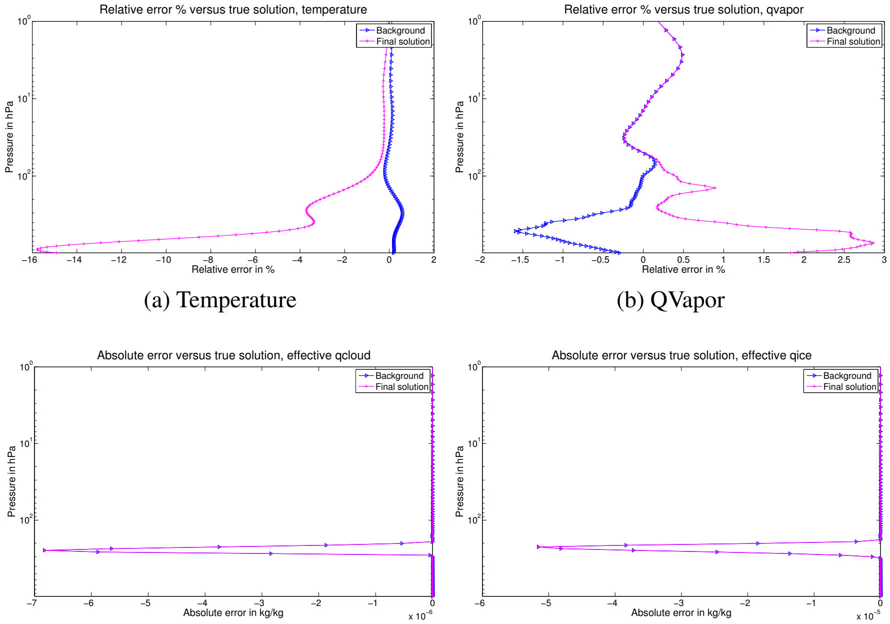
(a) Temperature

(b) QVapor

(c) Effective QCloud

(d) Effective QIce

Figure 4.35: L-BFGS reconstructed cloudy-sky profile error from starting clear



(a) Temperature

(b) QVapor

(c) QCloud

(d) QIce

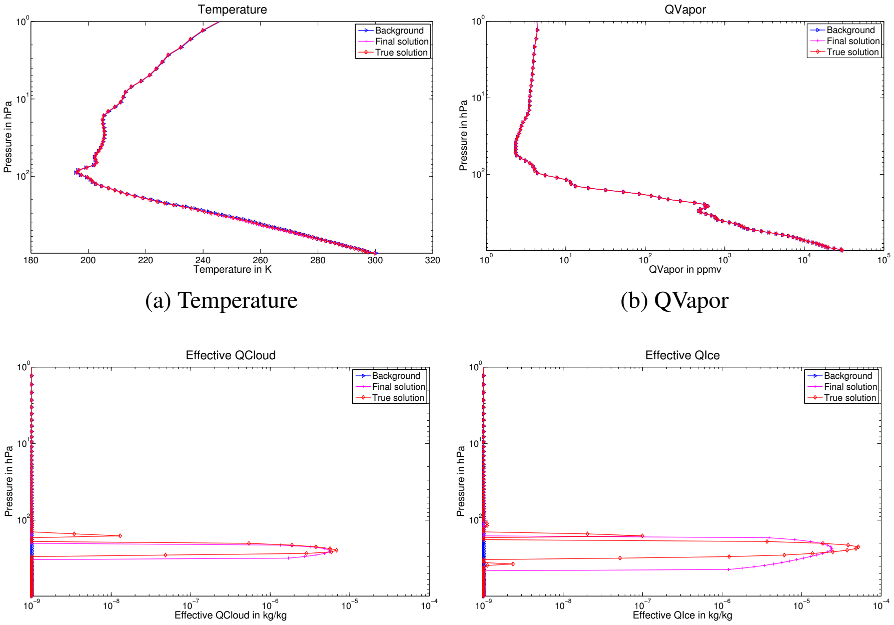
Figure 4.36: CG-Descent reconstructed cloudy-sky profile error from starting clear

### 4.6.2 Alternate approach

As we have seen above, when clouds are not present in the observations, starting from a cloudy sky is not a hindrance; when spurious clouds are present in the background, it can be difficult to remove them; and finally that one cannot start from a clear profile and hope to reconstruct a cloudy profile with RTTOV. One potential alternative is to start with a very small cloud at all layers, so that the adjoint code for scattering is invoked, but at the same time the effective clouds will be small enough that they will not influence the actual brightness temperature much. However, since RTTOV only allows the specification of clouds on half of the layers, and AIRS channels are designed to peak in sensitivity in the lower atmosphere, we might eliminate clouds (i.e. start clear) in both the mesosphere and the upper stratosphere, at the possible risk of sacrificing rare phenomenon such as mesospheric noctilucent clouds ([67]). The details of this experiment are shown in figure 4.9, and the results are shown in figures 4.37 to 4.41.

Table 4.9: Cloudy-sky starting with half minimum clouds experimental setup

| T $\sigma$ % | QVapor $\sigma$ % | QCloud $\sigma$ % | QIce $\sigma$ % | Cloud fraction % | $\sigma_{\text{obs}}$ | $\mu$ |
|---|---|---|---|---|---|---|
| .1 | .1 | 10 | 10 | 10 | 10 | 1 |



(a) Temperature

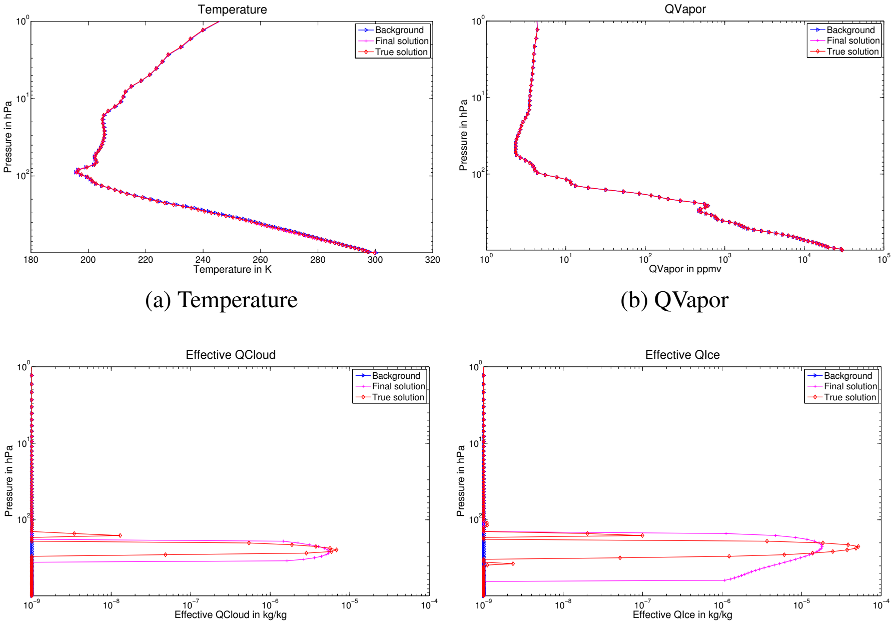
(b) QVapor

(c) QCloud

(d) QIce

Figure 4.37: LMBM reconstructed cloudy-sky profile starting with half minimum clouds

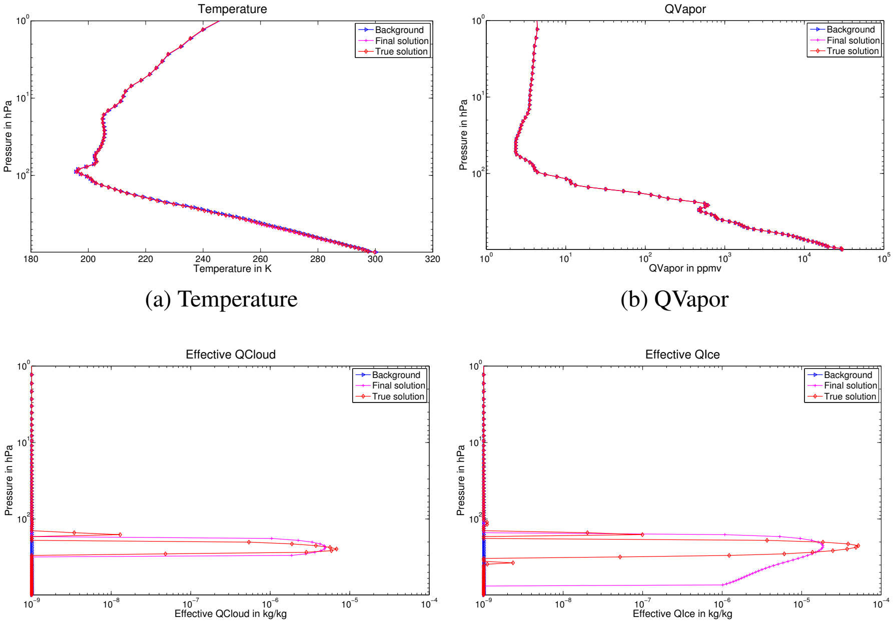
As these results show, all three algorithms were able to reconstruct some measure of the main cloud starting from a profile with half of the clouds at $10^{-12}$ kg/kg, although the cloud in all three



(a) Temperature (b) QVapor

(c) Effective QCloud (d) Effective QIce

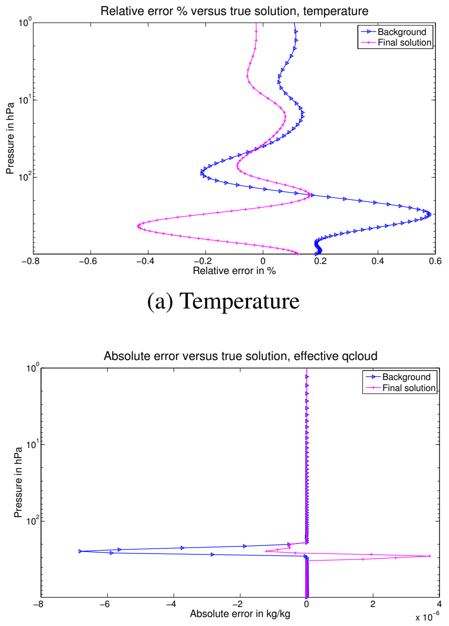
Figure 4.38: L-BFGS reconstructed cloudy-sky profile starting with half minimum clouds



(a) Temperature (b) QVapor

(c) QCloud (d) QIce

Figure 4.39: CG-Descent reconstructed cloudy-sky profile starting with half minimum clouds



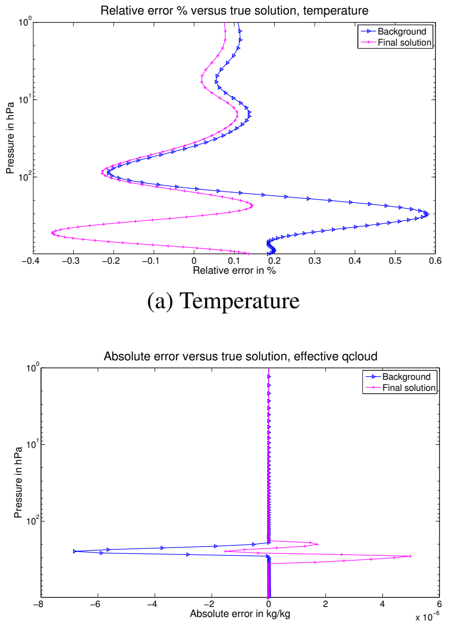
(a) Temperature

(b) QVapor

(c) QCloud

(d) QIce

Figure 4.40: LMBM reconstructed cloudy-sky profile error starting with half minimum clouds



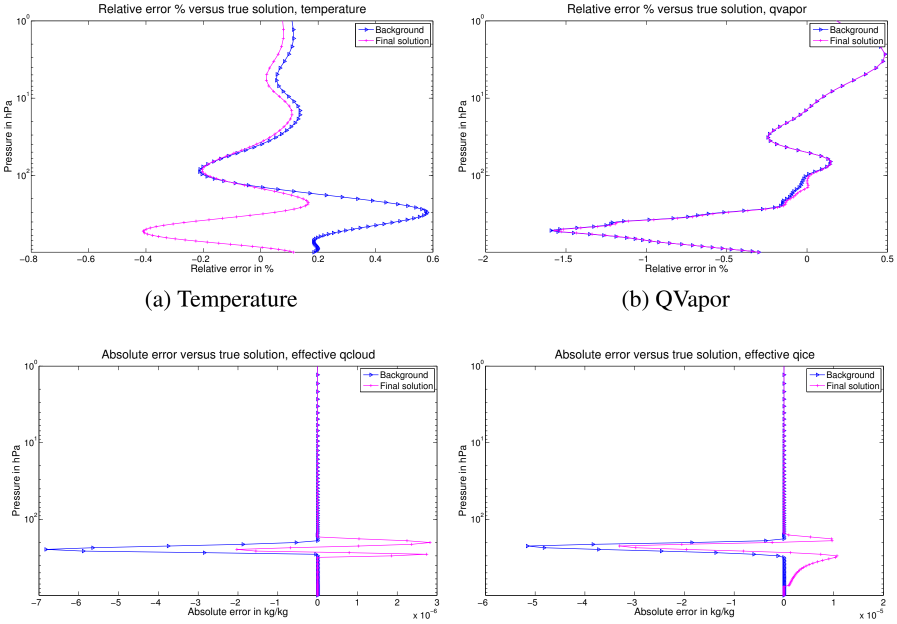
(a) Temperature

(b) QVapor

(c) Effective QCloud

(d) Effective QIce

Figure 4.41: L-BFGS reconstructed cloudy-sky profile error starting with half minimum clouds



(a) Temperature

(b) QVapor

(c) QCloud

(d) QIce

Figure 4.42: CG-Descent reconstructed cloudy-sky profile error starting with half minimum clouds

cases, especially L-BFGS and CGD, were far too thick without enough intensity. This is another case of ill-posedness, whereby a thin, strong cloud explains the observations as well as a thick weak cloud. Nonetheless, considering the fact that the background contained no relevant cloud information, the simulated AIRS observations had errors, and heuristic background error covariance lengths were used, these results are impressive.

As is clear from the overall profile, the cloud tops from all three methods more or less perfectly agrees with the location of the main cloud, and our algorithm was even able to reconstruct the mixed-phase nature of the cirrus cloud to a remarkable extent. While the profile below the main cloud is wrong, the IR observations would have no way of knowing this as the cloud is so optically thick that nothing below is visible to the top of atmosphere. Thus we can consider that this method is a complete success for all three methods.

The minimization history for this test case is shown in figure 4.43. As with most of the tests so far, both LMBM and L-BFGS are competitive, with CG-Descent being slightly less effective.

### 4.6.3 Error measurements

In order to complement the somewhat subjective appraisals of each test case given above, this section gives more quantitative descriptions of the error in the final solution.

Table 4.10 shows the reduction in root-mean-squared-error (RMSE) versus the background achieved for each case. Negative values mean that the RMSE increased.

The RMSE of a variable $*$ is defined by

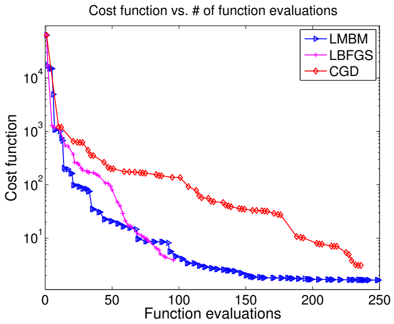


Figure 4.43: Optimization history for cloudy-sky profile starting with half minimum clouds

$$\mathrm{RMSE}_{\ast}^{(k)} = \sqrt{\frac{\left(\ast_{\text{exact}}^{(k)} - \ast^{(k)}\right)^{\mathrm{T}} \left(\ast_{\text{exact}}^{(k)} - \ast^{(k)}\right)}{N}} \tag{4.25}$$

where $N$ is the size of the variable $\ast$.

Table 4.10: RMSE reduction, in percentage from background, of each test case

| **Type** | **Case** | **Method** | **T %** | QVapor % | QCloud % | QIce % |
|---|---|---|---|---|---|---|
| clear | start cloudy | LMBM | 360.01 | 5.79 | Inf | Inf |
| clear | start cloudy | LBFGS | 355.99 | 7.05 | Inf | Inf |
| clear | start cloudy | CGD | 354.00 | 5.44 | Inf | Inf |
| cloudy | test case 1 | LMBM | 36.86 | 0.46 | 806.45 | 185.98 |
| cloudy | test case 1 | LBFGS | -0.43 | 0.02 | 2253.37 | 15.85 |
| cloudy | test case 1 | CGD | -0.49 | 0.02 | 1812.66 | 19.97 |
| cloudy | test case 2 | LMBM | 79.95 | 1.36 | 61.04 | 45.30 |
| cloudy | test case 2 | LBFGS | 58.28 | 0.69 | 48.55 | -9.27 |
| cloudy | test case 2 | CGD | 10.34 | 0.07 | 114.13 | -1.32 |
| cloudy | test case 3 | LMBM | 477.97 | 235.74 | 75.29 | 373.91 |
| cloudy | test case 3 | LBFGS | 504.91 | 429.85 | 194.14 | 490.14 |
| cloudy | test case 3 | CGD | 356.86 | 63.31 | 129.87 | 618.42 |
| cloudy | start clear | LMBM | -85.60 | -97.17 | 0.00 | 0.00 |
| cloudy | start clear | LBFGS | -96.10 | -91.08 | 0.00 | 0.00 |
| cloudy | start clear | CGD | -96.16 | -75.17 | 0.00 | 0.00 |
| cloudy | st. half min | LMBM | 43.11 | 1.77 | 112.57 | 85.61 |
| cloudy | st. half min | LBFGS | 58.40 | 0.14 | 29.33 | 53.58 |
| cloudy | st. half min | CGD | 41.83 | 0.42 | 88.04 | 56.23 |

These results show that in all cases except test case 3 and cloudy start clear, LMBM was able to attain the best reduction in RMSE. When it converged, L-BFGS also achieved impressive reductions in RMSE, especially in test case 3; however, L-BFGS did not converge on the difficult test case 1.

CGD behaved in a similar fashion to L-BFGS, but with generally slightly less impressive RMSE results.

While RMSE is one error measure, because infrared is only able to provide information above the optical cloud-tops, the RMSE contribution below the cloud is basically irrelevant. Another measure that is more suitable and familiar in the field of meteorology is cloud-top pressure and cloud-top temperature. These two fields, along with their respective absolute errors, are shown in table 4.11. These two fields are calculated as the first layer when the cumulative cloud mixing ratio reaches $10^{-4}$ kg/kg or higher.

Table 4.11: Cloud-top pressure (CTP) and cloud-top temperature (CTT) and their respective absolute errors (true - final). Surface means that no clouds were detected.

| **Type** | **Case** | **Method** | **CTP** (hPa) | **CTT** (K) | **CTP Err** (hPa) | **CTT Err** (K) |
|---|---|---|---|---|---|---|
| clear | start cloudy | LMBM | Surface | 297.66 | 0.00 | 0.05 |
| clear | start cloudy | LBFGS | Surface | 297.61 | 0.00 | 0.10 |
| clear | start cloudy | CGD | Surface | 297.65 | 0.00 | 0.05 |
| cloudy | test case 1 | LMBM | 18.58 | 204.91 | 193.44 | 19.06 |
| cloudy | test case 1 | LBFGS | 12.65 | 209.27 | 199.38 | 14.70 |
| cloudy | test case 1 | CGD | 12.65 | 209.27 | 199.38 | 14.70 |
| cloudy | test case 2 | LMBM | 200.99 | 221.99 | 11.04 | 1.98 |
| cloudy | test case 2 | LBFGS | 212.03 | 224.72 | 0.00 | -0.76 |
| cloudy | test case 2 | CGD | 200.99 | 222.19 | 11.04 | 1.78 |
| cloudy | test case 3 | LMBM | 212.03 | 223.71 | 0.00 | 0.26 |
| cloudy | test case 3 | LBFGS | 212.03 | 223.48 | 0.00 | 0.49 |
| cloudy | test case 3 | CGD | 212.03 | 222.94 | 0.00 | 1.03 |
| cloudy | start clear | LMBM | Surface | 296.10 | -774.04 | -72.13 |
| cloudy | start clear | LBFGS | Surface | 252.42 | -774.04 | -28.46 |
| cloudy | start clear | CGD | Surface | 255.05 | -774.04 | -31.08 |
| cloudy | st. half min | LMBM | 212.03 | 224.19 | 0.00 | -0.22 |
| cloudy | st. half min | LBFGS | 200.99 | 221.73 | 11.04 | 2.23 |
| cloudy | st. half min | CGD | 200.99 | 221.80 | 11.04 | 2.16 |

This chart shows that for all but test case 1 and cloudy start clear, all three methods were able to find the cloud top pressure within a single layer (11 hPa) and 1 - 2 K for cloud top pressure.

## 4.7 Conclusions

In this chapter we evaluated the behavior of LMBM, L-BFGS, and the non-linear conjugate gradient CG-Descent method on the realistic problem of all-sky infrared satellite penalized 1D-Var data assimilation using RTTOV version 10 with multiple scattering. The problem itself is highly non-linear, with a sharp transition between clear and cloudy skies, and the penalty term introduces a mild discontinuity in the first derivative at the regression limits, although this discontinuity is not seen as having much impact on this particular problem after the first few iterations once the minimization algorithms focus on the feasible region. When started from clear-sky observations, the algorithms were all able to effectively clear out the clouds; in the case of cloudy-sky observations, all of the methods were also able to reduce the error generated from the background. For the difficult test case 1, with a perturbation of 100% (1e-2 kg/kg) in the cloud background, all three methods failed to reconstruct the desired profile, and both CG-Descent and L-BFGS terminated due to an error in the line search. As shown, starting from a clear sky, a cloudy profile will not be created; therefore, a method whereby a minimum cloudy profile is chosen was able to fully reconstruct the desired cloud with impressive accuracy.

All three optimization methods achieved a good deal of success in minimizing the cost function. However, when considering performance, RMSE reduction, and cloud-top errors for this problem, LMBM emerged as the best choice, followed closely by L-BFGS, with CG-Descent in a distant third. This may have been due in part to the CG-Descent optimization settings that were used with only minimal changes for this problem; further tweaking of the various parameters within CG-Descent, as in [5], may have made it more competitive. However, as seen from test case 1, when difficult cloud-clearing is required, LMBM is better suited to achieve the desired goal while both L-BFGS and CG-Descent fail in their line search routines. This is seen as a strength of LMBM's theoretically guaranteed descent line-search ([78]) even for problems that are not technically non-smooth but contain sharp transitions that may appear as discontinuities to optimization algorithms.

As noted in [194] and [105], LMBM is not as numerically stable as L-BFGS (nor CG-Descent, most likely), as LMBM requires double precision real numbers while L-BFGS can be operate well in single precision. It is thus recommended that before using LMBM in an operational setting that the main enhancements of LMBM – namely the bundling of sub-gradients, null-steps, and guaranteed descent line-search – be ported to the core numerically-stable code of L-BFGS. This step is highly recommended before any attempt is made to use LMBM operationally.

RTTOV version 10 is a sophisticated operational radiative transfer model which can be used for solving the all-sky IR radiance data assimilation problem. There are currently several undesirable features of RTTOV for this task, however. First of all, the fact that only two cloud types are allowable at each layer prevents all of the available species from a detailed microphysical scheme such as the WSM 6-class microphysics scheme ([92]) from being used. In addition, the limitation that only half of the layers can specify clouds is arbitrary and cumbersome to handle (fortunately this limitation will be removed in version 10.2, due to be released in 2012 [88]). Finally, because RTTOV is parameterized based on cumulus, stratus, and cirrus rather than cloud vapor, snow and ice crystals, rain and graupel along with their corresponding distributions, much of the detail that could be used by a full multiple scattering model is lost. Specification of the clouds based on microphysical properties would allow much finer detail of cloud properties to be derived and thus assimilated, especially when cloudy IR was combined with microwave observations.

In this chapter we also saw the importance of correctly choosing both a background and a background error covariance model. This problem was ill-posed as evidenced by the fact that several drastically different profiles were found by the optimization algorithms – in particular, test case 3, where QVapor was used incorrectly to compensate for the presence of a cloud. Having a background has its pros and cons, however, as a bad choice of a background or background error covariance could inadvertently hold the solution back. This was especially evident in the test case where a cloudy profile was started from a clear background, and as a result the gradient of the cost function forced the optimization algorithms to remain clear where clouds were needed. Therefore, an approach where the background was set to have half of the levels contain a very small amount of clouds – needed to activate the RTTOV cloudy adjoint but not overly influencing the brightness temperatures – was highly advantageous. This method, in conjunction with the enhanced LMBM, is therefore recommended for solving the all-sky IR radiance data assimilation problem. Updating RTTOV or another model to remove some or all of the limitations mentioned above would also be most beneficial.

# CHAPTER 5

# NON-SMOOTH OBSERVATION OPERATORS: SHALLOW WATER EQUATIONS

## 5.1 Introduction

In this chapter, we investigate data assimilation of the shallow water equations with discontinuous observation operators in order to compare the performance of large-scale non-smooth optimization methods. We compare and contrast variational (3D-Var and 4D-Var) approaches, ensemble/probabilistic (EnKF and LETKF) methods, as well as the Maximum Likelihood Ensemble Filter (MLEF) ([250], [248]) hybrid ensemble/variational data assimilation method. In light of [252] showing that MLEF can be derived without a differentiability requirement for the prediction model or the observation operator, we investigate the non-smooth optimization properties of MLEF. As in the previous chapter, we also compare and contrast the results obtained by using the L-BFGS, LMBM, and CG-Descent methods for large-scale non-smooth optimization within 4D-Var and MLEF. However, given the derivation of EnKF and LETKF and the heavy reliance on the linearity assumption of the observation operator, one would not expect these methods to perform well for these types of tests. At the time of writing, this work (without the comparison of 3D-Var, EnKF, and LETKF and without the conjugate gradient) has been accepted for publication in [194] and is now in early view.

MLEF has been tested with the shallow water equations in [251], [209] and [65], and the optimal control of the shallow water equations with a linear observation operator has been studied in e.g. [41]. Appel studied sensitivities for discontinuous fluid flows in [9]. The fully non-convex non-smooth variational data assimilation problem has been investigated in [144], [90], [235], [101] and [12] on highly simplified problems, and Levy et al. ([128]) recently investigated a physical-based approach for potentially discontinuous optimal interpolation of sea ice data assimilation for a model with 15 physical control variables. The approach in this chapter suggests techniques suitable for more general and larger-scale data assimilation problems, although combining these techniques is certainly an intriguing possibility.

As stated in the introduction, the smooth optimization quasi-Newton limited memory BFGS algorithm (L-BFGS), long used in data assimilation (e.g. [242], [91]), has recently been found to possess properties of a non-smooth optimization algorithm in [130], [129], and [185]. This method may offer promise for large-scale non-smooth optimal control problems.

We conduct our tests using a series of closely related optimal control problems with observation operators containing varying degrees of non-smoothness. We run these tests on all of our data assimilation methods with each of the optimization algorithms used above in order to make contrasts and comparisons.

## 5.2 Shallow Water Equations Model

We begin by describing the model that will be used in our optimal control problem.

Consider the limited area shallow water equation model as detailed in [215].

$$\frac{\partial u}{\partial t} = -u\frac{\partial u}{\partial x} - v\frac{\partial u}{\partial y} + fv - \frac{\partial \phi}{\partial x} \tag{5.1}$$

$$\frac{\partial v}{\partial t} = -u\frac{\partial v}{\partial x} - v\frac{\partial v}{\partial y} - fu - \frac{\partial \phi}{\partial y} \tag{5.2}$$

$$\frac{\partial \phi}{\partial t} = -\frac{\partial u\phi}{\partial x} - \frac{\partial v\phi}{\partial y} \tag{5.3}$$

where $u$ and $v$ are the two components of the horizontal velocity in m/s, $\phi = gh$ is the geopotential field in $\mathrm{m}^2/\mathrm{s}^2$, $h$ is the free surface height in m, and $f$ is the Coriolis factor in $\mathrm{s}^{-1}$.

The initial conditions used were based on those in [70], namely a channel on a $\beta$ plane of length $L$ and depth $D$, with $h$ given by

$$\begin{array}{rcl} h(x,y) & = & h_0 + h_1 \tanh\left(\frac{9(y-y_0)}{2D}\right) \\ & + & h_2 \,\mathrm{sech}^2\left(\frac{9(y-y_0)}{2D}\right)\sin\left(\frac{2\pi x}{L}\right) \end{array} \tag{5.4}$$

where $h_0 = 2000$ m, $h_1 = -220$ m, $h_2 = 133$ m, $L = 6000$ km, $D = 4400$ km, and $y_0 = D/2$.

From equation (5.5), the initial conditions are derived through geostrophic balance by the relation

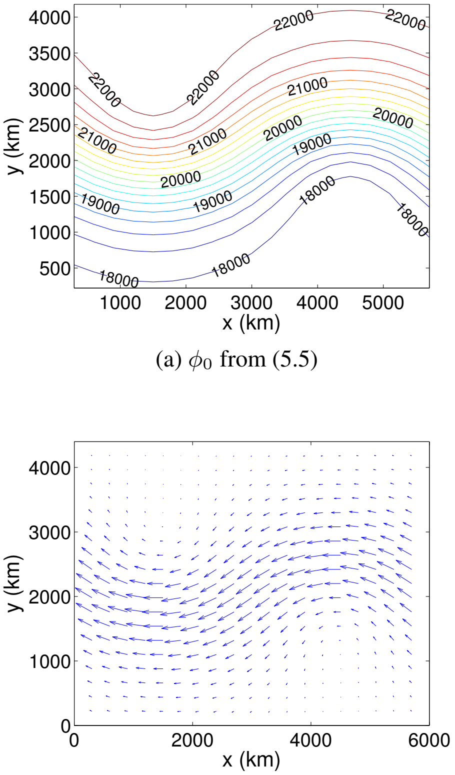
$$\begin{array}{rcl} \phi_0(x,y) & = & gh(x,y) \\ u_0(x,y) & = & -\frac{g}{f}\frac{\partial h}{\partial y}(x,y) \\ v_0(x,y) & = & \frac{g}{f}\frac{\partial h}{\partial x}(x,y) \end{array} \tag{5.5}$$

where $g$ = 10 m $\mathrm{s}^{-2}$ and $f = 10^{-4}$ $\mathrm{s}^{-1}$.

This model is discretized using the second-order quadratic conservation advective scheme detailed in [70] referred to as "scheme F." The space and time increments are $\Delta x = 300$ km, $\Delta y = 220$ km, and $\Delta t = 600$ s, respectively, resulting in a mesh comprising 21 × 21 spatial grid points. The model is integrated for 80 time steps, i.e. a window of assimilation of 13 hours 20 minutes in model time.

These initial conditions are shown in figure 5.1.

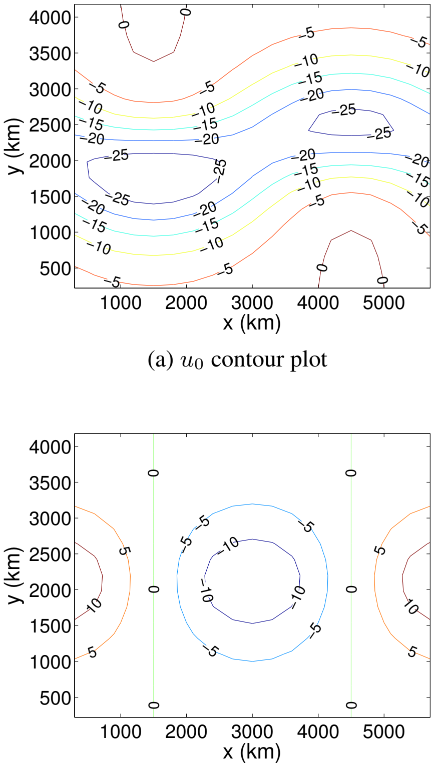
Because in this chapter the $u$ and $v$ velocity components will have separate observation operator components as described in section 5.3.2, the contour plots for the initial values of these two fields are shown in figure 5.2



(a) $\phi_0$ from (5.5)

(b) Wind field from equations (5.5). Arrows are scaled by a factor of 1 km

Figure 5.1: Initial conditions $(u_0, v_0, \phi_0)$



(a) $u_0$ contour plot

(b) $v_0$ contour plot

Figure 5.2: Contour plot for $u_0$, $v_0$

The boundary conditions are given by a rigid wall homogeneous Neumann condition along the south and north boundaries and a wrapping periodic conditions along the east/west boundary. In other words,

$$
\begin{array}{rcl}
u(x_l, y, t) & = & u(x_r, y, t) \\
v(x_l, y, t) & = & v(x_r, y, t) \\
\phi(x_l, y, t) & = & \phi(x_r, y, t) \\
\frac{\partial u}{\partial y}(x, y_t, t) & = & 0 \\
\frac{\partial u}{\partial y}(x, y_b, t) & = & 0 \\
v(x, y_t, t) & = & 0 \\
v(x, y_b, t) & = & 0 \\
\frac{\partial \phi}{\partial y}(x, y_t, t) & = & 0 \\
\frac{\partial \phi}{\partial y}(x, y_b, t) & = & 0
\end{array}
\qquad (5.6)
$$

where $x_l$, $x_r$, $y_t$, $y_b$ are locations of the left, right, top and bottom boundaries, respectively.

## 5.3 Experimental setup

We now consider data assimilation of the shallow water equations with a discontinuous observation operator detailed below.

Starting from the exact initial conditions given in (5.5) and boundary conditions in (5.6), the model is evolved forward in time. Observations of the model state $(u, v, \phi)$ are taken at each time step and every spatial grid point using an observation operator and are then perturbed with uncorrelated Gaussian noise of mean 0 and standard deviation $\sigma_{u_{obs}}$, $\sigma_{v_{obs}}$ and $\sigma_{\phi_{obs}}$, respectively. The exact initial conditions are perturbed with correlated Gaussian noise of mean 0 and covariance matrix $B$. The perturbed initial conditions are the *background* and the problem is to optimally reconstruct the exact solution by using background and observations.

We will use our data assimilation algorithms outlined in chapter 3 to solve this problem.

In this research, observations at all grid points are available for each time step – perhaps the best possible scenario for data assimilation. This removes the issue of sparsity of observations from the experimental setup in order to focus on the impact of non-smooth observation operators. Thus, assuming there are $N, M$ non-boundary grid points in the $x$ and $y$ direction, respectively, $N_{\text{state}} = N_{\text{obs}} = 3NM$.

### 5.3.1 Observation error covariance matrix

For this experiment, the observation error covariance matrix $R$ is taken to be diagonal. Thus, $R^{-1}$ is the diagonal inverse matrix with

$$R^{-1}_{i,i} = \begin{cases} 1/\sigma^2_{u_{obs}} & 1 \le i \le MN \\ 1/\sigma^2_{v_{obs}} & MN+1 \le i \le 2MN \\ 1/\sigma^2_{\phi_{obs}} & 2MN+1 \le i \le 3MN \end{cases} \tag{5.7}$$

In this experiment, $\sigma_{u_{obs}} = \sigma_{v_{obs}} = 1$ m/s and $\sigma_{\phi_{obs}} = 12$ m$^2$/s$^2$ were chosen based on approximate geostrophic considerations. A sample of this uncorrelated noise at time $t_0$ is shown in figure 5.3.

### 5.3.2 Non-Smooth observation operator

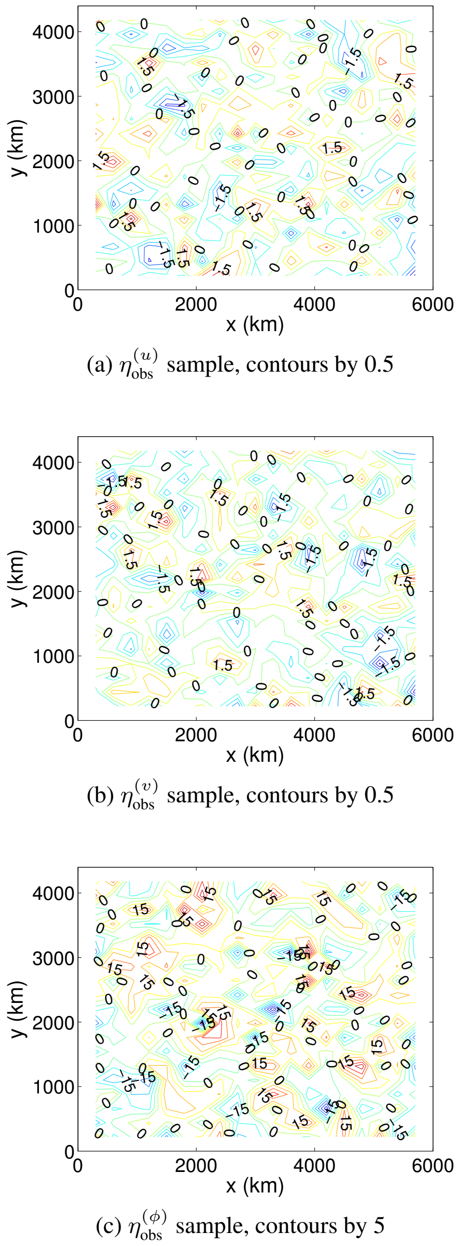
In this section we detail an observation operator with varying levels of non-smoothness in its components. This operator is not based on physical considerations but rather chosen solely to demonstrate the behavior of the optimization algorithms.

The observation operator $\mathcal{H}$ is given by

$$\mathcal{H}(x_i) = \begin{cases} \mathcal{H}_1(u_i) & 1 \le i \le MN \\ \mathcal{H}_2(v_{i-MN}) & MN+1 \le i \le 2MN \\ \mathcal{H}_3(\phi_{i-2MN}) & 2MN+1 \le i \le 3MN \end{cases} \tag{5.8}$$

$$\mathcal{H}_1(u_i) = \begin{cases} u_i^3/u_{min}^2 & u_i < u_{min} \\ u_i^2/u_{max} & u_i \ge u_{max} \\ u_i & \text{else} \end{cases} \tag{5.9}$$

$$\mathcal{H}_2(v_i) = \begin{cases} \log(v_i + \delta) & v_i \ge 0 \\ \log(-v_i + \delta) & v_i < 0 \end{cases} \tag{5.10}$$



(a) $\eta_{\text{obs}}^{(u)}$ sample, contours by 0.5

(b) $\eta_{\text{obs}}^{(v)}$ sample, contours by 0.5

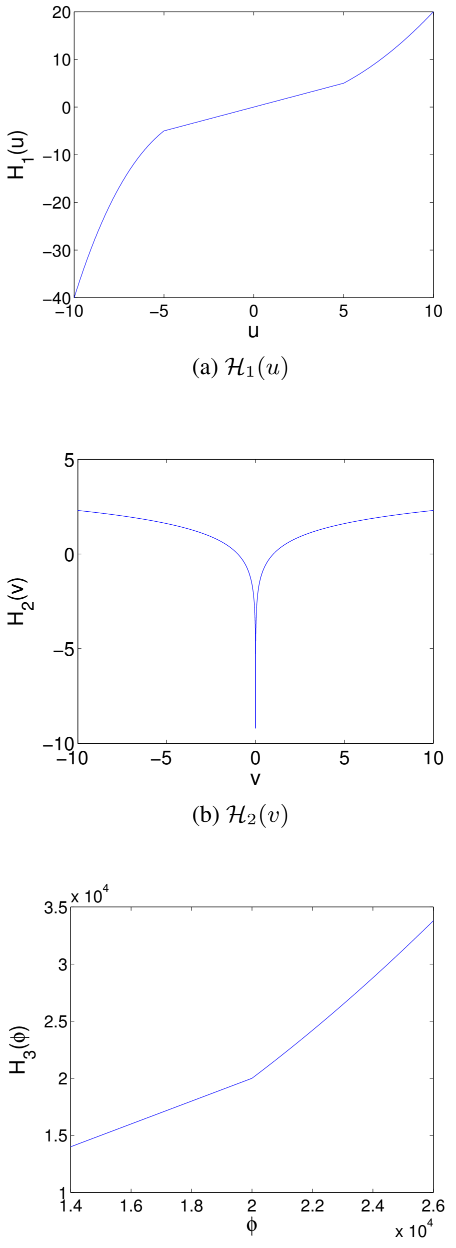
(c) $\eta_{\text{obs}}^{(\phi)}$ sample, contours by 5

Figure 5.3: Observation error sample for $u$, $v$, and $\phi$ at $t_0$

$$\mathcal{H}_3(\phi_i) = \left\{ \begin{array}{rl} \phi_i & \phi_i < H_{max} \\ \phi_i^2/H_{max} & \phi_i \geq H_{max} \end{array} \right. \tag{5.11}$$

Here, $u_{min} = -5$ m/s, $u_{max}$ = 5 m/s, and $H_{max} = 20000$ m.

These components of the observation operator are shown in figure 5.4, and the observation operator of the initial state plus observational noise is shown in figure 5.5.

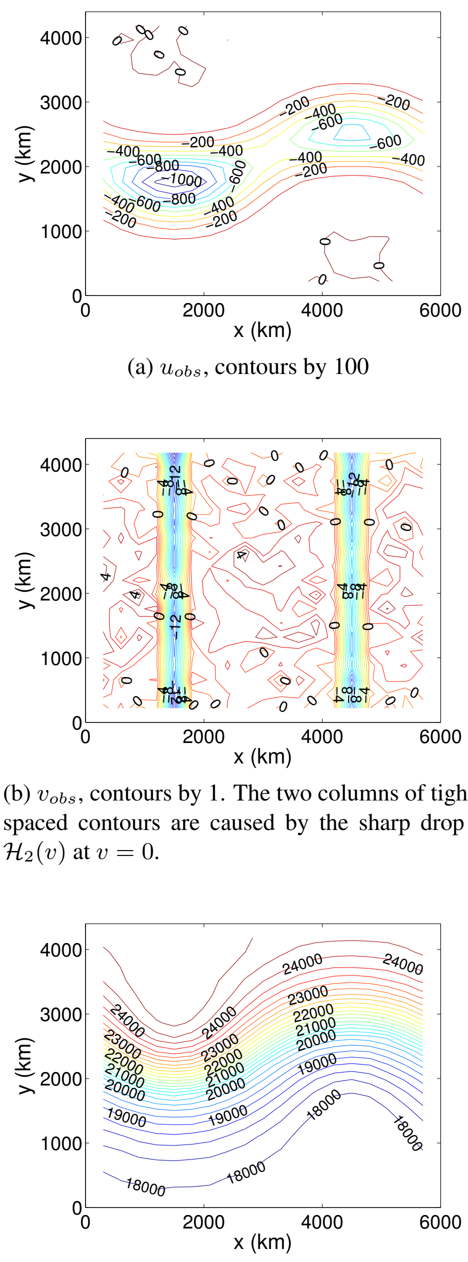


(a) $\mathcal{H}_1(u)$

(b) $\mathcal{H}_2(v)$

(c) $\mathcal{H}_3(\phi)$

Figure 5.4: Observation operator



(a) $u_{obs}$, contours by 100

(b) $v_{obs}$, contours by 1. The two columns of tightly spaced contours are caused by the sharp drop in $\mathcal{H}_2(v)$ at $v = 0$.

(c) $\phi_{obs}$, contours by 250

Figure 5.5: Obs. sample for $u$, $v$, and $\phi$ at $t_0$

Due to the kinks in the observation operators, this becomes a non-smooth optimization problem. The discontinuity in the piecewise derivative of $\mathcal{H}_1$ through $\mathcal{H}_3$ are shown in figure 5.6. Note that the discontinuity becomes progressively more acute in $\nabla\mathcal{H}_3$, $\nabla\mathcal{H}_1$, and $\nabla\mathcal{H}_2$, giving flexibility for testing the behavior of non-smooth optimization algorithms. In $\nabla\mathcal{H}_2$, the parameter $\delta$ controls the size of the discontinuity. All of these functions are locally Lipschitz continuous, and the best global Lipschitz constant for $\mathcal{H}_2$ is $1/\delta$.

### 5.3.3 Background error covariance matrix

In this work we use an exact background error covariance matrix to perturb the background $x_b$. The perturbation to the background vector is given by

$$\begin{aligned} u_b &= u_0 + \eta_u \\ v_b &= v_0 + \eta_v \\ \phi_b &= \phi_0 + \eta_\phi \end{aligned} \tag{5.12}$$

where $(u_i, v_i, \phi_i)$ are the exact initial conditions given in (5.5), $\eta_u \sim N(0, \Sigma_u)$, $\eta_v \sim N(0, \Sigma_v)$, and $\eta_\phi \sim N(0, \Sigma_\phi)$ where $\Sigma_u$, $\Sigma_v$, $\Sigma_\phi \in \mathbb{R}^{NM\times NM}$ are the covariance matrices of $(u, v, \phi)$. Thus, $B$ is the block matrix

$$B = \begin{pmatrix} \Sigma_u & 0 & 0 \\ 0 & \Sigma_v & 0 \\ 0 & 0 & \Sigma_\phi \end{pmatrix} \tag{5.13}$$

and

$$B^{-1} = \begin{pmatrix} \Sigma_u^{-1} & 0 & 0 \\ 0 & \Sigma_v^{-1} & 0 \\ 0 & 0 & \Sigma_\phi^{-1} \end{pmatrix} \tag{5.14}$$

$\Sigma_u = \sigma_u^2\Sigma$, $\Sigma_v = \sigma_v^2\Sigma$, and $\Sigma_\phi = \sigma_\phi^2\Sigma$ are created by the exponential squared kernel of

$$\Sigma = B^{1/2}B^{\mathrm{T}/2} \tag{5.15}$$

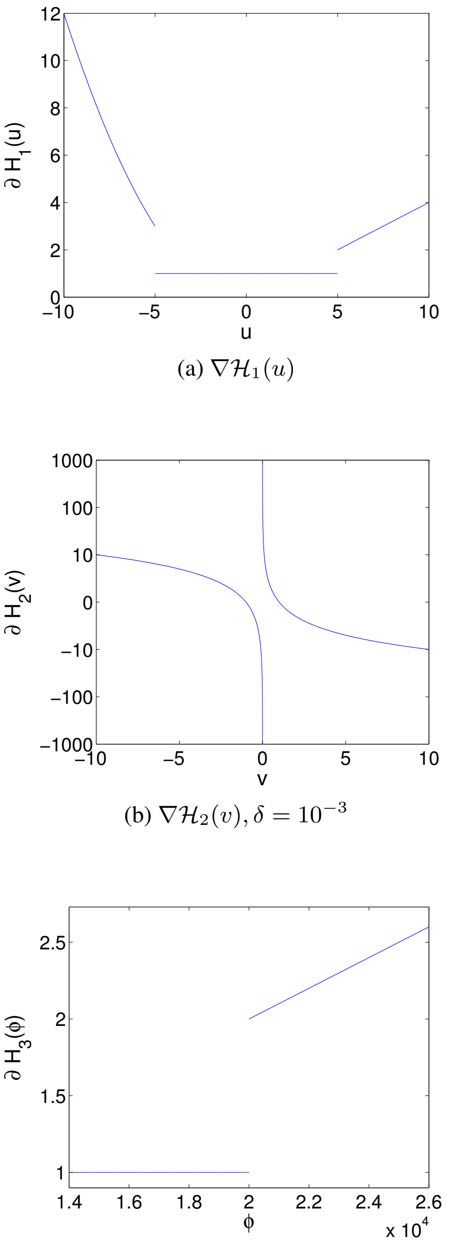
and

$$B^{1/2}_{i,j} = B^{1/2}_{j,i} = \exp(-r_{i,j}^2/L^2) \tag{5.16}$$

Here, $r$ represents the Euclidean distance between two points in the grid, i.e. $r_{i,j}^2 = \Delta x_{i,j}^2 + \Delta y_{i,j}^2$ where $\Delta x_{i,j}$ and $\Delta y_{i,j}$ are the $x$– and $y$–coordinate distances, in meters, between the two points with global indexes $i$ and $j$, respectively. $L$ represents the correlation length, in meters, required for the correlation between two points to reach $1/e \approx 0.36788$.

The normal perturbation $\eta_u$ is created with the transformation $\eta_u = \Sigma_u^{1/2} Z_u$, where $Z_u \in \mathbb{R}^{NM}$ and the components of $Z_{u_i} \sim N(0, 1)$ are i.i.d. standard normal variables. $\eta_v$ and $\eta_\phi$ are created the same manner with $\Sigma_v$, $\Sigma_\phi$, $Z_v$ and $Z_\phi$ uncorrelated standard normal variables, respectively.

Based on the results in [251], a correlation length of $L = 7000$ km and background perturbation magnitudes of $\sigma_u = \sigma_v = 20$ m and $\sigma_\phi = 200$ m$^2$/s$^2$ were chosen. As discussed below, the background used in the data assimilation experiments are found by taking an ensemble average over 96 realizations of $(u_b, v_b, \phi_b)$. The actual perturbation versus the exact solution used as the background is shown in figure 5.7.

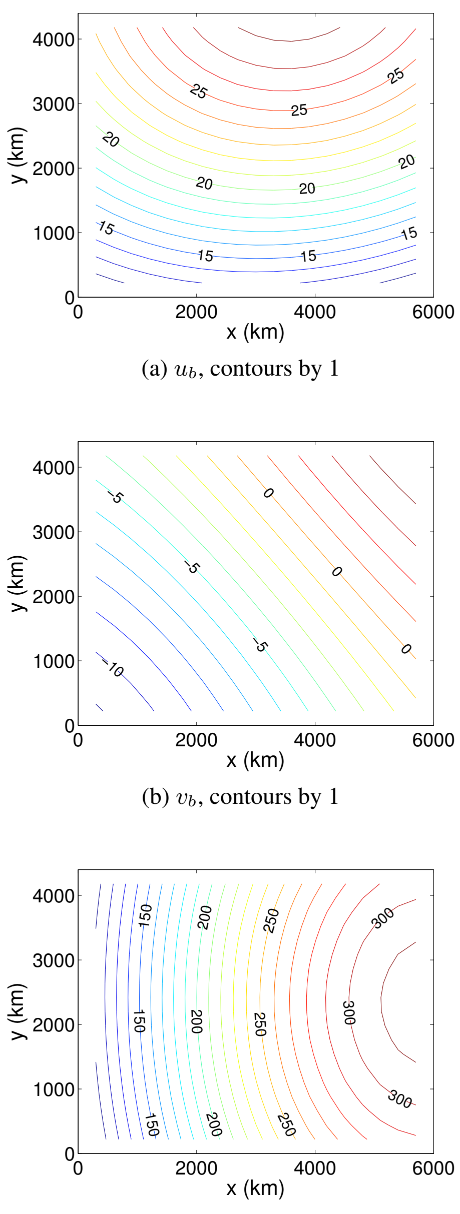


(a) $\nabla\mathcal{H}_1(u)$

(b) $\nabla\mathcal{H}_2(v), \delta = 10^{-3}$

(c) $\nabla\mathcal{H}_3(\phi)$

Figure 5.6: Piecewise derivative of $\mathcal{H}$



(a) $u_b$, contours by 1

(b) $v_b$, contours by 1

(c) $\phi_b$, contours by 10

Figure 5.7: $\eta = x_b - x_{\text{true}}$

### 5.3.4 Transformed cost function

As in the previous chapter for 1D-Var, we once again have a matrix $B$ that has, in theory, full rank, but is numerically rank deficient due to the effect of having a much larger correlation length than the grid spacing. We once again apply the change of variables transformation $\delta_b(x) = x - x_b = B^{1/2}z$ and use $z$ as the control variable. Since $B^{1/2}B^{\mathrm{T}/2} = B$, this removes the necessity of obtaining the inverse from the cost function, so that the 3D-Var cost function (3.7) becomes

$$J(z) = \frac{1}{2}z^{\mathrm{T}}z + \frac{1}{2}\,\delta_{y_k}(z'(z))^{\mathrm{T}}\,R^{-1}\delta_{y_k}(z'(z)) \tag{5.17}$$

where $z'(z) = B^{1/2}z + x_b$, and $\delta_{y_k}(z'(z)) = y_k - \mathcal{H}(B^{1/2}z + x_b)$ at the model time step $k$, while the subgradient becomes

$$\nabla_z J(z) = z - B^{1/2}\left(\frac{\partial \mathcal{H}}{\partial z'}\right)^{\mathrm{T}} R^{-1}\delta_{y_k}(z'(z)) \tag{5.18}$$

The transformation is similar for 4D-Var, with the cost function given by

$$J(z) = \frac{1}{2}z^{\mathrm{T}}z + \frac{1}{2}\sum_{k=0}^{NT}\delta_{y_k}(x^{(k)}(z'(z)))^{\mathrm{T}}\,R^{-1}\delta_{y_k}(x^{(k)}(z'(z))) \tag{5.19}$$

where $x^{(k)}(z'(z)) = \mathcal{M}(x^{(k-1)}(z'(z)), x^{(0)}(z'(z)) = z'(z)$, and the subgradient becomes

$$\nabla_z J(z) = z - \sum_{k=0}^{NT} B^{1/2}\left(\frac{\partial x^{(k)}}{\partial z'}\right)^{\mathrm{T}}\left(\frac{\partial \mathcal{H}}{\partial x^{(k)}}\right)^{\mathrm{T}} R^{-1}\delta'_{y_k} \tag{5.20}$$

### 5.3.5 Design

In order to compare our various methods in the presence of non-smooth observation operators, the following procedure is used:

1) The initial condition for $h$ is listed in equation (5.4). From this, $u_0$, $v_0$, and $\phi_0$ are created by (5.5).

2) These conditions are evolved forward in time by solving (5.1) through (5.3) to create $x_{\text{exact}} = (u_{\text{exact}}, v_{\text{exact}}, \phi_{\text{exact}})$ at times $t = 0, \ldots, NT$.

3) Observations: Gaussian noise, as detailed in section 5.3.1, is added to $\mathcal{H}\left(x_{\text{exact}}\right)$ to create the observations

Ensemble methods (including MLEF):

4a) Ensemble members: each member of the initial ensemble is created by sampling from a correlated Gaussian random variable with mean $x_{\text{exact}}$ and covariance matrix $B$ as discussed in section 5.3.3.

5a) Control/mean state: the mean of the ensemble created in 4a) is used as the control state

Variational methods:

4b) Background value: the MLEF initial control state from 5a) is used as the background value

5b) Background error covariance matrix: the matrix $B$ used in 4a) to create the ensemble is used. The matrix $B^{1/2}$ is found using an eigenvalue decomposition.

All methods:

6) The experiment is then run and the RMSE is taken at each time step versus $(u_{\text{exact}}, v_{\text{exact}}, \phi_{\text{exact}})$ as detailed in section 5.3.7.

### 5.3.6 Experiments

Within the context above, we now design three numerical experiments to test the performance of L-BFGS and LMBM within the data assimilation frameworks in order to assess their performance in data assimilation in the presence of non-differentiable observation operators of increasing difficulty. The experimental setup is shown in table 5.1. Experiment 1 is the most favorable experimental setup for data assimilation with all linear (and thus differentiable) observation operators. Experiment 2 possesses only a "slight" non-smoothness in the observation operator for $\phi$, experiment 3 has a discontinuity in both $u$ and $\phi$, while experiment 4 constitutes the most difficult case with a sharp discontinuity in the observation operator for $v$ and the same operator as experiment 3 for $u$ and $\phi$.

Table 5.1: Experimental setup

| **Experiment #** | $u$ **obs op** | $v$ **obs op** | $\phi$ **obs op** |
|---|---|---|---|
| 1 | linear | linear | linear |
| 2 | linear | linear | $\mathcal{H}_3$ |
| 3 | $\mathcal{H}_1$ | linear | $\mathcal{H}_3$ |
| 4 | $\mathcal{H}_1$ | $\mathcal{H}_2$ | $\mathcal{H}_3$ |

### 5.3.7 Success criteria

We now define our success criteria for the experiments listed in section 5.3.6.

To judge the quality of the assimilation results, we use the root mean squared error (RMSE) of the calculated solution versus the exact solution. In the same fashion as equation (4.25), the RMSE for cycle $(k)$ is calculated as follows:

$$\begin{aligned}
\mathrm{RMSE}_u^{(k)} &= \sqrt{\frac{\left(u_{\text{exact}}^{(k)}-u^{(k)}\right)^{\mathrm{T}}\left(u_{\text{exact}}^{(k)}-u^{(k)}\right)}{NM}} \\
\mathrm{RMSE}_v^{(k)} &= \sqrt{\frac{\left(v_{\text{exact}}^{(k)}-v^{(k)}\right)^{\mathrm{T}}\left(v_{\text{exact}}^{(k)}-v^{(k)}\right)}{NM}} \\
\mathrm{RMSE}_\phi^{(k)} &= \sqrt{\frac{\left(\phi_{\text{exact}}^{(k)}-\phi^{(k)}\right)^{\mathrm{T}}\left(\phi_{\text{exact}}^{(k)}-\phi^{(k)}\right)}{NM}}
\end{aligned} \tag{5.21}$$

As is common in data assimilation experiments, success is judged by the assimilation achieving an RMSE that is lower than both the observation and background errors. In this case, the expected RMSE from simply using the observations is much lower than that the expected RMSE of the background. Thus, in order for the data assimilation procedure to be considered a success, the RMSE must reach a level lower than the observational noise $(\sigma_{u_{obs}}, \sigma_{v_{obs}}.\sigma_{\phi_{obs}})$ from section 5.3.1.

### 5.3.8 Optimization settings

As in the previous chapter, in this section we present the optimization settings used for each of our three algorithms. The LMBM optimization settings are shown in table 5.2, the settings for L-BFGS are shown in table 5.3, and the CG-Descent settings are shown in table 5.4. Again details for these parameters can be found in [105], [136], and [82], respectively.

Table 5.2: LMBM optimization settings

| Parameter | Value | Parameter | Value |
|---|---|---|---|
| NA | 4 | MCU | 9 |
| MC | 9 | NW | Default |
| RPAR(1-2) | $10^{-4}$ | IPAR(1) | 1 |
| RPAR(3) | 0 | IPAR(2) | 300 |
| RPAR(4-5) | $10^{-16}$ | IPAR(3) | 300 |
| RPAR(6) | 1 | IPAR(4) | 5 |
| RPAR(7) | $10^{-8}$ | IPAR(6) | 0 |
| RPAR(8) | 1 | IPAR(7) | 1 |

Table 5.3: L-BFGS optimization settings

| Parameter | Value | Parameter | Value |
|---|---|---|---|
| MSave | 9 | $\epsilon$ | $10^{-16}$ |
| Max iter | 300 | diagco | false |

Table 5.4: CG-Descent optimization settings

| Parameter | Value | Parameter | Value |
|---|---|---|---|
| grad_tol | $10^{-16}$ | $\delta$ | $10^{-1}$ |
| $\sigma$ | 0.9 | eps | $10^{-3}$ |
| $\gamma$ | 0.66 | $\rho$ | 2.0 |
| $\eta$ | $10^{-2}$ | psi0 | $10^{-2}$ |
| psi1 | $10^{-1}$ | psi2 | 2.0 |
| QuadCutOff | $10^{-5}$ | StopFact | $10^{-4}$ |
| AWolfeFac | $10^{-3}$ | restart_fac | 1.0 |
| maxit_fac | 0.1 | feps | $10^{-5}$ |
| Qdecay | 0.7 | nexpand | 50 |
| nsecant | 50 | PertRule | true |
| StopRule | true | AWolfe | true |
| Step | false | debug | false |

Note that for MLEF, LMBM benefited from a larger maximum allowable step-size (`RPAR(8)` = $10^3$) than that used for 3D-Var or 4D-Var. The other parameters were the same. A similar tweaking of parameters for MLEF was not thought to be of benefit to either CG-Descent or L-BFGS as they have adaptive step-size parameters. As discussed in [105], selecting the `RPAR(8)` parameter is an important part of tuning LMBM that is problem dependent; as MLEF and 3/4D-Var use differently scaled cost functions, it makes sense that this parameter would behave differently for these problems. This a priori knowledge of the maximum step-size may give LMBM somewhat of an advantage over the other methods; however, based it is not anticipated that this parameter plays a determining role in a comparison of these method's performance.

The other parameters were cursorily tested for stability versus the optimization results, and all were found to be fairly stable with the exception of `MC`/`MCU` for LMBM and `MSave` for L-BFGS. These two parameters serve approximately the same purpose in their respective algorithms; namely this parameter represents the number of correction pairs of the inverse generalized Hessian to store. It was found that on this problem, these values need to be above 7 to achieve decent results, while values between 8-20 gave nearly identical results as those presented below for both LMBM and L-BFGS.

## 5.4 Numerical results

In this section we present the numerical results for our data assimilation methods with the experimental setup detailed in section 5.3.

### 5.4.1 Timings

In the sections that following, timings are presented for each method from start until finish. These timings are from a Macintosh Intel Core 2 Duo 2.13 GHz Processor with 2 GB of RAM running OS X 10.6.8 (64-bit). The code was compiled using `gfortran` with an optimization level of `-O3`. If the method in question failed to converge or diverged for a given experiment, a marker of either $^*$ or $^\dagger$, respectively, is given next to the timing in seconds. Failure to converge means that the one or more variables in the assimilated state spent more than 10% of the model time steps outside of the success region defined in section 5.3.7; divergence means that one or assimilated variables more spent more than 10% of the model time steps above the level of background error in RMSE for that variable.

In the tables that follow the code was run ten times, and the average time is presented in seconds. The standard deviation presented is the sample standard deviation, i.e. a correction of $n - 1$ was used inside the square-root.

### 5.4.2 Ensemble filter results

For EnKF and LETKF, the four experiments from section 5.3.6 are run with an ensemble of 96 members started from the procedure described in section 5.3.5. The $\delta$ parameter for $\mathcal{H}_2$ is set to $\delta = 10^{-4}$, and the covariance localization length for LETKF, discussed in chapter 3 was set to 7000 km.

The results for EnKF and LETKF are shown in figures 5.9 and 5.10, respectively. As evidenced by experiment 1, both EnKF and LETKF are able to handle linear observation operators quite nicely, as expected. Surprisingly, however, both EnKF and LETKF have nearly identical results for experiment 2, which involves both a slight non-linearity as well as non-smoothness. After the first iteration, both EnKF and LETKF were able to remain within the successful region. These results suggest that EnKF and LETKF can both be used with the observation operator is only mildly non-linear or non-smooth. However, as evidenced by experiment 3 and 4, when the non-smoothness is strong these methods both diverge.

The timings for EnKF and LETKF are shown in table 5.5 in accordance with section 5.4.1.

Table 5.5: Timings for EnKF and LETKF. A $^*$ denotes that the method failed to converge, while $^\dagger$ denotes that the method diverged.

| Method | Obs. operator | Opt. algorithm | Time in sec | Std. dev. |
|---|---|---|---|---|
| EnKF | 1 | – | 9.54 | 0.1865 |
| EnKF | 2 | – | 9.47 | 0.01728 |
| EnKF | 3 | – | $9.51^\dagger$ | 0.04586 |
| EnKF | 4 | – | $9.57^\dagger$ | 0.01895 |
| LETKF | 1 | – | 7.69 | 0.05028 |
| LETKF | 2 | – | 7.73 | 0.01278 |
| LETKF | 3 | – | $7.93^\dagger$ | 0.1260 |
| LETKF | 4 | – | $1.90^\dagger$ | 0.05719 |

### 5.4.3 3D-Var results

The four experiments from section 5.3.6 are run for the cost function (5.17) with the setup described in section 5.3.5. The assimilated state at time $k$ is moved forward in time using the model operator $\mathcal{M}$, and this becomes the new background at time $k+1$. As with the ensemble case, the $\delta$ parameter for $\mathcal{H}_2$ is set to $\delta = 10^{-4}$.

The results for LMBM, L-BFGS, and CG-Descent are shown in figure 5.11 through 5.13. As these results show, all three methods were able to handle the first two experiments, although only LMBM and L-BFGS were able to handle experiment 3. None of the optimization methods were able to handle the fourth case successfully with the 3D-Var algorithm, although at least LMBM and CG-D are able to reduce the RMSE from the level of the background while L-BFGS diverges on this test case. These results suggest that the fourth experiment with straight 3D-Var – where the covariance model is static, and no covariance localization or preconditioning of the cost function is applied – is too difficult for these algorithms to handle.

The timings for 3D-Var is shown in table 5.6 in accordance with section 5.4.1.

### 5.4.4 4D-Var results

For 4D-Var, the four experiments from section 5.3.6 are run for the cost function (5.19) with the setup described in section 5.3.5. As with the previous cases, $\delta = 10^{-4}$.

The results of using 4D-Var are shown in figure 5.14 to 5.16. LMBM is able to converge for all four experiments, while CG-Descent only converges for experiments 1 and 2. As shown in figure 5.8, the reason is because of the excessive function evaluations that force termination after 300

Table 5.6: Timings for 3D-Var. A $^*$ denotes that the method failed to converge, while $^\dagger$ denotes that the method diverged.

| Method | Obs. operator | Opt. algorithm | Time in sec | Std. dev. |
|---|---|---|---|---|
| 3D-Var | 1 | L-BFGS | 7.08 | 0.116 |
| 3D-Var | 1 | LMBM | 21.10 | 1.440 |
| 3D-Var | 1 | CG-D | 11.44 | 0.6081 |
| 3D-Var | 2 | L-BFGS | 10.33 | 0.5984 |
| 3D-Var | 2 | LMBM | 20.69 | 0.6462 |
| 3D-Var | 2 | CG-D | 20.59 | 1.030 |
| 3D-Var | 3 | L-BFGS | 19.67 | 0.5915 |
| 3D-Var | 3 | LMBM | 38.48 | 0.8057 |
| 3D-Var | 3 | CG-D | 37.73$^*$ | 1.298 |
| 3D-Var | 4 | L-BFGS | 30.00$^\dagger$ | 0.7822 |
| 3D-Var | 4 | LMBM | 41.61$^*$ | 1.114 |
| 3D-Var | 4 | CG-D | 40.95$^*$ | 1.219 |

function evaluations. For the fourth experiment, L-BFGS does not successfully reduce the RMSE below that expected only from observations, and thus has failed on this challenging non-smooth case. Changing the line search from strong to weak Wolfe conditions, as suggested in [130], does not remedy the situation. LMBM, however, is able to handle this situation with the same level of accuracy as the other cases.

The timings for 4D-Var are shown in table 5.7.

Table 5.7: Timings for 4D-Var. A $^*$ denotes that the method failed to converge, while $^\dagger$ denotes that the method diverged.

| Method | Obs. operator | Opt. algorithm | Time in sec | Std. dev. |
|---|---|---|---|---|
| 4D-Var | 1 | L-BFGS | 3.83 | 0.129 |
| 4D-Var | 1 | LMBM | 3.87 | 0.126 |
| 4D-Var | 1 | CG-D | 3.67 | 0.134 |
| 4D-Var | 2 | L-BFGS | 3.94 | 0.0127 |
| 4D-Var | 2 | LMBM | 4.03 | 0.126 |
| 4D-Var | 2 | CG-D | 3.86 | 0.131 |
| 4D-Var | 3 | L-BFGS | 4.31 | 0.0200 |
| 4D-Var | 3 | LMBM | 4.51 | 0.104 |
| 4D-Var | 3 | CG-D | 4.26 | 0.0692 |
| 4D-Var | 4 | L-BFGS | 4.94$^*$ | 0.213 |
| 4D-Var | 4 | LMBM | 4.92 | 0.0455 |
| 4D-Var | 4 | CG-D | 4.91$^*$ | 0.0557 |

The performance of the LMBM, L-BFGS, and CG-Descent methods (measured in terms of cost function value versus number of cost function evaluations) is shown in figure 5.8. The results demonstrate that the more challenging non-smooth experiments require more iterations, especially for L-BFGS and CG-Descent.

### 5.4.5 MLEF results

For MLEF, 96 ensemble members are used with the experimental setup detailed in section 5.3.5 to test experiments 1-4 using MLEF. Recall from chapter 3 that MLEF minimizes the cost function given in equation (3.48) in the ensemble space rather than state space; thus there are only 96 members of the control variable.

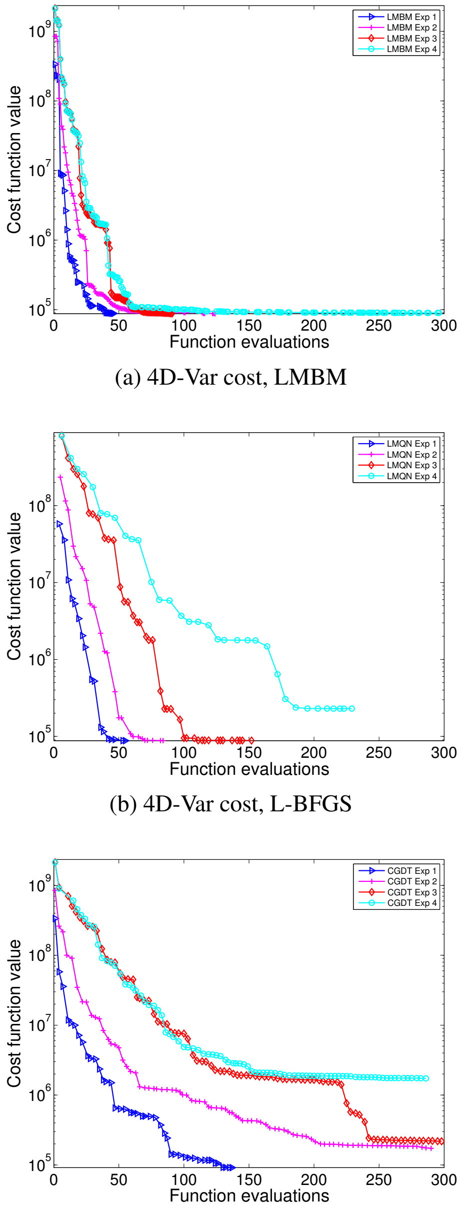
The RMSE for the four experiments using MLEF and the smooth optimization algorithm are shown for the $u$ and $v$ components of the velocity field along with the geopotential field $\phi$ in figures 5.17 through 5.19. The $\delta$ parameter is again set to $\delta = 10^{-4}$.

The results show that both CG-Descent and L-BFGS are able to handle the first three experiments with MLEF, only failing to converge on the final difficult experiment. LMBM, as with 4D-Var, is able to handle all four experiments. These results demonstrate, as predicted by theory in [250], MLEF can handle slightly non-smooth cases even with an algorithm originally designed for smooth optimization in place. However, the RMSE from experiment 4 shows that MLEF with the smooth L-BFGS and CG-Descent algorithms have difficulty with a highly non-smooth data assimilation case.

The timings for MLEF are shown in table 5.8.

Table 5.8: Timings for MLEF. A $^*$ denotes that the method failed to converge, while $^\dagger$ denotes that the method diverged.

| Method | Obs. operator | Opt. algorithm | Time in sec | Std. dev. |
|---|---|---|---|---|
| MLEF | 1 | L-BFGS | 13.05 | 0.05416 |
| MLEF | 1 | LMBM | 15.94 | 0.2420 |
| MLEF | 1 | CG-D | 19.61 | 0.1104 |
| MLEF | 2 | L-BFGS | 13.12 | 0.01653 |
| MLEF | 2 | LMBM | 16.01 | 0.2337 |
| MLEF | 2 | CG-D | 20.07 | 0.1238 |
| MLEF | 3 | L-BFGS | 14.03 | 0.02670 |
| MLEF | 3 | LMBM | 17.62 | 0.5439 |
| MLEF | 3 | CG-D | 24.34 | 0.1119 |
| MLEF | 4 | L-BFGS | 15.24$^*$ | 0.1109 |
| MLEF | 4 | LMBM | 29.74 | 0.3023 |
| MLEF | 4 | CG-D | 49.86$^*$ | 0.1674 |

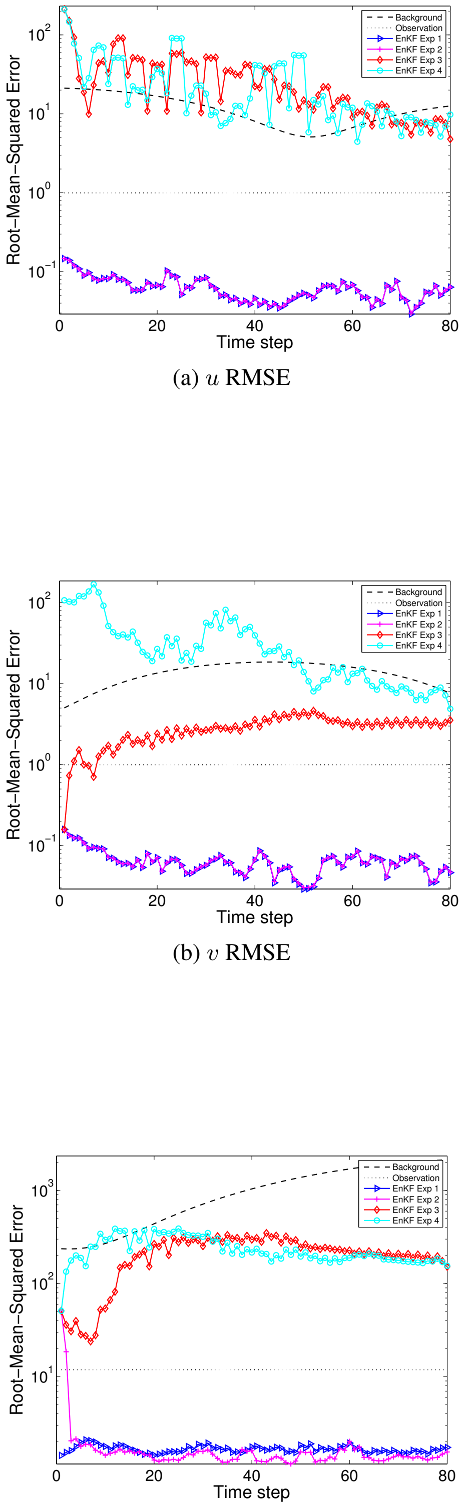


(a) 4D-Var cost, LMBM

(b) 4D-Var cost, L-BFGS

(c) 4D-Var cost, CG-Descent

Figure 5.8: 4D-Var cost history. The $x$-axis shows the number of cost function evaluations required for the minimization, while the $y$-axis shows the cost function achieved at that point in the optimization.

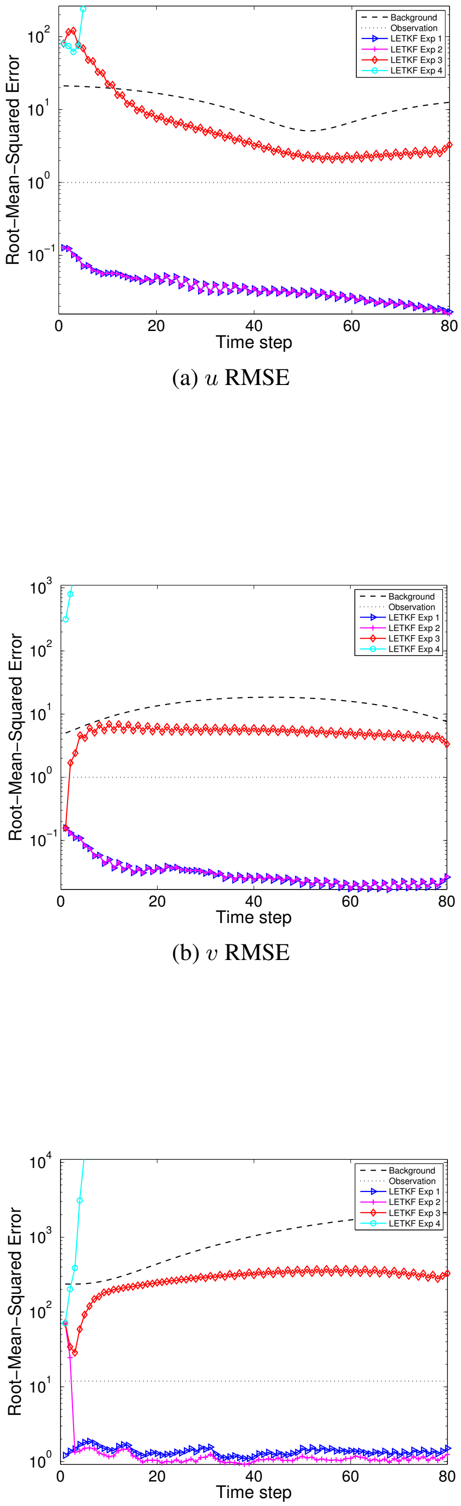


(a) $u$ RMSE

(b) $v$ RMSE

(c) $\phi$ RMSE

Figure 5.9: EnKF

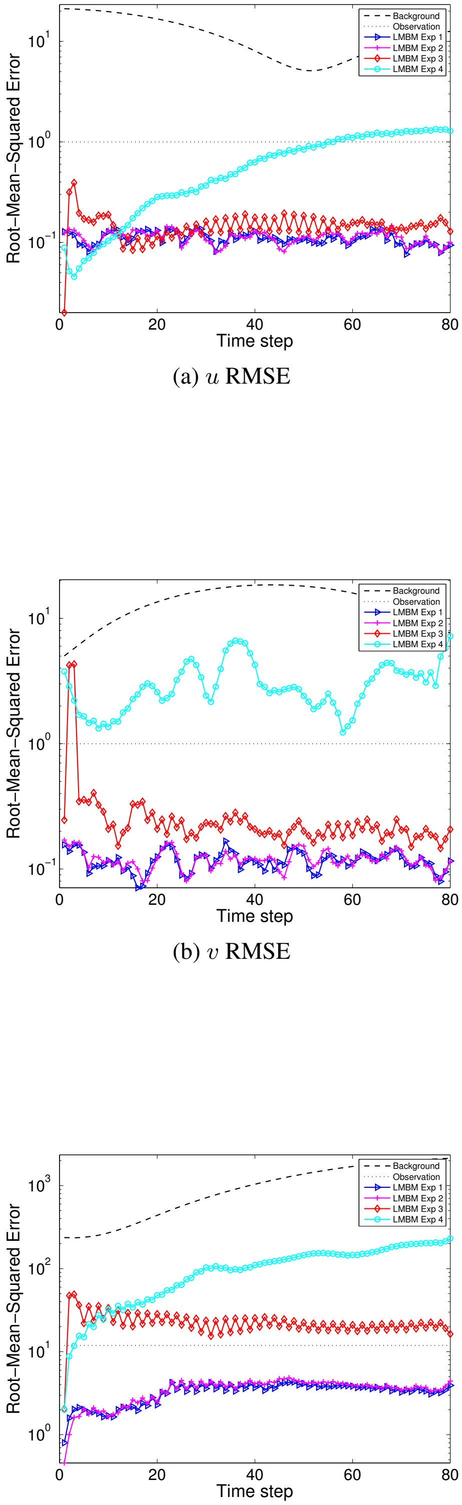


(a) $u$ RMSE

(b) $v$ RMSE

(c) $\phi$ RMSE

Figure 5.10: LETKF

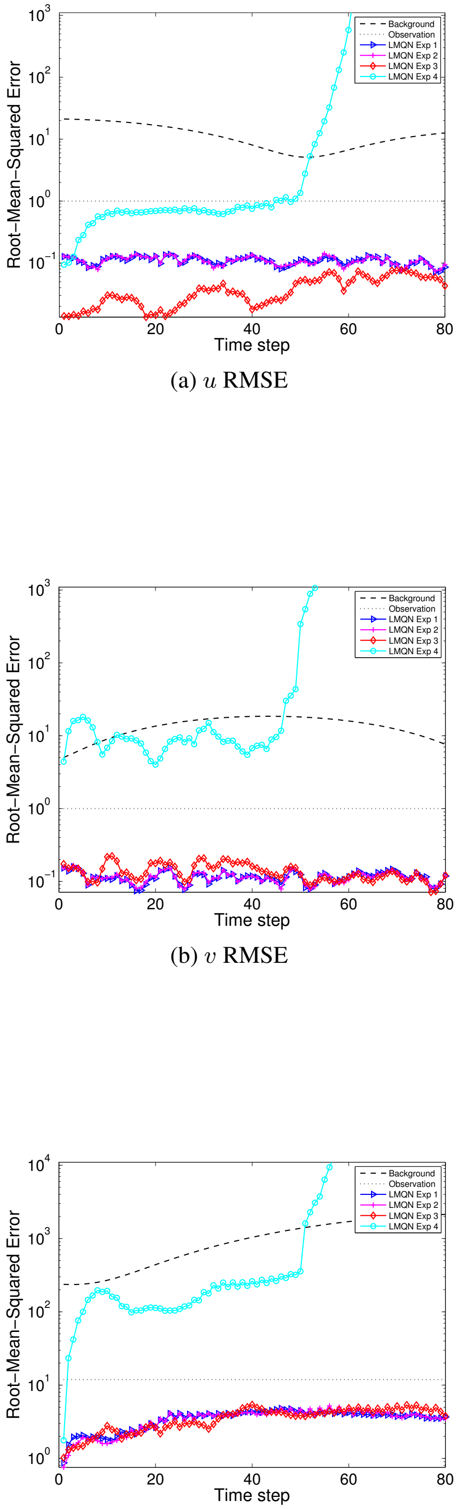


(a) $u$ RMSE

(b) $v$ RMSE

(c) $\phi$ RMSE

Figure 5.11: 3D-Var, LMBM

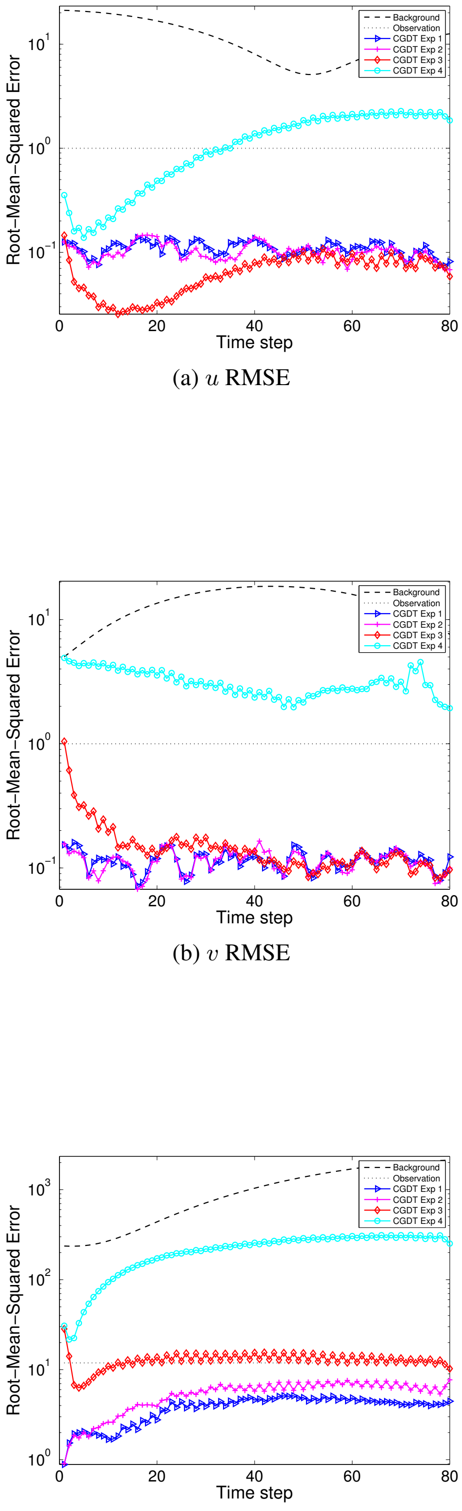


(a) $u$ RMSE

(b) $v$ RMSE

(c) $\phi$ RMSE

Figure 5.12: 3D-Var, L-BFGS

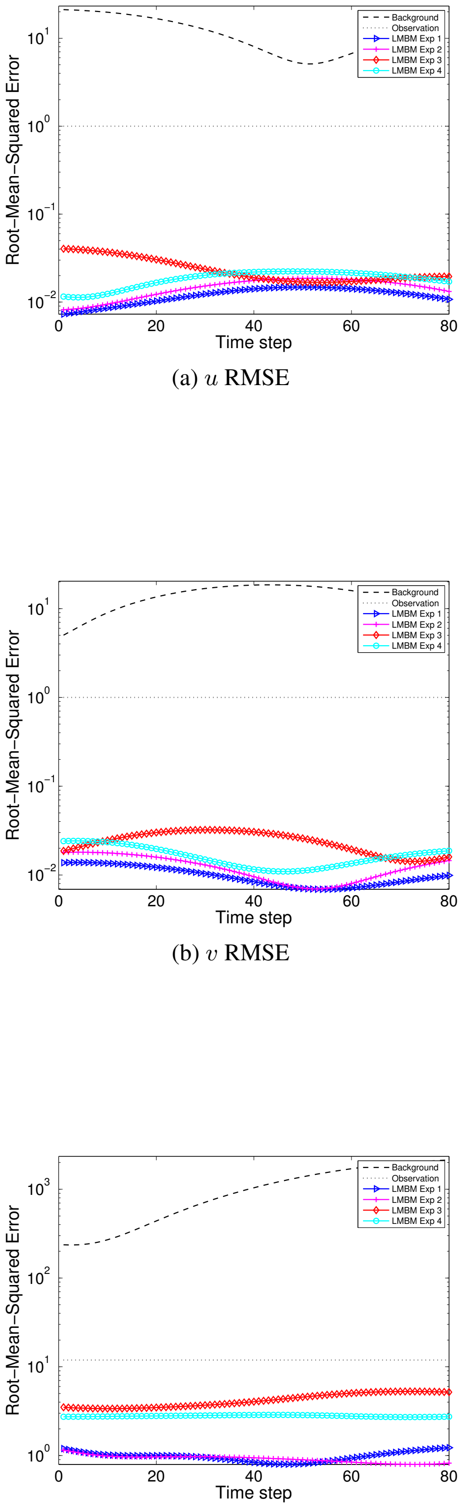


(a) $u$ RMSE

(b) $v$ RMSE

(c) $\phi$ RMSE

Figure 5.13: 3D-Var, CG-Descent

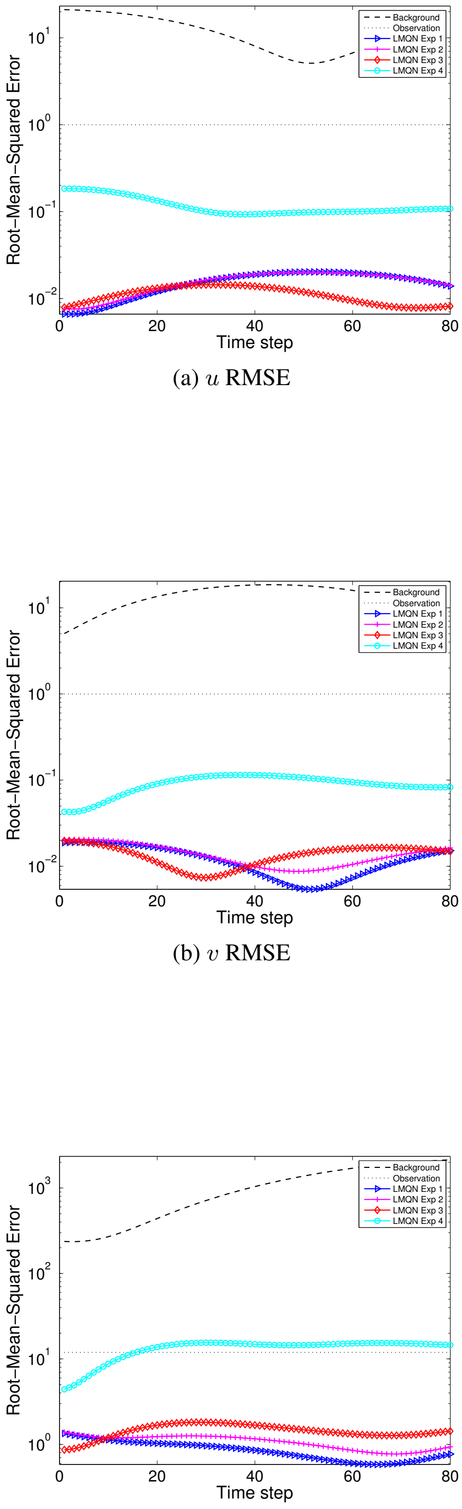


(a) $u$ RMSE

(b) $v$ RMSE

(c) $\phi$ RMSE

Figure 5.14: 4D-Var, LMBM

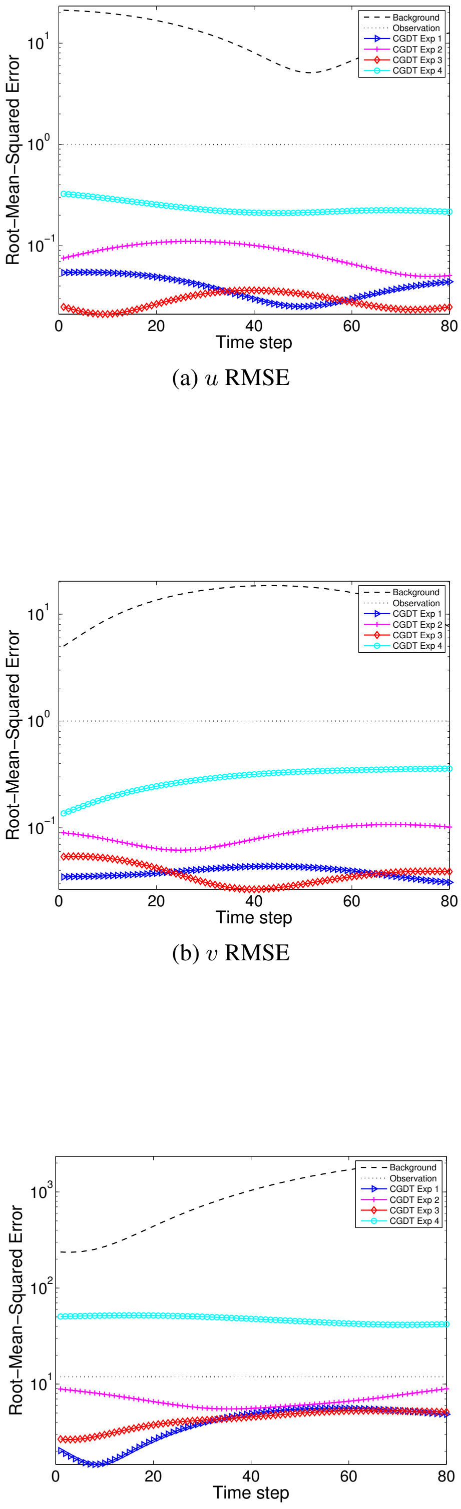


(a) $u$ RMSE

(b) $v$ RMSE

(c) $\phi$ RMSE

Figure 5.15: 4D-Var, L-BFGS

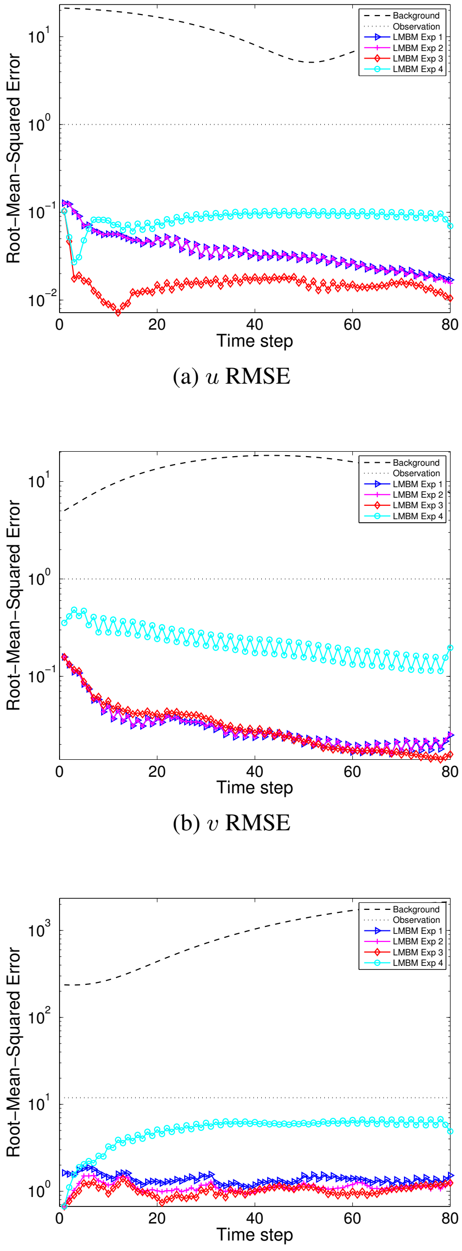


(a) $u$ RMSE

(b) $v$ RMSE

(c) $\phi$ RMSE

Figure 5.16: 4D-Var, CG-Descent

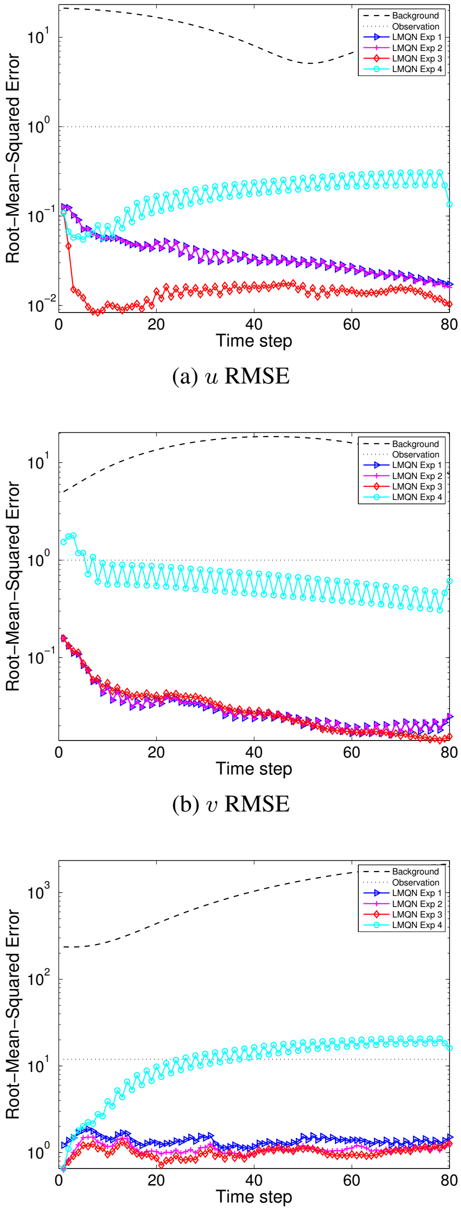


(a) $u$ RMSE

(b) $v$ RMSE

(c) $\phi$ RMSE

Figure 5.17: MLEF, LMBM

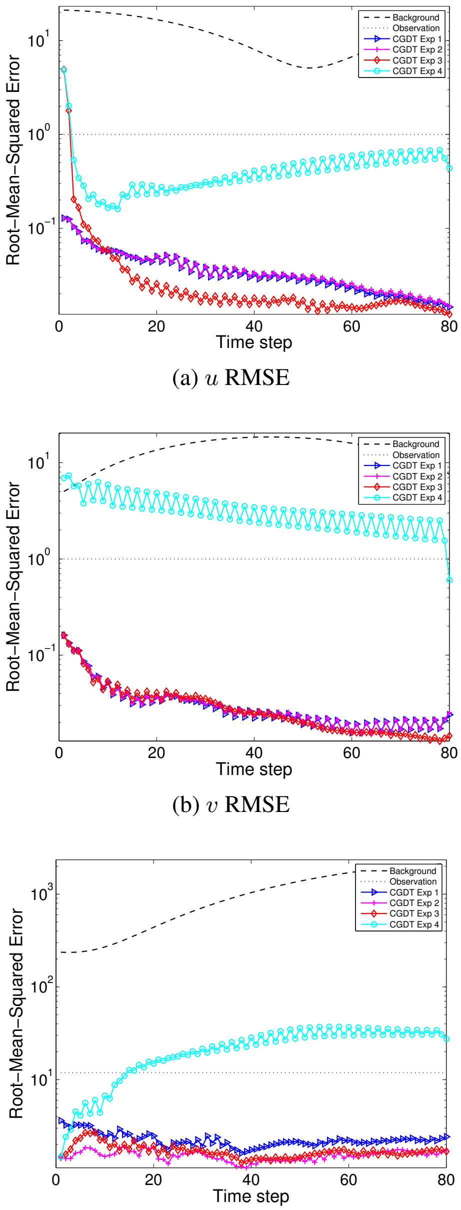


(a) $u$ RMSE

(b) $v$ RMSE

(c) $\phi$ RMSE

Figure 5.18: MLEF, L-BFGS



(a) $u$ RMSE

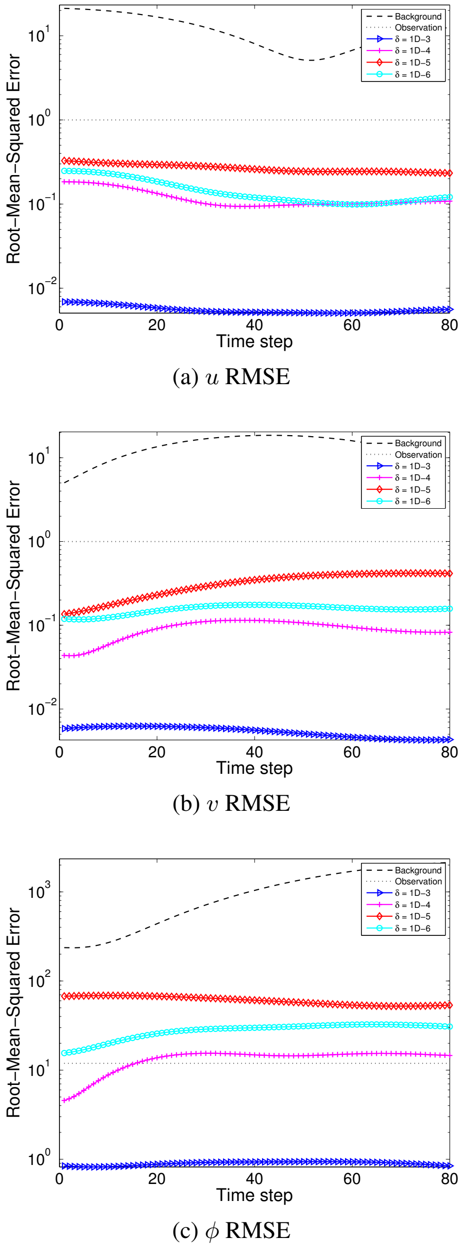
(b) $v$ RMSE

(c) $\phi$ RMSE

Figure 5.19: MLEF, CG-Descent

### 5.4.6 Varying the delta parameter

By adjusting the parameter $\delta$, we can control the Lipschitz parameter of the observation operator $\mathcal{H}_2$, thus increasing the difficulty of the non-smooth optimization. The results of varying $\delta$ for 4D-Var with L-BFGS and LMBM are shown in figures 5.20 and 5.21, respectively. These results show that 4D-Var with LMBM can successfully handle even the case where $\delta = 10^{-8}$.

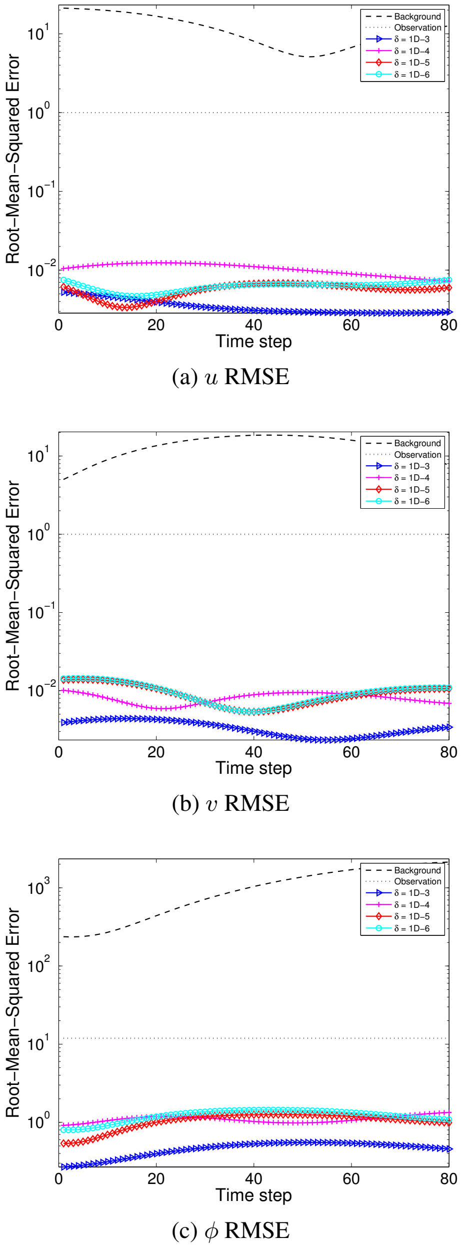


(a) $u$ RMSE

(b) $v$ RMSE

(c) $\phi$ RMSE

Figure 5.20: 4D-Var: impact of $\delta$ on RMSE vs data assimilation cycle for experiment 4, L-BFGS



(a) $u$ RMSE

(b) $v$ RMSE

(c) $\phi$ RMSE

Figure 5.21: 4D-Var: impact of $\delta$ on RMSE vs data assimilation cycle for experiment 4, LMBM

### 5.4.7 Final solution error

The error in the final solution versus the exact solution found by LMBM for experiment 4 is shown in figure 5.22 for $\delta = 10^{-4}$. The error is small and has evolved in time away from the smooth background error in figure 5.7.

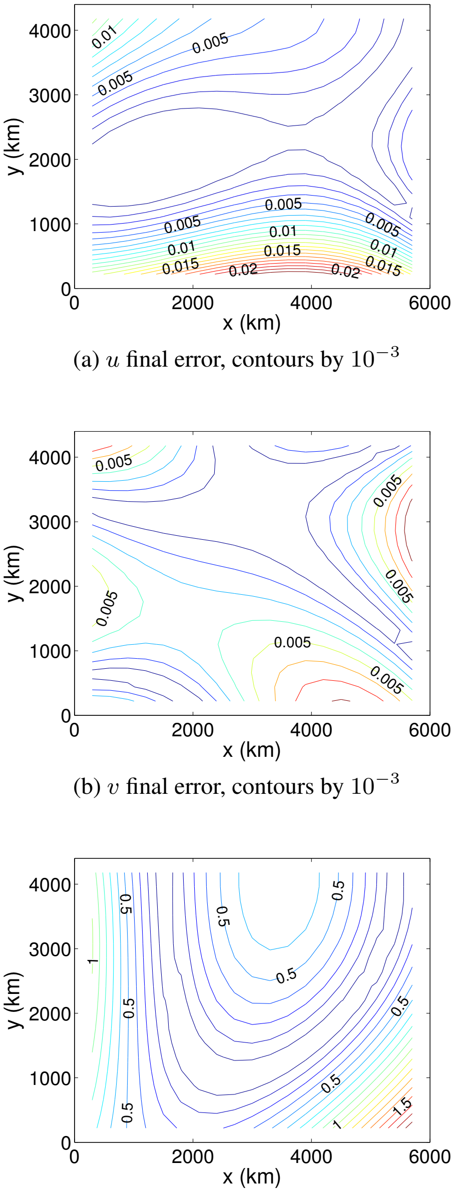
## 5.5 Conclusions

In this research, we tested the impact of non-differentiable observation operators on the data assimilation of a limited-area shallow water equations model. By simply replacing the gradient of the cost function with an arbitrary sub-gradient, all of the methods tested are able to assimilate the non-smooth observations to varying degrees of success with a smooth optimization algorithm, especially when the non-smoothness is not severe, as is the case of experiment 2. However, most methodologies encounter difficulties with the more sharply non-smooth experiments 3 and 4. This difficulty can be remedied in both MLEF and 4D-Var with the use of an algorithm specifically designed for non-smooth optimization, which in this research was the limited memory bundle algorithm (LMBM). With an unsophisticated 3D-Var approach, however, even LMBM was not able to salvage the situation.

As the timings in the above sections demonstrate, 4D-Var is the fastest approach of all the methods tested on this problem. In addition, the best RMSE results are also achieved with 4D-Var, especially when paired with LMBM. The reason for the superiority in terms of both speed and error of 4D-Var is that all observations are considered at once, and the model adjoint is used to guide the best trajectory. Only one minimization procedure takes place for the initial conditions rather than a minimization algorithm or statistical inference being performed at each time step.

The algorithms other than 4D-Var are not able to incorporate observations beyond a single timestep. As these data assimilation algorithms do not use the model as a constraint, this can lead to solutions that satisfy the minimization algorithm but are unphysical, and in particular can create shocks as the transition between two successive states is not smooth. If the optimization at each time step is successful, this leads to oscillations that dampen out. However, if these errors are large, they begin to accumulate as the background for the next iteration includes these errors, and therefore the data assimilation begins to diverge.

While the benefits of using the model adjoint in 4D-Var are large, the model adjoint itself can be very difficult to develop and maintain. Each time the code of the model evolves, the adjoint also must be updated and kept in sync; these issues can prove very challenging from a software development standpoint, and for complex models, developing, testing and maintaining these adjoints represent a major investment. Therefore, one major strength of MLEF is that, unlike 4D-Var, neither the adjoint/tangent linear model of either $\mathcal{M}$ nor $\mathcal{H}$ is required. Likewise, LETKF also does not require an adjoint, while EnKF and 3D-Var only require the adjoint of $\mathcal{H}$. The results of MLEF, approximately equal to those of 4D-Var, may be more than sufficient for solving problems with highly non-smooth observation operators, while EnKF, LETKF, and 3D-Var (even with LMBM) as implemented were not able to solve the difficult experiment 4. Thus, 4D-Var and MLEF are the only two methodologies recommended for steep discontinuities. One downside of MLEF is the additional computational time needed to compute the finite differences between $\mathcal{H}$ acting on the ensembles, as discussed in chapter 3, while a downside of 4D-Var is the need for a both the model and observation operator adjoints. The choice of methodology for this type of problem may come down to whether



(a) $u$ final error, contours by $10^{-3}$

(b) $v$ final error, contours by $10^{-3}$

(c) $\phi$ final err, contours by $10^{-1}$

Figure 5.22: Final error of the computed vs. exact solution

computational time or development time and effort is a larger concern.

With regards to the optimization algorithms, LMBM, L-BFGS, and CG-Descent were all tested. L-BFGS handled the first three non-smooth cases well for all methods, but failed on the more difficult experiment 4 for $\delta < 10^{-3}$. This translates into an observation operator with a Lipschitz constant greater than 1000. This suggests L-BFGS performs well as a non-smooth optimization algorithm when the Lipschitz constants are not extreme. CG-Descent, which has no theoretical non-smooth properties, did not fare as well, failing to converge for experiment 2–4 for 4D-Var, 3–4 for 3D-Var, and experiment 4 for MLEF. The true hero, though not without faults, was LMBM, which benefited from paying careful attention to line search and convergence issues, enabling it to perform successfully in practice far beyond the range in which L-BFGS fails for all methods but 3D-Var. The use of "null-steps" which do not progress the optimization algorithm but only add additional information about the function allow LMBM to handle such difficult cases. In addition, the use of a modified line-search and avoiding convergence criterion based on small gradients ensures the global convergence of LMBM. While it is possible to have some measure of success without paying attention to these issues, as shown the adverse effects on the data assimilation become increasingly apparent with larger Lipschitz constants. A globally convergent line search such as the null-step approach used by LMBM is thus recommended for non-smooth data assimilation with large Lipschitz constants.

While data assimilation of non-smooth observation operators using this model – with control variables on the order of $10^3$ – was successful, it remains to be seen if similar results may be obtained in the case of data assimilation using realistic non-smooth observation operators and an actual operational weather prediction model with the number of variables on the order of $10^7$. Continued research in this area is needed in order to be of practical benefit to operational weather prediction centers and other large-scale data assimilation optimal control problems. Combining the satellite assimilated 1D-Var results from the previous chapter with MLEF and 4D-Var is the most logical next step in this process.

# CHAPTER 6

# CONCLUSIONS

In this dissertation, two difficult cases in variational data assimilation were studied in order to compare and contrast the L-BFGS, CG-Descent, and LMBM optimization algorithms.

The assimilation of all-sky infrared radiances is an important outstanding problem in numerical weather prediction with a highly non-linear transition between clear and cloudy regimes that complicates the data assimilation significantly. By testing data assimilation of the RTTOV observation operator, we see that each of the optimization methods tested have their strengths and weaknesses; LMBM is able to handle difficult highly non-linear cloudy-sky cases, while L-BFGS, being better behaved numerically, exhibited better performance on cases where the variables of temperature and water vapor, which have a less highly non-linear relationship to the observation operator, were primarily modified rather than the cloud species. The performance of CG-Descent lagged behind LMBM and L-BFGS somewhat on the investigated problems.

The second problem is a mathematically challenging problem: non-smooth observation operators, i.e. observation operators that have a "kink" in them. Mathematically, this translates to a discontinuity in the first derivative, perhaps arising from different parameterizations or phase changes in the model space. This problem was demonstrated using the shallow water equations, a model appropriate to tracking the geopotential height of the earth's atmosphere along with the velocity at the free surface. Here, traditional optimization algorithms such as CG-Descent and L-BFGS are out of their original comfort-zones – the performance of these algorithms relies primarily upon the theory of sub-gradients. Because of this, an additional algorithm, known as the Limited-Memory Bundle Method, a descendant of the L-BFGS method specifically designed for non-smooth optimization, was employed. When the discontinuity in the derivative is not severe, both CG-Descent and L-BFGS perform well. However, when the Lipschitz constant – a measure of the size of the discontinuity in the derivative – becomes very large, both CG-Descent and L-BFGS fail, while LMBM is able to handle these cases quite well for all but the 3D-Var data assimilation methodology. The main issue for the smooth-based methodologies is the failure of their line search, while LMBM utilizes a guaranteed descent strategy through bundling sub-gradient information.

In this problem, we also compared and contrasted EnKF, LETKF, 3D-Var, 4D-Var, and MLEF for four non-smooth experiments. EnKF and LETKF, as expected, were not able to handle highly non-smooth cases, since they have no opportunity to iteratively revisit the tangent linear hypothesis which will be invalid precisely on the types of problems investigated; however, surprisingly, they were able to handle mild non-linearity and non-smoothness well. An unsophisticated 3D-Var implementation faired slightly better than EnKF and LETKF, but was not able to handle the most difficult non-smooth case even with LMBM. Thus the choice for highly non-smooth data assimilation problems appears to come down to 4D-Var or MLEF, with the main determining factor in the decision being whether accurate model and observation operator adjoints are available. On the problem tested here, 4D-Var exhibits superior computational performance; however, in realistic cases such as data assimilation with WRF, where a full-physics, reliable model adjoint is not reliable, MLEF is very likely a superior choice.

LMBM is not quite ready for "prime-time" as it exhibits numerical instability. However, by modifying the numerically stable and mature code of L-BFGS to include the innovations of LMBM, it is expected that improved performance can be achieved. It is thus recommended that before attempting to use LMBM operationally that additional investment into the numerical stability be made. However, even with the code as it currently stands, LMBM is a strong competitor for highly non-linear data assimilation cases such as the assimilation of all-sky IR radiances and stands alone as a realistic optimization algorithm for large scale non-smooth optimization.

### 6.0.1 Next steps

The next steps for this research include developing/testing a radiative transfer model and adjoint appropriate for all-sky data assimilation that integrates all microphysical information that is available to a modern NWP model such as WRF. This model should be able to handle both microwave and infrared observations as well as process multiple scattering in an efficient way. RTTOV version 10.2 may well meet these requirements. Additionally, the numerical stability of LMBM will be improved by porting it to the L-BFGS codebase. Finally, a practical and accurate method for determining background error covariance must be developed. Once these tasks are accomplished, the integration between 1D-Var and either MLEF and 4D-Var can be utilized to fully solve the all-sky data assimilation problem with both infrared and microwave. It is anticipated that infrared and microwave cloud data will be mutually beneficial to each other. When little to no microwave or infrared satellite data is thrown away and is instead efficiently utilized, we will have come a long way towards the task of accurately predicting the weather and climate.

Furthermore, as I have implemented most of the main modern data assimilation techniques as part of this dissertation, I would like to continue to develop an open-source framework for data assimilation, uncertainty quantification, and numerical weather prediction that I began with Dr. Emil Constantinescu at Argonne National Labs known as Æolus, named after the Greek god of the wind. Using a high-level XML framework to describe work-flows in a system- and even model-independent way, this project would no doubt be highly beneficial to promoting the adoption of wind-powered green energy and advancing the adoption of data assimilation techniques across the world.

# BIBLIOGRAPHY

[1] Nimbus program history: earth-resources research satellite program. http://atmospheres.gsfc.nasa.gov/uploads/files/Nimbus_History.pdf, October 2004.

[2] A. Aksoy, D. C. Dowell, and C. Snyder. A multicase comparative assessment of the ensemble Kalman filter for assimilation of radar observations. Part I: storm-scale analyses. *Monthly Weather Review*, 137:1805–1824, June 2009.

[3] A. B. Akvilonova, A. Y. Basharinov, A. K. Gorodetskiy, A. S. Gurvich, M. S. Krylova, B. G. Kutuza, D. T. Matveyev, and A. P. Orlov. Cloud parameters measured from the Cosmos-384 satellite. *Izvestiya, Atmospheric and Oceanic Physics*, 9:187–189, 1973.

[4] S. C. Albers, J. A. McGinley, D. L. Birkenheuer, and J. R. Smart. The Local Analysis and Prediction System (LAPS): analysis of clouds, precipitation, and temperature. *Weather and Forecasting*, 11(3):273–287, 1996.

[5] A. K. Alekseev, I. M. Navon, and J. L. Steward. Comparison of advanced large-scale minimization algorithms for the solution of inverse ill-posed problems. *Optimization Methods and Software*, 24(1):63–87, 2009.

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

[7] N. Andrei. Numerical comparison of conjugate gradient algorithms for unconstrained optimization. *Studies in Informatics and Control*, 16(4):333–352, 2007.

[8] K. Aonashi and G. S. Liu. Direct assimilation of multichannel microwave brightness temperatures and impact on mesoscale numerical weather prediction over the TOGA COARE domain. *Journal of the Meteorological Society of Japan*, 77(3):771–794, 1999.

[9] J. R. Appel. *Sensitivity calculations for conservation laws with application to discontinuous fluid flows*. Ph. D. thesis, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, 1997.

[10] H. H. Aumann, M. T. Chahine, C. Gautier, M. D. Goldberg, E. Kalnay, L. M. McMillin, H. Revercomb, P. W. Rosenkranz, W. L. Smith, D. H. Staelin, L. L. Strow, and J. Susskind. AIRS/AMSU/HSB on the Aqua mission: design, science objectives, data products, and processing systems. *IEEE Transactions on Geoscience and Remote Sensing*, 41(2):253–264, February 2003.

[11] R. T. Austin and G. L. Stephens. Retrieval of stratus cloud microphysical parameters using millimeter-wave radar and visible optical depth in preparation for CloudSat. I-Algorithm formulation. *Journal of Geophysical Research. D. Atmospheres*, 106:28, 2001.

[12] C. Bardos and O. Pironneau. Data assimilation for conservation laws. *Methods and Applications of Analysis*, 12(2):103, 2005.

[13] A. Y. Basharinov. Determination of geophysical parameters from data on thermally-induced radio emission obtained with the Kosmos 243. *Academy of Sciences USSR Earth Sciences Section Doklady*, 188:1273–1276, 1969.

[14] P. Bauer. Over-ocean rainfall retrieval from multisensor data of the Tropical Rainfall Measuring Mission. Part I: design and evaluation of inversion databases. *Journal of Atmospheric and Oceanic Technology*, 18(8):1315–1330, August 2001.

[15] P. Bauer, P. Amayenc, C. D. Kummerow, and E. A. Smith. Over-ocean rainfall retrieval from multisensor data of the Tropical Rainfall Measuring Mission. Part II: algorithm implementation. *Journal of Atmospheric and Oceanic Technology*, 18(11):1838–1855, November 2001.

[16] P. Bauer, T. Auligne, W. Bell, A. Geer, V. Guidard, S. Heilliette, M. Kazumori, M.-J. Kim, E. H.-C. Liu, A. P. McNally, B. Macpherson, K. Okamoto, R. Renshaw, and L.-P. Riishojgaard. Satellite cloud and precipitation assimilation at operational NWP centres. *Quarterly Journal of the Royal Meteorological Society*, 137(661):1934–1951, October 2011.

[17] P. Bauer, A. J. Geer, P. Lopez, and D. Salmond. Direct 4D-Var assimilation of all-sky radiances. Part I: implementation. *Quarterly Journal of the Royal Meteorological Society*, 136(652):1868–1885, October 2010.

[18] P. Bauer, P. Lopez, A. Benedetti, D. Salmond, and E. Moreau. Implementation of 1D+4D-Var assimilation of precipitation-affected microwave radiances at ECMWF. I: 1D-Var. *Quarterly Journal of the Royal Meteorological Society*, 132(620):2277–2306, October 2006.

[19] P. Bauer, P. Lopez, D. Salmond, A. Benedetti, S. Saarinen, and M. Bonazzola. Implementation of 1D+4D-Var assimilation of precipitation-affected microwave radiances at ECMWF. II: 4D-Var. *Quarterly Journal of the Royal Meteorological Society*, 132(620):2307–2332, October 2006.

[20] P. Bauer, E. Moreau, and S. Di Michele. Hydrometeor retrieval accuracy using microwave window and sounding channel observations. *Journal of Applied Meteorology*, 44:1016–1032, July 2005.

[21] G. M. Bayler, R. M. Aune, and W. H. Raymond. NWP cloud initialization using GOES sounder data and improved modeling of nonprecipitating clouds. *Monthly Weather Review*, 128:3911–3920, November 2000.

[22] R. S. Bell and O. Hammon. Sensitivity of fine-mesh rainfall and cloud forecasts to the initial specification of humidity. *Meteorological Magazine MTMGA 5*, 118(1404), 1989.

[23] A. Benedetti, G. L. Stephens, and T. Vukicevic. Variational assimilation of radar reflectivities in a cirrus model. I: model description and adjoint sensitivity studies. *Quarterly Journal of the Royal Meteorological Society*, 129(587):277–300, January 2003.

[24] A. Benedetti, G. L. Stephens, and T. Vukicevic. Variational assimilation of radar reflectivities in a cirrus model. II: optimal initialization and model bias estimation. *Quarterly Journal of the Royal Meteorological Society*, 129(587):301–319, January 2003.

[25] S. Bielli and F. Roux. Initialization of a cloud-resolving model with airborne Doppler radar observations of an oceanic tropical convective system. *Monthly Weather Review*, 127:1038–1055, June 1999.

[26] C. H. Bishop, B. J. Etherton, and S. J. Majumdar. Adaptive sampling with the ensemble transform Kalman filter. Part I: theoretical aspects. *Monthly Weather Review*, 129(3):420–436, March 2001.

[27] J. Blum, F. X. Le Dimet, and I. M. Navon. Data assimilation for geophysical fluids. In P. G. Ciarlet, R. Temam, and J. Tribbia, editors, *Computational Methods for the Atmosphere and the Oceans*, volume 14 of *Handbook of Numerical Analysis*, pages 385–442. October 2008.

[28] M. Bocquet. Ensemble Kalman filtering without the intrinsic need for inflation. *Nonlin. Processes Geophys*, 18:735–750, 2011.

[29] M. Bocquet and L. Wu. Bayesian design of control space for optimal assimilation of observations. Part II: asymptotic solutions. *Quarterly Journal of the Royal Meteorological Society*, 2011. In early view.

[30] M. Bocquet, L. Wu, and F. Chevallier. Bayesian design of control space for optimal assimilation of observations. Part I: consistent multiscale formalism. *Quarterly Journal of the Royal Meteorological Society*, 2011. In early view.

[31] J. F. Bonnans, J. C. Gilbert, C. Lemarechal, and C. A. Sagastizabal. *Numerical optimization: theoretical and practical aspects*. Springer-Verlag New York Inc, 2006.

[32] S. -A Boukabara, F.-Z. Weng, and Q.-H. Liu. Passive microwave remote sensing of extreme weather events using NOAA-18 AMSUA and MHS. *IEEE Transactions on Geoscience and Remote Sensing*, 45(7):2228–2246, July 2007.

[33] G. Burgers, P. J. van Leeuwen, and G. Evensen. Analysis scheme in the ensemble Kalman filter. *Monthly weather review*, 126(6):1719–1724, 1998.

[34] R. H. Byrd, J. Nocedal, and R. B. Schnabel. Representations of quasi-Newton matrices and their use in limited memory methods. *Mathematical Programming*, 63(1-3):129–156, January 1994.

[35] D. L. Cadet. Mean fields of precipitable water over the Indian Ocean during the 1979 summer monsoon from TIROS-N soundings and FGGE data. *Tellus B*, 35B(5):329–345, November 1983.

[36] O Caumont, V. Ducrocq, E. Wattrelot, G. Jaubert, and S. Pradier-vabre. 1D+3DVar assimilation of radar reflectivity data: a proof of concept. *Tellus A*, 62(2):173–187, March 2010.

[37] A. Caya, J. Sun, and C. Snyder. A comparison between the 4DVAR and the ensemble Kalman filter techniques for radar data assimilation. *Monthly Weather Review*, 133:3081–3094, November 2005.

[38] M. T. Chahine. Remote sounding of cloudy atmospheres. I: the single cloud layer. *Journal of Atmospheric Sciences*, 31:233–243, 1974.

[39] M. T. Chahine. Remote sounding of cloudy atmospheres. II: multiple cloud formations. *Journal of Atmospheric Sciences*, 34:744–757, 1977.

[40] M. T. Chahine, H. H. Aumann, and F. W. Taylor. Remote sounding of cloudy atmospheres. III: experimental verifications. *Journal of Atmospheric Sciences*, 34:758–765, 1977.

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

[42] F. Chevallier, P. Lopez, A. Tompkins, M. Janiskova, and E. Moreau. The capability of 4D-Var systems to assimilate cloud-affected satellite infrared radiances. *Q. J. R. Meteorol. Soc.*, 130:917–932, 2004.

[43] F. Chevallier and J. F. Mahfouf. Evaluation of the Jacobians of infrared radiation models for variational data assimilation. *Journal of Applied Meteorology*, 40(8):1445–1461, August 2001.

[44] J. C. Chiu and G. W. Petty. Bayesian retrieval of complete posterior PDFs of oceanic rain rate from microwave observations. *Journal of applied meteorology and climatology*, 45(8):1073–1095, 2006.

[45] S. E. Cohn. An introduction to estimation theory. In *Data assimilation in meteorology and oceanography: theory and practice: a collection of papers presented at the WMO Second International Symposium on Assimilation of Observations in Meteorology and Oceanography, 13-17 March 1995, Tokyo, Japan*, volume 75, page 147. Meteorological Society of Japan, 1997.

[46] P. Courtier, E. Andersson, W. Heckley, D. Vasiljevic, M. Hamrud, A. Hollingsworth, F. Rabier, M. Fisher, and J. Pailleux. The ECMWF implementation of three-dimensional variational assimilation (3D-Var). I: formulation. *Quarterly Journal of the Royal Meteorological Society*, 124(550):1783–1807, July 1998.

[47] N. E. Davidson and K. Puri. Tropical prediction using dynamical nudging, satellite-defined convective heat sources, and a cyclone bogus. *Monthly Weather Review*, 120:2501–2522, November 1992.

[48] M. Desbois, G. Seze, and G. Szejwach. Automatic classification of clouds on METEOSAT imagery: Application to high-level clouds. *Journal of Applied Meteorology*, 21:401–412, March 1982.

[49] S. Di Michele, F. S. Marzano, A. Mugnai, A. Tassa, and J. P. V. P. Baptista. Physically based statistical integration of TRMM microwave measurements for precipitation profiling. *Radio Science*, 38:16 PP., June 2003.

[50] S. Di Michele, A. Tassa, A. Mugnai, F. S. Marzano, P. Bauer, and J. P. V. P. Baptista. Bayesian algorithm for microwave-based precipitation retrieval: description and application to TMI measurements over ocean. *IEEE Transactions on Geoscience and Remote Sensing*, 43(4):778– 791, April 2005.

[51] L. J. Donner. An initialization for cumulus convection in numerical weather prediction models. *Monthly Weather Review*, 116:377–385, February 1988.

[52] D. C. Dowell, L. J. Wicker, and C. Snyder. Ensemble Kalman filter assimilation of radar observations of the 8 may 2003 Oklahoma City supercell: Influences of reflectivity observations on storm-scale analyses. *Monthly Weather Review*, 139:272–294, January 2011.

[53] V. Ducrocq, J. P. Lapore, J. L. Redelsperger, and F. Orain. Initialization of a fine-scale model for convective-system prediction: A case study. *Quarterly Journal of the Royal Meteorological Society*, 126(570):3041–3065, October 2000.

[54] E.E. Ebert, J.E. Janowiak, and C. Kidd. Comparison of near-real-time precipitation estimates from satellite observations and numerical models. *Bulletin of the American Meteorological Society*, 88(1):47–64, 2007.

[55] S. D. Eckermann, K. W. Hoppel, L. Coy, J. P. McCormack, D. E. Siskind, K. Nielsen, A. Kochenash, M. H. Stevens, C. R. Englert, W. Singer, and M. Hervig. High-altitude data assimilation system experiments for the northern summer mesosphere season of 2007. *Journal of Atmospheric and Solar-Terrestrial Physics*, 71(3-4):531–551, March 2009.

[56] R. M. Errico, P. Bauer, and J. F. Mahfouf. Issues regarding the assimilation of cloud and precipitation data. *Journal of the Atmospheric Sciences*, 64(11):3785–3798, November 2007.

[57] R. M. Errico, G. Ohring, P. Bauer, B. Ferrier, J. F. Mahfouf, J. Turk, and F. Z. Weng. Assimilation of satellite cloud and precipitation observations in numerical weather prediction models: Introduction to the JAS special collection. *Journal of the Atmospheric Sciences*, 64(11):3737–3741, November 2007.

[58] G. Evensen. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res*, 99(10):10143–10162, 1994.

[59] G. Evensen. *Data assimilation: the ensemble Kalman filter*. Springer Verlag, 2009.

[60] J. R. Eyre. Inversion of cloudy satellite sounding radiances by nonlinear optimal estimation. I: theory and simulation for TOVS. *Quarterly Journal of the Royal Meteorological Society*, 115(489):1001–1026, July 1989.

[61] J. R. Eyre. Inversion of cloudy satellite sounding radiances by nonlinear optimal estimation. II: application to TOVS data. *Quarterly Journal of the Royal Meteorological Society*, 115(489):1027–1037, July 1989.

[62] J. R. Eyre, G. A. Kelly, A. P. McNally, E. Andersson, and A. Persson. Assimilation of TOVS radiance information through one-dimensional variational analysis. *Quarterly Journal of the Royal Meteorological Society*, 119(514):1427–1463, October 1993.

[63] L. Fillion and R. Errico. Variational assimilation of precipitation data using moist convective parameterization schemes: A 1D-Var study. *Monthly Weather Review*, 125:2917–2942, November 1997.

[64] L. Fillion and J.-F. Mahfouf. Coupling of moist-convective and stratiform precipitation processes for variational data assimilation. *Monthly Weather Review*, 128:109–124, January 2000.

[65] S. J. Fletcher and M. Zupanski. A study of ensemble size and shallow water dynamics with the maximum likelihood ensemble filter. *Tellus A*, 60(2):348–360, March 2008.

[66] I. Fukumori and P. Malanotte-Rizzoli. An approximate Kalman filter for ocean data assimilation; an example with an idealized Gulf Stream model. *Journal of Geophysical Research*, 1994.

[67] R. R. Garcia. Dynamics, radiation, and photochemistry in the mesosphere: Implications for the formation of noctilucent clouds. *Journal of Geophysical Research*, 94(D12):14605–14, 1989.

[68] A. J. Geer, P. Bauer, and P. Lopez. Direct 4D-Var assimilation of all-sky radiances. Part II: assessment. *Quarterly Journal of the Royal Meteorological Society*, 136(652):1886–1905, October 2010.

[69] L. Giglio. A passive microwave technique for estimating rainfall and vertical structure information from space. Part I: algorithm description. *Journal of Applied Meteorology*, 1994.

[70] A. Grammeltvedt. A survey of finite-difference schemes for the primitive equations for a barotropic fluid. *Monthly Weather Review*, 97(5):384, 1969.

[71] M. Grecu, W. S. Olson, and E. N. Anagnostou. Retrieval of precipitation profiles from multiresolution, multifrequency active and passive microwave observations. *Journal of Applied Meteorology*, 43(4):562–575, April 2004.

[72] T. J. Greenwald, R Hertenstein, and T. Vukicevic. An All-Weather observational operator for radiance data assimilation with mesoscale forecast models. *American Meteorological Society*, 130:1882–1897, July 2002.

[73] T. J. Greenwald, T. Vukicevic, L. D. Grasso, and T. Vonder Haar. Adjoint sensitivity analysis of an observational operator for visible and infrared cloudy-sky radiance assimilation. *Quarterly Journal of the Royal Meteorological Society*, 130(597):685–705, January 2004.

[74] M. D Gunzburger. *Perspectives in flow control and optimization*. Society for Industrial Mathematics, 2003.

[75] Y-R. Guo, Y-H. Kuo, J. Dudhia, D. Parsons, and C. Rocken. Four-dimensional variational data assimilation of heterogeneous mesoscale observations for a strong convective case. *Monthly Weather Review*, 128:619–643, March 2000.

[76] J.-H. Ha, H.-W. Kim, and D.-K. Lee. Observation and numerical simulations with radar and surface data assimilation for heavy rainfall over central korea. *Advances in Atmospheric Sciences*, 28:573–590, May 2011.

[77] M. Haarala, K. Miettinen, and M. M. Makela. New limited memory bundle method for large-scale nonsmooth optimization. *Optimization Methods and Software*, 19(6):673, 2004.

[78] N. Haarala, K. Miettinen, and M. M. Makela. Globally convergent limited memory bundle method for large-scale nonsmooth optimization. *Mathematical Programming*, 109(1):181–205, April 2006.

[79] Z. Haddad, E. Im, S. L. Durden, S. Hensley, and M. H. Alves. Stochastic filtering of rain profiles using radar, surface-referenced radar, or combined radar/radiometer measurements. *J. APPL. METEOR*, 35:229—242, 1996.

[80] Z. S. Haddad, E. Im, and S. L. Durden. Optimal estimation of rain-rate profiles from single-frequency radar echoes. *Journal of Applied Meteorology*, 35(2):214–228, February 1996.

[81] Z. S. Haddad, E. A. Smith, C. D. Kummerow, T. Igushi, M. R. Farrar, S. L. Durden, M. Alves, and W. S. Olson. The TRMM "Day-1" radar/radiometer combined rain-profiling algorithm. *Journal of the Meteorological Society of Japan*, 75(4):799–809, 1997.

[82] W. W. Hager and H. Zhang. Algorithm 851: CG_DESCENT, a conjugate gradient method with guaranteed descent. *ACM Transactions on Mathematical Software (TOMS)*, 32(1):113–137, 2006.

[83] W. W. Hager and H. Zhang. A new conjugate gradient method with guaranteed descent and an efficient line search. *SIAM Journal on Optimization*, 16(1):170–192, 2006.

[84] T. M. Hamill, R. P. D'Entremont, and J. T. Buntin. A description of the air force real-time nephanalysis model. *Weather and forecasting*, 7(2):288–306, 1992.

[85] W. Heckley, G. Kelly, and M. Tiedtke. On the use of satellite-derived heating rates for data assimilation within the tropics. *Monthly weather review*, 118:1743–1757, 1990.

[86] A. K. Heidinger, C. O'Dell, R. Bennartz, and T. Greenwald. The successive-order-of-interaction radiative transfer model. Part I: model development. *Journal of applied meteorology and climatology*, 45(10):1388–1402, October 2006.

[87] S. Heilliette and L. Garand. A practical approach for the assimilation of cloudy infrared radiances and its evaluation using AIRS simulated observations. *Atmosphere-Ocean*, 45(4):211–225, 2007.

[88] J. Hocking. Bugs in cloudy IR simulations, August 2011. NWP SAF forums, http://www.nwpsaf.eu/forum/viewtopic.php?f=8&t=38&p=83#p83.

[89] J. Hocking, P. Rayer, R. Saunders, M. Matricardi, and A. Geer. RTTOV v10 users guide. Tech. rep., European Center for Medium Range Weather Forecasting, EUMETSAT, January 2011.

[90] C. Homescu and I. M. Navon. Optimal control of flow with discontinuities. *Journal of Computational Physics*, 187(2):660–682, May 2003.

[91] Y. Honda, M. Nishijima, K. Koizumi, Y. Ohta, K. Tamiya, T. Kawabata, and T. Tsuyuki. A pre-operational variational data assimilation system for a non-hydrostatic model at the Japan Meteorological Agency: Formulation and preliminary results. *Quarterly Journal of the Royal Meteorological Society*, 131(613):3465–3475, October 2005.

[92] S.-Y. Hong and J.-O.-J. Lim. The WRF single-moment 6-class microphysics scheme (WSM6). *J. Korean Meteor. Soc*, 42(2):129–151, 2006.

[93] A.-Y. Hou and S.-Q. Zhang. Assimilation of precipitation information using column model physics as a weak constraint. *Journal of the Atmospheric Sciences*, 64:3865–3878, November 2007.

[94] P. L Houtekamer and H. L Mitchell. Data assimilation using an ensemble Kalman filter technique. *Monthly Weather Review*, 126(3):796–811, March 1998.

[95] B. R. Hunt, E. J. Kostelich, and I. Szunyogh. Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter. *Physica D: Nonlinear Phenomena*, 230(1-2):112–126, June 2007.

[96] M. Janiskova, J. F. Mahfouf, and J. J. Morcrette. Preliminary studies on the variational assimilation of cloud-radiation observations. *Quarterly Journal of the Royal Meteorological Society*, 128(586):2713–2736, October 2002.

[97] M. Janiskova and J. J. Morcrette. Investigation of the sensitivity of the ECMWF radiation scheme to input parameters using the adjoint technique. *Quarterly Journal of the Royal Meteorological Society*, 131(609):1975–1995, 2005.

[98] M. Janiskova, J. N. Thepaut, and J. F. Geleyn. Simplified and regular physical parameterizations for incremental four-dimensional variational assimilation. *Monthly Weather Review*, 127:26–45, January 1999.

[99] M. Jardak, I. Navon, and M. Zupanski. Comparison of ensemble data assimilation for the shallow water equations model in the presence of nonlinear observation operators. *Quarterly Journal of the Royal Meteorological Society*, 2011. Accepted with revisions.

[100] A. H. Jazwinski. *Stochastic processes and filtering theory*. Academic Press, 1970.

[101] Zhu Jiang, Masafumi Kamachi, and Zhou Guangqing. Nonsmooth optimization approaches to VDA of models with on/off parameterizations: Theoretical issues. *Advances in Atmospheric Sciences*, 19(3):405–424, May 2002.

[102] R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45, 1960.

[103] E. Kalnay. *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge Univ. Press, Cambridge, electronic version edition, 2003.

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

[105] N. Karmitsa. LMBM - FORTRAN subroutines for large-scale nonsmooth minimization: User's manual. TUCS Technical Report 856, Turku Centre for Computer Science, Turku, 2007.

[106] N. Karmitsa, A. Bagirov, and M. M. Makela. Empirical and theoretical comparisons of several nonsmooth minimization methods and software. Technical Report 959, Turku Centre for Computer Science, Turku, October 2009.

[107] A. Kasahara, J.-I. Tsutsui, and H. Hirakuchi. Inversion methods of three cumulus parameterizations for diabatic initialization of a tropical cyclone model. *Monthly Weather Review*, 124:2304–2321, October 1996.

[108] T. Kawabata, T. Kuroda, H. Seko, and K. Saito. A cloud-resolving 4DVAR assimilation experiment for a local heavy rainfall event in the Tokyo metropolitan area. *Monthly Weather Review*, 139:1911–1931, June 2011.

[109] T. Kawabata, H. Seko, K. Saito, T. Kuroda, K. Tamiya, T. Tsuyuki, Y. Honda, and T. Wakazuki. An assimilation and forecasting experiment of the Nerima heavy rainfall with a cloud-resolving nonhydrostatic 4-Dimensional variational data assimilation system. *Journal of the Meteorological Society of Japan*, 85(3):255–276, 2007.

[110] K. C. Kiwiel. *Methods of descent for nondifferentiable optimization*. Springer-Verlag Berlin, 1985.

[111] K.C. Kiwiel. Proximity control in bundle methods for convex nondifferentiable minimization. *Mathematical Programming*, 46(1):105–122, 1990.

[112] S. E. Koch, A. Aksakal, and J. T. McQueen. The influence of mesoscale humidity and evapotranspiration fields on a model forecast of a cold-frontal squall line. *Monthly Weather Review*, 125:384–409, March 1997.

[113] K. Koizumi, Y. Ishikawa, and T. Tsuyuki. Assimilation of precipitation data to the JMA mesoscale model with a four-dimensional variational method and its impact on precipitation forecasts. *SOLA*, 1:45–48, 2005.

[114] T. N. Krishnamurti, H. S. Bedi, W. Heckley, and K. Ingles. Reduction of the spinup time for evaporation and precipitation in a spectral model. *Monthly Weather Review*, 116:907–920, April 1988.

[115] T. N Krishnamurti, H. S Bedi, and K. Ingles. Physical initialization using SSM/I rain rates. *Tellus A*, 45(4):247–269, August 1993.

[116] T. N Krishnamurti, J.-S. Xue, H. S Bedi, K. Ingles, and D. Oosterhof. Physical initialization for numerical weather prediction over the tropics. *Tellus B*, 43(4):53–81, September 1991.

[117] T.N. Krishnamurti, K. Ingles, S. Cocke, T. Kitade, and R. Pasch. Details of low latitude medium range numerical weather prediction using a global spectral model. II: effects of orography and physical initialization. *Journal of the Meteorological Society of Japan*, 62(4):613–649, 1984.

[118] C. Kummerow and L. Giglio. A passive microwave technique for estimating rainfall and vertical structure information from space. Part II: applications to SSM/I data. *Journal of Applied Meteorology*, 33(1):19–34, 1994.

[119] C. Kummerow, Y. Hong, W. S. Olson, S. Yang, R. F. Adler, J. McCollum, R. Ferraro, G. Petty, D.-B. Shin, and T. T. Wilheit. The evolution of the goddard profiling algorithm (GPROF) for rainfall estimation from passive microwave sensors. *Journal of Applied Meteorology*, 40:1801–1820, November 2001.

[120] C. Kummerow, W. S Olson, and L. Giglio. A simplified scheme for obtaining precipitation and vertical hydrometeor profiles from passive microwave sensors. *IEEE Transactions on Geoscience and Remote Sensing*, 34(5):1213–1232, September 1996.

[121] Y-H. Kuo, X. Zou, and W. Huang. The impact of global positioning system data on the prediction of an extratropical cyclone: an observing system simulation experiment. *Dynamics of Atmospheres and Oceans*, 27(1-4):439–470, January 1998.

[122] L. Lavanant, N. Fourrie, A. Gambacorta, G. Grieco, S. Heilliette, F. I. Hilton, M. J. Kim, A. P McNally, H. Nishihata, E. G. Pavelin, and F. Rabier. Comparison of cloud products within IASI footprints for the assimilation of cloudy radiances. *Quarterly Journal of the Royal Meteorological Society*, 2011. In early view.

[123] F. X. Le Dimet. Une etude generale d'analyse objective variationnelle des champs meteorologiques. Technical Report 28, Universite de Clermond, Aubiere, France, 1980.

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

[125] T. S. L'Ecuyer and G. L. Stephens. An estimation-based precipitation retrieval algorithm for attenuating radars. *Journal of Applied Meteorology*, 41(3):272–285, March 2002.

[126] J.-H. Lee, H.-H. Lee, Y.-H. Choi, H.-W. Kim, and D.-K. Lee. Radar data assimilation for the simulation of mesoscale convective systems. *Advances in Atmospheric Sciences*, 27:1025–1042, August 2010.

[127] C. Lemarechal. An extension of Davidon methods to non differentiable problems. *Nondifferentiable optimization*, pages 95–109, 1975.

[128] G. Levy, M. Coon, G. Nguyen, and D. Sulsky. Physically-based data assimilation. *Geoscientific Model Development Discussions*, 3(2):517–540, April 2010.

[129] A. S. Lewis and M. L. Overton. Behavior of BFGS with an exact line search on nonsmooth examples. technical report. Technical report, Optimization Online, 2008. Submitted to SIAM J. Optimization.

[130] A. S. Lewis and M. L. Overton. Nonsmooth optimization via BFGS. Technical report, Optimization Online, 2008. Submitted to SIAM J. Optimization.

[131] J. Li, H.-L. Huang, C.-Y. Liu, P. Yang, T. J. Schmit, H.-L. Wei, E. Weisz, L. Guan, and W. P. Menzel. Retrieval of cloud microphysical properties from MODIS and AIRS. *Journal of Applied Meteorology*, 44(10):1526–1543, October 2005.

[132] J. Li and H. Liu. Improved hurricane track and intensity forecast using single field-of-view advanced IR sounding measurements. *Geophysical Research Letters*, 36:4 PP., June 2009.

[133] J. Li, W. P. Menzel, and A. J. Schreiner. Variational retrieval of cloud parameters from GOES sounder longwave cloudy radiance measurements. *Journal of Applied Meteorology*, 40(3):312–330, March 2001.

[134] X.-L. Li and J. R. Mecikalski. Assimilation of the dual-polarization Doppler radar data for a convective storm with a warm-rain radar forward operator. *Journal of Geophysical Research*, 115:16 PP., August 2010.

[135] Y. Lin, P. S. Ray, and K. W. Johnson. Initialization of a modeled convective storm using Doppler radar-derived fields. *Monthly Weather Review*, 121:2757–2775, October 1993.

[136] D. C. Liu and J. Nocedal. On the limited memory BFGS method for large scale optimization. *Math. Program.*, 45(3):503–528, 1989.

[137] P. Lopez. Cloud and precipitation parameterizations in modeling and variational data assimilation: A review. *Journal of the Atmospheric Sciences*, 64(11):3766–3784, November 2007.

[138] P. Lopez. Direct 4D-Var assimilation of NCEP stage IV radar and gauge precipitation data at ECMWF. *Monthly Weather Review*, 139:2098–2116, July 2011.

[139] P. Lopez and P. Bauer. "1D+4DVAR" assimilation of NCEP stage-IV radar and gauge hourly precipitation data at ECMWF. *Monthly weather review*, 135(7):2506–2524, 2007.

[140] A. C. Lorenc. Analysis methods for numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society*, 112(474):1177–1194, October 1986.

[141] B. Macpherson, B. J. Wright, W. H. Hand, and A. J. Maycock. The impact of MOPS moisture data in the U.K. meteorological office mesoscale data assimilation scheme. *Monthly Weather Review*, 124:1746–1766, August 1996.

[142] J. Mailhot, C. Chouinard, R. Benoit, M. Roch, G. Verner, J. Cote, and J. Pudykiewicz. Numerical forecasting of winter coastal storms during CASP: evaluation of the regional finite-element model. *Atmosphere-Ocean*, 27(1):24–58, 1989.

[143] M. M. Makela. Survey of bundle methods for nonsmooth optimization. *Optimization methods and software*, 17(1):1–29, 2002.

[144] M. M. Makela and P. Neittaanmaki. *Nonsmooth optimization: analysis and algorithms with applications to optimal control*. World Scientific, Singapore, 1992.

[145] J. Manobianco, S. Koch, V. M. Karyampudi, and A. J. Negri. The impact of assimilating Satellite-Derived precipitation rates on numerical simulations of the ERICA IOP 4 cyclone. *Monthly Weather Review*, 122:341–365, February 1994.

[146] V. Marecal and J. F. Mahfouf. Variational retrieval of temperature and humidity profiles from TRMM precipitation data. *Monthly Weather Review*, 128:3853–3866, November 2000.

[147] F. S Marzano, A. Mugnai, G. Panegrossi, N. Pierdicca, E. A Smith, and J. Turk. Bayesian estimation of precipitating cloud parameters from combined measurements of spaceborne microwave radiometer and radar. *IEEE Transactions on Geoscience and Remote Sensing*, 37(1):596–613, January 1999.

[148] H. Masunaga and C. D. Kummerow. Combined radar and radiometer analysis of precipitation profiles for a parametric retrieval algorithm. *Journal of Atmospheric and Oceanic Technology*, 22(7):909–929, July 2005.

[149] M. Matricardi. The inclusion of aerosols and clouds in RTIASI, the ECMWF fast radiative transfer model for the infrared atmospheric sounding interferometer. Technical Report 474, ECMWF, July 2005.

[150] A. P McNally. A note on the occurrence of cloud in meteorologically sensitive areas and the implications for advanced infrared sounders. *Quarterly Journal of the Royal Meteorological Society*, 128(585):2551–2556, October 2002.

[151] A. P McNally. The direct assimilation of cloud-affected satellite infrared radiances in the ECMWF 4D-Var. *Quarterly Journal of the Royal Meteorological Society*, 135(642):1214–1229, July 2009.

[152] I. Meirold-Mautner, C. Prigent, E. Defer, J. R. Pardo, J.-P. Chaboureau, J.-P. Pinty, M. Mech, and S. Crewell. Radiative transfer simulations using mesoscale cloud model outputs: Comparisons with passive microwave and infrared satellite observations for midlatitudes. *Journal of the Atmospheric Sciences*, 64:1550–1568, May 2007.

[153] W. P. Menzel, W. L. Smith, and T. R. Stewart. Improved cloud motion wind vector and altitude assignment using VAS. *Journal of Climate and Applied Meteorology*, 22:377–384, March 1983.

[154] G. A. Mills. The sensitivity of a numerical prognosis to moisture detail in the initial state. *Australian Meteorological Magazine*, 31(2), 1983.

[155] G. A. Mills and N. E. Davidson. Tropospheric moisture profiles from digital IR satellite imagery - system description and analysis/forecast impact. *Australian Meteorological Magazine*, 35:109–118, 1987.

[156] T. Miyoshi, S. Yamane, and T. Enomoto. Localizing the error covariance by physical distances within a local ensemble transform Kalman filter (LETKF). *SOLA*, 3(0):89–92, 2007.

[157] T. Montmerle, A. Caya, and I. Zawadzki. Simulation of a midlatitude convective storm initialized with bistatic Doppler radar data. *Monthly Weather Review*, 129:1949–1967, August 2001.

[158] A. Mugnai, E. A. Smith, and G. J. Tripoli. Foundations for statistical/physical precipitation retrieval from passive microwave satellite measurements. Part II: emission-source and generalized weighting-function properties of a time-dependent cloud-radiation model. *Journal of Applied Meteorology*, 32(1):17–39, January 1993.

[159] I. M. Navon. Data assimilation for numerical weather prediction. In S.K. Park and L. Xu, editors, *Data assimilation for atmospheric, oceanic and hydrologic applications*, pages 21–81. Springer Verlag, 2009.

[160] I. M. Navon, X. Zou, J. Derber, and J. Sela. Variational data assimilation with an adiabatic version of the NMC spectral model. *Monthly Weather Review*, 120(7):1433–1446, July 1992.

[161] J. Nocedal. Updating quasi-Newton matrices with limited storage. *Mathematics of computation*, 35(151):773–782, 1980.

[162] J. Nocedal and S. Wright. *Numerical Optimization*. Springer, 2nd edition, July 2006.

[163] D. C. Norquist. Alternative forms of humidity information in global data assimilation. *Monthly Weather Review*, 116:452–471, February 1988.

[164] C. W. O'Dell, A. K. Heidinger, T. Greenwald, P. Bauer, and R. Bennartz. The successive-order-of-interaction radiative transfer model. Part II: model performance and applications. *Journal of Applied Meteorology and Climatology*, 45(10):1403–1413, October 2006.

[165] W. S. Olson, C. D. Kummerow, G. M. Heymsfield, and L. Giglio. A method for combined passive-active microwave retrievals of cloud and precipitation profiles. *Journal of Applied Meteorology*, 35(10):1763–1789, 1996.

[166] E. Olvovsky. *Novel gradient-type Optimization Algorithms for Extremely Large-Scale Nonsmooth Convex Optimization*. Ph. D. thesis, University of Technion, Israel, January 2005.

[167] T. N. Palmer, R. Gelaro, J. Barkmeijer, and R. Buizza. Singular vectors, metrics, and adaptive observations. *Journal of the Atmospheric Sciences*, 55:633–653, February 1998.

[168] Thomas Pangaud, Nadia Fourrie, Vincent Guidard, Mohamed Dahoui, and Florence Rabier. Assimilation of AIRS radiances affected by mid- to Low-Level clouds. *Monthly Weather Review*, 137(12):4276–4292, December 2009.

[169] E. G Pavelin, S. J English, and J. R Eyre. The assimilation of cloud-affected infrared satellite radiances for numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society*, 134(632):737–749, April 2008.

[170] D. J. Perkey. A description and preliminary results from a fine-mesh model for forecasting quantitative precipitation. *Monthly Weather Review*, 104:1513–1526, December 1976.

[171] R. Polkinghorne and T. Vukicevic. Data assimilation of cloud-affected radiances in a cloud-resolving model. *Monthly Weather Review*, 139:755–773, March 2011.

[172] K. Puri and M. J. Miller. The use of satellite data in the specification of convective heating for diabatic initialization and moisture adjustment in numerical weather prediction models. *Monthly weather review*, 118(1):67–93, 1990.

[173] M. Rajeevan, A. Kesarkar, S. B. Thampi, T. N. Rao, B. Radhakrishna, and M. Rajasekhar. Sensitivity of WRF cloud microphysics to simulations of a severe thunderstorm event over southeast India. In *Annales Geophysicae*, volume 28, pages 603–619, 2010.

[174] W. H. Raymond, W. S. Olson, and G. Callan. Diabatic forcing and initialization with assimilation of cloud water and rainwater in a forecast model. *Monthly Weather Review*, 123:366–382, February 1995.

[175] C. D. Rodgers. *Inverse methods for atmospheric sounding: Theory and practice*, volume 2 of *Series on atmospheric, oceanic and planetary physics*. World Scientific, Singapore, January 2000.

[176] R. R. Rogers. *A Short Course in Cloud Physics*. Butterworth-Heinemann, Waltham, Massachusetts, USA, 3rd edition, 1986.

[177] L. S. Rothman, I. E. Gordon, A. Barbe, D. C Benner, P. F. Bernath, M. Birk, V. Boudon, L. R. Brown, A. Campargue, and J. P Champion. The HITRAN 2008 molecular spectroscopic database. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 110(9-10):533–572, 2009.

[178] A. Routray, U. C. Mohanty, S. R. H. Rizvi, D. Niyogi, K. K. Osuri, and D. Pradhan. Impact of Doppler weather radar data on numerical forecast of Indian monsoon depressions. *Quarterly Journal of the Royal Meteorological Society*, 136(652):1836–1850, October 2010.

[179] R. Saunders, M. Matricardi, A. Geer, P. Rayer, O. Embury, and C. Merchant. RTTOV9 science and validation plan. Tech. rep., ECMWF, EUMETSAT, March 2008.

[180] H. Schramm and J. Zowe. A version of the bundle idea for minimizing a nonsmooth function: Conceptual idea, convergence analysis, numerical results. *SIAM Journal on Optimization*, 2:121, 1992.

[181] A. J. Schreiner, K. I. Strabala, D. A. Unger, W. P. Menzel, G. P. Ellrod, and J. L. Pellet. A comparison of ground and satellite observations of cloud cover. *Bulletin of the American Meteorological Society*, 74:1851–1861, October 1993.

[182] C. J. Seaman, M. Sengupta, and T. H. Vonder Haar. Mesoscale satellite data assimilation: impact of cloud-affected infrared observations on a cloud-free initial model state. *Tellus A*, 62(3):298–318, May 2010.

[183] N. Z. Shor, K. C. Kiwiel, and A. Ruszcaynski. Minimization methods for non-differentiable functions. 1985.

[184] Randhir Singh, C. M. Kishtawal, and P. K. Pal. Impact of ATOVS radiance on the analysis and forecasts of a mesoscale model over the indian region during the 2008 summer monsoon. *Pure and Applied Geophysics*, July 2011.

[185] A. Skajaa. *Limited Memory BFGS for Nonsmooth Optimization*. M.S. thesis, New York University, Courant Institute of Mathematical Science, January 2010.

[186] W. Skamarock, J. B. Klemp, J. Dudhia, D. O. Gill, D. M. Barker, M. G. Duda, X. Huang, W. Wang, and J. G. Powers. A description of the advanced research WRF version 3. Tech. Rep. Tech Notes-475+ STR, NCAR, 2008.

[187] E. A. Smith, H. J. Cooper, X.-W. Xiang, A. Mugnai, and G. J. Tripoli. Foundations for statistical-physical precipitation retrieval from passive microwave satellite measurements. Part I: brightness-temperature properties of a time-dependent cloud-radiation model. *Journal of Applied Meteorology*, 31(6):506–531, June 1992.

[188] W. L. Smith and C. M. R. Platt. Comparison of satellite-deduced cloud heights with indications from radiosonde and ground-based laser measurements. *Journal of Applied Meteorology*, 17:1796–1802, 1978.

[189] W. L. Smith, H. M. Woolf, P. G. Abel, C. M. Hayden, M. Chalfant, and N. Grody. NIMBUS-5 sounder data processing system: measurement characteristics and data reduction procedures. Tech. Memo. NESS 57, NOAA, 1974.

[190] W. L. Smith, D. K. Zhou, H. L. Huang, H. E. Revercomb, A. M. Larar, and C. Barnett. Ultra high spectral satellite remote sounding - results from aircraft and satellite measurements. In *14th International TOVS Study Conference, NASA, Beijing*, 2005.

[191] C. Snyder and F.-Q. Zhang. Assimilation of simulated Doppler radar observations with an ensemble Kalman filter. *Monthly Weather Review*, 131:1663–1677, August 2003.

[192] G. L. Stephens. *Remote Sensing of the Lower Atmosphere: An Introduction.* Oxford University Press, USA, illustrated edition, April 1994.

[193] G. L. Stephens and C. D. Kummerow. The remote sensing of clouds and precipitation from space: A review. *Journal of the Atmospheric Sciences*, 64(11):3742–3765, November 2007.

[194] J. L Steward, I. M Navon, M. Zupanski, and N. Karmitsa. Impact of non-smooth observation operators on variational and sequential data assimilation for a limited-area shallow-water equation model. *Quarterly Journal of the Royal Meteorological Society*, 2011. Early view.

[195] J. L. Steward, I. M. Navon, M. Zupanski, and N. Karmitsa. Non-smooth optimization in the direct 1D-Var assimilation of cloudy-sky infrared radiances using RTTOV version 10. 2011. In preparation.

[196] A. M. Stuart. Inverse problems: a bayesian perspective. *Acta Numerica*, 19:451–559, 2010.

[197] S. Sugimoto, N. A. Crook, J.-Z. Sun, Q.-N. Xiao, and D. M. Barker. An examination of WRF 3DVAR radar data assimilation on its capability in retrieving unobserved variables and forecasting precipitation through observing system simulation experiments. *Monthly Weather Review*, 137:4011–4029, November 2009.

[198] J.-Z. Sun and N. A. Crook. Dynamical and microphysical retrieval from Doppler radar observations using a cloud model and its adjoint. Part I: model development and simulated data experiments. *Journal of the Atmospheric Sciences*, 54:1642–1661, June 1997.

[199] J.-Z. Sun and N. A. Crook. Dynamical and microphysical retrieval from Doppler radar observations using a cloud model and its adjoint. Part II: retrieval experiments of an observed Florida convective storm. *Journal of the Atmospheric Sciences*, 55:835–852, March 1998.

[200] M. D. E Szyndel, A. D Collard, and J. R Eyre. A simulation study of 1D variational cloud retrieval with infrared satellite data from multiple fields of view. *Quarterly Journal of the Royal Meteorological Society*, 130(599):1489–1503, April 2004.

[201] O. Talagrand and P. Courtier. Variational assimilation of meteorological observations with the adjoint vorticity equation. I: theory. *Quarterly Journal of the Royal Meteorological Society*, 113(478):1311–1328, 1987.

[202] M. K. Tippett, J. L. Anderson, C. H. Bishop, T. M. Hamill, and J. S. Whitaker. Ensemble square root filters. *Monthly Weather Review*, 131(7):1485–1490, July 2003.

[203] A. M Tompkins and M. Janiskova. A cloud scheme for data assimilation: Description and initial tests. *Quarterly Journal of the Royal Meteorological Society*, 130(602):2495–2517, October 2004.

[204] M.-J. Tong and M. Xue. Ensemble Kalman filter assimilation of [Doppler] radar data with a compressible nonhydrostatic model: OSS experiments. *Monthly Weather Review*, 133:1789–1807, July 2005.

[205] R. E. Treadon, H.-L. Pan, W.-S. Wu, Y. Lin, W. S. Olson, and R. J. Kuligowski. Global and regional moisture analyses at NCEP. In *Proc. ECMWF/GEWEX Workshop on Humidity Analysis,*, pages 33–47, Reading, United Kingdom, European Centre for Medium-Range Weather Forecasts,, 2002.

[206] T. Tsuyuki. Variational data assimilation in the tropics using precipitation data. Part II: 3D model. *Monthly Weather Review*, 124:2545–2561, November 1996.

[207] T. Tsuyuki, K. Koizumi, and Y. Ishikawa. The JMA mesoscale 4D-Var system and assimilation of precipitation and moisture data. In *Proc. ECMWF/GEWEX Workshop on Humidity Analysis*, pages 59–67. European Centre for Medium-Range Weather Forecasts, 2002.

[208] O. M. Turpeinen, L. Garand, R. Benoit, and M. Roch. Diabatic initialization of the canadian regional finite-element (RFE) model using satellite data. Part i: Methodology and application to a winter storm. *Monthly Weather Review*, 118:1381–1395, July 1990.

[209] B. Uzunoglu, S. J. Fletcher, M. Zupanski, and I. M. Navon. Adaptive ensemble reduction and inflation. *Quarterly Journal of the Royal Meteorological Society*, 133(626):1281–1294, 2007.

[210] P. J Van Leeuwen and G. Evensen. Data assimilation and inverse methods in terms of a probabilistic formulation. *Monthly Weather Review*, 124:2898–2913, 1996.

[211] J. Verlinde and W. R. Cotton. Fitting microphysical observations of nonsteady convective clouds to a numerical model: An application of the adjoint technique of data assimilation to a kinematic model. *Monthly weather review*, 121(10):2776–2793, 1993.

[212] J. Vlcek and L. Luksan. Globally convergent variable metric method for nonconvex nondifferentiable unconstrained minimization. *Journal of Optimization Theory and Applications*, 111(2):407–430, October 2001.

[213] T. Vukicevic, T. Greenwald, M. Zupanski, D. Zupanski, T. Vonder Haar, and A. S. Jones. Mesoscale cloud state estimation from visible and infrared satellite radiances. *Monthly Weather Review*, 132(12):3066–3077, December 2004.

[214] T. Vukicevic, M. Sengupta, A. S. Jones, and T. V Haar. Cloud-resolving satellite data assimilation: Information content of IR window observations and uncertainties in estimation. *Journal of the Atmospheric Sciences*, 63(3):901–919, 2006.

[215] Z. Wang, I. M. Navon, F. X. Dimet, and X. Zou. The second order adjoint analysis: Theory and applications. *Meteorology and Atmospheric Physics*, 50(1-3):3–20, 1992.

[216] H.-L. Wei, P. Yang, J. Li, B. A. Baum, H.-L. Huang, S. Platnick, Y.-X. Hu, and L. Strow. Retrieval of semitransparent ice cloud optical thickness from atmospheric infrared sounder (AIRS) measurements. *IEEE Transactions on Geoscience and Remote Sensing*, 42(10):2254–2267, October 2004.

[217] E. Weisz, J. Li, J.-L. Li, D. K. Zhou, H.-L. Huang, M. D. Goldberg, and P. Yang. Cloudy sounding and cloud-top height retrieval from AIRS alone single field-of-view radiance measurements. *Geophysical Research Letters*, 34:5 PP., June 2007.

[218] F. Z. Weng. Advances in radiative transfer modeling in support of satellite data assimilation. *Journal of the Atmospheric Sciences*, 64(11):3799–3807, November 2007.

[219] F. Z. Weng and Q. H. Liu. Satellite data assimilation in numerical weather prediction models. Part I: forward radiative transfer and Jacobian modeling in cloudy atmospheres. *Journal of the Atmospheric Sciences*, 60(21):2633–2646, 2003.

[220] F.-Z. Weng, T. Zhu, and B.-H. Yan. Satellite data assimilation in numerical weather prediction models. Part II: uses of rain-affected radiances from microwave observations for hurricane vortex analysis. *American Meteorological Society*, 64(11):3910–3925, November 2007.

[221] S. S. Weygandt, A. Shapiro, and K. K. Droegemeier. Retrieval of model initial fields from single-Doppler observations of a supercell thunderstorm. Part I: Single-Doppler velocity retrieval. *Monthly weather review*, 130(3):433–453, 2002.

[222] S. S. Weygandt, A. Shapiro, and K. K. Droegemeier. Retrieval of model initial fields from single-Doppler observations of a supercell thunderstorm. Part II: thermodynamic retrieval and numerical prediction. *Monthly weather review*, 130(3):454–476, 2002.

[223] Jeffrey S. Whitaker and Thomas M. Hamill. Ensemble data assimilation without perturbed observations. *Monthly Weather Review*, 130(7):1913–1924, July 2002.

[224] M. Wiedner, C. Prigent, J.R. Pardo, O. Nuissier, J.P. Chaboureau, J.P. Pinty, and P. Mascart. Modeling of passive microwave responses in convective situations using output from mesoscale models: Comparison with TRMM/TMI satellite observations. *J. Geophys. Res*, 109(D6):1–13, 2004.

[225] S. W. Wolcott and T. T. Warner. A moisture analysis procedure utilizing surface and satellite data. *Monthly Weather Review*, 109:1989–1998, September 1981.

[226] P. Wolfe. A method of conjugate subgradients for minimizing nondifferentiable functions. *Nondifferentiable optimization*, pages 145–173, 1975.

[227] B. Wu, J. Verlinde, and J.-Z. Sun. Dynamical and microphysical retrievals from Doppler radar observations of a deep convective cloud. *Journal of the Atmospheric Sciences*, 57:262–283, January 2000.

[228] X.-Q. Wu and W. L. Smith. Assimilation of ERBE data with a nonlinear programming technique to improve Cloud-Cover diagnosis. *Monthly Weather Review*, 120:2009–2024, September 1992.

[229] Q. Xiao, X. Zou, and Y-H. Kuo. Incorporating the SSM/I-derived precipitable water and rainfall rate into a numerical model: A case study for the Erica IOP-4 cyclone. *Monthly Weather Review*, 128:87–108, January 2000.

[230] Q.-N. Xiao, Y.-H. Kuo, J.-Z. Sun, W.-C. Lee, D. M. Barker, and E. Lim. An approach of radar reflectivity data assimilation and its assessment with the inland QPF of typhoon Rusa (2002) at landfall. *Journal of Applied Meteorology and Climatology*, 46:14–22, January 2007.

[231] Q. Xu. Generalized adjoint for physical processes with parameterized discontinuities. Part I: basic issues and heuristic examples. *Journal of the Atmospheric Sciences*, 53:1123–1142, April 1996.

[232] S. C. Yang, E. Kalnay, and B. Hunt. Weight interpolation for efficient data assimilation with the local ensemble transform Kalman filter. *Quarterly Journal of the Royal Meteorological Society*, 135(638):251–262, 2009.

[233] J. Zhang. *Moisture and diabatic initialization based on radar and satellite observations*. Ph. D. thesis, University of Oklahoma, Norman, OK, 1999.

[234] J. Zhang, F. Carr, and K. Brewster. ADAS cloud analysis. In *Preprints, 12th Conf. on Numerical Weather Prediction, Phoenix, AZ, Amer. Meteor. Soc*, pages 185–188, 1998.

[235] S. Zhang, X. Zou, J. E. Ahlquist, I. M. Navon, and J. G. Sela. Use of differentiable and non-differentiable optimization algorithms for variational data assimilation with discontinuous cost functions. *Monthly Weather Review*, 128:4031–4044, 2000.

[236] K. Zhao, M. Xue, and W.-C. Lee. Assimilation of GBVTD-retrieved winds from single-Doppler radar for short-term forecasting of super typhoon Saomai (0608) at landfall. *Quarterly Journal of the Royal Meteorological Society*, November 2011.

[237] D.-K. Zhou, W. L. Smith, X. Liu, A. M. Larar, H.-L. Huang, J. Li, M. J. McGill, and S. A. Mango. Thermodynamic and cloud parameters retrieval using infrared spectral data. *Geophys. Res. Lett*, 32(L15):805–1, 2005.

[238] D.-K. Zhou, W. L. Smith Sr, X. Liu, A. M. Larar, S. A. Mango, and H.-L. Huang. Physically retrieving cloud and thermodynamic parameters from ultraspectral IR measurements. *Journal of the atmospheric sciences*, 64(3):969–982, 2007.

[239] Y. P. Zhou, W.-K. Tao, A. Y. Hou, W. S. Olson, C.-L. Shie, K.-M. Lau, M.-D. Chou, X. Lin, and M. Grecu. Use of high-resolution satellite observations to evaluate cloud and precipitation statistics from cloud-resolving model simulations. Part I: south China sea monsoon experiment. *Journal of the Atmospheric Sciences*, 64:4309–4329, December 2007.

[240] Y.-Q. Zhu and I. M Navon. Impact of parameter estimation on the performance of the FSU global spectral model using its full-physics adjoint. *Monthly Weather Review*, 127:1497–1517, 1999.

[241] X. Zou and Y-H. Kuo. Rainfall assimilation through an optimal control of initial and boundary conditions in a limited-area mesoscale model. *Monthly Weather Review*, 124:2859–2882, December 1996.

[242] X. Zou, I. M. Navon, and F. X. Le Dimet. Incomplete observations and control of gravity waves in variational data assimilation. *Tellus A*, 44(4):273–296, August 1992.

[243] X. Zou and Q.-N. Xiao. Studies on the initialization and simulation of a mature hurricane using a variational bogus data assimilation scheme. *Journal of the Atmospheric Sciences*, 57(6):836–860, March 2000.

[244] X. L. Zou, I. M. Navon, and J. G. Sela. Variational data assimilation with moist threshold processes using the NMC spectral model. *Tellus A*, 45(5):370–387, 1993.

[245] D. Zupanski. The effects of discontinuities in the Betts-Miller cumulus convection scheme on four-dimensional variational data assimilation. *Tellus A*, 45(5):511–524, October 1993.

[246] D. Zupanski and F. Mesinger. Four-dimensional variational assimilation of precipitation data. *Monthly Weather Review*, 123:1112–1127, April 1995.

[247] D. Zupanski, S.-Q. Zhang, M. Zupanski, A.-Y. Hou, and S.-H. Cheung. A prototype WRF-based ensemble data assimilation system for dynamically downscaling satellite precipitation observations. *Journal of Hydrometeorology*, 12:118–134, February 2011.

[248] D. Zupanski and M. Zupanski. Model error estimation employing an ensemble data assimilation approach. *Monthly Weather Review*, 134(5):1337–1354, May 2006.

[249] D. Zupanski, M. Zupanski, L. D. Grasso, R. Brummer, I. Jankov, D. Lindsey, M. Sengupta, and M. Demaria. Assimilating synthetic GOES-R radiances in cloudy conditions using an ensemble-based method. *International Journal of Remote Sensing*, 32(24):9637–9659, 2011.

[250] M. Zupanski. Maximum likelihood ensemble filter: theoretical aspects. *Monthly Weather Review*, 133:1710–1726, 2005.

[251] M. Zupanski, S. J. Fletcher, I. M. Navon, B. Uzunoglu, R. P. Heikes, D. A. Randall, T. D. Ringler, and D. Daescu. Initiation of ensemble data assimilation. *Tellus A*, 58(2):159–170, March 2006.

[252] M. Zupanski, I. M. Navon, and D. Zupanski. The maximum likelihood ensemble filter as a non-differentiable minimization algorithm. *Quarterly Journal of the Royal Meteorological Society*, 134(633):1039–1050, 2008.

# BIOGRAPHICAL SKETCH

My name is Jeffrey Lawrence Steward, and I was born on September 9th, 1980 in Denver, Colorado to Lawrence and Suzanne Steward. At the age of 8, a kind-hearted volunteer from Hewlett-Packard in Fort Collins, Colorado, showed me how to program in QBasic on an Apple 2E, and I have been hooked ever since. At the age of 16, while other teenagers asked for music CDs or car paraphernalia, I asked for and happily received the Borland C++ 4.5 Compiler from my supportive father. As a sophomore at the University of Colorado studying computer science, I was incredibly fortunate to land an internship at IBM in Boulder, and after this, due to the booming economy of the dot-com bubble, I found I was able to find an excellent job even without a degree. I went to work as a programmer in industry, first at small dot-coms and then at large companies such as Disney and Yahoo. After the bubble burst, I wanted to resume my schooling, but found the thought of learning computer science a chore, so I went back and completed my education, this time following my other passion of Chinese Language and Literature at the University of Florida. After briefly going back to industry, seeking more challenge and freedom, I enrolled in the Florida State University to pursue my Ph. D. under the direction of the wonderful and brilliant professor Ionel Michael Navon in the field of computational science – the study of *using* computers to solve problems in science rather than the study of computers themselves. This truly was one of the best decisions in my life, for I learned that my true passion, besides languages, was to be found in using computational tools and applied mathematics to solve open problems in science. In the future I hope to continue to find new and challenging problems, especially in the domain of optimal control of fluid dynamical systems, and seek new and creative solutions in order to contribute to science and society as a whole.