LAPPEENRANTA
UNIVERSITY OF TECHNOLOGY

Harri Auvinen

INVERSION AND ASSIMILATION METHODS WITH APPLICATIONS IN GEOPHYSICAL REMOTE SENSING

Acta Universitatis
Lappeenrantaensis
375

Supervisor   Professor Heikki Haario
             Faculty of Technology
             Department of Mathematics and Physics
             Lappeenranta University of Technology
             Finland

Reviewers   Professor Heikki Järvinen
             Finnish Meteorological Institute
             Finland

             Professor Markku Lehtinen
             University of Oulu
             Sodankylä Geophysical Observatory
             Finland

Opponent   Professor Heikki Järvinen
             Finnish Meteorological Institute
             Finland

To Jaana, Joonas and Jessica

Preface

The first part of this work was carried out between 1998 and 2002 at the former Geophysical Research Division (GEO) of the Finnish Meteorological Institute (FMI); the OSIRIS/Odin[1] project continues under Earth Observation Unit of FMI. Since then the focus of the work shifted to assimilation research, supported by the Graduate School of Inverse Problems at the University of Helsinki, and also by the Centre of Excellence in Inverse Problems, performed at Department of Mathematics and Physics, Lappeenranta University of Technology.

I would like to express my gratitude to my supervisor, Professor Heikki Haario, the Head of the Department of Mathematics and Physics at Lappeenranta University of Technology, for all his guidance, ideas and encouragement during my career as a researcher. In fact, without Heikki's contribution my profession would be something totally different. Furthermore I am grateful to Tuomo Kauranne of Lappeenranta University of Technology and Associate Professor Johnathan M. Bardsley of the University of Montana, USA, for their useful ideas and suggestions.

I wish to tank my supervisor during my FMI years, Professor Erkki Kyrölä. Furthermore, I thank all my former colleagues at FMI, especially Liisa Oikarinen[2], Johanna Tamminen, Simo Tukiainen, Seppo Hassinen, Annika Seppälä, Teemu Siiskonen, Marko Laine, Petri Toivanen and Jouni Polkko. Special thanks go to Risto Pellinen, the former Head of the GEO, and Tuija Pulkkinen the present Head of Earth Observation Unit. The spirit of GEO still inspires and reminds me about my years there.

I would like to thank all my colleagues and friends at the Department of Mathematics and Physics in Lappeenranta, especially Antti Solonen, Tapio Leppälampi, Jere Heikkinen and Virpi Junttila for creating a pleasant working atmosphere and for interesting discussions. Furthermore, I thank Antti Solonen for his contribution of coding chi-square test routines used in this work.

I also wish to thank the reviewers Professor Markku Lehtinen and Professor Heikki Järvinen for giving valuable suggestions for the thesis.

Finally I wish to express my gratitude to my spouse Jaana Huuskonen for her invaluable support and resilience during this work.

Lappeenranta, December 2009

Harri Auvinen

---

[1]Odin is a Swedish-led satellite project funded jointly by Sweden (SNSB), Canada (CSA), Finland (Tekes) and France (CNES).

[2]Deceased 27th April 2002.

Abstract

Harri Auvinen

INVERSION AND ASSIMILATION METHODS WITH APPLICATIONS IN GEOPHYSICAL REMOTE SENSING

Stratospheric ozone can be measured accurately using a limb scatter remote sensing technique at the UV-visible spectral region of solar light. The advantages of this technique includes a good vertical resolution and a good daytime coverage of the measurements. In addition to ozone, UV-visible limb scatter measurements contain information about NO_2, NO_3, $OClO$, BrO and aerosols. There are currently several satellite instruments continuously scanning the atmosphere and measuring the UV-visible region of the spectrum, e.g., the Optical Spectrograph and Infrared Imager System (OSIRIS) launched on the Odin satellite in February 2001, and the Scanning Imaging Absorption SpectroMeter for Atmospheric CartograpHY (SCIAMACHY) launched on Envisat in March 2002. Envisat also carries the Global Ozone Monitoring by Occultation of Stars (GOMOS) instrument, which also measures limb-scattered sunlight under bright limb occultation conditions. These conditions occur during daytime occultation measurements.

The global coverage of the satellite measurements is far better than any other ozone measurement technique, but still the measurements are sparse in the spatial domain. Measurements are also repeated relatively rarely over a certain area, and the composition of the Earth's atmosphere changes dynamically. Assimilation methods are therefore needed in order to combine the information of the measurements with the atmospheric model.

In recent years, the focus of assimilation algorithm research has turned towards filtering methods. The traditional Extended Kalman filter (EKF) method takes into account not only the uncertainty of the measurements, but also the uncertainty of the evolution model of the system. However, the computational cost of full blown EKF increases rapidly as the number of the model parameters increases. Therefore the EKF method cannot be applied directly to the stratospheric ozone assimilation problem.

The work in this thesis is devoted to the development of inversion methods for satellite instruments and the development of assimilation methods used with atmospheric models.

3D-Var	3-Dimensional Variational data assimilation
4D-Var	4-Dimensional Variational data assimilation
BFGS	Broyden-Flether-Goldfarb-Shanno
CCD	Charge coupled device
CFC	Chloro-fluoro-carbons
CNES	Centre National d'Etudes Spatiales
CDIPI	Combining Differential-Integral approach involving the Picard Iterative approximation
CSA	Canadian Space Agency
DNA	Deoxyribonucleic acid
DOAS	Differential Optical Absorption Spectroscopy
DU	Dobson unit
ECMWF	European Centre for Medium-range Weather Forecast
EKF	Extended Kalman filter
EnKF	Ensemble Kalman filter
EOS	Earth Observing System
ESA	European Space Agency
FLKS	Fixed-lag Kalman smoother
FMI	Finnish Meteorological Institute
FOV	Field of view
GEO	Geophysical Research Division
GOMOS	Global Ozone Monitoring by Occultation of Stars
IR	Infrared
KF	Kalman filter
LBFGS	Limited memory Broyden-Flether-Goldfarb-Shanno
LIMBTRAN	Pseudo-spherical model for limb-viewing geometry
LORE	Limb Ozone Retrieval Experiment
LOS	Line-of-sight
MAP	Maximum *a posteriori*
MCMC	Monte Carlo Markov chain
MOP	Modified Onion Peeling

NASA	National Aeronautics and Space Administration
OSIRIS	Optical Spectrograph and Infrared Imager System
OMI	Ozone Measuring Instrument
POAM	Polar Ozone and Aerosol Measurement
RTM	Radiative transfer model
SBUV	Solar Backscattered Ultraviolet
SCIAMACHY	Scanning Imaging Absorption SpectroMeter for Atmospheric CartograpHY
SME	Solar Mesospheric Explorer
SOLSE	Shuttle Ozone Limb Sounding Experiment
STP	Standard temperature and pressure
TOMS	Total Ozone Mapping Spectrometer
UV	Ultraviolet
VIS	Visible
VKF	Variational Kalman filter
VKS	Variational Kalman smoother
WMO	World Meteorological Organization

CONTENTS

List of the original articles and the author's contribution

This thesis consist of an introductory part and four original refereed articles in scientific journals. The articles and the author's contributions in this thesis are summarized below.

   I   **Auvinen, H., L. Oikarinen, and E. Kyrölä**, Inversion algorithms for recovering minor species densities from limb scatter measurements at UV-visible wavelengths, *J. Geophys. Res., 107* (D13), 4172, doi:10.1029/2001JD000407, 2002.

  II **Tukiainen, S., S. Hassinen, A. Seppälä, H. Auvinen, E. Kyrölä, J. Tamminen, C. S. Haley, N. Lloyd, and P. T. Verronen**, Description and validation of a limb scatter retrieval method for Odin/OSIRIS, *J. Geophys. Res., 113*, D04308, doi:10.1029/2007JD008591, 2008.

 III **H. Auvinen, J. M. Bardsley, H. Haario and T. Kauranne**, The variational Kalman filter and an efficient implementation using limited memory BFGS, *International Journal on Numerical methods in Fluids*, doi:10.1002/fld.2153, 2009.

 IV **Auvinen H., Bardsley J. M., Haario H., Kauranne T.**, Large-scale Kalman filtering using the limited memory BFGS method. *Electronics Transactions on Numerical Analysis*, Vol. 35, pp. 217-233, 2009.

H. Auvinen is the principal author of the articles **I** and **III**. The author has developed together with L. Oikarinen and E. Kyrölä the concept of the Modified Onion Peeling (MOP) inversion method for limb scatter measurements, introduced in article **I**. Furthermore, the author has implemented and optimized the forward model of the method used in articles **I** and **II**, excluding the multiple scattering database. The author has developed the concept of the variational Kalman filter (VKF) method in company with H. Haario, T. Kauranne and J. M. Bardsley, (article **III**). Furthermore, the author has implemented the VKF method together with J. M. Bardsley. The author has computed all of the results presented in articles **I** and **III**, and most of the results presented in article **IV**. The author has translated the Lorenz'95 code, used in articles **III** and **IV**.

The overview of the thesis is organized as follows. Chapter 2 introduces the research field, the objectives and the scientific contribution of the thesis. The inversion method of the OSIRIS instrument and the retrieval results are introduced in Chapter 3, which summarizes the publications (**I** and **II**). In Chapter 4, the assimilation methods, namely the variational Kalman filter, variational Kalman smoother and LBFGS-KF are introduced (publications **III** and **IV**). The conclusions of this work are presented in Chapter 5.

Introduction

Acquiring accurate information on the concentrations of various atmospheric constituents is one of the key objectives of atmospheric research. Ozone is one of the crucial components in the atmosphere, since stratospheric ozone provides a shield against solar ultraviolet (UV) radiation. UV light is electromagnetic radiation with a wavelength range from 200 nm to 400 nm. UV exposure causes DNA damage, and therefore, it is an environmental human carcinogen. The toxic effects of UV light motivates atmospheric scientists to measure the amount of ozone variations and to detect ozone holes. Furthermore, information about the trend of the total amount of ozone has played an important role in political decisions that aim to conserve the Earth's atmosphere.

G. M. B. Dobson $(1889 - 1976)$ is one of the pioneers in a field of ozone research. One Dobson unit (DU), which is named after him, is defined to be a 0.01 mm thick layer of ozone at standard temperature and pressure (STP). An average amount of ozone over a certain area at STP would form a layer with a thickness of approximately 3 mm, which corresponds to 300 DU. Dobson developed the Dobson spectrophotometer to measure the amount of ozone in the atmosphere from the ground. The instrument measures the intensity of the solar electromagnetic radiation at four wavelengths, two of which are contributed by ozone.

The first empirical indications of ozone loss were documented in the 1980's, when scientists found a large ozone hole over Antarctica. The drop in stratospheric ozone was locally very dramatic: roughly a one-third loss compared to the average of the 1960's. At first, scientists thought that there was a problem with their instruments and the findings were even omitted as faulty data. It was measurements from the *satellite instrument* Solar Backscatter Ultraviolet/Total Ozone Mapping Spectrometer (SBUV/TOMS) onboard NASA's Nimbus-7 that confirmed the existence of the ozone hole [*Farman et. al.*, 1985].

The production and reduction reactions of ozone in the upper atmosphere are mainly caused by UV radiation, together with several other mechanisms (see [*Chapman S.*, 1930]). The anthropogenic release of CFC gases is involved in the excessive destruction of stratospheric ozone. The process is roughly the following. First, UV-radiation *hv* at wavelength $\lambda < 230$ nm disintegrates the chloro-fluoro-carbons (CFC) molecules and releases chlorine atoms:

$$CF_2Cl_2 + hv \quad \rightarrow \quad CF_2Cl + Cl \tag{2.1}$$

$$CFCl_3 + hv \quad \rightarrow \quad CFCl_2 + Cl. \tag{2.2}$$
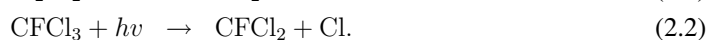
The reaction of chlorine atoms with ozone molecules produces chlorine monoxide and oxygen, which leads furthermore to O_3 depletion reactions

$$Cl + O_3 \rightarrow ClO + O_2 \tag{2.3}$$

$$ClO + O_3 \rightarrow Cl + 2O_2 \tag{2.4}$$

$$ClO + O \rightarrow Cl + O_2. \tag{2.5}$$

At the end, the chlorine atom is free to start the process all over again. This process continues until chlorine atoms react with methane or nitrogen dioxide, which form inactive molecules. It is estimated that a single chlorine atom is able to react with $100,000$ ozone molecules. The ozone-destroying chemical process, from which only the rudiments are described above, requires cold conditions that lead to heterogeneous reactions on ice particles. It is therefore natural that the first ozone hole appeared over the Antarctica. The first publication that indicated the kinetics above was published in [*Crutzen, P. J.*, 1974], roughly 10 years before the empirical confirmation.

During the most recent decades, several remote sensing techniques have been demonstrated to be suitable for measuring atmospheric ozone. These methods use UV, visible, infrared (IR), submillimeter and microwave regions of the electromagnetic spectrum from different sources. This work concentrates on a limb scatter technique at the UV-visible spectral region of solar light. In addition to ozone, the UV-visible limb scatter measurements contain information about NO_2, NO_3, $OClO$, BrO and aerosols. Typically, many of these constituents are measured simultaneously by using a suitable data processing algorithm.

In Finland, the Finnish Meteorological Institute (FMI) has contributed significantly to several projects that aim to measure particularly stratospheric ozone. The work started in the late 1980's with the Global Ozone Monitoring by Occultation of Stars (GOMOS) instrument program and continued a few years later with the Optical Spectrograph and Infrared Imager System (OSIRIS) instrument project. GOMOS is onboard the European Space Agency's (ESA) Envisat satellite, launched in March 2002. OSIRIS was launched in February 2001 onboard the Odin satellite. In addition, FMI manages several other satellite programs, such as ESA's Ozone Measuring Instrument (OMI) program.

The global coverage of satellite measurements is naturally far better than any other ozone measurement technique, but still the measurements are sparse in the spatial domain. Measurements are also repeated relatively rarely over a certain area. The ozone remote sensing task is, in principle, a tomography problem. However, the composition of the Earth's atmosphere changes dynamically, so assimilation methods are needed in order to combine the information of the measurements and the atmospheric model.

In recent years, the focus of assimilation algorithm research has turned towards the use of filtering methods. The traditional Extended Kalman filter (EKF) method provides the estimate of forecast error covariance and takes into account not only the uncertainty of the measurements, but also the uncertainty of the evolution model of the system. However, the computational cost of full-blown EKF increases rapidly as the number of the model parameters increases. Therefore, the EKF method cannot be applied directly to the assimilation problem of stratospheric ozone.

The aim of this work is to present inversion methods for satellite instruments, as well as methods to assimilate the measurement results in global dynamic chemistry models of the atmosphere.

2.1 Remote sensing techniques of the Earth's atmosphere

There are various important criteria for ozone measurement methods, such as the vertical resolution, long-term stability and coverage of the measurements. Basically, there are three different classes of ozone related UV-visible satellite instruments, cathegorized by the measurement geometry:

i. Occultation instruments measure the transmission of the sun/moon/stellar light through the atmosphere.

ii. Limb-viewing instruments measure transmitted and scattered sunlight through the atmosphere, Figure (2.1). This geometry occurs also during daytime occultations, since the sun illuminates the atmosphere.

iii. Nadir viewing instruments measure transmitted and backscattered sunlight from the Earth's surface, with a relatively wide viewing angle.
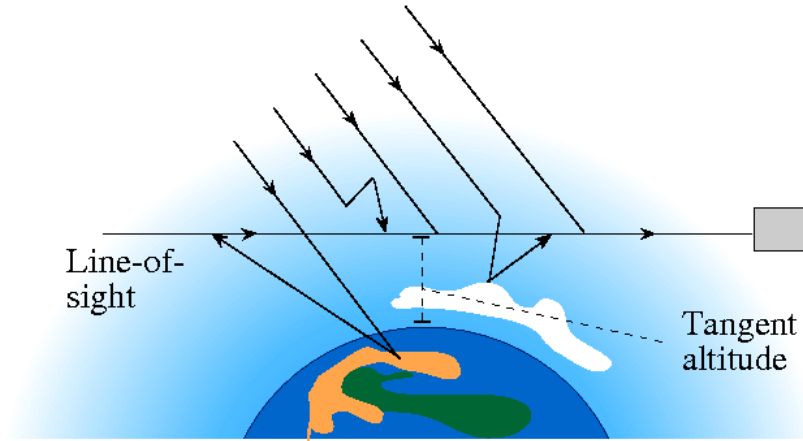
Figure 2.1: The limb scatter measurement geometry. Measured solar light contains single and multiple scattered effects. The latter part may include contributions from reflection at the Earth's surface and/or clouds.

For occultation instruments, the measurement geometry, at least during nighttime, is simple, since the multiple scattering component illustrated in Figure (2.1) vanishes. The remaining *transmission* of the stellar light along the line-of-sight (LOS) can be modelled by Beer's law as follows:

$$T(\boldsymbol{\rho}, \lambda, \ell) = \frac{I_{obs}}{I_0} = \exp\left\{ -\sum_{j=1}^{J} \alpha_j(\lambda) \int_{\ell} \rho_j(s)ds \right\}, \tag{2.6}$$

where $\alpha_j(\lambda)$ is the so-called cross section at wavelength λ, which is known and typical for each constituent, and $\int_{\ell} \rho_j(s)ds$ is the integrated line density over ray path ℓ for constituent j. The contribution of different constituents is summed over J. The reference spectrum I_0 is measured above the atmosphere with the same instrument.

The inverse problem for occultation instruments is to estimate the densities $\boldsymbol{\rho}$ in (2.6) of the atmosphere locally in space and time, with given measurements at different tangent altitudes.

A simplified approximation for limb scatter measurements is the single scattering scenario

$$I^{SS}(\boldsymbol{\rho}, \lambda, \ell^{\dagger}) = I^{\text{Sun}}(\lambda) \int_{\ell^{\dagger}} T^{ps}(\boldsymbol{\rho}, \lambda, s) T^{op}(\boldsymbol{\rho}, \lambda, s) P[\lambda, \mathbf{r}(s), \theta(s)] k^s[\lambda, \mathbf{r}(s)] ds. \qquad (2.7)$$

Here s is a distance along the LOS of the detector, measured from the detector. The line-of-sight $\ell^{\dagger}$ is specified by both the detector position and its viewing direction. $T^{ps}(\boldsymbol{\rho}, \lambda, s)$ is the transmittance of the ray path from the Sun to the scattering point $\mathbf{r}(s)$ at wavelength λ, and similarly $T^{op}(\boldsymbol{\rho}, \lambda, s)$ is the transmittance along the line-of-sight $\ell^{\dagger}$ from the scattering point to the detector, computed as in (2.6). The solar irradiance incident on the atmosphere is denoted by $I^{\text{Sun}}(\lambda)$. It arrives at the LOS at angle $\theta(s)$. Due to refraction, the scattering angle $\theta(s)$ varies slightly along the LOS.

The scattering phase function $P(\lambda, \mathbf{r}, \theta)$ is a weighted sum of phase functions for molecular and particle scattering:

$$P(\lambda, \mathbf{r}, \theta) = P^m(\lambda, \mathbf{r}, \theta) \frac{k^m(\lambda, \mathbf{r})}{k^s(\lambda, \mathbf{r})} + P^a(\lambda, \mathbf{r}, \theta) \frac{k^a(\lambda, \mathbf{r})}{k^s(\lambda, \mathbf{r})}. \qquad (2.8)$$

Function $k^m(\lambda, \mathbf{r})$ is the coefficient for scattering from molecules. It is equal to the product of the spectral scattering cross section and the local density, $k^m(\lambda, \mathbf{r}) = \sigma^m(\lambda, \mathbf{r}) \rho^m(\lambda, \mathbf{r})$. Similarly, $k^a(\lambda, \mathbf{r})$ is the coefficient for scattering from aerosols. The total scattering coefficient $k^s(\lambda, \mathbf{r})$ at wavelength λ and point $\mathbf{r}$ is $k^s(\lambda, \mathbf{r}) = k^m(\lambda, \mathbf{r}) + k^a(\lambda, \mathbf{r})$. The phase function for molecular scattering can be taken to be independent of $\mathbf{r}$, and it depends on the scattering angle θ as

$$P^m(\theta) = \frac{3}{16\pi}(1 + \cos^2 \theta), \qquad (2.9)$$

where we have ignored the depolarization of Rayleigh scattering (which makes P^m also independent on λ). The aerosol scattering phase function can be simulated by the Henyey-Greenstein phase function

$$P^a(\lambda, \mathbf{r}, \theta) = \frac{1}{4\pi} \frac{1 - g^2}{(1 + g^2 - 2g\cos\theta)^{3/2}}. \qquad (2.10)$$

Parameter g is the asymmetry parameter of the phase function. It generally depends on both λ and $\mathbf{r}$. A representative value for background stratospheric aerosols is $g = 0.75$.

However, the single scattering model is not very accurate at wavelengths longer than approximately 310 nm. The amount of multiple scattering depends on the wavelength, the reflectivity of the underlaying surface of the Earth, and amount of aerosols along the measurement. A heavy aerosol load in the stratosphere, or a strongly reflecting lower surface, can increase the multiple scattering effect significantly. The amount of multiple scattering also depends on the solar angles of the measurement. Especially the relative amount of multiple scattering decreases as the solar zenith angle increases. In the OSIRIS measurement geometry, multiple scattering constitutes $10-40\%$ of the total intensity at visible wavelengths [*Oikarinen et al.*, 1999]. Figure 2.2 shows the UV-visible spectrum of limb radiance at tangent altitudes $10, 15, 20$, and 30 km. The spectra were simulated by the Monte Carlo model Siro [*Oikarinen et al.*, 1999] using the U.S. Standard Atmosphere 1976 neutral density profile, the U.S. Standard ozone concentration profile, the MODTRAN background

stratospheric aerosol model [*Berk et al.*, 1989], and the surface albedo of $A = 0.3$. The solar zenith angle at the LOS tangent point was 80 deg and the azimuth angle was 90 deg.

A key issue for constituent retrieval from limb scattering measurements is the demonstration of the ability to retrieve with acceptable accuracy the vertical profile of ozone and possibly other gases. The main obstacle is to understand how to handle the multiple scattering component of the signal [*Oikarinen et al.*, 1999].

The simulation of limb scatter measurements is complicated, since it requires solving the radiative transfer equation with multiple scattering in a spherical atmosphere. Some radiative transfer models (RTMs) that can be applied to limb-viewing have been developed. The model introduced by *Herman et. al.* [1994] uses a Gauss-Seidel iteration scheme to solve the RTM in a spherical atmosphere. A pseudo-spherical model for limb-viewing geometry (LIMBTRAN) has been developed by *Griffioen and Oikarinen* [2000], which makes use of a plane-parallel finite-difference or matrix operator doubling and adding model. A spherical RTM called "Combining Differential-Integral approach involving the Picard Iterative approximation" (CDIPI) has been introduced by *Rozanov et. al.* [2000]. In addition, Monte Carlo radiative transfer models have been used to simulate limb radiance [*Collins et. al.*, 1972; *Marchuk et. al.*, 1980; *Oikarinen et al.*, 1999].

Although the models listed above have been developed with computational efficiency especially in mind, solving the RTM at each wavelength and tangent altitude at each iteration step of data inversion is very time consuming.
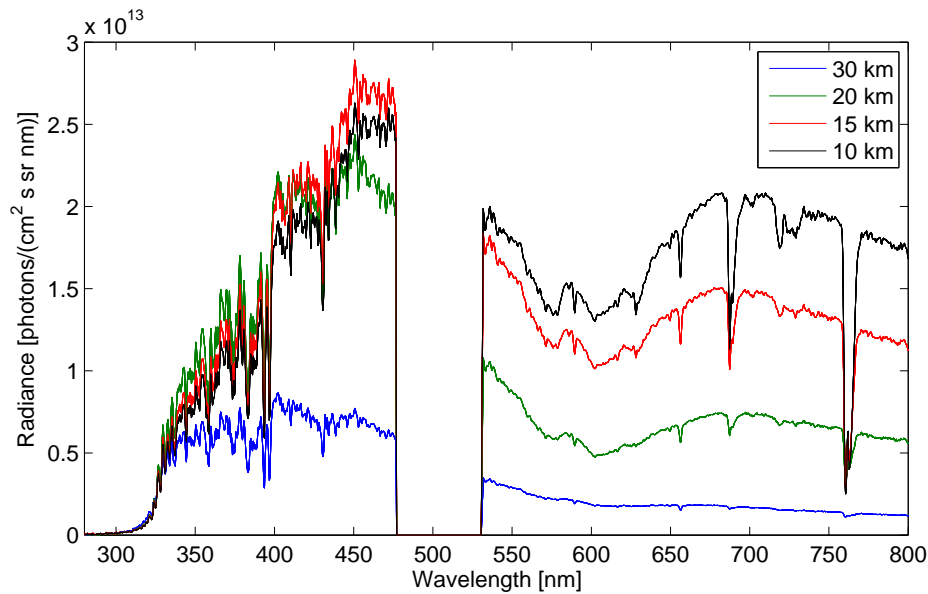
Figure 2.2: Measured limb radiance spectrums of the OSIRIS instrument at tangent altitudes $10, 15, 20$ and 30 km. The contribution of the ozone is visible at the UV-region and at the Chappuis-band. Note that OSIRIS does not record the spectra between 470 and 530 nm, and the wavelengths beyond 680 nm are not used in the data inversion process.

The advantage of nadir viewing instruments is a good global coverage, but vertical profiles of the constituents are difficult to retrieve. Occultation measurements, on the other hand, provide a good

vertical resolution, but the global coverage is uneven. Continuous limb scanning measurements combine both good global coverage and good vertical resolution.

The first spaceborne UV-visible limb scatter measurements were made by instruments onboard the Solar Mesospheric Explorer (SME) [*R*usch et al., 1984]. The retrieval of stratospheric ozone profiles with this technique has been demonstrated by the Shuttle Ozone Limb Sounding Experiment (SOLSE) and the Limb Ozone Retrieval Experiment (LORE) instruments on the Space Shuttle flight STS-87 in 1997 [*McPeters et al.*, 2000; *Flittner et al.*, 2000].
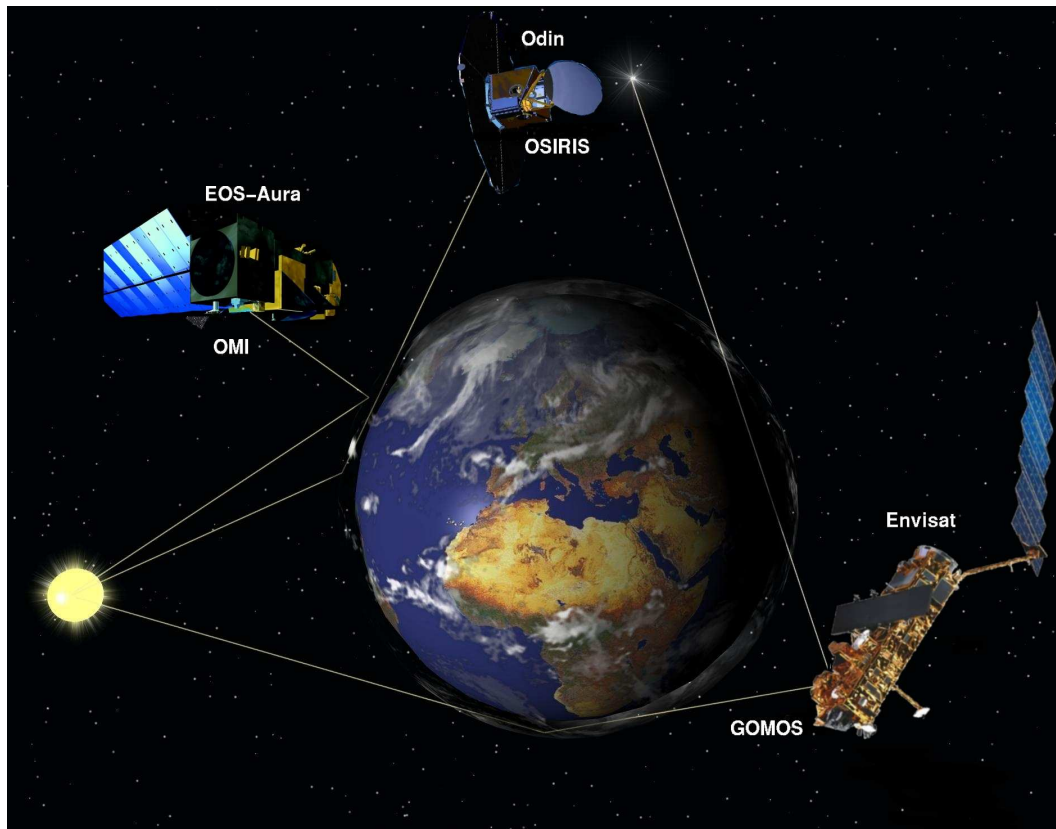
Figure 2.3: OSIRIS, GOMOS and OMI instruments. Courtesy of Seppo Hassinen, Finnish Meteorological Institute, Swedish Space Corporation, ESA and NASA.

At present, there are several limb-scanning satellite instruments using the UV-visible region of the spectrum, e.g. the Optical Spectrograph and Infrared Imager System (OSIRIS) [*Llewellyn et al.*, 1997] launched on the Odin satellite in February 2001 and the Scanning Imaging Absorption SpectroMeter for Atmospheric CartograpHY (SCIAMACHY) [*Bovensmann et. al.*, 1999] launched on Envisat in March 2002. Envisat also carries the Global Ozone Monitoring by Occultation of Stars (GOMOS) instrument [*Bertaux et. al.*, 1991], which measures also limb-scattered sunlight under bright limb occultation conditions. These conditions occur during daytime occultation measurements.

The Finnish Meteorological Institute (FMI) has been involved in the GOMOS project from the very begining. The work started at FMI with forward model simulations already during the year 1988 and continued with the GOMOS instrument proposal one year later (1989). Furthermore, FMI has contributed significantly to the development of the inverse methods for GOMOS and OSIRIS instruments. The work with the OSIRIS instrument started in the mid 90's. The present thesis research is a part of this program; in collaboration with FMI, the author has developed the concept of the Modified Onion Peeling (MOP) method, as will be described more in detail in Chapter 3.

The orbit of Odin is sun-synchronous, with the ascending node at 18.00, with an altitude of about 600 km and a period of 96 min. During 24 h there are about 15 orbits. OSIRIS scans the limb either in a continuous or in a stepwise manner with tangent altitudes of the measurements $7-60$ km (alternatively $7-120$ km). For Odin's aeronomic mode the instrument nominally scans in the orbit plane, but it is possible to scan off-track by up to ± 32 deg. Each scan takes about $70-150$ s, giving about $40-60$ scans per orbit. The restituted pointing accuracy in the aeronomy mode is 1.2 arc-minutes, which is equivalent to 1 km uncertainty in the vertical at the tangent point. As OSIRIS scans the limb, the tangent point moves 0.75 km/s in the vertical direction. During measurements in the orbit plane the tangent point moves horizontally 1.7 km/s towards or away from the spacecraft depending on the scanning direction (upward or downward scan). The satellite itself moves at 7.6 km/s relative to Earth's surface. This leads to a net movement of $400-1300$ km during one scan.

The OSIRIS instrument has a UV-visible spectrograph, covering the wavelength ranges of $280-470$ and $530-800$ nm, and 3 fixed 30 nm wide IR channels centered to 1260 nm, 1270 nm, and 1560 nm. The spectrograph has a charge-coupled device (CCD) matrix detector. The slit of the OSIRIS spectrograph is oriented along the Earth limb. The instrument has a telescope of aperture 10 cm^2. Spectral resolution of OSIRIS is 1 nm in the UV-visible. The spectrograph field of view (FOV) is 0.02 deg $\times$ 0.75 deg leading to a vertical coverage of 1 km and a horizontal coverage of 40 km across the line-of-sight. The detector integration time can be varied, typically values between 0.1 s and 5 s are used.

The orbit of Envisat, the platform of SCIAMACHY and GOMOS, is also a Sun synchronous polar orbit but with an ascending node at 10.00. The orbital altitude is around 800 km and the orbit period is 101 min. During 24 h there are about 15 orbits.

The GOMOS instrument has a UV-visible range spanning $250-675$ nm and there are also two infrared channels at $756-773$ nm and at $926-952$ nm. All channels have CCD detectors. The slit of GOMOS is oriented perpendicular to the Earth limb. Spectral resolution of the instrument is 1.2 nm in the UV-visible and 0.2 nm in the IR. GOMOS looks at setting stars during night and day. During day time measurements the CCDs record not only the star signal but also solar light scattered at the limb of the atmosphere just below and above the star. The spectral resolution of the limb scatter signal is ~ 5 nm, which is worse than the resolution of the stellar spectrum due to relatively large slit of the GOMOS spectrometer. The vertical coverage of one exposure is about 2 km and the horizontal coverage is 35 km. The stars can be occulted in the azimuth limits -10 deg to 90 deg with respect to the anti-velocity vector. The occultations in the orbit plane cover the altitude region $15-120$ km and the vertical slice during the integration time of 0.5 s is 1.7 km. Occultations off the orbital plane take more time and the vertical resolution is better. The tangent point movement in the orbital plane occultations is ~ 10 km in latitudinal direction and ~ 30 km in longitudinal direction. In the side occultations lasting 175 s it is ~ 1300 km and ~ 800 km, respectively.

In the limb mode SCIAMACHY will scan the atmosphere from the ground to about 100 km in the direction of the velocity vector of the satellite. SCIAMACHY has 8 one-dimensional Reticon

array detectors. These cover the spectral region $240 - 2380$ nm (a few gaps included) at a spectral resolution varying from 0.22 nm to 1.48 nm. The detecctor integration time is 1 s. The vertical resolution of SCIAMACHY is 3 km and the horizontal resolution is 100 km.

The OMI instrument onboard the EOS-Aura satellite measures the backscattered solar radiation from the Earth's surface through the atmosphere in nadir geometry. The instrument contains an imaging spectrograph with a wavelength range of $270 - 500$ nm. The spectral resolution of the instrument is about 0.5 nm. The viewing angle of the telescope is 144 deg, which corresponds to a 2600 km wide sweep of the Earth's surface. This geometry enables measurements with a daily global coverage. The depolarization of the instrument is taken care using a scrambler. Afterwards, the signal is split into two channels. The UV-channel covers $270 - 380$ nm and the VIS channel covers with overlap $350 - 500$ nm.

2.2   Scientific contribution of the thesis

This thesis concentrates on two primary objectives: to develop inversion methods for limb scatter measurements and assimilation methods for large scale problems. The motivation of this work was originally to develop an inversion method for the data processing of the Optical Spectrograph and Infrared Imager System (OSIRIS) satellite instrument. The inversion method developed can also be applied to GOMOS (Global Ozone Monitoring by Occultation of Stars) bright limb measurements. The final aim is to assimilate the measurements with atmospheric models in order to estimate global distributions of the constituent concentrations. Since the atmospheric models are extremely high dimensional, the traditional assimilation methods, such as the Kalman filter, cannot be directly applied. This motivates to develop assimilation methods, which are suitable for large scale problems.

The thesis consist of four original publications (**I-IV**). Papers **I** and **II** concentrate on the inversion method of the OSIRIS instrument while papers **III** and **IV** describe assimilation methods for large scale problems. The more detailed scientific contributions of the individual papers are as follows. Publ. **I** introduces the Modified Onion Peeling (MOP) inversion method for the OSIRIS data processing. The paper includes a description of the forward model of the inversion method, which uses *a priori* information of the neutral density of the atmosphere. The results of the paper have been created with simulated data, since the work was done prior to the launch of the satellite. Publ. **II** describes necessary updates of the method after the launch of the satellite. The presented inversion and validation results are produced using real measurements of the OSIRIS and GOMOS instruments. Publ. **IV** introduces optimization based assimilation methods, which approximate the standard Kalman filter (KF) and the Extended Kalman filter (EKF) by replacing matrix inversions with a limited memory Broyden-Flether-Goldfarb-Shanno (LBFGS) optimizations. This reduces the computational cost significantly compared to the Kalman filter. The effectiveness of the method is demonstrated in linear and non-linear test cases. Publ. **III** introduces the variational Kalman filter (VKF) method, which also approximates KF and EKF. The derivation of the method is based on the Bayesian approach, and the traditional computation of the Kalman gain and the matrix inversions are avoided. The state estimate produced by the filter is computed iteratively using the limited memory BFGS optimization method. The effectiveness of the VKF method is tested again with linear and non-linear test cases. The convergence of BFGS/LBFGS methods is demonstrated and tested numerically.

An inversion method for the OSIRIS satellite instrument

Instruments measuring transmitted and/or scattered sunlight in the UV-visible wavelength range have the advantage that molecules responding at this region are insensitive to the atmospheric temperature and pressure. Furthermore, the need of absolute knowledge of the external source of light can be avoided by using the relative measurement principle. The absolute knowledge can be archived by directly measuring, for example, the solar irradiance spectrum, but it is difficult to make these measurements with the required stability, [e.g. *World Meterological Organization*, 1988]. The relative measurement principle is a very natural approach for occultation measurements. Solar occultation measurements have good signal-to-noise ratios, but their global coverage is poor and the termination geometry is an additional source of uncertainty for diurnally varying species.

The UV-visible wavelength band of the OSIRIS instrument carries information of several atmospheric constituents, including ozone (O_3), nitrogen dioxide (NO_2), nitrogen trioxide (NO_3), chlorine dioxide ($OClO$) and bromine monoxide (BrO). These constituents leave their absorption fingerprints in the solar spectrum, which are scattered by the neutral molecules of the atmosphere. Stratospheric aerosols also contribute to the shape of the observed spectrum.

As the measurement principle of the remote sensing instruments is indirect, data inversion methods are needed to extract the information from the physical measurements. Several inversion methods have been developed to invert the density of the quantities from the OSIRIS measurements. One of those is the Triplet method, developed by *Flittner et. al.* [2000] and *McPeters et. al.* [2000]. The method is adapted to the OSIRIS limb scatter measurements in [*Von Savigny et al.*, 2003]. The Triplet method uses wavelength triplets in the Chappuis absorption band near 600 nm to retrieve stratospheric ozone. According to *Petelina et. al.* [2004], the validated altitude range for the OSIRIS triplet ozone is $15-32$ km.

The DOAS (Differential Optical Absorption Spectroscopy) method is a widely used approach to retrieve several different atmospherical constituents. The basic DOAS approach was proposed by *Platt* [1994] and the method was applied to simulated limb scatter measurements by *McDade et. al.* [2002] and *Strong et. al.* [2002]. According to *Haley et. al.* [2003], the DOAS technique was applied to retrieve ozone from OSIRIS data and found good agreement with the Triplet method.

The DOAS method has also been used to retrieve stratospheric NO_2 from the OSIRIS data. *Haley et. al.* [2004] and *Sioris et. al.* [2003] used slightly different DOAS variants and obtained quite consistent results. They also performed preliminary validation against sonde and POAM III measurements, proving that the method is feasible.

3.1   Modified Onion Peeling inversion method

The general idea of the so-called onion peeling inversion methods is to divide the atmosphere into separate layers. Within each layer, the densities of the different constituents are assumed to be constant. In a standard onion peeling approach applied, for example, to occultation data, one assumes that a measurement depends only on constituent densities in and above the layer which contains the tangent altitude of the measurement. The previous assumption is quite natural for nighttime occultations, since the measurement contains mainly the transmission through the atmosphere, and the multiple scattering component of the signal can be neglected. At first, the densities in the uppermost layer are inverted using the data from the uppermost measurement. Then the densities are inverted layer by layer from the top of the atmosphere downwards using the already inverted densities for the layers above. In this way, we can construct the vertical profiles of trace constituents of the atmosphere.

In the scattered sunlight case, where the measured signal includes contributions from the lower atmosphere, we cannot, in principle, use the onion peeling method. However, it is possible to use the onion peeling approach in an iterative manner. In every iteration of the peeling process, we obtain a new better approximation for the lower atmosphere. Usually $2-3$ iterations of the peeling is enough to receive accurate results.

The theoretical basis of the Modified Onion Peeling (MOP) inversion method, introduced in Publ. **I**, is a Bayesian approach. Using a flat *a priori* distribution and assuming Gaussian measurement noise, the solution reduces to a simple weighted least-squares fit to the data, see e.g. *Rodgers* [2000]. Non-linear problems, such as the one related to OSIRIS, require an iterative fitting procedure.

It is advantageous not to use directly measured radiances but the ratio

$$R_{\mathrm{obs}}(\lambda, j) = \frac{I_{\mathrm{obs}}(\lambda, j)}{I_{\mathrm{obs}}^{\mathrm{ref}}(\lambda)}, \tag{3.1}$$

where $I_{\mathrm{obs}}(\lambda, j)$ are measured radiances at tangent heights j and $I_{\mathrm{obs}}^{\mathrm{ref}}(\lambda)$ is a reference measurement from the same scan at a high tangent altitude. The radiance is a function of the wavelength λ. We have chosen to use the first measurement below 50 km as the reference. It would be possible to try other tangent heights, as well, but the values around 50 km seem to yield the best results in practice. It is already high enough to exclude spectral fingerprints from minor trace gases (such as NO_2, $OClO$, and BrO), making the reference spectrum easier to model. Furthermore, straylight contamination in OSIRIS increases as a function of tangent height, making high altitude measurements less unreliable to use [*Llewellyn et al.*, 2004].

The use of the so-called transfer spectrum (3.1) is useful because it diminishes systematic errors due to surface albedo, clouds, and polarization [*Flittner et al.*, 2000; *Oikarinen*, 2001]. It also reduces errors due to imperfect instrument calibration.

The modelled transfer spectrum is defined as

$$R_{\mathrm{mod}}(\lambda, j, \boldsymbol{\rho}) = \frac{I_{\mathrm{mod}}(\lambda, j, \boldsymbol{\rho})}{I_{\mathrm{mod}}^{\mathrm{ref}}(\lambda, \boldsymbol{\rho}_{\mathrm{ref}})}, \tag{3.2}$$

where $I_{\mathrm{mod}}(\lambda, j, \boldsymbol{\rho})$ are modeled radiances and $I_{\mathrm{mod}}^{\mathrm{ref}}(\lambda, \boldsymbol{\rho}_{\mathrm{ref}})$ is a model reference spectrum. The gas density profiles $\boldsymbol{\rho}$ are adjusted iteratively, and after every iteration a new and better agreement is obtained between (3.1) and (3.2).

A typical background atmosphere is assumed when we calculate the model reference spectrum. Obviously, even the best estimate differs from the true state of the atmosphere and typically produces a systematic bias to the retrieved profiles. This effect is studied in more detail in [**I** - **II**].

3.2 Multiple scattering correction of the forward model

Taking account of the multiple scattering effects is a crucial part of limb scatter retrieval methods. In the limb scatter geometry, multiple scattering can constitute 10–50% of the observed radiance at visible wavelengths [*Oikarinen et al.*, 1999]. Its proportion of the total radiance is strongly dependent on the wavelength, as the ratio of multiple to total scattering increases steeply at wavelengths greater than 310 nm [*Oikarinen et al.*, 1999]. In addition, the multiple scattering contribution depends on the tangent height, surface albedo, solar angles, and composition of the atmosphere itself. For these reasons, a mere single scattering radiative transfer model is not generally satisfactory to describe scattering and absorption effects in limb scatter problems. However, taking the multiple scattering effects into account complicates limb scatter problems significantly and certainly increases computational costs, which leads to some kind of compromise between the modeling accuracy and the available computation time.

During the fitting procedure in each layer, we have to use a few (usually $3-10$) iterations, and as many forward model calls, before a suitable optimization algorithm finds good enough agreement between the model and the measurement. A full 3-D radiative transfer model operating in a multiple scattering mode would be too slow to use, and hence, we must seek faster solutions. One way would be to reduce the total number of wavelengths used (from around 300 to only a few) in the first peeling loop. This solution would not exploit the whole bandwidth of the OSIRIS instrument, and the altitude range of the ozone retrieval would shrink.

A practical approach, included in the original publication [**I**], is to use a single scattering forward model during the fitting iterations and include multiple scattering effects using precalculated look-up tables. The single scattering forward model built into the MOP inversion module solves the radiative transfer by numerical integration and is computationally efficient to run.

The look-up tables used in the MOP inversion contain modeled (single and multiple scattered) radiances calculated as a function of tangent height, solar angles, season, albedo, and latitude. The look-up tables are produced using the LIMBTRAN [*Griffioen and Oikarinen*, 2000] forward model.

In order to take advantage of the precalculated look-up tables, we divide the modeled transfer spectra (3.2) into two parts:

$$R_{\mathrm{mod}}(\lambda, j, \boldsymbol{\rho}) = \frac{I_{\mathrm{mod}}(\lambda, j, \boldsymbol{\rho})}{I_{\mathrm{mod}}^{\mathrm{ref}}(\lambda, j, \boldsymbol{\rho}_{\mathrm{ref}})} = \frac{I_{\mathrm{mod}}^{\mathrm{ss}}(\lambda, j, \boldsymbol{\rho})}{I_{\mathrm{mod}}^{\mathrm{ref}}(\lambda, \boldsymbol{\rho}_{\mathrm{ref}})} M(\lambda, j), \tag{3.3}$$

where $I_{\mathrm{mod}}^{\mathrm{ss}}(\lambda, j, \boldsymbol{\rho})$ is the dynamic single scattering term (2.7) which is adjusted iteratively during the fitting procedure. The model reference radiance $I_{\mathrm{mod}}^{\mathrm{ref}}(\lambda, \boldsymbol{\rho}_{\mathrm{ref}})$ in the denominator of (3.3) is also calculated with LIMBTRAN (including multiple scattering). The second term on the right side, $M(\lambda, j)$, is the static part which comes from the look-up tables and is kept fixed during the iterations. This correction term is defined as the modeled multiple scattering radiance (total radiance) divided by the corresponding single scattering radiance:

$$M(\lambda, j) = \frac{I_{\mathrm{ms}}(\lambda, j, \boldsymbol{\rho}_{\mathrm{prior}})}{I_{\mathrm{ss}}(\lambda, j, \boldsymbol{\rho}_{\mathrm{prior}})}, \tag{3.4}$$

where $\boldsymbol{\rho}_{\mathrm{prior}}$ are the constituent densities of the standard atmosphere used in the LIMBTRAN radiance simulations.

3.3   Spectral fitting between the measurements and the forward model

At every layer, assuming that the measured transfer spectra (3.1) are independent, the sum of squared residuals is defined as

$$\chi_j^2(\boldsymbol{\rho}) = (R_{\mathrm{mod}}(\boldsymbol{\lambda}, j, \boldsymbol{\rho}) - R_{\mathrm{obs}}(\boldsymbol{\lambda}, j))^{\mathrm{T}} \mathbf{C}^{-1} (R_{\mathrm{mod}}(\boldsymbol{\lambda}, j, \boldsymbol{\rho}) - R_{\mathrm{obs}}(\boldsymbol{\lambda}, j)), \qquad (3.5)$$

where $\mathbf{C}$ is a covariance matrix, which includes contributions from the measurement and modeling errors. The errors at different wavelengths are assumed to be uncorrelated, which leads to a diagonal covariance matrix. The modeling error describes our inability to model limb scatter observations perfectly, mainly due to multiple scattering in the atmosphere. The modeling error is estimated as a function of wavelength and altitude using the Monte Carlo radiative transfer model Siro [*Oikarinen et al.*, 1999].

The fitting problem (3.5) is solved using an initial guess for the densities and the Levenberg-Marquardt algorithm [*Levenberg*, 1944; *Marquardt*, 1963; *Gill et al.*, 1981] to find the best fit. The Levenberg-Marquardt algorithm is commonly used in non-linear curve fitting problems. It finds the minimum of (3.5) by combining techniques of gradient descent and the inverse-Hessian optimization. The algorithm also provides error estimates for the fitted parameter values. *A posteriori* distributions of the MOP method and the correlation between individual constituents has been studied with the Monte Carlo Markov chain (MCMC) method by *Auvinen et. al.* [1999]. The MCMC studies provide useful information for algorithm development research, but the computational cost is much larger compared to the operational inversion algorithm.

The residuals of the fit increase at lower altitudes, but generally, a good consistency can be found between the model and the measurement; see [**II**]. Completely flawless agreement is very difficult to achieve because the wavelength band used is relatively wide (over 400 nm) and the radiance is governed by numerous wavelength-dependent phenomena.

Limb scatter measurements include relatively low noise, but the model is unable to describe observations perfectly. Because the modelling of the atmosphere is a very complex problem indeed, it is even evident that the forward model lacks some processes, or that they have been taken into account in a too simplified way. These factors can be, for example, a missing constituent, incorrect crosssections (e.g. an uncertain temperature profile), too simplified an aerosol model, or an incorrect albedo model. Furthermore, insufficient modelling of the diurnal effects and errors due to geometry simplifications may contribute to the inversion results. The (current) forward model presents our best understanding of the physics behind the observations. The clear structures in the residuals indicate that there is still work to do in order to improve the model in the future.

3.4   Improved NO$_2$ retrieval of the original MOP method

The original idea, introduced in Publ. [**I**], was to use the whole spectral range of OSIRIS and retrieve all of the desired trace gas densities simultaneously. However, this approach seemed to work only with simulated data. When inverting real OSIRIS data, the NO$_2$ retrievals were usually of poor quality although the inverted ozone profiles were proper. The original inversion method often produced a bias of several hundred percent to the NO$_2$ profiles below 30 km compared with the results from the other OSIRIS retrieval algorithms or NO$_2$ measurements by other instruments [*Tukiainen*, 2006].

Using the whole spectrum for the fit has both advantages and disadvantages. When a large wavelength band is used, the information content can naturally be maximized, but on the other hand the modelling issues become more critical. Since ozone is a strong absorber, minor modeling uncertainties are not as crucial for ozone retrieval as they are for minor absorbers such as NO$_2$. As there are uncertainties in the OSIRIS modeling, we noticed that the NO$_2$ retrieval benefits from using a shorter wavelength band where the signal to noise ratio is more optimal for the NO$_2$ retrieval. Also the band should be short enough that the wavelength dependent modeling errors are not dominating the retrieval.

In the present version of MOP, described in Publ. [**II**], the retrieval of major absorbers and scatterers (ozone, air, aerosols) is separated from the retrieval of minor ones (NO$_2$, and others). For NO$_2$, the wavelength band of $430-450$ nm is selected due to a strong NO$_2$ absorption fingerprint in this region. Furthermore, ozone as well as other species absorb weakly in this band, and the region is also free of strong Fraunhofer lines. In theory, the Fraunhofer lines should be canceled out when we apply Eq. (3.1), but a small residual may still remain. This is the tilt effect recognized by *Sioris et. al.* [2003]. It basically arises from the different spectral slopes of the radiances at different tangent heights and from the finite spectral resolution of the instrument. Thus it is safest to avoid strong Fraunhofer lines where the effect is largest. In this wavelength band, all of the 50 available wavelengths were used in the spectral fitting.

Because of the distinct fitting windows for the strong and weak absorbers, we have to run two separate peeling loops. During the first peeling loop, we retrieve only ozone, aerosol, and neutral air. The NO$_2$ profile is summoned from a climatology (U.S. Standard Atmosphere, 1976). During the second loop, we retrieve NO$_2$. The ozone profile, retrieved from the first peeling loop, is now fixed, but aerosols and neutral air are again allowed to vary freely to obtain a good NO$_2$ fit.

This kind of iterative solving of the parameters is justified, because the use of the fixed NO$_2$ profile in the first peeling loop has little effect on the outcomes. Nevertheless, it is not totally insignificant. The use of the NO$_2$ climatology in the first peeling loop seems to result in roughly a 2% bias at the ozone peak. It is possible to remove this bias by adding a third peeling loop and retrieve ozone, neutral air, and aerosols again with the retrieved NO$_2$ profile, but this would double the required computing time. The third peeling loop is possible to implement later in the future, if more computing resources are provided to the operative OSIRIS processing.

3.5   Inversion results of the stratospheric ozone

The comparison between GOMOS instrument nighttime and OSIRIS daytime ozone profiles indicates a very good agreement, as reported in Publ. [**II**]. This demostrates that both measurement techniques and inversion approaches are valid and strengthen each other, since the instruments together provide daytime and nighttime measurements. Figure 3.1 illustrates comparison results of the measurements during the year 2003 over the northern hemisphere between the latitudes $30 - 60$ N. The coincidence criterion for the individual matches in time was less than 24 h, in the latitude direction less than $1°$, and in the longitude direction less than $2°$.
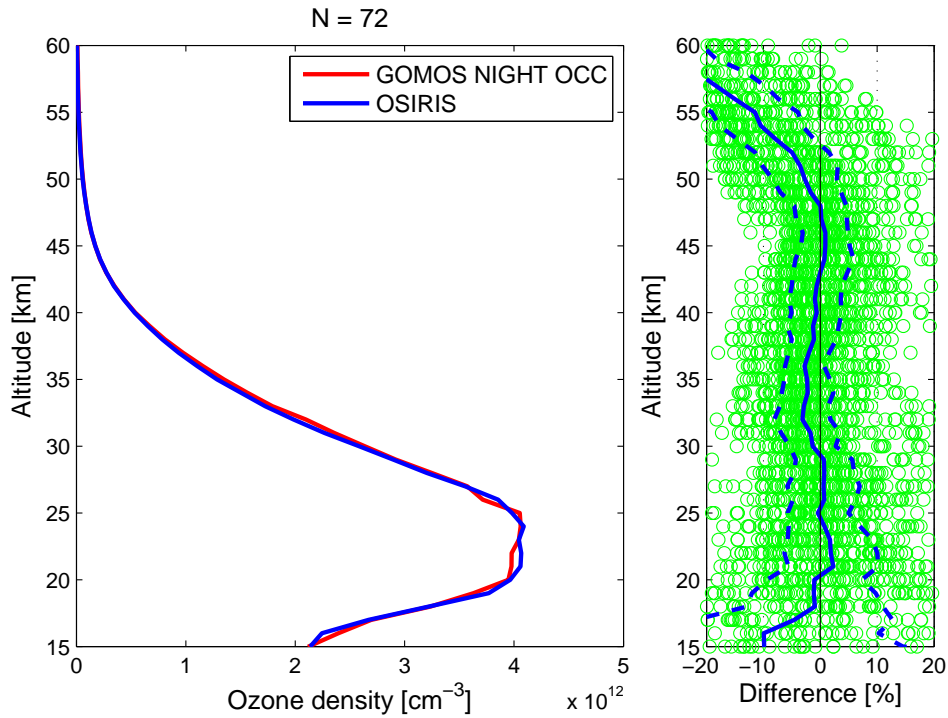
Figure 3.1:   Comparison of retrieved ozone profiles of the GOMOS and OSIRIS measurements. The blue and red curves on the left panel present the medians of 72 coincidences. The blue solid curve on the right panel is the median of the individual relative differences defined as 100%(OSIRIS-GOMOS)/GOMOS. The blue dashed lines in right panel around the median present the semi-interquartile deviation (SID), where the area between the ±1 SID lines includes 50% of the data points. Courtesy of Simo Tukiainen, the principal author of the article **II**, FMI.

Figure 3.2 contains two individual samples of retrieved ozone concentrations as a function of the latitude and altitude from different locations and times. The upper panel illustrates an inverted ozone distribution along the flight path of OSIRIS on the 26th of October, 2001. This sample of results contains 20 scans. The measurements are analyzed with the original version of the MOP method, introduced in [**I**]. The ozone hole is located over the Antarctica between lat -47 and lat -82. The

values below 15 km are less reliable, since the uncertainty of the model increases rapidly after this limit. This is also indicated by the error estimates of the MOP method. The lower panel shows typical variation of the ozone layer altitude as a function of latitude. This set of measurements contains 24 scans. The measurements are made on the 29th of January, 2009, under normal conditions of stratospheric ozone. At mid latitudes, the peak values of the ozone layer are around 25 km; when moving towards the south, the altitude of the peak decreases. The measurements are processed with the updated version of the MOP method [**II**]. The OSIRIS instrument onboard the Odin satellite continues measuring the atmosphere after eight successful years.
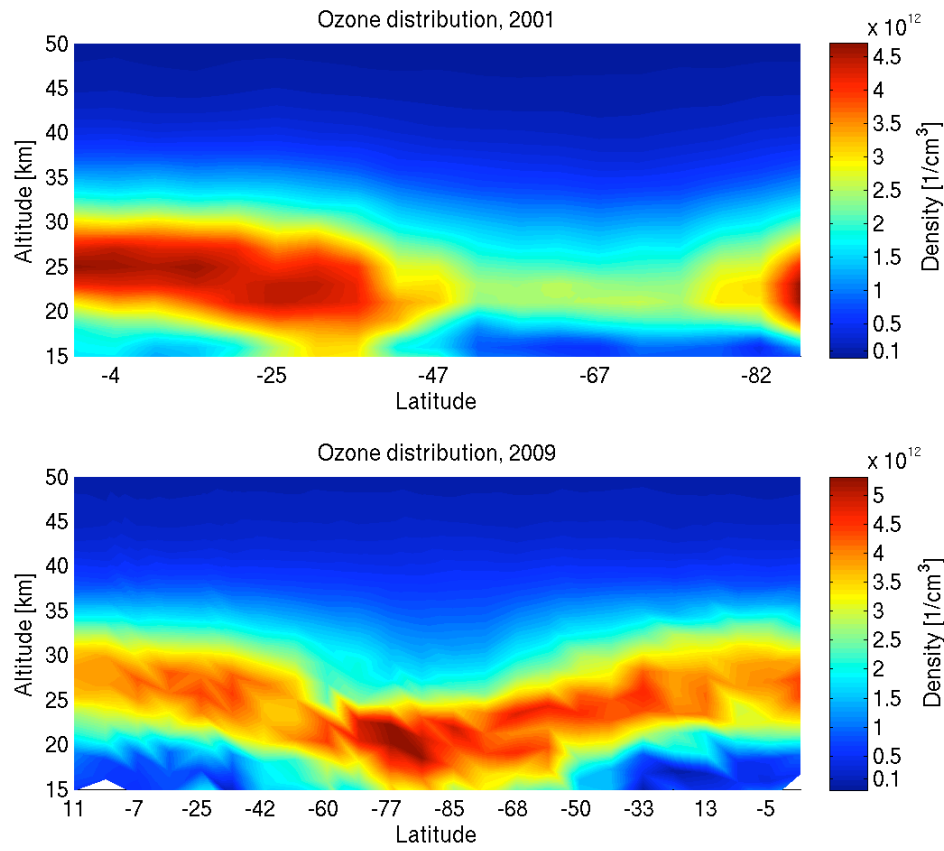
Figure 3.2: The upper panel illustrates inverted ozone distribution using OSIRIS measurements on the 26th of October, 2001, over the Antarctica. The ozone hole is located between lat -47 and lat -82. The lower panel shows typical variation of the ozone layer altitude as a function of latitude. The measurements are made on the 29th of January, 2009, under normal conditions of stratospheric ozone.

New assimilation methods: VKF, VKS and LBFGS-KF

In the field of state space estimation and data assimilation, the Kalman filter (KF) and the extended Kalman filter (EKF) are among the most reliable methods used. However, KF and EKF require the storage of, and operations with, matrices of the size $n \times n$, where n is the dimension of the state space. Furthermore, both methods include inversion operations for $m \times m$ matrices, where m is the dimension of the observation space. Due to this fact, there are several interesting application areas, such as ozone satellite data assimilation, where the standard formulation of KF or EKF is impractical to implement.

Various versions of KF and EKF have been proposed to reduce their computational complexity for large-dimensional problems. The Reduced Rank Kalman filter or Reduced Order extended Kalman filter [see e.g. *Dee D.P.*, 1990, *Cane et. al.*, 1996, *Voutilainen et. al.*, 2007, *Fisher M.*, 1998, *Gejadze et. al.*, 2008, *Tian et. al.*, 2008, *Veersé et al.*, 2000] project the dynamical state vector of the model onto a low dimensional subspace. The success of the approach depends on a judicious choice of the reduction operator. Moreover, since the reduction operator is typically fixed in time, the dynamics of the system may not be correctly captured; for more details, see [*Michael Fisher and Erik Andersson*, 2001].

There exist various Ensemble Kalman filter (EnKF) algorithms—first proposed by *Evensen* [1994]—that are widely used in the field of data assimilation. The idea behind these methods is to form an ensemble of state vectors that represent the state estimate covariance. Each of the members of the ensemble is then propagated forward in time by the full nonlinear evolution model in order to approximate component-wise covariances of the prediction error. EnKF can be used in large-scale data assimilation problems because it is highly parallelizable.

In Publ. **IV**, we have shown how high dimensional KF and EKF may be carried out approximatively using the limited memory Broyden-Flether-Goldfarb-Shanno (LBFGS) optimization algorithm. The resulting methods are effective and exhibit low storage and computational cost characteristics. In Publ. **III**, we introduced an alternative approximation, the variational Kalman filter (VKF), for KF and EKF. Furthermore in Publ. **III**, we introduced a variational Kalman smoother (VKS) method to approximate the fixed-lag Kalman smoother (FLKS) method. In the variational approach, we solve an equivalent maximum *a posteriori* optimization problem using LBFGS, which replaces the explicit computation and use of the Kalman gain matrix in order to obtain state estimates and covariance approximations.

The idea of using the LBFGS method in variational data assimilation is not new, see e.g. [*Yang et. al*, 1996, *Michael Fisher and Philippe Courtier*, 1995, *Veersé, F.*, 1999a, *Veersé, F.*, 1999b, *Veersé et al.*, 2000, *Gejadze et. al.*, 2008, *Tian et. al.*, 2008]. In many of these references, the LBFGS Hessian or inverse Hessian is used as a preconditioner, and even as an approximate error covariance matrix for the background term in variational data assimilation. However, in the VKF method presented here, the LBFGS method is further used in order to propagate effectively the state estimate covariance information forward in time.

4.1   The Kalman filter

Consider the following coupled system of discrete, linear, stochastic difference equations

$$\mathbf{x}_k = \mathbf{M}_k \mathbf{x}_{k-1} + \varepsilon_k^p, \tag{4.1}$$

$$\mathbf{y}_k = \mathbf{K}_k \mathbf{x}_k + \varepsilon_k^o. \tag{4.2}$$

In the first equation, $\mathbf{x}_k$ denotes the $n \times 1$ state vector of the system at time k; $\mathbf{M}_k$ is the $n \times n$ linear evolution operator; and ε_k^p is an $n \times 1$ random vector representing the prediction error and is assumed to characterize errors in the model and in the corresponding numerical approximations. In the second equation, $\mathbf{y}_k$ denotes the $m \times 1$ observed data vector; $\mathbf{K}_k$ is the $m \times n$ linear observation operator; and ε_k^o is an $m \times 1$ random vector representing the observation error. The error terms are assumed to be independent and normally distributed with a zero mean and with covariance matrices $\mathbf{C}_{\varepsilon_k^p}$ and $\mathbf{C}_{\varepsilon_k^o}$, respectively.

The task is to estimate $\mathbf{x}_k$ and its error covariance $\mathbf{C}_k$ at time point k given $\mathbf{y}_k$, $\mathbf{K}_k$, ε_k^o, $\mathbf{M}_k$, ε_k^p and estimates $\mathbf{x}_{k-1}^{est}$ and $\mathbf{C}_{k-1}^{est}$ of the state and covariance at time point $k-1$. The Kalman filter is the standard approach for such problems. It has the form

The Kalman filter algorithm

Step 0: Select initial guess $\mathbf{x}_0^{est}$ and covariance $\mathbf{C}_0^{est}$, and set $k = 1$.
Step 1: Compute the evolution model estimate and covariance:
  (i) Compute $\mathbf{x}_k^p = \mathbf{M}_k \mathbf{x}_{k-1}^{est}$;
  (ii) Compute $\mathbf{C}_k^p = \mathbf{M}_k \mathbf{C}_{k-1}^{est} \mathbf{M}_k^{\mathrm{T}} + \mathbf{C}_{\varepsilon_k^p}$.
Step 2: Compute Kalman filter estimate and covariance:
  (i) Compute the Kalman Gain $\mathbf{G}_k = \mathbf{C}_k^p \mathbf{K}_k^{\mathrm{T}} (\mathbf{K}_k \mathbf{C}_k^p \mathbf{K}_k^{\mathrm{T}} + \mathbf{C}_{\varepsilon_k^o})^{-1}$;
  (ii) Compute the Kalman filter estimate $\mathbf{x}_k^{est} = \mathbf{x}_k^p + \mathbf{G}_k (\mathbf{y}_k - \mathbf{K}_k \mathbf{x}_k^p)$;
  (iii) Compute the estimate covariance $\mathbf{C}_k^{est} = \mathbf{C}_k^p - \mathbf{G}_k \mathbf{K}_k \mathbf{C}_k^p$.
Step 3: Update $k := k + 1$ and return to Step 1.

A nonlinear extension of KF, known as the extended Kalman filter (EKF), is obtained when (4.1), (4.2) are replaced by

$$\mathbf{x}_k = \mathcal{M}(\mathbf{x}_{k-1}) + \varepsilon_k^p, \tag{4.3}$$

$$\mathbf{y}_k = \mathcal{K}(\mathbf{x}_k) + \varepsilon_k^o, \tag{4.4}$$

where $\mathcal{M}$ and $\mathcal{K}$ are possibly nonlinear functions. EKF is obtained by the following modification of the KF algorithm: in Step 1, (i) use the nonlinear model $\mathbf{x}_k^p = \mathcal{M}(\mathbf{x}_k^{est})$ to compute the prior, but

employ the linearized approximations,

$$\mathbf{M}_k = \frac{\partial \mathcal{M}(\mathbf{x}_{k-1}^{est})}{\partial \mathbf{x}} \quad \text{and} \quad \mathbf{K}_k = \frac{\partial \mathcal{K}(\mathbf{x}_k^p)}{\partial \mathbf{x}}, \tag{4.5}$$

for the covariance calculations, and otherwise employ the same formulae as above.

We note that $\mathbf{M}_k$ and $\mathbf{K}_k$ can be computed or estimated in a number of ways. For example, the numerical scheme that is used in the solution of either the evolution or the observation model defines a tangent linear code (see, e.g., [*F.-X. LeDimet and O. Talagrand*, 1986]), which can be used to compute (4.5). A common, but also more computationally expensive approach is to use finite differences to approximate (4.5).

The Kalman filter is expensive to implement due to the fact that it is necessary to store $n \times n$ matrices and invert $m \times m$ matrices at each step. Our task is to overcome these limitations. We make the assumptions that multiplications by the evolution and observation matrices $\mathbf{M}_k$ and $\mathbf{K}_k$ and by the covariance matrices $\mathbf{C}_{\varepsilon_k^p}$ and $\mathbf{C}_{\varepsilon_k^o}$ are efficient, both in terms of storage and CPU time. Additional computational challenges arise for a sufficiently large n due to the storage requirements for $\mathbf{C}_k^{est}$, which becomes a full matrix as the filter proceeds in time. The same is also true for $\mathbf{C}_k^p$. However, given that

$$\mathbf{C}_k^p = \mathbf{M}_k \mathbf{C}_k^{est} \mathbf{M}_k^{\mathrm{T}} + \mathbf{C}_{\varepsilon_k^p}, \tag{4.6}$$

storage issues are restricted to those for $\mathbf{C}_k^{est}$; typically the matrix $\mathbf{C}_{\varepsilon_k^p}$ is assumed to be diagonal.

4.2 The limited memory BFGS Kalman filter (LBFGS-KF) method

It is well known that the BFGS algorithm provides an approximation for both the Hessian and the inverse Hessian of a cost function. We use this fact to approximate matrix operations needed in KF/EKF.

A quadratic optimization problem can be formulated as:

$$\arg\min_{\mathbf{u}} \frac{1}{2} \langle \mathbf{A}\mathbf{u}, \mathbf{u} \rangle - \langle \mathbf{b}, \mathbf{u} \rangle, \tag{4.7}$$

where the given variables are matrix $\mathbf{A}$ and vector $\mathbf{b}$.

First we apply limited memory Broyden-Flether-Goldfarb-Shanno (LBFGS) to make the Kalman filter more efficient and describe LBFGS-KF method. The idea is to follow Kalman formulas, but replace the matrix inversion operations with artificial optimization problems.

A low storage approximation of $\mathbf{C}_k^{est}$ can be obtained by applying the LBFGS algorithm to the problem of minimizing (4.7) with $\mathbf{A} = \mathbf{C}_k^{est}$ and $\mathbf{b} = \mathbf{0}$. The LBFGS matrix $\mathbf{B}_\nu^{-1}$ is then a low storage approximation of $(\mathbf{C}_k^{est})^{-1}$ and formulas for $\mathbf{C}_k^{est}$ from [*Byrd et. al.*, 1994] can be used.

Additionally, when m is sufficiently large, the computation of $(\mathbf{K}_k \mathbf{C}_k^p \mathbf{K}_k^{\mathrm{T}} + \mathbf{C}_{\varepsilon_k^o})^{-1}(\mathbf{y}_k - \mathbf{K}_k \mathbf{x}_k^p)$ that is required in Step 2, (ii) of the Kalman filter iteration will be prohibitively expensive. For the approximation of $(\mathbf{K}_k \mathbf{C}_k^p \mathbf{K}_k^{\mathrm{T}} + \mathbf{C}_{\varepsilon_k^o})^{-1}(\mathbf{y}_k - \mathbf{K}_k \mathbf{x}_k^p)$, we set $\mathbf{A} = \mathbf{K}_k \mathbf{C}_k^p \mathbf{K}_k^{\mathrm{T}} + \mathbf{C}_{\varepsilon_k^o}$ and $\mathbf{b} = \mathbf{y}_k - \mathbf{K}_k \mathbf{x}_k^p$ in (4.7) and apply the LBFGS algorithm to the problem of minimizing (4.7).

The LBFGS Kalman filter method can now be presented as the following algorithm:

The LBFGS Kalman filter (LBFGS-KF)

Step 0: Select initial guess $\mathbf{x}_0^{est}$ and covariance $\mathbf{B}_\# = \mathbf{C}_0^{est}$, and set $k = 0$.
Step 1: Compute the evolution model estimate and covariance:
   (i) Compute $\mathbf{x}_k^p = \mathbf{M}_k \mathbf{x}_k^{est}$;
   (ii) Define $\mathbf{C}_k^p = \mathbf{M}_k \mathbf{B}_\# \mathbf{M}_k^{\mathrm{T}} + \mathbf{C}_{\boldsymbol{\varepsilon}_k^p}$.
Step 2: Compute the Kalman filter estimate and covariance:
   (i) Define $\mathbf{A} = (\mathbf{K}_k \mathbf{C}_k^p \mathbf{K}_k^{\mathrm{T}} + \mathbf{C}_{\boldsymbol{\varepsilon}_k^o})$ and $\mathbf{b} = \mathbf{y}_k - \mathbf{K}_k \mathbf{x}_k^p$ in (4.7) and compute
   the LBFGS approximations $\mathbf{B}_*$ of $\mathbf{A}^{-1}$ and $\mathbf{u}_*$ of $\mathbf{A}^{-1}\mathbf{b}$.
   (ii) Compute the LBFGS-KF estimate $\mathbf{x}_{k+1}^{est} = \mathbf{x}_k^p + \mathbf{C}_k^p \mathbf{K}_k^{\mathrm{T}} \mathbf{u}_*$;
   (iii) Define $\mathbf{A} = \mathbf{C}_k^p - \mathbf{C}_k^p \mathbf{K}_k^{\mathrm{T}} \mathbf{B}_* \mathbf{K}_k \mathbf{C}_k^p (\approx \mathbf{C}_{k+1}^{est})$ and $\mathbf{b} = \mathbf{0}$ in (4.7) and compute
   the LBFGS approximation $\mathbf{B}_\#$ of $\mathbf{C}_{k+1}^{est}$.
Step 3: Update $k := k + 1$ and return to Step 1.

All operations with the $\mathbf{C}_k^{est}$ and $\mathbf{A}^{-1}$ are done using the LBFGS formulas. As a result, LBFGS-KF is much less memory and computationally intensive than KF making its use on large-scale problems more feasible. Specifically, the storage requirements for the LBFGS estimate of $\mathbf{C}_k^{est}$ are on the order of $2n\ell + 4n$, where ℓ is the number of stored vectors in LBFGS (typically $10-50$), rather than $n^2 + 4n$ [*Jorge Nocedal and Stephen Wright*, 1999], and the computational cost for both obtaining and using this estimate is the order n. Furthermore, the inversion of the $m \times m$ matrix $\mathbf{K}_k \mathbf{C}_k^p \mathbf{K}_k^{\mathrm{T}} + \mathbf{C}_{\boldsymbol{\varepsilon}_k^o}$ is carried out in order m operations and its storage requirements are on the order of $2m\ell + 4m$ rather than $m^2 + 4m$ [*Jorge Nocedal and Stephen Wright*, 1999].

In the first example considered in the Publ. **IV**, LBFGS-KF and KF are compared and it is noted that LBFGS-KF is roughly 10 times faster, in terms of CPU time, than KF when applied to the same problem. Moreover, using our MATLAB implementation, LBFGS-KF can be used on significantly larger-scale problems.

As we have mentioned, in our implementations of KF and LBFGS-KF, the covariance matrices $\mathbf{C}_{\boldsymbol{\varepsilon}_k^p}$ and $\mathbf{C}_{\boldsymbol{\varepsilon}_k^o}$ are taken to be diagonal. This is not a necessary requirement. More structured covariances can be used, containing important *a priori* information [*Lorenc A. C.*, 2003]. However, in order to maintain the computational efficiency and low storage requirements of LBFGS-KF, $\mathbf{C}_{\boldsymbol{\varepsilon}_k^p}$ and $\mathbf{C}_{\boldsymbol{\varepsilon}_k^o}$ must be comparable to $\mathbf{M}_k$, $\mathbf{B}_\#$ and $\mathbf{K}_k$, $\mathbf{B}_*$, respectively, in terms storage requirements and the computational cost required for their multiplication.

4.3   The variational Kalman filter method

Bayes' Theorem can be used to formulate the Kalman filter as a sequential maximum *a posteriori* estimation. To see this, we recall Bayes' formula

$$p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y}) = \frac{p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})p_{\mathbf{x}}(\mathbf{x})}{p_{\mathbf{y}}(\mathbf{y})}, \tag{4.8}$$

where $\mathbf{x}$ is the vector of unknowns, $\mathbf{y}$ the measurements, $p_{\mathbf{x}}$ denotes the prior density, and $p_{\mathbf{y}|\mathbf{x}}$ is the density of the likelihood function. The maximum *a posteriori* (MAP) estimate is obtained by maximizing (4.8). Equivalently, one can minimize

$$\ell(\mathbf{x}|\mathbf{y}) := -\log p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}) - \log p_{\mathbf{x}}(\mathbf{x}). \tag{4.9}$$

For the linear model (4.2) at time k, the function ℓ assumes the form

$$\ell(\mathbf{x}|\mathbf{y}_k) = \frac{1}{2}(\mathbf{y}_k - \mathbf{K}_k\mathbf{x})^{\mathrm{T}}\mathbf{C}_{\boldsymbol{\varepsilon}_k^o}^{-1}(\mathbf{y}_k - \mathbf{K}_k\mathbf{x}) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_k^p)^{\mathrm{T}}(\mathbf{C}_k^p)^{-1}(\mathbf{x} - \mathbf{x}_k^p), \qquad (4.10)$$

where $\mathbf{C}_{\boldsymbol{\varepsilon}_k^o}$ and $\mathbf{C}_k^p$ are the covariance matrices of the measurement noise $\boldsymbol{\varepsilon}_k^o$ and of the prior $\mathbf{x}_k^p$, respectively. The Kalman filter estimate and its covariance $\mathbf{x}_k^{est}$ and $\mathbf{C}_k^{est}$ are precisely the minimizer and inverse Hessian of $\ell(\mathbf{x}|\mathbf{y}_k)$, respectively.

The advantage of the variational Kalman filter is that it uses the use of an optimization algorithm to minimize $\ell(\mathbf{x}|\mathbf{y}_k)$ in (4.10), together with computing the inverse of *a priori* covariance for the next time step. For large-scale problems, this can be very advantageous.

In particular, we advocate using the limited memory BFGS algorithm (LBFGS) for the minimization problem (4.9). Given specific choices for initial guess, stopping criteria, and number of stored vectors, LBFGS will yield estimates of both $\mathbf{x}_k^{est}$ and $\mathbf{C}_k^{est}$. The storage requirement for the covariance approximation—which we denote by $\mathbf{B}_k^{\#}$—is $2rn$, where r is the number of stored LBFGS vectors (typically on the order of 10), and multiplication by $\mathbf{B}_k$ is order n. So LBFGS provides the minimum and the inverse Hessian of $\ell(\mathbf{x}|\mathbf{y}_k)$ in (4.10).

However, it is $(\mathbf{C}_k^p)^{-1}$ that is needed in the optimization problem in the next VKF iteration (see (4.10)). This problem is solved by using the same approximation $\mathbf{C}_k^P = \mathbf{M}_k\mathbf{C}_{k-1}^{est}\mathbf{M}_k^{\mathrm{T}} + \mathbf{C}_{\boldsymbol{\varepsilon}_k^p}$ as in EKF, but $(\mathbf{C}_k^p)^{-1}$ is computed directly by LBFGS. For this purpose we apply LBFGS a second time to an auxiliary optimization problem (4.7), where $\mathbf{A} = \mathbf{M}_k\mathbf{B}_{k-1}^{\#}\mathbf{M}_k^{\mathrm{T}} + \mathbf{C}_{\boldsymbol{\varepsilon}_k^p}$ and $\mathbf{b}$ is the zero vector. This gives an approximation $\mathbf{B}_k^*$ of $(\mathbf{C}_k^p)^{-1}$.

The variational Kalman filter algorithm is summarized as follows:

The variational Kalman filter algorithm

    Step 0: Select initial guess $\mathbf{x}_0^{\#}$ and covariance $\mathbf{B}_0^{\#} = \mathbf{C}_0^{est}$, and set $k = 1$.
    Step 1: Compute the evolution model estimate and covariance:
        (i) Compute $\mathbf{x}_k^p = \mathbf{M}_k\mathbf{x}_{k-1}^{\#}$;
        (ii) Define $\mathbf{C}_k^p = \mathbf{M}_k\mathbf{B}_{k-1}^{\#}\mathbf{M}_k^{\mathrm{T}} + \mathbf{C}_{\boldsymbol{\varepsilon}_k^p}$;
        (iii) Compute LBFGS approximation $\mathbf{B}_k^*$ of $(\mathbf{C}_k^p)^{-1}$;
    Step 2: Compute variational Kalman filter and covariance estimates:
        (i) Minimize $\ell(\mathbf{x}|\mathbf{y}_k) = (\mathbf{y}_k - \mathbf{K}_k\mathbf{x})^{\mathrm{T}}(\mathbf{C}_{\boldsymbol{\varepsilon}_k^o})^{-1}(\mathbf{y}_k - \mathbf{K}_k\mathbf{x}) + (\mathbf{x} - \mathbf{x}_k^p)^{\mathrm{T}}\mathbf{B}_k^*(\mathbf{x} - \mathbf{x}_k^p)$
        using LBFGS, and define $\mathbf{x}_k^{\#}$ and $\mathbf{B}_k^{\#}$ to be the LBFGS minimizer and
        inverse Hessian approximations;
    Step 3: Update $k := k + 1$ and return to Step 1.

Note that in Step 1, (ii) and Step 2, (i) the optimizations are quadratic and therefore only quadratic LBFGS is needed. For practical applications, a judicious choice of the initial inverse Hessian is needed in order to obtain accurate results efficiently. In the numerical examples of this work, we have used $\mathbf{B}_0^{-1} = \beta\mathbf{I}$ with β chosen so that $\beta\mathbf{I}$ approximates the diagonal of the covariance matrix of interest. For more discussion on the choice of $\mathbf{B}_0^{-1}$ (the preconditioner) see [*Jorge Nocedal and Stephen Wright*, 1999, *Veersé et al.*, 2000].

4.4   The nonlinear variational Kalman filter method

As in the case of EKF, we need a linearization to propagate the covariance information from one observation time to the next. However, the direct linearization as in EKF is impractical for large dimensions. Rather, in the case of a non-linear evolution model we should use the adjoint operator, if available, in Step 1, (ii) of VKF. Furthermore, if the adjoint operator is coded in an implicit form, *i.e.* in the software of the model, we get full benefit from a limited memory presentation of $\mathbf{C}_k^{est}$. This is the situation in many operational codes for weather forecasting.

Supposing that the linearization $\mathbf{M}_k$ of $\mathcal{M}_k$ is available, then Step 2, (i) can written in the VKF algorithm as:

Step 2: Compute the variational Kalman filter estimate and covariance:
(i) Minimize $\ell(\mathbf{x}|\mathbf{y}_k) = (\mathbf{y}_k - \mathcal{K}(\mathbf{x}))^{\mathrm{T}}(\mathbf{C}_{\boldsymbol{\varepsilon}_k^o})^{-1}(\mathbf{y}_k - \mathcal{K}(\mathbf{x})) + (\mathbf{x} - \mathbf{x}_k^p)^{\mathrm{T}}\mathbf{B}_k^*(\mathbf{x} - \mathbf{x}_k^p)$
using LBFGS, and define $\mathbf{x}_k^{\#}$ and $\mathbf{B}_k^{\#}$ to be the LBFGS minimizer and inverse Hessian approximations;

Especially, if the *tangent linear* $\mathbf{M}_k^{tl}$ and corresponding *adjoint code* $\mathbf{M}_k^*$ [*F.-X. LeDimet and O. Talagrand*, 1986] are available for the evolution model $\mathcal{M}$, Step 1, (ii) can be written as:

Step 1: Compute the evolution model estimate and covariance:
(ii) Define $\mathbf{C}_k^p = \mathbf{M}_k^{tl}\mathbf{B}_{k-1}^{\#}\mathbf{M}_k^* + \mathbf{C}_{\boldsymbol{\varepsilon}_k^p}$;

This feature of the method is one of the major advantages compared to EKF, since the time consuming linearization of the evolution model can be avoided. The computational cost of a single tangent linear and adjoint evaluations are both roughly double the cost of a single nonlinear evolution model evaluation. Inside VKF the inversion of the above $\mathbf{C}_k^p$ matrix requires around $15-60$ tangent linear and adjoint code evaluations. So, for large n, the total cost is a fraction of the traditional finite difference evaluation that requires n model evaluations.

Similar possibilities exist for use the tangent linear code $\mathbf{K}_k^{tl}$ of the observation model $\mathcal{K}$ in Step 2, (i):

Step 2: Compute the evolution model estimate and covariance:
(i) Minimize $\ell(\mathbf{x}|\mathbf{y}_k) = (\mathbf{y}_k - \mathbf{K}_k^{tl}\mathbf{x})^{\mathrm{T}}(\mathbf{C}_{\boldsymbol{\varepsilon}_k^o})^{-1}(\mathbf{y}_k - \mathbf{K}_k^{tl}\mathbf{x}) + (\mathbf{x} - \mathbf{x}_k^p)^{\mathrm{T}}\mathbf{B}_k^*(\mathbf{x} - \mathbf{x}_k^p)$
using LBFGS, and define $\mathbf{x}_k^{\#}$ and $\mathbf{B}_k^{\#}$ to be the LBFGS minimizer and inverse Hessian approximations;

In EKF the computational advantage of using $\mathbf{K}_k^{tl}$ instead is lost, since the Kalman gain is typically a full matrix.

4.5   The variational Kalman smoother method

Next we introduce a variational Kalman smoother (VKS) method, which can be used afterwards to smooth the results of the variational Kalman filter. The idea is to simulate a fixed-lag Kalman smoother (FLKS) method and take full benefit from the limited memory covariance approximation form of the VKF method. In general, such post-processing improves the quality of the VKF results.

The VKF method provides an estimate $\mathbf{x}_k^{est}$ and a corresponding limited memory approximation of the covariance matrix $\mathbf{C}_k^{est}$ after each time step k. In VKS, we use these results from the previous $[k_0, k_0+1, ..., k]$ time steps, where the parameter $k_0 = k - lag$ determines the length of the time interval. In case of linear evolution model $\mathbf{M}_k$, we couple the results together by using the following cost function:

$$J(\mathbf{x}_{k_0}) = \sum_{t=k_0}^{k} (\mathbf{M}_t \mathbf{x}_{k_0} - \mathbf{x}_t^{est})^{\mathrm{T}} (\mathbf{C}_t^{est})^{-1} (\mathbf{M}_t \mathbf{x}_{k_0} - \mathbf{x}_t^{est}), \tag{4.11}$$

where $\mathbf{M}_t \mathbf{x}_{k_0}$ is a model trajectory from $\mathbf{x}_{k_0}$. The minimization of the cost function is done by using the 4d-Var method (see [*Derber J.*, 1989; *F.-X. LeDimet and O. Talagrand*, 1986]), using LBFGS.

In the nonlinear case, the evolution model $\mathcal{M}_t$ is used instead of $\mathbf{M}_t$ in the cost function formulation (4.11). Furthermore, the gradient of (4.11) can be computed efficiently by using the adjoint of the evolution model, but in principle, the linearization of $\mathcal{M}_t$ can be used again as well. Since the smoothing process improves the accuracy of the estimate at time k_0, it is possible to outperform EKF in retrospective analysis.

During VKF iterations, the inverse Hessian limited memory BFGS formula is used to represent $\mathbf{C}_k^{est}$. In the VKS cost function (4.11) we instead need the inverse of $\mathbf{C}_t^{est}$. In practice this detail is handled by using the direct Hessian limited memory BFGS formula (see, e.g., [*Jorge Nocedal and Stephen Wright*, 1999]). The direct Hessian limited memory BFGS formula provides the $(\mathbf{C}_t^{est})^{-1}$ required.

## 4.6   Monitoring the quality of the BFGS approximation

The quality of approximations produced by Quasi-Newton methods to Hessian matrices, such as the covariance matrices in our case, has been studied at least since 1970 [*Greenstadt J.*, 1970]. Typically this is carried out by monitoring the matrix norm of the difference between an approximation and a known Hessian matrix.

As discussed above, the Kalman filter is basically a statistical procedure, that repeatedly applies the Bayes rule to create the distribution of the state vector. Therefore it is justified to employ the chi-square test to monitor the goodness of the LBFGS updates as approximations of the covariances of known distributions. The chi–square distribution provides a scalar valued test for multinormality: if $\mathbf{x}$ is an n–dimensional Gaussian random vector with zero mean and covariance C, then $\mathbf{x}^{\mathrm{T}} \mathbf{C}^{-1} \mathbf{x} \sim \chi_n^2$. The idea is to monitor how well the approximative covariances fulfill this test, when a sample of vectors $\mathbf{x}$ has been generated using a 'true' covariance matrix.

The true covariance $\mathbf{C}^{true}$ is adopted from the EKF process of the Lorenz95 case (for more details, see [*Lorenz E. N.*, 1996]), with 100 variables. Then we sample sample a set of vectors $\mathbf{x}_i$ from $N(0, \mathbf{C}^{true})$, $i = 1, ..., m$. Next, we compute the approximations $\mathbf{C}_{iter}^{-1}$ of the inverse of the

covariance using the LBFGS optimization with an increasing number of iterations, and calculate $qtest(i) := \mathbf{x}_i^{\mathrm{T}} \mathbf{C}_{iter}^{-1} \mathbf{x}_i$, for every i. These values are compared with the chi-square significance values at, e.g., the $0.5, 0.75$ and 0.95 levels, to see which percentage of the vectors sampled from the true distribution lie inside those confidence regions of the respective approximative multinormal distributions. It is known that for a linear-quadratic problem the BFGS algorithm forms an exact inverse Hessian with full n updates, so the test values are expected to reach the true limits with n iterations. The test are perfomed with both the limited memory BFGS and full BFGS, but with the LBFGS method we stop at the convergence limit, when the stopping criterion has been chosen to be close to machine precision. The full BFGS runs are performed with the LBFGS code, by keeping all search directions in the memory. During LBFGS runs we keep all other search directions in memory except the first one.

In the test examples, we use $m = 1000$ samples and repeat the procedure 10 times to get the mean values of the results.
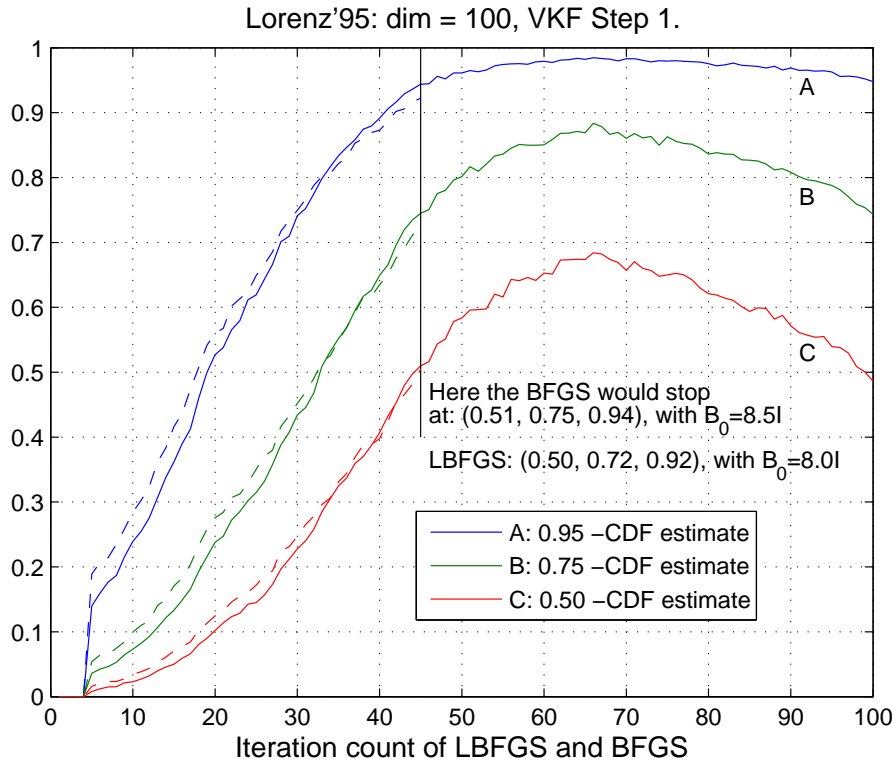
Figure 4.1:   Plot of the chi-square test confidence limit estimates of the and BFGS (-) and LBFGS (- -) as a function of the iteration count.

In the above test we allow the BFGS optimization method to continue iteration beyond the normal stopping point. If the typical stopping criteria are used, the LBFGS method will terminate after

about 45 iterations in these cases, with 100 variables. We also perform a true limited memory chi-square test with LBFGS, where we stop the optimization process at the convergence limit.

The results for the test are given in Figure 4.1. With too few updates, the calculated values are below the chi-square test values, so the approximate covariance gives a too narrow distribution. With increasing updates, the test values are first exceeded, then we can see how the correct values are reached with full updates. The choice of the initial inverse matrix B_0 naturally has an impact on the process. To demonstrate this, we use slightly different B_0 values: $B_0 = 8.5\mathbf{I}$, with BFGS and $B_0 = 8.0\mathbf{I}$, with LBFGS.

Conclusions

The Modified Onion Peeling (MOP) inversion method for OSIRIS limb scatter measurements is introduced and validated in Publ. [**I-II**]. The advantages of the MOP method includes efficient use of measured data, since the whole spectral range of the instrument is utilized to invert the vertical profiles of ozone, air and aerosol densities. In fact, the MOP method is the first proposed inversion method for UV-visible limb scatter measurements that uses a large number of data points simultaneously. Therefore, the retrievals of these constituents may be estimated up to 70 km in altitude. Above this limit, the straylight contamination of the OSIRIS measurements causes bias for the inversion/retrieval results. Without the straylight contamination of the OSIRIS measurements, it could be possible to invert the constituent densities even at higher altitudes.

However, for NO_2 retrieval, it is crucial to use a narrow wavelength band. In principle, this approach can be easily expanded to retrieve other minor trace gases, such as BrO and $OClO$, by selecting another wavelength window and adding an extra peeling iteration. The underlying challenge is that the concentrations of BrO and $OClO$ are an order of magnitude smaller than that of NO_2 and the absorption fingerprints are more easily masked by noise. In the MOP method, the measured limb scatter signal, containing the multiple scattering effect, is modeled efficiently using a single scattering model with a pre-calculated look-up-table database. The forward model used for the inversion must yield fast solutions. However, in order to perform some case specific studies, it is also possible to use a more time consuming forward model, such as the Siro.

Validation against other satellite instruments demonstrated the capability of the MOP inversion method in practice. Good agreement was found between OSIRIS daytime and GOMOS nighttime ozone profiles [Publ. **II**]. The median of the relative individual differences is less than 5% between 21 and 45 km. Above 45 km, the diurnal variation of ozone prevented comparisons between instruments.

In order to combine the information from satellite measurements with an atmospheric model, two new assimilation methods are proposed to approximate the Kalman filter and extended Kalman filter.

The standard implementations of KF and EKF become exceedingly time and memory consuming as the dimension of the underlying state space increases. Several variants of KF and EKF have been proposed to reduce the dimension of the system, thus making implementation in high dimensions possible. The Reduced Rank Kalman filter or Reduced Order extended Kalman filter project the dy-

namical state vector of the model onto a lower dimensional subspace. The success of this approach depends on the choice of the reduction operator. Moreover, since the reduction operator is typically fixed in time, they may exhibit "covariance leaks". This is natural, since a nonlinear system does not generally leave any fixed linear subspace invariant.

The assimilation methods developed in this thesis use the limited memory Broyden-Flether-Goldfarb-Shanno (LBFGS) minimization method in order to circumvent the computational and memory issues of standard KF and EKF. For KF, we simply employ the limited memory optimization techniques for matrix operations, but otherwise use the standard KF formulae. In the VKF variant of EKF, the idea is to seek the minimum of the underlying objective function directly, whiteout explicit use or computation of the Kalman formulae. The variational Kalman filter uses the LBFGS method to compute the state estimate and the corresponding covariance matrix much in the same way as is done in the well known 3D-Var assimilation method. The innovation of VKF is in how we compute the inverse of the *a priori* covariance for the next time step of the filter. This is done by applying LBFGS again to the very same matrix expression as is used in EKF. However, the explicit matrices are never created, but all the operations are carried out by the approximative low memory representations produced by LBFGS. This allows us to carry out filtering in very high dimensional problems. For large-scale problems, such as the assimilation of ozone satellite data, this can be very advantageous.

But nothing comes without a price. There are two main features that might limit the use of VKF. The first is the possible inaccuracy of the limited memory approximation of the inverse covariance matrix. Our results in Publ. **III** show that for typical test cases, good results can be achieved, provided a preconditioning of the LBFGS iterations is implemented. We believe that this will be the situation with real assimilation problems, as well. The second point concerns the use of the linearized EKF formula of the covariance for truly high dimensional situations; in practice, it requires that the linearization is performed using the tangent linear and adjoint codes. These have to be tailor-made, which is often a formidable task. However, in many important application fields, notably in numerical weather forecasting, these codes are already available and can be directly combined with the VKF filtering approach.

Auvinen, H., L. Oikarinen, E. Kyrölä, J. Tamminen and G. W. Leppelmeier, Inversion Algorithms for OSIRIS and GOMOS Bright-Limb Background Term. *ESAMS'99, European Symposium on Atmospheric Measurements from Space*, Vol. WPP-161, pp. 271-276, 1999.

Berk, A., L. S. Bernstein, and D. C. Robertson, MODTRAN: A moderate resolution model for LOWTRAN 7, *AFGL Tech. Rep. F19628-86-C-0079,* 38, Air Force Geophys. Lab., Hanscom Air Force Base, Mass., 1989.

Bertaux, J. L., G. Megie, T. Widemann, E. Chassefiere, R. Pellinen, E. Kyrölä, S. Korpela, and P. Simon, Monitoring of ozone trend by stellar occultations: The GOMOS instrument, *Adv. Space Res., 11*(3), 237-242, 1991.

Bovensmann, H., J. P. Burrows, M. Buchwitz, J. Frerick, S. Noel, V. V. Rozanov, K. V. Chance, and A. P. H. Goede, SCIAMACHY: Mission objectives and measurement modes, *J. Atmos. Sci., 56,* pp. 127–150, 1999.

Byrd, R. H., J. Nocedal, and R. B. Schnabel, *Representations of Quasi-Newton Matrices and Their Use in Limited Memory Methods*, Math. Programming, 63(1-3), pp. 129–156, 1994.

Cane M. A., R. N. Miller, B. Tang, E. C. Hackert and A. J. Busalacchi, *Mapping tropical Pacific sea level: data assimilation via reduced state Kalman filter*. J. Geophys. Res., Vol. 101, pp. 599–617, 1996.

Chapman S., *A theory of upper atmosphere*, Quart. J. Roy. Meteor. Soc., Vol. 3, 103, 1930.

Collins, D. G., W. G. Blättner, M. B. Wells, and H. G. Horak Backward Monte Carlo calculations of the polarization characteristics of the radiation emerging from spherical-shell atmospheres, *Appl. Opt., 11,* pp. 2684–2696, 1972.

Crutzen, P. J., Estimates of possible future ozone reductions from continued use of fluoro-chloro-methanes (CF2Cl2, CFCl3), Geophys. Res. Lett., 1(5), pp. 205–208, 1974.

Dee D. P., *Simplification of the Kalman filter for meteorological data assimilation* Quart. J. Roy. Meteor. Soc., vol. 117, pp. 365–384, 1990.

Derber J.: A variational continuous assimilation technique. *Mon. Weather Rev.* **117** pp. 2437-2446, 1989.

Evensen G., *Sequential data assimilation with a non-linear quasi-geostrophic model using Monte Carlo methods to forecast error statistics*. J. Geophys. Res., Vol. 99, pp. 10143–10162, 1994.

Fisher M., *Development of a simplified Kalman filter*. ECMWF Technical Memorandum 260, 1998.

Michael Fisher and Erik Andersson, *Developments in 4D-Var and Kalman Filtering*, ECMWF Technical Memorandum 347, 2001.

Michael Fisher and Philippe Courtier, *Estimating the covariance matrices of analysis and forecast error in variational data assimilation*, ECMWF Technical Memorandum 220, 1995.

Farman J. C., Gardiner G. and Shanklin J. D., Large losses of total ozone in Antarctica reveal sesonal ClO_x/NO_x interaction, Nature, *315* pp. 207–210, 1985.

Flittner, D. E., P. K. Bhartia, and B. M. Herman, O_3 profiles retrieved from limb scatter measurements: Theory, *Geophys. Res. Lett.*, *27*, 2601–2604, doi:10.1029/1999GL011343, 2000.

Greenstadt J., *Variations on Variable-Metric Methods*. Math. Comput. 24, pp. 1-22, 1970.

Gejadze, I. Yu., F.-X. Le Dimet and V. Shutyaev, *On analysis error covariances in variational data assimilation*. SIAM J. Sci. Comput. 30, pp. 1847-1874, 2008.

Gill, P. R., W. Murray, and M. H. Wright (Eds.), *Practical Optimization*, Academic Press, 1981.

Griffioen, E., and L. Oikarinen, LIMBTRAN: A pseudo three-dimensional radiative transfer model for the limb-viewing imager OSIRIS on the Odin satellite, *J. Geophys. Res.*, *105*(D24), 29,717–29,730, doi:10.1029/2000JD900566, 2000.

Haley, C. S., C. von Savigny, S. Brohede, C. E. Sioris, I. C. McDade, E. J. Llewellyn, and D. P. Murtagh, A comparison of methods for retrieving stratospheric ozone profiles from OSIRIS limb-scatter measurements., *Adv. Space Res.*, *34*, pp. 769–774, 2003.

Haley, C. S., S. M. Brohede, C. E. Sioris, E. Griffioen, D. P. Murtagh, I. C. McDade, P. Eriksson, E. J. Llewellyn, A. Bazureau, and F. Goutail, Retrieval of stratospheric O_3 and NO_2 profiles from Odin Optical Spectrograph and Infrared Imager System (OSIRIS) limb-scattered sunlight measurements., *J. Geophys. Res.*, *109*, doi:10.1029/2004JD004588, 2004.

Herman, B. M., A. Ben-David, and K. J. Thome, Numerical technique for solving the radiative transfer equation for a spherical shell atmosphere, *Appl. Opt., 33,* pp. 1760–1770, 1994.

LeDimet, F.-X. and O. Talagrand: Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects. *Tellus A series,* Vol. 38, pp. 97-110, 1986.

Levenberg, K., A method for the solution of certain problems in least squares, *Quart. Appl. Math.*, *2*, pp. 164–168, 1944.

Llewellyn, E. J., N. D. Lloyd, D. A. Degenstein, R. L. Gattinger, S. V. Petelina, A. E. Bourassa, J. T. Wiensz, E. V. Ivanov, I. C. McDade, B. H. Solheim, J. C. McConnell, C. S. Haley, C. von Savigny, C. E. Sioris, C. A. McLinden, E. Griffioen, J. Kaminski, W. F. J. Evans, E. Puckrin, K. Strong, V. Wehrle, R. H. Hum, D. J. W. Kendall, J. Matsushita, D. P. Murtagh, S. Brohede, J. Stegman, G. Witt, G. Barnes, W. F. Payne, L. Piché, K. Smith, G. Warshaw, D.-L. Deslauniers, P. Marchand, E. H. Richardson, R. A. King, I. Wevers, W. McCreath, E. Kyrölä, L. Oikarinen, G. W. Leppelmeier, H. Auvinen, G. Mègie, A. Hauchecorne, F. Lefèvre, J. de La Nöe, P. Ricaud, U. Frisk, F. Sjoberg, F. von Schéele, and L. Nordh, The OSIRIS instrument on the Odin spacecraft, *Can. J. Phys*, *82*, 411–422, doi:10.1139/P04-005, 2004.

OSIRIS - an application of tomography for absorbed emissions in remote sensing, in *Applications of Photonic Technology 2,* edited by G. A. Lampropoulos and R. A. Lessard, pp. 627–632, Plenum Press, New York, 1997.

Lorenc A. C., *Modelling of error covariances by 4D-Var data assimilation*, Q. J. R. Meterol. Soc. 129, pp. 3167-3182, 2003.

Lorenz E. N., *Predictability: A problem partly solved*, Proc. Seminar on Predicability, Vol. 1, ECMWF, Reading, Berkshire, UK, pp. 1-18, 1996.

Marchuk, G. I., G. A. Mikhailov, M. A. Nazaraliev, R. A. Darbinjan, B. A. Kargin, and B. S. Elepov *The Monte Carlo Methods in Atmospheric Optics,* 208 pp., Springer, New York, 1980.

Marquardt, D., An Algorithm for Least-Squares Estimation of Nonlinear Parameters, *J. Soc. Indust. Appl. Math.*, *11*, pp. 431–44, 1963.

McDade, I. C., K. Strong, C. S. Haley, J. Stegman, D. P. Murtagh, and E. J. Llewellyn, A method for recovering stratospheric minor species densities from the Odin/OSIRIS scattered-sunlight measurements, *Can. J. Phys.*, *80*, pp. 395–408, 2002.

McPeters, R. D., S. J. Janz, E. Hilsenrath, T. L. Brown, D. E. Flittner, and D. F. Heath, The retrieval of O_3 profiles from limb scatter measurements: Results from the Shuttle Ozone Limb Sounding Experiment, *Geophys. Res. Lett.*, *27*, pp. 2597–2600, 2000.

Jorge Nocedal and Stephen Wright, *Numerical Optimization*, Springer, 1999.

Oikarinen, L., E. Sihvola, and E. Kyrölä, Multiple scattering radiance in limb-viewing geometry, *J. Geophys. Res.*, *104*(D24), 31,261–31,274, doi:10.1029/1999JD900969, 1999.

Oikarinen, L., Polarization of light in UV-visible limb radiance measurements, *J. Geophys. Res.*, *106*(D2), pp. 1533–1544, doi:10.1029/2000JD900442, 2001.

Petelina, S. V., E. J. Llewellyn, D. A. Degenstein, N. D. Lloyd, R. L. Gattinger, C. S. Haley, C. von Savigny, E. Griffioen, I. C. McDade, W. F. J. Evans, D. P. Murtagh, and J. de La Noe, Comparison of the Odin/OSIRIS stratospheric ozone profiles with coincident POAM III and ozonesonde measurements, *Geophys. Res. Lett.*, *31*, doi:10.1029/2003GL019299, 2004.

Platt, U., Differential optical absorption spectroscopy (DOAS), Air monitoring by spectroscopic techniques, *Chem. Anal. Ser.*, *127*, pp. 27–84, 1994.

Rodgers, C. D., *Inverse Methods for Atmospheric sounding: Theory and Practice*, World Scientific, Singapore, 2000.

Rozanov, A., V. Rozanov, and J. P. Burrows, A numerical radiative transfer model for a spherical planetary atmosphere: Combined differential-integral approach involving the Picard iterative approximation, *J. Quant. Spectrosc. Radiat. Transfer*, 2000.

Rusch, D. W., G. H. Mount, C. A. Barth, R. J. Thomas, and M. T. Callan, Solar mesosphere explorer ultraviolet spectrometer: Measurement of ozone in the 1.0-0.1 mbar region, *J. Geophys. Res.*, *89*, 11,677–11,687, 1984.

Sioris, C. E., Haley, C. S., McLinden, C. A., von Savigny, C., McDade, I. C., McConnell, J. C., Evans, W. F. J., Lloyd, N. D., Llewellyn, E. J., Chance, K. V., Kurosu, T. P., Murtagh, D. P., Frisk, U., Pfeilsticker, K., Bosch, H., Weidner, F., Strong, K., Stegman, J., and Megie, G., Stratospheric profiles of nitrogen dioxide observed by Optical Spectrograph and Infrared Imager System on the Odin satellite, *J. Geophys. Res.*, 108(D7), 4215, doi:10.1029/2002JD002672, 2003.

Strong, K., B. M. Joseph, R. Dosanjh, I. C. McDade, C. A. McLinden, J. C. McConnell, S. J., D. P. Murtagh, and E. J. Llewellyn, Retrieval of vertical concentration profiles from OSIRIS UV-visible limb spectra, *Can. J. Phys*, *80*, 409–434, 2002.

Tian, X., Xie, Z. and Dai, A., An ensemble-based explicit four-dimensional variational assimilation method. J. Geophys. Res., Vol. 113, pp. D21124 (1-13), 2008.

Tukiainen, S., *Development and Validation of the OSIRIS inversion module*, Master's Thesis at the Helsinki University of Technology, Laboratory of Space Technology, 2006.

Veersé, F., *Variable-storage quasi-Newton operators as inverse forecast/analysis error covariance matrices in variational data assimilation*. INRIA Report No 3685, April 26, 1999.

Veersé, F., *Variable-storage quasi-Newton operators for modeling error covariances*. In Proceedings of the Third WMO Irternational Symposium on Assimilation of Observations in Meteorology and Oceanography, 7-11 June 1999, Quebec City, Canada, World Meteorological Organization, Geneva, 1999.

Veersé, F., Auroux, D. and Fisher, M., *Limited-memory BFGS diagonal predonditioners for a data assimilation problem in meteorology*. Optimization and Engineering, 1, pp. 323-339, 2000.

Von Savigny, C., C. S. Haley, C. E. Sioris, I. C. McDade, E. J. Llewellyn, D. Degenstein, W. F. J. Evans, R. L. Gattinger, E. Griffioen, E. Kyrölä, N. D. Lloyd, J. C. McConnell, C. A. McLinden, G. Mégie, D. P. Murtagh, B. Solheim, and K. Strong, Stratospheric ozone profiles retrieved from limb scattered sunlight radiance spectra measured by the OSIRIS instrument on the Odin satellite, *Geophys. Res. Lett.*, *30*, doi:10.1029/2002GL016401, 2003.

Voutilainen A., T. Pyhälahti, K. Kallio, J. Pulliainen, H. Haario, J. Kaipio *A filtering Approach for Estimating Lake Water Quality from Remote Sensing data*. Int. J. Appl. Earth Observation & Geoinformation, ISSN 0303-2434, Vol. 9, 1, pp. 50-64, February 2007.

WMO, Scientific assessment of stratospherosc ozone, *World meteorological Organization Global ozone research and monitoring project*, Report No. 18, Geneva, 1988.

Yang W., Navon I. M., Courtier P., *A new Hessian preconditioning method applied to variational data assimilation experiments using NASA general circulation models*. Mon. Weather Rev. 124, pp. 1000-1017, 1996.

PUBL. I

Inversion algorithms for recovering minor species densities from limb scatter measurements at UV-visible wavelengths

Harri Auvinen, Liisa Oikarinen,[1] and Erkki Kyrölä

Geophysical Research Division, Finnish Meteorological Institute, Helsinki, Finland

[1] Stratospheric ozone can be measured with good global coverage and good vertical resolution by continuous scanning of the limb of the sunlit atmosphere. In the near future there will be several satellite instruments exploiting this limb-scanning method using the UV-visible region of the spectrum, e.g., the Optical Spectrograph and Infrared Imaging System (OSIRIS) launched on the Odin satellite in February 2001, and the Scanning Imaging Absorption Spectrometer for Atmospheric Chartography launched on Envisat in March 2002. Envisat also carries the Global Ozone Monitoring by Occultation of Stars instrument, which will measure limb-scattered sunlight under bright limb occultation conditions. In this paper we present an inversion method to retrieve vertical ozone profiles from limb scatter measurements. The method uses a modified onion-peeling approach. Multiple scattering is taken into account by a precalculated total to single-scattering radiance ratios tabulated as a function of wavelength, tangent altitude, and several other relevant parameters. The sensitivity of the retrieval method is studied using the OSIRIS instrument as an example. Constituent retrieval errors are estimated by applying the method to simulated OSIRIS measurements. *INDEX TERMS:* 3260 Mathematical Geophysics: Inverse theory; 0394 Atmospheric Composition and Structure: Instruments and techniques; 0340 Atmospheric Composition and Structure: Middle atmosphere—composition and chemistry; *KEYWORDS:* Odin/OSIRIS, trace gas retrieval, inversion method, onion peeling, limb scattering, multiple scattering radiative transfer.

1. Introduction

[2] Acquiring accurate information on concentrations of various atmospheric constituents is one of the key objectives of atmospheric research. Criteria that can be used to judge various measurement methods are vertical resolution, long-term stability, and global coverage of the measurements. The need for little or no a priori information can also be seen as an advantage of a method.

[3] Instruments measuring transmitted and/or scattered sunlight in the ultraviolet and visible regions have the advantage that molecules responding at these wavelengths are not very sensitive to temperature and pressure. On the other hand, the retrieval process requires either absolute knowledge of the external source of light or a retrieval approach that makes use of relative measurements. The absolute knowledge can be achieved by directly measuring, for example, the solar irradiance spectrum, but as has been demonstrated in the past, it is difficult to make these measurements with the required stability [e.g., *World Meteorological Organization*, 1988]. The relative measurement principle is demonstrated directly in occultation measurements. The solar occultation measurements have good signal-to-noise ratios, but their global coverage is poor

and the terminator geometry is an additional source of uncertainty for diurnally varying species. The stellar occultation method, on the other hand, suffers from poor signal-to-noise ratios but delivers better vertical resolution and improved global coverage.

[4] A new proposed measurement method is based on limb-scattered solar radiation (Figure 1). It exploits the high luminosity of the Sun and delivers a good vertical resolution and better global coverage than occultations. The method does not directly require absolute knowledge of the solar irradiance since the retrieval method can be based on ratios of measurements at high and low tangent heights. Therefore it partially achieves the self-calibration principle used by the occultation measurements. An inherent difficulty in this method is the relatively high amount of multiple-scattering radiance in the signal, which is difficult to include in the forward model used in the retrieval method. Multiple scattering is also dependent on the albedo of Earth's surface and cloud cover.

[5] The limb-scattering technique has been used in the Ultraviolet Spectrometer on the Solar Mesospheric Explorer to measure mesospheric ozone [*Rusch et al.*, 1984]. Retrieval of stratospheric ozone profiles by this technique has been demonstrated recently by the Shuttle Ozone Limb Sounding Experiment (SOLSE) and the Limb Ozone Retrieval Experiment (LORE) flown on shuttle flight STS-87 in 1997 [*McPeters et al.*, 2000; *Flittner et al.*, 2000]. Other satellite missions using this technique include the Optical Spectrograph and Infrared Imaging System

---

[1]Deceased 27 April 2002.
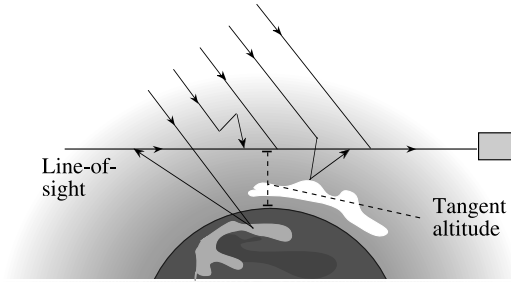
Figure 1. The limb scatter measurement. Solar light can be single scattered or multiple scattered. The latter part may include contributions from reflection at Earth's surface or from clouds.

(OSIRIS) [*Llewellyn et al.*, 1997] on board the Odin satellite (launched in February 2001) and the Scanning Imaging Absorption Spectrometer for Atmospheric Chartography (SCIAMACHY) [*Bovensmann et al.*, 1999] on Envisat (launched in March 2002). The Global Ozone Monitoring by Occultation of Stars (GOMOS) instrument [*Bertaux et al.*, 1991], also on Envisat, will use the stellar occultation to retrieve ozone profiles, but its measurements, made under bright limb conditions, also provide spectra of scattered sunlight.

[6] A key issue for constituent retrieval from limb-scattering measurements is the demonstration of the ability to retrieve the vertical profile of ozone and possibly other gases with acceptable accuracy. The main obstacle is to understand how to handle the multiple-scattering component of the signal [*Oikarinen et al.*, 1999]. The purpose of this paper is to present an inversion method for the limb measurements. The effectiveness of the method is demonstrated using simulated observations. In section 2 we present some characteristics of the OSIRIS, SCIAMACHY, and GOMOS limb measurements.

[7] In section 3 we discuss radiative transfer modeling of the limb scatter measurement. In sections 4 and 5 we present the inversion method, and in section 6 we study the accuracy of the method using simulated measurements.

2. Limb-Viewing Instruments

[8] In the coming years, a number of satellite instruments using limb-scattered sunlight measurements are planned. The first will be OSIRIS on Odin. Odin is designed for a lifetime of 2 years. The orbit of Odin is sun-synchronous, with the ascending node at 18.00, an altitude of about ~600 km, a period of 96 min, and ~15 orbits per day. OSIRIS will scan the limb either in a continuous or in a stepwise manner from tangent altitudes of 60 km (alternatively 120 km) to 7 km. For Odin's aeronomic mode the instrument nominally scans in the orbit plane, but it is possible to scan off track by up to ±32°. Figure 2 shows the illumination geometry of the OSIRIS instrument in different months. Each scan takes ~70–150 s, giving ~40–60 scans per orbit. The restituted pointing accuracy in the aeronomy mode is 1.2 arc min, which is equivalent to 1 km of

uncertainty in the vertical at the tangent point. As OSIRIS scans the limb, the tangent point moves 0.75 km/s in the vertical. During measurements in the orbit plane the tangent point moves horizontally 1.7 km/s toward or away from the spacecraft depending on the scanning direction (upward or downward scan). The satellite itself moves at 7.6 km/s relative to Earth's surface. This leads to a net movement of 400–1300 km during one scan.

[9] The OSIRIS instrument has a UV-visible spectrograph, covering the wavelength range 280–800 nm and 3 fixed 30-nm-wide IR channels near 1260, 1270, and 1560 nm. The spectrograph has a CCD matrix detector. The slit of the OSIRIS spectrograph is oriented along the Earth limb. The instrument has a telescope aperture of 10 cm². Spectral resolution of OSIRIS is 1 nm in the UV visible. The spectrograph field of view is 0.02° × 0.75°, leading to a vertical coverage of 1 km and a horizontal coverage of 40 km across the line of sight. The detector integration time can be varied. Typically, values between 0.1 and 5 s are used.

[10] In this paper we have simulated data and have retrieved geophysical parameters using the instrumental parameters of OSIRIS and the satellite parameters of Odin. We expect that the retrieval method developed in this way can also be applied to the GOMOS sunlit occultation measurements and the SCIAMACHY limb measurements in the UV-visible spectral region, which have different spectral and spatial resolution to OSIRIS.

3. Limb Radiance

[11] The simulation of limb scatter measurements requires solving the radiative transfer equation with multiple scattering in a spherical atmosphere. Some radiative transfer models (RTMs) that can be applied to limb viewing have been developed. The model of *Herman et al.* [1994] uses a
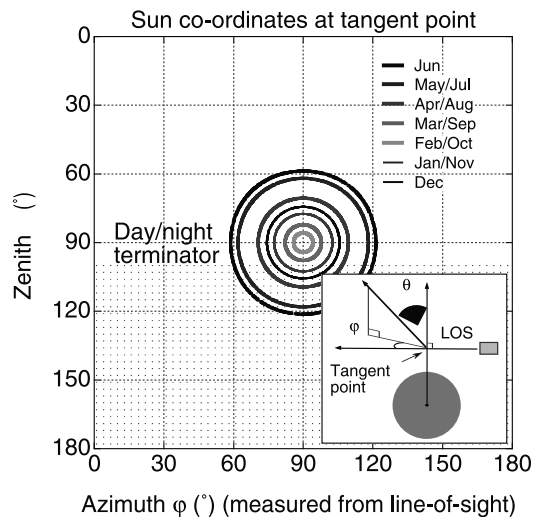
Figure 2. Solar illumination geometry of OSIRIS measurements at different months of the year. LOS, line of sight.

Gauss-Seidel iteration scheme to solve the RTM in a spherical atmosphere. *Rozanov et al.* [2000] have developed a spherical RTM called Combining Differential-Integral approach involving the Picard Iterative approximation (CDIPI) and *Griffioen and Oikarinen* [2000] have developed a pseudospherical model for limb-viewing geometry (LIMBTRAN), which makes use of a plane-parallel finite difference or matrix operator doubling and adding model. In addition, Monte Carlo radiative transfer models have been used to simulate limb radiance [*Collins et al.*, 1972; *Marchuk et al.*, 1980; *Oikarinen et al.*, 1999].

[12] Although the models listed above have been developed with computational efficiency especially in mind, solving the RTM at each wavelength and tangent altitude at each iteration step of data inversion is very time consuming. A faster approximation is the single-scattering scenario

$$I^{SS}(\lambda, 1) = I^{Sun}(\lambda)\int_1 T^{ps}(\lambda, s)T^{op}(\lambda, s)P[\lambda, \mathbf{r}(s), \theta(s)]k^s[\lambda, \mathbf{r}(s)]ds.$$

(1)

Here s is a distance along the line of sight (LOS) of the detector measured from the detector. The LOS $\mathbf{l}$ is specified by both the detector position and its viewing direction. $T^{ps}(\lambda, s)$ is the transmittance of the path from the Sun to the scattering point $\mathbf{r}(s)$ at wavelength λ, and $T^{op}(\lambda, s)$ is the transmittance of the path from the scattering point to the detector. The solar irradiance incident on the atmosphere is denoted by $I^{Sun}(\lambda)$. It arrives at the LOS at angle $\theta(s)$. Because of refraction the scattering angle $\theta(s)$ varies slightly along the LOS.

[13] The scattering phase function $P(\lambda, \mathbf{r}, \theta)$ is a weighted sum of phase functions for molecular and particle scattering

$$P(\lambda, \mathbf{r}, \theta) = P^m(\lambda, \mathbf{r}, \theta)\frac{k^m(\lambda, \mathbf{r})}{k^s(\lambda, \mathbf{r})} + P^a(\lambda, \mathbf{r}, \theta)\frac{k^a(\lambda, \mathbf{r})}{k^s(\lambda, \mathbf{r})}.$$

(2)

Function $k^m(\lambda, \mathbf{r})$ is the coefficient for scattering from molecules. It is equal to the product of the spectral-scattering cross section and the local density, $k^m(\lambda, \mathbf{r}) = \sigma^m(\lambda, \mathbf{r})\rho^m(\lambda, \mathbf{r})$. Similarly, $k^a(\lambda, \mathbf{r})$ is the coefficient for scattering from aerosols. The total scattering coefficient $k^s(\lambda, \mathbf{r})$ at wavelength λ and point $\mathbf{r}$ is $k^s(\lambda, \mathbf{r}) = k^m(\lambda, \mathbf{r}) + k^a(\lambda, \mathbf{r})$. The phase function for molecular scattering can be taken to be independent of $\mathbf{r}$, and it depends on the scattering angle θ as

$$P^m(\theta) = \frac{3}{16\pi}(1 + \cos^2\theta),$$

(3)

where we have ignored the depolarization of Rayleigh scattering (which makes P^m also independent of λ). The aerosol-scattering phase function can be simulated by the Henyey-Greenstein phase function

$$P^a(\lambda, \mathbf{r}, \theta) = \frac{1}{4\pi}\frac{1 - g^2}{(1 + g^2 - 2g\cos\theta)^{1.5}}.$$

(4)

Parameter g is the asymmetry parameter of the phase function. It generally depends on both λ and $\mathbf{r}$. A
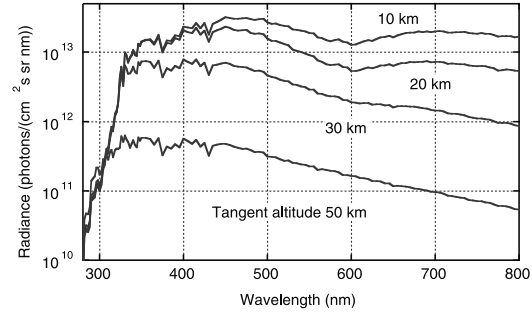
Figure 3. Limb intensity spectra simulated for tangent altitudes 10, 20, 30, and 50 km.

representative value for background stratospheric aerosols is $g = 0.75$.

[14] The single-scattering model is not very accurate at wavelengths longer than ~310 nm. The amount of multiple scattering depends on wavelength, the reflectivity of the underlying surface, and aerosol load during the measurement. A heavy aerosol load in the stratosphere, or a strongly reflecting lower surface, can increase the ratio of multiple to total scattering significantly. The amount of multiple scattering also depends on the measurement geometry. Especially, the relative amount of multiple scattering decreases as the solar zenith angle increases. In the OSIRIS measurement geometry, multiple scattering constitutes 10–40% of the total intensity at visible wavelengths [*Oikarinen et al.*, 1999]. Figure 3 shows the UV-visible spectrum of limb radiance at tangent altitudes 10, 20, 30, and 50 km. The spectra were simulated by the Monte Carlo model Siro [*Oikarinen et al.*, 1999] using the U.S. Standard Atmosphere 1976 neutral density profile, U.S. Standard ozone concentration profile, the MODTRAN background stratospheric aerosol model [*Berk et al.*, 1989], rural tropospheric aerosol model of MODTRAN (visibility 23 km), and surface albedo of $A = 0.3$. The solar zenith angle at the LOS tangent point was 80°, and the azimuth angle was 90°.

[15] In the following we have ignored several details of radiative transfer modeling that have to be taken into account in the analysis of measurement data but which do not directly affect the inversion scheme. These include refractive bending, polarization of light [*Oikarinen*, 2001], and nonhomogeneity of the surface albedo [*Oikarinen*, 2002].

4. Inversion Method

4.1. Approach

[16] Data in a limb inversion problem consist of radiance spectra and related pointing data. In the present work we divide the data into individual limb scans. Each scan is a set of consecutive measurements covering the altitude region in a monotonic manner. In the inversion one scan is inverted at a time. The benefit of this approach is that the volume of data to be handled in the inversion can be minimized. In addition, we can approximate the atmosphere to be, at least

in the first order, locally and spherically symmetric (see section 2 for tangent point movements).

[17] The goal of the inversion is to retrieve vertical profiles of those constituents that contribute to the apparent optical depth in the measured spectral range. Experiments performed using the Limb and Occultation Measurement Simulator [*Kyrölä et al.*, 1999] have shown that O_3, NO_2, $OClO$, BrO, neutral density, and aerosols are potential candidates for retrieval attempts in the UV-visible region.

[18] The basic data that the inversion requires are the measured radiances $I^{meas}(\lambda, l_j)$ from one scan (the subscript j runs over the different tangent altitudes). In order to release requirements on absolute calibration, it is advantageous not to use radiances directly but to use the following radiance ratios:

$$F(\lambda, j) = \frac{I^{meas}(\lambda, j)}{I^{meas}_{ref}(\lambda, ref)}, \qquad (5)$$

where j references to a measurement tangent altitude. We call $F(\lambda, j)$ the transfer spectrum. The reference spectrum I^{meas}_{ref} is a measurement made at high tangent altitude by the same instrument. In principle we could also use a solar reference spectrum, but then the self-calibration feature of the method would disappear. We remark that the reference layer does not have to be the uppermost layer in our inversion process. The use of ratio (5) also diminishes the effects due to surface albedo and clouds [*Flittner et al.*, 2000] and polarization [*Oikarinen*, 2001]. The modeled transfer spectrum is following:

$$M(\lambda, j, \rho) = \frac{I^{model}(\lambda, j, \rho)}{I^{model}_{ref}(\lambda, \rho_{ref})}. \qquad (6)$$

Parameters ρ in I^{model} are the unknown constituent densities to be fitted. For calculation of I^{model}_{ref} we need a priori information of densities ρ_{ref}. In our present work we have neglected multiplication with solar spectrum and convolution with instrument function in the calculation of I^{model} and I^{model}_{ref} to save computation time. However, it is possible to add this feature later if needed. According to our simulations, the most important point in modeling I^{model}_{ref} is to find the correct signal level. The shape of the radiance spectrum does not change rapidly at tangent altitudes of 40–60 km. For example, the spectral shape of simulated measurements is practically unchanged when tangent altitude is varied by 1–2 km. The signal level of the measurements is basically controlled by neutral density. Neutral density profiles concurrent to the measurements can be obtained from meteorological analysis, e.g., by the European Centre for Medium-Range Forecasts (ECMWF), at an accuracy of ~0.5% at altitude 10 km and 2–5% at altitude 60 km. If the signal level of the simulated reference spectrum is shifted from the true one, we will get a systematic shift in the retrieved neutral density profile. To avoid this, we have to add one step to our inversion algorithm. The idea is to compare retrieved profile with the a priori profile in lower atmosphere (10–20 km), where the a priori information has good quality, and to remove the shift in lower atmosphere by updating

the a priori information of the upper atmosphere (and I^{model}_{ref}) iteratively.

[19] The problem of inverting the measurements (5) to concentration profiles can be approached in several ways. Data from one scan can be inverted in one step (a so-called global fit), or we can invert each exposure separately. We can use all spectral data or only selected wavelengths. An important choice is to decide if we are going to invert all feasible atmospheric parameters from the data or fix some parameters on the basis of some of a priori data. For example, as noted above, reasonable a priori data can be found for neutral density. In addition, temperature profiles required for accurate modeling of absorption cross sections can be obtained a priori.

[20] The major contribution to limb radiance often comes from near the tangent point of the LOS [*Flittner et al.*, 2000]. This fact has led us to try the onion-peeling approach [see, e.g., *Chu and McCormick*, 1979] in data inversion. As we noticed in section 3, the radiative transfer modeling can be very time consuming in limb geometry. The onion-peeling approach allows us to divide the inversion problem to smaller parts, and radiative transfer modeling becomes lighter and faster than, e.g., in a global-fit approach. In the spectral domain we use the whole spectral range of the data and invert all constituents simultaneously.

[21] The method used for SOLSE/LORE data [*Flittner et al.*, 2000] differs in two respects from our method. First, instead of inverting the radiance measured at different wavelengths (or ratios (5)) directly, they use a ratio of radiance measured at selected wavelength pairs or triplet. *Flittner et al.* [2000] use only a few wavelengths in the UV-visible range (LORE has only 5 narrow-wavelength bands), whereas we use as many spectral points as the available computer time permits. In addition, they aim to retrieve only ozone and to try to minimize perturbations of other constituents. We aim to retrieve all contributing constituents. Second, *Flittner et al.* [2000] use a priori information in the solution, i.e., the so-called optimal estimation method. We have tried to minimize all a priori information effects in our solution. However, the a priori information for neutral density is also in use in our solution.

4.2. Modified Onion-Peeling Method

[22] In onion peeling, we assume that the atmosphere consists of layers. Within each layer the densities of the different constituents are assumed to be constant. In a standard onion-peeling approach applied, for example, to occultation data, one assumes that a measurement depends only on constituent densities in and above the layer which contains the tangent altitude of the measurement. In the first step of onion peeling, the densities in the uppermost layer are retrieved using the uppermost measurement. Then the densities in the lower layers are solved from the top to the bottom using the already retrieved densities for the layers above.

[23] In the scattered sunlight case, where the measured signal also includes contributions from the lower atmosphere we cannot, in principle, use the onion-peeling method. However, it is possible to use the onion-peeling approach in an iterative manner. In every iteration of the

peeling process we obtain a new, better approximation for the lower atmosphere. Usually 2–3 iterations of the peeling is enough to receive accurate results.

[24] Let us now take a closer look at the modified onion-peeling method. In this approach we use measurements in the following way. We construct the layer structure so that the tangent point of exactly one measurement lies in each layer. Hence the number of layers is the same as the number of measurements. Typically, the number of measurements is 40–60 in one scan.

[25] We then try to find constituent densities for each layer one by one by using measurements from the top down. The iterative use of onion peeling minimizes the effect of the initial guess densities in the lower layers. In the first peeling iteration we simultaneously retrieve vertical profiles for neutral density, aerosols, and ozone. After the first peeling iteration we have information of the atmosphere which can be used as a priori information of the lower atmosphere during the second peeling iteration. In the second peeling iteration we try to retrieve simultaneously minor trace gases like NO_2, OClO, and BrO. If the density of any of these minor trace gases gets negative values during the retrieval process, we draw the conclusion that we cannot detect this constituent, and therefore we are no longer trying to fit that gas in the corresponding layer. This feature of the method can be used to select dynamically the constituents that we are going to retrieve during the operational use of the method.

[26] To start the peeling, we have to select the uppermost measurement to use and to give an initial guess for the densities in the corresponding layer. At first we retrieve constituent densities for the uppermost layer using a non-linear least squares fitting method. The peeling step is to find constituent densities for the next layer by using the already fitted density values in the upper atmosphere. For starting values in each layer we use the densities just retrieved for the previous layer or, if available, we use the densities fitted in the previous iteration of the peeling in the current layer. When we repeat the peeling step, we will simultaneously get a vertical profile for each of the constituents.

4.3. Fitting Inside a Layer

[27] In each layer we have to solve a nonlinear inverse problem. The traditional approach to inverse problems is to seek for the solution which minimizes the sum of squared residuals χ^2, i.e., the least squares solution [see, e.g., *Menke*, 1989].

[28] Let us next derive an expression of the sum of squared residuals. Let $\rho_i(j)$ be the density of constituent i in layer j, and let $F(\lambda, j)$ be the observed transfer spectrum at tangent height j, which also includes random noise. Using matrix notation in spectral dimension, the χ^2 term for tangent height j is

$$\chi_j^2 = [M(\lambda,j,\rho) - F(\lambda,j)]C^{-1}[M(\lambda,j,\rho) - F(\lambda,j)], \quad (7)$$

where $F(\lambda, j)$ is the spectrum measured at the tangent altitude corresponding to layer j and $M(\lambda, j, \rho)$ is the forward modeled transfer spectrum at different wavelengths and given constituent densities ρ at different layers. For OSIRIS

the number of unknown quantities ρ per layer is 3–6. In this work we further assume that there is no correlation in the noise between different detector pixels. Therefore the covariance matrix of the measurement error C is a diagonal matrix.

[29] The detector actually measures photons (i.e., discrete counts), and hence the measurement error is Poisson distributed with variance proportional to the magnitude of the measured signal. Owing to the great amount of detected photons, the noise due to photon-counting statistics can be assumed to be normally distributed with variance proportional to the number of photons in the atmospheric signal and the estimate of the dark current, which is subtracted from the signal before inversion. In addition, readout noise has to be added (square-wise) to the noise estimate. In OSIRIS, readout noise is very small compared with the other noise sources.

[30] The nonlinear least squares solution (i.e., the minimum of χ_j^2) can be found by using an iterative routine. We have applied the Levenberg-Marquardt method [see, e.g., *Press et al.*, 1992] to the OSIRIS inversion problem. The method is based on combining the steepest descent and the inverse Hessian methods. In addition to the fitted parameter values the Levenberg-Marquardt method returns error estimates which are computed by assuming that the χ^2 term can be linearized around the minimum value.

5. Multiple-Scattering Corrections

[31] In this section we will introduce a method to approximate the multiple-scattering effect. Our aim is to approximate the multiple-scattering effect by using the single-scattering model and ratio spectrum look-up tables. Alternatively, if a fast multiple-scattering algorithm is available, the ratio spectra can be calculated during the inversion process. *Flittner et al.* [2000] have also used a ratio spectrum approach to take into account multiple scattering in their retrieval method.

[32] Let $M_s(\lambda, j, \rho)$ be the simulated transfer spectrum at altitude j and at given constituent densities ρ obtained using the single-scattering model and similarly, let $M(\lambda, j, \rho)$ be the simulated transfer spectrum obtained using the full radiative transfer model, including the multiple-scattering effect with same constituent densities. By dividing the spectrum $M(\lambda, j, \rho)$ by $M_s(\lambda, j, \rho)$, we obtain a ratio spectrum

$$R(\lambda,j,\rho) = \frac{M(\lambda,j,\rho)}{M_s(\lambda,j,\rho)}. \quad (8)$$

[33] The ratio spectrum method gives us a fast way to approximate the multiple-scattering effect in OSIRIS data inversion. Hence the model can be written as follows:

$$M(\lambda,j,\rho) = M_s(\lambda,j,\rho)R(\lambda,j,\rho). \quad (9)$$

The set of $R(\lambda, j, \rho)$ is kept fixed during the onion peeling. When fitting constituent densities inside a layer, χ_j^2 is minimized by adjusting the density values in that layer alone using the single-scattering model. Ratio spectra depend strongly on wavelength and altitude, but the density
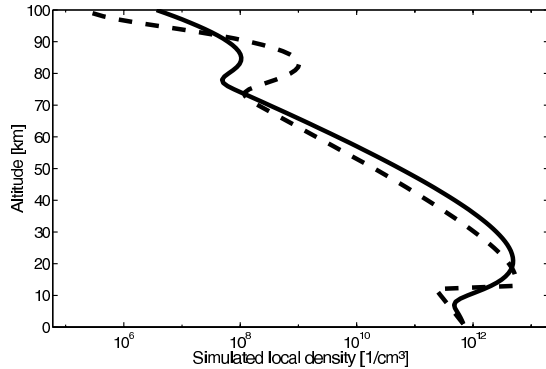
Figure 4. Different ozone density profiles used in the simulations. The solid line is the U.S. Standard ozone profile, and the dashed line is the modified ozone profile.

dependence is weak enough to allow reasonably accurate results by the modified onion-peeling method. After each peeling iteration we have a new vertical profile for each constituent, and we can reselect (or recalculate) better-matching ratio spectra to use in the next peeling iteration. This scheme converges rapidly, and results are shown later in this paper. The ratio spectrum must be tabulated as a function of tangent altitude, solar zenith angle, and solar azimuth angle. These parameters can be calculated from the instrument position and pointing information. The ratio spectrum also depends on the albedo of Earth's surface, which is not retrieved directly from the measurement but is assumed to be known from other sources. For example, the cloud optical depth, cloud top altitude, and snow cover at the time and location of an OSIRIS measurement can be obtained from the ECMWF. A database of the albedo of Earth's surface has been collected, e.g., by the Surface and Atmospheric Radiation Budget working group of the Clouds and the Earth's Radiant Energy System project of NASA. It is needed to calculate R for a set of different neutral density profiles. For example, in the case of OSIRIS, the neutral density profile of the lower atmosphere can be obtained off line at reasonable accuracy from the ECMWF. During the reselection of ratio spectra we need to have R tabulated for a selection of standard ozone and aerosol profiles.

[34] To show how the density of O_3, the principal constituent to be retrieved from OSIRIS measurements, affects the ratio spectrum, let us compare R calculated using two different ozone density profiles. We have simulated R by the Siro model with the same atmospheric parameters that we used in section 3 to calculate the absolute intensity spectra of Figure 3. The original U.S. Standard ozone profile and a modified ozone density profile are shown in Figure 4. The corresponding ratio spectra are shown in Figure 5. The noise in the spectra comes from the Monte Carlo simulation. It has a standard deviation of 0–0.5% in the UV and 0.2–1.8% in the visible region. The difference between these ratio spectra is small enough to allow our use of the ratio spectra simulated by using the modified ozone profile in the first peeling iteration. In the

next section we present some inversion results based on these ratio spectra.

6. Results

[35] In this section we present some results of the accuracy of OSIRIS data inversion obtained by applying the inversion method to simulated measurement data. The inversion scheme presented in sections 4.2, 4.3, and 5 can be summarized by the following algorithm.

6.1. Step 1

[36] Initialize the first guess of the ratio spectra and simulate the reference spectrum by using a priori values.

6.2. Step 2

[37] Repeat the peeling two times (first step). At the first time, fit only ozone, aerosols, and neutral density; at the second time, fit also minor absorbers. The profiles retrieved in this step are used as inversion starting values in step 3.

  1. Select constituents to retrieve (see section 4.2 for more details).
  2. Working from top down for every tangent height j, solve the nonlinear least squares problem (7) by using the Levenberg-Marquardt method and by adjusting only the densities in layer j. The multiple-scattering effect is taken into account by (9), where the ratio spectrum (8) is based on a priori knowledge.
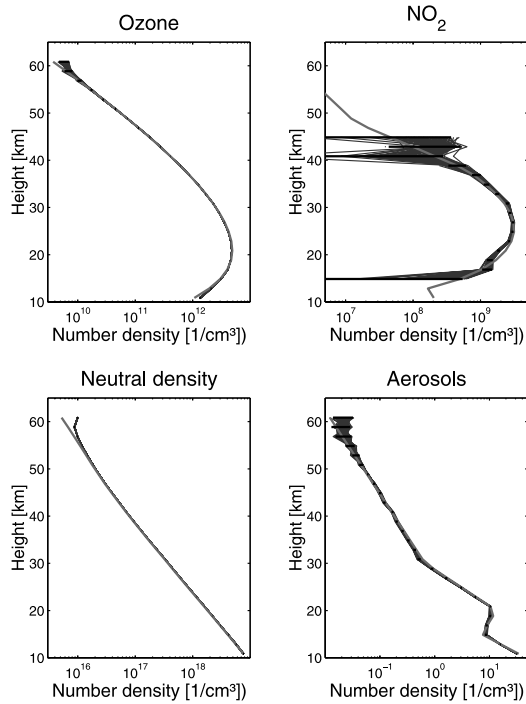
6.3. Step 3

[38] Iterate 1–2 times.
  1. Compare retrieved neutral density profile with the corresponding ECMWF profile in lower atmosphere
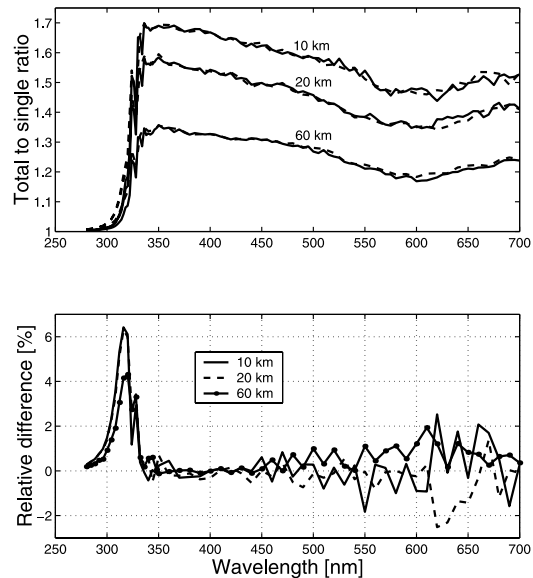
Figure 5. Simulated ratio spectra. The solid line shows spectra simulated using the U.S. Standard ozone profile, and the dashed line shows spectra simulated using the modified ozone profile.

Figure 6. One hundred samples of the OSIRIS data inversion with exact ratio spectra. The red line shows the true profile, and each blue line corresponds to one sample of the inversion results. See color version of this figure at back of this issue.

(10–20 km) and adjust retrieved neutral density profile by scaling with a constant factor if necessary.

2. Resimulate the reference spectrum in (6) by using the possible adjusted profile of neutral density and retrieved profiles for other constituents.

3. Reselect the ratio spectra from look-up tables using the profiles retrieved in the previous iteration of the peeling.

4. Repeat the peeling two times (as in step 2).

[39] The measurement data was simulated by the radiative transfer model Siro. The same simulation parameters as in sections 3 and 5 were used (here we used the U.S. Standard ozone profile). The simulation contains O_3, NO_2, neutral density, and aerosols. The solar zenith angle at the LOS tangent point is set to 80°, and the azimuth angle is set to 90°. Earth's surface was assumed to have a constant albedo of $A = 0.3$. The data consist of one scan with 45 measurements from tangent altitude 98–10 km at 2-km intervals. The number of wavelength points used in data inversion was ∼100. Random noise with a signal-to-noise ratio representative of true OSIRIS measurements was added to the radiance spectra.

[40] Let us first study the sensitivity of the modified onion-peeling method in general by using exact ratio spectra (8). Hence the ratio spectra are assumed to be known, and therefore we do not need to use the step 3 of the previous algorithm. In this case the retrieval will not

suffer from any modeling error. The pattern of random noise in the data was varied in order to get samples of the inversion results. As long as we keep the data, the ratio spectra and simulated reference spectrum fixed the inversion, and results do not variate even if we change the initial guess densities a lot. This shows that the method is not sensitive to the inversion starting values and that the Levenberg-Marguardt method is capable of finding the true minimum of the χ_f^2. In Figure 6 we have 100 samples of the result of OSIRIS data inversion. The corresponding mean relative error of different constituents is shown in Figure 7. As Figure 6 shows, the deviations of the inversion results are quite small. The onion-peeling method is often criticized for the propagation of errors from the uppermost altitude layers to the lower layers. However, we see no evidence of error propagation in the inversion results of Figure 6. We are not using any a priori information of the layers above the uppermost layer, and therefore there is a systematic shift in the uppermost layer, where the inversion was started. This shift combines the contributions of the upper atmosphere and the uppermost layer. We notice that the shift dies out quickly when the peeling continues downward.
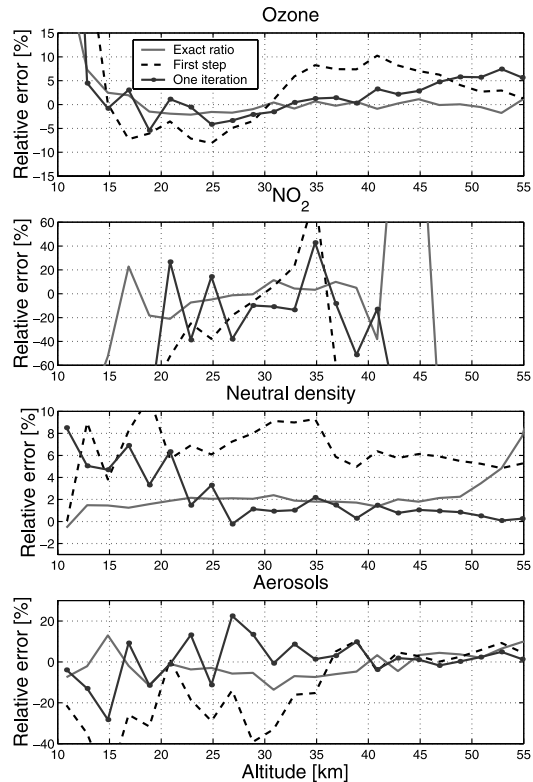
Figure 7. Relative error of the inversion results. The red line is the exact ratio spectra. The dashed line is the relative error after the first step, and the blue line is the relative error after the first iteration in the retrieval with unknown ratio spectra. See color version of this figure at back of this issue.

[41] Let us next study the more realistic case where the exact ratio spectra are not known. In the first step we use ratio spectra which are calculated using modified density profiles. Before starting the first iteration of the inversion, we recalculate the ratio spectra using the profiles retrieved from the first step (or select new ratio spectra from the look-up tables on the basis of the retrieved profiles).

[42] In Figure 8 we have the true density profiles of the constituents used in data simulation, the retrieved profiles after the first step, and the profiles retrieved in the first iteration of the peeling. The data inversion took ~5 min on a Linux workstation with a 266 MHz processor. The results of the ozone retrieval are quite promising. The retrieval accuracy is better than 5% throughout the stratosphere. The ozone profile converges well toward the true profile. By using more wavelength points, we should receive even better accuracy (OSIRIS spectra have a total of 1353 wavelength pixels). This should especially increase the retrieval accuracy of NO_2. Using more wavelength points would, of course, increase the computation time of the data inversion.

[43] We notice that even if we recalculated ratio spectra instead of using look-up tables, the number of function calls of the multiple-scattering model would be much smaller than in case of using a complete multiple-scattering forward model throughout the inversion. However, the retrieval algorithm could be easily modified to use a full multiple-scattering forward model instead of the ratio spectrum approximation if, in the future, we have more computational power and a fast and accurate multiple-scattering model available. Even then it is advantageous to retrieve the first approximation of the atmospheric parameters using the ratio spectra and a much faster single-scattering forward model.

7. Conclusions

[44] Limb scatter measurements are providing a new way to monitor the chemical composition of the stratosphere. High signal-to-noise ratios and good vertical resolution are the characteristics of the method. On the other hand, the measured signal includes a relatively large multiple-scattering component. The difficulty with the multiple-scattering component is that it is very time consuming to compute, and it is also dependent both on the albedo of the Earth's surface and the cloud cover beneath the measurement line of sight.

[45] In this paper we have presented an inversion approach for limb-scattering measurements that delivers reasonably accurate results for ozone profile retrieval. The basic inversion approach is a combination of the standard onion-peeling method in the spatial domain and the spectrally global nonlinear inversion in the spectral domain. The onion peeling in the limb scatter case is not an exact solution for the discretized spatial problem, but combined with iterative steps, it has turned out to be a very effective solution. The multiple-scattering component has been taken into account using iterations. All of the multiple-scattering information has been included in the total versus single-scattering ratios calculated as a function of various geophysical parameters.

[46] The performance of the presented inversion method was studied using simulated measurements. The analysis suggests that ozone concentration in the stratosphere is
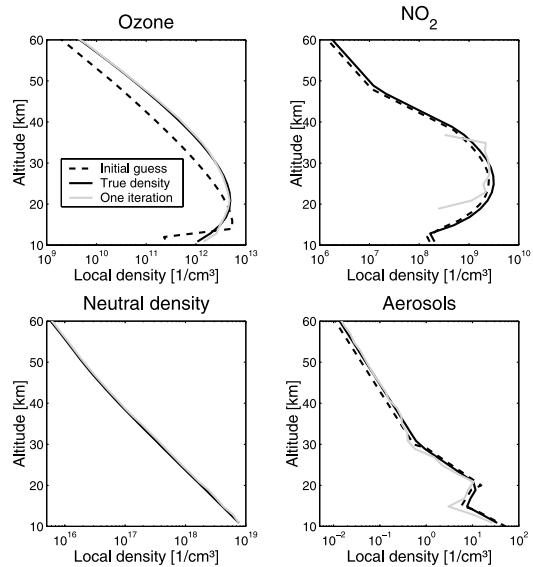
Figure 8. Convergence of the inversion of O_3, NO_2, neutral density, and aerosol profiles. The solid line is the true density, the dashed line is the initial guess used in the calculation of ratio spectra, the dotted line is the profile retrieved in the first step, and the grey line is the profile retrieved in the first iteration.

retrieved at an accuracy of 5% or better. It should be noted, however, that some complications of the processing of true measurement data were neglected in this study. The final error budget of the OSIRIS instrument, for example, consists of errors arising from random measurement noise, error in a priori information of the surface albedo, errors in modeling multiple scattering (inaccurate ratio spectra), errors modeling the optical response of aerosols, polarization sensitivity of the instrument, and uncertainty in instrument-pointing knowledge. We estimate that the first five of these error sources together cause an ozone retrieval error of ~2% to at worst 10% in the middle stratosphere. Similar to SOLSE/LORE [*Flittner et al.*, 2000], pointing uncertainty is the most serious source of error for OSIRIS. The pointing uncertainty of ±1 km can cause an ozone retrieval error of 10−15% in the middle stratosphere. Because the neutral density profile is known quite well a priori, a more accurate pointing can hopefully be obtained from the radiance measurements alone. For example, *Janz et al.* [1996] have developed a method which uses the radiance profile at wavelengths where the atmosphere is almost nonabsorbing to recover the altitude vector of the limb scan. This method is estimated to return the altitude at ±0.5 km accuracy. Many of these errors have only weak dependence on wavelength, that is, they modulate the radiance spectra by an almost constant factor. Therefore they will bias the retrieval of aerosol parameters but do not affect ozone dramatically. The work reported here and the earlier work on SOLSE/LORE retrievals by *McPeters et al.* [2000] and *Flittner et al.* [2000] give us confidence that the

retrieval problem for limb-scattering instruments is surmountable.

References

Berk, A., L. S. Bernstein, and D. C. Robertson, MODTRAN: A moderate resolution model for LOWTRAN 7, *AFGL Tech. Rep. F19628-86-C-0079*, 38 pp., Air Force Geophys. Lab., Hanscom Air Force Base, Mass., 1989.

Bertaux, J. L., G. Megie, T. Widemann, E. Chassefiere, R. Pellinen, E. Kyrölä, S. Korpela, and P. Simon, Monitoring of ozone trend by stellar occultations: The GOMOS instrument, *Adv. Space Res.*, *11*(3), 237–242, 1991.

Bovensmann, H., J. P. Burrows, M. Buchwitz, J. Frerick, S. Noel, V. V. Rozanov, K. V. Chance, and A. P. H. Goede, SCIAMACHY: Mission objectives and measurement modes, *J. Atmos. Sci.*, *56*, 127–150, 1999.

Chu, W. P., and M. P. McCormick, Inversion of stratospheric aerosol and gaseous constituents from spacecraft solar extinction data in the 0.38–1.0 μm wavelength region, *Appl. Opt.*, *18*, 1404–1413, 1979.

Collins, D. G., W. G. Blttner, M. B. Wells, and H. G. Horak, Backward Monte Carlo calculations of the polarization characteristics of the radiation emerging from spherical-shell atmospheres, *Appl. Opt.*, *11*, 2684–2696, 1972.

Flittner, D. E., P. K. Bhartia, and B. M. Herman, O_3 profiles retrieved from limb scatter measurements, *Theory*, *27*, 261–264, 2000.

Griffioen, E., and L. Oikarinen, LIMBTRAN: A pseudo three-dimensional radiative transfer model for the limb-viewing imager OSIRIS on the Odin satellite, *J. Geophys. Res.*, *105*, 29,717–29,730, 2000.

Herman, B. M., A. Ben-David, and K. J. Thome, Numerical technique for solving the radiative transfer equation for a spherical shell atmosphere, *Appl. Opt.*, *33*, 1760–1770, 1994.

Janz, S. J., E. Hilsenrath, D. Flittner, and D. Heath, Rayleigh scattering attitude sensor, *SPIE Proc.*, *2831*, 146–153, 1996.

Kyrölä, E., J. Tamminen, L. Oikarinen, E. Sihvola, P. Verronen, and G. W. Leppelmeier, LIMBO-Limb and Occultation Measurement Simulator, in *Proceedings of ESAMS '99-European Symposium on Atmospheric Measurements From Space, Rep. WPP-161*, pp. 487–493, Eur. Space Agency, Paris, 1999.

Llewellyn, E. J., et al., OSIRIS—An appplication of tomography for absorbed emissions in remote sensing, in *Applications of Photonic Technology 2*, edited by G. A. Lampropoulos and R. A. Lessard, pp. 627–632, Plenum, New York, 1997.

Marchuk, G. I., G. A. Mikhailov, M. A. Nazaraliev, R. A. Darbinjan, B. A. Kargin, and B. S. Elepov, *The Monte Carlo Methods in Atmospheric Optics*, 208 pp., Springer-Verlag, New York, 1980.

McPeters, R. D., S. J. Janz, E. Hilsenrath, T. L. Brown, D. E. Flittner, and D. F. Heath, The retrieval of O_3 profiles from limb scatter measurements: Results from the Shuttle Ozone Limb Sounding Experiment, *Geophys. Res. Lett.*, *27*, 2597–2600, 2000.

Menke, W., *Geophysical Data Analysis: Discrete Inverse Theory*, Academic, San Diego, Calif., 1989.

Oikarinen, L., Polarization of light in UV-visible limb radiance measurements, *J. Geophys. Res.*, *106*, 1533–1544, 2001.

Oikarinen, L., Effect of surface albedo variations on UV-visible limb scattering measurements of the atmosphere, *J. Geophys. Res.*, *107*, 10.1029/2001JD001492, in press, 2002.

Oikarinen, L., E. Sihvola, and E. Kyrölä, Multiple scattering radiance in limb-viewing geometry, *J. Geophys. Res.*, *104*, 31,261–31,274, 1999.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in FORTRAN, The Art of Scientific Computing*, Clarendon, Oxford, UK, 1992.

Rozanov, A., V. Rozanov, and J. P. Burrows, A numerical radiative transfer model for a spherical planetary atmosphere: Combined differential-integral approach involving the Picard iterative approximation, *J. Quant. Spectrosc. Radiat. Transfer*, in press, 2000.

Rusch, D. W., G. H. Mount, C. A. Barth, R. J. Thomas, and M. T. Callan, Solar mesosphere explorer ultraviolet spectrometer: Measurement of ozone in the 1.0–0.1 mbar region, *J. Geophys. Res.*, *89*, 11,677–11,687, 1984.

World Meteorological Organization, Scientific assessment of stratospheric ozone, *Rep. 18*, Global Ozone Res. Monit. Proj., Geneva, 1988.

———————————

H. Auvinen and E. Kyrölä, Geophysical Research Division, Finnish Meteorological Institute, P.O. Box 503, Helsinki, FIN-00101, Finland. (harri.auvinen@fmi.fi; erkki.kyrola@fmi.fi)
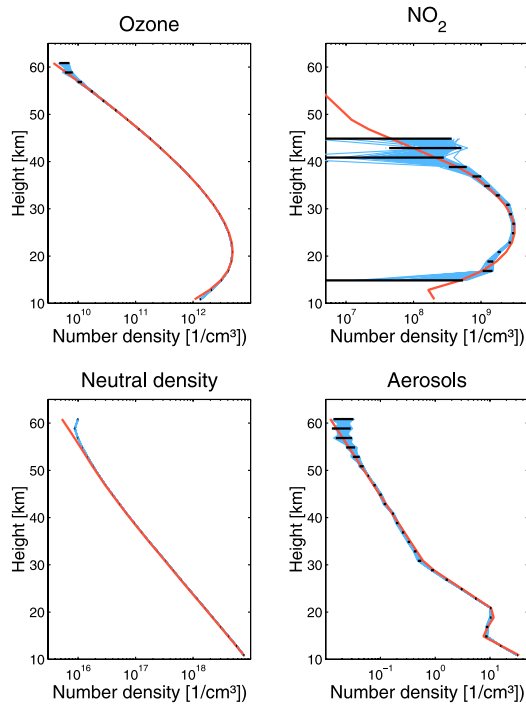
Figure 6. One hundred samples of the OSIRIS data inversion with exact ratio spectra. The red line shows the true profile, and each blue line corresponds to one sample of the inversion results.
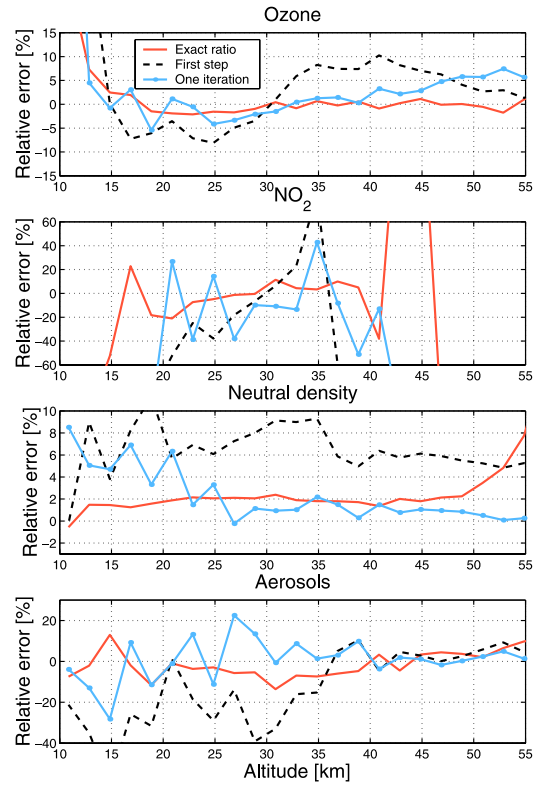
Figure 7. Relative error of the inversion results. The red line is the exact ratio spectra. The dashed line is the relative error after the first step, and the blue line is the relative error after the first iteration in the retrieval with unknown ratio spectra.

PUBL. II

Description and validation of a limb scatter retrieval method for Odin/OSIRIS

S. Tukiainen,[1] S. Hassinen,[1] A. Seppälä,[1] H. Auvinen,[2] E. Kyrölä,[1] J. Tamminen,[1] C. S. Haley,[3] N. Lloyd,[4] and P. T. Verronen[1]

[1] In this paper we present the Modified Onion Peeling (MOP) inversion method, which is for the first time used to retrieve vertical profiles of stratospheric trace gases from Odin/OSIRIS limb scatter measurements. Since the original publication of the method in 2002, the method has undergone major modifications discussed here. The MOP method now uses a spectral microwindow for the NO_2 retrieval, instead of the wide UV-visible band used for the ozone, air, and aerosol retrievals. We give a brief description of the algorithm itself and show its performance with both simulated and real data. Retrieved ozone and NO_2 profiles from the OSIRIS measurements were compared with data from the GOMOS and HALOE instruments. No more than 5% difference was found between OSIRIS daytime and GOMOS nighttime ozone profiles between 21 and 45 km. The difference between OSIRIS and HALOE sunset NO_2 mixing ratio profiles was at most 25% between 20 and 40 km. The neutral air density was compared with the ECMWF analyzed data and around 5% difference was found at altitudes from 20 to 55 km. However, OSIRIS observations yield as much as 80% greater aerosols number density than GOMOS observations between 15 and 35 km. These validation results indicate that the quality of MOP ozone, NO_2, and neutral air is good. The new version of the method introduced here is also easily expanded to retrieve additional species of interest.

1. Introduction

[2] OSIRIS (Optical Spectrograph and Infrared Imager System) is one of the two instruments on board the Swedish Odin satellite, launched in February 2001 *Llewellyn et al.* [2004]. The spectrograph part of the instrument measures limb-scattered solar light (radiance) in the wavelength region of 280–800 nm with around 1 nm spectral resolution. Odin scans toward the Earth's limb from 7 to 110 km through a controlled nodding motion. The effective Field of View (FOV) is 1–2 km due to motion of the 1 km Instantaneous Field of View (IFOV) during the integration period. In turn, the FOV is sampled discretely with 1–3 km vertical spacing. OSIRIS is the first dedicated satellite instrument to measure continuously the vertical composition of the atmosphere using the limb scatter technique and by recording the full spectrum from UV to visible wavelengths with a good spectral resolution.

[3] The UV-visible wavelength band of OSIRIS carries information of several atmospheric trace gases including ozone (O_3), nitrogen dioxide (NO_2), nitrogen trioxide (NO_3), chlorine dioxide (OClO), and bromine monoxide (BrO). Trace gases leave their absorption fingerprints in the solar spectrum, which is scattered by the neutral molecules of the atmosphere. Stratospheric aerosol loading and direct emissions from excited molecules also contribute to the shape of the observed spectrum.

[4] In addition to OSIRIS, several other spaceborne instruments have been deployed in recent decades to observe chemical composition of the middle atmosphere. Occultation and limb-viewing instruments have been used to measure vertical profiles, while nadir instruments provide mainly total column abundances of the compounds. The first spaceborne UV-visible limb scatter measurements were done by instruments on board SME (Solar Mesospheric Explorer) in the beginning of the 1980s [*Rusch et al.*, 1983, 1984]. The SOLSE and LORE instruments on the Space Shuttle flight STS-87 further proved the limb scatter technique to be feasible [*McPeters et al.*, 2000; *Flittner et al.*, 2000]. Examples of more recent space instruments, capable of doing UV-visible limb scatter measurements, are SCIA-MACHY [*Bovensmann et al.*, 1999], GOMOS [*Kyrölä et al.*, 2004], and SAGE III [*Rault*, 2005]. As the measurement principle of remote sensing instruments is always indirect,

[1]Finnish Meteorological Institute, Earth Observation, Helsinki, Finland.
[2]Lappeenranta University of Technology, Lappeenranta, Finland.
[3]Centre for Research in Earth and Space Science, York University, Canada.
[4]Department of Physics and Engineering Physics, University of Saskatchewan, Saskatoon, Canada.

data inversion methods are needed to extract information from the physical measurements.

[5] Several inversion methods have been developed to retrieve trace gas densities from the OSIRIS measurements. The so-called Triplet method was developed by *Flittner et al.* [2000] and *McPeters et al.* [2000] and was adapted to the OSIRIS limb scatter measurements by *von Savigny et al.* [2003]. The Triplet method uses wavelength triplets in the Chappuis absorption band near 600 nm to retrieve stratospheric ozone. According to *Petelina et al.* [2004], the validated altitude range for the OSIRIS triplet ozone is 15–32 km.

[6] The DOAS (Differential Optical Absorption Spectroscopy) method is one widely used approach for the retrieval of several different species. The basic DOAS approach was proposed by *Platt* [1994] and the method was applied to simulated limb scatter measurements by *McDade et al.* [2002] and *Strong et al.* [2002]. *Haley et al.* [2003] applied the DOAS technique to retrieve ozone from OSIRIS data and found good agreement with the Triplet method.

[7] The DOAS method has been also used to retrieve stratospheric NO_2 from the OSIRIS data. *Haley et al.* [2004] and *Sioris et al.* [2003] used slightly different DOAS variants and obtained quite consistent results. They also made preliminary validation against sonde and POAM III measurements proving the method feasible.

[8] The Triplet method and the various DOAS algorithms exploit only a small fraction of the available UV-visible spectrum. There have also been efforts to analyze the limb scatter spectra without reduction to differential structures. For example, *Kaiser et al.* [2004] used this principle to retrieve satellite pointing from limb scatter measurements. In addition, *Rohen et al.* [2005] retrieved mesospheric ozone using the Hartley band in ∼240–310 nm.

[9] The method used in this study is the Modified Onion Peeling (MOP) method. It was originally developed by *Auvinen et al.* [2002], but the method has later undergone significant modifications discussed in Section 2. In particular, the NO_2 retrieval part has been revised since the original publication of the method. The MOP method should offer at least one advantage compared with the other retrieval techniques: The advantage of the MOP method, compared with the other retrieval techniques, is that using information from the whole spectral band of the OSIRIS instrument, we are able to retrieve ozone by one method between 15 and 70 km.

[10] After the original MOP method was modified, we tested the sensitivity of the new version using simulated data (Section 3). Finally, we validated the outcome of the MOP method using real OSIRIS data. Inverted vertical profiles were compared with data from other instruments measuring the middle atmosphere (Section 4).

2. MOP Inversion Method

[11] The general idea of onion peeling inversion methods is to divide the atmosphere into separate layers and solve the inversion problem layer by layer from top of the atmosphere downward. This way we can construct the vertical profiles of different trace gases. In the MOP method, as described by *Auvinen et al.* [2002], we can use any number of

available wavelengths and try to retrieve various species simultaneously.

[12] The theoretical basis of the MOP inversion method is a Bayesian approach. However, using a flat a priori distribution and assuming Gaussian noise, the solution reduces to a simple weighted least squares fit to the data [e.g., *Rodgers*, 2000]. Non-linear problems, such as the one related to OSIRIS, require an iterative fitting procedure.

[13] Following *Auvinen et al.* [2002], it is advantageous not to use directly measured radiances but the ratio

$$R_{obs}(\lambda, j) = \frac{I_{obs}(\lambda, j)}{I_{obs}^{ref}(\lambda)}, \tag{1}$$

where $I_{obs}(\lambda, j)$ are measured radiances at tangent heights j and $I_{obs}^{ref}(\lambda)$ is a reference measurement from the same scan at high tangent altitude. Radiance is a function of wavelength λ. We have chosen to use the first measurement below 50 km as the reference. It would be possible to try other tangent heights as well, but that around 50 km seems to yield the best results in practice. It is already high enough to exclude spectral fingerprints from minor trace gases (such as NO_2, $OClO$, and BrO) making it easier to model. Furthermore, straylight contamination in OSIRIS increases as a function of tangent height making high altitude measurements less unreliable to use [*Llewellyn et al.*, 2004].

[14] The use of the so-called transfer spectrum (1) is useful because it diminishes systematic errors due to surface albedo, clouds, and polarization [*Flittner et al.*, 2000; *Oikarinen*, 2001]. It also reduces errors due to imperfect instrument calibration.

[15] The modeled transfer spectrum is defined as

$$R_{mod}(\lambda, j) = \frac{I_{mod}(\lambda, j, \boldsymbol{\rho})}{I_{mod}^{ref}(\lambda, \boldsymbol{\rho}_{ref})}, \tag{2}$$

where $I_{mod}(\lambda, j, \boldsymbol{\rho})$ are modeled radiances and $I_{mod}^{ref}(\lambda, \boldsymbol{\rho}_{ref})$ is a model reference spectrum. The gas density profiles $\boldsymbol{\rho}$ are adjusted iteratively, and after every iteration a new and better agreement is obtained between (1) and (2).

[16] A typical background atmosphere is assumed when we calculate the model reference spectrum. Obviously, even the best estimate differs from the true state of the atmosphere and typically produces systematic bias to the retrieved profiles. This effect is studied later in Section 3.

2.1. Spectral Fitting

[17] At every layer, assuming that the measured transfer spectra (1) are independent, the sum of squared residuals is defined as

$$\chi^2 = (R_{mod} - R_{obs})^T C^{-1} (R_{mod} - R_{obs}), \tag{3}$$

where C is a covariance matrix. The covariance matrix includes contributions from the measurement and modeling errors. The errors at different wavelengths are assumed to be uncorrelated, which leads to a diagonal covariance matrix. The modeling error describes our inability to model limb scatter observations perfectly, mainly because of multiple scattering of the atmosphere (see Section 2.3). The modeling error is estimated as a function of wavelength
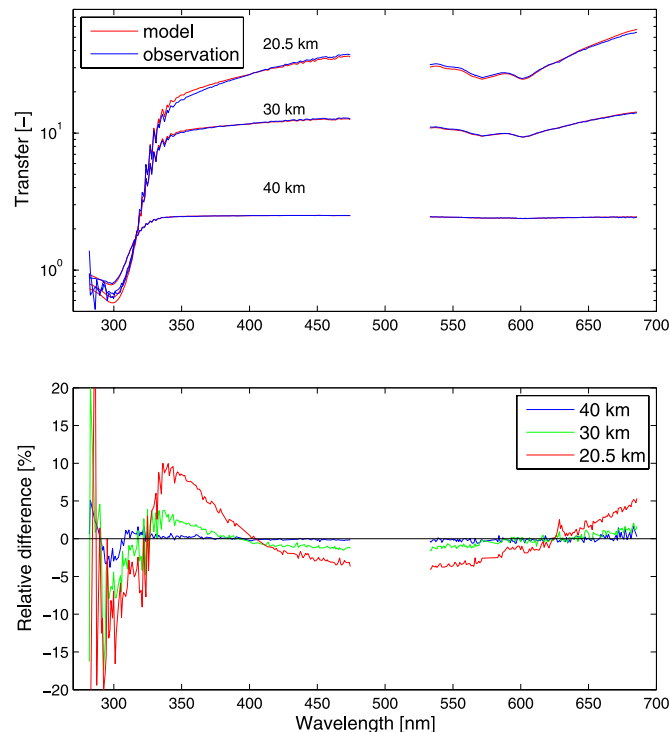
Figure 1. OSIRIS spectral fits in the UV-visible region after the first peeling loop (upper panel) and the corresponding residuals (lower panel). The NO_2 profile is kept fixed during the iterations while ozone, air density, and aerosols are allowed to vary. Note that OSIRIS does not record the spectra between 470 and 530 nm and the wavelengths beyond 680 nm are not used in the fitting.

and altitude using the Monte Carlo radiative transfer model Siro [*Oikarinen et al.*, 1999].

[18] The fitting problem (3) is solved using an initial guess for the densities and the Levenberg-Marquardt algorithm [*Levenberg*, 1944; *Marquardt*, 1963; *Gill et al.*, 1981] to find the best fit. The Levenberg-Marquardt algorithm is commonly used in non-linear curve fitting problems. It finds the minimum of (3) by combining techniques of gradient descent and the inverse-Hessian optimization. The algorithm also provides error estimates for the fitted parameter values.

[19] Figure 1 presents examples of the spectral fits after the first peeling loop. The relative differences increase at lower altitudes, but generally a good consistency can be found between the model and the measurement. Completely flawless agreement is very difficult to achieve because the wavelength band used is relatively wide (over 400 nm) and the radiance is governed by numerous wavelength-dependent phenomena.

[20] Limb scatter measurements include relatively small noise, but the model is unable to describe observations perfectly. Because the modeling of the atmosphere is a very complex problem indeed, it is possible (or even evident) that the model lacks some processes, or that they have been taken account in too simplified way. These factors can be,

for example, a missing gas, incorrect cross sections (e.g., an uncertain temperature profile), too simplified aerosol model, or incorrect albedo model. Furthermore, an insufficient modeling of the diurnal effects and errors due to geometry simplifications may contribute. The (current) model presents our best understanding of the physics behind the observations. The clear structures in the residuals indicate that there is still work to do in order to improve the model in the future.

[21] In addition, the measurement data contain errors and therefore the model is not able to describe observations exactly. The spectral fit to the OSIRIS data shown in Figure 1 was done with a model including an accurate single scattering model and an approximate multiple scattering correction. The model will be described more extensively in Section 2.3.

2.2. Improved NO_2 Retrieval

[22] The original idea by *Auvinen et al.* [2002] was to use the whole spectral range of OSIRIS and retrieve all the desired trace gas densities simultaneously. However, this approach seemed to work only with simulated data. When inverting real OSIRIS data, the NO_2 retrievals were usually of poor quality although the inverted ozone profiles were proper. The old retrieval often produced a bias of several
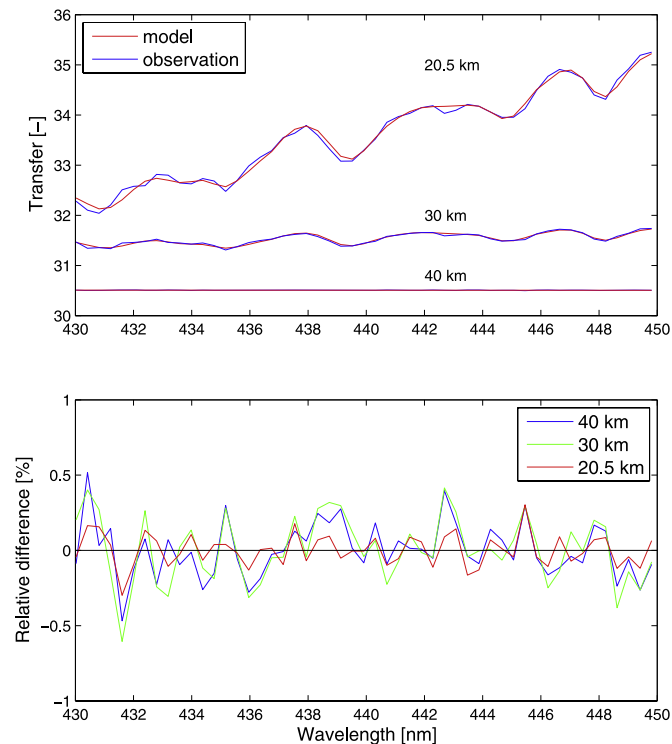
Figure 2. Fits in the NO_2 retrieval band after the second peeling loop (upper panel) and the corresponding residuals (lower panel). The 30 km and 40 km spectra are scaled to fit in the figure with the 20.5 km spectra. The ozone profile (retrieved from the first peeling loop) is kept fixed during the iterations while NO_2, air density and aerosols are allowed to vary.

hundreds of percent to the NO_2 profiles below 30 km compared with the results from the other OSIRIS retrieval algorithms or NO_2 measurements by other instruments [*Tukiainen*, 2006].

[23] Using the whole spectrum for the fit has both advantages and disadvantages. When a large wavelength band is used, the information content can naturally be maximized, but on the other hand the modeling issues become more critical. Since ozone is a strong absorber, minor modeling uncertainties are not as crucial for the ozone retrieval as they are for minor absorbers like NO_2. As there are uncertainties in the OSIRIS modeling, we noticed that the NO_2 retrieval benefits from using a shorter wavelength band where the signal-to-noise ratio is more optimal for the NO_2 retrieval. Also the band should be short enough that the wavelength dependent modeling errors are not dominating the retrieval.

[24] In the present version of MOP, the retrieval of major absorbers and scatterers (ozone, air, aerosols) is separated from the retrieval of minor ones (NO_2, and others). For NO_2, the wavelength band of 430–450 nm is selected due to strong NO_2 absorption fingerprint in this region. Furthermore, ozone as well as other species absorb weakly in this band and it is also free of strong Fraunhofer lines. In theory,

the Fraunhofer lines should cancel out when we apply equation (1), but a small residual may still be left over. This is the tilt effect recognized by *Sioris et al.* [2003]. It basically arises from the different spectral slopes of the radiances at different tangent heights and from the finite spectral resolution of the instrument. Thus it is safest to avoid strong Fraunhofer lines where the effect is largest. In this wavelength band, all of the 50 available wavelengths were used in the spectral fitting.

[25] Because of the distinct fitting windows for the strong and weak absorbers, we have to run two separate peeling loops. During the first peeling loop, we retrieve only ozone, aerosol, and neutral air. The NO_2 profile is summoned from a climatology (U.S. Standard Atmosphere 1976). During the second loop, we retrieve NO_2. The ozone profile, retrieved from the first peeling loop, is now fixed, but aerosols and neutral air are again allowed to vary freely to obtain a good NO_2 fit. Figure 2 shows NO_2 fits from a single scan at three different tangent heights.

[26] This kind of iterative solving of the parameters is justified, because the use of the fixed NO_2 profile in the first peeling loop has little effect on the outcomes. Nevertheless, it is not totally insignificant. The use of the NO_2 climatology in the first peeling loop seems to result in roughly 2%

bias at the ozone peak. This bias is possible to eliminate by adding a third peeling loop and retrieve ozone, neutral air, and aerosols again with the retrieved NO_2 profile, but this would double the required computing time. With just two peeling loops, it already takes a couple of minutes to invert one scan (AMD Athlon 1800+ CPU). This means that we need around six months to process the current OSIRIS data set (2001–2007) with the present computing facilities for OSIRIS at Finnish Meteorological Institute. The third peeling loop is possible to implement later in the future, if more computing resources are provided to the operative OSIRIS processing.

2.3. Multiple Scattering Correction

[27] Taking account the multiple scattering effects is a crucial part of limb scatter retrieval methods. Therefore the multiple scattering correction of the MOP method is given here, though the correction method has remained the same since the previous paper by *Auvinen et al.* [2002].

[28] In the limb scatter geometry, multiple scattering can constitute 10–50% of the observed radiance at visible wavelengths [*Oikarinen et al.*, 1999]. Its proportion of the total radiance is strongly dependent on wavelength as the ratio of multiple to total scattering increases steeply at wavelengths greater than 310 nm [*Oikarinen et al.*, 1999]. In addition, the multiple scattering contribution depends on the tangent height, surface albedo, solar angles, and composition of the atmosphere itself. For these reasons, a mere single scattering radiative transfer model is not generally satisfactory to describe scattering and absorption effects in limb scatter problems. However, taking the multiple scattering effects into account complicates limb scatter problems significantly and certainly increases computational costs, which leads to some kind of compromise between the modeling accuracy and the available computation time.

[29] During the fitting procedure in each layer, we have to use a few (usually 2–10) iterations, and as many forward model calls, before the Levenberg-Marquard algorithm finds good enough agreement between the model and the measurement. A full 3-D radiative transfer model operating in multiple scattering mode would be too slow to use and hence we must seek faster solutions. One way would be to reduce the total number of wavelengths used (from around 300 to only a few) in the first peeling loop. This solution would not exploit the whole bandwidth of the OSIRIS instrument and the altitude range of the ozone retrieval would shrink.

[30] A practical approach, discussed in the original publication by *Auvinen et al.* [2002], is to use a single scattering forward model during the fitting iterations and include multiple scattering effects using precalculated look-up tables. The single scattering forward model built into the MOP-inversion module solves the radiative transfer by numerical integration and is computationally efficient to run.

[31] The look-up tables used in the MOP inversion contain modeled (single- and multiple scattered) radiances calculated as a function of tangent height, solar angles, season, albedo, and latitude. The look-up tables are produced using the LIMBTRAN [*Griffioen and Oikarinen*, 2000] forward model.

[32] In order to take advantage of the precalculated look-up tables, we divide the modeled transfer spectra (2) into two parts:

$$R_{\mod} = \frac{I_{\mod}^{\mathrm{ref}}}{I_{\mod}} = \frac{I_{\mod}^{\mathrm{ss}}(\lambda, j, \rho)}{I_{\mod}^{\mathrm{ref}}(\lambda, \rho_{\mathrm{ref}})} M(\lambda, j), \qquad (4)$$

where $I_{\mod}^{\mathrm{ss}}(\lambda, j, \rho)$ is the dynamic single scattering term which is adjusted iteratively during the fitting procedure. The model reference radiance $I_{\mod}^{\mathrm{ref}}(\lambda, \rho_{\mathrm{ref}})$ in the denominator of (4) is also calculated with LIMBTRAN (including multiple scattering). The second term on the right side, $M(\lambda, j)$, is the static part which comes from the look-up tables and is kept fixed during the iterations. This correction term is defined as the modeled multiple scattering radiance (total radiance) divided by the corresponding single scattering radiance:

$$M(\lambda, j) = \frac{I_{\mathrm{ms}}\left(\lambda, j, \rho_{\mathrm{prior}}\right)}{I_{\mathrm{ss}}\left(\lambda, j, \rho_{\mathrm{prior}}\right)}. \qquad (5)$$

where ρ_{prior} are the gas densities for the standard atmosphere used in the LIMBTRAN radiance simulations.

3. Sensitivity Study

[33] The calculation of the model reference spectrum $I_{\mod}^{\mathrm{ref}}(\lambda, \rho_{\mathrm{ref}})$ requires at least ozone and neutral air profiles up to the upper limit of the LIMBTRAN atmosphere (90 km). Other species can be ignored as they have a negligible impact on the radiance at ∼50 km. The effect of possibly incorrect ozone and neutral air profiles in the model reference was studied using single scattering simulations and a 50 km reference tangent height.

[34] Radiances were first simulated using the internal single scattering kernel of the MOP algorithm for a given atmospheric composition. These simulated radiances were then used as an input, and the MOP algorithm was run to resolve the original number densities. The model reference spectrum was also simulated with the single scattering model, but with slightly modified ozone and air profiles to study the effect on the outcomes.

[35] The results from the simulations are shown in Figure 3 and Figure 4. We clearly see how the modified ozone density profile in the model reference creates biased ozone profiles after inversion at the upper layers (Figure 3). This bias diminishes rapidly, and is already negligible below the 45 km altitude. The effect on the NO_2, air, and aerosol profiles is not as large. Instead, if the neutral air profile of the reference differs from truth, it creates a roughly equal difference in the inverted air product (Figure 4). The aerosol product is also affected considerably, but NO_2 and ozone errors are only a few percent at most. It should be noticed from Figure 4 that the MOP neutral air and aerosols are anticorrelated.

[36] As a summary, if the ozone estimate in the reference calculation is wrong, it will produce a bias in the retrieved ozone profile, but the effect is large only at the few uppermost layers. An incorrect air density estimate, on the other hand, creates a corresponding bias in the retrieved air
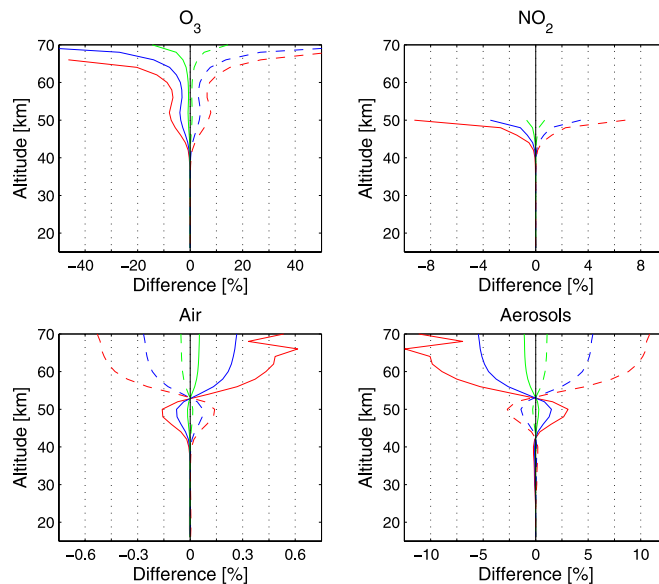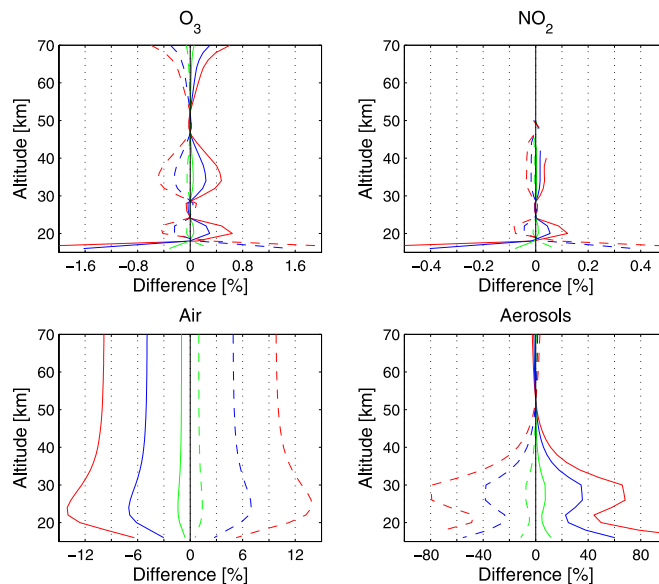
Figure 3. Effect of ozone density in the model reference calculation. The plots present the relative difference between the true and the inverted profiles. Ozone densities in the model reference were modified by −10% (red), −5% (blue), −1% (green), +1% (green dashed), +5% (blue dashed), and +10% (red dashed) compared with the true state of the atmosphere. The simulation was run with single scattering and using 50 km as the reference altitude.

density and can also affect the retrieved aerosol profile by tens of percent.

[37] A faulty ozone or air density profile in the model reference calculation is not the only uncertainty in the

retrieval. There are several other possible sources of error such as incorrect albedo or aerosol model, imprecise satellite pointing, and polarization sensitivity. The impact of these on the limb scatter ozone retrievals has been studied

Figure 4. Same as Figure 3 but for the effect of air densities in the model reference.
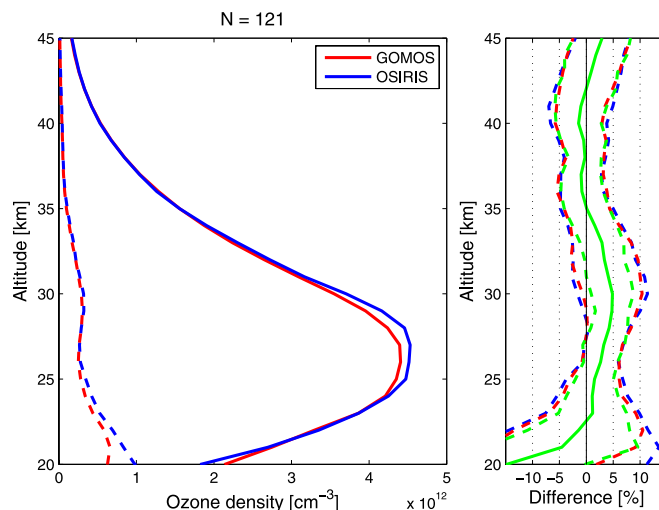
Figure 5. Comparison of GOMOS nighttime and OSIRIS daytime ozone profiles in the stratosphere between latitudes 35°S and 35°N. The coincidence criteria for the individual matches in time was less than one day, in latitude less than 1°, and in longitude less than 1°. The blue and red curves on the left present the medians of 121 coincidences and the dashed curves are the corresponding standard deviations. The green solid curve is the median of the individual relative differences defined as Comparison of OSIRIS and GOMOS aerosols between latitudes 35°S and 35°N. Coincidence limits for the individual matches are the same than in Figure 5. The blue and red curves present the medians of 110 coincidences and the dashed curves are the corresponding standard deviations. The green solid curve is the median of the individual relative differences defined as (OSIRIS-GOMOS)/GOMOS100%. The green dashed curves around the median show the semi-interquartile deviation of the differences. The blue and red dashed curves around the median show the corresponding deviations of the OSIRIS and GOMOS profiles.

comprehensively by *Loughman et al.* [2005], *von Savigny et al.* [2005], and *Oikarinen* [2001].

4. Validation of Profiles

[38] A preliminary validation of ozone, NO_2, neutral air, and aerosol profiles was carried out using OSIRIS data from the year 2003. The NO_2 profiles are now much more reasonable than with the earlier version of the MOP method. The ozone, air, and aerosol profiles also seem to be realistic.

4.1. Ozone

[39] The diurnal variation of stratospheric ozone is insignificant below around 45 km [*Brasseur and Solomon*, 2005]. Thus we can compare day and nighttime profiles with similar geolocation if the time difference of measurements is on the order of days. Ozone is also generally more stable at low and midlatitudes than in polar regions.

[40] OSIRIS ozone profiles inverted using the MOP method were compared with the GOMOS (Global Ozone Monitoring by Occultation of Stars) nighttime profiles between 20 and 45 km. GOMOS is a stellar occultation instrument on board the ESA's Envisat satellite, launched in February 2002 [*Kyrölä et al.*, 2004]. The vertical sampling resolution of GOMOS is 0.5–1.7 km and the vertical resolution of GOMOS ozone profiles is 2–3 km depending on the altitude. According to *Meijer et al.* [2004], there is

less than 5% bias in the GOMOS nighttime ozone profiles. GOMOS nighttime measurements are very accurate and there is less than 5% bias in the ozone profiles.

[41] Coincidences were selected from the year 2003 using a latitude band from 35°S to 35°N. The latitude and longitude difference of the 121 matches was less than 1°, and the time difference less than one day. The local time of the OSIRIS coincidences is close to either 6 am or 6 pm.

[42] Figure 5 shows the result of the comparison. Median profiles of both distributions are plotted with the corresponding standard deviations. The green solid curve on the right panel shows the median of the individual relative differences. The individual differences were defined as (OSIRIS-GOMOS)/GOMOS100%. The green dashed lines around the median present the semi-interquartile deviation (SID):

$$SID = \frac{|Q_3 - Q_1|}{2}, \tag{6}$$

where Q_1 and Q_3 are the 25th and 75th percentiles. The area between the ±1 SID lines includes 50% of the data points. The blue dashed lines around the median present the relative semi-interquartile deviation of the retrieved OSIRIS profiles (i.e., the natural variability of the atmosphere). See also the upper panels of Figure 6 for examples of typical individual comparisons between OSIRIS and GOMOS ozone profiles.
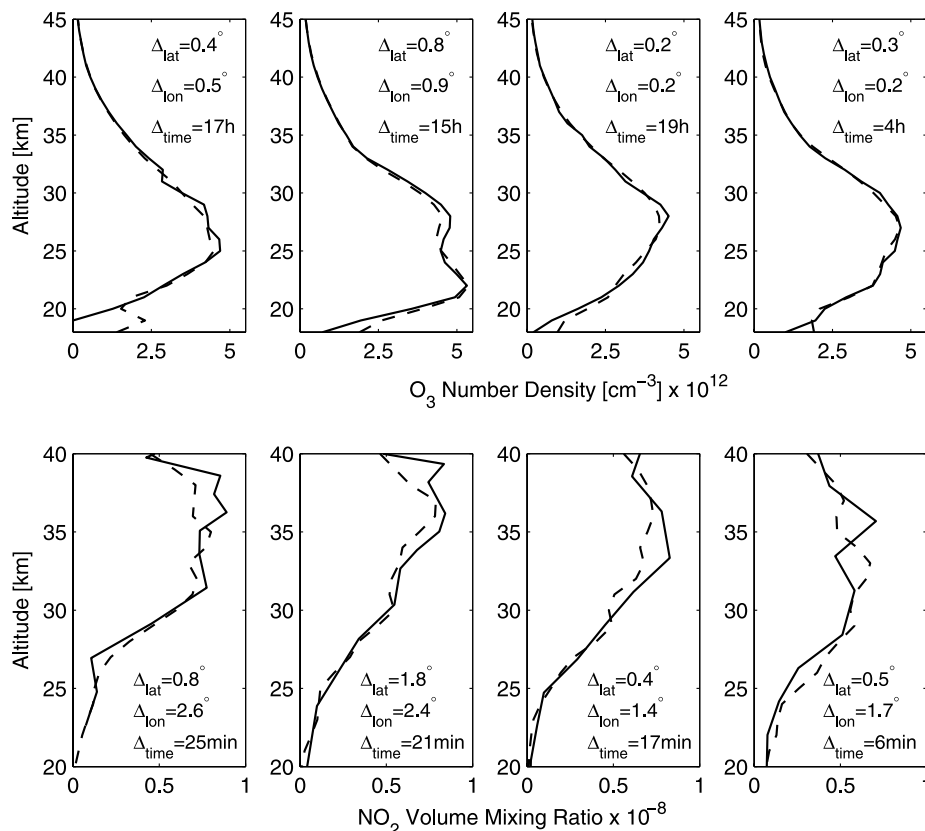
Figure 6. Upper panels: typical individual ozone profile comparisons between OSIRIS (solid) and GOMOS (dashed). Lower panels: typical individual NO_2 profile comparisons between OSIRIS (solid) and HALOE (dashed).

[43] The mean could be also used instead of the median, but more attention should be paid to flag abnormal data points. In practice, there was no significant difference between the two (mean or median). The median of the individual relative differences was less than 5% between 21 and 45 km proving that the two different instruments provide very consistent observations of stratospheric ozone. The median difference increases up to 15% at the lowest layer (OSIRIS measuring less ozone than GOMOS).

4.2. Nitrogen Dioxide

[44] The behavior of stratospheric NO_2 is characterized by strong diurnal variation. The NO_2 concentration decreases rapidly in the morning as the sun rises. A similar, but positive, change happens when the sun sets in the evening. Thus if we compare NO_2 measurements with the solar zenith angles close to 90°, the difference from the diurnal cycle should be minimized as well as possible.

[45] OSIRIS NO_2 measurements were compared with the sunset measurements of the HALOE (The Halogen Occultation Experiment) instrument [*Russell et al.*, 1993]. HALOE was a solar occultation instrument launched in 1991 and it was operational until November 2005. The

HALOE retrieval gives the NO_2 volume mixing ratio while the outcomes of the MOP method are number densities. Therefore OSIRIS NO_2 densities were converted into mixing ratios using retrieved neutral air profiles. The vertical resolution of HALOE NO_2 profiles is about 2–3 km, which is similar to the resolution of OSIRIS profiles inverted with the MOP method.

[46] The coincidence criteria for a comparable OSIRIS and HALOE measurement in time was less than 30 min, in latitude less than 2°, and in longitude less than 3°. The solar zenith angles of the OSIRIS measurements were chosen to be between 85° and 90°. With this kind of differences in time and space, we found 28 OSIRIS and HALOE matches. Thus we get reasonable statistics for the validation and the profiles still describe roughly the same air mass. The coincidences were found between latitudes 15°S and 50°N from the year 2003. In all these cases, the OSIRIS measurement comes first in the time domain. The local times of the OSIRIS coincidences are between 5.30 pm and 6.15 pm.

[47] As we compare NO_2 profiles from solar occultation and limb scatter instruments, we are forced to accept a small difference in the solar zenith angles. This will inevitably produce some difference to the results, but the effect should
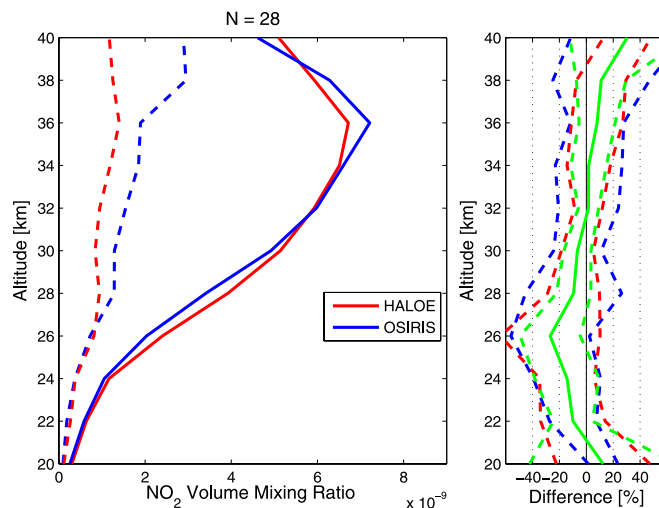
Figure 7. Comparison of HALOE sunset measurements and OSIRIS NO_2 profiles in the stratosphere. The coincidence criteria for the individual matches was in time less than 30 min, in latitude less than $2°$, and in longitude less than $3°$. The blue and red curves on the left present the medians of 28 coincidences and the dashed curves are the corresponding standard deviations. The green solid curve is the median of the individual relative differences defined as (OSIRIS-HALOE)/HALOE100%. The green dashed curves around the median show the semi-interquartile deviation of the differences. The blue and red dashed curves around the median show the corresponding deviations of the OSIRIS and HALOE profiles.

be only small because the solar zenith angles of the OSIRIS profiles used in this study are close to $90°$. The other source of error is the impact of the lack of horizontal homogeneity on limb scatter retrievals [*McLinden et al.*, 2006]. The solar zenith angle may vary along the line of sight causing species with diurnal cycle to vary as well, and the horizontal homogeneity assumption fails. This effect is generally largest during twilight and below 25 km. The diurnal effects due to changing solar zenith angle along the line of sight are not taken account in the MOP retrieval. For occultation instruments, the problem is less complicated and the HALOE retrieval does account for it. This may explain some of the differences between OSIRIS and HALOE NO_2 below 25 km, but the total effect is rather hard to quantify.

[48] Figure 7 shows the comparison of the 28 OSIRIS and HALOE NO_2 profiles between 20 and 40 km. The observations of the instruments are quite consistent and there is no particular bias between the instruments. The median difference of the individual profiles is usually less than 20%, which is a quite good agreement. OSIRIS seems to measure more NO_2 than HALOE at the peak and less than HALOE at the lower stratosphere. The lower panels of Figure 6 show typical individual comparisons between OSIRIS and HALOE NO_2 profiles.

4.3. Air

[49] Scattering from the neutral molecules of the atmosphere follows Rayleigh theory with approximately λ^{-4} dependency. This basically determines the signal level of the observed radiative transfer spectrum. Air density profiles retrieved with the MOP method were compared with

the analysis data from European Center for Medium-Range Weather Forecast (ECMWF). The MOP air profiles are from the first peeling loop, inverted together with ozone and aerosols. The profiles were selected randomly covering all latitudes from the year 2003.

[50] Figure 8 shows results of the neutral air comparison. The profiles were compared for altitudes between 20 and 65 km. The median of the relative individual differences is around 5% at almost all altitudes. OSIRIS appears to be biased low compared with the ECMWF data. The comparison result indicates that the neutral air retrieval of the MOP method is also quite accurate.

4.4. Aerosols

[51] Stratospheric aerosols contribute to radiance through scattering and absorption. The aerosol modeling in the MOP method is done using λ^{-1} scattering cross section and well-known Henyey-Greenstein phase function. This is the classical, but not very sophisticated, way to model stratospheric aerosols.

[52] Figure 9 shows the comparison of GOMOS and OSIRIS aerosol number density profiles. The MOP aerosol profiles are from the first peeling loop. OSIRIS seems to observe as much as 80% more aerosols than GOMOS. The shapes of the median profiles are rather similar but the reason for the bias has yet to be worked out.

[53] These results indicate (partly expected) difficulties in the validation of the MOP aerosol product. As noticed earlier, the outcome of the MOP aerosol retrieval is sensitive to the model reference spectrum. It should also be noted that the GOMOS retrieval method uses a second-order polyno-
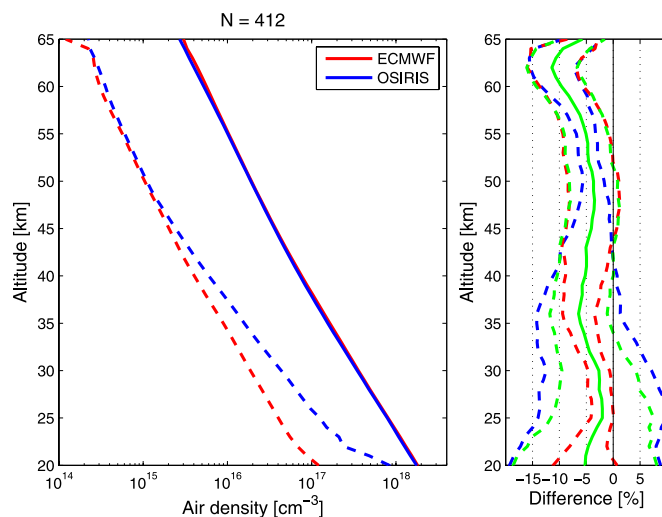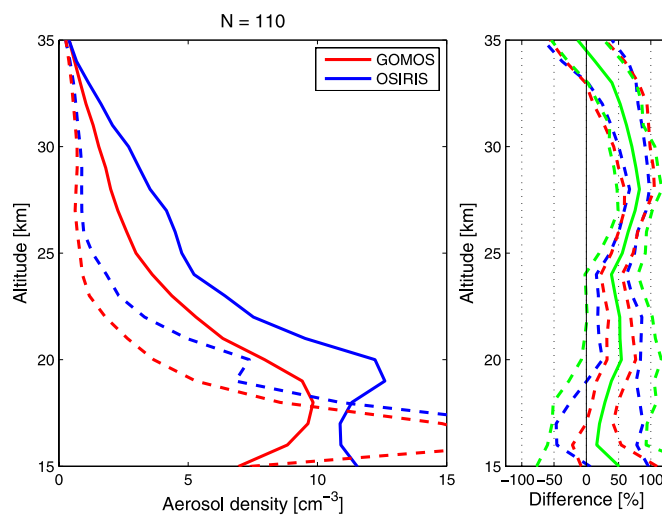
Figure 8. Comparison of OSIRIS neutral air density and ECMWF analysis data. The blue and red curves on the left present the medians of 412 coincidences and the dashed curves are the corresponding standard deviations. The green solid curve is the median of the individual relative differences defined as (OSIRIS-ECMWF)/ECMWF100%. The green dashed curves around the median show the semi-interquartile deviation of the differences. The blue and red dashed curves around the median show the corresponding deviations of the OSIRIS and ECMWF profiles.

mial aerosol model instead of the λ^{-1} form used here. This polynomial model could be also tried with OSIRIS as it might explain some of the differences. For more informa-

tion about GOMOS aerosols, see *Vanhellemont et al.* [2005a, 2005b].

[54] However, the underlying difficulty is the unknown (aerosol) particle shape and size distribution in the atmo-

Figure 9. Comparison of OSIRIS and GOMOS aerosols between latitudes 35°S and 35°N. Coincidence limits for the individual matches are the same than in Figure 5. The blue and red curves present the medians of 110 coincidences and the dashed curves are the corresponding standard deviations. The green solid curve is the median of the individual relative differences defined as (OSIRIS-GOMOS)/ GOMOS100%. The green dashed curves around the median show the semi-interquartile deviation of the differences. The blue and red dashed curves around the median show the corresponding deviations of the OSIRIS and GOMOS profiles.

sphere, which makes the problem notoriously difficult. We are forced to do simplifications in the retrieval, and in the end we still lack a reliable reference for the basis of the validation. The aerosol model selection itself is a difficult decision. One promising idea is to apply Bayesian model selection by using a Markov chain Monte Carlo technique as proposed by *Laine et al.* [2007].

5. Conclusions

[55] In this study, the MOP inversion method was used to retrieve ozone, NO_2, neutral air, and aerosol densities from the OSIRIS limb scatter measurements. Compared with the other limb scatter retrieval methods, there are a few unique aspects in MOP. The MOP method uses a wide spectral band and several hundreds wavelengths to retrieve ozone, air, and aerosol number densities. Thus the maximum amount of information from the data is used, while the use of a priori information is minimized. The same method is also used for the whole altitude range of the retrievals.

[56] A sensitivity study with the simulated data showed that the upper part of the retrieved ozone profile becomes biased if the ozone estimate in the model reference spectrum is incorrect. The aerosol retrieval is quite sensitive to a faulty neutral air estimate in the reference calculation and this can lead to errors up to tens of percents. Roughly equal error in the retrieved neutral air is also produced at all altitudes. However, errors in the NO_2 profile due to incorrect model reference are always insignificant.

[57] Validation against other satellite instruments demonstrated the strength of the MOP method in practice. A good agreement was found between OSIRIS daytime and GOMOS nighttime ozone profiles. The median of the relative individual differences is less than 5% between 21 and 45 km. Above 45 km the diurnal variation of ozone prevented comparisons between the instruments. The quality of this mesospheric part of the ozone retrieval should be confirmed in the future.

[58] OSIRIS NO_2 profiles between 20 and 40 km were consistent with the HALOE profiles despite the challenging twilight conditions for OSIRIS. The comparison between solar occultation and limb viewing instruments is always difficult because the NO_2 concentration experiences a large transition near sunrise and sunset.

[59] The MOP neutral air retrieval seems to result in roughly 5% negative bias compared with the ECMWF analysis data. On the other hand, there is a $20-80\%$ positive bias in the MOP aerosol number densities compared with the GOMOS aerosols. The anticorrelation between neutral air and aerosols, which is apparent from the sensitivity study in Section 3 (see, e.g., Figure 4), could explain most of the observed differences. Thus it could be possible to get better aerosol results by fixing the MOP neutral air density to the ECMWF data, or by setting some interval for the MOP air density values to vary around the ECMWF values. This issue remains to be solved in the future studies.

[60] Ozone, neutral air, and aerosols are retrieved using the whole spectral range of the instrument, so the retrievals may go up to 70 km altitude (corresponding to the upper limit of most OSIRIS scans). For the NO_2 retrieval, it was crucial to use a narrow wavelength band. This approach is easily expanded to retrieve also other trace gases such as BrO and OClO by selecting another wavelength window and adding an extra peeling iteration loop. The retrieved ozone and NO_2 profiles may then be used in this iteration. The underlying challenge is that the concentrations of BrO and OClO are an order of magnitude smaller than that of NO_2 and the absorption fingerprint is easily masked by noise. However, *Krecl et al.* [2006] have shown that the retrieval of OClO should be possible.

[61] It should also be noted that with the current Level 1 data we do not find any altitude shift between OSIRIS and GOMOS ozone profiles. GOMOS uses a star tracker to achieve a pointing accuracy of some tens of meters. Thus it provides a good reference to validate the pointing performance of Odin/OSIRIS.

References

Auvinen, H., L. Oikarinen, and E. Kyrölä (2002), Inversion algorithms for recovering minor species densities from limb scatter measurements at UV-visible wavelengths, *J. Geophys. Res.*, *107*(D13), 4172, doi:10.1029/2001JD000407.

Bovensmann, H., J. P. Burrows, M. Buchwitz, J. Frerick, S. Noel, V. V. Rozanov, K. V. Chance, and A. P. H. Goede (1999), SCIAMACHY: Mission objectives and measurement modes, *J. Atmos. Sci.*, *56*, 127– 150, doi:10.1175/1520-0469.

Brasseur, G. P., and S. Solomon (2005), *Aeronomy of the Middle Atmosphere*, 3rd revised and enlarged ed., Springer, Dordrecht.

Flittner, D. E., P. K. Bhartia, and B. M. Herman (2000), O_3 profiles retrieved from limb scatter measurements: Theory, *Geophys. Res. Lett.*, *27*, 2601–2604, doi:10.1029/1999GL011343.

Gill, P. R., W. Murray, and M. H. Wright (Eds.) (1981), *Practical Optimization*, Academic Press.

Griffioen, E., and L. Oikarinen (2000), LIMBTRAN: A pseudo three-dimensional radiative transfer model for the limb-viewing imager OSIRIS on the Odin satellite, *J. Geophys. Res.*, *105*(D24), 29,717– 29,730, doi:10.1029/2000JD900566.

Haley, C. S., C. von Savigny, S. Brohede, C. E. Sioris, I. C. McDade, E. J. Llewellyn, and D. P. Murtagh (2003), A comparison of methods for retrieving stratospheric ozone profiles from OSIRIS limb-scatter measurements, *Adv. Space Res.*, *34*, 769–774.

Haley, C. S., S. M. Brohede, C. E. Sioris, E. Griffioen, D. P. Murtagh, I. C. McDade, P. Eriksson, E. J. Llewellyn, A. Bazureau, and F. Goutail (2004), Retrieval of stratospheric O_3 and NO_2 profiles from Odin Optical Spectrograph and Infrared Imager System (OSIRIS) limb-scattered sunlight measurements, *J. Geophys. Res.*, *109*, D16303, doi:10.1029/2004JD004588.

Kaiser, J., C. von Savigny, K.-U. Eichmann, S. Noel, H. Bovensmann, and J. Burrows (2004), Satellite-pointing retrieval from atmospheric limb-scattering of solar UV-B radiation, *Can. J. Phys.*, *82*, doi:10.1139/P04-071.

Krecl, P., C. S. Haley, J. Stegman, S. M. Brohede, and G. Berthet (2006), Retrieving the vertical distribution of stratospheric OClO from Odin/OSIRIS limb-scattered sunlight measurements, *Atmos. Chem. Phys*, *6*, 1879–1894.

Kyrölä, E., et al. (2004), GOMOS on Envisat: An overview, *Adv. Space Res.*, *33*, 1020–1028.

Laine, M., J. Tamminen, E. Kyrölä, and H. Haario (2007), *Aerosol model selection and uncertainity modelling by RJMCMC technique*, Envisat ACVE-3 workshop proceedings.

Levenberg, K. (1944), A method for the solution of certain problems in least squares, *Q. Appl. Math.*, *2*, 164–168.

Llewellyn, E. J., N. D. Lloyd, D. A. Degenstein, et al. (2004), The OSIRIS instrument on the Odin spacecraft, *Can. Phys, J.*, *82*, 411–422, doi:10.1139/P04-005.

Loughman, R. P., D. E. Flittner, B. M. Herman, P. K. Bhartia, E. Hilsenrath, and R. D. McPeters (2005), Description and sensitivity analysis of a limb

scattering ozone retrieval algorithm, *J. Geophys. Res.*, *110*, D19301, doi:10.1029/2004JD005429.

Marquardt, D. (1963), An algorithm for least-squares estimation of nonlinear parameters, *J. Soc. Indust. Appl. Math.*, *11*, 431–441.

McDade, I. C., K. Strong, C. S. Haley, J. Stegman, D. P. Murtagh, and E. J. Llewellyn (2002), A method for recovering stratospheric minor species densities from the Odin/OSIRIS scattered-sunlight measurements, *Can J. Phys.*, *80*, 395–408.

McLinden, C. A., C. S. Haley, and C. E. Sioris (2006), Diurnal effects in limb scatter observations, *J. Geophys. Res.*, *111*, D14302, doi:10.1029/2005JD006628.

McPeters, R. D., S. J. Janz, E. Hilsenrath, T. L. Brown, D. E. Flittner, and D. F. Heath (2000), The retrieval of O_3 profiles from limb scatter measurements: Results from the shuttle ozone limb sounding experiment, *Geophys. Res. Lett.*, *27*, 2597–2600.

Meijer, Y. J., et al. (2004), Pole-to-pole validation of ENVISAT/GOMOS ozone profiles using data from ground-based and balloon-sonde measurements, *J. Geophys. Res.*, *109*, D23305, doi:10.1029/2004JD004834.

Oikarinen, L. (2001), Polarization of light in UV-visible limb radiance measurements, *J. Geophys. Res.*, *106*(D2), 1533–1544, doi:10.1029/2000JD900442.

Oikarinen, L., E. Sihvola, and E. Kyrölä (1999), Multiple scattering radiance in limb-viewing geometry, *J. Geophys. Res.*, *104*(D24), 31,261–31,274, doi:10.1029/1999JD900969.

Petelina, S. V., et al. (2004), Comparison of the Odin/OSIRIS stratospheric ozone profiles with coincident POAM III and ozonesonde measurements, *Geophys. Res. Lett.*, *31*, L07104, doi:10.1029/2003GL019299.

Platt, U. (1994), Differential optical absorption spectroscopy (DOAS), air monitoring by spectroscopic techniques, *Chem. Anal. Ser.*, *127*, 27–84.

Rault, D. F. (2005), Ozone profile retrieval from Stratospheric Aerosol and Gas Experiment (SAGE III) limb scatter measurements, *J. Geophys. Res.*, *110*, D09309, doi:10.1029/2004JD004970.

Rodgers, C. D. (2000), *Inverse Methods for Atmospheric sounding: Theory and Practice*, World Scientific, Singapore.

Rohen, G., C. Savigny, M. Sinnhuber, E. J. Llewellyn, J. W. Kaiser, C. H. Jackman, M.-B. Kallenrode, J. Schröter, K.-U. Eichmann, H. Bovensmann, and J. P. Burrows (2005), Ozone deflection during the solar proton events of October/November 2003 as seen by SCIAMACHY, *J. Geophys. Res.*, *110*, A09S39, doi:10.1029/2004JA010984.

Rusch, D. W., G. H. Mount, J. M. Zawodny, C. A. Barth, G. J. Rottman, R. J. Thomas, G. E. Thomas, R. W. Sanders, and G. W. Lawrence (1983), Temperature measurements in the earth's stratosphere using a limb scanning visible light spectrometer, *Geophys. Res. Lett.*, *10*, 261–264.

Rusch, D. W., G. H. Mount, C. A. Barth, R. J. Thomas, and M. T. Callan (1984), Solar Mesosphere Explorer ultraviolet spectrometer: Measure-

ments of ozone in the 1.0–0.1 mbar region, *J. Geophys. Res.*, *89*, 11,677–11,678.

Russell, J. M., L. L. Gordley, J. H. Park, S. R. Drayson, W. D. Hesketh, R. J. Cicerone, A. F. Tuck, J. E. Frederick, J. E. Harries, and P. J. Crutzen (1993), The halogen occultation experiment, *J. Geophys. Res.*, *98*(D6), 10,777–10,797, doi:10.1029/93JD00799.

Sioris, C. E., C. S. Haley, C. A. McLinden, and et al. (2003), Stratospheric profiles of nitrogen dioxide observed by optical spectrograph and infrared imager system on the Odin satellite, *J. Geophys. Res.*, *108*(D7), 4215, doi:10.1029/2002JD002672.

Strong, K., B. M. Joseph, R. Dosanjh, I. C. McDade, C. A. McLinden, J. C. S. J. McConnell, D. P. Murtagh, and E. J. Llewellyn (2002), Retrieval of vertical concentration profiles from OSIRIS UV-visible limb spectra, *Can. J. Phys.*, *80*, 409–434.

Tukiainen, S. (2006), *Development and Validation of the OSIRIS inversion module*, Master's Thesis at the Helsinki University of Technology, Laboratory of Space Technology.

Vanhellemont, F., et al. (2005a), A first comparison of GOMOS aerosol extinction retrievals with other measurements, *aisr*, *36*, 894–898, doi:10.1016/j.asr.2005.04.094.

Vanhellemont, F., et al. (2005b), A 2003 stratospheric aerosol extinction and PSC climatology from GOMOS measurements on Envisat, *Atmos. Chem. Phys.*, *5*, 2413–2417.

von Savigny, C., et al. (2003), Stratospheric ozone profiles retrieved from limb scattered sunlight radiance measured by the OSIRIS instrument on the Odin satellite, *Geophys. Res. Lett.*, *30*(14), 1755, doi:10.1029/2002GL016401.

von Savigny, C., I. C. McDade, E. Griffioen, C. S. Haley, C. E. Sioris, and E. J. Llewellyn (2005), Sensitivity studies and first validation of stratospheric ozone profile retrievals from Odin/OSIRIS observations of limb-scattered solar radiation, *Can. J. Phys.*, *83*, 957–972, doi:10.1139/P05-041.

————————

H. Auvinen, Lappeenranta University of Technology, P.O. Box 20, FI-53851, Lappeenranta, Finland.

C. S. Haley, Centre for Research in Earth and Space Science (CRESS), York University, Toronto, ON M3J 1P3, Canada.

S. Hassinen, E. Kyrölä, A. Seppälä, J. Tamminen, S. Tukiainen, and P. T. Verronen, Finnish Meteorological Institute, P.O. Box 503, FI-00101, Helsinki, Finland. (simo.tukiainen@fmi.fi)

N. Lloyd, Institute of Space and Atmospheric Studies, University of Saskatchewan, Saskatoon, SK S7N 5E2, Canada.

PUBL. III

The variational Kalman filter and an efficient implementation using limited memory BFGS

H. Auvinen[1,*,†], J. M. Bardsley[2], H. Haario[1] and T. Kauranne[1]

[1]*Department of Mathematics and Physics, Lappeenranta University of Technology, Lappeenranta, Finland*
[2]*Department of Mathematical Sciences, University of Montana, Missoula, MT 59812, U.S.A.*

SUMMARY

In the field of state space estimation and data assimilation, the Kalman filter (KF) and the extended Kalman filter (EKF) are among the most reliable methods used. However, KF and EKF require the storage of, and operations with, matrices of size $n \times n$, where n is the size of the state space. Furthermore, both methods include inversion operations for $m \times m$ matrices, where m is the size of the observation space. Thus, KF methods become impractical as the dimension of the system increases. In this paper, we introduce a variational Kalman filter (VKF) method to provide a low storage, and computationally efficient, approximation of the KF and EKF methods. Furthermore, we introduce a variational Kalman smoother (VKS) method to approximate the fixed-lag Kalman smoother (FLKS) method. Instead of using the KF formulae, we solve the underlying maximum *a posteriori* optimization problem using the limited memory Broyden–Fletcher–Goldfarb–Shanno (LBFGS) method. Moreover, the LBFGS optimization method is used to obtain a low storage approximation of state estimate covariances and prediction error covariances. A detailed description of the VKF and VKS methods with LBFGS is given. The methodology is tested on linear and nonlinear test examples. The simulated results of the VKF method are presented and compared with KF and EKF, respectively. The convergence of BFGS/LBFGS methods is tested and demonstrated numerically. Copyright © 2009 John Wiley & Sons, Ltd.

*Correspondence to: H. Auvinen, Department of Mathematics and Physics, Lappeenranta University of Technology, Lappeenranta, Finland.
†E-mail: harri.auvinen@lut.fi

1. INTRODUCTION

Several variants of the Kalman filter (KF), (see [1]), and the extended Kalman filter (EKF) have been proposed to reduce their computational complexity for large-dimensional problems. The reduced rank KF or reduced-order EKF (see, e.g. [2–8]) project the dynamical state vector of the model onto a low-dimensional subspace. The success of the approach depends on a judicious choice of the reduction operator. Moreover, since the reduction operator is typically fixed in time, the dynamics of the system may not be correctly captured; see [9] for more details.

There exist various Ensemble Kalman Filter (EnKF) algorithms—first proposed in [10]—that are widely used in the field of data assimilation. The idea behind these methods is to form an ensemble of state vectors that represent the state estimate covariance. Each of the members of the ensemble is then propagated forward in time by the full nonlinear evolution model in order to approximate componentwise covariances of prediction error. EnKF can be used on large-scale data assimilation problems because it is highly parallelizable.

In [11], we have shown how high-dimensional KF and EKF may be carried out approximatively using the limited memory Broyden–Fletcher–Goldfarb–Shanno (LBFGS) optimization algorithm. The resulting methods were effective and exhibited low storage and computational cost characteristics. In this paper, we introduce an alternative approximation, the variational Kalman filter (VKF), for KF and EKF. Furthermore, we introduce a variational Kalman smoother (VKS) method to approximate the fixed-lag Kalman smoother (FLKS) method. In the variational approach, we solve an equivalent maximum *a posteriori* (MAP) optimization problem using LBFGS, which replaces the explicit computation and use of the Kalman gain matrix, in order to obtain state estimates and covariance approximations.

The idea of using the LBFGS method in variational data assimilation is not new (see e.g. [6–8, 12–15]). In many of these references, the LBFGS Hessian or inverse Hessian is used as a preconditioner, and even as an approximate error covariance matrix for the background term in variational data assimilation. However, in the VKF method presented here, the LBFGS method is further used for matrix inversion, in order to propagate effectively the state estimate covariance information forward in time.

The paper is organized as follows. We introduce the notations used in this paper in Section 2, with discussion of the KF. In Section 3, we present our methods for approximating the KF and FLKS. The convergence of LBFGS/BFGS methods is studied numerically in Section 4. To test and compare these methods, we present results from a number of numerical experiments in Section 5, and we end with discussion and conclusions in Sections 6 and 7, respectively.

2. THE KALMAN FILTER

Consider the following coupled system of discrete, linear, stochastic difference equations:

$$\mathbf{x}_k = \mathbf{M}_k \mathbf{x}_{k-1} + \boldsymbol{\varepsilon}_k^p \tag{1}$$

$$\mathbf{y}_k = \mathbf{K}_k \mathbf{x}_k + \boldsymbol{\varepsilon}_k^o \tag{2}$$

In the first equation, $\mathbf{x}_k$ denotes the $n \times 1$ state vector of the system at time k; $\mathbf{M}_k$ is the $n \times n$ linear evolution operator; and $\boldsymbol{\varepsilon}_k^p$ is an $n \times 1$ random vector representing the prediction error and is

assumed to characterize errors in the model and in the corresponding numerical approximations. In the second equation, $\mathbf{y}_k$ denotes the $m \times 1$ observed data vector; $\mathbf{K}_k$ is the $m \times n$ linear observation operator; and $\boldsymbol{\varepsilon}_k^o$ is an $m \times 1$ random vector representing the observation error. The error terms are assumed to be independent and normally distributed, with zero mean and with covariance matrixes $\mathbf{C}_{\boldsymbol{\varepsilon}_k^p}$ and $\mathbf{C}_{\boldsymbol{\varepsilon}_k^o}$, respectively.

The task is to estimate $\mathbf{x}_k$ and its error covariance $\mathbf{C}_k$ at time point k given $\mathbf{y}_k$, $\mathbf{K}_k$, $\boldsymbol{\varepsilon}_k^o$, $\mathbf{M}_k$, $\boldsymbol{\varepsilon}_k^p$ and estimates $\mathbf{x}_{k-1}^{\text{est}}$ and $\mathbf{C}_{k-1}^{\text{est}}$ of the state and covariance at time point $k-1$. The KF is the standard approach taken for such problems. It has the form:

The Kalman filter algorithm

Step 0: Select initial guess $\mathbf{x}_0^{\text{est}}$ and covariance $\mathbf{C}_0^{\text{est}}$, and set $k=1$.
Step 1: Compute the evolution model estimate and covariance:

 (i) Compute $\mathbf{x}_k^p = \mathbf{M}_k \mathbf{x}_{k-1}^{\text{est}}$.
 (ii) Compute $\mathbf{C}_k^p = \mathbf{M}_k \mathbf{C}_{k-1}^{\text{est}} \mathbf{M}_k^{\mathrm{T}} + \mathbf{C}_{\boldsymbol{\varepsilon}_k^p}$.

Step 2: Compute KF estimate and covariance:

  (i) Compute the Kalman gain $\mathbf{G}_k = \mathbf{C}_k^p \mathbf{K}_k^{\mathrm{T}} (\mathbf{K}_k \mathbf{C}_k^p \mathbf{K}_k^{\mathrm{T}} + \mathbf{C}_{\boldsymbol{\varepsilon}_k^o})^{-1}$.
 (ii) Compute the KF estimate $\mathbf{x}_k^{\text{est}} = \mathbf{x}_k^p + \mathbf{G}_k (\mathbf{y}_k - \mathbf{K}_k \mathbf{x}_k^p)$.
(iii) Compute the estimate covariance $\mathbf{C}_k^{\text{est}} = \mathbf{C}_k^p - \mathbf{G}_k \mathbf{K}_k \mathbf{C}_k^p$.

Step 3: Update $k := k+1$ and return to Step 1.

A nonlinear extension of KF, known as the EKF, is obtained when (1), (2) are replaced by

$$\mathbf{x}_k = \mathcal{M}(\mathbf{x}_{k-1}) + \boldsymbol{\varepsilon}_k^p \tag{3}$$

$$\mathbf{y}_k = \mathcal{K}(\mathbf{x}_k) + \boldsymbol{\varepsilon}_k^o \tag{4}$$

where $\mathcal{M}$ and $\mathcal{K}$ are possibly nonlinear functions. EKF is obtained by the following modification of the KF algorithm: in Step 1 (i) use the nonlinear model $\mathbf{x}_k^p = \mathcal{M}(\mathbf{x}_k^{\text{est}})$ to compute the prior, but employ the linearized approximations,

$$\mathbf{M}_k = \frac{\partial \mathcal{M}(\mathbf{x}_{k-1}^{\text{est}})}{\partial \mathbf{x}} \quad \text{and} \quad \mathbf{K}_k = \frac{\partial \mathcal{K}(\mathbf{x}_k^p)}{\partial \mathbf{x}} \tag{5}$$

for the covariance calculations, and otherwise employ the same formulae as above.

We note that $\mathbf{M}_k$ and $\mathbf{K}_k$ can be computed or estimated in a number of ways. For example, the numerical scheme that is used in the solution of either the evolution or the observation model defines a tangent linear code (see, e.g. [16]), which can be used to compute (5). A common, but also more computationally expensive, approach is to use finite differences to approximate (5).

The KF is expensive to implement due to the fact that it is necessary to store $n \times n$ matrices and invert $m \times m$ matrices at each step. Our task is to overcome these limitations. For this we will need a variational formulation of the KF that is set forth in the following section. We make the

(reasonable) assumptions that multiplication by the evolution and observation matrices $\mathbf{M}_k$ and $\mathbf{K}_k$ and by the covariance matrices $\mathbf{C}_{\varepsilon_k^p}$ and $\mathbf{C}_{\varepsilon_k^o}$ is efficient, both in terms of storage and CPU time.

3. THE VKF METHOD

In a general assimilation problem, where n is the size of state space and m is the size of the observation space, the standard formulation of the KF and the EKF requires the storage and multiplication of $n \times n$ matrices and the inversion of $m \times m$ matrices. Thus, the computational cost of KF/EKF increases rapidly as the size of the problem becomes large. Owing to this fact there are several interesting assimilation problems, where the standard formulation of KF or EKF is impractical to implement.

Bayes' Theorem can be used to formulate the KF as a sequential MAP iteration. To see this, we recall Bayes' formula

$$p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}) = \frac{p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})\, p_{\mathbf{x}}(\mathbf{x})}{p_{\mathbf{y}}(\mathbf{y})} \tag{6}$$

where $\mathbf{x}$ is the vector of unknowns, $\mathbf{y}$ the measurements, $p_{\mathbf{x}}$ denotes the prior density, and $p_{\mathbf{y}|\mathbf{x}}$ is the density of $\mathbf{y}$ given $\mathbf{x}$ the likelihood function. The MAP estimate is obtained by maximizing (6). Equivalently, one can minimize

$$\ell(\mathbf{x}|\mathbf{y}) := -\log p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}) - \log p_{\mathbf{x}}(\mathbf{x}) \tag{7}$$

For the linear model (2) at time k, the function ℓ assumes the form

$$\ell(\mathbf{x}|\mathbf{y}_k) = \tfrac{1}{2}(\mathbf{y}_k - \mathbf{K}_k\mathbf{x})^{\mathrm{T}} \mathbf{C}_{\varepsilon_k^o}^{-1}(\mathbf{y}_k - \mathbf{K}_k\mathbf{x}) + \tfrac{1}{2}(\mathbf{x} - \mathbf{x}_k^p)^{\mathrm{T}}(\mathbf{C}_k^p)^{-1}(\mathbf{x} - \mathbf{x}_k^p) \tag{8}$$

where $\mathbf{C}_{\varepsilon_k^o}$ and $\mathbf{C}_k^p$ are the covariance matrices of the measurement noise ε_k^o and of the prior $\mathbf{x}_k^p$, respectively. The KF estimate and its covariance $\mathbf{x}_k^{\mathrm{est}}$ and $\mathbf{C}_k^{\mathrm{est}}$ are precisely the minimizer and inverse Hessian of $\ell(\mathbf{x}|\mathbf{y}_k)$, respectively.

The advantage of the variational formulation of the KF is that it suggests the use of an optimization algorithm for computing estimates of $\mathbf{x}_k^{\mathrm{est}}$ and $\mathbf{C}_k^{\mathrm{est}}$. For large-scale problems, this can be very advantageous, [17].

In particular, we advocate using the limited memory BFGS algorithm (LBFGS) for the minimization problem (7). Given specific choices of initial guess, stopping criteria, and number of stored vectors, LBFGS will yield estimates of both $\mathbf{x}_k^{\mathrm{est}}$ and $\mathbf{C}_k^{\mathrm{est}}$. The storage requirement for the covariance approximation—which we denote $\mathbf{B}_k^{\#}$—is $2rn$, where r is the number of stored LBFGS vectors (typically on the order of 10), and multiplication by $\mathbf{B}_k$ is order n. See the Appendix for the details of the LBFGS algorithm for a quadratic cost function, as well as the limited memory formulas for $\mathbf{B}_k^{\#}$ and $(\mathbf{B}_k^{\#})^{-1}$.

However, it is $(\mathbf{C}_k^p)^{-1}$ that is needed in the optimization problem in the next VKF iteration (see (8)). Thus, we apply LBFGS a second time to an auxiliary optimization problem

$$\arg\min_{\mathbf{u}} \tfrac{1}{2}\langle \mathbf{A}\mathbf{u}, \mathbf{u}\rangle - \langle \mathbf{b}, \mathbf{u}\rangle \tag{9}$$

where $\mathbf{A} = \mathbf{M}_k \mathbf{B}_{k-1}^{\#} \mathbf{M}_k^{\mathrm{T}} + \mathbf{C}_{\varepsilon_k^p}$ and $\mathbf{b}$ is the zero vector. This gives an approximation $\mathbf{B}_k^*$ of $(\mathbf{C}_k^p)^{-1}$.

We summarize the VKF algorithm as follows:

The VKF algorithm

Step 0: Select initial guess $\mathbf{x}_0^\#$ and covariance $\mathbf{B}_0^\# = \mathbf{C}_0^{\text{est}}$, and set $k = 1$.

Step 1: Compute the evolution model estimate and covariance:

(i) Compute $\mathbf{x}_k^p = \mathbf{M}_k \mathbf{x}_{k-1}^\#$.

(ii) Define $\mathbf{C}_k^p = \mathbf{M}_k \mathbf{B}_{k-1}^\# \mathbf{M}_k^{\text{T}} + \mathbf{C}_{\boldsymbol{\varepsilon}_k^p}$.

(iii) Compute LBFGS approximation $\mathbf{B}_k^*$ of $(\mathbf{C}_k^p)^{-1}$.

Step 2: Compute VKF and covariance estimates:

(i) Minimize $\ell(\mathbf{x}|\mathbf{y}_k) = (\mathbf{y}_k - \mathbf{K}_k \mathbf{x})^{\text{T}} (\mathbf{C}_{\boldsymbol{\varepsilon}_k^o})^{-1} (\mathbf{y}_k - \mathbf{K}_k \mathbf{x}) + (\mathbf{x} - \mathbf{x}_k^p)^{\text{T}} \mathbf{B}_k^* (\mathbf{x} - \mathbf{x}_k^p)$ using LBFGS, and define $\mathbf{x}_k^\#$ and $\mathbf{B}_k^\#$ to be the LBFGS minimizer and inverse Hessian approximations;

Step 3: Update $k := k + 1$ and return to Step 1.

Note that in Step 1 (ii) and Step 2 (i) the optimizations are quadratic and therefore only quadratic LBFGS is needed, (see the Appendix for details). For practical applications, a judicious choice of the initial inverse Hessian is needed in order to obtain accurate results efficiently. In the numerical examples of this work, we have used $\mathbf{B}_0^{-1} = \beta \mathbf{I}$ with β chosen so that $\beta \mathbf{I}$ approximates the diagonal of the covariance matrix of interest. For more discussion on the choice of $\mathbf{B}_0^{-1}$ (the preconditioner) see [8, 18].

3.1. The nonlinear VKF method

As in the case of EKF, we need a linearization to propagate the covariance information from one observation time to the next. However, the direct linearization as in EKF is impractical for large dimensions. Rather, in the case of a non-linear evolution model we should use the adjoint operator, if available, in Step 1 (ii) of VKF. Furthermore, if the adjoint operator is coded in an implicit form, that is, in the software of the model, we get full benefit from a limited memory presentation of $\mathbf{C}_k^{\text{est}}$. This, indeed, is the situation in many operational codes for weather forecasting.

Supposing that the linearization $\mathbf{M}_k$ of $\mathcal{M}_k$ is available, then Step 2 (i) can written in the VKF algorithm as:

Step 2: Compute the VKF estimate and covariance:

(i) Minimize $\ell(\mathbf{x}|\mathbf{y}_k) = (\mathbf{y}_k - \mathcal{K}(\mathbf{x}))^{\text{T}} (\mathbf{C}_{\boldsymbol{\varepsilon}_k^o})^{-1} (\mathbf{y}_k - \mathcal{K}(\mathbf{x})) + (\mathbf{x} - \mathbf{x}_k^p)^{\text{T}} \mathbf{B}_k^* (\mathbf{x} - \mathbf{x}_k^p)$ using LBFGS, and define $\mathbf{x}_k^\#$ and $\mathbf{B}_k^\#$ to be the LBFGS minimizer and inverse Hessian approximations;

Especially, if the *tangent linear* $\mathbf{M}_k^{\text{tl}}$ and corresponding *adjoint code* $\mathbf{M}_k^*$ [16] are available for the evolution model $\mathcal{M}$, Step 1 (ii) can be written as:

Step 1: Compute the evolution model estimate and covariance:

(ii) Define $\mathbf{C}_k^p = \mathbf{M}_k^{\text{tl}} \mathbf{B}_{k-1}^\# \mathbf{M}_k^* + \mathbf{C}_{\boldsymbol{\varepsilon}_k^p}$;

This feature of the method is one of the major advantages compared with EKF, since the time-consuming linearization of the evolution model can be avoided. The traditional linearization

requires n evolution model function calls, but inside VKF the required number of tangent linear and adjoint code evaluations is around 15–60, which is far less than n in large-scale problems.

Similar possibilities exist for using the tangent linear code $\mathbf{K}_k^{\text{tl}}$ of the observation model $\mathscr{K}$ in Step 2 (i):

Step 2: Compute the evolution model estimate and covariance:

(i) Minimize $\ell(\mathbf{x}|\mathbf{y}_k) = (\mathbf{y}_k - \mathbf{K}_k^{\text{tl}}\mathbf{x})^{\text{T}}(\mathbf{C}_{\boldsymbol{\varepsilon}_k^o})^{-1}(\mathbf{y}_k - \mathbf{K}_k^{\text{tl}}\mathbf{x}) + (\mathbf{x} - \mathbf{x}_k^p)^{\text{T}}\mathbf{B}_k^*(\mathbf{x} - \mathbf{x}_k^p)$ using LBFGS, and define $\mathbf{x}_k^\#$ and $\mathbf{B}_k^\#$ to be the LBFGS minimizer and inverse Hessian approximations;

In EKF the computational advantage of using $\mathbf{K}_k^{\text{tl}}$ instead is lost, since the Kalman gain is typically a full matrix.

3.2. The VKS method

Next we introduce a VKS method, which can be used afterwards to smooth the results of the VKF. The idea is to simulate an FLKS method and take full benefit from the limited memory covariance approximation form of the VKF method. In general, such post-processing improves the quality of the VKF results.

The VKF method provides an estimate $\mathbf{x}_k^{\text{est}}$ and a corresponding limited memory approximation of the covariance matrix $\mathbf{C}_k^{\text{est}}$ after each time step k. In VKS, we use these results from the previous $[k_0, k_0+1, \ldots, k]$ time steps, where the parameter $k_0 = k - lag$ determines the length of the time interval. In case of linear evolution model $\mathbf{M}_k$, we couple the results together by using the following cost function:

$$J(\mathbf{x}_{k_0}) = \sum_{t=k_0}^{k} (\mathbf{M}_t\mathbf{x}_{k_0} - \mathbf{x}_t^{\text{est}})^{\text{T}}(\mathbf{C}_t^{\text{est}})^{-1}(\mathbf{M}_t\mathbf{x}_{k_0} - \mathbf{x}_t^{\text{est}}) \tag{10}$$

where $\mathbf{M}_t\mathbf{x}_{k_0}$ is a model trajectory from $\mathbf{x}_{k_0}$. The minimization of the cost function is done by using the 4d-Var method (see [16, 19]), using LBFGS.

In the nonlinear case, the evolution model $\mathscr{M}_t$ is used instead of $\mathbf{M}_t$ in the cost function formulation (10). Furthermore, the gradient of (10) can be computed efficiently by using the adjoint of the evolution model, but in principle, the linearization of $\mathscr{M}_t$ can be used again as well. As the smoothing process improves the accuracy of the estimate at time k_0, it is possible to outperform EKF in retrospective analysis.

During VKF iterations, the inverse Hessian limited memory BFGS formula is used to represent $\mathbf{C}_k^{\text{est}}$. In the VKS cost function (10) we instead need the inverse of $\mathbf{C}_t^{\text{est}}$. In practice this detail is handled by using the direct Hessian limited memory BFGS formula (see, e.g. [18] and the Appendix). The direct Hessian limited memory BFGS formula provides the $(\mathbf{C}_t^{\text{est}})^{-1}$ required.

4. MONITORING THE QUALITY OF THE BFGS APPROXIMATION

The quality of approximations produced by quasi-Newton methods of Hessian matrices, such as the covariance matrices in our case, has been studied at least since 1970 [20]. Normally this is carried out by monitoring the matrix norm of the difference between an approximation and a known Hessian matrix.

As discussed above, the KF is a statistical procedure, that repeatedly applies the Bayes rule to create the distribution of the state vector. We will therefore employ the chi-square test to monitor the goodness of the LBFGS updates as approximations of the covariances of known distributions. The chi-square distribution provides a scalar valued test for multinormality: if $\mathbf{x}$ is an n-dimensional Gaussian random vector with zero mean and covariance $\mathbf{C}$, then $\mathbf{x}^T \mathbf{C}^{-1} \mathbf{x} \sim \chi_n^2$. We monitor how well the approximative covariances fulfill this test, when a sample of vectors $\mathbf{x}$ has been generated using a 'true' covariance matrix.

We start with a covariance $\mathbf{C}^{\text{true}}$ and sample a set of vectors $\mathbf{x}_i$ from $N(0, \mathbf{C}^{\text{true}})$, $i = 1, \ldots, m$. Next, we compute the approximations $\mathbf{C}_{\text{iter}}^{-1}$ of the inverse of the covariance using the LBFGS optimization with an increasing number of iterations, and calculate $qtest(i) := \mathbf{x}_i^T \mathbf{C}_{\text{iter}}^{-1} \mathbf{x}_i$, for every i. These values are compared with the chi-square significance values at, for example, the $0.5, 0.75$, and 0.95 levels, to see which percentage of the vectors sampled from the true distribution lies inside those confidence regions of the respective approximative multinormal distributions. It is known that for a linear-quadratic problem the BFGS algorithm forms an exact inverse Hessian with full n updates; hence, the test values are expected to reach the true limits with n iterations. Accordingly, we perform tests with both the limited memory BFGS and full BFGS, but with the LBFGS method we stop at the convergence limit, when the stopping criterion has been chosen to be close to machine precision. The full BFGS runs are performed with the LBFGS code, by keeping all search directions in the memory. During LBFGS runs we keep all other search directions in memory except the first one.

In the test examples, we use $m = 1000$ samples and repeat the procedure 10 times to get the mean values of the results.

We consider two different cases: in the first example, the true covariance $\mathbf{C}_i^{\text{true}} \in \mathbb{R}^{m \times m}$ is defined with singular values proportional to $1/k$, where $k = 4, 5, \ldots, m+4$. In the second test, the covariance is adopted from the EKF process of the Lorenz95 case (for more details, see the next Section), with 100 variables.

In the above tests we allow the BFGS optimization method to continue iteration beyond the normal stopping point. If the typical stopping criteria are used, the LBFGS method will terminate after about 45 iterations in these cases, with 100 variables. We also perform a true limited memory chi-square test with LBFGS, where we stop the optimization process at the convergence limit.

The results for the first test case are given in Figure 1. With too few updates, the calculated values are below the chi-square test values, hence, the approximate covariance gives a too narrow distribution. With increasing updates, the test values are first exceeded, then we can see how the correct values are reached with full updates. The choice of the initial inverse matrix B_0 naturally has an impact on the process. To demonstrate this, we use slightly different B_0 values: $B_0 = 1.8\mathbf{I}$, with BFGS and $B_0 = 1.55\mathbf{I}$, with LBFGS.

Next we adopt the covariance $\mathbf{C}^{\text{est}}$ from the EKF process in the Lorenz'95 case, with 100 variables. Then we perform the chi-square test over the VKF Step 1 (ii) where we approximate $(\mathbf{C}^p)^{-1} = (\mathbf{M}\mathbf{C}^{\text{est}}\mathbf{M}^T + \mathbf{C}_\varepsilon)^{-1}$ by using BFGS and LBFGS methods. The corresponding results for the second test example, the Lorenz95 case, are plotted in Figure 2. The initial inverse Hessian $B_0 = 8.5\mathbf{I}$, with BFGS and $B_0 = 8.0\mathbf{I}$, with LBFGS.

We can see that BFGS method relative quickly finds the right confidence limits, with a suitable initial inverse Hessian. After the normal stopping point of the optimization process BFGS exceeds its designated confidence limits, but afterwards converges to the right values as the iteration count reaches the dimension of the problem. We note, that for practical applications the most important
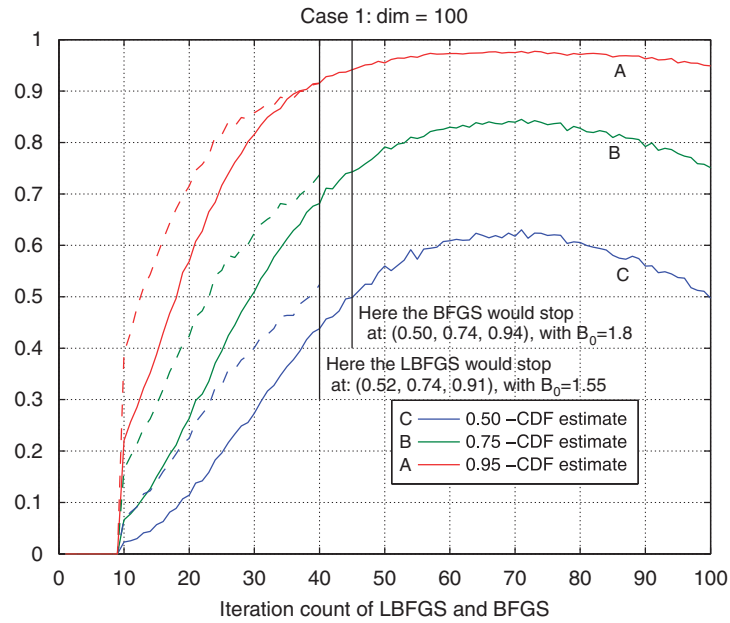
Figure 1. Plot of the chi-square test confidence limit estimates of the BFGS (-) and LBFGS (- -)
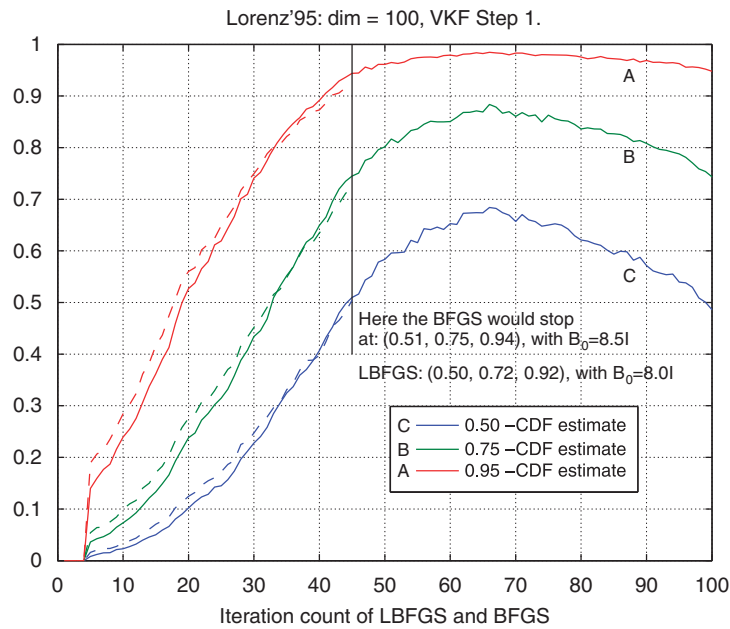as a function of the iteration count.

Figure 2. Plot of the chi-square test confidence limit estimates of the BFGS (-) and LBFGS (- -)
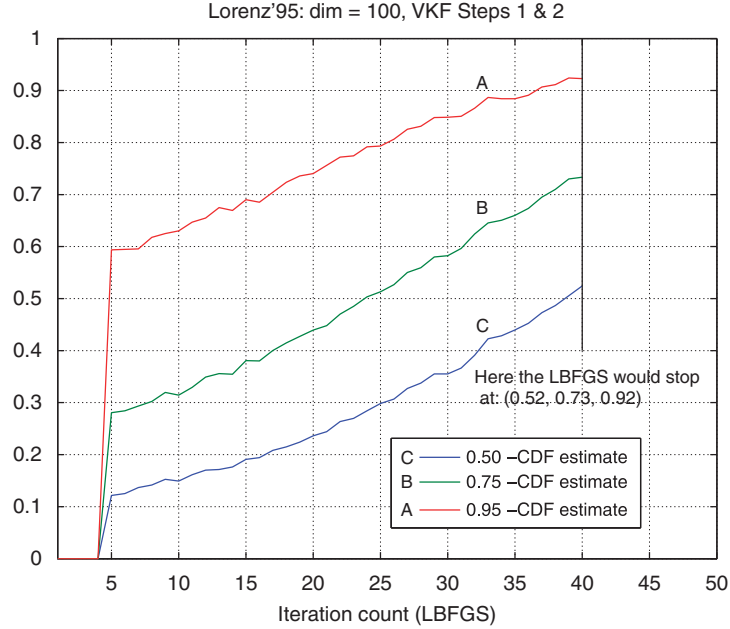as a function of the iteration count.

Figure 3. Plot of the chi-square test confidence limit estimates for LBFGS (-) over the VKF steps. The LBFGS estimates are plotted as a function of the iteration count.

values are located near the normal stopping point of the optimization process, where the method seems to provide reasonably accurate results.

Next, we study the accuracy of the covariance approximation over one full VKF time step, that is, both Step 1 and Step 2. For simplicity, we use the same number of iterations during both steps. Again, we take the covariance $\mathbf{C}^{\text{est}}$ and also $\mathbf{x}^p$ from the EKF process in the Lorenz'95 case, with 100 variables. First we perform VKF Step 1 (ii) where we approximate $(\mathbf{C}^p)^{-1} = (\mathbf{MC}^{\text{est}}\mathbf{M}^{\text{T}} + \mathbf{C}_\varepsilon)^{-1}$ by using LBFGS method. Then we use this approximation in VKF Step 2 (i) to compute the VKF covariance approximation with the LBFGS method using initial inverse Hessian $B_0 = 0.12\mathbf{I}$. In order to analyze the accuracy of the approximated covariance $\mathbf{C}_{\text{iter}}^{-1}$, we compute the corresponding operations with the EKF method, namely EKF Steps 1 (ii), Step 2 (i) and Step 2 (iii) to obtain $\mathbf{C}^{\text{true}}$. Finally, we perform similar chi-square tests for these covariances as earlier. The results are shown in Figure 3. They indicate that the LBFGS method provides reasonably accurate approximations after 40 iterations already. Please see the Appendix for analytical result about the convergence rate of LBFGS.

5. NUMERICAL EXPERIMENTS

In this section, we test the VKF on two examples.

5.1. An example with a large-scale linear evolution model

The first example is intended to test a large dimensional situation. We consider the following forced heat equation model

$$\frac{\partial x}{\partial t} = -\frac{\partial^2 x}{\partial u^2} - \frac{\partial^2 x}{\partial v^2} + \alpha \exp\left[-\frac{(u-2/9)^2 + (v-2/9)^2}{\sigma^2}\right] \tag{11}$$

where x is a function of u and v over the domain $\Omega = \{(u,v)|0 \leqslant u, v \leqslant 1\}$ and $\alpha \geqslant 0$. We generate synthetic data using (11) with $\alpha > 0$ and assume that the evolution model is given by (11) with $\alpha = 0$, which gives a model bias. The problem can be made arbitrarily large scale via a sufficiently fine spatial discretization. However, the well-behaved nature of solutions of (11) calls for further experiments with a different test case.

We discretize the model (11) using a uniform $N \times N$ computational grid and the standard finite difference schemes of both the time and spatial derivatives. This gives the time-stepping equation $\mathbf{x}_{k+1} = \mathbf{M}\mathbf{x}_k + \mathbf{f}$, where $\mathbf{M} = \mathbf{I} - \Delta t \mathbf{L}$. Here $\mathbf{L}$ is given by the standard finite difference discretization of the two-dimensional Laplacian operator with homogeneous Dirichlet boundary conditions, Δt is chosen to guarantee stability, and $\mathbf{f}$ is the constant vector determined by the evaluation of the forcing term in (11) at each of the points of the computational grid. We define $\mathbf{K}_k = \mathbf{K}$ for all k in (2), where $\mathbf{K}$ is the full weighting matrix, which has the following grid representation

$$\frac{1}{16}\begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}$$

Such an observation matrix could model, for example, an array of square heat sensors on the bottom of a metal plate that have dimension $2/N \times 2/N$ with the edges aligned with the grid lines and equally spaced at $n^2/64$ locations.

We first generate synthetic data using the stochastic equations

$$\mathbf{x}_{k+1} = \mathbf{M}\mathbf{x}_k + \mathbf{f} + N(\mathbf{0}, (0.5\sigma_{\mathrm{ev}})^2\mathbf{I}) \tag{12}$$

$$\mathbf{y}_{k+1} = \mathbf{K}\mathbf{x}_{k+1} + N(\mathbf{0}, (0.8\sigma_{\mathrm{obs}})^2\mathbf{I}) \tag{13}$$

with $\alpha = \frac{3}{4}$ in (11) and where σ_{ev}^2 and σ_{obs}^2 are chosen so that the signal to noise ratios, defined by $\|\mathbf{x}_0\|^2 / n^2 \sigma_{\mathrm{ev}}^2$ and $\|\mathbf{K}\mathbf{x}_0\|^2 / n^2 \sigma_{\mathrm{obs}}^2$, respectively, are both 50. The initial condition used for the data generation is

$$[\mathbf{x}_0]_{ij} = \exp[-((u_i - 1/2)^2 + (v_j - 1/2)^2)]$$

where (u_i, v_j) is the ijth grid point.

For the implementation of KF, we used the biased models

$$\mathbf{x}_{k+1} = \mathbf{M}\mathbf{x}_k + N(\mathbf{0}, \sigma_{\mathrm{ev}}^2\mathbf{I})$$

$$\mathbf{y}_{k+1} = \mathbf{K}\mathbf{x}_{k+1} + N(\mathbf{0}, \sigma_{\mathrm{obs}}^2\mathbf{I})$$

with initial conditions $\mathbf{x}_0 = \mathbf{0}$ and $\mathbf{C}_0^{\mathrm{est}} = 0.001\mathbf{I}$ in Step 0 of the filter. We compare the results obtained with VKF and KF, where $n = 2^j$ with j taken to be the largest positive integer so that
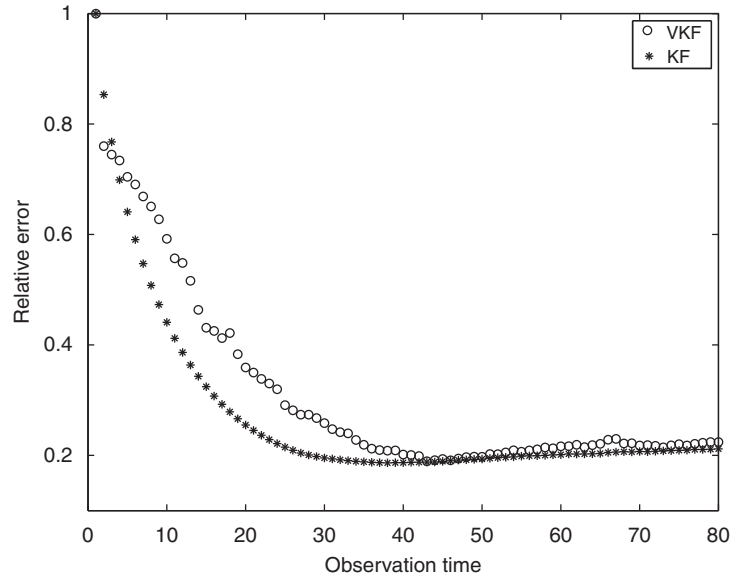
Figure 4. Relative error curves for KF ($*$) and VKF (o). The horizontal axis
represents the time of the observations.

memory issues do not arise in the MATLAB implementation for the standard KF. For the computer
on which the simulations were done (a laptop with 2G RAM memory and a 2.13 GHz processor)
the largest such j was 5, making $N = 32$. We note that in our implementation of the LBFGS
method, we have chosen to take only 10 LBFGS iterations with 9 saved vectors.

In our first test, the goal is only to show that the results obtained with VKF are comparable to
those obtained with KF. For this purpose, we present a plot in Figure 4 of the relative error vector,
which has kth component

$$[\textbf{relative\_error}]_k := \frac{\|\mathbf{x}_k^{\text{est}} - \mathbf{x}_k\|}{\|\mathbf{x}_k\|}$$

for both VKF and the standard KF. We see that results obtained using the two approaches yield
quite similar relative error curves. Inside the VKF method, the initial inverse Hessian parameters
during approximation of $\mathbf{C}_k^{\text{est}}$ and $(\mathbf{C}_k^p)^{-1}$ were $\mathbf{B}_0^{-1} = \mathbf{I}$ and $\mathbf{B}_0^{-1} = 4000\mathbf{I}$, respectively for all k.
At the beginning of the filtering period KF provides more accurate results, but later the difference
decreases. Both curves eventually begin to increase once the forcing term, which is not used in the
state space model in KF, has a prominent effect on the data; in early iterations, it is overwhelmed
by the diffused initial temperature. We also mention that in the large number of test runs we did
using this model, our implementation of the VKF was on average about 10 times faster than the
standard KF.

For more comparisons, we perform similar tests using different values for σ_{ev}^2 and σ_{obs}^2, namely,
so that the signal to noise ratios mentioned above are both 10. The same initial inverse Hessian
parameters were used inside the VKF method as in the previous test. The relative error curves are
exhibited in Figure 5. In this case, VKF provides better results at the beginning of the filtering
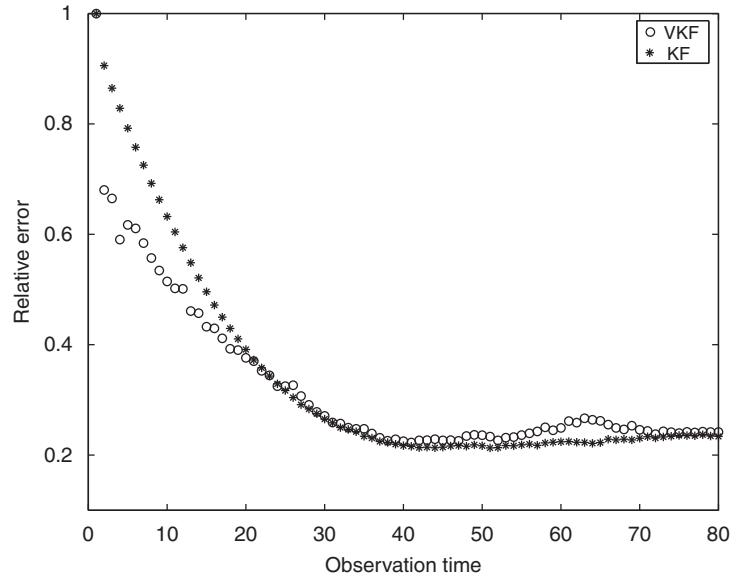
Figure 5. Relative error curves for KF (∗) and VKF (o). The horizontal
axis represents the observation time.

period. This might be due to a regularization effect, implicitly implemented via the use of a truncated LBFGS algorithm.

Lastly, we define σ^2_{ev} and σ^2_{obs} to be as in the original experiment, but take $\alpha = 2$. This has the effect that the state space model used within VKF and KF is less accurate. In this case the initial inverse Hessian parameters were $\mathbf{B}_0^{-1} = \mathbf{I}$ for $\mathbf{C}_k^{est}$ and $\mathbf{B}_0^{-1} = 2000\mathbf{I}$ for $(\mathbf{C}_k^p)^{-1}$. We now obtain the solution curves appearing in Figure 6. It seems that as the underlying evolution becomes less accurate, while the noise level remains moderately low, KF provides better results.

Satisfactory results can also be obtained for much larger-scale problems. To show this, we take $j = 8$ which gives $N = 256$, $n = 65\,536$. Hence, the number of unknowns in this problem is then $65\,536$. Otherwise, we fix the parameter values to be the same as above. However, the stability condition of the time-stepping scheme requires a much smaller step now. The respective error plot is given in Figure 7. We cannot include any error curve for the standard KF because memory issues prevent it on our computer for both $N = 128\,(n = 16\,384)$ or $N = 256\,(n = 65\,536)$. However, in this case we compare results with the LBFGS-KF method [11]. We note that the results here for VKF are rather similar to those presented for LBFGS-KF in [11].

5.2. An example with a small-scale, nonlinear evolution model

Our second example produces chaotic, unpredictable behavior. We consider the non-linear Lorenz'95 model introduced and analyzed in [21, 22], given by

$$\frac{\partial x^i}{\partial t} = (x^{i+1} - x^{i-2})x^{i-1} - x^i + 8, \quad i = 1, 2, \ldots, 40 \tag{14}$$
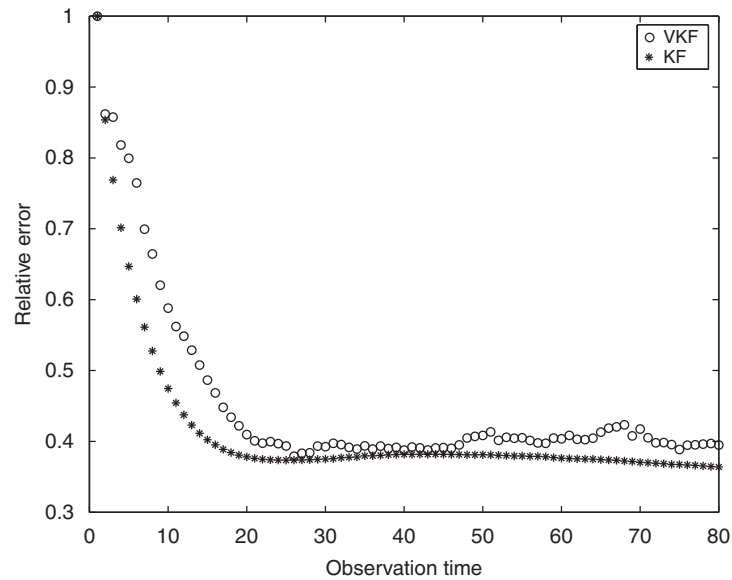
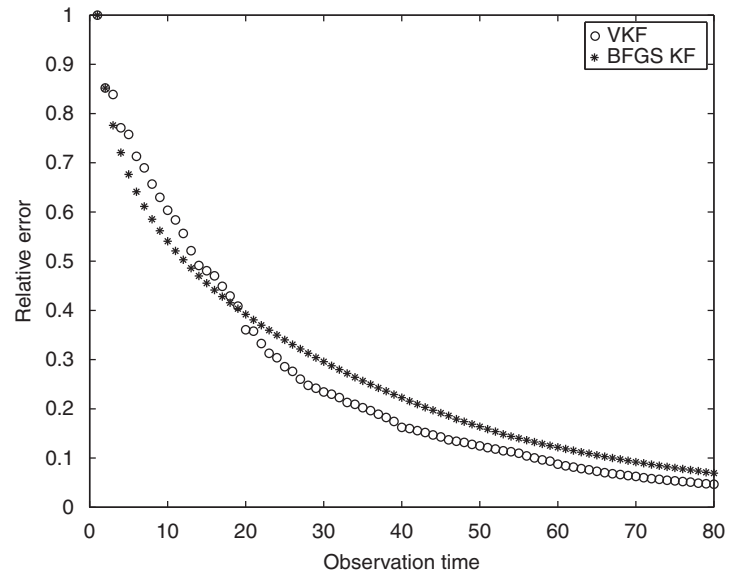Figure 6. Relative error curves for KF (∗) and VKF (o). The horizontal axis represents the observation time.

Figure 7. Relative error curves for LBFGS-KF (∗) and VKF (o). The horizontal axis represents the observation time.

with periodic state space variables, that is, $x^{-1} = x^{n-1}$, $x^0 = x^n$ and $x^{n+1} = x^1$. In the present tests we use the dimension $n = 40$. The model shares many characteristics with realistic atmospheric models (cf. [22]), and is often used as a test case for various weather forecasting schemes.

Next, we apply EKF, VKF, and VKS to the problem of estimating the state variables from data generated using the nonlinear, chaotic evolution model (14). The data were generated by integrating the model using a fourth-order Runge–Kutta (RK4) method with time step $\Delta t = 0.025$. The discussion in [22] suggests that when using (14) as a test example for weather forecasting algorithms, the characteristic time scale is such that the above Δt corresponds to 3 h. The 'truth' was generated by taking 42 920 time steps of the RK4 method, that is, 5365 days. The initial state for the data generation was $x^{20} = 8 + 0.008$ and $x^i = 8$ for all $i \neq 20$.

The observed data are then computed using these true data. In particular, after a 365 day long initial period, the true data is observed at every other time step and at the last 3 grid points in each set of 5; that is, the observation matrix is $m \times n$, with nonzero entries

$$[\mathbf{K}]_{rs} = \begin{cases} 1, & (r,s) \in \{(3j+i, 5j+i+2) \mid i=1,2,3, \, j=0,1,\ldots,7\} \\ 0 & \text{otherwise} \end{cases}$$

The observation error is simulated using Gaussian noise $N(\mathbf{0}, (0.15\,\sigma_{\text{clim}})^2\mathbf{I})$ where σ_{clim} is a standard deviation of the model state used in climatological simulations, $\sigma_{\text{clim}} := 3.6414723$. The data generation codes were written in MATLAB and were transcribed by us from the `scilab` codes written by the author of [23].

For the application of EKF and VKF, we employ the coupled system

$$\mathbf{x}_{k+1} = \mathcal{M}(\mathbf{x}_k) + N(\mathbf{0}, (0.05\,\sigma_{\text{clim}})^2\mathbf{I}) \tag{15}$$

$$\mathbf{y}_{k+1} = \mathbf{K}\mathbf{x}_{k+1} + N(\mathbf{0}, (0.15\,\sigma_{\text{clim}})^2\mathbf{I}) \tag{16}$$

where $\mathcal{M}(\mathbf{x}_k)$ is obtained by taking two steps of the RK4 method applied to (14) from $\mathbf{x}_k$ with time step 0.025. We note that this coincides with the data generation scheme, if the noise term is removed and the above initial condition is used. Owing to the fact that $\mathcal{M}$ is a nonlinear function, EKF must be used (see Equations (3) and (4)). As $\mathcal{K} := \mathbf{K}$ in (4) is linear, $\mathbf{K}_k = \mathbf{K}$ for all k in (5). However, a linearization of the nonlinear evolution function $\mathcal{M}$ is required. The computation of $\mathbf{M}_k$ in (5) is performed by a routine in one of the `scilab` codes mentioned above, adopted for our use in MATLAB.

The initial condition is defined by $[\mathbf{x}_{t0}]_i = [\mathbf{x}_{t0}^{\text{true}}]_i + N(0, (0.3\,\sigma_{\text{clim}})^2)$ for all i, and the initial covariance was taken to be $\mathbf{C}_0^{\text{est}} = (0.13\,\sigma_{\text{clim}})^2\mathbf{I}$. In our implementation of the LBFGS method within VKF, we computed 15 iterations with 14 saved vectors and the initial inverse Hessian parameters were $\mathbf{B}_0^{-1} = 0.15\mathbf{I}$ for $\mathbf{C}_k^{\text{est}}$ and $\mathbf{B}_0^{-1} = 10\mathbf{I}$ for $(\mathbf{C}_k^p)^{-1}$.

To present the accuracy of the state estimates $\mathbf{x}_k^{\text{est}}$ obtained by EKF, VKF, and VKS we plot the vector with components

$$[\mathbf{rms}]_k = \sqrt{\tfrac{1}{40} \|\mathbf{x}_k^{\text{est}} - \mathbf{x}_k^{\text{true}}\|^2} \tag{17}$$

in Figure 8. We observe that all the three methods yield comparable results.

In order to compare the forecasting abilities of the two approaches, we compute the following forecast statistics at every 8th observation. Take $j \in \mathscr{I} := \{8i \mid i = 1, 2, \ldots, 100\}$ and define

$$[\mathbf{forcast\_error}_j]_i = \tfrac{1}{40} \|\mathcal{M}_{4i}(\mathbf{x}_j^{\text{est}}) - \mathbf{x}_{j+4i}^{\text{true}}\|^2, \quad i = 1, \ldots, 20 \tag{18}$$
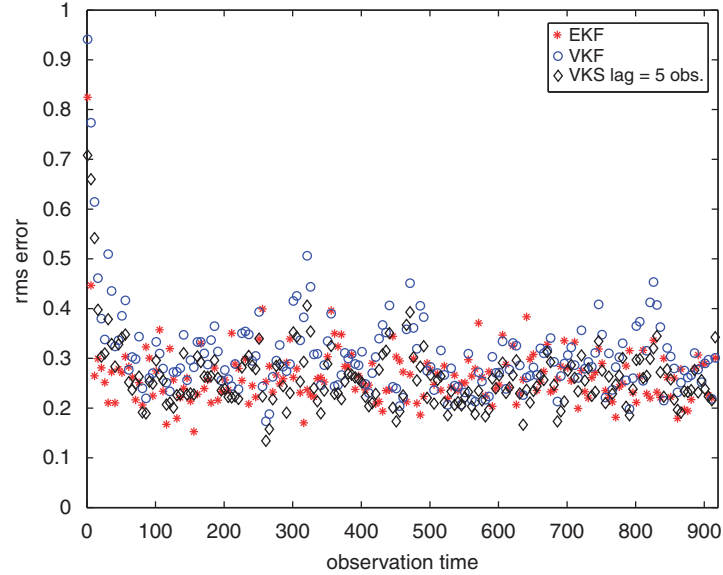
Figure 8. Plot of residual mean-square error for VKF (o), EKF (∗) and VKS (◇) applied to (14).

where $\mathscr{M}_n$ denotes a forward integration of the model by n time steps with the RK4 method. Thus, this vector gives a measure of forecast accuracy given by the respective filter estimate up to 80 time steps, or 10 days out. We may define the forecast skill vector

$$[\mathbf{forecast\_skill}]_i = \frac{1}{\sigma_{\mathrm{clim}}} \sqrt{\frac{1}{100} \sum_{j \in \mathscr{I}} [\mathbf{forecast\_error}_j]_i}, \quad i = 1, \ldots, 20 \tag{19}$$

plotted in Figure 9. The results show that the forecasting skills of EKF and VKF are quite similar. This suggests that the quality of the VKF estimates is as high as those obtained using EKF. The forecast of VKS method is computed lag = 5 observation times afterwards and therefore it actually is not a forecast, but it demonstrates the improvement of the accuracy of retrospective analysis. Figure 9 also illustrates the fact that the Lorenz95 model (14) is truly chaotic.

In our test cases a linear or linearized model matrix $\mathbf{M}_k$ has been available. This may not true in many important examples. But in numerical weather forecasting, for example, a tangent linear code often is available, providing an efficient way to compute the matrix vector product $\mathbf{M}_k\mathbf{x}$.

6. DISCUSSION

Fisher and Courtier [13] adopt several low-rank corrections derived from a 3D-Var minimization with LBFGS to approximate background and forecast error covariances, and also use them for preconditioning the minimizations. They observe that such methods perform reasonably well, but that in a 2D weather model, even 52 leading singular vectors fail to capture most of the difference between the variances of background and analysis error. Veersé *et al.* [8] adopt diagonal
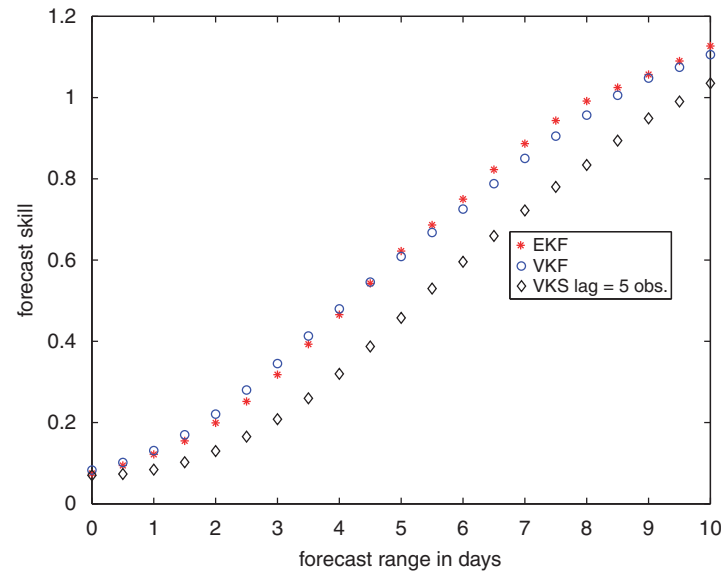
Figure 9. Plot of forecast skill vector for VKF (o) and EKF (∗) applied to (14). This plot also contains a retrospective forecast skill of the VKS (◇) method.

preconditioning based on LBFGS, too, and also apply it to the inverse Hessian approximation. They do not involve the Kalman formulas that would propagate the covariance in time.

VKF is related in many ways to various reduced rank KFs, see [2], [5], or reduced-order KFs and proper orthogonal decomposition (POD) filters, see [7]. Like these methods, VKF uses a low-dimensional subspace to approximate the forecast and analysis error covariance matrices.

Dee [2] and Fisher [5] select a suitable subspace for the low-rank correction beforehand and keep it over the assimilation window. Tian *et al.* [7] use a POD of an ensemble for spanning the four-dimensional search basis of an ensemble method and save considerably in the dimension of the control space. They are able to carry out a minimization over a set of entire 4D model trajectories, but they keep the same ensemble for the entire assimilation window, unlike VKF.

A popular alternative to variational data assimilation is the family of many different EnKF algorithms, first proposed in [10]. EnKFs avoid the need to use tangent linear and adjoint codes, and rely instead on generating an ensemble of state vectors that are then propagated with the full nonlinear evolution model forward in time.

Ensemble methods are relatively simple to implement, although many advanced versions use sophisticated manipulation to glean information about error covariance from the ensemble. Gejadze *et al.* [6] have implemented an ensemble method with a BFGS search direction and gradient basis. Their formulation of the problem is control theoretic, just like in 4D-Var, which renders the control problem to an initial state control, unlike the final state control formulation adopted by VKF and EKF alike. Their method is capable of handling strong nonlinearities in the model and produce very good approximate analysis error covariance matrices. VKF is even more similar to the Maximum Likelihood Ensemble Filter MLEF of Zupanski [24] in its principle of maximizing the likelihood in Step 2 (i) of VKF. MLEF, like VKF, follows the mode of the set of search directions, rather than

the mean. However, unlike VKF, it uses a fixed set of search directions throughout the assimilation window.

EnKFs like many Reduced Rank KFs tend to suffer from covariance leaks. This means that if the size of the sample, or the dimension of the reduced basis, is small compared with the dimension of the state vector, progressively less and less of true model error covariance stays in the subspace spanned by the ensemble, or the low-rank subspace, see [9]. The reasons to this phenomenon are not fully understood but it seems plausible that non-linear system dynamics simply do not leave any low-dimensional sub-space invariant, even if a basis of that subspace is transformed with the nonlinear model. VKF, on the other hand, compensates for this defect by discontinuously updating its subspace by picking up new 3D-Var and 4D-Var minimization directions every time there are new observations to process, or at any pre-defined time intervals.

The computational complexity of VKF closely parallels that of 4D-Var, and even that of EnKFs. In its Step 1 (i) VKF simply computes a short forecast with the full nonlinear model. In Step 1 (ii) an artificial 4D-Var minimization is carried out with fixed tangent linear and adjoint models. This is otherwise similar to the inner minimization loop of an incremental formulation of 4D-Var, see [25], except that the minimization is formulated as a final state control problem. It can therefore be expected to converge in some 50 iterations in an operational weather model. Since the approximate error covariance matrix is assembled using the LBFGS update formula in the course of the minimization, there is no extra cost compared with incremental 4D-Var, apart from conducting the minimization with the LBFGS method, as opposed to the method of Conjugate Gradients, which requires that a set of search directions and gradients is stored and manipulated.

In Step 2, the essence of VKF involves just a single 3D-Var minimization in Step 2 (i) again using LBFGS. We can therefore conclude that the computational complexity of VKF is similar to the classical 4D-Var method, see [16], where there is only a single minimization loop that uses LBFGS minimization, instead of the nested loop structure of incremental 4D-Var.

VKF also mimics the algorithmic structure of 4D-Var rather faithfully. It involves a 4D-Var-type sequence of steps, including computing the tangent linear and adjoint models during every forecast step. Linearization does not need to be repeated, since the time-dependent minimization is carried out with the linearized model by definition in Step 1 (ii). Repeated linearizations may be needed in Step 2 (i), but this will involve only the observation operator, as this step mimics 3D-Var and the forecast model is not involved. If the tangent linear and adjoint models are available because of a previously adopted 4D-Var, VKF will be relatively straightforward to implement.

Parallelization is an issue not discussed in the current paper, but please see [26] for some analysis. The formulation of VKF used here is a serial algorithm because of the sequential minimizations involved, although the forecast model can obviously be run in parallel, just as in current operational practice. In this respect, too, VKF is a close relative to current implementations of 4D-Var. However, it is possible to replace Step 1 (ii) by propagating 3D-Var minimization directions only forward in time, as described in [26]. Another possibility is to adopt an Ensemble version of VKF.

Some recent Ensemble Kalman methods have adopted a hybrid approach, where an EnKF is used to produce a dynamic error covariance matrix, while the state is transported with a 4D-Var method to retain its smoothness [27]. An earlier hybrid method combines a 3D-Var minimization with an EnKF [28]. In hybrid methods, the error covariance matrix is a weighted combination of an EnKF error covariance matrix and a static background error covariance matrix. In such a case, the ensemble can be computed in parallel, but the variational assimilation remains sequential.

VKF can be seen as a hybrid method, too, but a deterministic one. It possesses characteristics very similar to hybrid EnKF methods, and the parallel version discussed in [26] behaves computationally in a manner very close to the hybrid method of [27].

Hybrid methods involving 4D-Var will need a tangent linear and adjoint model, just like VKF. As to operational implementation, we feel that VKF is well worth implementing if the NWP model used has already been provided with a tangent linear and an adjoint code. In this case, the modifications needed to the analysis suite are mostly just some matrix and vector algebra. If 4D-Var has not been implemented, and the tangent linear and adjoint codes have not been produced, EnKFs will definitely be easier to implement and may better meet the needs of such NWP centers.

7. CONCLUSIONS

The standard implementations of KF and EKF become exceedingly time and memory consuming as the dimension of the underlying state space increases. Several variants of KF and EKF have been proposed to reduce the dimension of the system, thus making implementation in high dimensions possible. The reduced rank KF or reduced order EKF (see, e.g. [2–8, 14, 15]) project the dynamical state vector of the model onto a lower-dimensional subspace. The success of this approach depends on a judicious choice of the reduction operator. Moreover, since the reduction operator is typically fixed in time, they can suffer from 'covariance leaks' [9]. A typical cause to this is that a nonlinear system does not generally leave any fixed linear subspace invariant.

In this paper, we propose the use of the limited memory BFGS (LBFGS) minimization method in order to circumvent the computational complexity and memory issues of standard KF and EKF. In particular, we replace the $n \times n$, where n is the dimension of the state space, covariance matrices within KF and EKF with low storage approximations obtained using LBFGS. The large-scale matrix inversions required in KF and EKF implementations are also approximated using LBFGS. We call the resulting method the VKF. In order to test these methods, we consider two test cases: a large-scale linear and a small-scale nonlinear one. The VKF is applied in the large-scale linear case and is shown to be effective. In fact, our method exceeds the speed of standard KF by an order of magnitude, and yields comparable results when both methods can be applied. Furthermore, it can be used on much larger-scale problems. In the nonlinear, small scale case, VKF and VKS are implemented and are also shown to give results that are comparable to those obtained using standard EKF. We believe that these results suggest that our approach deserves further consideration for large-scale linear and non-linear data assimilation problems. We note that in truly high-dimensional non-linear cases we need a tangent linear and corresponding adjoint code of the evolution model in order to get full benefits of the VKF method. In many important application fields, for example, in numerical weather forecasting, such codes are already available both for the evolution model and for the observation model.

APPENDIX A

For completeness, we present the LBFGS method for a quadratic minimization and the limited memory formulations for the Hessian and inverse Hessian matrices. The LBFGS Minimization

Algorithm for $q(\mathbf{u}) = \frac{1}{2}\langle \mathbf{Au}, \mathbf{u}\rangle - \langle \mathbf{b}, \mathbf{u}\rangle$ reads as

$v := 0$;
$\mathbf{u}_0 :=$ initial guess;
$\mathbf{B}_0^{-1} :=$ initial inverse Hessian approximation;
begin quasi-Newton iterations
  $\mathbf{g}_v := \nabla q(\mathbf{u}_v) = \mathbf{Au}_v - \mathbf{b}$;
  $\mathbf{B}_v^{-1} :=$ LBFGS approximation to $\mathbf{A}^{-1}$;
  $\mathbf{v}_v = \mathbf{B}_v^{-1}\mathbf{g}_v$;
  $\tau_v = \langle \mathbf{g}_v, \mathbf{v}_v\rangle / \langle \mathbf{v}_v, \mathbf{Av}_v\rangle$;
  $\mathbf{u}_{v+1} := \mathbf{u}_v - \tau_v\mathbf{v}_v$;
end quasi-Newton iterations

A.1. The limited memory approximation for $\mathbf{A}^{-1}$

The BFGS matrix $\mathbf{B}_v^{-1}$ is computed using recursion

$$\mathbf{B}_{v+1}^{-1} = \mathbf{V}_v^{\mathrm{T}}\mathbf{B}_v^{-1}\mathbf{V}_v + \rho_v\mathbf{s}_v\mathbf{s}_v^{\mathrm{T}}$$

where

$$\mathbf{s}_v := \mathbf{u}_{v+1} - \mathbf{u}_v$$

$$\mathbf{d}_v := \nabla q(\mathbf{u}_{v+1}) - \nabla q(\mathbf{u}_v)$$

$$\rho_v := 1/\mathbf{d}_v^{\mathrm{T}}\mathbf{s}_v$$

$$\mathbf{V}_v := \mathbf{I} - \rho_v\mathbf{d}_v\mathbf{s}_v^{\mathrm{T}}$$

However, for large-scale problems the storage of the full matrix $\mathbf{B}_v^{-1}$ is infeasible, which motivates the limited storage version of the algorithm. At iteration v, suppose that the j vector pairs $\{\mathbf{s}_i, \mathbf{d}_i\}_{i=v-j}^{v-1}$ are stored. Then the LBFGS approximation of the inverse Hessian is given by

$$\begin{aligned}
\mathbf{B}_v^{-1} = {}& (\mathbf{V}_{v-1}^{\mathrm{T}}\ldots\mathbf{V}_{v-j}^{\mathrm{T}})\mathbf{B}_0^{-1}(\mathbf{V}_{v-j}\ldots\mathbf{V}_{v-1}) \\
& + \rho_{v-j}(\mathbf{V}_{v-1}^{\mathrm{T}}\ldots\mathbf{V}_{v-j+1}^{\mathrm{T}})\mathbf{s}_{v-j}\mathbf{s}_{v-j}^{\mathrm{T}}(\mathbf{V}_{v-j+1}\ldots\mathbf{V}_{v-1}) \\
& + \rho_{v-j+1}(\mathbf{V}_{v-1}^{\mathrm{T}}\ldots\mathbf{V}_{v-j+2}^{\mathrm{T}})\mathbf{s}_{v-j+1}\mathbf{s}_{v-j+1}^{\mathrm{T}}(\mathbf{V}_{v-j+2}\ldots\mathbf{V}_{v-1}) \\
& + \vdots \\
& + \rho_{v-1}\mathbf{s}_{v-1}\mathbf{s}_{v-1}^{\mathrm{T}}
\end{aligned} \tag{A1}$$

Assuming exact arithmetic and that $j = n$, we have that $\mathbf{u}_v$ converges to the unique minimizer of q in at most n iterations, and if n iterations are performed $\mathbf{B}_{n+1}^{-1} = \mathbf{A}^{-1}$ [18]. In the implementation in this paper, however, $j \ll n$ and LBFGS iterations are stopped once a prespecified maximum number of iterations or gradient norm stopping tolerance is reached.

It is proven in [29] that for quadratic minimization problems and exact line searches, LBFGS is equivalent to preconditioned conjugate gradient with fixed preconditioner $\mathbf{B}_0$. Thus, its convergence rate is given by [18]

$$\|\mathbf{u}_k - \mathbf{u}^*\|_{\tilde{\mathbf{A}}}^2 \leqslant \left(\frac{\lambda_{N-k} - \lambda_1}{\lambda_{N-k} + \lambda_1} \right)^2 \|\mathbf{u}_0 - \mathbf{u}^*\|_{\tilde{\mathbf{A}}}^2$$

where $\mathbf{u}^* = \mathbf{A}^{-1}\mathbf{b}$, $\tilde{\mathbf{A}} = \mathbf{B}_0^{1/2}\mathbf{A}\mathbf{B}_0^{1/2}$ is $N \times N$ with eigenvalues $\lambda_1 \leqslant \lambda_2 \leqslant \cdots \leqslant \lambda_N$, and $\|\mathbf{v}\|_{\tilde{\mathbf{A}}} = \mathbf{v}^T\tilde{\mathbf{A}}\mathbf{v}$.

A.2. A low storage approximation of $\mathbf{A}$

The required formulas are given in [30], and take the following form. Let

$$\mathbf{S}_v = [\mathbf{s}_{v-j}, \ldots, \mathbf{s}_{v-1}], \quad \mathbf{D}_v = [\mathbf{d}_{v-j}, \ldots, \mathbf{d}_{v-1}]$$

then

$$\mathbf{B}_v = \xi_v\mathbf{I} - [\xi_v\mathbf{S}_v \ \ \mathbf{D}_v] \begin{bmatrix} \xi_v\mathbf{S}_v^T\mathbf{S}_v & \mathbf{L}_v \\ \mathbf{L}_v^T & -\mathbf{D}_v \end{bmatrix}^{-1} \begin{bmatrix} \xi_v\mathbf{S}_v^T \\ \mathbf{D}^T \end{bmatrix} \tag{A2}$$

where $\mathbf{L}_v$ and $\mathbf{D}_v$ are the $j \times j$ matrices

$$(\mathbf{L}_v)_{i,j} = \begin{cases} \mathbf{s}_{v-j-1+i}^T\mathbf{d}_{v-j-1+j} & \text{if } i > j \\ 0 & \text{otherwise} \end{cases}$$

and

$$\mathbf{D}_v = \mathrm{diag}(\mathbf{s}_{v-j}^T\mathbf{d}_{v-j}, \ldots, \mathbf{s}_{v-1}^T\mathbf{d}_{v-1})$$

We note that when $\xi_v = 1$ for all v in (A2), we have an exact equality between $\mathbf{B}_v$ in (A2) and (A1). However, we have found that a more accurate Hessian approximation is obtained if, following [18], we use the scaling $\xi_v = \mathbf{d}_{v-1}^T\mathbf{d}_{v-1}/\mathbf{s}_{v-1}^T\mathbf{d}_{v-1}$ instead.

We note that the middle matrix in (A2) has size $2j \times 2j$, which is of reasonable size provided j is not too large, and its inversion can be carried out efficiently using a Cholesky factorization that exploits the structure of the matrix (for details see [30]).

REFERENCES

1. Kalman RE. A new approach to linear filtering and prediction problems. *Transactions of the ASME—Journal of Basic Engineering* 1960; **82**(Series D):35–45.
2. Dee DP. Simplification of the Kalman filter for meteorological data assimilation. *Quarterly Journal of the Royal Meteorological Society* 1990; **117**:365–384.
3. Cane MA, Miller RN, Tang B, Hackert EC, Busalacchi AJ. Mapping tropical Pacific sea level: data assimilation via reduced state Kalman filter. *Journal of Geophysical Research* 1996; **101**:599–617.
4. Voutilainen A, Pyhälahti T, Kallio K, Pulliainen J, Haario H, Kaipio J. A filtering approach for estimating lake water quality from remote sensing data. *International Journal of Applied Earth Observation and Geoinformation* 2007; **9**(1):50–64. ISSN: 0303-2434.
5. Fisher M. Development of a simplified Kalman filter. *ECMWF Technical Memorandum 260*, 1998.
6. Gejadze I-Y, Le Dimet FX, Shutyaev V. On analysis error covariances in variational data assimilation. *SIAM Journal on Scientific Computing* 2008; **30**:1847–1874.
7. Tian X, Xie Z, Dai A. An ensemble-based explicit four-dimensional variational assimilation method. *Journal of Geophysical Research* 2008; **113**:D21124-1–D21124-13.
8. Veersé F, Auroux D, Fisher M. Limited-memory BFGS diagonal predonditioners for a data assimilation problem in meteorology. *Optimization and Engineering* 2000; **1**:323–339.
9. Fisher M, Andersson E. Developments in 4D-var and Kalman filtering. *ECMWF Technical Memorandum 347*, 2001.
10. Evensen G. Sequential data assimilation with a non-linear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research* 1994; **99**:10143–10162.
11. Auvinen H, Bardsley JM, Haario H, Kauranne T. Large-scale Kalman filtering using the limited memory BFGS method, 2008; submitted.
12. Yang W, Navon IM, Courtier P. A new Hessian preconditioning method applied to variational data assimilation experiments using NASA general circulation model. *Monthly Weather Review* 1996; **124**:1000–1017.
13. Fisher Michael, Courtier Philippe. Estimating the covariance matrices of analysis and forecast error in variational data assimilation. *ECMWF Technical Memorandum 220*, 1995.
14. Veersé F. Variable-storage quasi-Newton operators as inverse forecast/analysis error covariance matrices in variational data assimilation. *INRIA Report No 3685*, 26 April 1999.
15. Veersé F. Variable-storage quasi-Newton operators for modeling error covariances. *Proceedings of the Third WMO International Symposium on Assimilation of Observations in Meteorology and Oceanography*, Que., Canada. World Meteorological Organization: Geneva, 7–11 June 1999.
16. LeDimet F-X, Talagrand O. Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects. *Tellus A Series* 1986; **38**:97–110.
17. Auvinen H, Haario H, Kauranne T. Optimal approximation of Kalman filtering with a temporally local 4D-Var in operational weather forecasting. In *Proceedings of the 11th ECMWF Workshop on Use of High-performance Computing in Meteorology*, Zwiefelhofer W, Mozdzynski G (eds). World Scientific: Singapore, 2005.
18. Nocedal Jorge, Wright Stephen. *Numerical Optimization*. Springer: Berlin, 1999.
19. Derber J. A variational continuous assimilation technique. *Monthly Weather Review* 1989; **117**:2437–2446.
20. Greenstadt J. Variations on variable-metric methods. *Mathematics of Computation* 1970; **24**:1–22.
21. Lorenz EN. Predictability: a problem partly solved. *Proceedings of the Seminar on Predictability*, ECMWF, vol. 1, Reading, Berkshire, U.K., 1996; 1–18.
22. Lorenz EN, Emanuel KA. Optimal sites for supplementary weather observations: simulation with a small model. *Journal of Atmospheric Science* 1998; 399–414.
23. Leutbecher M. *A Data Assimilation Tutorial Based on the Lorenz-95 System*, European Centre for Medium-Range Weather Forecasts Web Tutorial. Available from: www.ecmwf.int/newsevents/training/lecture_notes/pdf_files/ASSIM/Tutorial.pdf.
24. Zupanski M. Maximum likelihood ensemble filter: theoretical aspect. *Monthly Weather Review* 2005; **133**:1710–1726.
25. Rabier F, Järvinen H, Klinker E, Mahfouh JF, Simmons A. The ECMWF operational implementation of four dimensional variational assimilation. Part I: experimental results with simplified physics. *Quarterly Journal of the Royal Meteorological Society* 2000; **126**:1143–1170.
26. Auvinen H, Haario H, Kauranne T. Variational Kalman filtering on parallel computers. In *Proceedings of the 12th ECMWF Workshop on Use of High-Performance Computing in Meteorology*, Zwiefelhofer W, Mozdzynski G (eds). World Scientific: Singapore, 2007.

27. Zhang F, Zhang M, Hansen JA. Coupling ensemble Kalman filter with four-dimensional data assimilation. *Advances in Atmospheric Sciences* 2009; **26**:1–8.
28. Hamill TM, Snyder C. A hybrid ensemble Kalman filter-3D variational analysis scheme. *Monthly Weather Review* 2000; **128**:2905–2919.
29. Nocedal J. Updating quasi-Newton matrices with limited storage. *Mathematics of Computation* 1980; **35**(151): 773–782.
30. Byrd RH, Nocedal J, Schnabel RB. Representations of quasi-Newton matrices and their use in limited memory methods. *Mathematical Programming* 1994; **63**(1–3):129–156.

PUBL. IV

LARGE-SCALE KALMAN FILTERING USING
THE LIMITED MEMORY BFGS METHOD[*]

H. AUVINEN[†], J. M. BARDSLEY[‡], H. HAARIO[†], AND T. KAURANNE[†]

Abstract. The standard formulations of the Kalman filter (KF) and extended Kalman filter (EKF) require the storage and multiplication of matrices of size $n \times n$, where n is the size of the state space, and the inversion of matrices of size $m \times m$, where m is the size of the observation space. Thus when both m and n are large, implementation issues arise. In this paper, we advocate the use of the limited memory BFGS method (LBFGS) to address these issues. A detailed description of how to use LBFGS within both the KF and EKF methods is given. The methodology is then tested on two examples: the first is large-scale and linear, and the second is small scale and nonlinear. Our results indicate that the resulting methods, which we will denote LBFGS-KF and LBFGS-EKF, yield results that are comparable with those obtained using KF and EKF, respectively, and can be used on much larger scale problems.

Key words. Kalman filter, Bayesian estimation, large-scale optimization

AMS subject classifications. 65K10, 15A29

1. Introduction. The Kalman filter (KF) for linear dynamical systems and the extended Kalman filter (EKF) for nonlinear but smoothly evolving dynamical systems are popular methods for use on state space estimation problems. As the dimension of the state space becomes very large, as is the case, for example, in numerical weather forecasting, the standard formulations of KF and EKF become computationally intractable due to matrix storage and inversion requirements.

Computationally efficient variants of KF and EKF have been proposed for use on such large-scale problems. The Reduced Rank Kalman Filter or Reduced Order extended Kalman filter (see, e.g., [4, 7, 30]) project the dynamical state vector of the model onto a lower dimensional subspace. The success of the approach depends upon a judicious choice of the reduction operator. Moreover, since the reduction operator is typically fixed in time, the dynamics of the system may not be correctly captured; see [9] for more details.

In the context of numerical weather forecasting, a great deal of attention has been given to the filtering problem. The current state of the art is 4D-Var (see, e.g., [9, 23]), which utilizes a variational formulation of an initial value estimation problem [11, 14, 17]. 4D-Var has been shown to be identical to a Kalman smoother when the model is assumed to be perfect [16]. The resulting quadratic minimization problem is very large-scale (10^4-10^7 unknowns) and so efficient numerical optimization methods are needed. Similar to the methods in the previous paragraph, the partial orthogonal decomposition is used in [5] to reduce the dimensionality of the 4D-Var minimization problem. A more standard approach is to implement a preconditioned conjugate gradient method [8, 12, 22, 26]. In this context, a number of different preconditioners have been tested.

In this paper, we take a different approach. In particular, we focus our attention on the Kalman filter itself, using the limited memory BFGS (LBFGS) [22] iterative method for the required large-scale matrix storage and inversion within KF and EKF. More specifically, suppose $\mathbf{Ax} = \mathbf{b}$ is a system, with symmetric positive definite matrix $\mathbf{A}$, that requires solution,

[†]Department of Mathematics and Physics, Lappeenranta University of Technology, Lappeenranta, Finland ({`harri.auvinen,heikki.haario,tuomo.kauranne`}`@lut.fi`).

[‡]Department of Mathematical Sciences, University of Montana, Missoula, Montana 59812, USA (`bardsleyj@mso.umt.edu`).

and or a low storage approximation of $\mathbf{A}^{-1}$ is needed, within the KF or the EKF algorithms. In such cases, LBFGS is a natural choice, since it generates both a sequence of approximations of $\mathbf{A}^{-1}\mathbf{b}$, and a sequence of symmetric positive definite matrices $\{\mathbf{B}_k^{-1}\}$ approximating $\mathbf{A}^{-1}$. This choice is further supported by the result that $\mathbf{A}^{-1}\mathbf{b}$ is reached within a finite number of LBFGS iterations (assuming exact arithmetic) [20, 21], and that if k and full memory BFGS is used $\mathbf{B}_{n+1}^{-1} = \mathbf{A}^{-1}$ [22, Chapter 8]. A result regarding the accuracy of the LBFGS Hessian approximations to $\mathbf{A}^{-1}$ is also given in [21]; it depends upon the eigenvalues of $\mathbf{A}^{1/2}\mathbf{B}^k\mathbf{A}^{1/2}$; see Appendix A. The use of LBFGS and its success in the examples that we consider can be motivated by the fact that the covariance matrices being approximated are approximately low rank. In applications of interest [6], covariance information is contained in slowly varying, low dimensional subspaces, making accurate low-rank approximations possible.

The idea of using the LBFGS method in variational data assimilation is not new; see, e.g., [10, 13, 17, 25, 27, 28, 29, 31]. In many of these references, the LBFGS Hessian or inverse Hessian is used as a preconditioner for conjugate gradient iterations, and even as an approximate error covariance matrix for the background term in 3D- and 4D-Var variational data assimilation. However, in the method presented here, the LBFGS method is further used for matrix inversion, in order to propagate effectively the state estimate covariance information forward in time. Moreover, we apply our methodology to the Kalman filter itself, not to the variation formulation used by the 3D- and 4D-Var methods [17]. LBFGS can also be incorporated in a fully variational formulation of the Kalman filter; see [2].

As has been stated, the approach presented here uses the LBFGS algorithm directly within the context of the Kalman filter. The equivalence of LBFGS and a certain preconditioned conjugate gradient method (see [21] and Appendix A) suggests that our approach and those cited above are similar. One advantage of the cited approaches, however, is that they can be incorporated into existing 3D- and 4D-Var codes used in practice.

An application of a similar methodology that could be used in conjunction with 3D- and 4D-Var is presented in [1, 2]. The aim of the current paper is to demonstrate the use of LBFGS within the standard (non-variational) formulation of the linear or extended Kalman filter.

The paper is organized as follows. We present KF and EKF in Section 2, and then in Section 3 we present LBFGS-KF and LBFGS-EKF. We test these methods with two numerical experiments in Section 4. Conclusions are then given in Section 5, and implementation details of the LBFGS algorithm are contained in Appendix A.

2. The Kalman filter. We consider the coupled system of discrete, linear stochastic difference equations given by

$$\mathbf{x}_k = \mathbf{M}_k\mathbf{x}_{k-1} + \varepsilon_k^p, \tag{2.1}$$

$$\mathbf{y}_k = \mathbf{K}_k\mathbf{x}_k + \varepsilon_k^o. \tag{2.2}$$

In the first equation, $\mathbf{x}_k$ denotes the $n \times 1$ state of the system at time k; $\mathbf{M}_k$ is the $n \times n$ linear evolution operator; and ε_k^p is a $n \times 1$ random vector known as the prediction error and is assumed to characterize errors in the model and corresponding numerical approximations. In the second equation, $\mathbf{y}_k$ denotes the $m \times 1$ observed data; $\mathbf{K}_k$ is the $m \times n$ linear observation operator; and ε_k^o is an $m \times 1$ random vector known as the observation error. The prediction error ε_k^p and observation error ε_k^o are assumed to be independent and normally distributed, with zero means and symmetric positive definite covariance matrices $\mathbf{C}_{\varepsilon_k^p}$ and $\mathbf{C}_{\varepsilon_k^o}$, respectively.

We assume, in addition, that we have in hand estimates of both the state $\mathbf{x}_{k-1}^{est}$ and its positive definite covariance matrix $\mathbf{C}_{k-1}^{est}$ at time $k - 1$. Moreover, we assume that $\mathbf{x}_{k-1}^{est}$, ε_k^p,

and ε_k^o are independent random vectors. The goal is to then estimate $\mathbf{x}_k^{est}$ and its covariance $\mathbf{C}_k^{est}$. In Bayesian terms, equation (2.2) provides the likelihood function in the estimation step, while (2.1) gives the prior. By the assumptions on ε_k^p and ε_k^o, the prior mean and the prior covariance matrix are directly obtained from (2.1),

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

The full negative-log posterior density given (up to an additive constant) by Bayes' theorem, takes the form

$$\ell(\mathbf{x}|\mathbf{y}_k) = \frac{1}{2}(\mathbf{y}_k - \mathbf{K}_k\mathbf{x})^T \mathbf{C}_{\boldsymbol{\varepsilon}_k^o}^{-1}(\mathbf{y}_k - \mathbf{K}_k\mathbf{x}) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_k^p)^T (\mathbf{C}_k^p)^{-1}(\mathbf{x} - \mathbf{x}_k^p),$$

and hence, we have

$$(2.3) \qquad\qquad \mathbf{x}_k^{est} = \arg\min_{\mathbf{x}} \ell(\mathbf{x}|\mathbf{y}_k),$$

$$(2.4) \qquad\qquad \mathbf{C}_k^{est} = \nabla^2 \ell(\mathbf{x}|\mathbf{y}_k)^{-1}.$$

Equations (2.3) and (2.4) motivate the variational Kalman filter, which is the subject of [1]. However, they can also be used to derive the Kalman filter. In particular, noting that (2.3) and (2.4) can be alternatively written (see [24] for detail)

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

where

$$\mathbf{G}_k = \mathbf{C}_k^p \mathbf{K}_k^{\mathrm{T}}(\mathbf{K}_k\mathbf{C}_k^p\mathbf{K}_k^{\mathrm{T}} + \mathbf{C}_{\boldsymbol{\varepsilon}_k^o})^{-1},$$

we have the following standard formulation of the Kalman filter.

The Kalman Filter
    Step 0: Select initial guess $\mathbf{x}_0^{est}$ and covariance $\mathbf{C}_0^{est}$, and set $k = 0$.
    Step 1: Compute the evolution model estimate and covariance:
        (i) Compute $\mathbf{x}_k^p = \mathbf{M}_k\mathbf{x}_{k-1}^{est}$;
        (ii) Compute $\mathbf{C}_k^p = \mathbf{M}_k\mathbf{C}_{k-1}^{est}\mathbf{M}_k^{\mathrm{T}} + \mathbf{C}_{\boldsymbol{\varepsilon}_k^p}$.
    Step 2: Compute the Kalman filter estimate and covariance:
        (i) Compute the Kalman Gain $\mathbf{G}_k = \mathbf{C}_k^p\mathbf{K}_k^{\mathrm{T}}(\mathbf{K}_k\mathbf{C}_k^p\mathbf{K}_k^{\mathrm{T}} + \mathbf{C}_{\boldsymbol{\varepsilon}_k^o})^{-1}$;
        (ii) Compute the Kalman filter estimate $\mathbf{x}_k^{est} = \mathbf{x}_k^p + \mathbf{G}_k(\mathbf{y}_k - \mathbf{K}_k\mathbf{x}_k^p)$;
        (iii) Compute the estimate covariance $\mathbf{C}_k^{est} = \mathbf{C}_k^p - \mathbf{G}_k\mathbf{K}_k\mathbf{C}_k^p$.
    Step 3: Update $k := k + 1$ and return to Step 1.
Note that it is typical to take the initial covariance $\mathbf{C}_k^{est}$ to be diagonal.

The extended Kalman filter (EKF) is the extension of KF when (2.1), (2.2) are replaced by

$$(2.5) \qquad\qquad \mathbf{x}_k = \mathcal{M}(\mathbf{x}_{k-1}) + \varepsilon_k^p,$$

$$(2.6) \qquad\qquad \mathbf{y}_k = \mathcal{K}(\mathbf{x}_k) + \varepsilon_k^o,$$

where $\mathcal{M}$ and $\mathcal{K}$ are (possibly) nonlinear functions. EKF is obtained by the following simple modification of the above algorithm: in Step 1 (i), use instead $\mathbf{x}^p = \mathcal{M}(\mathbf{x}_k^{est})$, and define

$$(2.7) \qquad\qquad \mathbf{M}_k = \frac{\partial \mathcal{M}(\mathbf{x}_{k-1}^{est})}{\partial \mathbf{x}}, \quad \text{and} \quad \mathbf{K}_k = \frac{\partial \mathcal{K}(\mathbf{x}^p)}{\partial \mathbf{x}}.$$

We note that $\mathbf{M}_k$ and $\mathbf{K}_k$ can be computed or estimated in a number of ways. For example, the numerical scheme that is used in the solution of either the evolution or the observation model defines a tangent linear code, which can be used to compute (2.7); see, e.g., [12, 15]. However, a more common, but also more computationally expensive, approach is to use finite differences to approximate (2.7).

3. Using LBFGS for large-scale Kalman filtering. When the model size n is large, the Kalman filter is known to be prohibitively expensive to implement. This motivates several alternative approaches—most notably the 4D-Var method—used in large-scale applications such as numerical weather forecasting [8, 9, 11, 12, 17, 23, 26] and oceanography [4]. We instead focus our attention on the Kalman filter itself, using the limited memory BFGS (LBFGS) [22] iterative method for the required large-scale matrix storage and inversion within KF and EKF.

First, we give a general description of the LBFGS method for minimizing

$$(3.1) \qquad\qquad q(\mathbf{u}) = \frac{1}{2}\langle \mathbf{A}\mathbf{u}, \mathbf{u} \rangle - \langle \mathbf{b}, \mathbf{u} \rangle,$$

where $\mathbf{A}$ is an $n \times n$ symmetric positive definite matrix and $\mathbf{b}$ is an $n \times 1$ vector. It is given by

The LBFGS method for quadratic minimization
$\nu := 0$;
$\mathbf{u}_0 :=$ initial guess;
$\mathbf{B}_0^{-1} :=$ initial inverse Hessian approximation;
begin quasi-Newton iterations
$\quad \mathbf{g}_\nu := \nabla q(\mathbf{u}_\nu) = \mathbf{A}\mathbf{u}_\nu - \mathbf{b}$;
$\quad \mathbf{v}_\nu = \mathbf{B}_\nu^{-1}\mathbf{g}_\nu$;
$\quad \tau_\nu = \langle \mathbf{g}_\nu, \mathbf{v}_\nu \rangle / \langle \mathbf{v}_\nu, \mathbf{A}\mathbf{v}_\nu \rangle$;
$\quad \mathbf{u}_{\nu+1} := \mathbf{u}_\nu - \tau_\nu \mathbf{v}_\nu$;
$\quad \mathbf{B}_\nu^{-1} :=$ LBFGS approximation to $\mathbf{A}^{-1}$;
end quasi-Newton iterations

In all of the examples considered in this paper, $\mathbf{B}_0$ was taken to be the identity matrix. The limited memory formulations for $\mathbf{B}_\nu^{-1}$, and corresponding formulas for $\mathbf{B}_\nu$, are found in Appendix append1. The stopping criteria for the LBFGS iterations is discussed in Section 4.

Some insight into the convergence properties of the LBFGS method can be obtained by an appeal to its connection with the well-known conjugate gradient (CG) method, which described in detail in [20, 21, 22, Section 9.1]. In particular, CG can be formulated as a *memoryless* BFGS method. Moreover, in the presence of exact arithmetic, LBFGS and iterates from a certain preconditioned CG iteration are identical [21], and hence finite convergence is guaranteed. Thus it seems reasonable to suspect that LBFGS will have convergence properties similar to that of CG, which are well-known and have been extensively studied. In particular, the early convergence of CG iterates within the dominant subspaces corresponding to the largest eigenvalues of the coefficient matrix is likely shared by LBFGS iterates.

Next, we describe how LBFGS was used to make the Kalman filter more efficient. We make the reasonable assumption that multiplication by the evolution and observation matrices $\mathbf{M}_k$ and $\mathbf{K}_k$, and by the covariance matrices $\mathbf{C}_{\boldsymbol{\varepsilon}_k^p}$ and $\mathbf{C}_{\boldsymbol{\varepsilon}_k^o}$, is efficient, both in terms of storage and CPU time. Additional computational challenges arise for sufficiently large n due to the storage requirements for $\mathbf{C}_k^{est}$, which becomes a full matrix as the iterations proceed. The same is also true for $\mathbf{C}_k^p$. However, given that

$$(3.2) \qquad\qquad \mathbf{C}_k^p = \mathbf{M}_k\mathbf{C}_k^{est}\mathbf{M}_k^{\mathrm{T}} + \mathbf{C}_{\boldsymbol{\varepsilon}_k^p},$$

storage issues are restricted to those for $\mathbf{C}_k^{est}$; typically the matrix $\mathbf{C}_{\boldsymbol{\varepsilon}_k^p}$ is assumed to be diagonal.

A low storage approximation of $\mathbf{C}_k^{est}$ can be obtained by applying the LBFGS algorithm to the problem of minimizing (3.1) with $\mathbf{A} = \mathbf{C}_k^{est}$ and $\mathbf{b} = \mathbf{0}$. The LBFGS matrix $\mathbf{B}_\nu^{-1}$ is then a low storage approximation of $(\mathbf{C}_k^{est})^{-1}$ and formulas for $\mathbf{C}_k^{est}$ from [3]—and also found in Appendix A.2—can be used.

Additionally, when m is sufficiently large the computation of $(\mathbf{K}_k\mathbf{C}_k^p\mathbf{K}_k^{\mathrm{T}} + \mathbf{C}_{\boldsymbol{\varepsilon}_k^o})^{-1}(\mathbf{y}_k - \mathbf{K}_k\mathbf{x}_k^p)$ that is required in Step 2, (ii) of the Kalman filter iteration will be prohibitively expensive.

For the approximation of $(\mathbf{K}_k\mathbf{C}_k^p\mathbf{K}_k^{\mathrm{T}} + \mathbf{C}_{\boldsymbol{\varepsilon}_k^o})^{-1}(\mathbf{y}_k - \mathbf{K}_k\mathbf{x}_k^p)$, we set $\mathbf{A} = \mathbf{K}_k\mathbf{C}_k^p\mathbf{K}_k^{\mathrm{T}} + \mathbf{C}_{\boldsymbol{\varepsilon}_k^o}$ and $\mathbf{b} = \mathbf{y}_k - \mathbf{K}_k\mathbf{x}_k^p$ in (3.1) and apply LBFGS to the problem of minimizing (3.1).

The LBFGS Kalman filter method can now be presented.

The LBFGS Kalman Filter (LBFGS-KF)

Step 0: Select initial guess $\mathbf{x}_0^{est}$ and covariance $\mathbf{B}_\# = \mathbf{C}_0^{est}$, and set $k = 0$.
Step 1: Compute the evolution model estimate and covariance:
    (i) Compute $\mathbf{x}_k^p = \mathbf{M}_k\mathbf{x}_k^{est}$;
    (ii) Define $\mathbf{C}_k^p = \mathbf{M}_k\mathbf{B}_\#\mathbf{M}_k^{\mathrm{T}} + \mathbf{C}_{\boldsymbol{\varepsilon}_k^p}$.
Step 2: Compute the Kalman filter estimate and covariance:
    (i) Define $\mathbf{A} = (\mathbf{K}_k\mathbf{C}_k^p\mathbf{K}_k^{\mathrm{T}} + \mathbf{C}_{\boldsymbol{\varepsilon}_k^o})$ and $\mathbf{b} = \mathbf{y}_k - \mathbf{K}_k\mathbf{x}_k^p$ in (3.1) and compute
    the LBFGS approximations $\mathbf{B}_*$ of $\mathbf{A}^{-1}$ and $\mathbf{u}_*$ of $\mathbf{A}^{-1}\mathbf{b}$.
    (ii) Compute the LBFGS-KF estimate $\mathbf{x}_{k+1}^{est} = \mathbf{x}_k^p + \mathbf{C}_k^p\mathbf{K}_k^{\mathrm{T}}\mathbf{u}_*$;
    (iii) Define $\mathbf{A} = \mathbf{C}_k^p - \mathbf{C}_k^p\mathbf{K}_k^{\mathrm{T}}\mathbf{B}_*\mathbf{K}_k\mathbf{C}_k^p(\approx \mathbf{C}_{k+1}^{est})$ and $\mathbf{b} = \mathbf{0}$ in (3.1) and compute
    the LBFGS approximation $\mathbf{B}_\#$ of $\mathbf{C}_{k+1}^{est}$ using (A.2).
Step 3: Update $k := k + 1$ and return to Step 1.

All operations with the $\mathbf{C}_k^{est}$ and $\mathbf{A}^{-1}$ are done using the LBFGS formulas; see Appendix A. As a result, LBFGS-KF is much less memory and computationally intensive than KF making its use on large-scale problems more feasible. Specifically, the storage requirements for the LBFGS estimate of $\mathbf{C}_k^{est}$ are on the order of $2n\ell + 4n$, where ℓ is the number of stored vectors in LBFGS (typically 10-20), rather than $n^2 + 4n$ [22, Section 9.1], and the computational cost for both obtaining and using this estimate is order n. Furthermore, the inversion of the $m \times m$ matrix $\mathbf{K}_k\mathbf{C}_k^p\mathbf{K}_k^{\mathrm{T}} + \mathbf{C}_{\boldsymbol{\varepsilon}_k^o}$ is carried out in order m operations and its storage requirements are on the order of $2m\ell + 4m$ rather than $m^2 + 4m$ [22].

The accuracy of the LBFGS covariance approximations is an important question. An analysis addressing this question in the similar variational setting is performed in [2]. We believe that the results of that analysis should be similar for LBFGS-KF. Thus, we choose not to repeat it here.

In the first example considered in the numerical experiments, LBFGS-KF and KF are compared and it is noted that LBFGS-KF is roughly 10 times faster, in terms of CPU time, than KF when applied to the same problem. Moreover, using our MATLAB implementation, LBFGS-KF can be used on significantly larger-scale problems.

As we have mentioned, in our implementations of KF and LBFGS-KF, the covariance matrices $\mathbf{C}_{\boldsymbol{\varepsilon}_k^p}$ and $\mathbf{C}_{\boldsymbol{\varepsilon}_k^o}$ are taken to be diagonal. This is not a necessary requirement. More structured covariances can be used, containing important prior information [17], however in order to maintain the computational efficiency and low storage requirements of LBFGS-KF, $\mathbf{C}_{\boldsymbol{\varepsilon}_k^p}$ and $\mathbf{C}_{\boldsymbol{\varepsilon}_k^o}$ must be comparable to $\mathbf{M}_k$, $\mathbf{B}_\#$ and $\mathbf{K}_k$, $\mathbf{B}_*$, respectively, in terms storage requirements and the computational cost required for their multiplication.

In the next section, we test the algorithm on two examples. The first is large-scale and linear, while the second is small-scale and nonlinear. The purpose of these experiments is

to demonstrate that LBFGS-KF and LBFGS-EKF are effective algorithms. We leave their comparison with other state of the art methods for approximate Kalman filtering [5, 8, 9, 11, 12, 14, 17, 23, 26] for a later paper.

4. Numerical experiments. In this section, we present numerical results that justify the use of LBFGS-KF. In particular, we apply the method to two examples. The first is sufficiently large-scale that the use of LBFGS-KF is justified. In particular, we assume the following forced heat equation evolution model

$$(4.1) \qquad \frac{\partial x}{\partial t} = -\frac{\partial^2 x}{\partial u^2} - \frac{\partial^2 x}{\partial v^2} + \alpha \exp\left[-\frac{(u - 2/9)^2 + (v - 2/9)^2}{\sigma^2} \right],$$

where x is a function of u and v over the domain $\Omega = \{(u, v) \mid 0 \leq u, v \leq 1\}$ and $\alpha \geq 0$. In our experiment, we will generate synthetic data using (4.1) with $\alpha > 0$ and assume that the evolution model is given by (4.1) with $\alpha = 0$, which gives a model bias. The problem can be made as large-scale as one wants via the choice of a sufficiently fine discretization of the domain Ω.

However, the well-behaved nature of solutions of (4.1)—in particular the fact that its solutions tend to a steady state—makes further experiments with a different test case a necessity. For this reason, we also test our method on a second example, which contains chaotic solutions, and hence has unpredictable behavior. In particular, we consider the simple non-linear model introduced and analyzed in [18, 19] and which is given by

$$(4.2) \qquad \frac{\partial x^i}{\partial t} = (x^{i+1} - x^{i-2})x^{i-1} - x^i + 8, \qquad i = 1, 2, \ldots, 40,$$

with periodic state space variables, i.e., $x^{-1} = x^{n-1}$, $x^0 = x^n$ and $x^{n+1} = x^1$, $n = 40$. Then (4.2) is a chaotic dynamical system (cf. [19]), which is desirable for testing purposes. As the model is computationally light and shares many characteristics with realistic atmospheric models (cf. [19]), it is commonly used for testing different data analysis schemes for weather forecasting.

4.1. An example with a large-scale linear evolution model. We perform our first experiments using model (4.1) using a uniform $N \times N$ computational grid and the standard finite difference discretization of both the time and spatial derivatives, which yields the following time stepping equation $\mathbf{x}_{k+1} = \mathbf{M}\mathbf{x}_k + \mathbf{f}$, where $\mathbf{M} = \mathbf{I} - \Delta t\mathbf{L}$. Here $\mathbf{L}$ is given by the standard finite difference discretization of the two-dimensional Laplacian operator with homogeneous Dirichlet boundary conditions, Δt is chosen to guarantee stability, and $\mathbf{f}$ is the constant vector determined by the evaluation of the forcing term in (4.1) at each of the points of the computational grid.

We define $\mathbf{K}_k = \mathbf{K}$ for all k in (2.2), where $\mathbf{K}$ is a matrix modeling an array of square sensors on the computational grid. Assuming that each sensor collects a weighted average of the state values in a 3×3 pixel region centered at every 8th pixel in both the x and y directions, $\mathbf{K}$ will have dimension $(n/64) \times n$. We assume, further, that the weighted average in the 3×3 region is defined by

$$\frac{1}{16} \left[\begin{array}{ccc} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{array} \right].$$

In our first test, we generate synthetic data using the linear stochastic equations

$$\mathbf{x}_{k+1} = \mathbf{M}\mathbf{x}_k + \mathbf{f} + N(\mathbf{0}, (0.5\sigma_{\text{ev}})^2\mathbf{I}),$$
$$\mathbf{y}_{k+1} = \mathbf{K}\mathbf{x}_{k+1} + N(\mathbf{0}, (0.8\sigma_{\text{obs}})^2\mathbf{I}),$$

with $\alpha = 3/4$ in (4.1) and where σ_{ev}^2 and σ_{obs}^2 are chosen so that the signal to noise ratios, defined by $\|\mathbf{x}_0\|^2/n^2\sigma_{\text{ev}}^2$ and $\|\mathbf{K}\mathbf{x}_0\|^2/n^2\sigma_{\text{obs}}^2$ respectively, are both 50. The initial condition used for the data generation was

$$[\mathbf{x}_0]_{ij} = \exp[-((u_i - 1/2)^2 + (v_j - 1/2)^2)],$$

where (u_i, v_j) is the ijth grid point.

In our implementation of KF, we used the biased models

$$(4.3) \qquad\qquad \mathbf{x}_{k+1} = \mathbf{M}\mathbf{x}_k + N(\mathbf{0}, \sigma_{\text{ev}}^2\mathbf{I}),$$

$$(4.4) \qquad\qquad \mathbf{y}_{k+1} = \mathbf{K}\mathbf{x}_{k+1} + N(\mathbf{0}, \sigma_{\text{obs}}^2\mathbf{I}),$$

with initial conditions $\mathbf{x}_0 = \mathbf{0}$ and $\mathbf{C}_0^{\text{est}} = 0.001\mathbf{I}$ in Step 0 of the filter. We compare the results obtained with the LBFGS-KF and KF, where $N = 2^j$ with j taken to be the largest positive integer so that memory issues do not arise in the MATLAB implementation for the standard KF. For the computer on which the simulations were done (a laptop with 2G RAM memory and a 1.8 GHz Core 2 Duo processor) the largest such j was 5, making $N = 32$ and $n = 1024$. We note that in our implementation of the LBFGS method within LBFGS-KF, we have chosen to take only 10 LBFGS iterations with 9 saved vectors. These choices may seem crude at first, however, more stringent stopping tolerances and/or a larger number of stored vectors did not appreciably affect the results for the examples that we considered.

The purpose of this test is to show that the results obtained with LBFGS-KF are comparable results with those obtained with KF. To do this, we present a plot in Figure 4.1 of the relative error vector, which has kth component

$$[\mathbf{relative\_error}]_k := \frac{\|\mathbf{x}_k^{\text{est}} - \mathbf{x}_k\|}{\|\mathbf{x}_k\|},$$

for both the LBFGS Kalman Filter and for the standard Kalman Filter. We see that results obtained using the two approaches yield similar, though not identical, relative error curves. Both curves eventually begin to increase once the forcing term, which is not used in the state space model in KF, has a prominent effect on the data; in early iterations, it is overwhelmed by the diffused initial temperature. We also mention that in the large number of test runs that we did using this large-scale model, our implementation of the LBFGS-KF was on average about 10 times faster than was the standard KF.

Additionally, in Figure 4.2, we present the filter estimates obtained from both KF and LBFGS-KF together with the true state values at time points 35 and 70. Note that in the early iterations of the filter, represented by time point 35, the filter does not detect the source because it is overwhelmed by the initial temperature and is not contained in the model (4.3), (4.4). However, the source is detected once the initial temperature has sufficiently dissipated.

For a thorough comparison, we perform the same test using values for σ_{ev}^2 and σ_{obs}^2 that yield signal-to-noise ratios of 10. The relative error curves in Figure 4.3 result. Interestingly, LBFGS-KF provides better results at the beginning of the filtering period than does KF. This can be explained, we believe, by the fact that a regularization of sorts is implicitly implemented via the use of a truncated LBFGS algorithm.

Finally, we choose σ_{ev}^2 and σ_{obs}^2 as in the original experiment, but take $\alpha = 2$, which has the effect of making the state space model that is used within LBFGS-KF and KF less accurate. When this is done, we obtain the solution curves appearing in Figure 4.4. Thus it seems that as the underlying evolution model becomes less accurate and the noise level remains moderately low KF provides better results
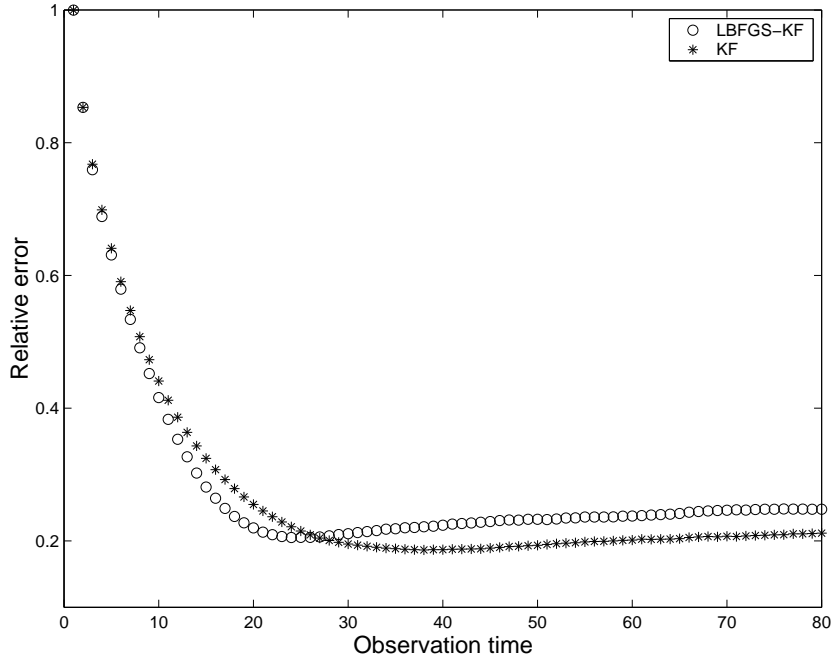
FIG. 4.1. *Relative error curves for KF ($*$) and LBFGS-KF (o). The horizontal axis represents the observation time.*

In order to show that satisfactory results can also be obtained for much larger scale problems, we take $j = 8$ which gives $N = 256$ and $n = 65536$. We take all other parameter values to be those of the original experiment. Note, however, that the stability condition of the time stepping scheme requires a much smaller time step for this problem. A relative error plot similar to those in the previous example is given in Figure 4.5. We do not include an error curve for the Kalman filter because memory issues prevent its implementation on our computer for either $N = 128$ ($n = 16384$) or $N = 256$ ($n = 65536$).

The previous large-scale example remains orders of magnitude smaller than the typical size of systems considered in practical weather models. We stopped at $N = 256$ because our experiments were performed on a laptop that could not handle a larger-scale problem. However, the discussion of computational cost and storage in the paragraph following the description of the LBFGS-KF algorithm suggests that it scales well with problem size. Thus the use of LBFGS-KF on much larger-scale problems should be feasible. Efficiency can be further improved if several time steps are allowed in the forward model for each Kalman filter iteration, much as is done in 4D-Var implementations. In addition, to the degree that LBFGS is parallelizable, LBFGS-KF will also be parallelizable.

4.2. An example with a small-scale, nonlinear evolution model. In our next example, we apply EKF and LBFGS-EKF to the problem of estimating the state variables from data generated using the nonlinear, chaotic evolution model (4.2). To generate the data, a time integration of the model was first performed using a fourth order Runge-Kutta (RK4) method with time-step $\Delta t = 0.025$. Analysis in [19] suggests that when (4.2) is used as a test example for weather forecasting data assimilation algorithms, the characteristic time scale is such that the above Δt corresponds to 3 hours, which we will use in what follows. It is also noted in
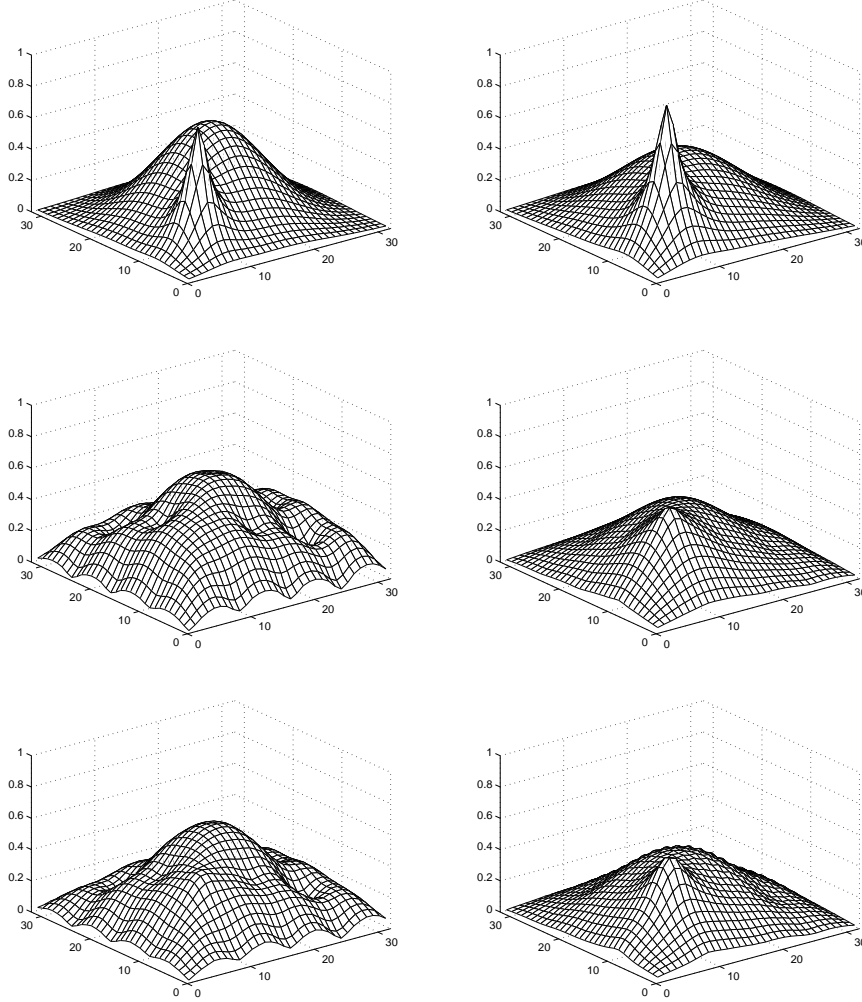
FIG. 4.2. *The top plots are of the true state at time points 35 and 70. The middle plots are of the Kalman filter estimates at time points 35 and 70. The bottom plots are of the LBFGS Kalman filter estimates at time points 35 and 70.*

[19] that for $\Delta t \leq 0.5$, the RK4 method is stable. The "true data" was generated by taking 42920 time steps of the RK4 method, which corresponds to 5365 days. The initial state at the beginning of the data generation was $x^{20} = 8 + 0.008$ and $x^i = 8$ for all $i \neq 20$.

The observed data is then computed using this true data. In particular, after a 365 day long initial period, the true data is observed at every other time step and at the last 3 grid points in each set of 5; that is, the observation matrix is $m \times n$ with nonzero entries

$$[\mathbf{K}]_{rs} = \begin{cases} 1 & (r,s) \in \{(3j+i, 5j+i+2) \mid i=1,2,3, \; j=0,1,\ldots,7\}, \\ 0 & \text{otherwise.} \end{cases}$$

The observation error is simulated using the Gaussian random vector $N(\mathbf{0}, (0.15\,\sigma_{\text{clim}})^2\mathbf{I})$
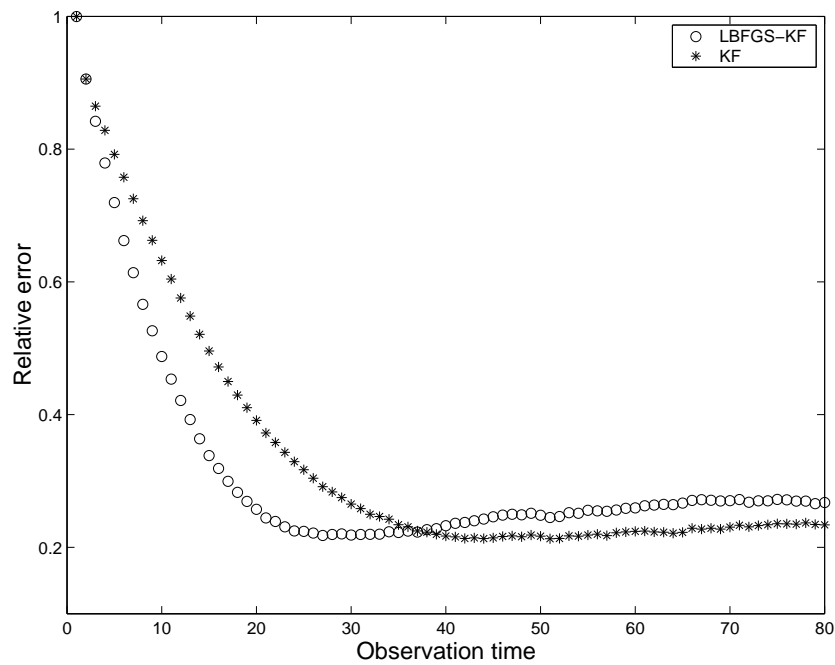
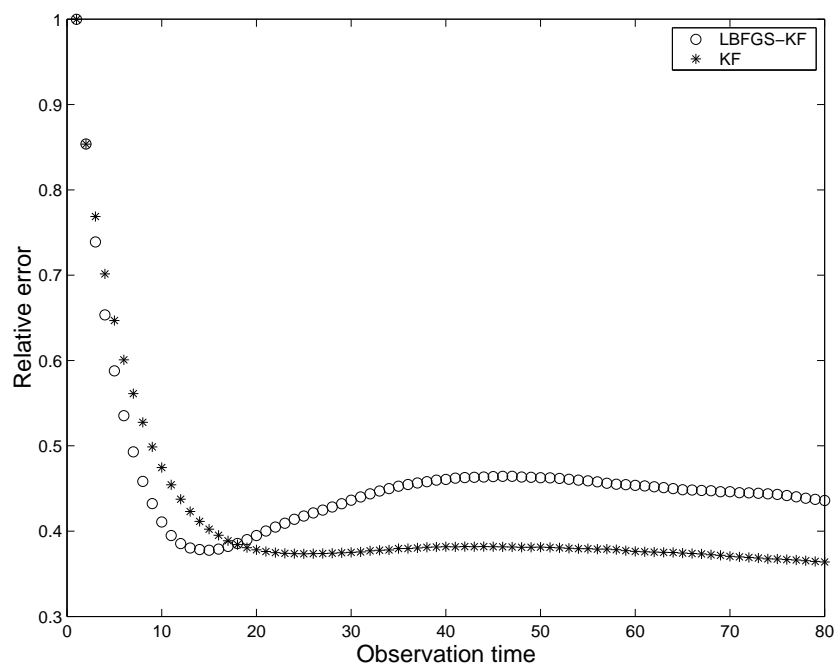FIG. 4.3. *Relative error curves for KF (∗) and LBFGS-KF (o). The horizontal axis represents the observation time.*

FIG. 4.4. *Relative error curves for KF (∗) and LBFGS-KF (o). The horizontal axis represents the observation time.*
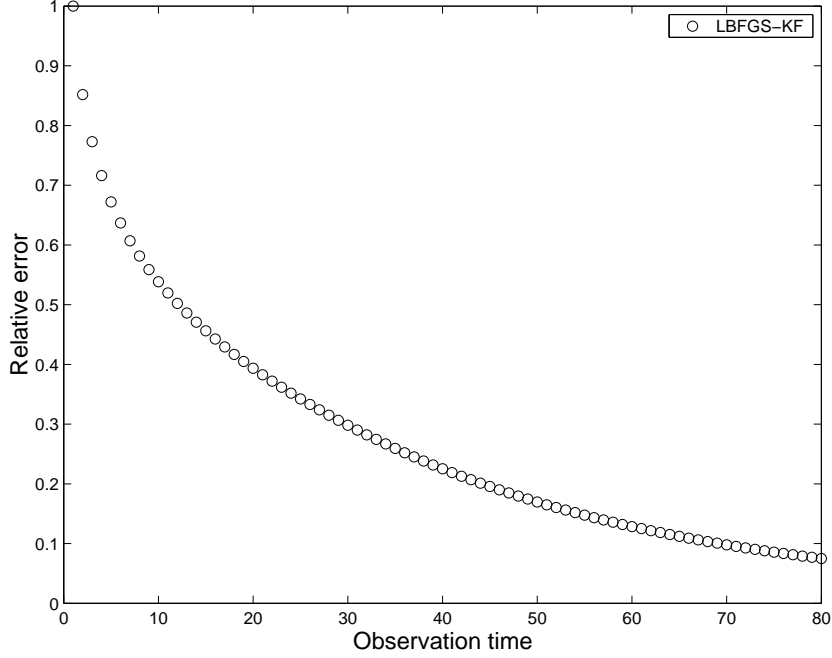
FIG. 4.5. *Relative error curves for LBFGS-KF. The horizontal axis represents the observation time.*

where σ_{clim} is a standard deviation of the model state used in climatological simulations, $\sigma_{\mathrm{clim}} := 3.6414723$. The data generation codes were written in MATLAB and were transcribed by us from the `scilab` codes written by the author of [15].

In our application of EKF and LFBGS-EKF, we assume the coupled stochastic system

$$(4.5) \qquad \mathbf{x}_{k+1} = \mathcal{M}(\mathbf{x}_k) + N(\mathbf{0}, (0.05\,\sigma_{\mathrm{clim}})^2 \mathbf{I}),$$

$$(4.6) \qquad \mathbf{y}_{k+1} = \mathbf{K}\mathbf{x}_{k+1} + N(\mathbf{0}, (0.15\,\sigma_{\mathrm{clim}})^2 \mathbf{I}),$$

where $\mathcal{M}(\mathbf{x}_k)$ is obtained by taking two steps of the RK4 method applied to (4.2) with initial condition $\mathbf{x}_k$ with time-step 0.025. We note that if the noise term is removed from (4.5) and the above initial condition is used, our data generation scheme results.

Due to the fact that $\mathcal{M}$ is a nonlinear function, EKF must be used; see (2.5) and (2.6). Since $\mathcal{K} := \mathbf{K}$ in (2.6) is linear, $\mathbf{K}_k = \mathbf{K}$ for all k in (2.7). However, a linearization of the nonlinear evolution function $\mathcal{M}$ is required. Fortunately, the computation of $\mathbf{M}_k$ in (2.7) is performed by a routine in one of the `scilab` codes mentioned above and that we have adapted for our use in MATLAB.

The initial condition used in implementation of both the EKF and LBFGS-EKF is defined by $[\mathbf{x}_{t0}]_i = [\mathbf{x}_{t0}^{\mathrm{true}}]_i + N(0, (0.3\,\sigma_{\mathrm{clim}})^2)$ for all i, and the initial covariance was taken to be $\mathbf{C}_0^{est} = (0.13\,\sigma_{\mathrm{clim}})^2 \mathbf{I}$. In our implementation of the LBFGS method within LBFGS-EKF, we computed 10 iterations with 9 saved vectors.

In order to analyze the accuracy of the state estimates $\mathbf{x}_k^{est}$ obtained by both EKF and LBFGS-EKF we plot the vector with components

$$(4.7) \qquad [\mathbf{rms}]_k = \sqrt{\frac{1}{40}\|\mathbf{x}_k^{est} - \mathbf{x}_k^{true}\|^2}$$
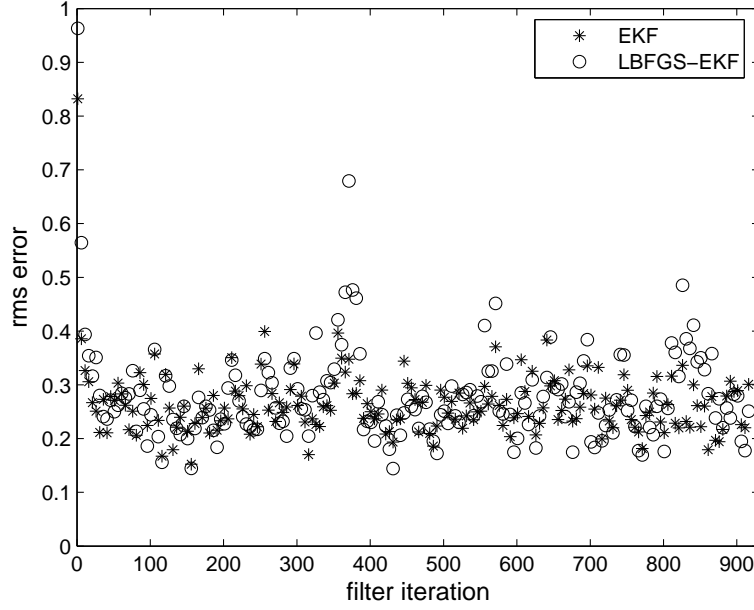
FIG. 4.6. *Plot of residual mean square error for LBFGS-EKF (o) and EKF (∗) applied to (4.2).*

in Figure 4.6. We can see that the two methods yield comparable results.

In order to compare the forecasting abilities of the two approaches, we compute the following forecast statistics at every 8th observation. Take $j \in \mathcal{I} := \{8i \mid i = 1, 2, \ldots, 100\}$ and define

$$(4.8) \qquad [\mathbf{forcast\_error}_j]_i = \frac{1}{40}\|\mathcal{M}_{4i}(\mathbf{x}_j^{est}) - \mathbf{x}_{j+4i}^{true}\|^2, \quad i = 1, \ldots, 20,$$

where $\mathcal{M}_n$ denotes a forward integration of the model by n time steps with the RK4 method. Thus this vector gives a measure of forecast accuracy given by the respective filter estimate up to 80 time steps, or 10 days out. This allows us to define the forecast skill vector

$$(4.9) \qquad [\mathbf{forecast\_skill}]_i = \frac{1}{\sigma_{\text{clim}}}\sqrt{\frac{1}{100}\sum_{j \in \mathcal{I}}[\mathbf{forecast\_error}_j]_i}, \quad i = 1, \ldots, 20,$$

which is plotted in Figure 4.7. The results show that the forecasting skill of the two methods is very similar, which suggests that on the whole, the quality of the LBFGS-EKF estimates is as high as those obtained using EKF. Figure 4.7 also illustrates the fact that the Lorenz 95 model (4.2) is truly chaotic.

In the test cases considered here, a linear or linearized model matrix $\mathbf{M}_k$ has been available. This is not true in important examples such as in numerical weather forecasting, where, on the other hand, a tangent linear code [14] is available that provides a means of computing the matrix vector product $\mathbf{M}_k\mathbf{x}$.

5. Conclusions. The standard implementations of KF and EKF become exceedingly time and memory intensive as the dimension of the underlying state space increases. Several
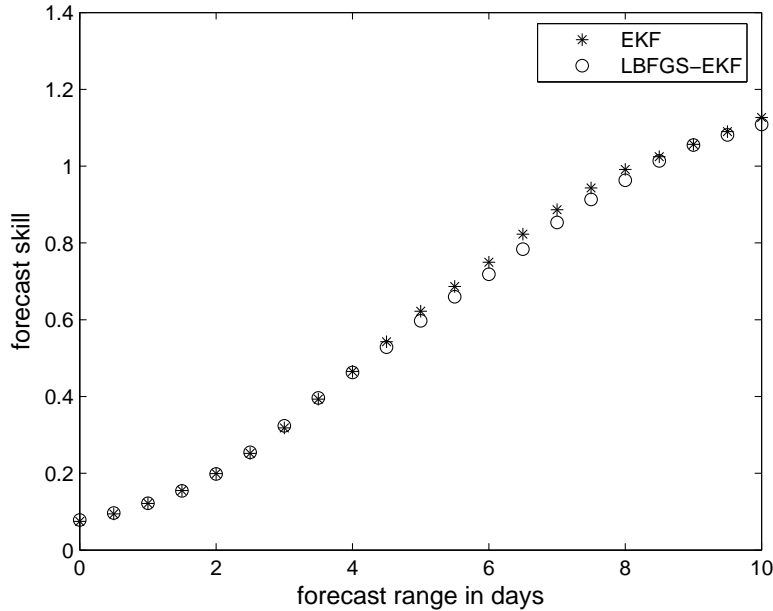
FIG. 4.7. *Plot of forecast skill vector for LBFGS-EKF (o) and EKF (∗) applied to (4.2).*

variants of KF and EKF have been proposed to reduce the dimension of the system, thus making implementation in high dimensions possible. The Reduced Rank Kalman Filter or Reduced Order extended Kalman filter (see, e.g., [4, 7, 30]) project the dynamical state vector of the model onto a lower dimensional subspace. The success of this approach depends on a judicious choice of the reduction operator. Moreover, since the reduction operator is typically fixed in time, they can suffer from "covariance leaks" [9]. A typical cause of this is that nonlinear systems do not generally leave any fixed linear subspace invariant.

In this paper, we propose the use of the limited memory BFGS (LBFGS) minimization method in order to circumvent the computational complexity and memory issues of standard KF and EKF. In particular, we replace the $n \times n$, where n is the dimension of the state space, covariance matrices within KF and EKF with low storage approximations obtain using LBFGS. The large-scale matrix inversions required in KF and EKF implementations are also approximated using LBFGS. The resulting methods are denoted LBFGS-KF and LBFGS-EKF, respectively.

In order to test these methods, we consider two test cases: large-scale linear and small scale nonlinear. LBFGS-KF is applied in the large-scale linear case and is shown to be effective. In fact, our method exceeds the speed of standard KF by an order of magnitude, and yields comparable results when both methods can be applied. Furthermore, it can be used on much larger scale problems. In the nonlinear, small scale case, LBFGS-EKF is implemented and is also shown to give results that are comparable to those obtained using standard EKF. We believe that these results suggest that our approach deserves further consideration.

The symmetric rank one (SR1) quasi-Newton method for minimizing (3.1) could be another attractive method for use within KF and EKF, since it also yields estimates of both the minimizer and inverse Hessian. The main drawback of using SR1, however, is that the

inverse Hessian approximations are not guaranteed to be positive definite.

Appendix A. In this appendix, we give a general description of the LBFGS method for minimizing

$$q(\mathbf{u}) = \frac{1}{2}\langle \mathbf{A}\mathbf{u}, \mathbf{u} \rangle - \langle \mathbf{b}, \mathbf{u} \rangle,$$

where $\mathbf{A}$ is an $n \times n$ symmetric positive definite matrix and $\mathbf{b}$ is an $n \times 1$ vector. It is given by.

The LBFGS method for quadratic minimization
$\nu := 0$;
$\mathbf{u}_0 :=$ initial guess;
$\mathbf{B}_0^{-1} :=$ initial inverse Hessian approximation;
begin quasi-Newton iterations
    $\mathbf{g}_\nu := \nabla q(\mathbf{u}_\nu) = \mathbf{A}\mathbf{u}_\nu - \mathbf{b}$;
    $\mathbf{v}_\nu = \mathbf{B}_\nu^{-1}\mathbf{g}_\nu$;
    $\tau_\nu = \langle \mathbf{g}_\nu, \mathbf{v}_\nu \rangle / \langle \mathbf{v}_\nu, \mathbf{A}\mathbf{v}_\nu \rangle$;
    $\mathbf{u}_{\nu+1} := \mathbf{u}_\nu - \tau_\nu \mathbf{v}_\nu$;
    $\mathbf{B}_\nu^{-1} :=$ LBFGS approximation to $\mathbf{A}^{-1}$;
end quasi-Newton iterations

A.1. The limited memory approximation for $\mathbf{A}^{-1}$. The BFGS matrix $\mathbf{B}_\nu^{-1}$ is computed using recursion

$$\mathbf{B}_{\nu+1}^{-1} = \mathbf{V}_\nu^{\mathrm{T}}\mathbf{B}_\nu^{-1}\mathbf{V}_\nu + \rho_\nu \mathbf{s}_\nu \mathbf{s}_\nu^{\mathrm{T}},$$

where

$$\begin{aligned}
\mathbf{s}_\nu &:= \mathbf{u}_{\nu+1} - \mathbf{u}_\nu, \\
\mathbf{d}_\nu &:= \nabla q(\mathbf{u}_{\nu+1}) - \nabla q(\mathbf{u}_\nu), \\
\rho_\nu &:= 1/\mathbf{d}_\nu^{\mathrm{T}}\mathbf{s}_\nu, \\
\mathbf{V}_\nu &:= \mathbf{I} - \rho_\nu \mathbf{d}_\nu \mathbf{s}_\nu^{\mathrm{T}}.
\end{aligned}$$

However, for large-scale problems the storage of the full matrix $\mathbf{B}_\nu^{-1}$ is infeasible, which motivates the limited storage version of the algorithm. At iteration ν, suppose that the j vector pairs $\{\mathbf{s}_i, \mathbf{d}_i\}_{i=\nu-j}^{\nu-1}$ are stored. Then we the LBFGS approximation of the inverse Hessian is given by

$$\begin{aligned}
\mathbf{B}_\nu^{-1} = &(\mathbf{V}_{\nu-1}^{\mathrm{T}} \cdots \mathbf{V}_{\nu-j}^{\mathrm{T}})(\mathbf{V}_{\nu-j} \cdots \mathbf{V}_{\nu-1}) \\
&+ \rho_{\nu-j}(\mathbf{V}_{\nu-1}^{\mathrm{T}} \cdots \mathbf{V}_{\nu-j+1}^{\mathrm{T}})\mathbf{s}_{\nu-j}\mathbf{s}_{\nu-j}^{\mathrm{T}}(\mathbf{V}_{\nu-j+1} \cdots \mathbf{V}_{\nu-1}) \\
&+ \rho_{\nu-j+1}(\mathbf{V}_{\nu-1}^{\mathrm{T}} \cdots \mathbf{V}_{\nu-j+2}^{\mathrm{T}})\mathbf{s}_{\nu-j+1}\mathbf{s}_{\nu-j+1}^{\mathrm{T}}(\mathbf{V}_{\nu-j+2} \cdots \mathbf{V}_{\nu-1}) \\
&+ \vdots \\
&+ \rho_{\nu-1}\mathbf{s}_{\nu-1}\mathbf{s}_{\nu-1}^{\mathrm{T}}.
\end{aligned}$$

(A.1)

Assuming exact arithmetic and that j, we have that $\mathbf{u}_\nu$ converges to the unique minimizer of q in at most n iterations, and if n iterations are performed $\mathbf{B}_{n+1}^{-1} = \mathbf{A}^{-1}$ [22]. In the implementation in this paper, however, $j << n$ and LBFGS iterations are truncated before convergence is obtained.

It is proven in [21] that for quadratic minimization problems and exact line searches LBFGS is equivalent to preconditioned conjugate gradient with fixed preconditioner $\mathbf{B}_0$. Thus its convergence rate is given by [22]

$$\|\mathbf{u}_k - \mathbf{u}^*\|_{\tilde{\mathbf{A}}}^2 \le \left(\frac{\lambda_{N-k} - \lambda_1}{\lambda_{N-k} + \lambda_1} \right)^2 \|\mathbf{u}_0 - \mathbf{u}^*\|_{\tilde{\mathbf{A}}}^2,$$

where $\mathbf{u}^* = \mathbf{A}^{-1}\mathbf{b}$, $\tilde{A} = \mathbf{B}_0^{1/2}\mathbf{A}\mathbf{B}_0^{1/2}$ is $N \times N$ with eigenvalues $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_N$ and $\|\mathbf{v}\|_{\tilde{\mathbf{A}}} = \mathbf{v}^T\tilde{\mathbf{A}}\mathbf{v}$.

A.2. A low storage approximation of A. The required formulas are given in [3] and take the following form. Let

$$\mathbf{S}_\nu = [\mathbf{s}_{\nu-j}, \ldots, \mathbf{s}_{\nu-1}], \qquad \mathbf{D}_\nu = [\mathbf{d}_{\nu-j}, \ldots, \mathbf{d}_{\nu-1}],$$

then

(A.2) $$\mathbf{B}_\nu = \xi_\nu\mathbf{I} - [\xi_\nu\mathbf{S}_\nu \quad \mathbf{D}_\nu] \left[\begin{array}{cc} \xi_\nu\mathbf{S}_\nu^T\mathbf{S}_\nu & \mathbf{L}_\nu \\ \mathbf{L}_\nu^T & -\mathbf{D}_\nu \end{array} \right]^{-1} \left[\begin{array}{c} \xi_\nu\mathbf{S}_\nu^T \\ \mathbf{D}^T \end{array} \right],$$

where $\mathbf{L}_\nu$ and $\mathbf{D}_\nu$ are the $j \times j$ matrices

$$(\mathbf{L}_\nu)_{i,j} = \left\{ \begin{array}{ll} \mathbf{s}_{\nu-j-1+i}^T\mathbf{d}_{\nu-j-1+j}, & \text{if } i > j, \\ 0, & \text{otherwise.} \end{array} \right.$$

and

$$\mathbf{D}_\nu = \text{diag}(\mathbf{s}_{\nu-j}^T\mathbf{d}_{\nu-j}, \ldots, \mathbf{s}_{\nu-1}^T\mathbf{d}_{\nu-1}).$$

We note that when $\xi_\nu = 1$ for all ν in (A.2), we have an exact equality between $\mathbf{B}_\nu$ in (A.2) and (A.1). However, we have found that a more accurate Hessian approximation is obtained if, following [22], we use the scaling $\xi_\nu = \mathbf{d}_{\nu-1}^T\mathbf{d}_{\nu-1}/\mathbf{s}_{\nu-1}^T\mathbf{d}_{\nu-1}$ instead.

We note that the middle matrix in (A.2) has size $2j \times 2j$, which is of reasonable size provided j is not too large, and its inversion can be carried out efficiently using a Cholesky factorization that exploits the structure of the matrix; see [3] for details.

REFERENCES

[1] H. AUVINEN, H. HAARIO, AND T. KAURANNE, *Optimal approximation of Kalman filtering with a temporally local 4D-Var in operational weather forecasting*, in Proceedings of the 11th ECMWF Workshop on Use of High-Performance Computing in Meteorology, W. Zwiefelhofer and G. Mozdzynski, eds., World Scientific, London, 2005.

[2] H. AUVINEN, J. M. BARDSLEY, H. HAARIO, AND T. KAURANNE, *The variational Kalman filter and an efficient implementation using limited memory BFGS*, Internat. J. Numer. Methods Fluids, to appear.

[3] R. H. BYRD, J. NOCEDAL, AND R. B. SCHNABEL, *Representations of quasi-Newton matrices and their use in limited memory methods*, Math. Program., 63 (1994), pp. 129–156.

[4] M. A. CANE, R. N. MILLER, B. TANG, E. C. HACKERT, AND A. J. BUSALACCHI, *Mapping tropical Pacific sea level: data assimilation via reduced state Kalman filter*, J. Geophys. Res., 101 (1996), pp. 599–617.

[5] Y. CAO, J. ZHU, I. M. NAVON, AND Z. LUO, *A reduced order approach to four-dimensional variational data assimilation using proper orthogonal decomposition*, Internat. J. Numer. Methods Fluids, 53 (2007), pp. 1571–1583.

[6] R. DALEY, *Atmospheric Data Analysis*, Cambridge University Press, Cambridge, UK, 1991.

[7] D. P. DEE, *Simplification of the Kalman filter for meteorological data assimilation* Q. J. R. Meteorol. Soc., 117 (1990), pp. 365–384.

[8] M. FISHER, *Minimization algorithms for variational data assimilation*, in Recent Developments in Numerical Methods for Atmospheric Modelling, Reading, UK, September 7–11, 2008, European Centre for Medium-Range Weather Forecasts, Reading, UK, 1998, pp. 364–385.

[9] M. FISHER AND E. ANDERSSON, *Developments in 4D-var and Kalman filtering*, European Centre for Medium-Range Weather Forecasts Technical Memorandum 347, Reading, UK, 2001.

[10] M. FISHER AND P. COURTIER, *Estimating the covariance matrices of analysis and forecast error in variational data assimilation*, European Centre for Medium-Range Weather Forecasts Technical Memorandum 220, Reading, UK, 1995.

[11] M. FISHER, M. LEUTBECHER, AND G. KELLEY, *On the equivalence between Kalman smoothing and weak-constraint four-dimensional variational data assimilation*, Q. J. R. Meteorol. Soc., 131 (2005), pp. 3235–3246.

[12] M. FISHER, J. NOCEDAL, Y. TRÉMOLET, AND S. J. WRIGHT, *Data assimilation in weather forecasting: a case study in PDE-constrained optimization*, Optim. Eng., 10 (2009), pp. 1389–4420.

[13] I. Y. GEJADZE, F.-X. LE DIMET, AND V. SHUTYAEV, *On analysis error covariances in variational data assimilation*, SIAM J. Sci. Comput., 30 (2008), pp. 1847–1874.

[14] F.-X. LEDIMET AND O. TALAGRAND, *Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects*, Tellus Ser. A, 38 (1986), pp. 97–110.

[15] M. LEUTBECHER, *A data assimilation tutorial based on the Lorenz-95 system*, European Centre for Medium-Range Weather Forecasts Web Tutorial, Reading, UK. Available at www.ecmwf.int/newsevents/training/meteorological_presentations/pdf/DA/ToyModel_notes.pdf.

[16] Z. LI AND M. NAVON, *Optimality of variational data assimilation and its relationship with the Kalman filter and smoother*, Q. J. R. Meteorol. Soc., 127 (2008), pp. 661–683.

[17] A. C. LORENC, *Modelling of error covariances by 4D-Var data assimilation*, Q. J. R. Meteorol. Soc., 129 (2003), pp. 3167–3182.

[18] E. N. LORENZ, *Predictability: A problem partly solved*, in Predictability of Weather and Climate, Tim Palmer and Renate Hagedorn, eds., Cambridge University Press, Cambridge, UK, 2006, pp. 40–58. Originally presented in a 1996 European Centre for Medium-Range Weather Forecasts workshop.

[19] E. N. LORENZ AND K. A. EMANUEL, *Optimal sites for supplementary weather observations: simulation with a small model*, J. Atmospheric Sci., (1998), pp. 399–414.

[20] L. NAZARETH, *A relationship between the BFGS and conjugate gradient algorithms and it implications for new algorithms*, SIAM J. Numer. Anal., 16 (1979), pp. 794–800.

[21] J. NOCEDAL, *Updating quasi-Newton matrices with limited storage*, Math. Comp., 35 (1980), pp. 773–782.

[22] J. NOCEDAL AND S. WRIGHT, *Numerical Optimization*, Springer, New York, 1999.

[23] F. RABIER, H. JÄRVINEN, E. KLINKER, J. F. MAHFOUF, AND A. SIMMONS, *The ECMWF operational implementation of four-dimensional variational assimilation. Part I: experimental results with simplified physics*, Q. J. R. Meteorol. Soc., 126 (2000), pp. 1143–1170.

[24] C. D. RODGERS, *Inverse Methods for Atmospheric Sounding: Theory and Practice*, World Scientific, London, 2000.

[25] X. TIAN, Z. XIE, AND A. DAI, *An ensemble-based explicit four-dimensional variational assimilation method*, J. Geophys. Res., 113 (2008), pp. D21124 (1–13).

[26] J. TSHIMANGA, S. GRATTON, A. T. WEAVER, AND A. SARTENAER, *Limited-memory preconditioners, with application to incremental four-dimensional variational data assimilation*, Q. J. R. Meteorol. Soc., 134 (2008), pp. 751–769.

[27] F. VEERSÉ, *Variable-storage quasi-Newton operators as inverse forecast/analysis error covariance matrices in variational data assimilation*, INRIA Report 3685, April 26, 1999.

[28] F. VEERSÉ, *Variable-storage quasi-Newton operators for modeling error covariances*, in Proceedings of the Third WMO International Symposium on Assimilation of Observations in Meteorology and Oceanography, 1999, Quebec City, Canada, World Meteorological Organization, Geneva.

[29] F. VEERSÉ, D. AUROUX, AND M. FISHER, *Limited-memory BFGS diagonal predonditioners for a data assimilation problem in meteorology*, Optim. Eng., 1 (2000), pp. 323–339.

[30] A. VOUTILAINEN, T. PYHÄLAHTI, K. KALLIO, J. PULLIAINEN, H. HAARIO, AND J. KAIPIO, *A filtering approach for estimating lake water quality from remote sensing data*, Int. J. Appl. Earth Observation & Geoinformation, 9 (2007), pp. 50–64.

[31] W. YANG, I. M. NAVON, AND P. COURTIER, *A new Hessian preconditioning method applied to variational data assimilation experiments using NASA general circulation models*, Monthly Weather Rev., 124 (1996), pp. 1000–1017.

ACTA UNIVERSITATIS LAPPEENRANTAENSIS

329.   VISKARI, KIRSI. Drivers and barriers of collaboration in the value chain of paperboard-packed consumer goods. 2008. Diss.

330.   KOLEHMAINEN, EERO. Process intensification: From optimised flow patterns to microprocess technology. 2008. Diss.

331.   KUOSA, MARKKU. Modeling reaction kinetics and mass transfer in ozonation in water solutions. 2008. Diss.

332.   KYRKI, ANNA. Offshore sourcing in software development: Case studies of Finnish-Russian cooperation. 2008. Diss.

333.   JAFARI, AREZOU. CFD simulation of complex phenomena containing suspensions and flow through porous media. 2008. Diss.

334.   KOIVUNIEMI, JOUNI. Managing the front end of innovation in a networked company environment – Combining strategy, processes and systems of innovation. 2008. Diss.

335.   KOSONEN, MIIA. Knowledge sharing in virtual communities. 2008. Diss.

336.   NIEMI, PETRI. Improving the effectiveness of supply chain development work – an expert role perspective. 2008. Diss.

337.   LEPISTÖ-JOHANSSON, PIIA. Making sense of women managers´ identities through the constructions of managerial career and gender. 2009. Diss.

338.   HYRKÄS, ELINA. Osaamisen johtaminen Suomen kunnissa. 2009. Diss.

339.   LAIHANEN, ANNA-LEENA. Ajopuusta asiantuntijaksi – luottamushenkilöarvioinnin merkitys kunnan johtamisessa ja päätöksenteossa. 2009. Diss.

340.   KUKKURAINEN, PAAVO. Fuzzy subgroups, algebraic and topological points of view and complex analysis. 2009. Diss.

341.   SÄRKIMÄKI, VILLE. Radio frequency measurement method for detecting bearing currents in induction motors. 2009. Diss.

342.   SARANEN, JUHA. Enhancing the efficiency of freight transport by using simulation. 2009. Diss.

343.   SALEEM, KASHIF. Essays on pricing of risk and international linkage of Russian stock market. 2009. Diss.

344.   HUANG, JIEHUA. Managerial careers in the IT industry: Women in China and in Finland. 2009. Diss.

345.   LAMPELA, HANNELE. Inter-organizational learning within and by innovation networks. 2009. Diss.

346.   LUORANEN, MIKA. Methods for assessing the sustainability of integrated municipal waste management and energy supply systems. 2009. Diss.

347.   KORKEALAAKSO, PASI. Real-time simulation of mobile and industrial machines using the multibody simulation approach. 2009. Diss.

348.   UKKO, JUHANI. Managing through measurement: A framework for successful operative level performance measurement. 2009. Diss.

349.   JUUTILAINEN, MATTI. Towards open access networks – prototyping with the Lappeenranta model. 2009. Diss.

350. LINTUKANGAS, KATRINA. Supplier relationship management capability in the firm´s global integration. 2009. Diss.

351. TAMPER, JUHA. Water circulations for effective bleaching of high-brightness mechanical pulps. 2009. Diss.

352. JAATINEN, AHTI. Performance improvement of centrifugal compressor stage with pinched geometry or vaned diffuser. 2009. Diss.

353. KOHONEN, JARNO. Advanced chemometric methods: applicability on industrial data. 2009. Diss.

354. DZHANKHOTOV, VALENTIN. Hybrid LC filter for power electronic drivers: theory and implementation. 2009. Diss.

355. ANI, ELISABETA-CRISTINA. Minimization of the experimental workload for the prediction of pollutants propagation in rivers. Mathematical modelling and knowledge re-use. 2009. Diss.

356. RÖYTTÄ, PEKKA. Study of a vapor-compression air-conditioning system for jetliners. 2009. Diss.

357. KÄRKI, TIMO. Factors affecting the usability of aspen (Populus tremula) wood dried at different temperature levels. 2009. Diss.

358. ALKKIOMÄKI, OLLI. Sensor fusion of proprioception, force and vision in estimation and robot control. 2009. Diss.

359. MATIKAINEN, MARKO. Development of beam and plate finite elements based on the absolute nodal coordinate formulation. 2009. Diss.

360. SIROLA, KATRI. Chelating adsorbents in purification of hydrometallurgical solutions. 2009. Diss.

361. HESAMPOUR, MEHRDAD. Treatment of oily wastewater by ultrafiltration: The effect of different operating and solution conditions. 2009. Diss.

362. SALKINOJA, HEIKKI. Optimizing of intelligence level in welding. 2009. Diss.

363. RÖNKKÖNEN, JANI. Continuous multimodal global optimization with differential evolution-based methods. 2009. Diss.

364. LINDQVIST, ANTTI. Engendering group support based foresight for capital intensive manufacturing industries – Case paper and steel industry scenarios by 2018. 2009. Diss.

365. POLESE, GIOVANNI. The detector control systems for the CMS resistive plate chamber at LHC. 2009. Diss.

366. KALENOVA, DIANA. Color and spectral image assessment using novel quality and fidelity techniques. 2009. Diss.

367. JALKALA, ANNE. Customer reference marketing in a business-to-business context. 2009. Diss.

368. HANNOLA, LEA. Challenges and means for the front end activities of software development. 2009. Diss.

369. PÄTÄRI, SATU. On value creation at an industrial intersection – Bioenergy in the forest and energy sectors. 2009. Diss.

370. HENTTONEN, KAISA. The effects of social networks on work-team effectiveness. 2009. Diss.