ABSTRACT

Title of dissertation: PROACTIVE QUALITY CONTROL BASED ON ENSEMBLE FORECAST SENSITIVITY TO OBSERVATIONS

Daisuke Hotta, Doctor of Philosophy, 2014

Dissertation directed by: Professor Eugenia Kalnay,
Department of Atmospheric and Oceanic Science

Despite recent major improvements in numerical weather prediction (NWP) systems, operational NWP forecasts occasionally suffer from an abrupt drop in forecast skill, a phenomenon called "forecast skill dropout." Recent studies have shown that the "dropouts" occur not because of the model's deficiencies but by the use of flawed observations that the operational quality control (QC) system failed to filter out. Thus, to minimize the occurrences of forecast skill dropouts, we need to detect and remove such flawed observations.

A diagnostic technique called Ensemble Forecast Sensitivity to Observations (EFSO) enables us to quantify how much each observation has improved or degraded the forecast. A recent study (Ota et al., 2013) has shown that it is possible to detect flawed observations that caused regional forecast skill dropouts by using EFSO with 24-hour lead time and that the forecast can be improved by not assimilating the detected observations.

Inspired by their success, in the first part of this study, we propose a new

QC method, which we call Proactive QC (PQC), in which flawed observations are detected 6 hours after the analysis by EFSO and then the analysis and forecast are repeated without using the detected observations. This new QC technique is implemented and tested on a lower-resolution version of NCEP's operational global NWP system. The results we obtained are extremely promising; we have found that we can detect regional forecast skill dropouts and the flawed observations after only 6 hours from the analysis and that the rejection of the identified flawed observations indeed improves 24-hour forecasts.

In the second part, we show that the same approximation used in the derivation of EFSO can be used to formulate the forecast sensitivity to observation error covariance matrix $\mathbf{R}$, which we call EFSR. We implement the EFSR diagnostics in both an idealized system and the quasi-operational NWP system and show that it can be used to tune the $\mathbf{R}$ matrix so that the utility of observations is improved.

We also point out that EFSO and EFSR can be used for the optimal assimilation of new observing systems.

Proactive Quality Control based on Ensemble Forecast Sensitivity to Observations

by

Daisuke Hotta

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2014

Advisory Committee:
Professor Eugenia Kalnay, Chair/Advisor
Professor Brian Hunt
Professor Kayo Ide
Professor Takemasa Miyoshi
Professor Konstantina Trivisa

Dedication

This dissertation is dedicated to my wife Wakana, my sister Miki, my father Hisashi and my mother Yoshiko.

My dedication also goes to Japan Meteorological Agency and its staff, especially those at the Numerical Prediction Division.

Special dedication goes to my former supervisor, Prof. Hisashi Nakamura of the University of Tokyo, and also to my hero and forever mentor, Prof. Takemasa Miyoshi of AICS/RIKEN and the University of Maryland.

Acknowledgments

This work would have been impossible without the great support from a number of people.

First of all, I would like to express my sincerest gratitude to Prof. Eugenia Kalnay. Ever since I started my career in the field of numerical weather prediction, my dream was to study under her guidance. Not many people have the luck to have their dreams come true. I am one of the few such privileged people. As I dreamed, my experience of pursuing a doctorate degree under her guidance was miraculous. Her constant encouragements and praises magically made my productivity higher than ever; if my advisor were not her, it would have taken no less than ten years to complete the dissertation. She not only academically guided my graduate study, but also helped me correct my bad habit of underestimating myself. Being not an eloquent writer like Cicero or Kenko Yoshida, I cannot find a way to describe in words how grateful I am to her. The best solution that I managed to find is: “she is my second mother.”

My dream came true not by sheer luck; the generous support from Prof. Miyoshi was fundamental in realizing this dream. However, this is just one of the many reasons why I am extremely grateful to him. He started supporting me even before I consider applying to the University of Maryland, and has always been both my academic mentor and my mental crutch. Although he resigned two years ago from being my official co-advisor, I still consider him to be my academic father.

I would like to express my deepest gratitude to my committee members, Pro-

fessors Eugenia Kalnay, Brian Hunt, Kayo Ide, Takemasa Miyoshi and Konstantina Trivisa for their careful examination of my dissertation and for constructive comments.

My graduate study at the University of Maryland was made possible by the support from the Japanese Government through the Long-term Fellowship Program. I would like to sincerely acknowledge the National Personnel Authority (NPA), Ministry of Land, Infrastructure, Transport and Tourism (MLIT), and Japan Meteorological Agency (JMA). I received tremendous help from many of my bosses at JMA. Among them, I would like to show my special gratitude to Mr. Ken'ichi Kuma, Mr. Yoshiaki Takeuchi, Mr. Ko Koizumu, Mr. Chiashi Muroi and Mr. Masanori Nishigata.

This study would have been impossible without generous support from JCSDA and also from Joint Polar Satellite System (JPSS). I am extremely grateful to its Director, Dr Sid Boukabara for allowing access to S4, the JCSDA's powerful supercomputer, and to Dr. James Jung for his kind help on setting up the quasi-operational experimentation system on S4. I would also like to sincerely appreciate Mr. Yoichiro Ota for providing us the EFSO and LETKF code, and Prof. Daryl Kleist and Dr. John Derber for constructive suggestions.

Our motivation comes from the fantastic work of the NCEP's "Dropout Team." I would especially like to thank Drs. Krishna Kummar, Jordan Alpert and Kevin Garrett for their kind support and stimulating discussion. I am also grateful to Drs. Carla Cardinali, Ricardo Todling, Daniel Holdaway and Rahul Mahajan and to Prof. Dacian Daescu for detailed discussions on forecast sensitivity studies.

For the material I presented in Appendix A, namely, semi-implicit modification to Lorenz N-cycle, I had a lot of help from Drs. Jeff Whitaker, Sajal Kar, and Jim Purser. Dr. Whitaker kindly provided me the Python code for plotting Figures A.1 and A.2. The idea for Figure A.3 was suggested by Dr. Kar. I would also like to thank Professors Eugenia Kalnay, Kayo Ide and Radu Balan for their guidance through AMSC663/AMSC664 classes.

I greatly thank Dr. Guo-Yuan Lien for always sitting behind my back in our shared office, and for always offering me interesting discussions.

AMSC is a truly wonderful interdisciplinary program with faculty and students from diverse backgrounds and specializations. I suppose managing such a diverse program is not an easy task. I would like to sincerely thank the Program Director Prof. Konstantina Trivisa and the Program Coordinator Ms. Alverda McCoy for their commitment and dedication.

Even with the world's best professors and colleagues, sometimes, graduate study can be stressful, especially with a tight schedule of only three years. For me, the best divertissement was to learn languages. I am grateful to the many friends whom I met at the weekly Coffee Chat of Language House of the University of Maryland, and also to the organizer of this fantastic event, Dr. Naime Yaramanoglu who also helped me with French. I would also like to thank my language partners, Ms. Terri Burger and Ms. Rachel Alade for helping me brush up my English, Ms. Te-Hsin Wu and Dr. Jen-Chien Chang for teaching me Mandarin, and Mr. Héctor Arthuro for teaching me Spanish. Without fun time I that spent with them, I would have had to be always distressed at any moment of my life.

Finally and most importantly, I would like to express my deepest gratitude to my beloved wife Wakana, but she modestly declined to be acknowledged.

Table of Contents

List of Tables

List of Figures

List of Abbreviations and Acronyms

3D-Var	Three-Dimensional Variational data assimilation system
4D-Var	Four-Dimensional Variational data assimilation system
ACARS	Aircraft Communications Addressing and Reporting System
ACC	Anomaly Correlation Coefficient
AFSR	Adjoint Forecast Sensitivity to $\mathbf{R}$
AGCM	Atmospheric General Circulation Model
AMSU-A	Advanced Microwave Sounding Unit A
AMV	Atmospheric Motion Vector
CMC	Canadian Meteorological Centre / Centre Météorologique Canadien
DA	Data Assimilation
DWD	Deutscher Wetterdienst (German Meteorological Service)
ECMWF	European Centre for Medium-range Weather Forecast
EFSO	Ensemble Forecast Sensitivity to Observation
EFSR	Ensemble Forecast Sensitivity to $\mathbf{R}$
EKF	Extended Kalman Filter
EnKF	Ensemble Kalman Filter
EnSRF	Ensemble Square Root Filter
FSO	Forecast Sensitivity to Observation
GEOS-5	Goddard Earth Observing System version 5
GFS	Global Forecasting System (NCEP's global NWP system or its forecast model)
GMAO	Global Modeling and Assimilation Office of NASA
GPS	Global Positioning System
GPSRO	GPS Radio Occlutation
GSI	Gridpoint Statistical Interpolation
IASI	Infrared Atmospheric Sounding Interferometer

JCSDA	Joint Center for Satellite Data Assimilation
JMA	Japan Meteorological Agency
LB04	Langland and Baker (2004)
LETKF	Local Ensemble Transform Kalman Filter
MHS	Microwave Humidity Sounder
MODIS	Moderate-Resolution Imaging Spectroradiometer
NASA	National Aeronautics and Space Agency
NCEP	National Centers for Environmental Prediction
NH	Northern Hemisphere
NMC	National Meteorological Center
NRL	Naval Research Laboratory
NWP	Numerical Weather Prediction
ODE	Ordinary Differential Equation
OI	Optimal Interpolation
OSE	Observing System Experiment
OSSE	Observing System Simulation Experiment
PDE	Partial Differential Equation
PDF	Probability Density Function
PIBAL	Pilot Balloon
PQC	Proactive Quality Control
PSAS	Physical Space Analysis Scheme
QC	Quality Control
RA filter	Robert-Asselin filter
RK4	4th-order Runge Kutta scheme
RMSE	Root Mean Square Error
S4	Supercomputer for Satellite Simulations and Data Assimilation Studies (JCSDA's cluster supercomputer)

SH	Southern Hemisphere
SPEEDY	Simplified parametrizations PrimitivE Equation DYnamics model
TRMM	Tropical Rainfall Measurement Mission
UKMO	United Kingdom MetOffice
VarQC	Variational Quality Control

Chapter 1: Introduction

1.1 Brief history of the development of Numerical Weather Prediction

Since the dawn of civilization, weather has always been beyond the reach of human comprehension. Inability of humanity to foresee the future of weather has led our ancestors to associate it with caprices of divine entities. The English word "thunder," for example, derives from "Thor," the name of a god in Scandinavian mythology; the Japanese word for thunder, "kami-nari," literally means roaring ("nari") of a god ("kami"). Until very recently, weather forecasting relied on religious fortunetelling or so-called "weather lore," namely, an accumulation of informally inherited, sometime superstitious, folklore related to the prediction of weather. As Roulstone and Norbury (2013) recounts, even in 19th century when astronomers were able to accurately calculate the orbit of a comet and the ebb and flow of ocean tides, or even to "predict" the existence of the (then-unknown) planet Neptune solely from the observational data and the first principles (*i.e.*, celestial mechanics), weather forecasters still depended on their subjective intuition. In stark contrast to this, modern weather forecasting is backed by rigid, scientific methods;

daily weather forecasts issued by national meteorological centers worldwide are now produced from numerical integration of the equations of fluid mechanics that govern the motion of the atmosphere, whose computations are carried out on state-of-the-art supercomputers. In this section, we briefly recount some of the key developments that paved the way to the emergence of modern weather forecasting to locate the role of this thesis in a broader historical perspective. Detailed historical narrations on the birth and initial development of modern weather forecasting can be found, for example, in Kalnay (2003, Chapter 1 and Appendix A), Roulstone and Norbury (2013) and Persson (2005a,b,c).

Weather prediction entered the realm of "hard science" at the beginning of 20th century when Vilhelm Bjerknes, a Norwegian physicist, fluid dynamicist and meteorologist, proposed to formulate weather prediction as a problem in physics and fluid dynamics. His proposal encouraged meteorologists and physicists to study the evolution of meteorological disturbances, not under the empirical framework of Natural History, but as an applied problem of laws of physics (Bjerknes, 1904). This paradigm shift fostered a number of important theoretical developments that led to the emergence of a new branch of meteorology, namely, "dynamic meteorology," that eventually became the foundation of modern meteorology. The establishment of dynamic meteorology as a precise quantitative discipline paved the road to the first, groundbreaking (although unsuccessful) attempt at "Numerical Weather Prediction" (NWP) conducted with hand-calculations by Lewis Fry Richardson in 1922. In NWP, the weather is viewed as a dynamical system that is governed by a set of hydrodynamic partial differential equations (PDEs); the evolution of the weather is thus predicted by numerical integration of the initial value problem. Richardson's first attempt at NWP unfortunately turned out to be a total failure, predicting an unrealistically large and rapid change in surface pressure when, in reality, the pressure barely changed. Since this failure, NWP had remained regarded as a mere, impracticable dream, until the advent of an electronic digital computer in 1940's; an interdisciplinary group of mathematicians, physicists, electrical engineers and meteorologists embarked on weather forecasting by means of numerical integration of the governing equations derived from fluid mechanics using the ENIAC, the first electronic digital computer. The first successful computer-generated weather forecast reported by Charney et al. (1950) marked the birth of NWP, a methodology that humanity finally acquired for the first time in the course of history by which objective predictions of weather were made possible based solely on observational data and the first principles (*i.e.*, the laws of physics). Following this epoch-making achievement by Charney et al. (1950), real-time operation of NWP by national meteorological administrations started in several countries, first in Sweden in 1954, followed shortly after by the United States (1955), then by Japan (1959) and the United Kingdom (1965).

The history of NWP since its birth is characterized by incessant succession of theoretical and technological innovations. Exponential growth of computing power that roughly follows the so-called Moore's Law allowed NWP centers to steadily increase the resolution of their forecast models. The first operational NWP models in the 1950's had only a single vertical layer and a horizontal resolution of about $300 \sim 500$ km with forecast area of continental scale (a total of only 300 grid points).

Over the last 50 years, the size of the problem solved by NWP systems has increased by more than million-fold; as of 2014, current operational NWP models typically predict the evolution of the entire atmosphere over the whole globe with ~ 100 vertical layers and ~ 20 km horizontal grid spacing, amounting to a total of $\sim 10^9$ grid points. Increase of resolution, together with adoption of governing equations with less and less approximation, and significant improvement in representation of parameterized physical processes, resulted in remarkable reduction of model errors.

As the model became more and more accurate, the accuracy of the initial conditions became increasingly important. NWP scientists' need for accurate initial conditions has given rise to a new, interdisciplinary branch of science: Data Assimilation (DA). In the early stages of NWP, initial conditions were produced by digitizing weather charts (*i.e.*, contour maps of meteorological quantities such as geopotential height or temperature at specific pressure level) that were manually (and subjectively) "analyzed" by human forecasters who were well trained in synoptic meteorology. To put the NWP into routine operation, however, it was necessary to automate this process (manual analysis was very laborious and time consuming) and also to attain stable quality that is independent of the skill of the forecaster in charge. The first automated objective "analysis" (*i.e.*, generation of initial conditions for NWP models) was produced by simple spatial polynomial interpolation of observations. A major difficulty in this approach, however, was that geographic distribution of observations is scarce and highly non-uniform: the number of available observations was much less than the number of grid points; there are relatively dense observations over the land, particularly near the densely-populated regions, such as North America, Europe and East Asia, but, over the oceans or less-populated, mountainous areas, observations are fewer or non-existent at all. Shortly after, however, early pioneers of NWP soon realized that short-range forecasts from previous analyses provided good estimate of the atmospheric state for such observation-scarce regions because the model can propagate information from observations in the upstream regions down toward data-scarce area. Thus, instead of producing analysis "from scratch" (*i.e.*, using only the information from latest observations) at every initial time, it is more advantageous to "correct" the short-range forecast from the previous run (which is now called "first-guess," "prior" or "background") by using the newly obtained information from observations (Bergthorsson and Döös, 1955). This allowed NWP scientists to formulate the problem under the framework of rigorous statistical estimation theory, notably Bayesian estimation, regarding the short-range forecast as the prior and the analysis as the posterior estimate. Theoretical advances thereafter and steady increase of computing power have led to successive improvements of DA methods. As of 2014, current operational DA systems of major NWP centers are based on sophisticated statistical methods, most notably, variational methods (3D-Var and 4D-Var), ensemble Kalman Filtering (EnKF) or a hybrid of both.

Innovations in observation technology have also significantly contributed to the improvement of NWP forecast skills. Most notably, exploitation of satellite-based remote-sensed data into the NWP systems has had considerable impacts in reducing forecast errors, as we show in later chapters (c.f., e.g., Figure 4.1). In fact, necessity of NWP scientists to effectively assimilate new types of observations was one of the main driving forces of theoretical advancements in DA; for example, the transition from Optimal Interpolation method (OI) to 3D-Var was motivated in part by necessity to directly assimilate satellite radiances rather than indirectly assimilate them via “retrievals.”

Today’s operational global NWP systems assimilate $\sim 10^6$ observations twice or four times each day to yield the best estimate of the atmospheric state in $\sim 10^9$-dimensional space. Global atmospheric DA is arguably the largest “inverse problem” ever solved; remarkably enough, this huge inverse problem is being solved on a routine basis every day, without ever being stopped, producing tremendous benefit to the society, economy and humanity.

As a consequence of these advancements, skills of NWP forecast have gone through dramatic improvement. For example, Simmons (2011) reports that, for the northern hemisphere (NH) extratropics, the skill of 6-day forecast achieved by the European Centre for Medium-range Weather Forecast (ECMWF) operational system at 2010 is as good as that of 3-day forecast of 1980’s operational system. The improvement is even larger for the southern hemisphere (SH) extratropics with 2010 7-day forecast beating 1980 3-day forecast. Similar dramatic improvements in forecast skill are also accomplished by other NWP centers.

1.2 Background: Importance of minimizing “forecast skill dropouts”

Although current operational NWP centers boast very high forecast skills on average, they occasionally suffer from abrupt drop of forecast performances. This phenomenon, denoted or referred to as “forecast skill dropout” by the National Centers for Environmental Prediction (NCEP) or “forecast bust” by ECMWF, is recently identified by major NWP centers as their major setback (e.g. Alpert et al., 2009; Kumar et al., 2009; Rodwell and Coauthors, 2013).

A typical example of such “forecast skill dropouts” is shown in Figure 1.1. The skills of 6-day forecast of several NWP centers, measured as the spatial anomaly correlation coefficient (ACC) of 500 hPa geopotential height field computed over Europe, normally fluctuate around 80% or higher. At around 6th of April, however, the scores of all centers abruptly drop and the debased forecast skills persist for several days. Forecast skill “dropouts” are highly undesirable because not only do they degrade the average performance of NWP forecast but also taint its reliability: forecasts with stable accuracy (*i.e.*, forecasts whose skills do not vary much over time) are easier to use for policy makers than those with unstable accuracy. Thus, even a single occurrence of a “dropout” can significantly lower the value of forecasts for users. Furthermore, as pointed out by Kumar et al. (2009) and Rodwell and Coauthors (2013), “dropouts” tend to occur in association with dynamic instability (baroclinic instability, in particular) of synoptic situations and thus tend to occur more frequently for high-impact weather events, which makes it even more important to reduce their occurrences.

In order to minimize the occurrences of “dropouts,” we need to first understand what causes them. Recent studies at NCEP (Alpert et al., 2009; Kumar et al., 2009) have shown that there are cases of “dropouts” that occur not because of the model’s deficiencies but because of the assimilation of flawed observations

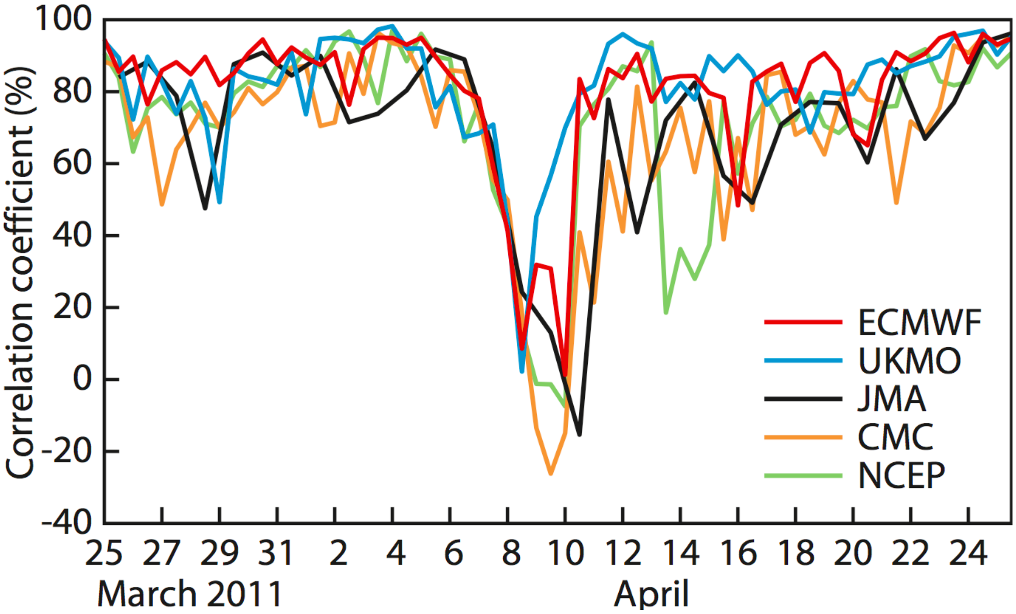


Figure 1.1: An example of a forecast skill "dropout": The day 6 forecast skill over Europe (35 °–75 °N, 12.5 °W – 42.5 °E) from the 5 major NWP centers are shown as time series over one-month period from March 25th to April 25th of the year 2011. Centers shown are ECMWF, UK Met Office (UKMO), Japan Meteorological Agency (JMA), Canadian Meteorological Centre / Centre Métórologique Canadien (CMC), and NCEP. The dates correspond to the initial time of the forecasts. The score shown is the spatial ACC of 500 hPa geopotential height. From Rodwell and Coauthors (2013).

that the operational quality control (QC) system failed to filter out: they focused on cases where NCEP's Global Forecasting System (GFS) suffered from "dropout" but ECMWF did not, and conducted hindcast experiments in which the initial conditions for the NCEP's GFS model are produced with the following procedure: first, a set of pseudo-observations are generated from analysis of ECMWF by applying observation operator to it. The pseudo-observations thus generated are then ingested to the Gridpoint Statistical Interpolation (GSI), the NCEP's global DA system, to produce initial conditions for the GFS model. When initiated from the initial conditions thus produced, the GFS model did not exhibit a drop of forecast skills, indicating that the mis-specification of the initial conditions, rather than the errors of the model, is responsible for the "dropouts." The major challenge in preventing "dropouts" resides thus in detecting the "flawed" observations that lead to erroneous initial conditions.

1.3 Ensemble Forecast Sensitivity to Observation (EFSO)

A diagnostic technique called Forecast Sensitivity to Observation (FSO) enables us to quantify how much *each* observation improved/degraded the forecast. It provides an estimate of how much *each* assimilated observation reduced or increased the forecast error measured with some specified norm. If the estimated change of forecast errors attributable to a specific observation is negative/positive, that means that its impact toward forecast improvement is positive/negative (N.B. the nomenclature is somewhat confusing; observations with positive impacts have negative values of FSO because they reduce the error, and vice versa). Thus, if the diagnostics is reliable, the "flawed" observations that cause forecast skill "dropouts" should manifest themselves as outliers whose impacts are abnormally negatively large. We propose to use an ensemble-based formulation of FSO, which we call EFSO, to identify such "flawed" observations. We anticipate that the accuracy of analyses and forecasts can be improved by repeating the data assimilation without using the identified observations (see next section). Detailed description of the EFSO is given in Section 2.2.

Promising results for the approach described above have already been obtained by a previous study: Ota et al. (2013) successfully implemented EFSO to a quasi-operational global ensemble DA system coupled with NCEP's GFS model and demonstrated that it is possible, for individual cases, to identify flawed observations that are responsible for forecast skill dropouts, by applying EFSO with 24-hour forecast lead time to relatively small horizontal regions. They report that, in all the 7 cases they examined, rejection of the identified defect observations actually improved the forecasts. Strikingly, in one of the examined cases, the regional forecast errors were reduced by as much as 30%.

Next section describes the overview of the proposed algorithm and the challenges to be explored before it is implemented to the operational system.

1.4 Proactive QC

In this thesis, we propose a new, simple QC scheme, which we denote "Proactive QC," based on the idea described in the previous section. The essence of the idea can be summarized as following: if we can identify the observations that significantly degraded forecast by using EFSO diagnostics, then we can improve the forecast by not using such deficient observations. Reflecting the simplicity of the idea, the algorithm is also very simple.

1.4.1 Overview of the algorithm

This section gives a brief overview of how the Proactive QC works. A more detailed description of the algorithm is given in Section 2.3.2.

Let 00h be the time for which Proactive QC is to be applied, and assume that the DA system has a 6-hour assimilation window (the cycling interval). The algorithm can be summarized as following:

1. Run regular DA cycle from time −06h to 00h.

2. Run regular DA cycle from time 00h to +06h.

3. Using the information available from Steps 1 and 2, detect horizontal regions where "forecast skill dropout" is likely to occur, for example where the 12-hour forecast from −06h is more accurate than the 6-hour forecast from 00h that used the observations.

4. If such regions are detected, perform 6-hour EFSO targeting at those regions

to identify "flawed" observations that are likely to have significantly degraded 6-hour forecast.

5. If such "flawed" observations are identified, repeat DA for time 00h without using the "flawed" observations.

6. Repeat Steps 1–5 shifting the time.

1.4.2 Challenges toward operational implementation

Although, as described in the previous section, promising results of this approach have been already obtained (Ota et al., 2013), several issues still remain to be addressed. First, it is not clear whether EFSO with forecast lead time as short as 6 hours is capable of detecting "flawed" observations: in the EFSO formulation proposed by Kalnay et al. (2012), forecast errors are verified against the analysis validating at the same time and date. Compared to forecast errors of very short-range such as 6 hours, errors of analysis against truth may not be negligibly small, which might make it difficult to accurately estimate forecast errors. The second, related issue is whether we can detect possible occurrences of regional "dropouts" in Step 3. after only 6 hours from the analysis. The third non-trivial question is what is the best criterion for rejection of observations given the 6-hour EFSO impacts of each observation: rejecting too many observations would lead to forecast degradation, but rejecting too few observations would make no difference. We have thus to design a method to determine the threshold that strikes best balance. The last and most important issue is whether the rejection of the "flawed" observations detected by EFSO really improves analysis and forecast, in particular, beyond the 6-hour lead time. We discuss these issues in more detail in Section 2.4 after introducing our mathematical approach.

In order to answer these questions, we conduct DA experiments using a lower-resolution version of NCEP's operational global DA system. A detailed description of the experimental setup is given in Chapter 3. The results of our experiments are described in Chapters 4–6.

1.5 Ensemble Forecast Sensitivity to Observation Covariance Matrix $\mathbf{R}$ (EFSR) and its application to tuning of $\mathbf{R}$

In DA methods currently adopted by most operational NWP systems, such as 3D-Var, 4D-Var or EnKF, information from background and observations are linearly combined with an "optimal weight." The "optimal weight" is determined, implicitly (in variational methods) or explicitly (in sequential methods including EnKF), based on the background- and observation- error covariances $\mathbf{B}$ and $\mathbf{R}$ (note that in EnKF methods, the background covariance, commonly denoted by $\mathbf{P}^b$ rather than by $\mathbf{B}$, is dynamically estimated from the perturbation of background ensemble). These covariances are parameters for DA systems that are externally prescribed and thus subject to empirical tuning. Since they determine the weight with which background and observations are combined, an accurate specification of these covariances is of vital importance. In EnKF methods, the structure of the background covariance matrix $\mathbf{P}_0^b$ is estimated as the sampled covariance of the ensemble of background fields; this allows us to fully take into account the flow-dependent structure of the errors. The magnitude of the background error covariance $\mathbf{P}_0^b$ is tuned by a technique called covariance inflation, whose adaptive methodologies are being intensively studied by a number of researchers. In this thesis, we focus on specification of observation error covariance **R**.

At present, the method most widely adopted by operational NWP systems for estimation of **R** is the statistical diagnostics proposed by Hollingsworth and Lönnberg (1986). This diagnostics assumes that the errors of observations and background are uncorrelated and that different observations are also uncorrelated (*i.e.*, **R** is diagonal), and diagnoses diagonal elements of **R** as the "jump" (discontinuity) at the diagonal element of the corresponding column of self-covariance matrix of observation innovation (O-B) vector. Expanding this idea, Talagrand (1999) showed that, if a DA system is optimal in the sense that **B** and **R** are correctly specified, then the (unknown) true **R** can be diagnosed as the expected value of the cross-covariance between the observation innovation (*i.e.*, observation minus background; O-B) and the observation residual (*i.e.*, observation minus analysis; O-A). Thus, if the two (prescribed and diagnosed) **R**'s disagree, that is an indication of mis-specification of **B** and **R**. Desroziers and Ivanov (2001) further extended this approach and proposed an iterative tuning procedure based on the consistency diagnostics; if the consistency test of Talagrand (1999) fails, then we can correct **R** iteratively by relaxing prescribed **R** to the diagnosed **R**.

Another approach to this problem recently proposed by Daescu (2008) and Daescu and Langland (2013) is to estimate the sensitivity of the forecast to the

R-matrix; using the adjoint method, they derived formulae similar to that of FSO by LB04 which estimate how a small change in each element of **R** would change the forecast errors measured with an arbitrary quadratic norm. With this sensitivity information, it is possible to tune the **R** matrix so that the forecast errors will be reduced.

In this thesis, we show that it is possible to formulate an ensemble equivalent of the forecast sensitivity to observation error covariance **R** of Daescu (2008) and Daescu and Langland (2013) based on EFSO formulation of Kalnay et al. (2012). We call our ensemble-based formulation by the acronym EFSR (ensemble forecast sensitivity to **R** matrix). We first apply this method to a simple, toy DA system to confirm the validity of our methodology. We then apply EFSR diagnostics to the lower-resolution version of operational system that we used to test "Proactive QC." Based on the EFSR results, we further perform a simple tuning experiment to see if EFSR-based tuning improves impacts from each observation type as desired.

1.6 Semi-implicit formulation of Lorenz N-cycle

The main focus of this thesis is on improvement of initial conditions for NWP models. NWP, being an initial value problem, can also be improved through improvement of the model. One way to improve a model is to adopt a more accurate time integration scheme. Despite recent advances in computational fluid dynamics, current NWP models, especially global models, still adopt a simple and rather inaccurate (only first-order) time integration scheme, namely, the semi-implicit version of the centered finite difference scheme (commonly referred to by "leapfrog") with a stabilization temporal filter (Robert, 1969, so-called Robert-Asselin (RA) filter;) which was introduced nearly half a century ago. One reason for the use of this old scheme is that, a stable, semi-implicit formulation, which allows longer timestepping beyond Courant-Friedrichs-Lewy (CFL) stability condition for the fast (but meteorologically unimportant) external gravity waves, is not known for higher-order schemes. In Appendix A, we propose a semi-implicit formulation for the so-called Lorenz N-cycle scheme (Lorenz, 1971) which is computationally as economical as the conventional leapfrog scheme but yet whose order of accuracy can be as high as 4-th order. The newly proposed formulation is applied and tested on a simplified, primitive equation atmospheric general circulation model (AGCM), whose results suggest that, for situations where taking long timestepping is not crucial, the new scheme can be an appealing alternative to the conventional leapfrog scheme. This work is presented in Appendix A.

1.7 Objectives

The main goal of this study is to develop a novel, fully flow-dependent "Proactive" QC method in which observations are rejected based on whether they actually degrade the forecast, and investigate if this method can improve operational NWP forecasts. In particular, we try to give definitive answers to the four key questions we posed in Section 1.4.2, namely:

1. Are 6 hours long enough for the detection of "flawed" observations?

2. How can we detect possible occurrences of "dropouts" after only 6 hours from analysis?

3. What is the best threshold for rejection of "flawed" observations? and

4. Does rejection of detected "flawed" observation really improve analysis and forecast?

We explore the answers to these questions by performing DA experiments and data denial experiments using a lower-resolution version of the NCEP's operational global NWP system.

The second, related goal is to investigate whether our new, ensemble-based EFSR diagnostics can be used to improve a DA system so that it can more effectively assimilate observations. We first validate our EFSR formulation by checking its consistency with respect to the adjoint-based formulation using a simple toy system. We then implement it in the lower-resolution version of the NCEP's operational global NWP system and perform an **R**-tuning experiment based on the EFSR diagnostics.

1.8 Outline

This thesis is structured as follows: the main subject of this thesis, Proactive QC, is described in Chapters 2 to 6. Chapter 2 reviews the algorithms of the currently operational QC methods and discusses their limitations. It then derives the ensemble and adjoint FSO formulation following Kalnay et al. (2012) and LB04, and presents the detailed algorithm of Proactive QC, followed by the discussion on the issues to be addressed before its operational implementation. Chapter 3 describes the forecast model, the DA system, and the observations used in our experiments, along with the details of the experimental setup. Chapter 4 examines EFSO's dependence on the forecast lead time and the verifying truth, and establishes the validity of using 6-hour EFSO for Proactive QC. Chapter 5 discusses how to improve the detection algorithm of regional forecast skill dropouts: it first introduces two different methods to divide the globe into smaller regions, and examines the statistics of regional forecast errors and EFSO values for each of the two methods, and then proposes an improved detection algorithm based on the obtained statistical relations. The first part of this thesis culminates with Chapter 6 which describes the results of the data denial experiments. It demonstrates that we can indeed improve the forecast by Proactive QC, in some cases with dramatic improvements.

The second subject of this thesis, EFSR, is described in Chapters 7 and 8. Chapter 7 derives the formulation of EFSR and verifies its effectiveness, along with that of its adjoint counterpart, through a series of idealized experiments using a simple toy system called Lorenz '96 system. Chapter 8 then applies the EFSR diagnostics to the real NWP system that we used in the experiments of Proactive QC, and performs a simple tuning experiment in which the observation error variances for some of the observation types are tuned based on the results of EFSR diagnostics. The validity of EFSR diagnostics is then verified by assessing if the EFSO impacts from the tuned observation types are improved by the tuning.

Finally, Chapter 9 synthesizes the findings of the present work, and describes future directions toward the improvement of NWP systems, in particular, how the power of EFSO technique can be used for new applications.

Appendix A presents our attempt at improving NWP through an improvement of the model rather than the DA system, more specifically, through an improvement of time integration scheme. The outline for Appendix A is given in the last paragraph of Section A.1.

Chapter 2: Proactive QC

2.1 Brief review of presently operational QC techniques and their limitations

Most DA methods assume that observation errors follow Gaussian distribution. They are thus not robust to observations with gross errors in the sense that, if an outlying observation with an abnormally large error is ingested, resulting analysis will also be abnormally inaccurate. Even assimilation of a single gross observation can have devastating effect, which can persist for several cycles due to propagation of errors through analysis-forecast cycles (e.g., see Figure 1.1). It is thus of vital importance for NWP systems to remove such outliers (observations with gross errors) from observational data set before they are fed to DA system. The process of removing such "bad" observations is called "Quality Control" (QC). This section gives a brief review of QC techniques adopted by current operational centers and points out some of their limitations. Since QC methods used by major operational centers follow more or less similar procedures, we describe the overview taking JMA's global system (JMA, 2013; Onogi, 1998) as an example.

QC is based on the following idea (e.g. Kalnay, 2003, Section 5.8): an observation is dubious if it significantly deviates from some assumed behavior or from some expected value. For example, it is reasonable to assume that vertical sounding observations should not deviate too much from hydrostatic balance, or that observations from ships should never be reported from land, etc. Observations that violate these assumptions are thus rejected during QC process. The expected values include climatology, interpolation from neighboring observations and the background (first guess from short-range forecast); observations are rejected by QC if their departures from these expected values exceed predefined thresholds.

For example, JMA's operational system implements the above idea with many layers of sequentially-executed processes which we describe in the next paragraph. Each of the processes takes output from the previous process and assigns a flag ("pass," "suspect" or "reject") to each observation based on some criteria. Data that are flagged with "pass" or "suspect" are passed to the next process, and the data that survived all these processes are finally injected to the main DA algorithm.

The QC processes at JMA can be divided into two parts; "internal QC" followed by "external QC". In the internal QC part, each observational data are quality-controlled without referring to external sources of information (hence the term "internal"). In the "external QC" part, each observational data are checked for its consistency with respect to the background or other observations in their neighborhood. The two parts consist of the following sequentially-processed steps:

Internal QC:

1. Blacklist check: some stations or instruments are registered in the blacklist

based on their “credit history” (past record of poor quality). Observations from such blacklisted stations or instruments are automatically rejected.

2. Climatological check: an observation is rejected in this step if it deviates beyond a reasonable range from its climatological record.

3. Trajectory check: observations from moving instruments are checked for consistency of their trajectories. For example, if a report from an aircraft says it moved 1,000 km in 10 minutes, or if an observation from a ship is located in the middle of a continent, such observations are highly incredible and thus are rejected during this step.

4. Temporal continuity and inter-element consistency check: surface observations from ground stations are dubious if, for example, temperature jumps by more than 20 K in 1 hour or oscillates with a very high frequency; such observations are rejected at this step.

5. Vertical consistency check: vertical sounding observations from radiosondes and aircraft are checked for vertical consistencies, including temperature lapse rate and hydrostatic balance.

External QC:

1. Gross error check: Departure d from background (O-B) is computed for each observation. An observation is flagged “pass” if $d \leq C_P$, “suspect” if $C_P < d \leq C_R$, and “reject” if $d > C_R$, where C_P and C_R are the thresholds given

separately for different types of observations. Observations flagged with "suspect" are then fed to the next "spatial consistency check"; those flagged with "pass" skip the next step and goes directly into the last "duplication check".

2. Spatial consistency check: In this step, a "minianalysis" based on optimal interpolation (OI) is performed for observations with "suspect" flag using only the "passed" observations in the vicinity of the observation in question (Lorenc, 1981, so-called "buddy check"). The departure of the suspect observation from this OI analysis (d_{OI}) is then computed. The suspect observations "revive" (are given "pass" flag) at this step if $|d - d_{OI}| < C_S$ for some thresold C_S.

3. Duplication check: Due to some transmission problems, it can happen that a same observation is reported in more than one records. Assimilating the same observations multiple times violates the assumption of DA algorithms and can degrade quality of analysis. It is thus important to remove duplicated observational data. Such duplicated observations are processed at this final step so that a same observation appears only once in the data set.

Other operational centers also adopt QC methods similar to the one described above; a unique feature of JMA's system, however, is that the thresholds C_P, C_R and C_S are allowed to vary dynamically depending on the dynamical condition of the atmosphere to account for flow-dependence of credibility of the background ("Dynamic QC"; Onogi, 1998): the underlying idea is that, if the atmospheric is in dynamically more active condition, then the background is likely to contain larger errors and thus, in such cases, the thresholds for rejection or suspicion should be relaxed. Onogi (1998) showed that spatial and temporal derivatives of the background fields, which represent the activeness of the atmospheric state, are well correlated with departures of observations from the background and proposed the scheme which adaptively assigns thresholds C_P, C_R and C_S so that they are larger when spatial or temporal derivatives of the background fields are larger.

A relatively new approach for dealing with the gross errors is to modify DA algorithms so that they are robust to observations with gross errors rather than removing them in a preprocessing step. "VarQC" (variational quality control; Anderson and Järvinen, 1999; Tavolato and Isaksen, 2010) is one of such approaches that can be applied to variational DA methods (3D-Var and 4D-Var). In VarQC, the assumed probability density function (PDF) of observation errors is modified from the usual Gaussian distribution to a superposition of two PDFs, one being the usual Gaussian distribution and the other being some PDF that accounts for gross errors. The advantages of using VarQC includes, for example, that, rather than using observations in an on/off manner, we can use observations in a continuous manner, giving less/more weights to more suspicious/credible observations, and that it allows for some flow-dependence in QC provided that the background error covariance be represented in a flow-dependent manner (which is true to some extent in 4D-Var in which the background error covariance is implicitly evolved by tangent-linear and adjoint models). Similar approach to EnKF DA systems are also being developed in recent studies (e.g., Roh et al., 2013).

It is widely acknowledged that current operational QC methods such as those reviewed here have dramatic positive impact on the accuracy of analysis and forecast. However, they still have some room for improvement. The biggest limitation is that, comparison of observations with respect to the background, which is the most significant part in the current QC procedures, can mistakenly screen out accurate observations when the quality of background is poor, a “ latent dropout” situation where correction of the background is particularly important. Flow-dependent techniques such as Dynamic QC of Onogi (1998) can partially alleviate this issue, but it has an inevitable side effect of potentially allowing flawed observations to pass QC, which can aggravate the situation.

DA is a process which brings back background that went away from the truth by pulling it back closer to observations. Thus, ideally, observation screening should be based on whether each observation is closer to the truth than the background is (if an observation is more distant from truth than the background is, it should be safer to keep the background intact). The difficulty here is that the truth is never knowable; we believe, however, that the impacts of observations on the error of short-term forecast could be used as good proxies because, if an observation is more distant from the truth than the background is, then the error of analysis introduced by assimilating it can be amplified during the course of forecast, thereby, allowing its negative impact on analysis to be detected.

2.2 EFSO

As described in the Introduction, EFSO constitutes an essential part of our "Proactive QC" algorithm. This section describes brief literature review of previous FSO (forecast sensitivity to observations) studies and derives the EFSO formulation following Kalnay et al. (2012). The relative advantages and disadvantages of adjoint-based and ensemble-based FSO are also discussed.

2.2.1 Brief literature review of FSO studies

In order to maximize the value of observations, it is important to understand how much different types of observations contribute to the improvement/degradation of NWP forecast. The importance of answering this question continues to grow as the number of observations available to DA systems increases, particularly because observations from remote-sensing instruments such as satellites and ground-based radars have increased, to a level at which even with the state-of-the-art supercomputer and computationally efficient algorithms, thinning or reduction of the ingested observational data is inevitable. Traditionally, this question was answered by conducting Observing System Experiments (OSEs) in which two sets of DA experiments are conducted, one which assimilates standard set of observations (control experiment), the other which either excludes or includes particular subset of observations. The impact of that particular subset of observations is then assessed by comparing results of the two experiments. Although an OSE can provide definitive answer to the above question based on fully nonlinear impacts of the observations, it has several major shortcomings: the first, practical point is that it is computationally extremely expensive. The second, methodological issue is that, if a new type of observations is added to the control system which already assimilates rich amount of observations, it becomes difficult to obtain statistically significant results because the abundance of the observations in the control experiment tend to obscure the impacts from the new observations.

A major breakthrough to this problem is the work by Langland and Baker (2004; LB04) who showed that it is possible, by using the adjoint technique, to estimate impacts of arbitrary subset of observation (in fact, of any single observation) onto the forecast, all at once, with only a single execution of reasonably economical computation. Several operational NWP centers, including Naval Research Laboratory (NRL; LB04), ECMWF (Cardinali, 2009), JMA (Ishibashi, 2010), National Aeronautics and Space Administration/Global Modeling and Assimilation Office (NASA/GMAO; Gelaro and Zhu, 2009; Holdaway et al., 2014) and UKMO (Lorenc and Marriott, 2013), soon adopted this powerful and economical diagnostic tool to monitor/assess impacts of different types of observations in their systems. This powerful diagnostic technique is now called FSO (forecast sensitivity to observations).

Adopted by numerous studies including those in operational NWP systems, the adjoint-based FSO of LB04 proved to be a powerful diagnostic method. Its practical applicability is somewhat limited, however, due to the requirement of an adjoint model, the development of which is notoriously difficult and laborious. An adjoint-free FSO was introduced by Liu and Kalnay (2008) who derived an ensemble-based FSO (EFSO) that can be applied to the Local Ensemble Transform Kalman Filter

(LETKF; Hunt et al., 2007; c.f., Section 3.4 of this thesis). Li et al. (2010) later fixed a minor error in Liu and Kalnay (2008) and showed that the corrected formulation improves the accuracy of estimation of observational impacts. Kalnay et al. (2012) devised a new, improved EFSO formulation which is more accurate and simpler to implement. Furthermore, unlike the former formulation which assumes that the DA uses the LETKF, the new EFSO is applicable to any form of EnKF. Ota et al. (2013) implemented the new EFSO to NCEP's Global Forecasting System (GFS) model coupled with the serial Ensemble Square Root Filter (EnSRF; Whitaker and Hamill, 2002) and gave comprehensive assessment of impacts from different types of observation. Sommer and Weissmann (2014) applied the new EFSO to a convective-scale DA system with a high-resolution limited area model and showed that EFSO-estimated impacts are consistent with the results of OSEs (data denial experiments), corroborating the validity of EFSO diagnostics.

While most previous FSO studies are primarily concerned with *statistical* features of observational impacts, such as the average impacts of each observation type or the percentage of observations that contribute positively or negatively to the forecast, Ota et al. (2013) successfully demonstrated that it is possible, for *individual cases*, to identify "flawed" observations that are responsible for forecast skill dropouts by applying EFSO with 24-hour forecast lead time to a relatively small horizontal region. This motivated us to investigate the possibility of "Proactive QC," the main subject of this thesis.

2.2.2 EFSO formulation

This section introduces the EFSO fomulation following Kalnay et al. (2012). The notation used throughout this thesis is also introduced in this section.

2.2.2.1 Notation

First, we make some remarks about our notational convention. In this thesis, we denote ensemble mean and perturbation of any variable by, respectively, an over bar and an uppercase letter; for example, the ensemble mean and perturbation of a variable $\mathbf{x}$ is represented by $\bar{\mathbf{x}}$ and $\mathbf{X}$, respectively. Vectors and matrices are represented, respectively, by lowercase and uppercase bold letters. Similarly, scalars are represented by a regular font. Superscripts a, b, f, o, t and v denote, respectively, analysis, background, forecast, observation, truth and verification. A letter with a two-character superscript preceded by δ denotes the difference of the two quantities represented by each character in the superscript; for example, $\delta\mathbf{y}^{ob}$ denotes $\mathbf{y}^{o} - \mathbf{y}^{b}$. Subscripts are used to represent valid time; if the subscript has a separator "|", as in $(\ )_{t_2|t_1}$, then it means that it is a forecast from time t_1 to time t_2. In this thesis, following notations are employed:

$n \in \mathbb{N}$: the dimension of the model space

$p \in \mathbb{N}$: the number of observations assimilated at each cycle

K: the ensemble size

$\bar{\mathbf{x}}_0^a \in \mathbb{R}^n$: ensemble mean analysis at time 0

$\bar{\mathbf{x}}_0^b = \bar{\mathbf{x}}_{0|-6}^f \in \mathbb{R}^n$: ensemble mean background at time 0

$\mathbf{x}_t^v \in \mathbb{R}^n$: verification state at time t

$\mathbf{y}_0^o \in \mathbb{R}^p$: observations assimilated at time 0

$\bar{\mathbf{y}}_0^b = \overline{H(\mathbf{x}_0^b)} \in \mathbb{R}^p$: ensemble mean background at time 0 in observation space

$\delta\bar{\mathbf{y}}_0^{ob} = \mathbf{y}_0^o - \bar{\mathbf{y}}_0^b \in \mathbb{R}^p$: observation innovation at time 0

$H : \mathbb{R}^n \to \mathbb{R}^p$: observation operator

$\mathbf{H} \in \mathbb{R}^{p\times n}$: Jacobian matrix of the observation operator H

$M_{t|0} : \mathbb{R}^n \to \mathbb{R}^n$: model integration from time 0 to t

$\mathbf{M}_{t|0} \in \mathbb{R}^{n\times n}$: tangent linear model integration from time 0 to t

$\mathbf{M}_{t|0}^T \in \mathbb{R}^{n\times n}$: adjoint model integration from time t to 0

$\mathbf{K} \in \mathbb{R}^{n\times p}$: Kalman gain matrix

$\mathbf{R} \in \mathbb{R}^{p\times p}$: observation error covariance matrix

2.2.2.2 Analysis equation

Now, consider an ensemble DA problem for time 0 and assume that the ensemble size of the system is K, the dimension of the model's state space is n, the number of observations is p and the assimilation interval is 6 hours. Our goal is to estimate how much the assimilation of each observation changes the error of t-hour forecast from initial time 0.

The analysis equation for time 0 is:

$$\bar{\mathbf{x}}_0^a - \bar{\mathbf{x}}_0^b = \mathbf{K}\delta\bar{\mathbf{y}}_0^{ob} \tag{2.1}$$

In Kalman Filter, the Kalman gain $\mathbf{K}$ can be represented using the analysis error covariance $\mathbf{P}_0^a$, as:

$$\mathbf{K} = \mathbf{P}_0^a \mathbf{H}^T \mathbf{R}^{-1} \tag{2.2}$$

In EnKF, the analysis error covariance $\mathbf{P}_0^a$ is approximated by the sampled covariance of analysis perturbation $\mathbf{X}_0^a$, giving:

$$\begin{aligned}\mathbf{K} &\approx \frac{1}{K-1}\mathbf{X}_0^a \mathbf{X}_0^{aT} \mathbf{H}^T \mathbf{R}^{-1} && (2.3)\\ &\approx \frac{1}{K-1}\mathbf{X}_0^a \mathbf{Y}_0^{aT} \mathbf{R}^{-1} && (2.4)\end{aligned}$$

where an approximation $\mathbf{H}\mathbf{X}_0^a \approx \mathbf{Y}_0^a$ (ensemble perturbation of analysis at time 0 in observation space) is used. As we will see later, this approximation turns out to be extremely powerful and plays a crucial role in our EFSO and EFSR derivation (see the next subsection and Section 7.3). Note that, in practical situations where the ensemble size K is smaller than the number of degrees of freedom of the system, the sampled covariance $\frac{1}{K-1}\mathbf{X}_0^a \mathbf{X}_0^{aT}$ must be localized to avoid sampling error. Using the above approximation, the analysis equation Eq. (2.1) can be approximated by:

$$\bar{\mathbf{x}}_0^a - \bar{\mathbf{x}}_0^b \approx \frac{1}{K-1}\mathbf{X}_0^a \mathbf{Y}_0^{aT} \mathbf{R}^{-1} \delta\bar{\mathbf{y}}_0^{ob} \tag{2.5}$$

2.2.2.3 Derivation of EFSO formulation

Now we proceed to deriving the EFSO formulation. The formulation presented here is identical to that of Kalnay et al. (2012), except that they assumed the verification state to be the analysis mean (see Eq. (2.6) and Eq. (2.7)), whereas, we allow it to be arbitrary.

In practical situations, true atmospheric state is unknowable; thus, in place of the truth state, we use some verification state $\mathbf{x}_t^v$ to estimate vectors of forecast errors:

$$\begin{aligned} \mathbf{e}^f_{t|0} &= \bar{\mathbf{x}}^f_{t|0} - \mathbf{x}^v_t && (2.6) \\ \mathbf{e}^f_{t|-6} &= \bar{\mathbf{x}}^f_{t|-6} - \mathbf{x}^v_t && (2.7) \end{aligned}$$

where $\mathbf{e}^f_{t|0}$ and $\mathbf{e}^f_{t|-6}$ represent, respectively, the errors of t-hour forecast from the analysis at time 0 and $(t+6)$-hour forecast from the analysis at time -6. These forecast error vectors are measured with a scalar metric defined by the following quadratic norm:

$$\begin{aligned} {e^f_{t|0}}^2 &= {\mathbf{e}^f_{t|0}}^T \mathbf{C} \mathbf{e}^f_{t|0} && (2.8) \\ {e^f_{t|-6}}^2 &= {\mathbf{e}^f_{t|-6}}^T \mathbf{C} \mathbf{e}^f_{t|-6} && (2.9) \end{aligned}$$

where $\mathbf{C} \in \mathbb{R}^{n \times n}$ is a positive definite weight matrix that defines the norm. In this thesis, we use dry or moist total energy norm in experiments with GFS model, and the identity matrix for experiments with 40-variable Lorenz '96 model (see Section 3.6 and Section 7.4.2).

The impact of assimilating the observations at time 0 onto the forecast at time

t can then be quantified by the difference of the two scalar errors:

$$\begin{aligned}
\Delta e^2 &= {e^f_{t|0}}^2 - {e^f_{t|-6}}^2 = {\mathbf{e}^f_{t|0}}^T \mathbf{C} \mathbf{e}^f_{t|0} - {\mathbf{e}^f_{t|-6}}^T \mathbf{C} \mathbf{e}^f_{t|-6} && (2.10)\\
&= \left(\mathbf{e}^f_{t|0} - \mathbf{e}^f_{t|-6}\right)^T \mathbf{C} \left(\mathbf{e}^f_{t|0} + \mathbf{e}^f_{t|-6}\right) && (2.11)
\end{aligned}$$

We now derive an expression for Δe^2 which can be interpreted as a sum of contributions from each observation:

$$\begin{aligned}
\Delta e^2 &= \left(\bar{\mathbf{x}}^f_{t|0} - \bar{\mathbf{x}}^f_{t|-6}\right)^T \mathbf{C} \left(\mathbf{e}^f_{t|0} + \mathbf{e}^f_{t|0}\right) \quad \because (2.6), (2.7) && (2.12)\\
&= \left(M_{t|0}(\bar{\mathbf{x}}^a_0) - M_{t|0}(\bar{\mathbf{x}}^b_0)\right)^T \mathbf{C} \left(\mathbf{e}^f_{t|0} + \mathbf{e}^f_{t|-6}\right) && (2.13)\\
&\approx \left(\mathbf{M}_{t|0}(\bar{\mathbf{x}}^a_0 - \bar{\mathbf{x}}^b_0)\right)^T \mathbf{C} \left(\mathbf{e}^f_{t|0} + \mathbf{e}^f_{t|-6}\right) && (2.14)\\
&= \left(\mathbf{M}_{t|0}\mathbf{K}\delta\bar{\mathbf{y}}^{ob}_0\right)^T \mathbf{C} \left(\mathbf{e}^f_{t|0} + \mathbf{e}^f_{t|-6}\right) \quad \because (2.1) && (2.15)\\
&\approx \frac{1}{K-1}\left(\mathbf{M}_{t|0}\mathbf{X}^a_0\mathbf{Y}^{aT}_0\delta\bar{\mathbf{y}}^{ob}_0\right)^T \mathbf{C} \left(\mathbf{e}^f_{t|0} + \mathbf{e}^f_{t|-6}\right) \quad \because (2.4) && (2.16)\\
&\approx \frac{1}{K-1}\left(\mathbf{X}^f_{t|0}\mathbf{Y}^{aT}_0\mathbf{R}^{-1}\delta\bar{\mathbf{y}}^{ob}_0\right)^T \mathbf{C} \left(\mathbf{e}^f_{t|0} + \mathbf{e}^f_{t|-6}\right) && (2.17)\\
\therefore \quad \Delta e^2 &= \delta\bar{\mathbf{y}}^{obT}_0 \frac{1}{K-1}\mathbf{R}^{-1}\mathbf{Y}^a_0\mathbf{X}^{fT}_{t|0}\mathbf{C} \left(\mathbf{e}^f_{t|0} + \mathbf{e}^f_{t|-6}\right) && (2.18)
\end{aligned}$$

where a linearization approximation $\mathbf{M}_{t|0}\mathbf{X}^a_0 \approx \mathbf{X}^f_{t|0}$ (ensemble perturbation of t-hour forecast from analysis at time 0) is used.

The above expression can be interpreted as an inner-product of the innovation vector $\delta\bar{\mathbf{y}}^{ob}_0$ and a sensitivity vector:

$$\Delta e^2 = \left\langle \delta\bar{\mathbf{y}}^{ob}_0, \frac{\partial(\Delta e^2)}{\partial \mathbf{y}} \right\rangle \tag{2.19}$$

where the sensitivity vector is

$$\frac{\partial(\Delta e^2)}{\partial \mathbf{y}} = \frac{1}{K-1}\mathbf{R}^{-1}\mathbf{Y}_0^a\mathbf{X}_{t|0}^{fT}\mathbf{C}\left(\mathbf{e}_{t|0}^f + \mathbf{e}_{t|-6}^f\right) \tag{2.20}$$

Contribution from a single observation, say the l-th element of the observation vector $\mathbf{y}_0^o$, can then be expressed as:

$$\left(\Delta e^2\right)\big|_{(\mathbf{y}_0^o)_l} = \left(\delta\bar{\mathbf{y}}_0^{ob}\right)_l \left(\frac{\partial(\Delta e^2)}{\partial \mathbf{y}}\right)_l \tag{2.21}$$

An important feature of this EFSO formulation is that it only requires standard output of ensemble DA system (the ensemble perturbation of forecast in state-space and that of analysis in observation space; see Eq. (2.18)). Also, the sensitivity vector can be computed by simple matrix-vector multiplications. Since neither matrix inversion nor eigenvalue decompositions is involved (note that $\mathbf{R}^{-1}$ is explicitly given and is usually diagonal in most DA systems), it is also computationally efficient and easy to implement.

Incidentally, as pointed out by Ota et al. (2013), the approximation Eq. (2.5) can be used to estimate how the analysis or forecast would change by not using some specific observations in the assimilation. Let $\delta\bar{\mathbf{y}}_0^{ob,\text{deny}} \in \mathbb{R}^p$ be a vector of observation innovation whose elements corresponding to the denied observations are identical to those in $\delta\bar{\mathbf{y}}_0^{ob}$ but all others are set to 0; for example, if the first and second elements of the observation vector $\mathbf{y}_0^o$ are to be denied, $\delta\bar{\mathbf{y}}_0^{ob,\text{deny}}$ is defined such that $\left(\delta\bar{\mathbf{y}}_0^{ob,\text{deny}}\right)_1 = \left(\delta\bar{\mathbf{y}}_0^{ob}\right)_1$, $\left(\delta\bar{\mathbf{y}}_0^{ob,\text{deny}}\right)_2 = \left(\delta\bar{\mathbf{y}}_0^{ob}\right)_2$ and $\left(\delta\bar{\mathbf{y}}_0^{ob,\text{deny}}\right)_l = 0$ for

$l = 3, \cdots, p$. Assume further that the analysis obtained by not using the denied observations can be approximated by the analysis obtained when those observations coincide with the corresponding background (*i.e.*, the innovation is zero). Let $\bar{\mathbf{x}}_0^{a,\text{deny}}$ be the analysis that would be obtained without using the denied observations. Then the analysis equation for $\bar{\mathbf{x}}_0^{a,\text{deny}}$ can be written as

$$\bar{\mathbf{x}}_0^{a,\text{deny}} \approx \mathbf{K}\left(\delta\bar{\mathbf{y}}^{ob} - \delta\bar{\mathbf{y}}_0^{ob,\text{deny}}\right) \tag{2.22}$$

Thus, from Eq. (2.1) and the approximate analysis equation Eq. (2.5), the change in *analysis* that would occur by not assimilating denied observations can be estimated by

$$\begin{aligned} \bar{\mathbf{x}}_0^{a,\text{deny}} - \bar{\mathbf{x}}_0^a &\approx -\mathbf{K}\delta\bar{\mathbf{y}}_0^{ob,\text{deny}} &(2.23)\\ &\approx -\frac{1}{K-1}\mathbf{X}_0^a\mathbf{Y}_0^{aT}\mathbf{R}^{-1}\delta\bar{\mathbf{y}}_0^{ob,\text{deny}} &(2.24)\end{aligned}$$

Similarly, by applying tangent linear approximation to the above equation, the change in *forecast* that would occur by not assimilating denied observations can be estimated by

$$\begin{aligned} \bar{\mathbf{x}}_{t|0}^{f,\text{deny}} - \bar{\mathbf{x}}_{t|0}^f &\approx \mathbf{M}_{t|0}\left(\bar{\mathbf{x}}_0^{a,\text{deny}} - \bar{\mathbf{x}}_0^a\right) \approx -\mathbf{M}_{t|0}\mathbf{K}\delta\bar{\mathbf{y}}_0^{ob,\text{deny}} &(2.25)\\ &\approx -\frac{1}{K-1}\mathbf{M}_{t|0}\mathbf{X}_0^a\mathbf{Y}_0^{aT}\mathbf{R}^{-1}\delta\bar{\mathbf{y}}_0^{ob,\text{deny}} &(2.26)\\ &\approx -\frac{1}{K-1}\mathbf{X}_{t|0}^f\mathbf{Y}_0^{aT}\mathbf{R}^{-1}\delta\bar{\mathbf{y}}_0^{ob,\text{deny}} &(2.27)\end{aligned}$$

(Note that the Equation (8) in Ota et al. (2013) is missing the factor $\frac{1}{K-1}$).

Furthermore, assuming that the observation error covariance matrix $\mathbf{R}$ is diagonal:

$$\mathbf{R} = \mathrm{diag}(\sigma_1^o, \sigma_2^o, \cdots, \sigma_p^o) \tag{2.28}$$

and by applying Jacobian of the observation operator $\mathbf{H}$ to Eq. (2.24), we have,

$$\begin{aligned}
\bar{\mathbf{y}}_0^{a,\mathrm{deny}} - \bar{\mathbf{y}}_0^a &:= H\left(\bar{\mathbf{x}}_0^{a,\mathrm{deny}}\right) - H\left(\bar{\mathbf{x}}_0^a\right) && (2.29)\\
&\approx \mathbf{H}\left(\bar{\mathbf{x}}_0^{a,\mathrm{deny}} - \bar{\mathbf{x}}_0^a\right) && (2.30)\\
&\approx -\frac{1}{K-1}\mathbf{H}\mathbf{X}_0^a\mathbf{Y}_0^{aT}\mathbf{R}^{-1}\delta\bar{\mathbf{y}}_0^{ob,\mathrm{deny}} && (2.31)\\
&= -\frac{1}{K-1}\mathbf{Y}_0^a\mathbf{Y}_0^{aT}\mathbf{R}^{-1}\delta\bar{\mathbf{y}}_0^{ob,\mathrm{deny}} && (2.32)
\end{aligned}$$

which, in turn, by taking derivative with respect to $\mathbf{y}_0^o$, becomes

$$S_{jl}^o := \frac{\partial\left(\bar{\mathbf{y}}_0^a\right)_l}{\partial\left(\mathbf{y}_0^o\right)_j} \approx \frac{1}{K-1}\cdot\frac{1}{\sigma_j^o}\sum_{i=1}^{k}\left\{\left(\mathbf{Y}_0^a\right)_{i,j}\cdot\left(\mathbf{Y}_0^a\right)_{i,l}\right\}. \tag{2.33}$$

This is identical to the ensemble formulation of analysis self-sensitivity matrix $\mathbf{S}^o$ given in Liu et al. (2009).

2.2.2.4 (Cross-)Covariance localization

In EnKF, when the ensemble size K is smaller than the number of degrees of freedom of the system, it is necessary to localize the sampled covariance to avoid noise from sampling errors. Our EFSO also needs localization in evaluating the cross-covariance $\frac{1}{K-1}\mathbf{Y}_0^a\mathbf{X}_{t|0}^{fT}$. With localization, Eq. (2.18) becomes:

$$\Delta e^2 \quad = \quad \delta\bar{\mathbf{y}}_0^{obT}\frac{1}{K-1}\mathbf{R}^{-1}\left[\rho \circ \left(\mathbf{Y}_0^a\mathbf{X}_{t|0}^{fT}\right)\right]\mathbf{C}\left(\mathbf{e}_{t|0}^f + \mathbf{e}_{t|-6}^f\right) \tag{2.34}$$

where the circle $\circ$ represents elementwise multiplication (Schur product) and $\rho \in \mathbb{R}^{p\times n}$ is a matrix of localization function whose (l, j)-element is a localization factor of the l-th observation onto the j-th grid point. Note that the localization function can be (in fact, should be) different from the one used in EnKF; the information from observations at time 0 is propagated and dispersed as the system evolves to time t, and the localization function in EFSO should account for this propagation. The question *"how should the localization function be propagated?"* is difficult to answer; ideally, perhaps, we should evolve the localization at the initial time by integrating it with Kolmogorov (Focker-Plank) equation associated with the model $M_{t|0}$ (e.g., Jazwinski, 1970, Section 4.6), but this computation is prohibitively expensive. For a simple system, we could use group velocities as a good proxy for speed of information propagation (Yoon et al., 2010), but for a complicated system such as comprehensive atmospheric models, even the computation of group velocity is not straightforward. Kalnay et al. (2012) and Ota et al. (2013) introduced a simple "moving localization" scheme which advects the center of the localization function by the horizontal winds of the analysis scaled by some tuning factor and showed that this scheme, despite being rather ad-hoc and simple, works well. Gasperoni and Wang (2013) devised an adaptive method based on a group filter technique (Anderson, 2007) and obtained promising results using a simplified 2-layer primitive equations system. In this thesis, we adopt the moving localization scheme of Ota et al. (2013).

2.2.2.5 Adjoint FSO formulation

The adjoint formulation uses Eq. (2.15) to evaluate the observational impacts Δe^2:

$$\Delta e^2 \quad = \quad \delta \mathbf{y}_0^{ob^T} \mathbf{K}^T \mathbf{M}_{t|0}^T \mathbf{C} \left(\mathbf{e}_{t|0}^f + \mathbf{e}_{t|-6}^f \right) \tag{2.35}$$

First, the vector $\mathbf{C} \left(\mathbf{e}_{t|0}^f + \mathbf{e}_{t|-6}^f \right)$ is integrated backward by the adjoint model $\mathbf{M}_{t|0}^T$ from time t to 0. Then, the adjoint of the DA system $\mathbf{K}^T$ is applied to the resulting vector to yield the adjoint sensitivity vector. Unlike EFSO, the adjoint FSO does not need localization. However, it does require the adjoint of the model and the DA system, which also introduces some drawbacks. Relative pros and cons of the adjoint and ensemble FSO are discussed in the next subsection.

2.2.3 Comparison of Adjoint and Ensemble FSO properties

As we can see from the derivations, adjoint and ensemble FSO are different approximations to the same problem. Thus, they should give consistent results. Kalnay et al. (2012) in fact showed that their results are consistent, given that the ensemble size K is large enough and the forecast lead time t is not too long. There are, however, some technical differences. In particular, they have different sources of shortcomings and limitations.

The adjoint FSO relies on the adjoint model $\mathbf{M}_{t|0}^{T}$ of the forecast model. For NWP models, the validity of the tangent-linear assumption quickly deteriorates as forecast lead time t increases, due to the intrinsic discontinuity of physical processes (e.g., Holdaway et al., 2014; Trémolet, 2004).

While the ensemble FSO does not require an adjoint model, the necessity to apply cross-covariance localization limits its applicability to longer forecast lead time t because an optimal way to propagate the localization function is not yet known.

Comparing the two methods from a computation and implementation aspect, the ensemble formulation has clear advantages. It is easier to implement because it does not require the adjoint model; it only requires standard output of EnKF. It is also faster and less expensive than the adjoint FSO since the formula Eq. (2.34) consists only of simple matrix multiplications.

In summary: the adjoint and ensemble FSO are equivalent (at least in theory); are both valid for short forecast lead times but due to different reasons; from a computational and technical perspective, ensemble FSO is more advantageous.

2.3 Proposed algorithm of Proactive QC

Our proposed "Proactive QC" exploits EFSO's capacity to detect observations that had detrimental effect on the forecast, as outlined in Section 1.4.1. This section describes the algorithm in more detail. Because our motivation stems from the work of Ota et al. (2013), we begin by reviewing their algorithm in detail.

2.3.1 Regional “dropout” attribution algorithm of Ota et al. (2013)

Ota et al. (2013) used EFSO with 24-hour lead time and the error norm targeted at some relatively small regions to successfully identify defect observations that caused significant drop of regional forecast skills. Below is a detailed description of their algorithm:

(1) “Regional dropout” detection

Divide the globe into small rectangular regions (roughly 30 °× 30 °in latitude-longitude grid; see Section 2.4.3), allowing overlaps by incrementing the longitude by 10 °and the latitude by 5 °. For each of the regions, compute 24-hour and 30-hour regional forecast errors defined by Eq. (2.8) and Eq. (2.9), choosing the total moist energy norm restricted to the target region as $\mathbf{C}$. From these regions, detect the occurrence of “regional dropouts” by testing if the region satisfies both of the following criteria:

1. The 24-hour forecast error $e^f_{24|0}$ is larger than its average over time by at least 1.7 times: $e^f_{24|0} > 1.7\ \overline{e^f_{24|0}}$.
2. The 24-hour forecast error $e^f_{24|0}$ is larger than the that of the 30-hour forecast started from the previous analysis $e^f_{24|-6}$ by at least 1.2 times: $e^f_{24|0} > 1.2\ e^f_{24|-6}$.

If two or more overlapping or adjacent regions are identified as “dropouts,” those areas are coalesced to form a single “dropout” region.

(2) “Flawed” observation type detection

For each "dropout" region detected in the previous step, perform 24-hour EFSO diagnostics with $\mathbf{C}$ being the moist total energy norm restricted to the target region. Then, find the types of observation that satisfy either of the following conditions:

1. The net impact from that type of observations has largest negative impact (largest positive EFSO value) among other types.
2. The sum of EFSO values of the observations of that type is positive and larger than one half of the total EFSO values (*i.e.*, the sum of EFSO values of all the assimilated observations).

(3) Selection of observations to be denied

For each of the types identified in the previous step as "flawed," choose observations to be denied in the next step by the following procedure:

1. For each vertical level (for non-radiance observations) or channel (for radiance observations from satellites), compute the sum of EFSO values within that level/channel, and sort levels/channels in descending order based on their sum of EFSO values. Select levels/channels until the total EFSO values of the selected levels/channels reaches the total EFSO value of all the observations of the type being processed. Levels/channels whose total EFSO values are larger than one half of the total EFSO value of the type being processed are also selected.
2. Divide the region into $10^\circ \times 10^\circ$sub-regions. For each of the sub-regions,

select observations to be denied by iterating over the selected levels/channels the following procedures:

Select observations whose EFSO value is larger than 10% of the largest absolute value of the EFSO values of the observations within the same level/channel and sub-region.

(4) Data denial experiment

Finally, repeat analysis without using the selected observations and run the forecast from the new analysis. Perform verification to confirm if rejection of the selected observations really improves the forecast.

2.3.2 Proactive QC algorithm

As outlined in Section 1.4.1, our proposed algorithm for Proactive QC is a straightforward extension from the algorithm of Ota et al. (2013) reviewed in detail in the previous subsection. The main difference is that, in stead of using EFSO with 24-hour lead time, we use EFSO with 6-hour lead time, so that it can be used in the operational systems without introducing too much delay. Below is a description of our proposed algorithm:

Let 00h be the initial time for which Proactive QC is to be applied. Let $t = 6$hours, the forecast lead time used in EFSO.

1. Run regular ensemble DA cycle from time -06h to 00h. Perform either ensemble or deterministic forecast for at least 12 hours to produce $\bar{\mathbf{x}}^f_{t|-6}$. This will be used to compute the forecast error $\mathbf{e}^f_{t|-6}$.

2. Run regular ensemble DA cycle from time 00h to +06h. Compute the analysis ensemble perturbation in observation space $\mathbf{Y}_0^a$. Again, perform either ensemble or deterministic forecast for at least 12 hours to produce $\bar{\mathbf{x}}_{t+6|0}^f$. This will be used in Proactive QC for the next cycle.

3. Apply "regional dropout" detection algorithm (see Chapter 5).

4. If "dropout regions" are detected, perform 6-hour EFSO targeting at those regions.

5. Apply "flawed observation" selection algorithm (see Section 6.3).

6. If "flawed" observations are identified, repeat analysis and forecast for time 00h without using the identified observations.

7. Repeat Steps. 1–6 shifting the time.

The important components of this algorithm are the "regional dropout" detection algorithm and the "flawed observation" selection algorithm. Detailed discussions on these algorithms are presented later in Chapter 5 and Section 6.3.

2.4 Issues to be resolved before operational implementation

This section expounds the issues to be addressed which we briefly described in Section 1.4.2.

2.4.1 Validity of Using 6 hours as the forecast lead time

First, a critical issue is whether a 6-hour lead time is long enough for capturing “flawed” observations: forecast errors of very short-range can be difficult to accurately estimate, if verified against analysis, because errors of analysis against truth may not be negligibly small compared to that of very short-range forecast. Furthermore, as Todling (2013) points out, analysis errors and forecast errors can be non-independent, especially when the lead time is short, which could make it even more difficult to accurately estimate forecast errors. It is thus important to carefully assess the validity of performing a 6-hour EFSO. Perhaps for this reason, since the first pioneering work of LB04, most FSO studies for global NWP systems, both adjoint-based and ensemble-based, have adopted 24 hours as the evaluation lead time. In Section 4.3 we show that 6-hour EFSO is in fact, at least qualitatively, consistent with the tried-and-true 24-hour EFSO.

2.4.2 Applicability to ensemble/variational hybrid DA systems

A growing trend in the development of operational NWP systems is to adopt a hybrid approach where a variational DA method (3D-Var or 4D-Var) partially takes in flow-dependent background error covariance from an ensemble of background fields which, in most formulations, is produced by an EnKF method (e.g., Lorenc, 2003). Most major NWP centers, including Canadian Meteorological Centre/Centre Météorologique Canadien (CMC), JMA, NASA/GMAO, NCEP and UKMO, are either adopting or developing hybrid DA methods in their operational systems (e.g., Buehner, 2005; Yoichiro Ota and Takashi Kadowaki, 2013, personal communication; Amal El Akkraoui and Ricardo Todling, 2013, personal communication; Wang et al., 2013; Kleist, 2012; Clayton et al., 2013). ECMWF has also just started testing a hybrid DA system which is based on a new hybrid scheme recently introduced by Penny (2014) where the gain matrix $\mathbf{K}$, rather than the background covariance $\mathbf{B}$ (or $\mathbf{P}^b$), is "hybridized," and has got preliminary promising results (Massimo Bonavita, 2014, personal communication). A new trend is to incorporate four-dimensionality (*i.e.*, asynchronicity of observations) into the hybrid (so-called 4DEnsVar; e.g., Kleist, 2012), which is expected to further improve the accuracy of 3D-Var based hybrid DA systems. For better or not, this trend is likely to persist for the foreseeable future. It is thus important to investigate whether EFSO is applicable to an EnKF within a hybrid DA system. A complication in applying EFSO to an EnKF within a hybrid system is that there are two different analyses, one from the variational part and the other from the EnKF part. Both can serve as the verifying truth $\mathbf{x}_t^v$ in evaluating EFSO, but no work has yet been done as to which is more appropriate or whether both are equally adequate. We address this issue in Section 4.4 by comparing two sets of EFSO impacts, one estimated using the analysis from vriational part as the verifying truth, the other estimated using that from EnKF part, and show that EFSO is not very sensitive to the choice of the verifying truth.

2.4.3 Division of the globe into sub-domains

Ota et al. (2013) proposed to detect *regional* forecast “dropouts” by dividing the globe into reasonably small regions. Here, how we should divide the globe is not a trivial question. A naïve way is to divide the globe into “rectangular” domains with equally-spaced intervals in latitude and longitude, for example, 30 °×30 ° rectangles; one may argue, however, that the resulting division is highly non-uniform, with domains near the Poles considerably smaller in area than those near the Equator. Ota et al. (2013) resolved this issue by adjusting the longitudinal spacing based on the latitude so that the areas of each region become as uniform as possible. For example, for the latitude interval 60 °N–90 °N, the longitudinal spacing is set to 60 °, while, for 15 °S–15 °N, it is only 15 °. Note that the latter, equatorial regions are shorter in longitude (about 1,667 km) than in latitude (3,333 km), which may not be desirable because, in the tropics where equatorial waves are trapped (e.g., Gill, 1982, Chapter 11), atmospheric disturbances tend to be zonally elongated (e.g., consider the so-called Matsuno-Gill pattern, Gill, 1980; Matsuno, 1966). One way to alleviate this issue is to adjust latitudinal spacing rather than longitudinal spacing; dividing the globe along the zeros of an isotropic spherical harmonic function $Y_{2m}^{m}(\lambda, \varphi)$ is a particularly advantageous option because any two neighboring anti-nodes have nearly equal distance and the areas of each cell are close to uniform (Eugenia Kalnay, 2013, personal communication; see Figure 5.1 and Table 5.1). In Chapter 5, we compare the regular, 30 °×30 ° rectangular domain decomposition with that based on the zeros of the isotropic spherical harmonics of total wavenumber 12 ($Y_{12}^{6}(\lambda, \varphi)$)

with regard to their ability to detect “regional dropouts”.

2.4.4 “Dropout” detection criteria

Ota et al. (2013) successfully detected “regional dropouts” with the criteria described in Section 2.3.1. However, the criteria is rather subjective and may not be optimal; it is also not clear whether “regional dropouts,” particularly those caused by the use of “flawed” observations, are detectable only from the information available 6 hours after the analysis. We explore this issue in Chapter 5.

2.4.5 Selection algorithm for the observations to be denied

Last and most importantly, it is not trivial how we should decide if an observation should be denied in the repeated analysis given the information from EFSO diagnostics. As we mentioned in Section 1.4.2, rejecting too many observations could be detrimental, but rejecting too few observations would have no impact at all. As we detailed in Section 2.3.1, Ota et al. (2013) carefully selected the observations to be rejected with a rather complicated, intricate algorithm; their tacit assumption is that observations with large negative impact should be clustered in localized region, both horizontally and vertically. It is not clear, however, if this assumption is justifiable with the actual data. In Section 6.3, we revisit this issue by examining the statistics and geographical distribution of EFSO impacts for individual cases.

2.5 Summary

In this chapter we briefly reviewed the current operational QC methods and discussed their limitations, and introduced the formulation of EFSO, a powerful diagnostic tool which is the main ingredient of our new QC scheme which we call Proactive QC (PQC). Its algorithm is then presented and the issues to be addressed are also discussed by critically reviewing the pioneering work of Ota et al. (2013). The later chapters present our answers to the issues we raised in the previous section.

Chapter 3: Experimental setup for Proactive QC experiments

3.1 Introduction

In this study, we examine the applicability of the Proactive QC method to the operational NWP systems using a lower-resolution version of the NCEP's operational global NWP system. This chapter describes the details of our experimental settings.

All our computations were performed on the Supercomputer for Satellite Simulations and Data Assimilation Studies (the "S4 supercomputer") of the Joint Center for Satellite Data Assimilation (JCSDA). The access to the S4 supercomputer was kindly provided by Dr. Sid Boukabara of JCSDA.

The NCEP's global NWP system that is operational since May 9th, 2011, along with several of its lower-resolution versions, are ported to S4 by Dr. James Jung of NCEP/JCSDA. With generous help from him, we modified the NCEP's global NWP system ported on the S4 supercomputer so that it fits our needs. The EFSO code and the LETKF code developed by Mr. Yoichiro Ota of JMA (see Section 3.4) were uploaded to the S4 supercomputer by Prof. Daryl Kleist of the University of Maryland (then at NCEP) with generous permission from Dr. John Derber of NCEP.

Since our experiments are an extension from Ota et al. (2013), a comparison of our experimental settings with those of their work should make it easy to grasp the overview; such comparison is summarized in Table 3.1.

3.2 The GFS model

The GFS model is the forecast model component of the NCEP's operational global NWP system. Its development as the operational model started in 1980 with the implementation at the National Meteorological Center (NMC, which later became NCEP) of a global spectral model based on the primitive equations (Sela, 1980). It has, ever since, always been one of the world's leading NWP models, with incessant continuous improvement in all aspects of NWP development.

The GFS model is a global spectral model in which the horizontal meteorological fields are internally represented by the coefficients of the spherical harmonics. For the vertical discretization, it adopts finite differencing in the σ-pressure hybrid coordinate system. For temporal discretization, it uses the classical Robert-Asselin filtered leapfrog scheme with semi-implicit treatment of the external gravity waves (Robert, 1969).

As of 2014, for the deterministic control forecast, the current operational GFS model adopts horizontal resolution of T574 (namely, the coefficients of the spherical harmonics are triangularly truncated at the total wavenumber of 574), which corresponds to grid spacing of about 28 km in the mid-latitude; for the ensemble forecasts, it adopts T254 resolution ($\sim$ 55 km). In the vertical, it has 64 layers with

	This study	Ota et al. (2013)
Forecast		
Forecast Model	GFS model	GFS model
Resolution (Deterministic)	T254L64	N/A
Resolution (Ensemble)	T126L64	T254L64
Analysis		
DA System	LETKF/3D-Var hybrid GSI	pure serial EnSRF
Member Size	80	80
Observations	same as the operational	same as the operational
Localization cut-off length	2,000 km (horizontal); 2 scale heights (vertical)	2,000 km (horizontal); 2 scale heights (vertical)
EFSO		
Verifying truth	GSI analysis and LETKF mean analysis	EnSRF mean analysis
Evaluation lead time	6, 12 and 24 hours	24 hours
Localization cut-off length	same as LETKF	same as serial EnSRF
Error norm	Dry and Moist Total Energy	Dry and Moist Total Energy
Period		
Spin-up	7 days from 2012-Jan-01-00Z to 2012-Jan-07-18Z	7 days from 2012-Jan-01-00Z to 2012-Jan-07-18Z
Statistical verification	31 days from 2012-Jan-08-00Z to 2012-Feb-07-18Z	31 days from 2012-Jan-08-00Z to 2012-Feb-07-18Z
Case studies	34 days from 2012-Jan-08-00Z to 2012-Feb-10-18Z	31 days from 2012-Jan-08-00Z to 2012-Feb-07-18Z

Table 3.1: Experimental settings for our Proactive QC study compared with Ota et al. (2013).

model top at 0.3 hPa, for both deterministic and ensemble forecasts.

In our experiments, however, we adopt a lower-resolution version of the GFS model, using T254 ($\sim$ 55 km) and T126 ($\sim$ 110 km) horizontal resolutions, respectively, for the deterministic and ensemble forecasts. The vertical resolution is the same as the operational system (64 layers). Halving the horizontal resolution reduces the I/O size and the necessary storage space by about a quarter, and the computing time by about an eighth, thus enabling much quicker executions and more experiments within the allocated resources.

3.3 The GSI 3D-Var based ensemble-variational hybrid data assimilation system

The DA system of the NCEP's currently operational global NWP system is the Gridpoint Statistical Interpolation (GSI) 3D-Var based ensemble-variational hybrid DA system (Kleist, 2012; Wang et al., 2013). It extends the previously operational GSI 3D-Var DA system (Kleist et al., 2009b; Wu et al., 2002) by introducing flow-dependence to the background error covariance using the background ensemble.

Let us first describe the traditional pure 3D-Var algorithm. 3D-Var solves the analysis equation

$$\delta\mathbf{x}^{ab} := \mathbf{x}_0^a - \mathbf{x}_0^b = \mathbf{K}\delta\mathbf{y}_0^{ob} \tag{3.1}$$

$$\text{with} \quad \mathbf{K} = \mathbf{B}\mathbf{H}^T\left(\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R}\right)^{-1} \tag{3.2}$$

by minimizing the cost function $J(\delta\mathbf{x})$:

$$\begin{aligned}
\delta\mathbf{x}^{ab} &= \arg\min \ J(\delta\mathbf{x}) && (3.3)\\
J(\delta\mathbf{x}) &= \frac{1}{2}\delta\mathbf{x}^T\mathbf{B}^{-1}\delta\mathbf{x} + \frac{1}{2}\left(\mathbf{y}_0^o - H(\mathbf{x}_0^b + \delta\mathbf{x})\right)^T\mathbf{R}^{-1}\left(\mathbf{y}_0^o - H(\mathbf{x}_0^b + \delta\mathbf{x})\right) && (3.4)
\end{aligned}$$

where $\delta\mathbf{x}^{ab}$ is the analysis increment, $\mathbf{x}_0^a$ and $\mathbf{x}_0^b$ are the analysis and the background, respectively, $\mathbf{B}$ is the (static) background error covariance, and the other notation is as defined in Section 2.2.2.1. In practice, the cost function may include the so-called "penalty term" or "constraint term" $J_c(\delta\mathbf{x})$ which imposes some constraints to the analysis such as dynamical balance (e.g., Kleist et al., 2009a; JMA, 2013, Section 2.5.5) or conservation of total mass (John Derber and Daryl Kleist, 2013, personal communication).

The currently operational GSI ensemble-variational hybrid 3D-Var extends the conventional 3D-Var by augmenting the control vector with the ensemble weight vectors $\{\boldsymbol{\alpha}_m \in \mathbb{R}^n, m = 1, \cdots, K\}$ where n is the dimension of the state space and $K = 80$ is the member size of the ensemble. Suppose we have a K-member ensemble perturbation of the background $\mathbf{X}^b = [\mathbf{x}'_1, \mathbf{x}'_2, \cdots, \mathbf{x}'_K]$. The hybrid 3D-Var obtains the analysis increment

$$\delta\mathbf{x}^{ab} = \delta\mathbf{x}^{\text{stat}} + \sum_{m=1}^{K}\left(\boldsymbol{\alpha}'_m \circ \mathbf{x}'_m\right), \tag{3.5}$$

where $\circ$ represents the elementwise multiplication (Schur product), by minimizing the new cost function $J\left(\delta\mathbf{x}, \boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_K\right)$ in the augmented space (Kleist, 2012;

Wang et al., 2013):

$$\begin{aligned}
\left(\delta\mathbf{x}^{\text{stat}}, \boldsymbol{\alpha}'_1, \cdots, \boldsymbol{\alpha}'_K\right) &= \arg \min_{(\delta\mathbf{x}, \boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_K)} J\left(\delta\mathbf{x}, \boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_K\right) && (3.6)\\
J\left(\delta\mathbf{x}, \boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_K\right) &= \beta_{\text{stat}}\, \frac{1}{2}\delta\mathbf{x}^T\mathbf{B}^{-1}\delta\mathbf{x} + \beta_{\text{ens}}\, \frac{1}{2}\sum_{m=1}^{K} \boldsymbol{\alpha}_m^T \mathbf{L}^{-1}\boldsymbol{\alpha}_m \\
&\quad + \frac{1}{2}\left(\mathbf{y}_0^o - H(\mathbf{x})\right)\mathbf{R}^{-1}\left(\mathbf{y}_0^o - H(\mathbf{x})\right) && (3.7)\\
\text{with} \qquad \mathbf{x} &:= \mathbf{x}_0^b + \delta\mathbf{x}^{\text{stat}} + \sum_{m=1}^{K}\left(\boldsymbol{\alpha}_m \circ \mathbf{x}'_m\right) && (3.8)
\end{aligned}$$

where β_{stat} and β_{ens}, whose reciprocals add to unity ($1/\beta_{\text{stat}} + 1/\beta_{\text{ens}} = 1$), are the parameters which determine how much weights are to be given to the static and ensemble covariances, and $\mathbf{L}$ is the localization matrix of the ensemble weights $\boldsymbol{\alpha}_m$.

The operational hybrid GSI uses the ensemble of background generated by the serial EnSRF (Whitaker and Hamill, 2002). For efficiency sake, however, we replace the default serial EnSRF with the LETKF (see next section) implemented in 2011 by Mr. Yoichiro Ota who was then detailed to NCEP from JMA (Ota, 2012, personal communication).

The operational hybrid GSI is a two-way coupled system where the 3D-Var takes in the flow-dependent background covariance from the ensemble, and the analysis ensemble is recentered to the deterministic analysis produced by the 3D-Var. Figure 3.1, adapted from Wang et al. (2013), shows the flowchart of the two-way coupled hybrid GSI. First, the EnKF (LETKF in our system) updates the analysis ensemble using the background ensemble and the observations. Concurrently, the 3D-Var (labeled “GSI-ACV”; ACV standing for “augmented control vector”) generates the control analysis using the control forecast (background) and the background ensemble along with the observations. Then, the analysis ensemble is recentered around the control analysis just generated from the 3D-Var part. Deterministic and ensemble forecast models (the GFS model) integrate the control analysis and the recentered analysis ensemble for 6 hours to produce the control and ensemble background, closing the cycle.

Note that in a hybrid system there are two different analyses; one is the control analysis from the variational part, and the other is the mean of the analysis ensemble (before the recentering) from the ensemble part. They both can be used as the verifying truth $\mathbf{x}_t^v$ when evaluating forecast errors with Eq. (2.6) or Eq. (2.7). In Section 4.4, we examine EFSO's dependence on the choice of the verifying truth.

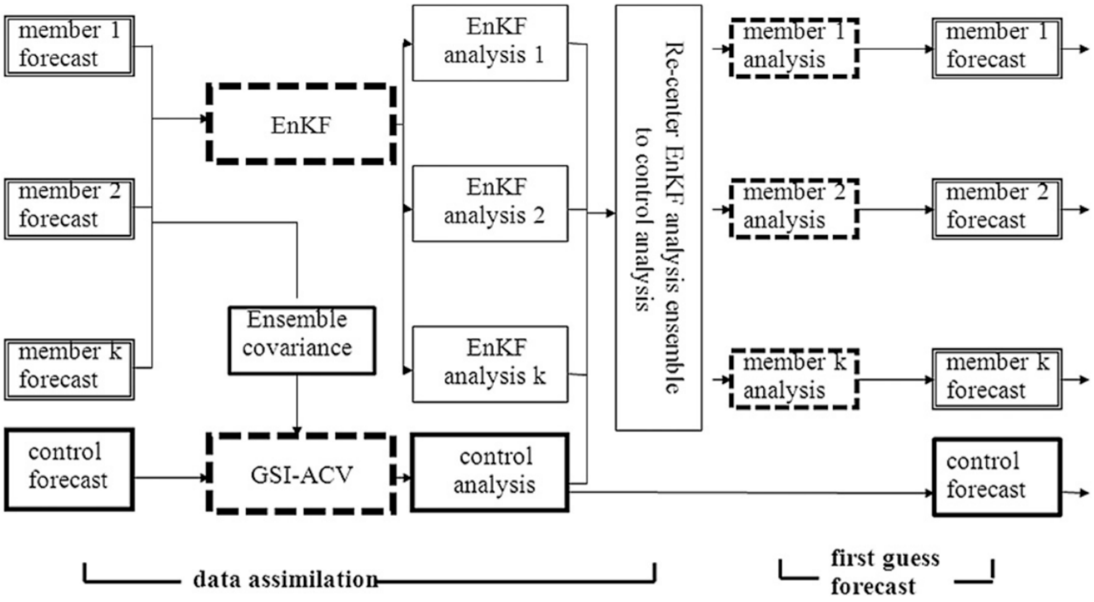


Figure 3.1: Schematical flow chart of the two-way coupled hybrid GSI system. This flow-chart is to be read from left to right. Adapted from Wang et al. (2013).

3.4 LETKF

As we mentioned in the previous section, although the operational hybrid GSI uses the serial EnSRF of Whitaker and Hamill (2002) for ensemble generation, for efficiency sake, we replaced it with the LETKF recently implemented at NCEP by Mr. Yoichiro Ota.

All EnKF methods can be classified into two categories: Perturbed Observation (PO) methods, and Square Root Filters (SRFs). SRF methods have less sampling errors than the PO methods, so most EnKF methods currently in use in geophysical fluid systems adopt the SRF methodology. In SRFs, the Kalman Filter equation for the analysis error covariance:

$$\mathbf{P}^a = (\mathbf{I} - \mathbf{K}\mathbf{H})\,\mathbf{P}^b \tag{3.9}$$

is solved by finding an ensemble transformation matrix (or a weight matrix) $\mathbf{W}$ that updates the background perturbation to the analysis perturbation

$$\mathbf{X}^a = \mathbf{X}^b\mathbf{W}, \tag{3.10}$$

so that it satisfies Eq. (3.9):

$$\frac{1}{K-1}\mathbf{X}^b\mathbf{W}\mathbf{W}^T\mathbf{X}^{bT} = (\mathbf{I} - \mathbf{K}\mathbf{H})\,\frac{1}{K-1}\mathbf{X}^b\mathbf{X}^{bT}. \tag{3.11}$$

Eq. (3.11) does not uniquely determine the weight matrix $\mathbf{W}$, and there are multiple ways to find $\mathbf{W}$ that satisfies it. The variant of ensemble SRF that is adopted by the operational hybrid GSI, the serial EnSRF of Whitaker and Hamill (2002), exploits the fact that, if the number of the assimilated observation is only one, then the weight matrix $\mathbf{W}$ takes a particularly simple form ($(\mathbf{I} - \mathbf{KH})$ scaled by a scalar factor that can be computed from σ^o and $\mathbf{y}^b$ corresponding to the single observation assimilated), and produces the analysis, serially assimilating one observation at a time, regarding analysis from previous assimilation as the background. The simple representation of the weight matrix $\mathbf{W}$ allows simplicity and ease of implementation, but having to process observations serially inevitably imposes limited parallelism, making the algorithm less efficient on massively parallel cluster computers (Miyoshi, 2006).

The LETKF solves the square root filtering problem Eqs. (3.9)–(3.11) differently. By interpreting $\mathbf{X}^b$ as a map that transforms variables from the ensemble space (the K-dimensional space spanned by each column vector of $\mathbf{X}^b$) to the physical space, the matrix

$$\tilde{\mathbf{P}}^a = \frac{1}{K-1}\mathbf{W}\mathbf{W}^T \tag{3.12}$$

can be interpreted as "the analysis error covariance matrix in the ensemble space" because, mapped back to the physical space, it yields the analysis error covariance $\mathbf{P}^a$ in the physical space:

$$\mathbf{P}^a = \mathbf{X}^b\tilde{\mathbf{P}}^a\mathbf{X}^{bT}. \tag{3.13}$$

Once the matrix $\tilde{\mathbf{P}}^a$ is obtained, by choosing the weight matrix $\mathbf{W}$ to be symmetric, the analysis perturbation $\mathbf{X}^a$ in physical space can be computed from Eq. (3.10) by

$$\mathbf{X}^a = \mathbf{X}^b \left\{ (K-1)\tilde{\mathbf{P}}^a \right\}^{1/2}. \tag{3.14}$$

Hunt et al. (2007) showed that, in the ensemble space, the analysis error covariance $\tilde{\mathbf{P}}^a$ can be computed as the following inverse matrix

$$\tilde{\mathbf{P}}^a = \left\{ (K-1)\mathbf{I} + \mathbf{Y}^{bT}\mathbf{R}^{-1}\mathbf{Y}^b \right\}^{-1} \tag{3.15}$$

where $\mathbf{Y}^b \in \mathbb{R}^{p\times K}$ is the background perturbation in the observation space. Since the matrix inverted in the above equation is $K \times K$, and the ensemble member size K is generally small, the computation of $\tilde{\mathbf{P}}^a$ is rather cheap. The LETKF further exploits the fact that information from one grid point to other grid points is spatially confined to its vicinity with some radius of information propagation, and performs the assimilation for each grid point, locally and independently. Because of the independence, the analysis for each grid point can be done completely in parallel, making the algorithm particularly suitable for a parallel computing environment.

In our experiments, we use the LETKF to provide the background ensemble to the GSI. The analysis resolution is half that of the operational system, namely, T126L64, and the member size is 80 (which is the same as in the operational system). As in the operational serial EnSRF, both covariance localization and covariance inflation are applied. The setup for localization and inflation is identical to the operational serial EnSRF with T254L64 resolution: the covariance is localized with the fifth-order polynomial localization function of Gaspari and Cohn (1999), both in the horizontal and the vertical, with the cut-off length of 2,000 km and 2.0 scale heights, respectively. These correspond, respectively, to the e-folding scale of 800 km and 0.8 scale height. For covariance inflation, both multiplicative and additive inflation are applied. The multiplicative inflation uses an adaptive algorithm of Zhang et al. (2004) and Whitaker and Hamill (2012) which inflates the posterior (analysis) covariance so that it is relaxed to the prior (background) variance multiplied by some fixed scaling factor, which, in the operational system and in our experiment, is set to 0.85. The posterior (analysis) covariance is further inflated with an additive inflation procedure that is similar in concept to the so-called National Meteorological Center (NMC) method of Parrish and Derber (1992): for each member, randomly pick a difference of 48-hour and 24-hour forecast which validate at the same date and time from a pre-computed inventory of NCEP's operational forecast, multiply it by a tuning factor, which we choose to be 0.32, and add it to the posterior (analysis) ensemble. It should be noted that, since the localization and inflation parameters we use in the lower-resolution T126 LETKF are optimized for use with T254 serial EnSRF, these choices may not be optimal. Nevertheless, the system worked without any problem during our experimentation.

3.5 Observations

The observations assimilated in our experiments are identical to those in the operational system. The observations are classified into two groups: non-radiance data and radiance data. In the NCEP's GSI system, non-radiance data are referred to as "Conventional" data. Radiance data are simply referred to by "Satellite" data. Note that, with this convention, satellite non-radiance observations, such as Atmospheric Motion Vector (AMV), surface wind estimates from scatterometers, or Global Positioning System Radio-Occultation (GPS-RO), are classified as "Conventional."

Each conventional observation is stored in the NCEP's PREPBUFR file format and is given an integer index called "Report type" that identifies the type of observation. The "report type" is referred to by "stattype" within the GSI code, so hereafter, we call them as "stattype." The list of "stattypes" operationally assimilated at NCEP can be found at their website (NCEP, 2013), which is reproduced in Tables 3.2 and 3.3.

In taking statistics, Ota et al. (2013) combined observations of several different stattypes into one group, as shown in the second column of Tables 3.2 and 3.3. As we will describe in Chapter 6, however, for Proactive QC, we will instead use raw stattypes because observations of different stattypes tend to have different error characteristics.

A list of satellite radiance data assimilated in our experiment is shown Table 3.4. In the later chapters, satellite data are referred to by the names given in this table.

3.6 EFSO setup

The EFSO impacts are computed for each observation using Eq. (2.34), for three different evaluation lead times t: 6 hours, 12 hours, and 24 hours. We also computed 0-hour EFSOs (*i.e.*analysis sensitivity to observations) but they are just for discussion's sake and we do not extensively describe their results. For covariance localization, we adopt the moving localization scheme of Ota et al. (2013). As the error norm, we adopt dry and moist total energy of Ehrendorfer et al. (1999): Let S be a domain on the globe (it can be the globe itself), and consider measuring the magnitude of a perturbation $\mathbf{x}'$ of the atmospheric state. The kinetic energy (KE), potential energy (PE) and moist energy (ME) of $\mathbf{x}'$ are defined, respectively, by:

$$\begin{aligned} e_{KE}^2 &= \frac{1}{2}\frac{1}{|S|}\int_S\int_0^1 \left(u'^2 + v'^2\right) \mathrm{d}\sigma\mathrm{d}S & (3.16) \\ e_{PE}^2 &= \frac{1}{2}\frac{1}{|S|}\int_S\left\{\left(\int_0^1 \frac{C_p}{T_r}T'^2\mathrm{d}\sigma\right) + \frac{R_dT_r}{P_r^2}P_s'^2\right\}\mathrm{d}S & (3.17) \\ e_{ME}^2 &= \frac{1}{2}\frac{1}{|S|}\int_S\int_0^1 \frac{L^2}{C_pT_r}q'^2\mathrm{d}\sigma\mathrm{d}S & (3.18) \end{aligned}$$

where $|S|$ is the area of the domain S, $\sigma = p/P_s$ is the vertical σ coordinate, u',v',T',q' and P_s' are, respectively, the zonal wind, meridional wind, temperature, specific humidity and surface pressure of the perturbation $\mathbf{x}'$, C_p is the specific isobaric heat capacity of the air, R_d is the gas constant of dry air, L is the latent heat of condensation per unit mass, and T_r and P_r are the reference temperature

Stattype	Type name used in Ota et al. (2013)	Description	Element
130	Aircraft	AIREP and PIREP	T_s
230	Aircraft	AIREP and PIREP	u, v
131	Aircraft	AMDAR	T_s
231	Aircraft	AMDAR	u, v
133	Aircraft	ACARS	T_s, q
233	Aircraft	ACARS	u, v
135	Aircraft	Canadian AMDAR	T_s
120	Radiosonde	Radiosonde	T_v, q, P_s
220	Radiosonde	Radiosonde	u, v
242	Satellite_Wind	AMV (cloud drift; below 850 hPa) from JMA-MTSAT	u, v
252	Satellite_Wind	AMV (cloud drift; above 850 hPa) from JMA-MTSAT	u, v
243	Satellite_Wind	AMV (cloud drift; below 850 hPa) from EUMETSAT	u, v
253	Satellite_Wind	AMV (cloud drift; above 850 hPa) from EUMETSAT	u, v
245	Satellite_Wind	AMV (cloud drift; all levels) from NESDIS-GOES	u, v
246	Satellite_Wind	AMV (WV cloud top; all levels) from NESDIS-GOES	u, v
250	Satellite_Wind	AMV (WV cloud top; all levels) from JMA-MTSAT	u, v
4,42 722 740 ∼746	GPSRO	GPS Radio Occultation	–
181	Land-Surface	SYNOP and METAR	P_s
281	Land-Surface	SYNOP and METAR	u, v
187	Land-Surface	METAR	P_s (inferred from altimeter setting)
287	Land-Surface	METAR	u, v

Table 3.2: List of assimilated non-radiance observation types (continued to Table 3.3). u, v, T_s, T_v, q and P_s represent, respectively, zonal wind, meridional wind, temperature, virtual temperature, specific humidity, and surface pressure.

Stattype	Type name used in Ota et al. (2013)	Description	Element
180	Marine-Surface	SHIP, BUOY, C-MAN and TIDE GAUGE	T_v, q, P_s
280	Marine-Surface	SHIP, BUOY, C-MAN and TIDE GAUGE	u, v
257	MODIS_Wind	MODIS/POES AMV (IR cloud drift; all levels) from AQUA and TERRA	u, v
258	MODIS_Wind	MODIS/POES AMV (WV cloud top ; above 600 hPa) from AQUA and TERRA	u, v
259	MODIS_Wind	MODIS/POES AMV (WV deep layer ; above 600 hPa) from AQUA and TERRA	u, v
290	ASCAT_Wind	ASCAT scatterometer surface wind over the ocean	u, v
221	PIBAL	Pilot Balloons	u, v
224	NEXRAD_Wind	wind from NEXRAD Radar	u, v
223	Profiler_Wind	NOAA Profiler Network (NPN) wind profiler	u, v
229	Profiler_Wind	Wind Profiler from PIBAL bulletin	u, v
132	Dropsonde	Flight-level reconnaissance and profile dropsonde	T_v, q
182	Dropsonde	Splash-level dropsonde over ocean	T_v, q, P_s
289	WINDSAT_Wind	ASCAT scatterometer over ocean (super observation)	u, v
112	TCVital	Pseudo surface pressure observations at tropical cyclone storm center	P_s
700 ∼721	Ozone	Ozone retrievals from satellite radiances	ozone

Table 3.3: List of assimilated non-radiance observation types (continued from Table 3.2). Unlike Ota et al. (2013), different stattypes are not consolidated.

Type (sensor name)	Description
AMSUA	Satellite microwave sounder radiances (from five satellites)
IASI	Satellite infrared hyperspectral sounder radiances
Aqua_AIRS	Satellite infrared hyperspectral sounder radiances
ATMS	Satellite microwave sounder radiances (from Suomi-NPP)
HIRS	Satellite infrared radiances (from two satellites)
MHS	Satellite microwave sounder radiances (from three satellites)
GOES	GOES infrared sounder radiances (GOES13 and 15)
SEVIRI	SEVIRI clear sky radiances

Table 3.4: List of assimilated satellite radiance data. Adapted from Ota et al. (2013).

and surface pressure. In our computation, we choose $T_r = 280$ K and $P_r = 1000$ hPa. As in Ota et al. (2013), the vertical integration is carried out from the surface up to the model top. The KE, PE and ME all have the dimension of J kg^{-1}. The dry total energy e^2_{DTE} and the moist total energy e^2_{MTE} are defined using the above as

$$e^2_{DTE} = e^2_{KE} + e^2_{PE} \tag{3.19}$$

$$e^2_{MTE} = e^2_{KE} + e^2_{PE} + e^2_{ME} \tag{3.20}$$

The matrix $\mathbf{C}$ in Eq. (2.8) and Eq. (2.9) is defined so that the quadratic form $\mathbf{x}'^T\mathbf{C}\mathbf{x}'$ evaluates to e^2_{DTE} or e^2_{MTE} defined above. Note that the use of EnKF allows a simple computation of the moist static energy, which is extremely difficult with an adjoint system.

3.7 Period of the experiments

The DA cycles are performed for 41 days from 00 UTC of January 1st, 2012 until 18 UTC of February 10th, 2012. The first cycle was started by assimilating the observations of 00 UTC of January 1st, 2012 to the control and ensemble background taken from the NCEP's operational system. In preparing the initial data, resolution transformation from the operational T574 (control) and T254 (ensemble) to T254 (control) and T126 (ensemble) was applied using the NCEP's `global_chgres` utility installed on the S4 supercomputer.

The first seven days from 00 UTC of January 1st, 2012 to 18 UTC of January 7th, 2012 are discarded from the verification and case studies. For comparison of statistical verification with Ota et al. (2013), we choose the period identical to that in Ota et al. (2013): 31 days from 00 UTC of January 8th, 2012 to 18 UTC of February 7th, 2012, with $31 \times 4 = 124$ samples in total. For case studies, we examine 34 days from 00 UTC of January 8th, 2012 to 18 UTC of February 10th, 2012, with $34 \times 4 = 136$ samples in total.

Chapter 4: EFSO's dependance on verifying truth and evaluation lead time

4.1 Introduction

As we discussed in Section 1.4.2 and Section 2.4, we first need to examine whether EFSO is applicable to the EnKF within a hybrid system, and how sensitive it is to the choice of the evaluation lead time and the verifying truth. These issues are addressed in this chapter. First, to check if EFSO also works on a hybrid DA system, we briefly examine the consistency of our results with previous studies, in particular Ota et al. (2013), who applied the same EFSO formulation to the same GFS model (but with different horizontal resolution) in a pure EnKF DA system, and Holdaway et al. (2014), who developed linear moist physics to the adjoint model of NASA/GMAO's Goddard Earth Observing System (GEOS-5) and applied the adjoint FSO to their GSI-based 4D-Var system. We then examine EFSO's sensitivity to the evaluation lead time by comparing some of their statistical features and then by comparing the EFSO impact of each observation for individual cases. Next we conduct similar assessment on EFSO's sensitivity to the choice of verifying truth (either the control analysis from the variational part or the ensemble mean analysis from the EnKF part). Our conclusion is that, somewhat to our surprise, EFSO results do not depend too much on these choices.

4.2 Comparison with previous FSO studies

Since our work is the first to apply EFSO to a hybrid DA system, it is important to make sure that EFSO also works with a hybrid system. In this section, we compare our results with previous FSO studies. We first compare our EFSO results with that of Ota et al. (2013). As we stated in Section 2.4, in our LETKF/3D-Var hybrid GSI, we have two choices for the verifying truth. Here, we show the EFSO impacts obtained by using the control analysis from GSI for verification. As we will see in Section 4.4, the EFSO results do not depend significantly on which verifying truth to use, so this choice does not affect our conclusion.

Panels (a) and (c) of Figure 4.1 show 24-hour EFSO impacts from each observation type defined in Tables 3.2, 3.3 and 3.4 measured with the (a) moist and (b) dry total energy norm, respectively, in units of J kg^{-1}. The impacts are evaluated for the whole globe and are averaged over the one-month period from 00 UTC of January 8th, 2012 to 18 UTC of February 7th (see Section 3.7), using all of the four initial times (00 UTC, 06 UTC, 12UTC and 18 UTC). For convenience, the corresponding figures from Ota et al. (2013) are reprinted in panels (b) and (d). In interpreting these figures, note that negative values of EFSO mean that the observations act to reduce the forecast errors and thus have positive impact. Despite the different configurations such as the horizontal resolutions and the DA methods used, our results are mostly consistent with those of Ota et al. (2013). In particular, the notable features pointed out by Ota et al. (2013), namely:

- AMSU-A (Advanced Microwave Sounding Unit A) contributes most positively, both in dry and moist norm, followed by Aircraft, Radiosonde and IASI (Infrared Atmospheric Sounding Interferometer).
- Ozone contributes slightly negatively with these error norms.
- All satellite radiance observations, especially the MHS (Microwave Humidity Sounder) and PIBAL (Pilot Ballon) exhibit reduced impacts with the dry total energy norm than with the moist total energy norm.
- GPSRO (Global Positioning System Radio Occultation) has almost equal impact on moist and dry error norms, indicating that it has minimal impact on moisture fields.

are all valid with our results as well. Two noticeable differences are that: (1) Satellite wind has less impact in our experiment than GPSRO, whereas, in Ota et al. (2013), the impacts from the two observation types are comparable, and (2) MODIS (Moderate-Resolution Imaging Spectroradiometer) wind has neutral impact in our experiment, whereas, in Ota et al. (2013), it has clear positive impact. It is difficult to answer what causes these discrepancies (and that is not in the scope of our study) but we speculate that suboptimality of thinning of these AMV (Atmospheric Motion Vector; Satellite wind and MODIS wind are both AMV) data in our experiment might be a possible factor: because AMV data are so densely observed, sometime exceeding the resolution of the models, the error correlations between nearby observations, which is neglected in most formulations of DA, can be non-negligible. Thinning of observations are thus necessary, and it has been shown by theoretical consideration that the thinning should be made stronger for a lower (coarser) model resolution (e.g. Liu and Rabier, 2002). Although the horizontal resolution of the ensemble DA is lower in our experiments (T126) than in Ota et al. (2013; T254), the two systems apply the same thinning to the AMV observations before they are ingested to the DA system. Since the thinning is optimized for the higher, operational resolution (T574 for control; T254 for ensemble), this could make the assimilation of AMV data less efficient in our lower-resolution system. Apart from these minor differences, our results are very consistent with Ota et al. (2013), which supports the validity of using EFSO with an EnKF within a hybrid DA system.

Next, we compare our results with an adjoint-based FSO, taking a recent result from Holdaway et al. (2014) as an example. Holdaway et al. (2014) extended NASA/GMAO's global NWP system by including moist physical processes to the tangent linear and adjoint model of the GEOS-5 model and showed that the inclusion of moist physics in the adjoint model improves the ability of the adjoint FSO to better reproduce the nonlinear forecast error reduction. Panels (a) and (b) of Figure 4.2, adapted from Holdaway et al. (2014), show the FSO impacts from each observation type evaluated with the adjoint formulation of LB04 with NASA/GMAO's GEOS-5 system, averaged over a one-month period from March 17th, 2012 to April 17th, 2012, using only analyses of 00 UTC. Different colors represent different con-

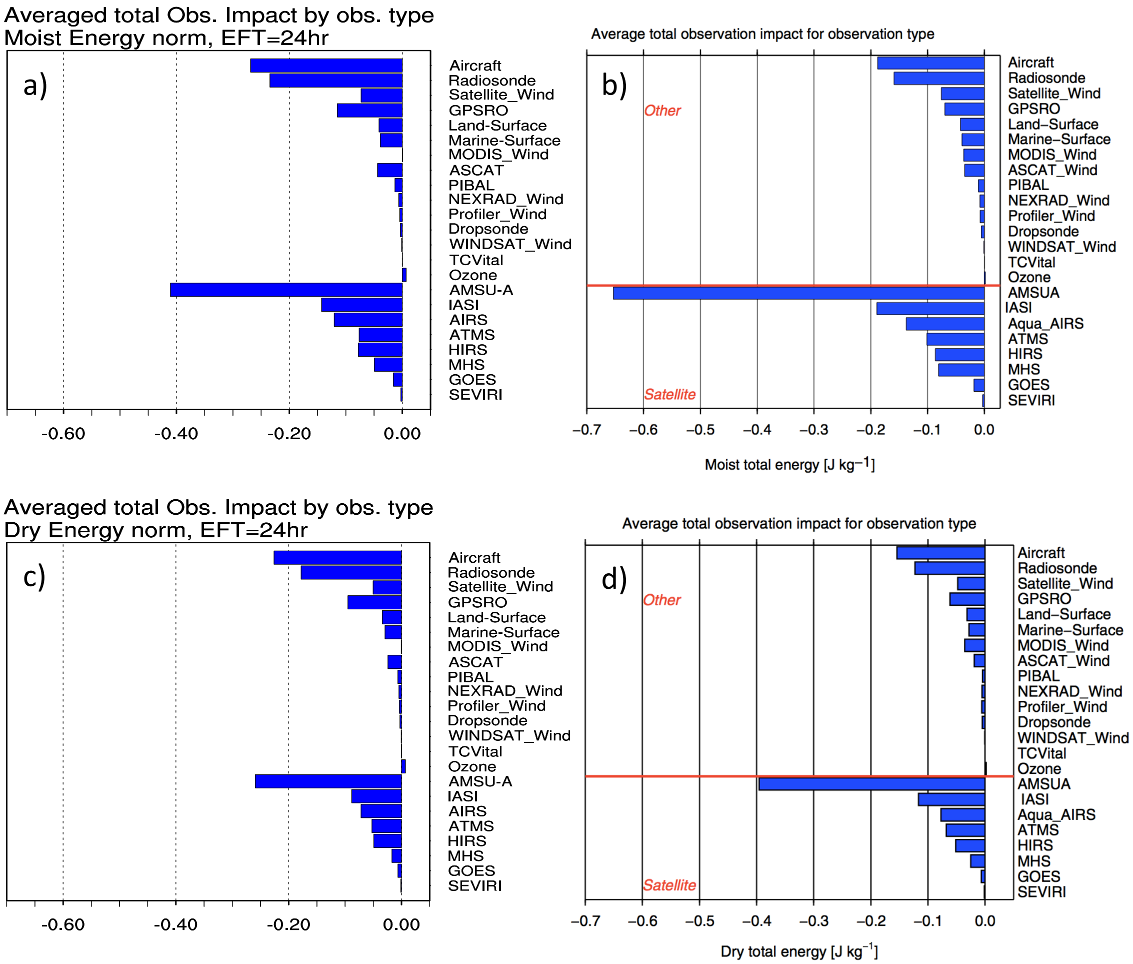


Figure 4.1: Comparison of our 24-hour EFSO impacts from each observation type with those of Ota et al. (2013). (a) 24-hour EFSO impacts (J kg^{-1}) from our experiment verified against the control analysis from GSI measured with the moist total energy norm averaged over a one-month period from 2012-Jan-08-00Z to 2012-Feb-07-18Z. (b) 24-hour EFSO impacts measured with the moist total energy norm from Ota et al. (2013) averaged over the same one-month period. (c) as in (a), but with the dry total energy norm. (d) as in (b), but with the dry total energy norm. Panels (b) and (d) are adapted from Ota et al. (2013).

figurations for the adjoint model and the error norm: the red bars in (a) and (b) show the impacts evaluated using the adjoint model with only dry physics measured with dry total energy norm (note that the red bars in (a) and (b) are identical, although they may look different due to the difference in the scales), blue bars in (a) represent the impacts evaluated using the adjoint model with dry and moist physics measured with dry total energy norm, and the green bars in (b) represent the impacts evaluated using the adjoint model with dry and moist physics measured with moist total energy norm. Note that, in Holdaway et al. (2014), the moist total energy norm e^2_{MTE} is defined as $e^2_{DTE} + 0.3e^2_{ME}$ rather than $e^2_{DTE} + e^2_{ME}$ as in our case, and the vertical integration is from the surface to $\sim$ 125 hPa, whereas, in our case, it is from the surface to the model top. To contrast their results with ours, we show, in panels (c) and (d), our 24-hour EFSO impacts measured with the dry and moist total energy norm, respectively. Unlike in panels (a) and (c) of Figure 4.1 where we used all samples from 00, 06, 12 and 18 UTC, here we averaged the impacts using only the samples from 00 UTC, to make them compatible with Holdaway et al. (2014). Since most radiosonde stations make observations only on 00 and 12 UTC (and not on 06 and 18 UTC), we can expect that taking samples only from 00 UTC makes the impact from radiosonde more significant, which is true in our case (compare panels (a) and (c) of Figure 4.1, respectively, with panels (d) and (c) of Figure 4.2). Our results and those of Holdaway et al. (2014) agree in that, regardless of the error norm used, AMSU-A and Radiosonde are the top two positively contributing types, followed by Aircraft and then by IASI and AIRS. If we focus on radiance data, both our results and those of Holdaway et al. (2014), and also Ota et al. (2013), all estimate that the impacts are ordered, from largest to smallest, as AMSU-A, IASI, AIRS, HIRS and MHS (note that the impacts from GOES and SEVIRI are not shown in Holdaway et al. (2014), perhaps because they are negligibly small).

Overall, our results are consistent with the previous FSO studies, which suggests that the EFSO is also applicable to the EnKF within a hybrid DA system. We next examine how sensitive the EFSO is to the choice of evaluation lead time.

4.3 Dependence on the evaluation lead time

Our proposed algorithm of Proactive QC relies on the ability of 6-hour EFSO to detect "flawed" observations. It is thus important to understand the characteristics of 6-hour EFSO. We investigate this issue in this section by comparing the EFSOs evaluated for different evaluation lead times.

4.3.1 Statistical properties

The EFSO impacts for each observation type evaluated for 6-, 12-, and 24-hour lead times, measured with the moist and dry total energy norms are summarized in Figure 4.3. In this figure, the control analysis from GSI is used as the verifying truth. As we stated in Section 1.4.2 and Section 2.4.1, we were not confident, before conducting the experiments, whether 6-hour EFSO works or not because the analysis errors against the truth, which is assumed sufficiently small in our EFSO formulation, might not be negligible when compared to 6-hour forecast errors. Despite our

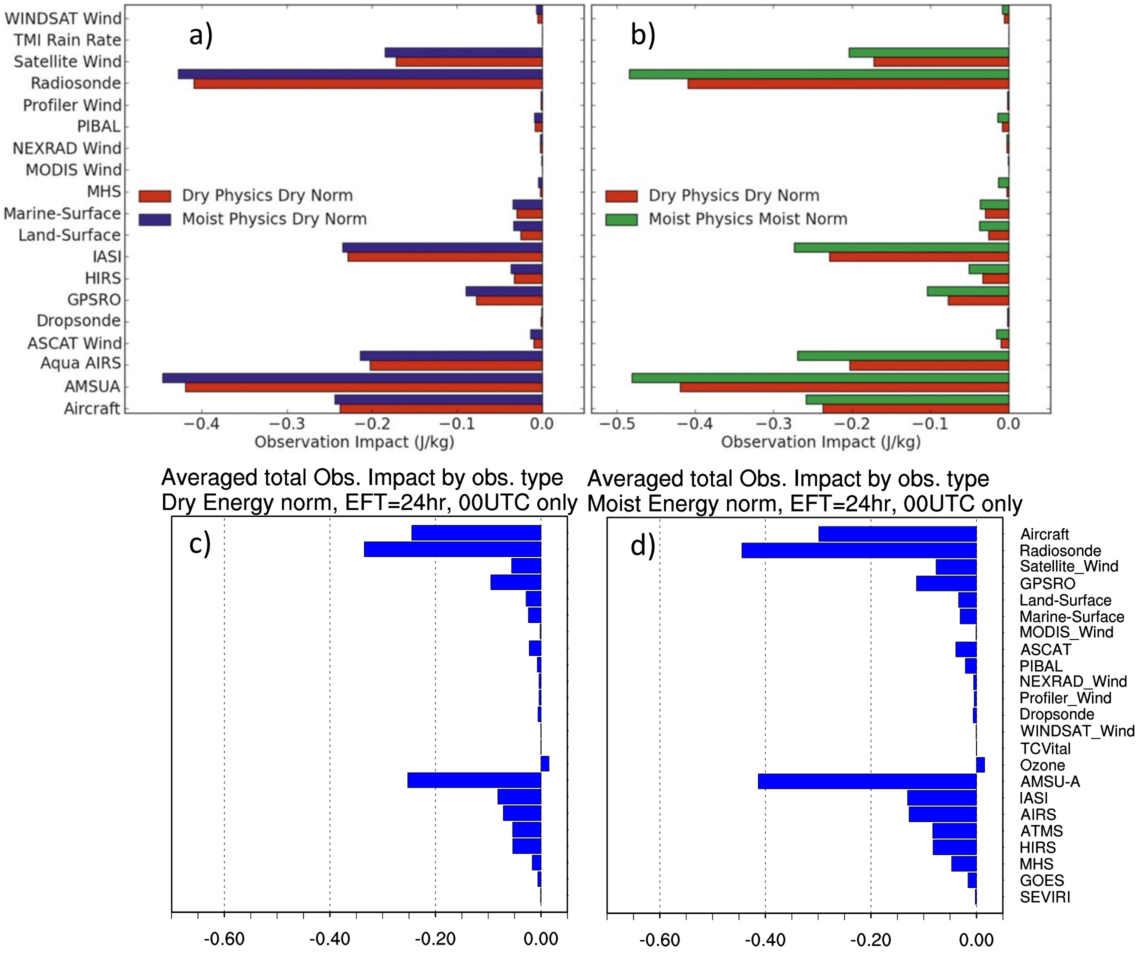


Figure 4.2: Comparison of the observational impacts to 24-hour forecast errors (J kg^{-1}) from our EFSO experiments with those from adjoint FSO of NASA/GMAO's GEOS-5 system. (a) 24-hour FSO impacts from each observation type evaluated with the adjoint formulation of LB04 with NASA/GMAO's GEOS-5 system. Shown are the impacts evaluated using the adjoint model with (red) only dry physics and (blue) dry and moist physics. Both impacts are measured with the dry total energy norm, and are averaged over a one-month period from March 17th, 2012 to April 17th, 2012, using only analyses of 00 UTC. (b) as in (a), but (green) the impacts measured with the moist total energy norm using the adjoint model with dry and moist physics, and (red) the impacts measured with the dry total energy norm using the adjoint model with only dry physics (identical to the red bar in the panel (a)). (c) As in Figure 4.1c, but averaged using only the analyses of 00 UTC. (d) As in Figure 4.1a, but averaged using only the analyses of 00 UTC. Panels (a) and (b) are adapted from Holdaway et al. (2014).

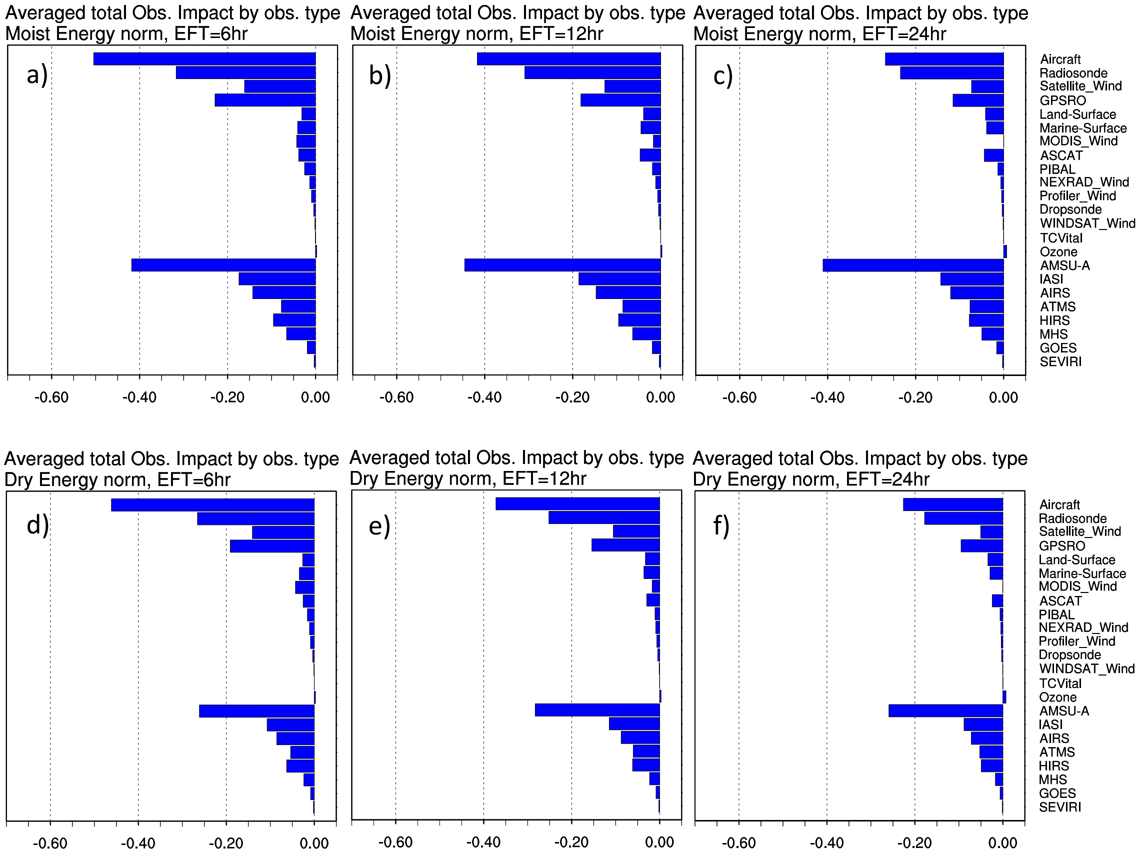


Figure 4.3: Comparison of EFSO impacts for different evaluation lead times. Top and bottom panels are, respectively, for the moist and dry total energy norm. Left, middle and right panels are, respectively, for 6-, 12- and 24-hour lead time. In all panels, the EFSO impacts are verified against the control GSI analysis. (a) as in Figure 4.1a, but for 6-hour lead time. (b) as in (a), but for 12-hour lead time. (c) as in (a), but for 24-hour lead time. (d) as in (a), but measured with the dry total energy norm. (e) as in (b), but measured with the dry total energy norm. (f) as in (c), but measured with the dry total energy norm.

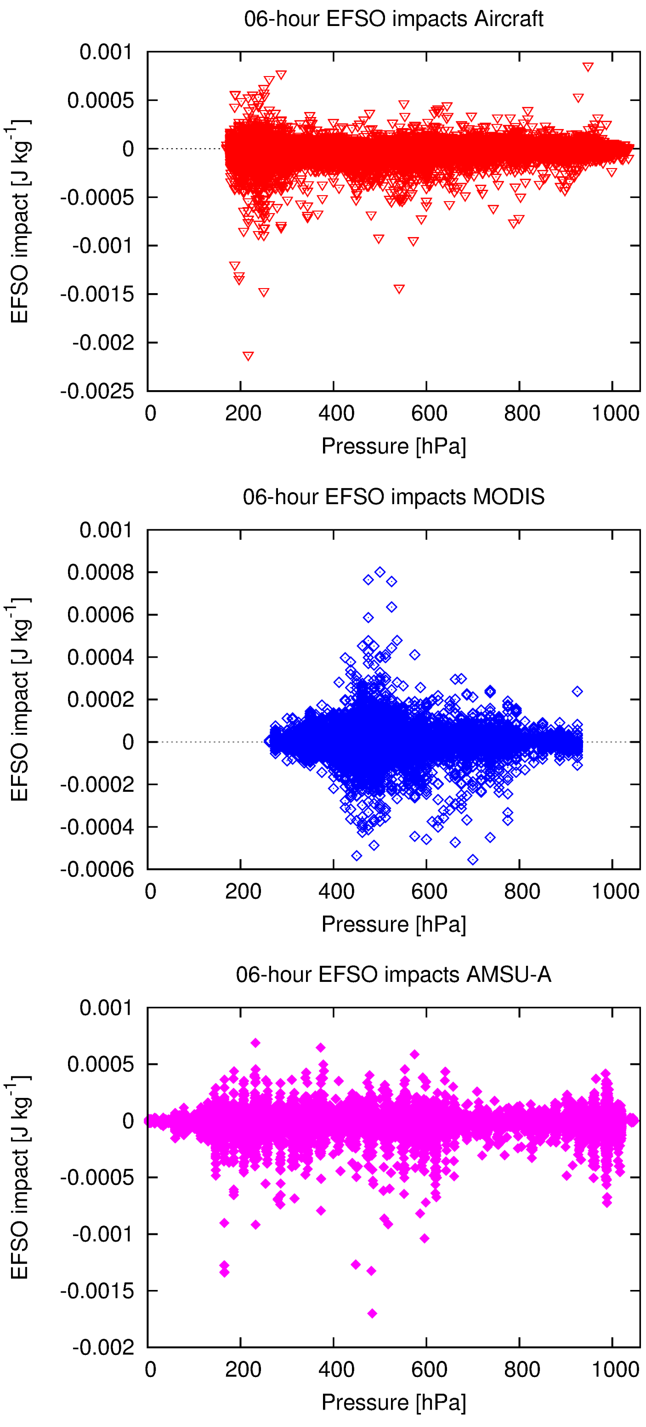


Figure 4.4: 6-hour EFSO impacts (J kg^{-1}) of each observation plotted against the height (in pressure; hPa). The impacts are evaluated for the entire globe with 6-hour lead time and the moist total energy norm, and are verified against the control GSI analysis. Shown are the observations of (a) Aircraft, (b) MODIS wind and (c) AMSU-A, assimilated at 18 UTC of February 1st, 2012. For AMSU-A, the pressure is defined for each channel as the maxima of the corresponding weighting function.

concerns, however, Figure 4.3 shows that the EFSO impacts with different forecast lead times are in fact highly consistent. In particular, relative importance of the satellite radiance observations (the bottommost 8 bars) does not differ much for different lead times. This is also true for the conventional observations except MODIS wind. Two features that draw our particular attention are: (1) the estimated impacts of all observation types decreases as the lead time increases, when, in fact, the forecast error grows, and (2) the decrease of the impacts with the lead time is modest for satellite radiances and surface observations (Land-Surface, Marine-Surface and ASCAT) but is large for other observation types, in particular, Aircraft and MODIS Wind. We speculate that both features can be explained by the limitation resulting from covariance localization (see Section 2.2.3): as the forecast time increases, the information from an observation is advected away from the location where it was observed, but the localization applied in EFSO fails to track it, resulting in diminished impact estimation. This is possibly a reason why the total estimated reduction of forecast errors (the sum of EFSO impacts from all observations) decreases as lead time gets longer.[1] The loss of impacts due to imperfect localization is stronger in the upper troposphere where the westerly jet prevails, and this explains why the impact from Aircraft observation, for example, weakens more quickly than surface observations as the lead time gets longer: the observations from aircraft have larger

[1] Another possible mechanism is the error saturation for a chaotic dynamical system (e.g., Dalcher and Kalnay, 1987): in a chaotic system, nonlinear error growth saturates at certain forecast lead time, due to the system's limited memory of initial states. At this saturation lead time, impacts from any observation becomes zero because, no matter how accurate the initial condition is, the forecast error reaches the climatological level. In the atmosphere, the dominant modes do not saturate until about 10–14 days, but for modes with small spatial scales, the saturation is much faster, as shown by Holdaway et al. (2014). Note that this feature cannot be captured by the adjoint-based FSO evaluation.

impacts in data-sparse (less populated) areas away from airports. Since aircraft fly over those areas at the cruising height just above the tropopause where the westerlies are strong, their impacts are prone to dilution due to covariance localization; the observations at surface level or lower troposphere, on the other hand, do not suffer much from such limitation.

Figure 4.4 supports our reasoning above. Panel (a) shows the height and the 6-hour EFSO impact of each aircraft observation assimilated at 18 UTC of February 1st, 2012. Aircraft observations are distributed from ground level up to the lower stratosphere ($\sim$ 170 hPa), but the observations are densest near the tropopause level (from about 350 hPa to 170 hPa). Observations with large EFSO impacts are also concentrated in this height range. MODIS wind observations (panel (b)) are densest in the middle to upper troposphere (between 600 hPa to 400 hPa). A peak of observations with large impacts is found at around $\sim$ 450 hPa level. Both Aircraft and MODIS wind have only limited fraction of observations near the ground where the wind is weak, making it particularly prone to the weakening of signal due to localization. On the other hand, AMSU-A (panel (c)) has a lot of channels that are sensitive to the lower troposphere (below 800 hPa), and the observations for such channels have large EFSO impacts, albeit not as large as those in the upper troposphere, making it rather robust against the weakening of signals by localization.

In the literature of FSO studies, there have been several discussions about how many percentage of the observations have beneficial impact onto the forecast, in particular, why so few does: all works on FSO published hitherto report that only slightly more than 50% of observations have positive FSO impact onto the forecast (e.g., Gelaro et al., 2010; Langland and Baker, 2004; Lorenc and Marriott, 2013; Todling, 2013). We should keep it in mind, however, that all these previous studies are based on FSO results with 24-hour lead time. According to Gelaro et al. (2010), a Monte-Carlo experiment for a scalar system conducted by Ehrendorfer (2007, see their Figure 2) and Mike Fisher (2006, personal communication), in which the background and observation error variance is perfectly specified and the background and the observation have comparable accuracy, suggests that 60%–65% of the observations should be beneficial to the analysis when the DA system is optimal. Citing this result, Gelaro et al. (2010) argued that the suboptimality of operational systems (*i.e.*, imperfect specification of background or observation error covariances) may be restricting the utility of the information from observations. Daescu and Todling (2009) argued that the imperfection of verifying truth makes the FSO estimation less reliable. Todling (2013), on the other hand, showed that the 0-hour FSO (*i.e.*, analysis sensitivity to observation) verified against the observations rather than the analysis and measured with the error norm evaluated in the observation space exhibits the fraction of beneficial observations that is close to the theoretical range (60%–65%) and argued that their operational system should be in fact nearly optimal. What he did not make clear, however, is why the results are so different between the 24-hour forecast sensitivity and the analysis (0-hour forecast) sensitivity. Lorenc and Marriott (2013), which was published almost simultaneously, proposed a mechanism that possibly answers this problem: they conducted a series of idealized Monte-Carlo experiments similar to that of Ehrendorfer (2007) but with a linear model with 8 different independent modes in which the growth rate of each mode, the background and observation error variances **B** and **R**, and the error of verifying state, are all allowed to vary, and suggested that, not only the suboptimality of DA system and the limited accuracy of verifying truth, as suggested by the previous studies, but also the difference of the growth rates of each mode of the forecast model, along with the lack of flow-dependence of **B**-matrix in the DA system, all contribute to the lowered fraction of beneficial observations. In view of these discussions, it would be interesting to look at how the percentage of beneficial observations changes with the evaluation lead time in our system.

Figure 4.5 shows the percentage of observations that are estimated to be beneficial with EFSO in our system, for the lead times of (a) 0 hour (*i.e.*, observational impact to *analysis*), (b) 6 hours, (c) 12 hours and (d) 24 hours. Consistent with the previous studies, the percentage of beneficial observations for 24-hour forecast (panel (d)) is only slightly above 50% for all observation types (except TCVital, whose statistics is not reliable due to the limited sample size of only 77). As the lead time decreases, however, more and more observations are estimated to be beneficial. The percentages of beneficial observations, all observation types combined, are 56%, 53%, 52% and 51%, respectively, for 0, 6, 12 and 24 hours. Following the argument of Lorenc and Marriott (2013), we can interpret this as follows:

> The atmosphere as a dynamical system has both growing and decaying modes. Suppose that an observation improves the analysis by significantly improving the decaying modes, but, at the same time, it slightly degrades the growing modes. For a forecast of a short period, the ob-

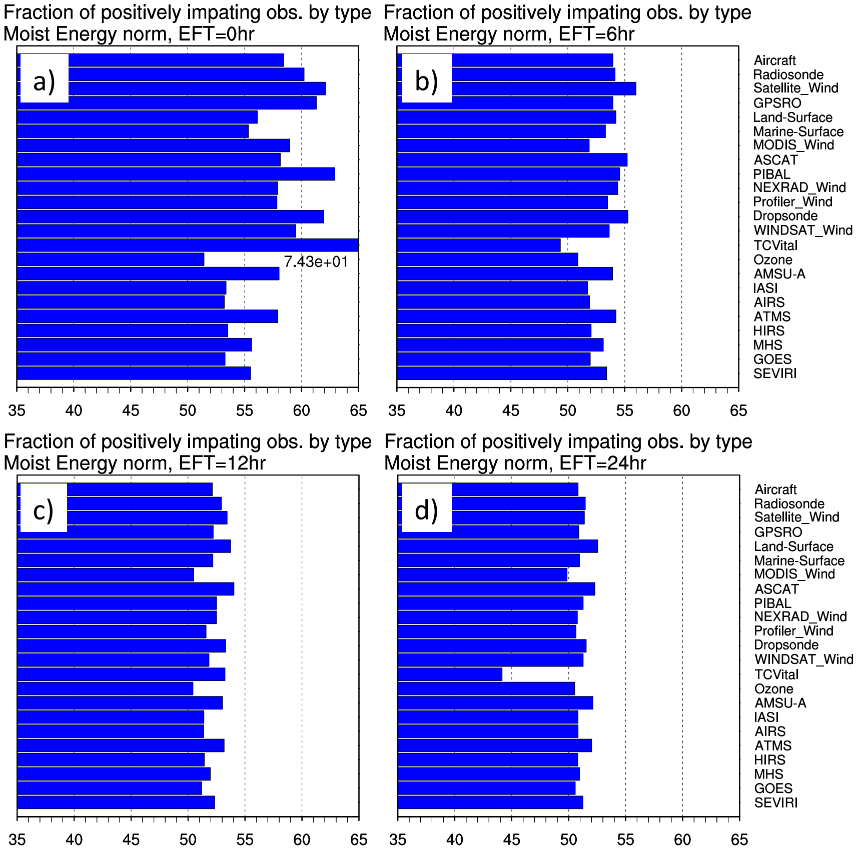


Figure 4.5: Percentage of beneficial observations (*i.e.*, the number of observations with positive impacts onto forecast divided by the total number of observations and then multiplied by 100) classified by the observation types. EFSO impacts are computed using the moist total energy norm with the control GSI analysis as the verifying truth. Shown are evaluated with lead time (a) 0 hours (*i.e.*, analysis sensitivity to observation), (b) 6 hours, (c) 12 hours, and (d) 24 hours. Statistics are taken for a one-month period from 2012-Jan-08-00Z to 2012-Feb-07-18Z, with total observation number of 218,025,941.

> servation could maintain its beneficial impact; eventually, however, the initial slight increase of the error in the growing modes will amplify and overwhelm the reduction of the error in the decaying modes, rendering the net impact of that observation negative.

Trevisan et al. (2010) showed that the analysis increment should be confined in the unstable subspace for this reason, and demonstrated with the Lorenz '96 system (c.f., Section 7.4.1) that the 4D-Var in which the analysis increment is sought within the unstable subspace performs better than the conventional 4D-Var.

4.3.2 Individual cases

In the previous subsection, we saw that the average EFSO impacts from each observation type are rather insensitive to the evaluation lead time. In this subsection, we examine if EFSO impacts for individual observations are also insensitive to the choice of the lead time by focusing on some of the cases we identify as "possible regional dropouts" (see Section 6.2).

As an example, we show, in Figure 4.6, the geographical and vertical distribution of EFSO impacts for the MODIS wind derived from the water vapor image by tracking movement of deep layer cloud (stattype 259; see Table 3.3), for one of the identified "regional dropout" cases. The initial date is 18 UTC of February 6th, 2012, and the area is the "rectangular" (in fact, "triangular") domain of [60 °N–90 °N]×[40 °E – 100 °E]. The EFSO impacts are measured with the moist total energy norm targeted in the region above and are verified against the control GSI analysis. Red and blue circles represent negative and positive impacts, respectively, and the size of each circle is proportional to the size of the impact. The left panels ((a) and (c)) show the results for 6-hour EFSO; the right panels ((b) and (d)) show the results for 24-hour EFSO.

In Proactive QC, we will exploit EFSO's ability to detect "flawed" observations, namely, the observations with large negative impacts. We are thus particularly interested in the consistency of positions and sizes of large red circles between the two lead times (left and right panels). By comparing the left panels ((a) and (c)) with the right panels ((b) and (d)) of Figure 4.6, we can observe that, the observations with large negative 24-hour EFSO impacts (large red points) are well collocated with those with large negative 6-hour EFSO impacts, both horizontally and vertically, and vice versa. This visual impression is supported by panel (b) of Figure 4.7 which shows the scatter plot of 6-hour and 24-hour EFSO for the case shown in Figure 4.6: the correlation between 6-hour and 24-hour EFSO is not very strong for observations with modest impacts (*i.e.*, dots near the origin), with some of the observations with positive 6-hour EFSO having negative 24-hour EFSO, but for the observations with large impacts which are of our particular interest, at least their signs are consistent. On the other hand, the correlation between 0-hour and 6-hour EFSO impacts (panel (a) of Figure 4.7) is much weaker, even with some negative correlation, suggesting that 0-hour EFSO (*i.e.*, analysis sensitivity to observations) cannot detect "flawed" observations.

The result for this particular case suggests that, even for individual observations, 6-hour EFSO and 24-hour EFSO are consistent to some extent, especially if we focus on observations with large negative impact. However, just looking at one case is not sufficient for drawing conclusions. In Table 4.1 we show how many of the observations whose 6-hour EFSO values are at least one standard deviation above the mean also have negative impact for the 24-hour forecast (*i.e.*, have positive 24-hour EFSO values). Except 3 out of 20 cases (the cases # 5, 11 and 14), the percentage is larger than 50%; the percentage is above 75% in as many as 8 cases (the cases # 1, 8, 12, 13, 17, 18, 19 and 20). This leads us to conclude that "flawed" observations can be detected by 6-hour EFSO as well.

4.4 Dependence on the choice of verifying truth

We now proceed to examine whether the choice of verifying truth significantly affects the EFSO results. We first compare statistical properties of the EFSO results verified against the ensemble mean analysis from LETKF with those verified against the control analysis from GSI. We then compare the two for individual observations. Our conclusion is that the choice of the verifying truth is not important.

4.4.1 Statistical properties

We first examine whether the time-averaged EFSO impacts from each observation type differ significantly by using different verifying truth. Figure 4.8 is the equivalent of Figure 4.3 verified against the ensemble mean analysis from LETKF. They show the time-averaged EFSO impacts from each observation type evaluated for different forecast lead times (6, 12 and 24 hours, from left to right) with dif-

Case #	Date	Latitude	Longitude	$> +\sigma$	"Hit"	%
1	2012-Jan-12-00Z	90 °S – 60 °S	110 °E – 140 °E	171	137	80.2
2	2012-Jan-12-18Z	60 °N – 90 °N	140 °E – 180 °	197	128	65.0
3	2012-Jan-13-06Z	60 °N – 90 °N	70 °W – 20 °W	427	282	66.0
4	2012-Jan-14-18Z	45 °N – 90 °N	120 °E – 150 °E	368	262	71.2
5	2012-Jan-15-18Z	60 °N – 90 °N	10 °E – 80 °E	858	278	32.4
6	2012-Jan-17-18Z	60 °N – 90 °N	50 °W – 0 °	516	283	54.8
7	2012-Jan-18-06Z	90 °S – 60 °S	70 °W – 30 °W	1,093	629	57.5
8	2012-Jan-18-18Z	45 °N – 90 °N	120 °E – 150 °E	284	225	79.2
9	2012-Jan-26-18Z	60 °N – 90 °N	40 °E – 80 °E	407	238	58.5
10	2012-Jan-27-00Z	60 °N – 90 °N	30 °E – 80 °E	177	105	59.3
11	2012-Jan-27-00Z	60 °N – 90 °N	20 °W – 10 °E	427	196	45.9
12	2012-Jan-28-18Z	60 °N – 90 °N	50 °E – 90 °E	462	368	79.7
13	2012-Feb-02-18Z	60 °N – 90 °N	40 °E – 110 °E	547	450	82.3
14	2012-Feb-03-00Z	60 °N – 90 °N	60 °E – 90 °E	315	18	8.1
15	2012-Feb-04-00Z	60 °N – 90 °N	40 °W – 10 °W	244	133	54.5
16	2012-Feb-05-12Z	90 °S – 60 °S	60 °W – 0 °	436	298	68.3
17	2012-Feb-06-18Z	60 °N – 90 °N	40 °E – 100 °E	582	497	85.4
18	2012-Feb-06-18Z	90 °S – 60 °S	60 °W – 10 °E	576	471	81.8
19	2012-Feb-09-06Z	60 °N – 90 °N	140 °W – 90 °W	1,592	1,268	79.6
20	2012-Feb-10-06Z	90 °S – 60 °S	50 °E – 80 °E	104	80	76.9

Table 4.1: Percentage of "large 6-hour EFSO" observations whose 24-hour EFSO values are positive, for the 20 cases identified as "regional dropouts" (c.f. Section 6.2). "Large 6-hour EFSO" observations are defined as the observations whose EFSO values exceed one standard deviation (σ) above the mean, the number of which is shown in the fifth column denoted by "$> +\sigma$." The number of observations among the "large 6-hour EFSO" observations whose 24-hour EFSO impacts are negative (*i.e.*, the value is positive) is shown in the sixth column (denoted by "Hit"), whose fraction (in percent) is shown in the rightmost column.

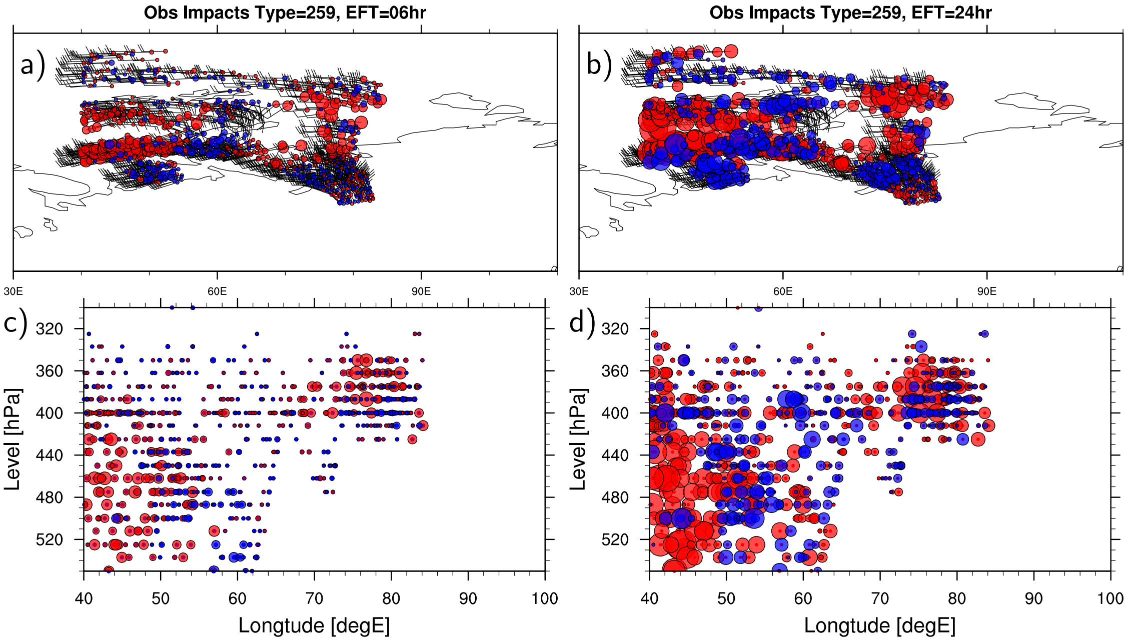


Figure 4.6: Geographical and vertical distributions of EFSO impacts for individual MODIS wind (stattype 259; see Table 3.3) observations on one of the "regional dropout" cases (18 UTC of February 6th, 2012 and the area of 60 °N – 90 °N, 40 °E – 100 °E). Shown are (a) horizontal distribution of 6-hour EFSO impacts (b) as in (a), but for 24-hour EFSO, (c) vertical distribution of 6-hour EFSO, and (d) as in (c), but for 24-hour EFSO. Red and blue circles represent, respectively, negative and positive impacts (*i.e.*, positive and negative EFSO values). The area encircled by each circle corresponds to the magnitude of the EFSO impact. Wind barbs in panels (a) and (b) represent the observed wind. Each MODIS wind observation is composed of a pair of observations, one for u (zonal wind) and the other for v (meridional wind), which are assimilated separately. Here, the impact for each MODIS observation is defined as the sum of the impacts from its u and v component. The EFSO is verified against the control GSI analysis and is measured with the moist total energy norm restricted to the above mentioned "rectangular region."

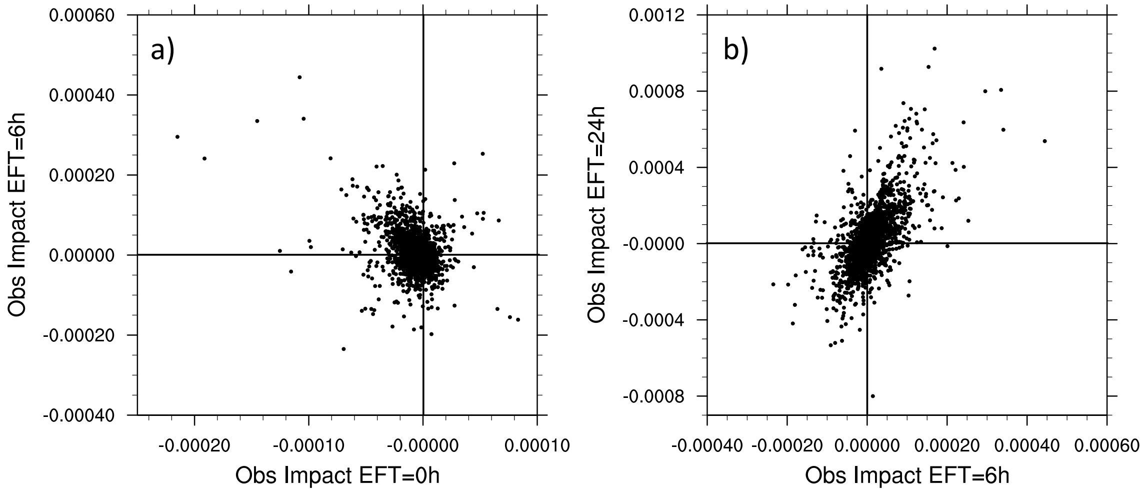


Figure 4.7: Scatter plot of (a) 0-hour and 6-hour EFSO, and (b) 6-hour and 24-hour EFSO (in J kg^{-1}) for the case and the observation type shown in Figure 4.6.

ferent error norms (moist and dry total energy norm, respectively, for the top and bottom panels). Similarity between any panel of the two figures is clearly evident. The only noticeable discrepancy is that Aircraft and MODIS wind have larger impact, especially with 6-hour lead time (panels (a) and (d)), when verified against the ensemble mean LETKF analysis (Figure 4.8) than against the control GSI analysis (Figure 4.3). Apart from this minor discrepancy, the two results are highly consistent, indicating the insignificance of the choice of the verifying truth.

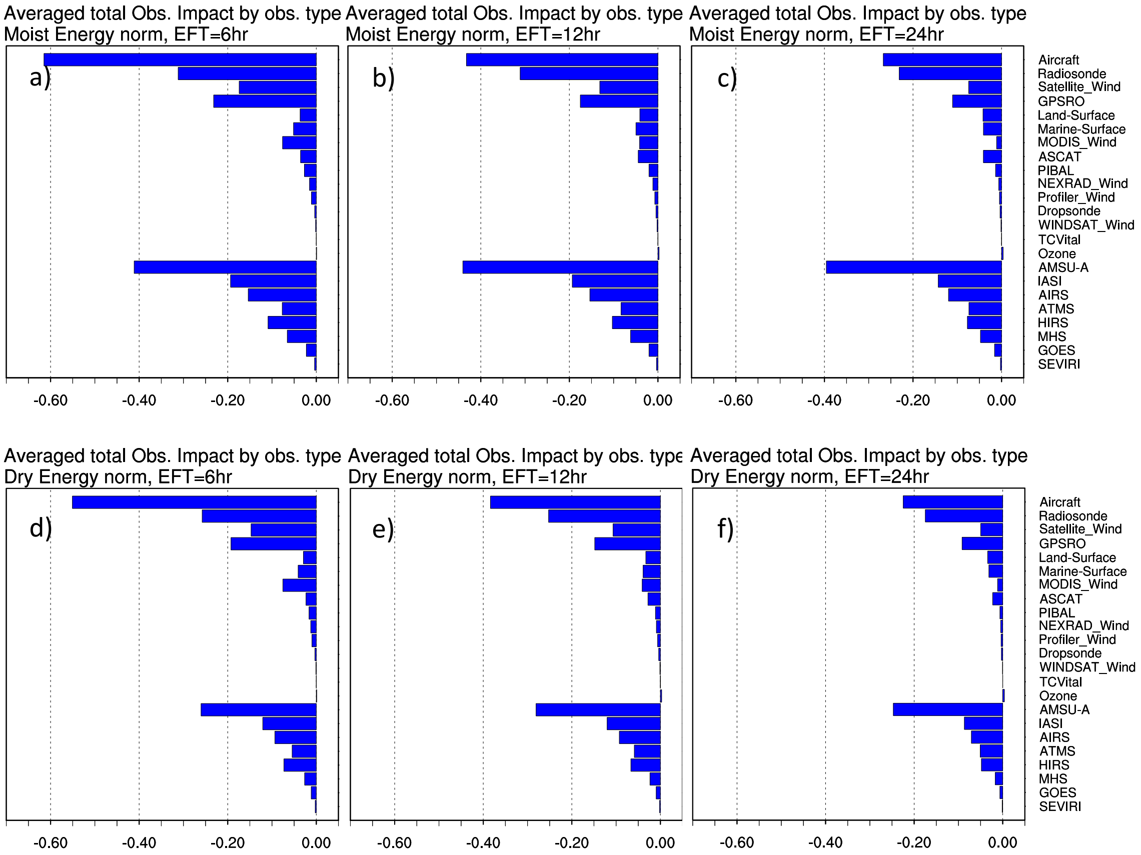


Figure 4.8: As in Figure 4.3, but for the EFSO impacts verified against the ensemble mean LETKF analysis.

4.4.2 Individual cases

Now we examine whether the consistency between EFSO verified against the two analyses is also robust for individual observations. Figure 4.9 shows the scatter plots of the two EFSO values (one verified against the control GSI analysis and the other verified against the ensemble mean LETKF analysis) for three different observation types (Aircraft, MODIS wind and AMSU-A) assimilated at a particular date and time of 18 UTC of February 1st, 2012. For any of the three observation types shown, we can observe clear, high correlation. The correlation is higher for 24-hour EFSO (right panels) than for 6-hour EFSO (left panels). The 6-hour EFSO of MODIS wind (panel (c)) exhibits some observations for which the two EFSO values do not agree, especially for large positive values (*i.e.*, negative impacts). Nevertheless, we can conclude that, overall, the EFSO diagnostics is not too sensitive to the choice of verification.

4.5 Summary

In this chapter, we first compared 24-hour EFSO from our experiments with previous FSO studies to check if the EFSO applied to a hybrid DA system yields reasonable results. Our results are in fact consistent with the previous studies, confirming the applicability of EFSO to a hybrid DA system. We then compared EFSOs of different forecast lead times and found that, somewhat to our surprise, EFSO with the lead time as short as 6 hours is overall consistent with that of the tried-and-true 24-hour lead time, especially if we focus on the EFSO of large negative

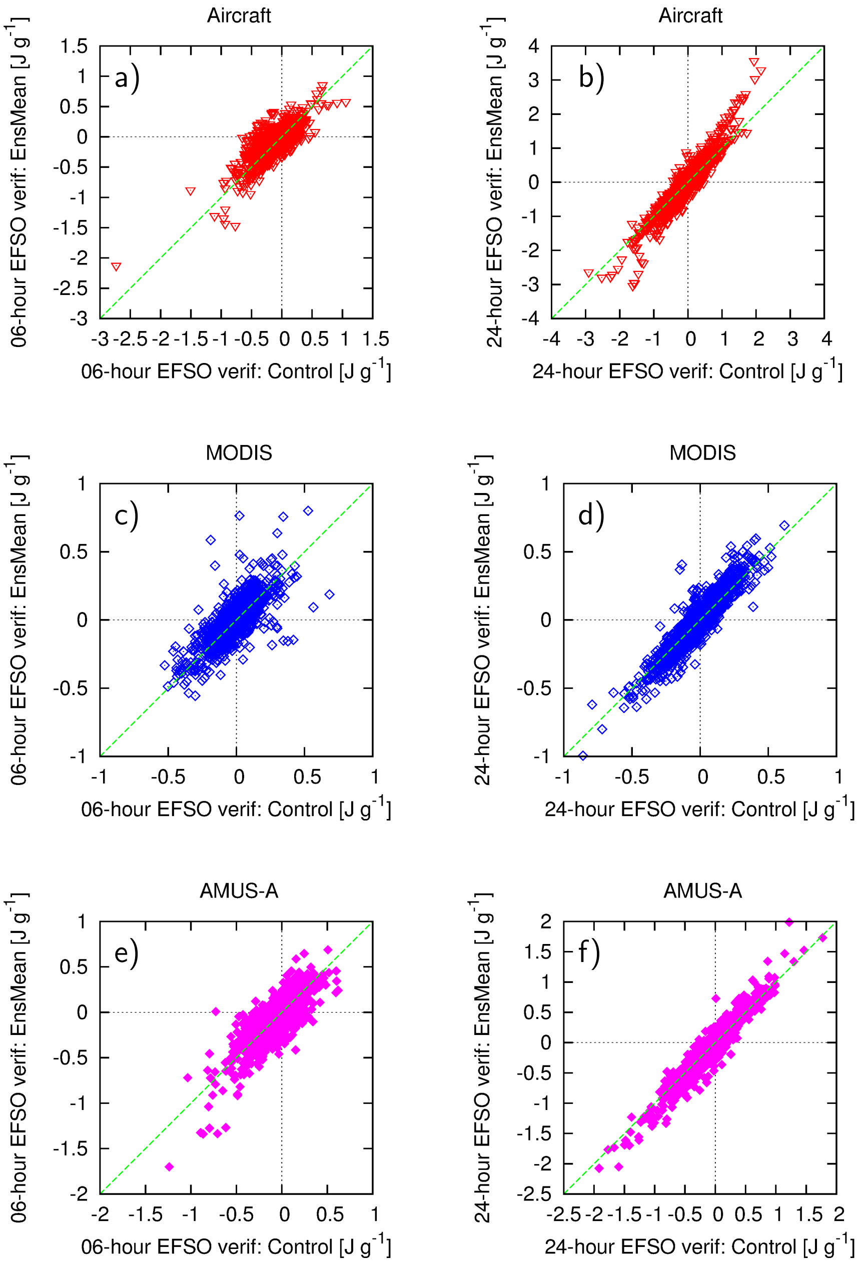


Figure 4.9: Scatter plots of the two EFSO values (in 10^{-3}J kg^{-1}) for individual observations, one verified against the control GSI analysis (x-axis) and the other verified against the ensemble mean LETKF analysis (y-axis), for (a,b) Aircraft observations, (c,d) MODIS wind observations and (e,f) AMSU-A observations. The left panels ((a), (c) and (e)) show the results for 6-hour EFSO; the right panels ((b), (d) and (f)) show the results for 24-hour EFSO. All EFSO values are measured globally with the moist total energy norm. The observations assimilated at 18 UTC of February 1st, 2012 are shown.

impacts. We then examined whether the choice of the verifying truth significantly affects the results of EFSO, and we found that the EFSO is rather insensitive to this choice, even for individual observations.

The fact that 6-hour and 24-hour EFSO can consistently identify observations with large negative impacts is of vital importance to us because this gives us hope that 6-hour EFSO can be used for Proactive QC.

Because the EFSO is found not to be too sensitive to the choice of the verifying truth, in later chapters we mainly use the control GSI analysis as the verifying truth: the control analysis is generally considered to be more accurate than the ensemble mean EnKF analysis because it is produced at a higher horizontal resolution and it combines the robustness of static background covariance with the flow-dependence from EnKF.

Chapter 5: Improvement of “regional dropout” detection algorithm

5.1 Introduction

In our proposed Proactive QC algorithm, we divide the globe into relatively small regions and apply an algorithm that detects possible “regional forecast dropouts” (Step 3. of the algorithm described in Section 1.4.1 and Section 2.3.2). Ideally, we should perform EFSO on all regions and see if there are any observations the rejection of which would reduce forecast errors. However, with ~ 400 regions per each analysis (see Section 5.2), this would be too computationally expensive and not feasible for an operational system, although EFSO is considerably computationally more efficient than the adjoint based FSO. Thus, we have to “screen out” regions for which rejection of a subset of observations is unlikely to improve the forecast, before performing EFSO. Following Ota et al. (2013), we examine criteria based on the following two measures: a) the error of the t-hour forecast initialized by the analysis normalized by its climatological mean, $e^f_{t|0}/\left\langle e^f_{t|0}\right\rangle$, which, hereafter we refer to simply as “normalized regional forecast errors,” and b) the ratio of the errors of the t-hour and $t+6$-hour forecasts validating at the same time, $e^f_{t|0}/e^f_{t|-6}$, which we refer to simply as “regional forecast error reduction by the analysis.” Here, the errors are measured with the moist total energy norm restricted to each region. The first criterion measures how much the regional forecast error is large or small compared to its "usual" value; the second criterion measures how much the assimilation of observations reduced (or increased) the regional forecast error. Ota et al. (2013) used the thresholds of 1.7 for a) and 1.2 for b) (c.f. Section 2.3.1), respectively, with the forecast lead time $t = 24$ hours, and identified 7 possible "regional dropout" cases, two of which were improved by more than 20% by rejecting the "flawed" observations that were identified by regional EFSO. While their achievement is a great success, there is still some room for improvement. First, as we discussed in Section 2.4.3, the way they divided the globe into smaller subdomains may not be optimal. Second, the thresholds for a) and b) are determined rather subjectively and may not be optimal. It is not clear whether they should be constant for all latitudes because the dynamics of the atmosphere have different characteristics in the extratropics and tropics. Furthermore, if the two quantities (normalized regional forecast errors, $e^f_{t|0}/\left\langle e^f_{t|0}\right\rangle$, and the regional forecast error reduction by the analysis, $e^f_{t|0}/e^f_{t|-6}$) are strongly correlated, then it would suffice to use just only one of them.

The issues we stated above are discussed in this chapter. In Section 5.2, we introduce two approaches for decomposing the globe. In Section 5.3 we examine the statistics of the normalized regional forecast errors $e^f_{t|0}/\left\langle e^f_{t|0}\right\rangle$ and the regional forecast error reduction by the analysis $e^f_{t|0}/e^f_{t|-6}$, first using samples from the entire globe, then limiting the samples to Northern Hemisphere (NH) extratropics, Southern Hemisphere (SH) extratropics, and the tropics. We will also examine the statistics with samples only from near the North and South Poles. Then, using these statistics, we pick-up ~ 200 cases in which we assume regional forecast dropouts are likely to occur. We then perform EFSO to each of them to estimate how much the forecast can be improved by not using the observations of the stattypes (see Section 3.5) or satellite sensors with net negative impacts. The selection of cases based on the the normalized regional forecast errors $e^f_{t|0}/\left\langle e^f_{t|0}\right\rangle$ and the regional forecast error reduction by the analysis $e^f_{t|0}/e^f_{t|-6}$ can be considered a success if, among the selected cases, there are cases for which the forecast is estimated to be significantly improved by not using the negatively-impacting observation types detected by EFSO. Further, we compare the estimated forecast improvements with the normalized regional forecast errors $e^f_{t|0}/\left\langle e^f_{t|0}\right\rangle$ and the regional forecast error reduction by the analysis $e^f_{t|0}/e^f_{t|-6}$ to see which of the two is the better statistical predictor for the estimated forecast improvements.

We note that the statistical analysis we present in this chapter is rather ad hoc and more like engineering than science; the optimal criteria could be different from one system to another, so we might have to re-tune them on a system-to-system basis.

5.2 Domain decomposition

As we discussed in Section 2.4.3, how best to divide the globe into smaller regions is not a trivial question. A naïve way, used by Ota et al. (2013), is to simply divide the globe into "rectangular" regions in latitude-longitude space, for example, $30^\circ \times 30^\circ$ "squares." One concern with this approach is that, the area of each region, or cell, varies with the latitude and becomes highly non-uniform. For example, the area of a cell on the 15 °S–15 °N band is larger than that on the 60 °N–90 °N band by more than 4 times. Ota et al. (2013) solved this issue by adjusting the longitude spacing so that the areas of each region become as uniform as possible. A possible drawback of this approach is that, in the tropics, the cell becomes meridionally elongated, when, in fact, the structures of tropical disturbances are zonally elongated. Here, we propose another approach in which the latitude spacing rather than longitude spacing is adjusted: we divide the globe by the zeros of the isotropic spherical harmonics with total wavenumber 12 ($Y_{12}^{6}(\lambda, \varphi)$). The advantage of this approach is that the areas of each cell become close to uniform (see Table 5.1 and the right panels of Figure 5.1). To allow for overlaps, we also use domain decomposition that is formed by zonal and meridional lines which connect anti-nodes of $Y_{12}^{6}(\lambda, \varphi)$. Also, overlaps in the zonal direction is allowed by shifting the longitude by 10 °. We call this domain decomposition "Y_{12}^{6} cells." The Y_{12}^{6} cells consist of a total of $(7 + 6) \times 36 = 468$ cells.

Along with the Y_{12}^{6} cells, we also examine the naïve "30 °×30 ° cells" where the globe is divided into 30 °×30 ° latitude-longitude cells, allowing overlaps by shifting the latitude and longitude, respectively, by 15 ° and 10 °. The 30 °×30 ° cells consist of a total of $(6 + 5) \times 36 = 396$ cells. The rationale for testing this division, despite the non-uniformity of the areas of each cell, is that the characteristic horizontal scales of meteorological disturbances decrease as the latitude goes higher: in the extratropics, the horizontal scale of disturbances are roughly determined by the

Rossby radius of deformation (e.g. Balgovind et al., 1983; Gill, 1982):

$$\lambda_R = \frac{\sqrt{gH}}{f} = \frac{\sqrt{gH}}{2\Omega \sin\varphi} \tag{5.1}$$

where g is the gravitational acceleration, H is the equivalent depth of the vertical mode, f is the Coriolis parameter, Ω is the angular velocity of Earth's rotation, and φ is the latitude (in radian). As the latitude increases, f also increases, thus decreasing λ_R. Moreover, the equivalent depth H also tends to become smaller in higher latitude because the static stability is generally stronger at higher latitudes. The tropopause level is also lower at the higher latitudes ($\sim$ 300 hPa near the Poles and $\sim$ 150 hPa at the Equator). For this reason, we guess that using larger regions in the lower latitude than in the higher latitude could be justifiable.

Figure 5.1 shows how the globe is divided into smaller cells by (top) the $30^\circ{\times}30^\circ$ cells and (bottom) Y_{12}^{6} cells, with (left) the orthograpic projection and (right) the cylindrical equidistant projection. The area of each cell in each latitude band is shown by the width of gray boxes on the right edge of the right panels. As we discussed, the areas of each cell in each latitude bands are highly nonuniform in $30^\circ{\times}30^\circ$ cells but close to uniform in Y_{12}^{6} cells. The specification of each latitude that divide the cells, along with the area of each cell in each latitude band, is shown in Table 5.1.

Latitude Specification

30 °×30 ° cells

minlat	maxlat	area (km^2)	minlat	maxlat	area (km^2)
60 °N	90 °N	2,748,630	45 °N	75 °N	5,370,630
30 °N	60 °N	7,715,440	15 °N	45 °N	9,600,990
0 °	30 °N	10,835,300	15 °S	15 °N	11,266,300
30 °S	0 °	10,835,300	45 °S	15 °S	9,600,990
60 °S	30 °S	7,715,440	75 °S	45 °S	5,370,630
90 °S	60 °S	2,748,630			

Y^6_{12} cells

minlat	maxlat	area (km^2)	minlat	maxlat	area (km^2)
42.4482 °N	90.0000 °N	6,743,580	33.6865 °N	54.0107 °N	5,366,600
24.8270 °N	42.4482 °N	5,440,880	16.5516 °N	33.6865 °N	5,805,080
8.2020 °N	24.8270 °N	6,006,440	0.0000 °	16.5516 °N	6,198,280
8.2020 °S	8.2020 °N	6,217,080	16.5516 °S	0.0000 °	6,198,280
24.8270 °S	8.2020 °S	6,006,400	33.6865 °S	16.5516 °S	5,805,080
42.4482 °S	24.8270 °S	5,440,880	54.0107 °S	33.6865 °S	5,366,600
90.0000 °S	42.4482 °S	6,743,580			

Table 5.1: Specification of the latitudes along with the area of each cell on each latitude band for (top) the 30 °×30 ° and (bottom) Y^6_{12} cells.

5.3 Statistics of the normalized regional forecast errors and the regional forecast error reduction by the analysis

In this section we examine how the normalized regional forecast errors $e^f_{t|0}/\left\langle e^f_{t|0}\right\rangle$ and the regional forecast error reduction by the analysis $e^f_{t|0}/e^f_{t|-6}$ are statistically distributed and whether they are correlated. Specifically, we are interested in whether the statistics are different for the two different divisions of the globe (30 °×30 ° and Y^6_{12}). We first look at the statistics with samples from the whole globe. Since

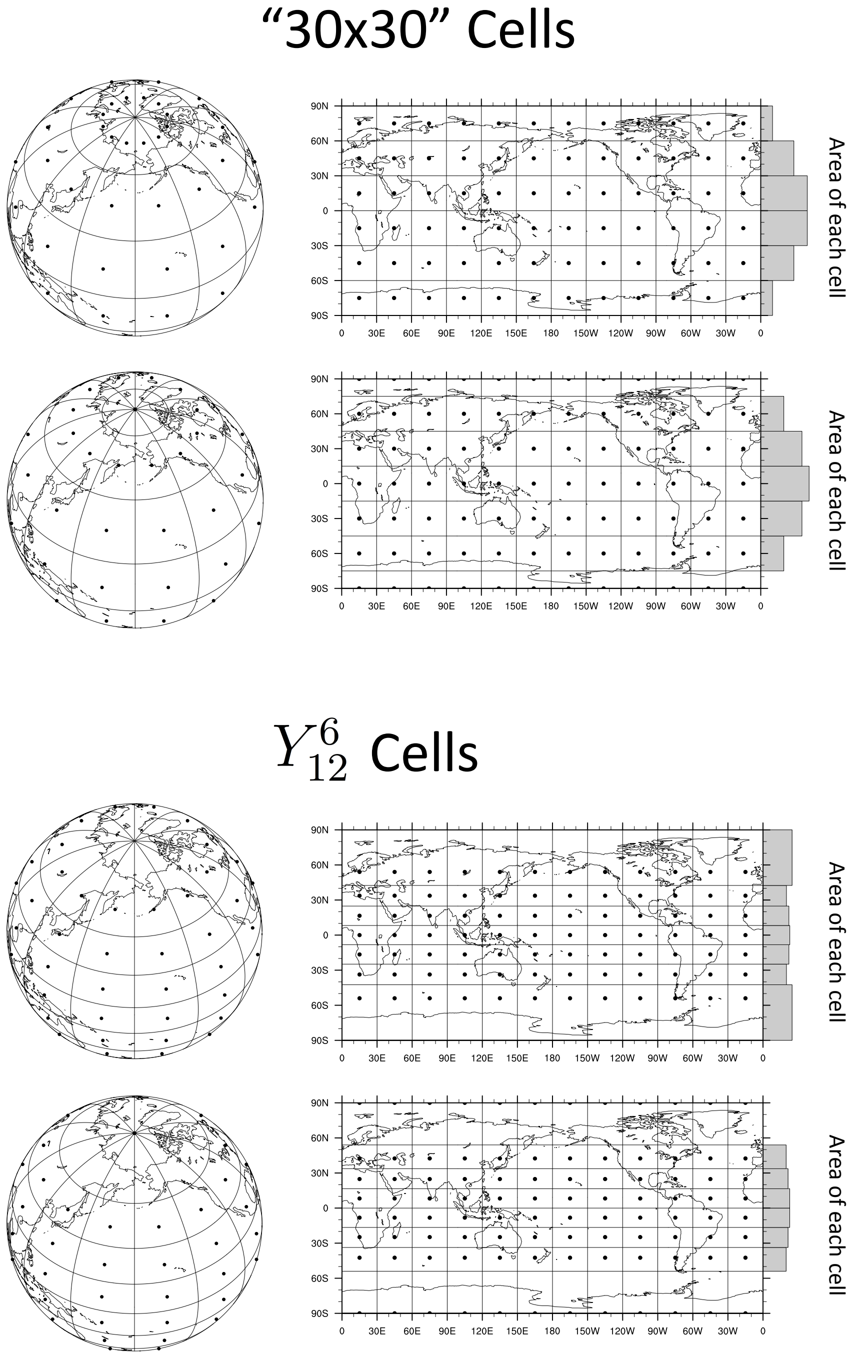
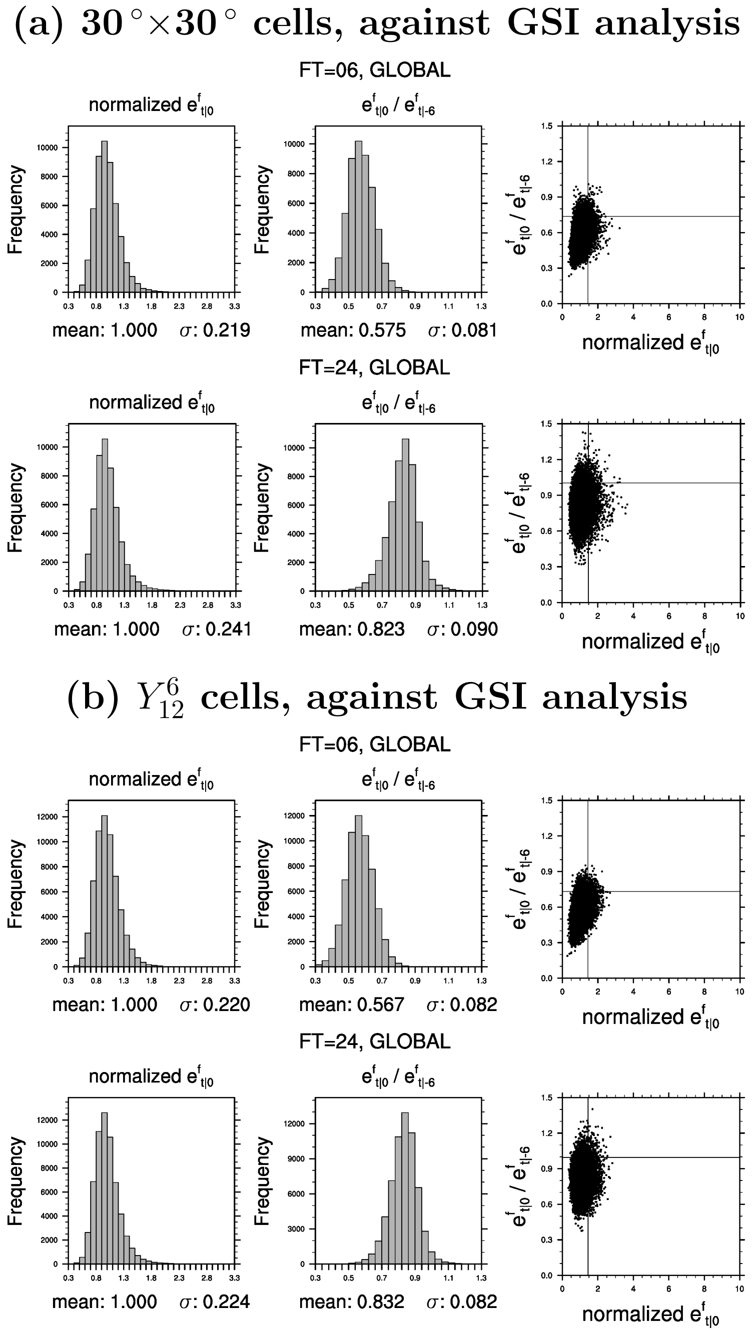


Figure 5.1: (Top) 30°×30° cells and (bottom) Y_{12}^{6} cells represented with (left) orthographic and (right) cylindrical equidistant projections. To allow for meridional overlaps, the two division approaches both have two different ways to segment the latitude. The area of each cell in each latitude band is shown by the width of gray boxes on the right edge of the right panels.

the dynamics of meteorological disturbances is different for different latitudes, the statistics may be also different. Thus, we also look at statistics with samples only from tropics, the NH and SH extratropics, and near the North Pole and the South Pole.

Figure 5.2a shows the statistics of the normalized regional forecast errors and the regional forecast error reduction by the analysis verified against the control GSI analysis for $30^\circ \times 30^\circ$ cells. The three panels in the upper row show, from left to right, the histogram of normalized regional forecast errors, the histogram of regional forecast error reduction by the analysis, and the scatter plot of the two, all for the lead time $t = 6$ hours; similarly, the lower three panels are for the lead time $t = 24$ hours. The sample mean and standard deviation (σ) are indicated at the bottom of each histogram. These statistics are also summarized in Table 5.2. The same plots for Y_{12}^{6} cells are shown in Figure 5.2b. Let us first look at the results for $30^\circ \times 30^\circ$ cells (Figure 5.2a). The normalized regional forecast errors, by definition, has mean 1 and its distribution is somewhat skewed to the right, both for $t = 6$ hours and 24 hours. An interesting observation is that the standard deviation is not very different for the two lead times, with $\sigma = 0.219$ for $t = 6$ hours and $\sigma = 0.241$ for $t = 24$ hours. The mean of the regional forecast error reduction by the analysis is 0.575 for $t = 6$ hours and 0.823 for $t = 24$ hours, both less than 1, meaning that the assimilation of observations on average act to reduce regional forecast errors. From the scatter plots, we can observe that, for both lead times, the dots are clustered in the upper left, indicating that if the regional t-hour forecast is less accurate than usual, then the forecast error is not reduced much by the analysis; equivalently, if



(a) 30°×30° cells, against GSI analysis

(b) Y^6_{12} cells, against GSI analysis

Figure 5.2: (a): Statistics of the normalized regional forecast errors and the regional forecast error reduction by the analysis for (top) 6-hour lead time and (bottom) 24-hour lead time. Errors are measured for each cell of the 30°×30° cells with the moist total energy norm and are verified against the control GSI analysis. The samples are taken from all cells (global). Shown are (left) the histogram of the normalized regional forecast errors, (middle) the histogram of regional forecast error reduction by the analysis, and (right) the scatter plot of normalized regional forecast errors (abscissa) and regional forecast error reduction by the analysis (ordinate). The vertical and horizontal lines in the scatter plot denote mean-plus-2σ levels. (b): As in (a), but for the Y^6_{12} cells.

the forecast error is significantly reduced by the analysis, then the regional forecast error is likely to be smaller than usual. However, there is no clear linear correlation between the two quantities. By comparing Figure 5.2a and Figure 5.2b, we find that, contrary to our initial expectation, the distributions do not depend on how we divide the globe.

Next we examine if the choice of the verifying truth affects the statistics. Figure 5.3 shows the statistics of regional forecast errors in the same format as in Figure 5.2. As in Figure 5.2, the results for the $30^\circ \times 30^\circ$ cells and the Y_{12}^6 cells do not differ much. Compared to Figure 5.2 in which the errors are verified against the control GSI analysis, the results for 6-hour lead time is somewhat different, with the mean of regional forecast error reduction by the analysis smaller when verified against the ensemble mean LETKF analysis (0.463 for $30^\circ \times 30^\circ$ cells and 0.461 for Y_{12}^6 cells) than when verified against the control GSI analysis (0.575 for $30^\circ \times 30^\circ$ cells and 0.567 for Y_{12}^6 cells). Also, there are several cases with very large normalized regional forecast errors which exceed 4 (rightmost panels). Those cases all occurred in the tropics.

The statistics (mean and standard deviation) with samples taken only from the NH extratropics (*i.e.*, north of 30°N) and the SH extratropics (*i.e.*, south of 30°S) are shown, respectively, in the second and third rows of Table 5.2a and Table 5.2b. As in the statistics with global sampling, there seems to be no noticeable difference between $30^\circ \times 30^\circ$ cells and Y_{12}^6 cells. The results for the tropics (between 30°S and 30°N; fourth rows) also seem to have no noticeable difference between $30^\circ \times 30^\circ$ cells and Y_{12}^6 cells. The scatter plots between normalized regional forecast

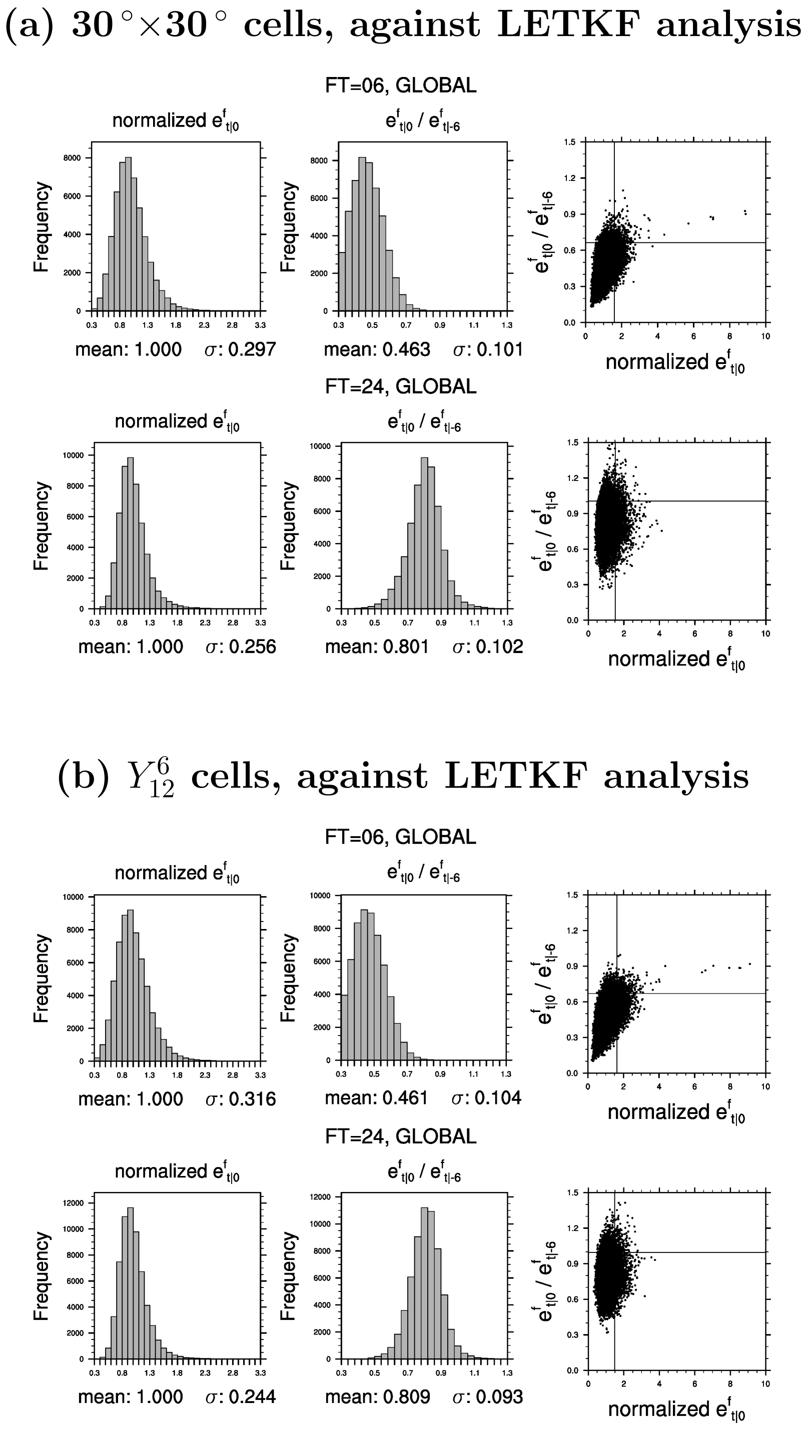
Summary of Statistics : (a)$30^\circ\times 30^\circ$ cells

verification		GSI				LEKTF			
		normalized $e^f_{t\vert 0}$		$e^f_{t\vert 0}/e^f_{t\vert -6}$		normalized $e^f_{t\vert 0}$		$e^f_{t\vert 0}/e^f_{t\vert -6}$	
lead time		6hrs.	24hrs.	6hrs.	24hrs.	6hrs.	24hrs.	6hrs.	24hrs.
Globe	mean	1.000	1.000	0.575	0.823	1.000	1.000	0.463	0.801
	std.dev.	0.219	0.241	0.081	0.090	0.297	0.256	0.101	0.102
NH	mean	1.000	1.000	0.603	0.808	1.000	1.000	0.476	0.783
	std.dev.	0.255	0.288	0.088	0.107	0.334	0.302	0.118	0.120
SH	mean	1.000	1.000	0.574	0.814	1.000	1.000	0.457	0.794
	std.dev.	0.225	0.281	0.071	0.104	0.269	0.295	0.085	0.122
Trop.	mean	1.000	1.000	0.558	0.839	1.000	1.000	0.459	0.815
	std.dev.	0.189	0.172	0.076	0.062	0.288	0.192	0.097	0.071
North Pole	mean	1.000	1.000	0.609	0.798	1.000	1.000	0.482	0.775
	std.dev.	0.278	0.334	0.093	0.120	0.347	0.342	0.115	0.134
South Pole	mean	1.000	1.000	0.584	0.799	1.000	1.000	0.452	0.782
	std.dev.	0.258	0.321	0.081	0.113	0.284	0.326	0.098	0.133

Summary of Statistics : (b) Y^6_{12} cells

verification		GSI				LEKTF			
		normalized $e^f_{t\vert 0}$		$e^f_{t\vert 0}/e^f_{t\vert -6}$		normalized $e^f_{t\vert 0}$		$e^f_{t\vert 0}/e^f_{t\vert -6}$	
lead time		6hrs.	24hrs.	6hrs.	24hrs.	6hrs.	24hrs.	6hrs.	24hrs.
Globe	mean	1.000	1.000	0.567	0.832	1.000	1.000	0.461	0.809
	std.dev.	0.220	0.224	0.081	0.082	0.316	0.244	0.104	0.093
NH	mean	1.000	1.000	0.600	0.808	1.000	1.000	0.472	0.782
	std.dev.	0.232	0.256	0.082	0.099	0.313	0.273	0.113	0.112
SH	mean	1.000	1.000	0.571	0.819	1.000	1.000	0.462	0.799
	std.dev.	0.209	0.262	0.064	0.096	0.257	0.279	0.078	0.111
Trop.	mean	1.000	1.000	0.559	0.840	1.000	1.000	0.459	0.816
	std.dev.	0.220	0.207	0.084	0.072	0.329	0.229	0.107	0.083
North Pole	mean	1.000	1.000	0.605	0.800	1.000	1.000	0.481	0.775
	std.dev.	0.214	0.243	0.078	0.097	0.289	0.256	0.103	0.109
South Pole	mean	1.000	1.000	0.573	0.809	1.000	1.000	0.457	0.789
	std.dev.	0.201	0.255	0.061	0.094	0.235	0.267	0.070	0.109

Table 5.2: Summary of the statistics (mean and standard deviation) with samples from different latitude ranges, for (a) $30^\circ\times 30^\circ$ cells and (b) Y^6_{12} cells of the normalized regional forecast errors $\mathbf{e}^f_{t|0}/\left\langle e^f_{t|0}\right\rangle$ and regional forecast error reduction by the analysis $e^f_{t|0}/e^f_{t|-6}$.



(a) $30^\circ\times30^\circ$ cells, against LETKF analysis

(b) Y_{12}^{6} cells, against LETKF analysis

Figure 5.3: As in Figure 5.2, but verified against the ensemble mean LETKF analysis.

errors and regional forecast error reduction by the analysis for the NH and SH extratropics and the tropics (not shown) also did not show any noticeable difference between $30^\circ \times 30^\circ$ and Y_{12}^6 cells.

Since the largest difference between the $30^\circ \times 30^\circ$ cells and Y_{12}^6 cells is at the polar areas, we can expect to see clearer contrast in the statistics for the two divisions by focusing on cells near the poles. With this anticipation, we also looked at the statistics with samples taken only from the cells that encircle the poles. The results are shown in the last two rows of Table 5.2. For both North and South Poles, the statistics are overall quite similar between $30^\circ \times 30^\circ$ cells and Y_{12}^6 cells. One noticeable feature is that, the standard deviations are larger in $30^\circ \times 30^\circ$ cells than in Y_{12}^6 cells, both for normalized regional forecast errors and regional forecast error reduction by the analysis and for 6-hour and 24-hour lead times, for both GSI and LETKF analysis. Here again, the scatter plots between normalized regional forecast errors and regional forecast error reduction by the analysis for the two polar regions (not shown) did not show any noticeable difference between $30^\circ \times 30^\circ$ and Y_{12}^6 cells.

The above statistical assessment was motivated by our anticipation that the distributions of the normalized regional forecast errors and the regional forecast error reduction by the analysis might have different characteristics for different latitude ranges, and for different methods to divide the globe, from our intuition that dynamics of meteorological disturbances are different for different latitudes. However, contrary to our expectation, we did not see clear distinction between different latitude ranges and different methods to divide the globe.

5.4 Correlation with the estimated forecast improvement

As we described in the previous section, we did not see clear differences for different latitude ranges in the statistics of normalized regional forecast errors and regional forecast error reduction by the analysis. Also, the dependence on the verifying truth seems not to be significant. For these reasons, we extracted cases to perform EFSO using the same threshold to all latitudes, and concentrated on verification with the control GSI analysis. The thresholds we applied are 2 standard deviation above the mean: for each of 6-hour and 24-hour lead times, and for each of $30^\circ \times 30^\circ$ and Y_{12}^{6} divisions, we extracted cases for which both normalized regional forecast errors and regional forecast error reduction by the analysis are above these thresholds. From the 34-day period (c.f. Section 3.7), these criteria, which hereafter we call by the "2σ-criteria," resulted in about $\sim$ 200 cases for each lead time and each globe division (Table 5.3), which is about 1–2 EFSOs per cycle and thus should be feasible in an operational system.

Number of the extracted cases with the 2σ criteria

lead time	$30^\circ \times 30^\circ$	Y_{12}^{6}
6 hours	219	266
24 hours	189	211

Table 5.3: The number of the cases extracted by the "2σ-criteria."

For each case, we performed regional EFSO and computed the "estimated regional forecast improvement" by the following procedure:

First, we compute the sum of t-hour regional EFSO impacts for each "stattype" (for non-radiance data; see Table 3.2 and Table 3.3) and each sensor (for radiance data; see Table 3.4). If there are any types whose net impacts are negative (*i.e.*, the values are positive), then we add them up and divide the sum by the regional t-hour forecast error $e^f_{t|0}$, and finally multiply it by 100. If there are no types with net negative impacts, the "estimated regional forecast improvement" is defined to be 0%.

As we described in Section 5.1, we expect some large percentage of the estimated forecast improvement to be found among the cases we selected with the 2σ-criteria.

In order to discern which of the following is the better "predictor" for forecast improvement, the normalized regional forecast errors $e^f_{t|0}/\left\langle e^f_{t|0}\right\rangle$ or the regional forecast error reduction by the analysis $e^f_{t|0}/e^f_{t|-6}$, we plot scatter diagrams between each of these two quantities and the estimated regional forecast improvement defined above. In interpreting these scatter plots, we are particularly interested in correlations between large values of the "predictand" (the estimated forecast improvements) and the "predictor" (the normalized regional forecast errors $e^f_{t|0}/\left\langle e^f_{t|0}\right\rangle$ or the regional forecast error reduction by the analysis $e^f_{t|0}/e^f_{t|-6}$).

Figure 5.4 shows the results for $30\,^\circ\times 30\,^\circ$ cells. The reassuring fact is that, there are quite a lot of cases with large estimated forecast improvements: for 6-hour (panels (a) and (c)) and 24-hour (panels (b) and (d)) lead time, we can find 15 (out of 219) and 13 (out of 189) cases, respectively, whose estimated forecast improvements are above 25% level. This assures us that, with the 2σ-criteria, we can effectively capture cases with large potential for forecast improvements by the denial of "flawed" observations.

Next, let us turn our attention to correlations. In the plot for normalized regional forecast errors with 6-hour lead time (panel (a)) it is difficult to discern any meaningful correlation. For 24-hour lead time (panel (b)) we can even see a negative correlation. In the plot for regional forecast error reduction by the analysis with 6-hour lead time, we can see some reasonable correlation, particularly if we focus on the right half of the plot where the regional forecast error reduction by the analysis is larger than ~ 0.85. We should not be too optimistic, however, because the sample size is rather small. For 24-hour lead time (panel (d)), again it is difficult to see any correlation.

Figure 5.5 shows the results for Y_{12}^{6} cells. In this case, the estimated forecast improvements are not very large (only 3 cases exceed 25% level, both for 6-hour and 24-hour lead times). It is also difficult to find any correlation for any of the panels, but normalized regional forecast errors with 6-hour lead time (panel (c)) can be considered the best (or "least bad") predictor because, for all cases with regional forecast error reduction by the analysis $e^{f}_{t|0}/e^{f}_{t|-6}$ less than 0.82, the estimated improvement is less than 15% and for all the cases whose estimated improvement is larger than 20%, the regional forecast error reduction by the analysis $e^{f}_{t|0}/e^{f}_{t|-6}$ is larger than 0.82.

5.5 Summary

To improve the "regional dropout" detection algorithm of Ota et al. (2013), in this chapter, we introduced two alternative globe divisions, the $30^\circ \times 30^\circ$ cells and

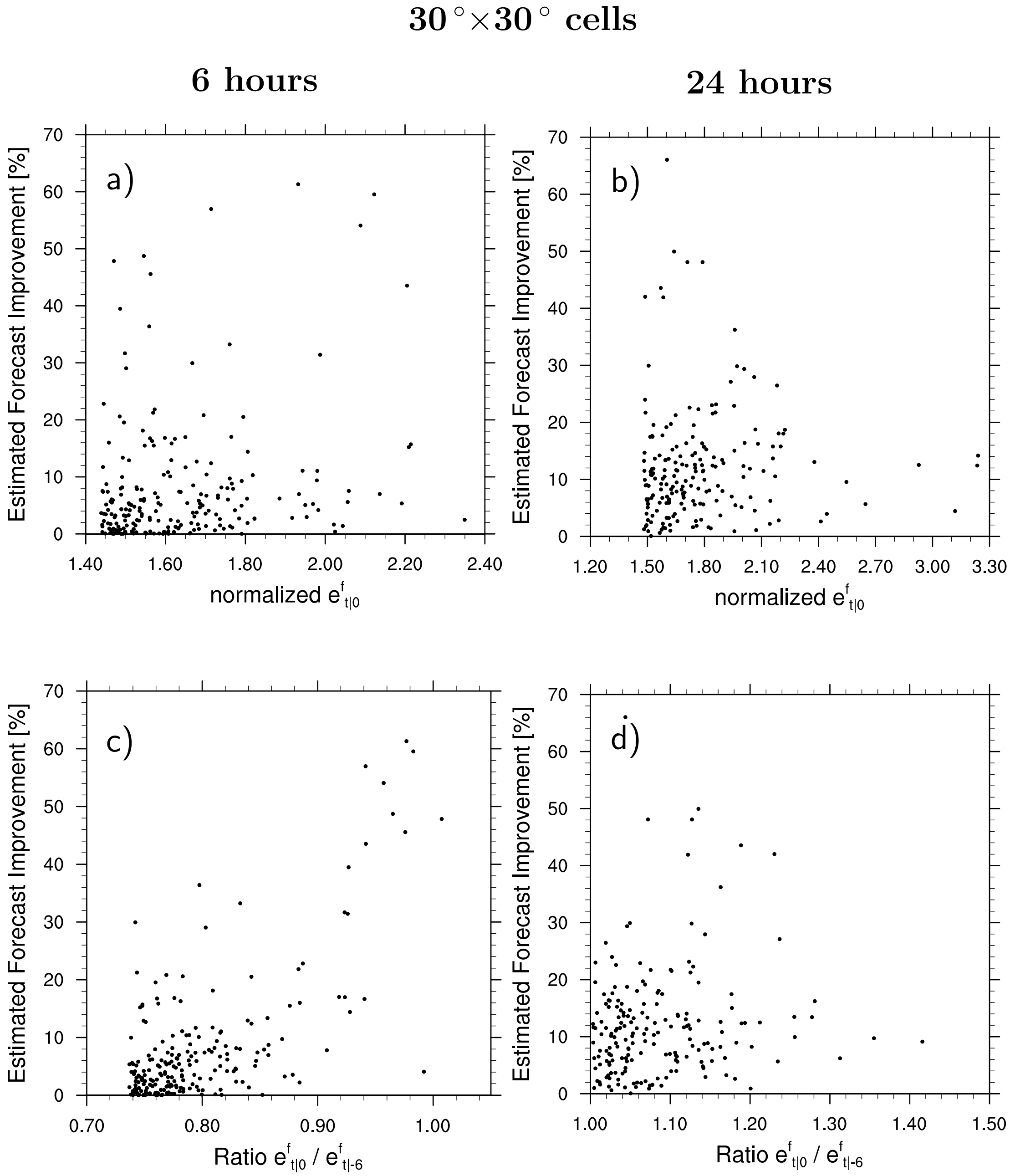


Figure 5.4: Scatter plots of the estimated forecast improvement (%) with (a,b) the normalized regional forecast errors, and (c,d) the regional forecast error reduction by the analysis. Panels on the left (a,c) and the right (b,d) are, respectively, for lead times of 6 hours and 24 hours. The samples are taken from the 30°×30° cells.

Y_{12}^{6} **cells**

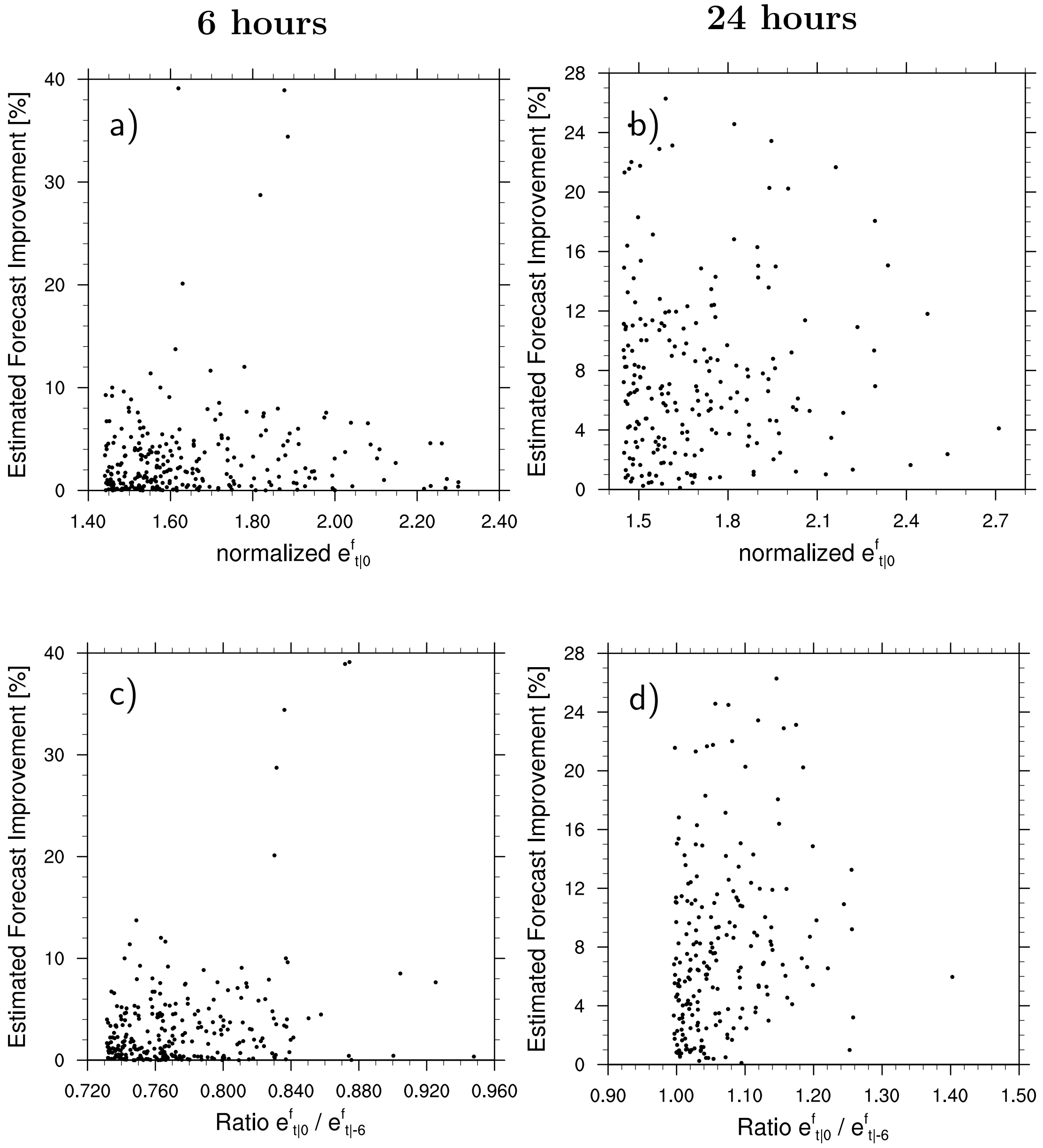


Figure 5.5: As in Figure 5.4, but for Y_{12}^{6} cells.

the Y_{12}^{6} cells, and examined the statistics of the two "predictors" for the estimated forecast improvement, the normalized regional forecast errors $e_{t|0}^{f}/\left\langle e_{t|0}^{f}\right\rangle$ and the regional forecast error reduction by the analysis $e_{t|0}^{f}/e_{t|-6}^{f}$ computed for the two globe divisions. Despite our expectation that the statistics should be different for the two globe divisions which are quite different in the uniformity of the areas of each cell, and also for different latitude ranges because the underlying dynamics is different, our results presented in Section 5.3 suggest that the statistics are not very different for these choices.

Based on this observation, we extracted ~ 200 cases, for each of the two lead times (6 and 24 hours) and for each of the $30^\circ\times30^\circ$ and Y_{12}^{6} divisions, by applying the "2σ-criteria" which select only the cases where normalized regional forecast errors $e_{t|0}^{f}/\left\langle e_{t|0}^{f}\right\rangle$ and the regional forecast error reduction by the analysis $e_{t|0}^{f}/e_{t|-6}^{f}$ both exceed its mean by more than 2 standard deviations. We then performed regional EFSO to see if the "2σ-criteria" can capture the "possible regional dropouts," namely, the cases in which the forecasts are expected to be improved by rejecting the observation types which are estimated by EFSO to have negatively impacted the forecast.

The results of regional EFSOs suggest that the "2σ-criteria" can reasonably capture the regional dropouts, especially with 6-hour lead time and the $30^\circ\times30^\circ$ cells, with about 7% (= 15/219) of all the selected cases exhibiting estimated forecast improvements exceeding 25%. We also found that the regional forecast error reduction by the analysis is better correlated with the estimated forecast improvements than the normalized regional forecast errors are, albeit with less confidence due to the limited sample size.

With the "2σ-criteria" applied to the $30^\circ \times 30^\circ$ cells and a 6-hour lead time, we were able to narrow down the regions to apply EFSO to only 1–2 regions per cycle, but yet retain 15 cases whose estimated forecast improvements are above 25%. Thus, the "2σ-criteria" seem to be a reasonable compromise for a need to capture as many regional forecast dropouts as possible with as few executions of EFSO computation as possible. High values of the estimated regional forecast improvements that are found in the extracted cases, however, do not necessarily guarantee that we can improve the forecast by not using the "flawed" observations detected by the EFSO: it can be guaranteed only by conducting a data denial experiment in which the detected "flawed" observations are removed from the analysis, and then assessing if the forecast really improves. In the next chapter, we conduct such data denial experiments and demonstrate that the forecasts can be improved this way.

Chapter 6: Data denial experiments

6.1 Introduction

In Chapter 4, we confirmed that EFSO can be applied not only to a pure EnKF DA system but also to an EnKF within a hybrid DA system and that 6-hour EFSO can yield results that are consistent with 24-hour EFSO. In Chapter 5, we further confirmed that potential "regional forecast dropouts" can be detected after only 6 hours from analysis. The rather prudent assessment we made in the previous two chapters has cleared all but the last one of the concerns we raised in Section 2.4. In this chapter, we finally address the last issue to be answered, *how best to choose the observations to be denied?*, by performing, for each potential "dropout" case, eight sets of data denial experiments with different data denial strategies. Here, by data denial experiments, we denote experiments in which the analyses are repeated without using the observations that are identified as "flawed" by EFSO diagnostics, and then the forecasts are repeated from the (supposedly improved) new analyses. If the forecasts from the new analyses are more accurate than those from the original (control) analyses, then that demonstrates that our Proactive QC can indeed improve the quality and reliability of the forecasts by minimizing the occurrences of regional "forecast skill dropouts."

Due to the limitation of computational resources, however, we cannot perform data denial experiments to all the 219 cases we extracted in Chapter 5. Thus, we first select 20 cases for which the data denial experiments are to be performed, with the criteria we describe in Section 6.2. The number of cases we examine, 20, is determined from the computational resources that we can afford: as we perform eight different data denial experiments for each of the 20 cases, the data denial experiments amount to a total of 160 analyses and forecasts, which is roughly equivalent in terms of computational cost to a one-month cycling experiment (including the spin-up period).

After presenting the 20 selected cases in Section 6.2, we examine how best to choose which and how much of the observations to deny. We first inspect the validity of the assumption tacitly made in the selection algorithm of Ota et al. (2013), *i.e.*, that the observations with large impacts should appear in localized regions (c.f., Section 2.4.5), by looking at geographical (horizontal) and vertical distribution of the EFSO impacts of individual observations. Then, we examine the statistical distribution of EFSO impacts from individual observations and propose eight strategies for selection of the observations to be denied. As we describe in Section 6.3, the eight strategies also include those based on 24-hour EFSO, by which we intend to see which of the 6-hour and 24-hour EFSO is more effective in improving the forecasts. In Section 6.4, we finally show the results of the data denial experiments, in particular, how the forecasts are improved (or degraded) by the denial of the "flawed" observations. As we show in Section 6.4, the results are extremely promising, with significant forecast improvements for all but two out of

20 cases.

6.2 Selection of cases

As we described in the previous section, data denial experiments are expensive and we cannot afford to perform them for all the 219 potential "regional dropout" cases we extracted in Chapter 5. For this reason, we picked up 20 notable cases from them with the following procedure:

1. Select the case with the largest "estimated forecast improvement" (which are estimated for 6-hour forecasts using 6-hour EFSO; see Section 5.4).

2. From the cases of the date of the selected case, search for regions that overlap or adjacent to the selected region. If such regions are found, merge them to form a single region. Put the merged region, along with the date, into the list of selected cases.

3. Exclude the selected regions from the "candidate" list and repeat Steps 1. and 2. to select the second case.

4. Repeat Steps 1.–3. until you get the 20th case.

The selected 20 cases are summarized in Tables 6.1 and 6.2. If two or more $30^\circ \times 30^\circ$ regions are merged, we performed regional 6-hour EFSO again on the merged region. The observation types that are identified by 6-hour EFSO as "flawed," namely, the observation types whose net EFSO impacts are negative (*i.e.*, positively valued), are shown in the fifth column. The "estimated improvements" we defined in Section 5.4 (the sum of the EFSO values from the observation types shown in the fifth column divided by the regional 6-hour forecast error measured with the moist total energy norm, $e^f_{t|0}$, then multiplied by 100) are also shown in the sixth column. If the target region shown in the third and fourth columns is formed by merging more than one $30\,^\circ\!\times\!30\,^\circ$ regions in the Step 2. of the above procedure, then the largest value among the un-merged $30\,^\circ\!\times\!30\,^\circ$ regions is shown. For comparison, we also performed 24-hour EFSO on the selected regions. The "flawed" observation types and the "estimated improvement" evaluated using 24-hour EFSO are shown in the seventh and eighth columns.

Among the 20 cases listed in Tables 6.1 and 6.2, the case #17 deserves special attention because it is exactly the case for which Ota et al. (2013) found $\sim$ 30% regional improvement by the rejection of MODIS wind observation by using 24-hour EFSO. As we will show in Section 6.4, we also obtained a significant forecast improvement for this case by the rejection of "flawed" observations. Consistent with their results, MODIS wind is identified as "flawed," both with 6-hour EFSO and 24-hour EFSO. The 24-hour "estimated improvement" for this case #17 (66%) is the largest in the 20 cases. Furthermore, (perhaps coincidentally), the region (60 °N–90 °N, 40 °E–100 °E) is exactly the same as what Ota et al. (2013) identified. The only difference from Ota et al. (2013) is that, with our method, Aircraft is also identified as "flawed" observation type, whereas, in Ota et al. (2013), only MODIS wind was judged "flawed."

We can observe from Tables 6.1 and 6.2 that all the potential major dropout cases occurred in the vicinity of the Poles, and that, for most of them, MODIS wind observations are identified as “flawed” by 6-hour EFSO. 24-hour EFSO also identified MODIS wind as “flawed,” albeit less frequent than 6-hour EFSO. This suggests that MODIS wind observations in the polar areas, either the observation itself or the way they are assimilated, such as the thinning or the prescribed error variance, might have had some problem for this particular period. This deduction is further corroborated by the results of data denial experiments we show in Section 6.4.

Case #	Date	Latitude	Longitude	6-hour EFSO Detected Types	Est. Imp.	24-hour EFSO Detected Types	Est. Imp.
1	Jan-12-00Z	90 °S–60 °S	110 °E–140 °E	257,258,259(MODIS)	16%	257,258,259(MODIS)	1.7%
2	Jan-12-18Z	60 °N–90 °N	140 °E–180 °	257,258,259(MODIS)	20%	130,133(Aircraft), 257,258,259(MODIS)	19%
3	Jan-13-06Z	60 °N–90 °N	70 °W– 20 °W	257,258,259(MODIS)	30%	130,131,133,230,231,233(Aircraft)	34%
4	Jan-14-18Z	45 °N–90 °N	120 °E–150 °E	4(GPSRO), 230,231(Aircraft), 257,258(MODIS)	10%	4(GPSRO), 130,131,133,230,231,233(Aircraft)	17%
5	Jan-15-18Z	60 °N–90 °N	10 °E– 80 °E	257,259(MODIS)	21%	131,133,231,233(Aircraft)	7%
6	Jan-17-18Z	60 °N–90 °N	50 °W– 0 °	4,42(GPSRO), 131,133(Aircraft), 257(MODIS)	18%	4,42,722–744(GPSRO), 130,131,133,230,231,233(Aircraft)	18%
7	Jan-18-06Z	90 °S–60 °S	70 °W– 30 °W	4,42(GPSRO), AMSU-A,IASI	13%	4,42,740–745(GPSRO), 257,258,259(MODIS)	25%
8	Jan-18-18Z	45 °N–90 °N	120 °E–150 °E	257,258,259(MODIS)	15%	257,258,259(MODIS)	7%
9	Jan-26-18Z	60 °N–90 °N	40 °E– 80 °E	257,258,259(MODIS)	36%	AMSU-A	5%
10	Jan-27-00Z	60 °N–90 °N	30 °E– 80 °E	4(GPSRO), 131,133,231(Aircraft), 180(Marine), 257,258,259(MODIS)	38%	131,133,231,233(Aircraft), 180(Marine), 257,258,259(MODIS)	22%

Table 6.1: List of the cases for which the data denial experiments are performed (continued on Table 6.2). The fifth column (labeled "6-hour EFSO detected types") shows the types of observations (stattype for conventional data and sensor name for radiance data; see Section 3.5) whose net 6-hour EFSO impacts for the region shown in the third and fourth columns are negative (*i.e.*, positive values). The sixth column (labeled "Est. Imp.") shows the "estimated improvement" defined in Section 5.4 computed for the 30 °×30 ° cell using 6-hour EFSO. If the target region is formed by coalescing two or more 30 °×30 ° regions, then the maximum value among them is shown. The seventh and eight columns are the same as the fifth and sixth columns but are based on 24-hour EFSO.

Case #	Date	Latitude	Longitude	6-hour EFSO Detected Types	Est. Imp.	24-hour EFSO Detected Types	Est. Imp.
11	Jan-27-00Z	60 °N–90 °N	20 °W– 10 °E	131,230(Aircraft), ASI,AIRS	11%	130,131,133,230,231,233(Aircraft)	12%
12	Jan-28-18Z	60 °N–90 °N	50 °E– 90 °E	257,258,259(MODIS)	48%	257,258,259(MODIS)	39%
13	Feb-02-18Z	60 °N–90 °N	40 °E–110 °E	120(Radiosonde), 257,258,259(MODIS)	61%	120,220(Radiosonde), 257,258,259(MODIS)	47%
14	Feb-03-00Z	60 °N–90 °N	60 °E– 90 °E	180,280(Marine), 257,258,259(MODIS)	10%	180(Marine), 257,258,259(MODIS)	11%
15	Feb-04-00Z	60 °N–90 °N	40 °W– 10 °W	42(GPSRO), 131(Aircraft)	11%	4,42,720,722,740(GPSRO), 120,220(Radiosonde), 130,131,230,231,233(Aircraft)	48%
16	Feb-05-12Z	90 °S–60 °S	60 °W– 0 °	257,259(MODIS)	33%	257,258,259(MODIS)	23%
17	Feb-06-18Z	60 °N–90 °N	40 °E–100 °E	131(Aircraft), 257,258,259(MODIS)	39%	131,133,231,233(Aircraft), 257,258,259(MODIS)	66%
18	Feb-06-18Z	90 °S–60 °S	60 °W– 10 °E	257,258,259(MODIS), 706–721(Ozone)	23%	4,740–745(GPSRO), 257,258,259(MODIS), 706–721(Ozone)	26%
19	Feb-09-06Z	60 °N–90 °N	140 °W– 90 °W	4,42(GPSRO), 257,258,259(MODIS), HIRS	47%	4,42,722,740,744(GPSRO), AMSU-A,HIRS,IASI,AIRS	42%
20	Feb-10-06Z	90 °S–60 °S	50 °E– 80 °E	257,259(MODIS)	10%	257,258,259(MODIS)	19%

Table 6.2: List of the cases for which the data denial experiments are performed. Continued from Table 6.1.

6.3 Selection of observations to be denied

In Sections 1.4.2, 1.7 and 2.4 we repeatedly posed the question: *given the information from 6-hour EFSO, how can we best choose the observations to be denied?* To answer this question, we first need to answer how we should "prioritize" the observation rejection; namely, we need to be able to answer, given any two observations, rejection of which of the two observations is more beneficial. Second, we need to answer how much we should reject: as we described in Section 1.4.2, rejecting too many observations would lead to forecast degradation, but rejecting too few observations would make little difference. Thus, we have to find a way to strike the best balance. This section addresses these questions.

Let us first examine the answer of Ota et al. (2013) to the first question. They answered this question by the data-selection algorithm which we described in detail in the Step (3) in Section 2.3.1. As we discussed in Section 2.4.5, this very intricate algorithm implicitly assumes that observations with large negative impact should be clustered in horizontally and vertically localized regions. It is not clear, however, if this assumption is justifiable with real data. Here, we investigate the validity of this assumption by looking at geographical and vertical distributions of the EFSO impacts of individual observations.

We have already shown a typical example of such distributions in Figure 4.6 when we discussed the consistency between 6-hour and 24-hour EFSO. Now, we examine the same figure from a different angle: *Are observations with large positive/negative impacts localized?*, or: *Are observations with positive impacts well* *separated from those with negative impacts?*

From Figure 4.6, we can see that, the observations with positive impacts and those with negative impacts are not well separated: for any observation with large negative impact, we can easily find observations with positive impacts in its vicinity, both vertically and horizontally. Visual inspections for other cases and other observation types (not shown), albeit rather subjective, all support our claim above. Our finding is consistent with Sommer and Weissmann (2014) who found, by applying the EFSO formulation of Kalnay et al. (2012) to a convective-scale regional LETKF DA system, that positively impacting observations are spatially intertwined closely with negatively impacting observations and thus are not localized. From the above inspection, we conclude that, it is more appropriate to choose the observations solely based on their EFSO values rather than to classify them based on their geographical and vertical locations.

Next, let us consider the second question: *how many should we reject?* To answer this question, we examine the statistical distribution of EFSO values, for each case and for each of the identified "flawed" observation types. For each observation type, we sort the observations based on their EFSO values and plot the EFSO values against the rank. In choosing the observations to be denied, we aim to reduce the forecast errors as much as possible by rejecting as few observations as possible. Thus, if we can find a "jump," *i.e.*, a steep slope or a discontinuity at which the EFSO value suddenly becomes large, it seems reasonable to put the threshold there.

Three typical examples of such plots are shown in Figure 6.1. The three panels

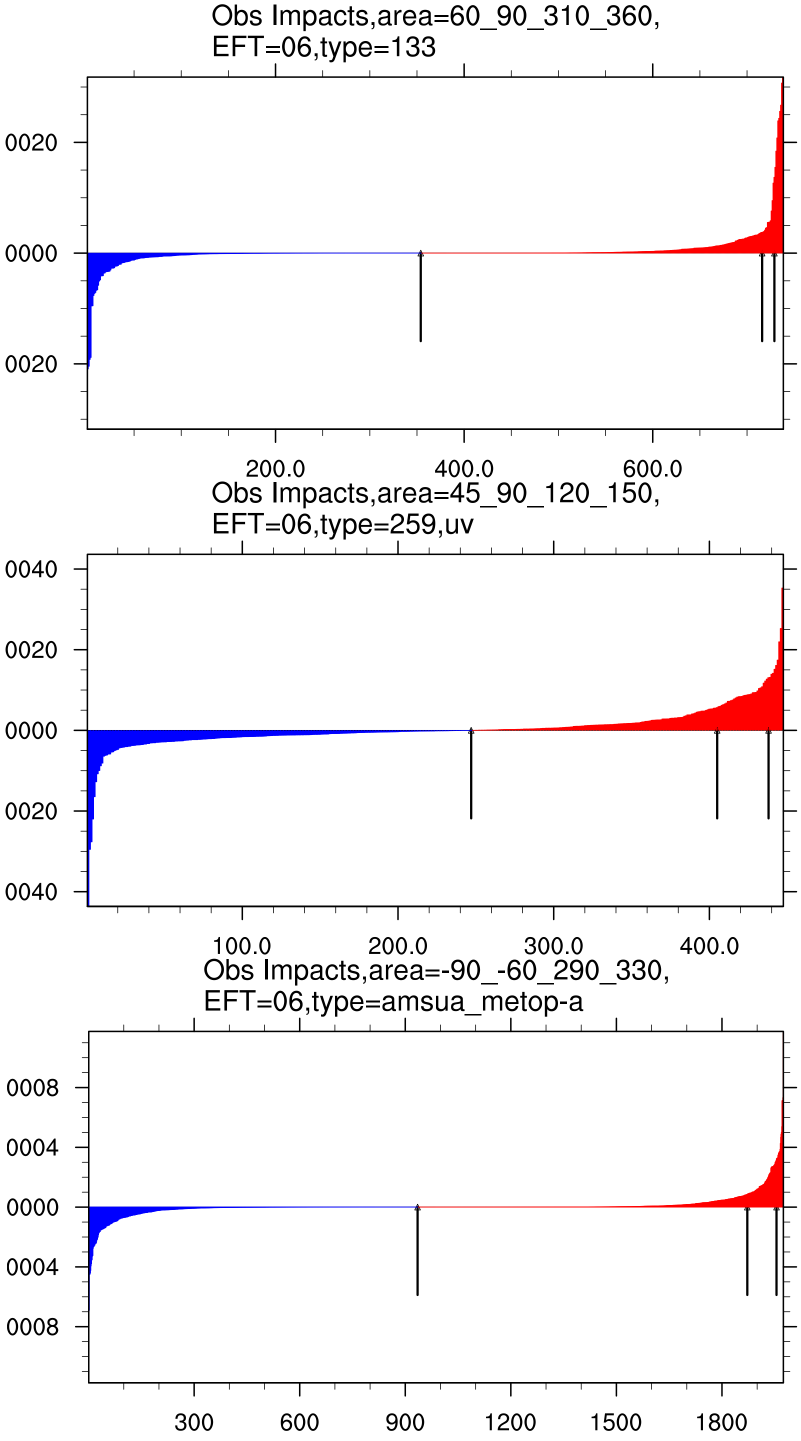


Figure 6.1: The 6-hour EFSO values of individual observations plotted against their ranks. Shown are (top) the observations of stattype 133 (Aircraft) for Case #6, (middle) the observations of stattype 259 (MODIS wind) for Case #8, and (bottom) AMSU-A observations from MetOp-A satellite for Case #7. Positive and negative impacts (or negative and positive values, respectively) are plotted with blue and red colors. The units are 10^{-3}J kg^{-1}. The three arrows represent, from left to right, the thresholds for rejection of "allneg," "one-sigma" and "netzero" criteria.

show the EFSO values of individual observations of (top) stattype 133 (Aircraft temperature and humidity observations reported by commercial airplanes through ACARS (Aircraft Communications Addressing and Reporting System)) for Case #6, (middle) stattype 259 (MODIS wind inferred from deep layer cloud of water vapor channel) for Case #8, and (bottom) AMSU-A observations from MetOp-A satellite for Case #7. Note that, as we mentioned in the caption of Figure 4.6, in computing EFSO statistics of wind observations, we combined the impacts from the zonal and meridional wind components (u and v, respectively), which are assimilated separately in our DA system.

In the top panel, we can locate a clear "jump" near the right edge of the plot. For this kind of distribution, it is easy to choose a threshold. Unfortunately, however, in most of the cases we examined, such clear "jumps" could not be found. In the middle and bottom panels of Figure 6.1, the EFSO values are distributed more continuously. For these distributions, it seems difficult to objectively determine the best threshold above which the observations can be regarded "outliers."

Since it seems difficult to objectively determine the threshold, we decided to try three simple criteria for determining thresholds and see which of them works best by performing data denial experiments to each of them. For comparison's sake, we also tried posing no threshold at all; namely, reject all observations whose types are judged "flawed" by EFSO. We call this criterion "allobs."

We summarize the three criteria, along with the "allobs" criterion we described above, in the following list:

allobs Remove all observations of the detected type within the target region, *regardless of the EFSO values of each observation*

allneg Remove all observations of the detected type within the target region *that had negative EFSO impacts (positive values)*

one-sigma Remove observations of the detected type within the target region *whose EFSO values were above the mean of the observations of the same type by at least one standard deviation (σ)*

netzero For each of the detected observation type, sort the observations of the detected type within the target region based on the EFSO impacts, and remove observations from the one with largest negative impacts (positive values) until the net impact of that type becomes zero (neutral)

The "netzero" criterion is a modest approach which just tries to prevent all observation types from collectively degrading the forecast. Ota et al. (2013) adopted this conservative approach and obtained successful results. This modest approach, however, may limit the full capacity of EFSO diagnostics. The idea behind our Proactive QC, or any conventional QC methods, is to detect and remove "outliers" which DA system cannot properly handle. With this idea in mind, we could determine which observations to reject based on whether each observations are "outliers" in terms of the statistics of their EFSO impacts, which led us to design the "one-sigma" criterion. The "allneg" criterion entirely trusts the results of EFSO and removes all observations that had negative impacts. Because EFSO diagnostics are subject to sampling errors, we can consider this approach to be a rather dangerous strategy because we can possibly reject observations that are actually helpful. For these reason, when we designed these criteria, we expected "one-sigma" criterion to be most successful.

In Figure 6.1, the threshold obtained by the "allneg," "one-sigma" and "netzero" criteria are shown by the vertical arrows. The leftmost one which points the rank at which the EFSO value becomes zero represent the thresold of "allneg"; the second leftmost one represents that of "one-sigma"; and the rightmost one represents that of "netzero." In a highly skewed distribution with clear discontinuity, the thresholds obtained by "netzero" and "one-sigma" criteria tend to be close to each other, as we can see from the top panel. Thus, for such cases, the two criteria are expected to yield similar results.

In the next section, we show results of data denial experiments in which the observations to be denied are determined by the four criteria introduced above. In order to compare the effectiveness of Proactive QC based on 6-hour EFSO (which we believe is feasible in an operational system) and that based on 24-hour EFSO (which would not be feasible in an operational real-time system but still could be exploited in a reanalysis), we also performed data denial experiments based on 24-hour EFSO.

The number of observations that are denied by each of the eight criteria are summarized in Table 6.3. As we can see by comparing the number of denied observations for "allobs" and "allneg," the "allneg" criteria denies about a half of the observations within the target region whose type is judged as "flawed." This is consistent with the fact that about a half of all the observations have negative impacts (c.f. the discussion in Figure 4.5). Note that, they are still a tiny portion of

Case	6-hour				24-hour			
#	allobs	allneg	one-sigma	netzero	allobs	allneg	one-sigma	netzero
1	1488	968	182	326	894	484	106	52
2	2292	1174	242	110	2421	1281	253	133
3	2842	1714	224	344	13776	7103	910	284
4	3827	2126	352	270	1498	778	108	22
5	3328	1714	246	118	480	247	36	36
6	9360	4430	230	22	4169	2014	109	24
7	31491	15690	867	67	10281	5694	785	385
8	3654	1816	320	138	3654	1978	318	130
9	2330	1510	296	510	2988	1509	88	5
10	3278	1720	204	89	4355	2256	181	52
11	27832	13726	375	32	28498	14160	510	75
12	3830	2282	526	462	3830	2102	436	152
13	6416	3936	720	908	6470	3310	388	60
14	481	234	34	26	183	64	1	2
15	966	508	23	11	16187	8489	747	287
16	6956	3544	522	174	6956	3634	556	212
17	5915	3326	616	415	7492	3857	679	355
18	6238	3276	622	366	5014	2643	448	324
19	8504	4678	749	809	39459	19624	660	38
20	1216	598	128	48	1216	668	116	84

Table 6.3: The number of denied observations for each case and each criterion.

the total number of the assimilated observations; our quasi-operational DA system assimilates about $\sim 3 \times 10^6$ observations per cycle. As we expected, the number of the denied observations generally decreases in the following order: allobs, allneg, one-sigma, netzero. Note that the three cases #1, #3 and #9 are exceptions; in these cases, more observations are denied in “netzero” than in “one-sigma.”

6.4 Results

6.4.1 Verification method

In the development of NWP systems, a common practice for measuring success of a new scheme is to compare standardized scores of global or hemispheric scales. Typical examples of such standardized scores include root mean square error (RMSE) or spatial anomaly correlation coefficient (ACC) computed for the globe, NH or SH extratropics, or the tropical belt (e.g., 20 °S–20 °N, 0 °–360 °) of some meteorological elements that are familiar to synoptic meteorologists such as 500 hPa geopotential height (Z500) or Mean Sea-Level Pressure (MSLP). We do not follow this practice, however, because we are interested to know whether we can minimize *regional* forecast failures by Proactive QC; the commonly used global or hemispheric scale scores would obscure local impact of data denial and hence may not be suitable for our verification purpose.

For the verification of the effectiveness of data denial, we compute the local "relative forecast improvement" which we define by the following:

First, we divide the globe into 10 °×10 ° patches. Then, for each of these 10 °×10 ° patches, we compute the forecast error measured with the moist total energy norm restricted to that 10 °×10 ° region verified against the control GSI analysis. We compute this scalar forecast error for each of the two forecasts, one initialized by the original analysis (before the data denial or Proactive QC), the other initialized by the new analysis obtained by rejecting the identified "flawed" observations, and denote them, respectively, by $e_{t|0}^{f,\text{beforeQC}}$ and $e_{t|0}^{f,\text{afterQC}}$. The relative forecast improvement for each 10 °×10 ° patch is defined using these two forecast errors as

$$\text{relative forecast improvement} := \frac{e_{t|0}^{f,\text{beforeQC}} - e_{t|0}^{f,\text{afterQC}}}{e_{t|0}^{f,\text{beforeQC}}} \times 100\ [\%] \tag{6.1}$$

Although our main focus is to capture local improvement of forecast, it is also important to make sure that the rejection of the observations does not make the forecast error measured at a larger scale worse. Thus, we also computed the "average improvement" which is also defined by Eq. (6.1) but with the scalar forecast errors $e_{t|0}^{f,\text{beforeQC}}$ and $e_{t|0}^{f,\text{afterQC}}$ computed for the NH extratropics (40 °N–90 °N, 0 °–360 °; if the target region is in the NH) or the SH extratropics (40 °S–90 °S, 0 °–360 °; if the target region is in the SH).

6.4.2 Case study (1): Examination of Case #17

We now proceed to show the results of data denial experiments. First, we show the results for an individual case, taking Case #17 as an example. This is the case for which Ota et al. (2013) obtained remarkable forecast improvement by the denial of MODIS wind observations. As they did, we also obtained great forecast improvement by the denial of "flawed" observations. This case is also one of the typical examples of our results and the features we will point out in this subsection are also valid for many of the other cases.

The relative forecast improvement for Case #17 obtained by the data denial experiments with the denial criteria based on 6-hour EFSO is shown in Figure 6.2.

In these contour plots, improvement and degradation of forecast are shown, respectively, with blue and red colors. As we can see from the "allobs" column, if we deny all observations of the types judged "flawed" by 6-hour EFSO (stattypes 257–259 (MODIS wind) in this case), the forecast is improved in some regions but is also degraded in other regions, although the degraded regions become smaller as the forecast lead time increases. Thus, rejecting all observations of "flawed" type regardless of the EFSO values of the individual observations is not a good strategy. If we remove only the observations that had negative impacts ("allneg"), we can effectively eliminate most of the degraded regions. Remarkably, with "allneg" criterion, the forecast improvement locally reaches as much as 48% for 24-hour forecast. The fact that "allneg" yields better forecast improvement than "allobs," and that it dramatically reduces the forecast degradation, proves the effectiveness of EFSO diagnostics. However, we still see some regions of degradation, for example north of Siberia close the North Pole at 12-hour lead time. By further restricting the denied observations by only denying observations whose EFSO values exceed the mean by more than one standard deviation ("one-sigma"), we can completely eliminate forecast degradation; however, this is achieved at the expense of diminished forecast improvement. If we more selectively restrict the denied observations ("netzero"), the forecast improvement becomes even smaller.

The results for data denial experiments based on 24-hour EFSO are shown in Figure 6.3. Basically the results are similar to those based on 6-hour EFSO. The biggest difference is that, for some reason, with 24-hour EFSO, the "one-sigma" criterion does not result in improvement or degradation. Interestingly, data denial based on 6-hour EFSO results in slightly better forecast improvement than that based on 24-hour EFSO, although the difference is very small. In the next subsection, we show a summary of how the forecast is locally improved or degraded in all the cases we examined. We demonstrate there that our success with Case #17 is not just sheer luck (that is, there are other cases that had comparably significant improvements).

6.4.3 Summary of the results

We summarize the results of data denial experiments in Tables 6.4 and 6.5 by showing, for each case and criterion for data denial, the largest positive value of the local relative improvement of 24-hour forecast ("max.imp."), the largest negative value of the local relative improvement of 24-hour forecast (or the largest relative forecast degradation ;"max.deg."), and the average 24-hour forecast improvement evaluated for the extratropics of the hemisphere in which the target region is located.

For "allneg," "one-sigma" and "netzero" criteria based on 6-hour EFSO, the average improvement ("avg.imp.") is positive in almost all cases. The only exceptions are "allneg" of Case #5 and #9, where we had degradation, respectively, of 0.2% and 0.4%. Because Proactive QC is designed to minimize the occurrences of *local* forecast failures, its impact is spatially localized and thus, if averaged over a large spatial extent such as the hemisphere, the impact becomes small. This is why the average improvement is at most $\sim 2\%$. Note that, although the improvement of the order of $\sim 0.2\%$ to $\sim 2\%$ might seem to be quite modest, in the development

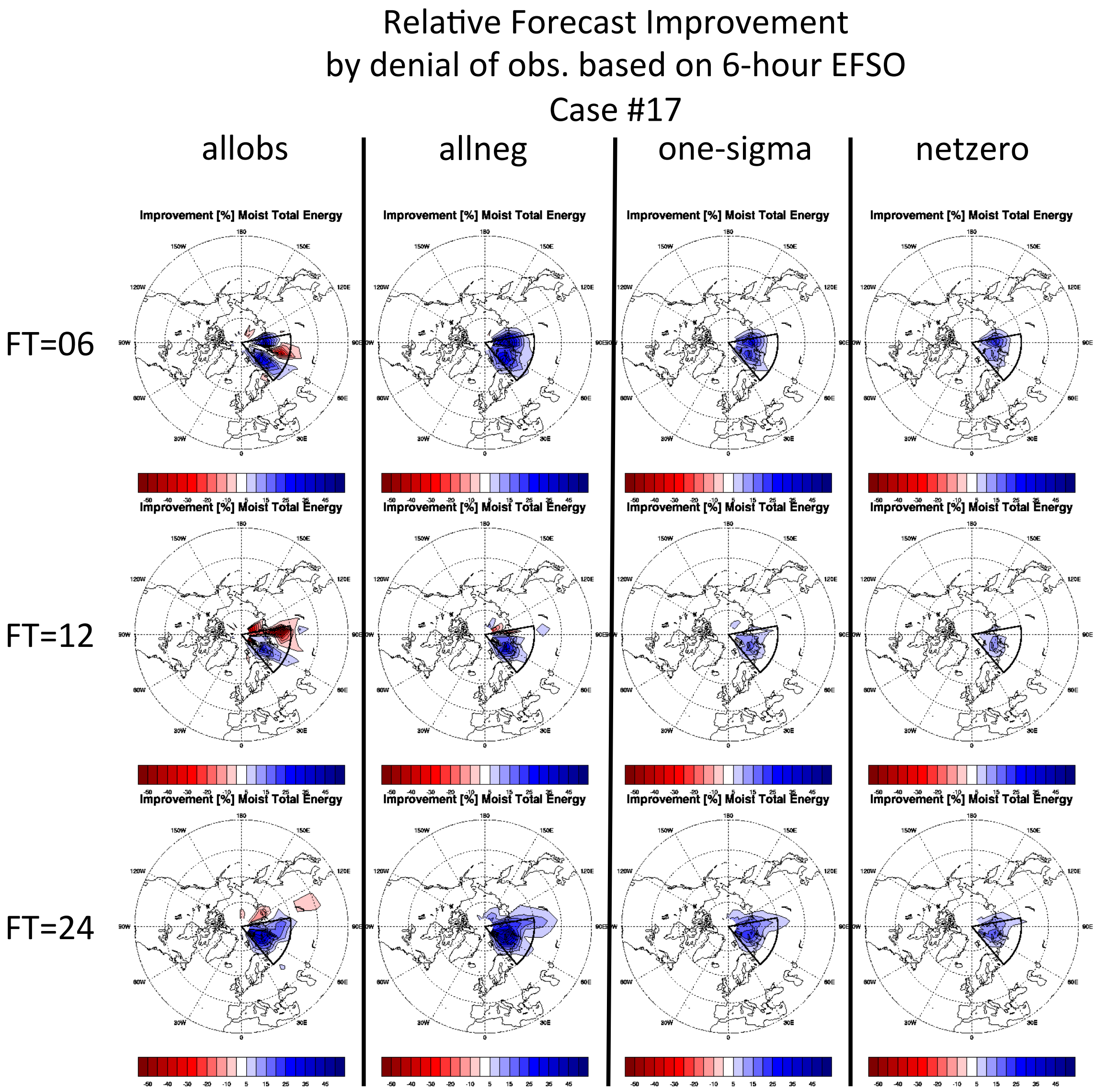


Figure 6.2: Relative forecast improvement for each of the four data rejection criteria based on 6-hour EFSO, for Case #17. Each column represents, from left to right, the "allobs," "allneg," "one-sigma" and "netzero" criteria. The first row represents the relative improvement of 6-hour forecast; the second and third rows represent, respectively, the improvement of 12 and 24 hour forecasts. Red colors represent forecast degradation; blue colors represent forecast improvement. The thick black cone represents the target region.

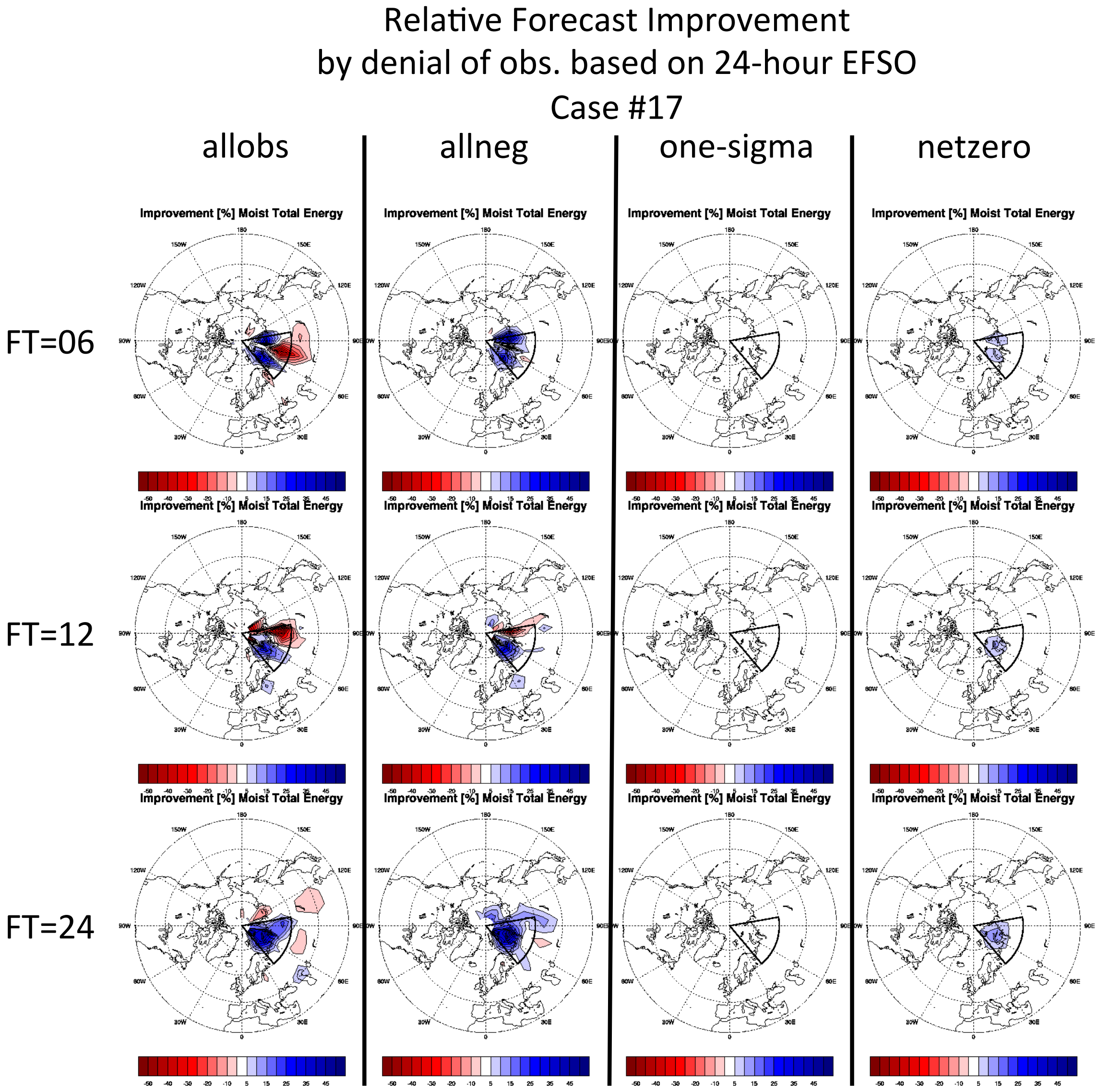


Figure 6.3: As in Figure 6.2, but with observation denial based on 24-hour EFSO.

of an operational NWP system, even such seemingly modest improvement is very difficult to obtain.

The features that we saw with the case study of Case #17 is also valid for most other cases, namely: (1) “allobs” exhibit both improvement and degradation, (2) “allneg” alleviates the the degradation seen in “allobs” and tend to show larger improvement, and (3) “one-sigma” and “netzero” further reduce the degradation, but they also tend to reduce the improvement. Also, except in Case #11 and #15, data denial based on 6-hour EFSO yielded better forecast improvement than that based on 24-hour EFSO.

Encouragingly, the large forecast improvement that we saw for Case #17 is not limited only to this particular case. For example, if we look at “allneg” criterion based on 6-hour EFSO, the cases #8, #12, #13, #16, #17, #18 and #19 all exhibit local maximum forecast improvement that exceed 30%. For these cases, “one-sigma” and “netzero” criteria also result in large maximum local forecast improvement ($\sim$ 20%). For all of these particularly successful cases, 6-hour EFSO identified MODIS wind as the flawed observation type. This suggests that either the observations from MODIS wind had anomalous errors or the way the DA system handles them is faulty.

Case		6-hour				24-hour			
#		allobs	allneg	one-sigma	netzero	allobs	allneg	one-sigma	netzero
1	max.imp.	12%	11%	4%	5%	12%	20%	0%	6%
	max.deg.	-9%	-1%	-1%	-1%	-9%	-1%	-1%	0%
	avg.imp.	0.0%	0.2%	0.1%	0.1%	0.0%	0.3%	-0.0%	0.1%
2	max.imp.	14%	11%	8%	4%	2%	10%	1%	2%
	max.deg.	-5%	-4%	-2%	0%	-45%	-5%	0.5%	0%
	avg.imp.	-0.1%	0.3%	0.2%	0.2%	-2%	0.1%	0%	0.1%
3	max.imp.	13%	7%	2%	4%	7%	12%	0%	2%
	max.deg.	-15%	-5%	-1%	-2%	-8%	-7%	0%	-3%
	avg.imp.	0.0%	0.2%	0.0%	0.0%	-0.1%	0.1%	0.0%	-0.1%
4	max.imp.	25%	27%	15%	13%	3%	4%	1%	0%
	max.deg.	-5%	-5%	-2%	-2%	-6%	-3%	-5%	0%
	avg.imp.	0.6%	0.7%	0.3%	0.2%	-0.3%	0.1%	-0.3%	-0.0%
5	max.imp.	15%	19%	23%	22%	12%	10%	1%	1%
	max.deg.	-32%	-81%	-30%	-13%	-78%	-21%	-1%	-1%
	avg.imp.	-0.2%	-0.2%	0.2%	0.3%	-1.3%	-0.4%	-0.0%	-0.0%
6	max.imp.	9%	15%	12%	3%	24%	9%	2%	3%
	max.deg.	-9%	-6%	-3%	-1%	-38%	-10%	-2%	-2%
	avg.imp.	0.0%	0.4%	0.3%	0.1%	0.0%	0.1%	0.0%	0.0%
7	max.imp.	17%	13%	2%	0%	19%	26%	0%	4%
	max.deg.	-9%	-5%	-3%	0%	-36%	-28%	0%	-1%
	avg.imp.	-0.0%	0.4%	0.0%	0.0%	0.3%	0.6%	0.0%	0.2%
8	max.imp.	41%	41%	21%	10%	41%	26%	0%	4%
	max.deg.	-18%	-14%	-5%	-2%	-18%	-10%	0%	-1%
	avg.imp.	0.9%	1.1%	0.8%	0.4%	0.9%	1.2%	-0.0%	0.2%
9	max.imp.	7%	8%	8%	8%	3%	5%	3%	3%
	max.deg.	-21%	-16%	-3%	-4%	-2%	-1%	-1%	-1%
	avg.imp.	-0.6%	-0.4%	0.0%	0.1%	-0.1%	0.1%	0.0%	0.0%
10	max.imp.	25%	19%	3%	6%	21%	17%	4%	2%
	max.deg.	-6%	-6%	-2%	0%	-5%	-12%	0%	0%
	avg.imp.	1.1%	0.7%	0.2%	0.2%	0.8%	0.8%	0%	0.2%

Table 6.4: Relative improvement or degradation of 24-hour forecast by the denial of observations. The first of the three rows of each case (labeled “max.imp.”) shows the largest positive value of the local “relative forecast improvement.” The second row (labeled “max.deg.”) shows the largest negative value of the local “relative forecast improvement” (*i.e.*, the largest local “relative forecast degradation”). The last row (labeled “avg.imp.”) shows the average forecast improvement evaluated for the extratropics of the hemisphere in which the target region is located. Continued on Table 6.5.

Case		6-hour				24-hour			
#		allobs	allneg	one-sigma	netzero	allobs	allneg	one-sigma	netzero
11	max.imp.	11%	9%	2%	3%	22%	15%	1%	2%
	max.deg.	-6%	-5%	-2%	0%	-5%	-6%	0%	0%
	avg.imp.	0.5%	0.3%	0.1%	0.1%	0.9%	0.9%	0.0%	0.2%
12	max.imp.	37%	39%	19%	19%	37%	38%	1%	12%
	max.deg.	-14%	-12%	-2%	-2%	-14%	-19%	0%	-6%
	avg.imp.	0.7%	0.7%	0.5%	0.5%	0.7%	0.4%	0.0%	0.2%
13	max.imp.	24%	30%	18%	19%	24%	26%	0%	8%
	max.deg.	-9%	-9%	-10%	-12%	-9%	-10%	0%	-6%
	avg.imp.	1.4%	0.8%	0.3%	0.4%	1.3%	1.1%	0.0%	0.1%
14	max.imp.	5%	3%	1%	1%	5%	3%	0%	0%
	max.deg.	0%	0%	0%	0%	0%	-1%	0%	0%
	avg.imp.	0.3%	0.1%	0.0%	0.1%	0.3%	0.1%	0.0%	0.0%
15	max.imp.	3%	1%	1%	1%	13%	35%	1%	7%
	max.deg.	-2%	-1%	-1%	-1%	-16%	-18%	-1%	-10%
	avg.imp.	0.1%	0.1%	-0.0%	0.0%	-0.1%	0.8%	0.0%	0.2%
16	max.imp.	27%	30%	23%	16%	30%	33%	1%	7%
	max.deg.	-15%	-21%	-4%	-2%	-20%	-43%	-1%	-1%
	avg.imp.	1.9%	1.8%	1.3%	0.7%	2.1%	1.2%	0.0%	0.3%
17	max.imp.	39%	48%	26%	20%	45%	51%	0%	15%
	max.deg.	-15%	-4%	-2%	-2%	-15%	-8%	-1%	-2%
	avg.imp.	0.8%	2.1%	1.2%	0.8%	0.7%	1.6%	-0.0%	0.5%
18	max.imp.	46%	46%	25%	21%	36%	47%	0%	20%
	max.deg.	-9%	-8%	-3%	-2%	-14%	-13%	-1%	-4%
	avg.imp.	2.4%	2.2%	1.0%	0.8%	1.6%	2.1%	-0.0%	0.6%
19	max.imp.	44%	37%	17%	14%	6%	8%	0%	2%
	max.deg.	-24%	-10%	-1%	-1%	-17%	-7%	0%	-1%
	avg.imp.	2.2%	2.2%	1.0%	1.0%	-0.2%	0.2%	0.0%	0.0%
20	max.imp.	12%	10%	5%	3%	12%	9%	1%	9%
	max.deg.	-3%	-1%	-1%	-1%	-3%	-2%	-2%	-1%
	avg.imp.	0.2%	0.3%	0.2%	0.0%	0.2%	0.2%	-0.0%	0.2%

Table 6.5: Continued from Table 6.4.

6.4.4 Case study (2): the unsuccessful case

We have seen, in the previous section, that in 18 out of 20 cases, we can in fact improve the forecast by denying observations that are identified by 6-hour EFSO as "flawed." This, we believe, is a remarkable achievement. However, in the two cases, #5 and #9, the denial of observations based on 6-hour or 24-hour EFSO failed to improve the forecast. It should be instructive to look into what happened in these cases. In this subsection, we examine the results for Case #5. We note that the results for Case #9 (not shown) exhibited similar features.

The patterns of relative 6-hour and 24-hour forecast improvement are shown, respectively, in Figure 6.4 and Figure 6.5. Let us first examine the results for the denial experiments based on 6-hour EFSO (Figure 6.4). By looking at the first row (FT=06), we can observe that, with "allneg," "one-sigma" and "netzero" criteria, the 6-hour forecast is actually improved within the target area (note that "allobs" does not use the information from EFSO of individual observations so we do not expect to gain any improvement with this method). Thus, the 6-hour EFSO is actually very accurate in the sense that the linear estimation from EFSO and the actual nonlinear forecast impact are very consistent. However, in "allneg," there is a forecast degradation of 8.7% at the 10 °×10 ° patch of (10 °E–20 °E, 70 °N–80 °N), which is just at the border of the boundary of the target area, although it is not clearly visible in this plot. As the forecast lead time gets longer, the improvement is attenuated (FT=12) and at FT=24, forecast degradation appears for "allneg." Note that, although 24-hour forecast is degraded with "allneg," the degradation can be avoided with the more conservative “one-sigma” and “netzero” criteria. We speculate that the forecast degradation observed in “allneg” is due to the EFSO’s sampling errors; because EFSO is estimated from a limited samples from the ensemble, it is inevitably subject to sampling errors and can mistakenly assign (perhaps small) negative impacts (positive value) to observations that are actually beneficial.

From Figure 6.5, we can observe that 24-hour EFSO is again very accurate in the sense that 24-hour forecast inside the target area is actually improved by the negatively contributing observations (see the panel of “allneg” and “FT=24”). What we did not expect is that, the forecast outside the target area is degraded at the same time. This means that the observations that are estimated by EFSO to have degraded the forecast within the target area was helping the forecast outside of the target area. Again, this forecast degradation can be avoided by adopting the more conservative “one-sigma” and “netzero” criteria.

The lesson we learn from this inspection is that, when we apply regional EFSO, we should keep it in mind that, it might assign negative impacts to observations that are beneficial for the forecast outside the target area.

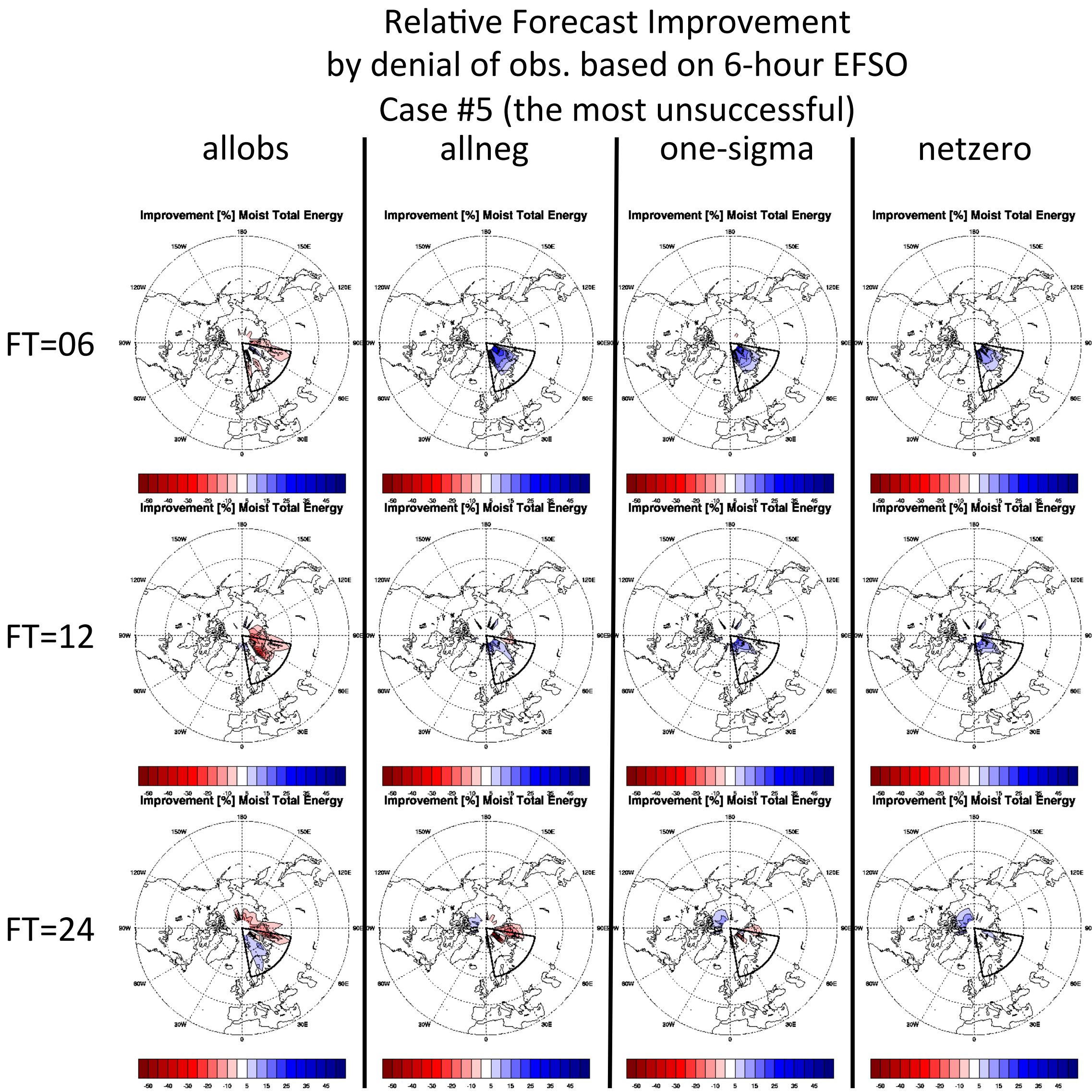

Relative Forecast Improvement
by denial of obs. based on 6-hour EFSO
Case #5 (the most unsuccessful)
allobs
allneg
one-sigma
netzero
FT=06
FT=12
FT=24
Improvement [%] Moist Total Energy
180
150W
150E
120W
120E
90W
90E
60W
60E
30W
30E
0
-50
-40
-30
-20
-10
5
15
25
35
45

Figure 6.4: As in Figure 6.2, but for Case #5.

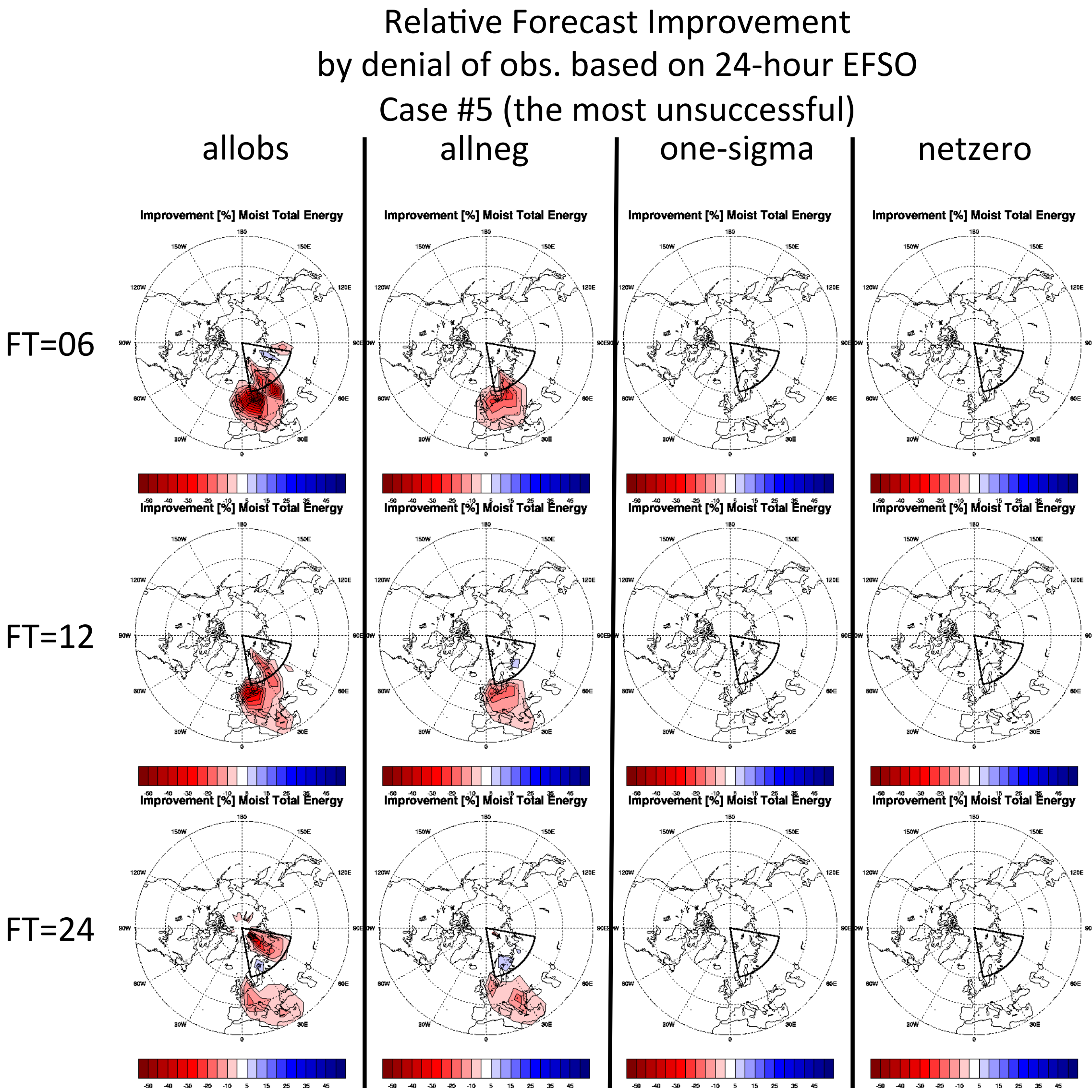

Relative Forecast Improvement
by denial of obs. based on 24-hour EFSO
Case #5 (the most unsuccessful)
allobs
allneg
one-sigma
netzero
FT=06
FT=12
FT=24
Improvement [%] Moist Total Energy
180
150W
150E
120W
120E
90W
90E
60W
60E
30W
30E
0
-50
-40
-30
-20
-10
5
15
25
35
45

Figure 6.5: As in Figure 6.3, but for Case #5.

6.5 Summary

The last and most important two questions we posed in Section 1.7, namely,

3. What is the best threshold for rejection of “flawed” observations? and
4. Does rejection of detected “flawed” observation really improve analysis and forecast?

are finally answered in this chapter. To answer these two questions, we performed data denial experiments to selected 20 cases with four different criteria for choosing the threshold for rejection of “flawed” observations. In the “allobs” criterion, all observations of the types that are identified by EFSO as “flawed” are rejected, regardless of the EFSO impacts of individual observations; in the “allneg” criterion, only the observations that had negative EFSO impacts are rejected; in the “onesigma” criterion, the observations whose EFSO values exceed the mean by at least one standard deviation are rejected; in the “netzero” criterion, observations are rejected, from the one with largest negative EFSO impact to that with the smallest, until the net impact from the selected observation type becomes neutral. These four criteria form a spectrum of “strictness,” with “allobs” the loosest and “netzero” the most selective. In order to compare the effectiveness of Proactive QC based on 6-hour EFSO (as we propose) and that based on 24-hour EFSO, we also performed data denial experiments with the criteria based on 24-hour EFSO as well.

The results we obtained are extremely encouraging; with the “allneg” criterion, we obtained hemisphere-scale forecast improvement in 18 out of 20 cases. Furthermore, in seven of the 18 successful cases, the local maximum forecast improvement reached over 30%. The forecast degradation found in the two unsuccessful cases can be made neutral if we adopt more conservative “one-sigma” or “netzero” criteria. Unfortunately, however, we found that the reduction of degradation also comes with reduction of improvement. Interestingly and encouragingly, we found that the data denial based on 6-hour EFSO is equally (or slightly more) effective in reducing 24-hour forecast errors compared to that based on 24-hour EFSO. This is interesting because it suggests that 6-hour EFSO, which estimates the observational impacts on 6-hour forecast, can provide more useful information on how to improve 24-hour forecast than 24-hour EFSO, which estimates the observational impacts on the same 24-hour forecast.

With these evidences, we believe that we can give affirmative answer to the question 4: rejection of “flawed” observations indeed improves analysis and forecast. In order to answer the question 3., we might need additional investigation. We confirmed that the forecast degradation can be suppressed by the “one-sigma” criterion but it sacrifices some of the achieved forecast improvement. Thus, we can expect that some looser threshold could still suppress the degradation and lower the reduction of forecast improvement. We point out, however, that this is a matter of compromise and the decision should ultimately be made in a subjective manner. Our personal preference is a rather conservative threshold which does not make the forecast worse anywhere; this is because, with Proactive QC, we are trying to enhance the reliability of NWP.

We believe that the results we obtained in this chapter are good enough for it to be considered for operational implementation. There still remain some technical issues to be addressed before the actual implementation to a high-resolution operational system. We will discuss more on such issues in Chapter 9.

As we mentioned in the previous section, the seven most successful cases all obtained significant forecast improvement by rejecting subsets of MODIS wind observations. This means that if our Proactive QC is implemented to the operational system, we can identify observation instruments that actually made forecast worse on a real-time basis. We can then use such information to improve the conventional off-line QC, for example, by improving the blacklist (c.f. Section 2.1). Or, we could even encourage developers of instrument to look into any possible faults of their algorithms by providing appropriate metadata to them. This motivates us to consider an additional application of EFSO and Proactive QC. We will discuss this subject in Chapter 9.

Chapter 7: Ensemble Forecast Sensitivity to observation error covariance (EFSR) I: Formulation and experiments with Lorenz '96 system

7.1 Introduction

As we outlined in Section 1.5, the observation error covariance **R** is an external parameter for DA systems that are typically prescribed somewhat empirically and subjectively. Thus, In order to optimized the performance of a DA system, the observation error covariance matrix **R** needs to be tuned. Daescu (2008) and Daescu and Langland (2013) showed that it is possible, with the adjoint technique analogous to the FSO of LB04, to derive an expression that estimates how a small change in each element of **R** matrix will change the forecast errors $e^f_{t|0}$. In this chapter, we show that an ensemble equivalent to the adjoint diagnostics of Daescu and Langland (2013) can be easily formulated by using the approximation proposed by Kalnay et al. (2012) in deriving the EFSO, which we described in Section 2.2.2. For succinctness, hereafter, we call this diagnostics FSR, short for Forecast Sensitivity to observation error covariance matrix **R**; we refer to the adjoint-based and ensemble-based FSR, respectively, by AFSR and EFSR.

We first derive the adjoint formulation of FSR following Daescu and Langland (2013), and present our ensemble formulation. We then test the effectiveness of the AFSR and EFSR diagnostics using a simple, toy system called Lorenz '96 model. Encouraged by the successful results with this toy system, we next implement and test the EFSR diagnostics using the lower-resolution version of the NCEP's operational global NWP system which we used for our Proactive QC study. Finally, we perform a simple **R**-tuning experiment based on the results of EFSR, and assess if the EFSR-based tuning improves the impacts of observations to the forecast by comparing EFSOs before and after the tuning. The setup and results of the quasi-operational experiments are described in Chapter 8.

7.2 Adjoint formulation following Daescu and Langland (2013)

In this section we derive a formulation of the forecast sensitivity to **R** matrix following Daescu and Langland (2013) using the notation we introduced in Chapter 2.

Consider a DA problem for time 0. Our goal is to obtain an approximate expression for how the t-hour forecast error $e^f_{t|0}$ defined by Eq. (2.8) changes with a small variation in the observation error covariance matrix from $\mathbf{R}$ to $\mathbf{R}+\mathbf{R}'$.

Since the Kalman gain matrix $\mathbf{K}$ can be written as

$$\mathbf{K} = \mathbf{P}^b_0\mathbf{H}^T\left(\mathbf{H}\mathbf{P}^b_0\mathbf{H}^T+\mathbf{R}\right)^{-1}, \tag{7.1}$$

the analysis equation Eq. (2.1) can be split into two equations

$$\begin{cases} \mathbf{x}_0^a - \mathbf{x}_0^b & = \mathbf{P}_0^b\mathbf{H}^T\mathbf{w}_0 \\ \left(\mathbf{H}\mathbf{P}_0^b\mathbf{H}^T + \mathbf{R}\right)\mathbf{w}_0 & = \delta\mathbf{y}_0^{ob} \end{cases} \tag{7.2}$$

where $\mathbf{w}_0$ is an intermediate analysis increment in the observation space. In some literature, this formulation is referred to as Physical Space Analysis Scheme (PSAS; Da Silva et al., 1995; Kalnay, 2003, Section 5.5.2). Now, let us introduce a variation in $\mathbf{R}'$ to the observation error covariance $\mathbf{R}$, and denote variations in other variables by $\mathbf{x}_0^{a\prime}$ and $\mathbf{w}_0{}'$. The analysis equations Eq. (7.2) then becomes

$$\begin{cases} \mathbf{x}_0^a + \mathbf{x}_0^{a\prime} - \mathbf{x}_0^b = \mathbf{P}_0^b\mathbf{H}^T\left(\mathbf{w}_0 + \mathbf{w}_0{}'\right) \\ \left(\mathbf{H}\mathbf{P}_0^b\mathbf{H}^T + \mathbf{R} + \mathbf{R}'\right)\left(\mathbf{w}_0 + \mathbf{w}_0{}'\right) = \delta\mathbf{y}_0^{ob} \end{cases} \tag{7.3}$$

By neglecting a second-order term $\mathbf{R}'\mathbf{w}_0'$ and by subtracting Eq. (7.2), we have,

$$\begin{cases} \mathbf{x}_0^{a\prime} = \mathbf{P}_0^b\mathbf{H}^T\mathbf{w}_0{}' \\ \left(\mathbf{H}\mathbf{P}_0^b\mathbf{H}^T + \mathbf{R}\right)\mathbf{w}_0{}' + \mathbf{R}'\mathbf{w}_0 \approx 0 \end{cases} \tag{7.4}$$

Thus, the change in the analysis $\mathbf{x}_0^{a\prime}$ caused by $\mathbf{R}'$ can be expressed as

$$\mathbf{x}_0^{a\prime} \approx \mathbf{P}_0^b\mathbf{H}^T\left\{-\left(\mathbf{H}\mathbf{P}_0^b\mathbf{H}^T + \mathbf{R}\right)^{-1}\right\}\mathbf{R}'\mathbf{w}_0 \tag{7.5}$$

$$= -\mathbf{K}\mathbf{R}'\mathbf{w}_0 \qquad \because (7.1) \tag{7.6}$$

Now we show that the intermediate analysis increment $\mathbf{w}_0$ has a simple expression.

By applying the Jacobian $\mathbf{H}$ of the observation operator H to the first of PSAS equations Eq. (7.2), we have

$$\begin{aligned} \mathbf{H}\left(\mathbf{x}_0^a - \mathbf{x}_0^b\right) &= \mathbf{H}\mathbf{P}_0^b\mathbf{H}^T\mathbf{w}_0 = \left(\mathbf{H}\mathbf{P}_0^b\mathbf{H}^T + \mathbf{R}\right)\mathbf{w}_0 - \mathbf{R}\mathbf{w}_0 && (7.7) \\ &= \delta\mathbf{y}_0^{ob} - \mathbf{R}\mathbf{w}_0 \quad \because \text{the second of PSAS equations (7.2)} && (7.8) \\ \therefore\ \mathbf{w}_0 &= \mathbf{R}^{-1}\left(\delta\mathbf{y}_0^{ob} - \mathbf{H}\left(\mathbf{x}_0^a - \mathbf{x}_0^b\right)\right) && (7.9) \\ &\approx \mathbf{R}^{-1}\left\{\delta\mathbf{y}_0^{ob} - \left(H(\mathbf{x}_0^a) - H(\mathbf{x}_0^b)\right)\right\} = \mathbf{R}^{-1}\delta\mathbf{y}_0^{oa} && (7.10) \end{aligned}$$

where $\delta\mathbf{y}_0^{oa} = \mathbf{y}_0^o - H(\mathbf{x}_0^a)$ is the misfit of observation with respect to the analysis. Thus, we can approximate $\mathbf{w}_0$ with the observation-minus-analysis (O-A) multiplied from left by the inverse of the observation error covariance $\mathbf{R}^{-1}$.

Finally, we derive an expression for how the forecast error $e_{t|0}^f$ defined by Eq. (2.8) changes by the variation $\mathbf{R}'$ in $\mathbf{R}$. The first-order approximation to the variation in $e_{t|0}^f$ is

$$\begin{aligned} e_{t|0}^{f\;\prime} &= (\mathbf{x}_0^{a\prime})^T \frac{\partial e_{t|0}^f}{\partial \mathbf{x}^a} = (-\mathbf{K}\mathbf{R}'\mathbf{w}_0)^T \frac{\partial e_{t|0}^f}{\partial \mathbf{x}^a} && \because (7.6) \qquad (7.11) \\ &= -\left(\mathbf{R}'\mathbf{w}_0\right)^T \mathbf{K}^T \frac{\partial e_{t|0}^f}{\partial \mathbf{x}^a} = -\left(\mathbf{R}'\mathbf{w}_0\right)^T \frac{\partial e_{t|0}^f}{\partial \mathbf{y}_0^o} && \because (2.1) \qquad (7.12) \end{aligned}$$

By taking derivative with respect to (i, j)-element of $\mathbf{R}'$, we have

$$\frac{\partial e_{t|0}^f}{\partial R_{i,j}} = -(\mathbf{w}_0)_j \left(\frac{\partial e_{t|0}^f}{\partial \mathbf{y}_0^o}\right)_i \tag{7.13}$$

One way to compute the forecast error sensitivity vector with respect to observations

$\frac{\partial e^f_{t|0}}{\partial \mathbf{y}^o_0}$ is to "reuse" it from the adjoint FSO of LB04 (Eq. (2.35)):

$$\begin{aligned} \frac{\partial e^f_{t|0}}{\partial \mathbf{y}^o_0} &= \frac{\partial \left(e^f_{t|-6} + \Delta e^2 \right)}{\partial \mathbf{y}^o_0} && (7.14) \\ &= \frac{\partial e^f_{t|-6}}{\partial \mathbf{y}^o_0} + \frac{\partial \left(\Delta e^2 \right)}{\partial \mathbf{y}^o_0} = 0 + \frac{\partial \left(\Delta e^2 \right)}{\partial \mathbf{y}^o_0} && (7.15) \\ &= \mathbf{K}^T \mathbf{M}^T_{t|0} \mathbf{C} \left(\mathbf{e}^f_{t|0} + \mathbf{e}^f_{t|-6} \right) && (7.16) \end{aligned}$$

Hereafter, we refer to this formulation as "AFSR-REUSE" (Adjoint FSR with the sensitivity gradient vector REUSEd from FSO computation). This AFSR-REUSE formulation has a particular advantage of being computationally economical, provided that the FSO is already computed, because we do not have to perform the adjoint computation again. One might argue, however, that this approximation may not be accurate because the formulation of LB04 for the sensitivity gradient $\frac{\partial \left(\Delta e^2 \right)}{\partial \mathbf{y}^o_0}$ is evaluated for two trajectories with different initial conditions at time 0, one being the analysis $\mathbf{x}^a_0$, and the other being the background $\mathbf{x}^b_0$, in order to yield a second-order estimate of the observational impacts Δe^2 (e.g. Todling, 2013). Daescu and Langland (2013) proposes instead to compute the gradient $\frac{\partial e^f_{t|0}}{\partial \mathbf{y}^o_0}$ by evaluating it at the analysis field $\mathbf{x}^a_0$ alone:

$$\begin{aligned} \frac{\partial e^f_{t|0}}{\partial \mathbf{y}^o_0} &= \mathbf{K}^T \frac{\partial e^f_{t|0}}{\partial \mathbf{x}^a_0} && \because (2.1) && (7.17) \\ &= \mathbf{K}^T \mathbf{M}^T_{t|0} \frac{\partial e^f_{t|0}}{\partial \mathbf{x}^f_{t|0}} && \because \ \mathbf{x}^f_{t|0} = M_{t|0} \left(\mathbf{x}^a_0 \right) && (7.18) \\ &= \mathbf{K}^T \mathbf{M}^T_{t|0} \cdot 2\mathbf{C}\mathbf{e}^f_{t|0} && \because (2.8) && (7.19) \end{aligned}$$

Hereafter, we refer to this formulation as “AFSR-NEW” (Adjoint FSR with the sensitivity gradient vector NEWly computed). This formulation requires a new adjoint computation.

7.2.1 Sensitivity to tuning factors

In tuning the observation covariance matrix $\mathbf{R}$, it is customary to classify observations into some groups for which we can assume that there is no inter-group correlations in the observation errors, and scale the error covariances within each group by a single common factor. Daescu and Langland (2013) also derived a formulation for forecast sensitivity to these scaling factors. Let the observations $\mathbf{y}_0^o$ be partitioned into I subgroups $\left\{\mathbf{y}_{0,i}^o, i = 1, \cdots, I\right\}$ where $\mathbf{y}_{0,i}^0 \in \mathbb{R}^{q_i}$, $\quad \sum_i q_i = p$, and scale each of them by the scaling factors $s_i^o,\ i = 1, \cdots, I$:

$$\mathbf{R}_i \to s_i^o \mathbf{R}_i \tag{7.20}$$

where $\{\mathbf{R}_i \in \mathbb{R}^{q_i \times q_i},\ i = 1, \cdots, I\}$ are the diagonal sub-blocks of $\mathbf{R}$ (note that we assume no correlation of errors between any two observations of different groups). The scaling factors s_i^o are positive non-dimensional scalars. Then, from Eq. (7.12) and Eq. (7.10), the forecast sensitivity to these scaling factors are

$$\begin{aligned}\frac{\partial e_{t|0}^f}{\partial s_i^o} &= -\left(\mathbf{R}_i \mathbf{w}_{0,i}\right)^T \frac{\partial e_{t|0}^f}{\partial \mathbf{y}_{0,i}^o} && (7.21)\\ &= -\left(\delta \mathbf{y}_{0\ ,i}^{oa}\right)^T \frac{\partial e_{t|0}^f}{\partial \mathbf{y}_{0,i}^o} && (7.22)\end{aligned}$$

where $\mathbf{w}_{0,i}$ is a sub-vector of $\mathbf{w}_0$ corresponding to $\mathbf{y}^o_{0,i}$. Eq. (7.22) means that, for any group of observations, the sensitivity of forecast errors to the scaling factor of observation error covariance of that group can be computed as the observation-minus-analysis (O-A) residual multiplied by forecast sensitivity gradient to the corresponding observation, summed up over all observations in that group (with the sign flipped), which is very simple to compute.

7.3 Ensemble formulation based on EFSO

The ensemble formulation immediately follows from the adjoint formulation and the Kalman gain approximation in Eq. (2.4). From equations Eq. (7.10) and Eq. (7.13), we have,

$$\frac{\partial e^f_{t|0}}{\partial R_{i,j}} = -\left(\mathbf{R}^{-1}\delta\bar{\mathbf{y}}^{oa}_0\right)_j \left(\frac{\partial e^f_{t|0}}{\partial \mathbf{y}^o_0}\right)_i \tag{7.23}$$

The sensitivity vector $\frac{\partial e^f_{t|0}}{\partial \mathbf{y}^o_0}$ can be expressed either by the "AFSR-REUSE" formulation Eq. (7.16) as

$$\begin{aligned} \frac{\partial e^f_{t|0}}{\partial \mathbf{y}^o_0} &= \left(\mathbf{M}_{t|0}\mathbf{K}\right)^T \mathbf{C}\left(\mathbf{e}^f_{t|0} + \mathbf{e}^f_{t|-6}\right) && (7.24) \\ &\approx \frac{1}{K-1}\mathbf{R}^{-1}\mathbf{Y}^a_0\mathbf{X}^{fT}_{t|0}\mathbf{C}\left(\mathbf{e}^f_{t|0} + \mathbf{e}^f_{t|-6}\right) \quad \because (2.4) \text{ or } (2.20) && (7.25) \end{aligned}$$

or by the“AFSR-NEW” formulation Eq. (7.19) as

$$\frac{\partial e^f_{t|0}}{\partial \mathbf{y}^o_0} = \left(\mathbf{M}_{t|0}\mathbf{K}\right)^T \cdot 2\mathbf{C}\mathbf{e}^f_{t|0} \tag{7.26}$$

$$\approx \frac{1}{K-1}\mathbf{R}^{-1}\mathbf{Y}^a_0\mathbf{X}^{fT}_{t|0} \cdot 2\mathbf{C}\mathbf{e}^f_{t|0} \quad \because (2.4)) \tag{7.27}$$

We refer to the former formulation Eq. (7.25) by “EFSR-REUSE” and the latter Eq. (7.25) by “EFSR-NEW.” Note that in deriving the above two formulations, we have used the approximations that we used in deriving EFSO (Eq. (2.18)). Here again, covariance localization is necessary if the ensemble size K is not sufficiently large, giving localized formulations

$$\frac{\partial e^f_{t|0}}{\partial \mathbf{y}^o_0} \approx \frac{1}{K-1}\mathbf{R}^{-1}\left[\rho \circ \left(\mathbf{Y}^a_0\mathbf{X}^{fT}_{t|0}\right)\right]\mathbf{C}\left(\mathbf{e}^f_{t|0} + \mathbf{e}^f_{t|-6}\right) \tag{7.28}$$

or

$$\frac{\partial e^f_{t|0}}{\partial \mathbf{y}^o_0} \approx \frac{1}{K-1}\mathbf{R}^{-1}\left[\rho \circ \left(\mathbf{Y}^a_0\mathbf{X}^{fT}_{t|0}\right)\right] \cdot 2\mathbf{C}\mathbf{e}^f_{t|0} \tag{7.29}$$

The forecast sensitivity to tuning factors Eq. (7.22) is also valid for these ensemble-based formulations, with the sensitivity vector replaced by either of the two expressions given above.

7.3.1 Sensitivity to the covariance inflation factor

It is interesting to note that we can estimate, from our EFSR formulation, the forecast sensitivity to the multiplicative covariance inflation factor. Consider scaling the background and the observation error covariances $\mathbf{P}^b$ and $\mathbf{R}$ with a same scalar scaling factor s':

$$\mathbf{P}^b \to s'\mathbf{P}^b, \quad \mathbf{R} \to s'\mathbf{R} \tag{7.30}$$

From the expression for Kalman gain matrix Eq. (7.1) we know that Kalman gain $\mathbf{K}$ does not change by this scaling. Thus, analysis $\mathbf{x}_0^a$ is kept unchanged and therefore, the forecast error $e_{t|0}^f$ also does not change by this scaling.[1]

Now, consider scaling $\mathbf{P}^b$ by s^b and the sub-matrices of $\mathbf{R}$, $\{\mathbf{R}_i,\ i = 1, \cdots, I\}$ by $\{s_i^o,\ i = 1, \cdots, I\}$. The variation of the forecast error $e_{t|0}^f{}'$ caused by the variations in scaling factors, $s^{b\prime}$ and $s_1^{o\prime}, \cdots, s_I^{o\prime}$, can be written as

$$e_{t|0}^f{}' = \frac{\partial e_{t|0}^f}{\partial s^b} s^{b\prime} + \sum_{i=1}^{I} \frac{\partial e_{t|0}^f}{\partial s_i^o} s_i^{o\prime} \tag{7.31}$$

As we saw in the previous paragraph, if the variations $s^{b\prime}$, $s_1^{o\prime}, \cdots, s_I^{o\prime}$ are all the same (let us denote it by s'), then the resulting change in the forecast error must

[1] Note that, although scaling $\mathbf{P}^b$ by a factor s and scaling $\mathbf{R}$ by a factor $1/s$ are equivalent for a single analysis, they may not be equivalent for a cycled system because scaling of $\mathbf{P}^b$ can accumulate over cycles, while scaling of $\mathbf{R}$ does not.

be zero. Thus,

$$0 = \frac{\partial e^f_{t|0}}{\partial s^b} s' + \sum_{i=1}^{I} \frac{\partial e^f_{t|0}}{\partial s^o_i} s' \tag{7.32}$$

$$\Rightarrow \quad \frac{\partial e^f_{t|0}}{\partial s^b} = -\sum_{i=1}^{I} \frac{\partial e^f_{t|0}}{\partial s^o_i} \tag{7.33}$$

Noting that the scaling factor for the background covariance s^b can be interpreted as a globally constant multiplicative inflation factor, we see that the equation Eq. (7.33) tells us that the forecast sensitivity to inflation factor can be estimated as the sum of forecast sensitivity to observation error covariance tuning factors (with the sign flipped). We comment that Daescu and Langland (2013) gives a proof of this equation by directly deriving the expression for $\frac{\partial e^f_{t|0}}{\partial s^b}$ in their Appendix.

7.4 Toy-model experiment with Lorenz '96 system

This section presents the experimental setup and the results of our experiment using the Lorenz '96 model. A brief discussion about the interpretation and implications of the results is presented in the next section. The code used in this experimentation was developed upon the experimental system which Mr. Yoichiro Ota of the JMA developed for an internal training purpose at the JMA led by Prof. Takemasa Miyoshi of the University of Maryland and RIKEN/AICS (then at JMA Hotta and Ota, 2011).

7.4.1 Lorenz '96 system

In our toy-system experiments, we use the Lorenz '96 model as the forecast model. It is a chaotic low-dimensional ODE system first introduced by Lorenz (1996) and Lorenz and Emanuel (1998) to address a fundamental question in predictability and DA study: *In a system with insufficient observations, how can we place a supplementary observation that optimally improves the predictability of the system?* In the literature of predictability and DA study, this model has been widely adopted as a benchmark system for new ideas.

The model is an N-dimensional ODE system defined by

$$\frac{\mathrm{d}x_j}{\mathrm{d}t} = x_j\left(x_{j+1} - x_{j-2}\right) - x_j + F, \quad j = 1, \cdots, N \tag{7.34}$$

with a cyclic boundary condition $x_{-1} = x_{N-1}, x_0 = x_N$ and $x_{N+1} = x_1$. The first term mimics the advection ($-\mathbf{u} \cdot \nabla\mathbf{u}$) of fluid mechanical equations; the second and third terms represent, respectively, dissipation (damping or diffusion) and external forcing, hence making the model a nonlinear, forced-dissipative system, a typical genre of chaotic systems. As in the original study by Lorenz and Emanuel (1998), for our study, we adopt $N = 40$ and $F = 8.0$. Note that, unlike Kalnay et al. (2012) and Liu and Kalnay (2008), who used different values of F for the nature run and DA cycles to account, to some extent, for model errors, we use the same parameter $F = 8.0$ for both the nature run and the forecast model (so-called "identical twin" or "perfect model" setting). The forecast model Eq. (7.34) is integrated by the standard fourth-order Runge-Kutta scheme with time step $\Delta t = 0.01$.

For DA, we adopt the LETKF (Hunt et al., 2007, see Section 3.4) with member size $K = 40$. Since the member size is equal to the dimension of the state space, we did not apply covariance localization. To avoid filter divergence, however, we applied multiplicative covariance inflation (Anderson, 2001) with a constant inflation parameter 1.15^2 at each assimilation cycle. This parameter is not tuned to an optimal value but, as we show in the next section, the system worked well. In fact, we could use Eq. (7.33) to tune it. The cycling interval (the assimilation window) is 0.05 in non-dimensional time. From an analogy with the then-operational NWP models in terms of e-folding time of perturbations, Lorenz and Emanuel (1998) proposes to dimensionalize the time by interpreting 0.2 in non-dimensional time as 24 hours. Hence, our cycling interval in dimensionalized time is 6 hours.

In our experiments, we assimilate observations available at every grid point (*i.e.*, H is the identity function). For the j-th grid point, the observations are generated for every analysis time by adding independent Gaussian pseudo-random numbers with the variance $\sigma_j^{o,\mathrm{true}^2}$. The pseudo-random numbers are generated with the Mersenne Twister algorithm (Matsumoto and Nishimura, 1998) using the Fortran 95 code developed and publicly distributed by Dr. Atsushi Ito of Nagoya University, Japan. The true observation error covariance can thus be assumed to be

$$\mathbf{R}^{\mathrm{true}} = \mathrm{diag}\left(\sigma_1^{o,\mathrm{true}^2}, \sigma_2^{o,\mathrm{true}^2}, \cdots, \sigma_{40}^{o,\mathrm{true}^2}\right) \tag{7.35}$$

Throughout the experiments, the observation error covariance prescribed in the DA system $\mathbf{R}$ is also assumed to be diagonal:

$$\mathbf{R} = \mathrm{diag}\left(\sigma_1^{o2}, \sigma_2^{o2}, \cdots, {\sigma_{40}^{o}}^2\right) \tag{7.36}$$

7.4.2 Experimental design

First, the nature (or the "truth") is produced by running the forecast model Eq. (7.34) from an initial condition randomly generated from the uniform distribution in $[0, 1]$. The nature run is integrated from time $t = 0$ to time $t = 730$ (which corresponds to 10 years in dimensionalized time), generating truth for 14,600 cycles.

The initial background ensemble at time $t = 0$ is generated by picking up 40 truth states at randomly chosen 40 distinctive dates. Each DA experiment is run 14,600 cycles (10 years) and the first 1,460 cycles (one year) are excluded from verification, regarding one year as a spin-up period.

To examine the ability of the adjoint and ensemble FSR diagnostics to detect mis-specification of observation error variances $\left\{{\sigma^{o2}_{j}},\ j = 1, \cdots, 40\right\}$, we conducted three pairs of "identical twin" experiments. Each pair consists of two DA cycle runs, one with correctly specified $\mathbf{R}$ (*i.e.*, identical to the truth; hereafter we refer to this by "correct-**R** run"), the other with incorrectly specified $\mathbf{R}$ (hereafter we refer to this by "incorrect-**R** run"). The true and specified observation error variances, along with the names of the experiments, are summarized in Table 7.1.

The SPIKE experiment is inspired by the experimental settings of Liu and Kalnay (2008) and Kalnay et al. (2012) who examined the capacity of EFSO to capture the negative impact from the observation at the 11-th grid point which have larger observation errors than the others. All observations but the one at the 11-th grid point have the error variance 0.2^2; at the 11-th grid point, it is 0.8^2. In the mis-specified DA run (the incorrect-$\mathbf{R}$ run), they are all prescribed by 0.2^2. With this experiment, we intend to see whether the adjoint or ensemble FSR diagnostics can detect the mis-specification of the error variance at the 11-th grid point to provide useful guidance on how to correct it. We also examine whether the FSR diagnostics do not signal "false alarm" when the specification of $\mathbf{R}$ is correct.

The STAGGERED experiment is designed to assess whether the FSR diagnostics are robust to cases where observations with different magnitude of errors are located close to each other. The true observation errors are 0.1^2 and 0.3^2, respectively, for odd- and even-numbered grid points. In the incorrect-$\mathbf{R}$ run, they are all prescribed by 0.2^2; we should thus reduce/increase the error variances at odd/even grid points.

The LAND-OCEAN experiment is inspired by Lorenz and Emanuel (1998) who simulated data-sparse "ocean" and data-rich "land" using the Lorenz '96 system. In our LAND-OCEAN experiment, however, the density of observations are the same for the "ocean" and the "land"; we, instead, vary the quality of observations, with the observations over the ocean being more accurate: satellite radiance observations are generally more accurate over the ocean than over the land because the surface condition is more uniform over the ocean than over the land. In this experiment, we mimic a situation where we assimilate only satellite radiance observations, whose accuracy is different depending on the surface condition. The observations are relatively more accurate over the "ocean" ($21 \leq j \leq 40$), with the error variance 0.1^2, while, over the "land" ($1 \leq j \leq 20$), the quality of observations are poorer, with the larger error variance 0.3^2. We assume that the NWP developers on this "virtual planet" are not yet aware of this non-uniformity in the observation quality and have been incorrectly using the constant observation error variance 0.2^2 in their DA system (we simulate this situation with the incorrect-$\mathbf{R}$ run) . They now use the FSR techniques to grasp the true magnitude of observation errors.

For each of the three experiments, we compute the FSR vector $\left(\frac{\partial e^f_{t|0}}{\partial \sigma^o_1{}^2}, \cdots, \frac{\partial e^f_{t|0}}{\partial \sigma^o_{40}{}^2}\right)$ with the four different methods, the adjoint based "AFSR-REUSE" (Eq. (7.16)) and "AFSR-NEW" (Eq. (7.19)), and the ensemble based "EFSR-REUSE" (Eq. (7.25)) and "EFSR-NEW" (Eq. (7.27)). As in the DA system, no covariance localization is performed for EFSR estimations. As the forecast lead-time, we adopt 24 hours (0.2 in nondimensional time). For evaluating forecast errors with Eq. (2.6), we use the analysis as the verifying truth $\mathbf{x}^v_t$. The quadratic error norm $\mathbf{C}$ in Eq. (2.8) is the identity matrix in $\mathbb{R}^{40\times 40}$ (*i.e.*, the Euclidian norm).

Name	True obs error variance	Prescribed error variance
SPIKE	$\sigma_j^{o,\text{true}2} = \begin{cases} 0.8^2 & j = 11 \\ 0.2^2 & j \neq 11 \end{cases}$	$\sigma_j^{o2} = 0.2^2$ everywhere
STAGGERED	$\sigma_j^{o,\text{true}2} = \begin{cases} 0.1^2 & j\text{: odd} \\ 0.3^2 & j\text{: even} \end{cases}$	$\sigma_j^{o2} = 0.2^2$ everywhere
LAND-OCEAN	$\sigma_j^{o,\text{true}2} = \begin{cases} 0.3^2 & \begin{array}{c} 1 \leq j \leq 20 \\ \text{(``land'')} \end{array} \\ 0.1^2 & \begin{array}{c} 21 \leq j \leq 40 \\ \text{(``ocean'')} \end{array} \end{cases}$	$\sigma_j^{o2} = 0.2^2$ everywhere

Table 7.1: The true and specified observation error variances for the three experiments performed using the Lorenz ’96 system.

7.4.3 Results

7.4.3.1 The SPIKE experiment

First, as a "sanity check," we show, in Figure 7.1, the analysis errors of the SPIKE experiment. The analysis errors are verified against the truth and their averages over the last 9 years of the cycling experiments are shown. The blue and red lines show, respectively, the analysis errors for the correct-**R** and incorrect-**R** runs.

In the vicinity of the 11-th grid point where the observation has larger errors, the analysis is considerably more accurate for the correct-**R** run (blue line) than for the incorrect-**R** run (red line). Even in the correct-**R** run, the "bad" observation at the 11-th grid point makes the analysis less accurate at that grid point than at elsewhere. Specifying the wrong observation error variance which is smaller than the true value exacerbates the situation by giving higher credence to the "bad" observation than it deserves. Nevertheless, the DA is still successful in the sense that the analysis is more accurate than the observations. This assures that "filter divergence" has not occurred in these DA cycles.

The top and bottom panels of Figure 7.2 show, respectively, the ensemble-based forecast sensitivity gradient to observation error variances, $\frac{\partial e^f_{t|0}}{\partial \sigma^{o\,2}_j}$, estimated using "EFSR-REUSE" and "EFSR-NEW" formulations (see Eqs. (7.23), (7.25) and (7.27)). Apart from the difference in magnitude, the two formulations give similar diagnostics. For the incorrect-**R** run, the two formulations both successfully signal us that we should increase the observation error variance for the "bad" observation at the 11-th grid point (note that a negative sensitivity gradient means that the forecast error decreases (and thus the forecast becomes more accurate) by increasing the observation error variance), while, for the correct-**R** run, they both give virtually zero sensitivity.

For the incorrect-**R** run, despite the fact that the observation error variances for the observations near the "bad" observation are correctly specified, the EFSR diagnostics, both 'the 'EFSR-REUSE" and "EFSR-NEW," tell us that we should decrease the observation error variances for them. Our interpretation for this is as follows:

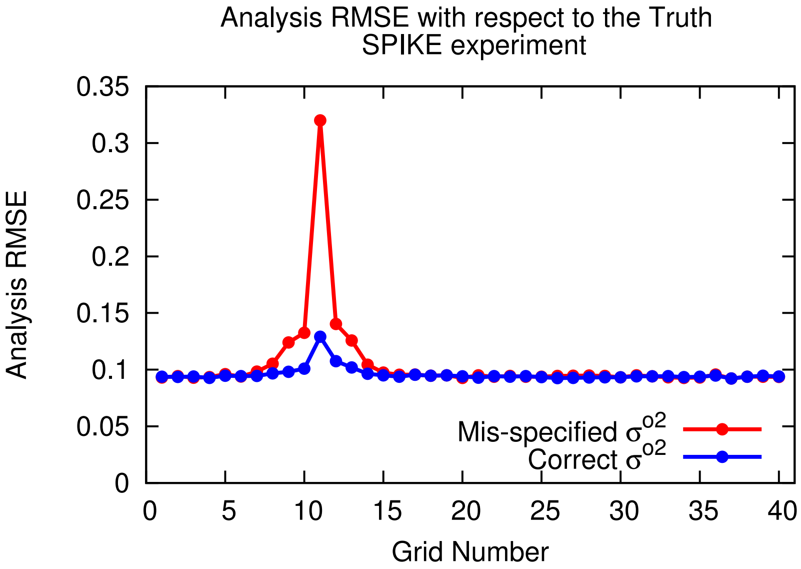


Figure 7.1: Analysis errors verified against the truth for the SPIKE experiment displayed as a function of grid number. The analysis errors for each grid point are averaged over the last 9 years of a 10-year DA cycling. The analysis errors are smaller when the observation error variances are correctly specified (blue line) than when they are mis-specified (red line).

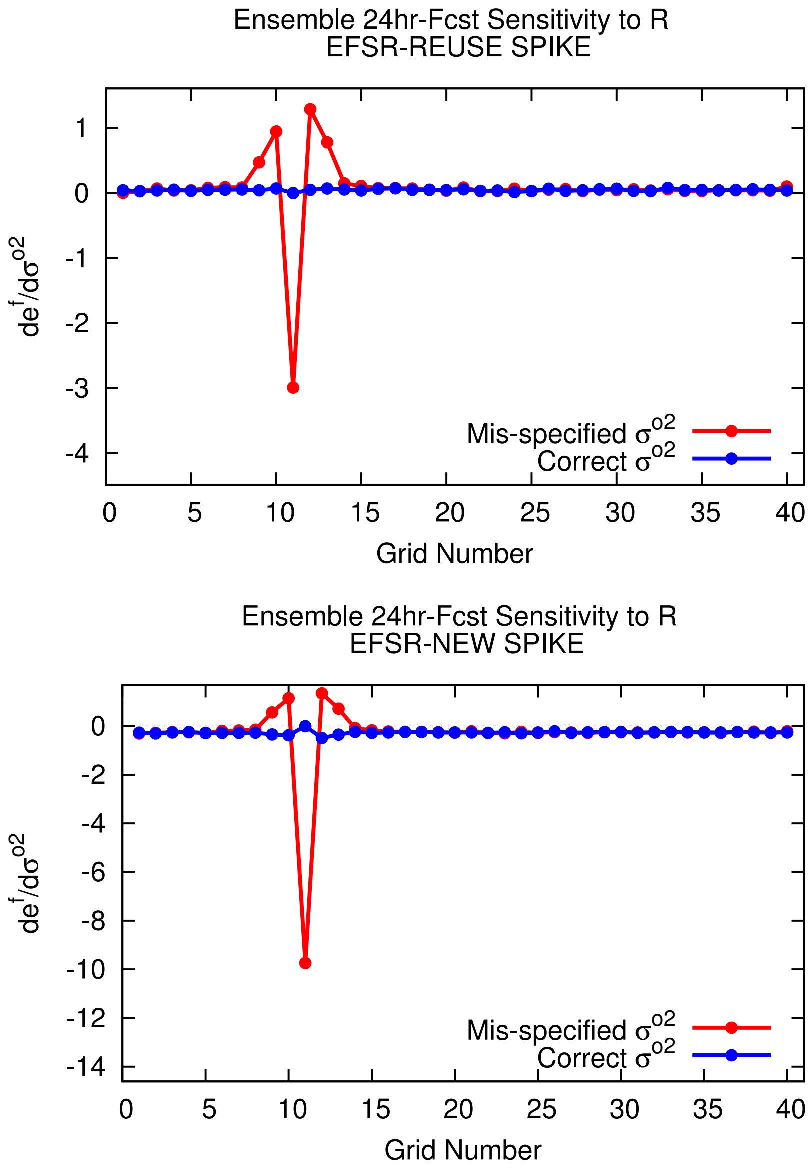


Figure 7.2: The forecast sensitivity gradient to observation error variances $\frac{\partial e^f_{t|0}}{\partial \sigma^{o2}_j}$ for the SPIKE experiment estimated using (top) "EFSR-REUSE" and (bottom) "EFSR-NEW" formulations. As in Figure 7.1, the blue and red lines represent, respectively, the DA runs with correctly specified observation error variances (correct-**R** run) and that with mis-specified observation error variances (incorrect-**R** run). As in Figure 7.1, the sensitivity gradients are averaged over the last 9 years of each DA run.

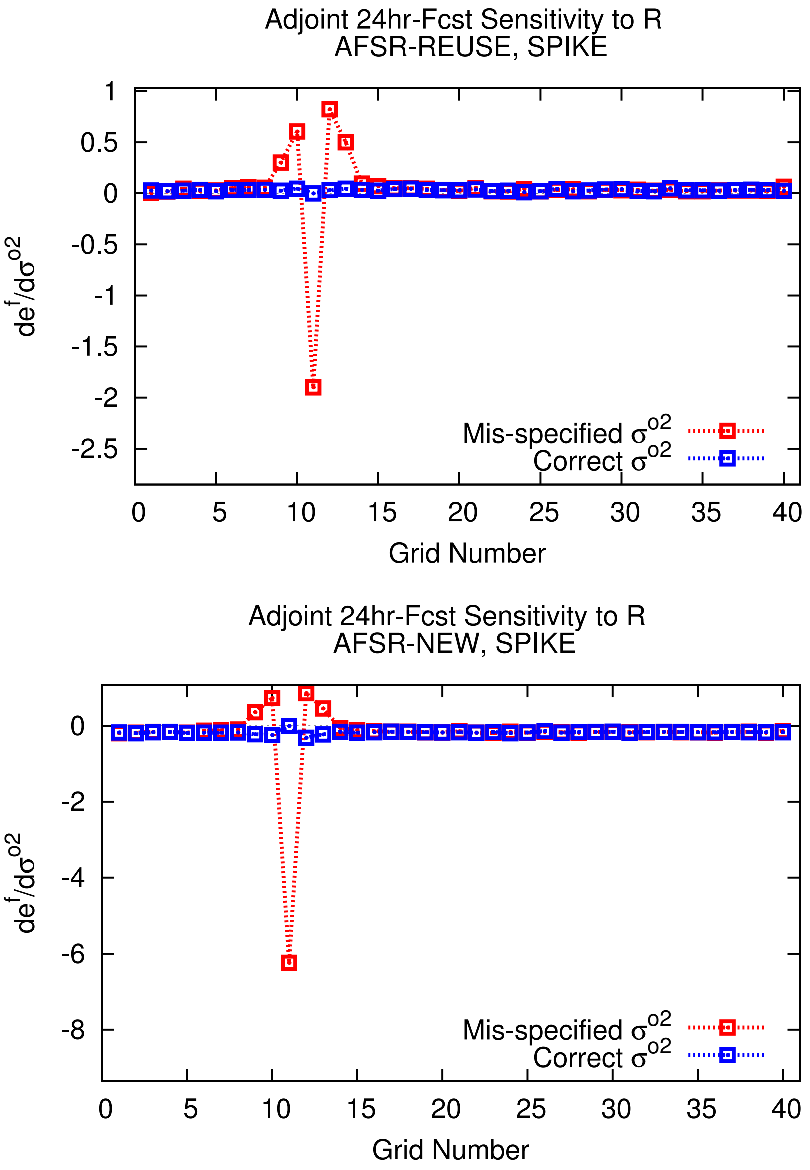


Figure 7.3: As in Figure 7.2, but for estimation using the adjoint FSR formulations: (top) “AFSR-REUSE” and (bottom) “AFSR-NEW.”

> The sensitivity gradient $\frac{\partial e^f_{t|0}}{\partial \sigma^{o2}_j}$, being a partial derivative, tells us how, for each index j, a small displacement in σ^{o2}_j would change the forecast error $e^f_{t|0}$ *if the other observation error variances* $\sigma^{o2}_l, l \neq j$ *are kept unchanged.* Thus, if there is an observation that makes the forecast worse, then we can make the forecast better by giving higher credit to (*i.e.*, decreasing observation error variances for) the adjacent, more accurate observations.

This raises us one concern: the FSR methods might not be a reliable diagnostic if accurate and inaccurate observations are located close to each other. This concern motivated us to try the STAGGERED experiment whose results are described in the next subsection.

Although the two formulations, "EFSR-REUSE" and "EFSR-NEW," give qualitatively similar results, the magnitude of the sensitivity gradient is sharply different, with the latter being larger. As we will show in later subsections, this feature is also observed in the other two experiments as well.

Figure 7.3 shows the result of FSR diagnostics using the adjoint formulations, "AFSR-REUSE" (top) and "AFSR-NEW" (bottom). They are quite consistent with their ensemble counterparts (Figure 7.2). Peculiarly, however, the adjoint results are smaller than the ensemble results by a factor of about 0.65. Again, as we will show in later subsections, this feature is also observed in the other two experiments as well. At this moment, we do not yet know how to interpret this discrepancy.

7.4.3.2 The STAGGERED experiment

Figure 7.4 shows the analysis errors of the STAGGERED experiments verified against the truth. As in the SPIKE experiment, the analysis is more accurate in the correct-**R** run than in the incorrect-**R** run. For both runs, the analysis is more accurate at the odd-numbered grid points where the observations are more accurate than at the even-numbered grid points where the observations are less accurate. Again, the fact that the analysis is more accurate than the observations, whose true error variances are 0.1^2 at the odd-numbered grid points and 0.3^2 at the even-numbered grid points, assures that filter divergence has not occurred in these DA runs.

Figure 7.5 shows the ensemble-based forecast sensitivity gradient to observation error variances, $\frac{\partial e^f_{t|0}}{\partial \sigma^{o\,2}_j}$, estimated using 'the 'EFSR-REUSE" and "EFSR-NEW" formulations. For the incorrect-**R** runs (red lines), the two formulations both state that we should decrease the observation error variances at the odd-numbered grid points where the prescribed observation error variances (0.2^2) are larger than their actual values (0.1^2; note that a positive gradient means that the forecast errors increases by increasing the observation error variances) and the opposite for the even-numbered grid points, which is exactly what we expect from a successful FSR diagnostics.

For the correct-**R** experiments (blue lines), ideally, we expect that if an FSR diagnostics is perfect, then it would state that we do not have to tune **R**, or at least, there should be no difference between the odd-numbered and the even-numbered grid points; our FSR diagnostics tell, however, that we should tune the observation error variances for the odd- and even-numbered grid points differently; furthermore, the two formulations, "EFSR-REUSE" and "EFSR-NEW," give mutually contradicting results: the "EFSR-REUSE" states that we should decrease/increase the observation error variances at the odd/even grid points, while, the "EFSR-NEW" states the opposite.

We would like to point out, however, that this issues may not be as serious as it may look: first, the magnitude of the sensitivity gradient is much smaller than that in the incorrect-**R** run. Also, if we compare them with the result of SPIKE experiment, it is virtually negligible. Moreover, as Figure 7.7 shows, if we compare the odd- and even-numbered grid points by using the forecast sensitivity to the *tuning factors* s_j^o $(j = 1, \cdots, 40)$ of the observation error variances defined by Eq. (7.22), the discrepancy among the odd- and even-numbered grid points become less sharp. This is because, in computing the sensitivity to the tuning factor $\frac{\partial e_{t|0}^f}{\partial s_j^o}$ from the sensitivity to the observation error variance $\frac{\partial e_{t|0}^f}{\partial \sigma_j^{o2}}$, we multiply it by the observation error variance. In particular, the "EFSR-NEW" formulation (bottom panel) gives almost flat profile for sensitivity to the tuning factors.

As in the SPIKE experiment, the adjoint and ensemble formulations are quite consistent (compare Figure 7.5 with Figure 7.6), but here again, the magnitude in the adjoint results is smaller than that in the ensemble results.

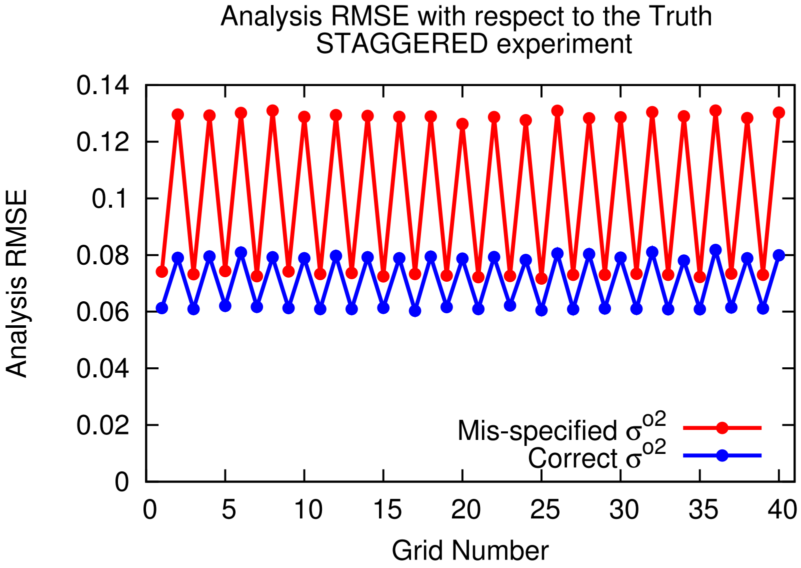


Figure 7.4: As in Figure 7.1, but for the STAGGERED experiment.

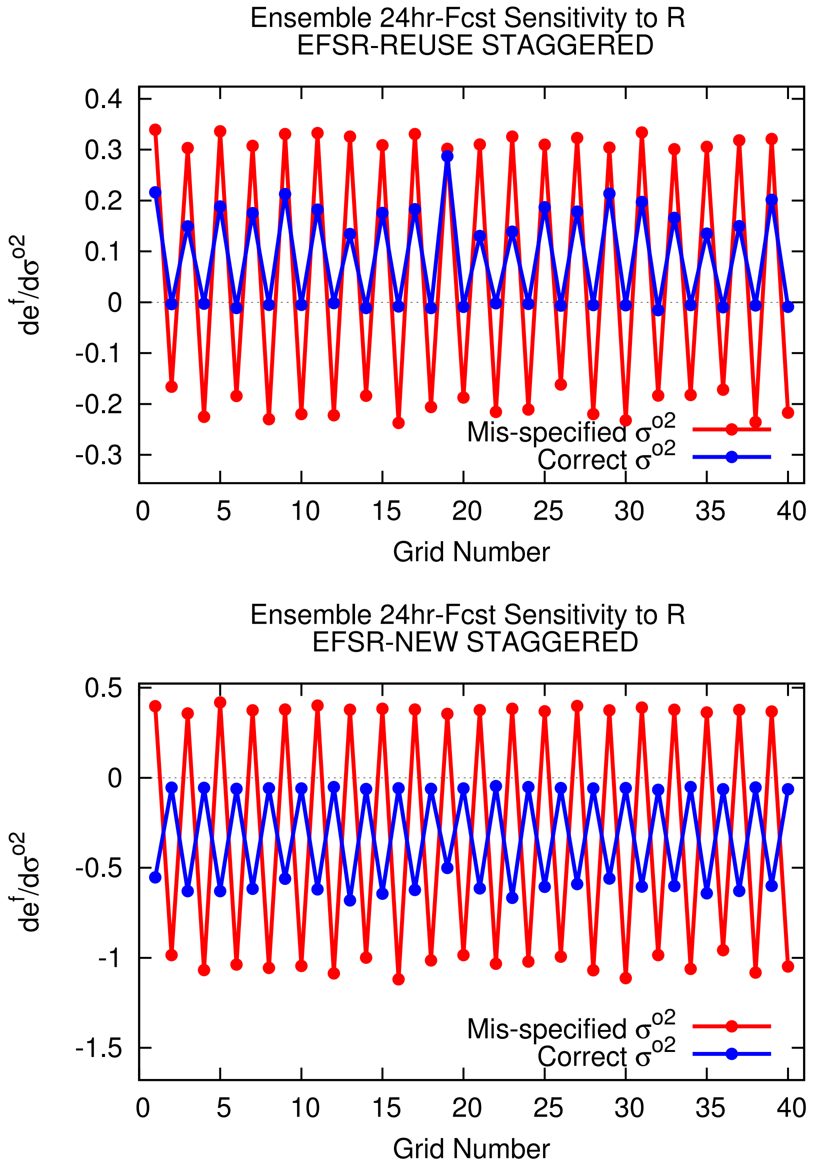


Figure 7.5: As in Figure 7.2, but for the STAGGERED experiment.

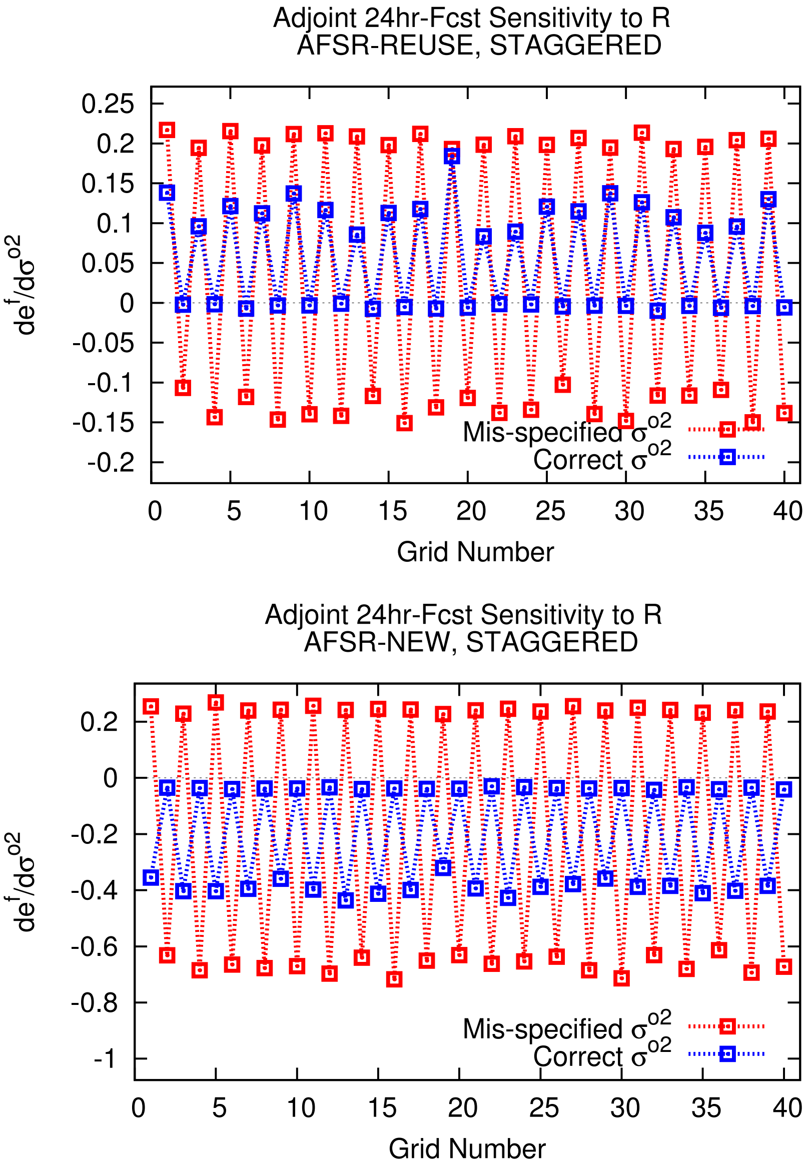


Figure 7.6: As in Figure 7.3, but for the STAGGERED experiment.

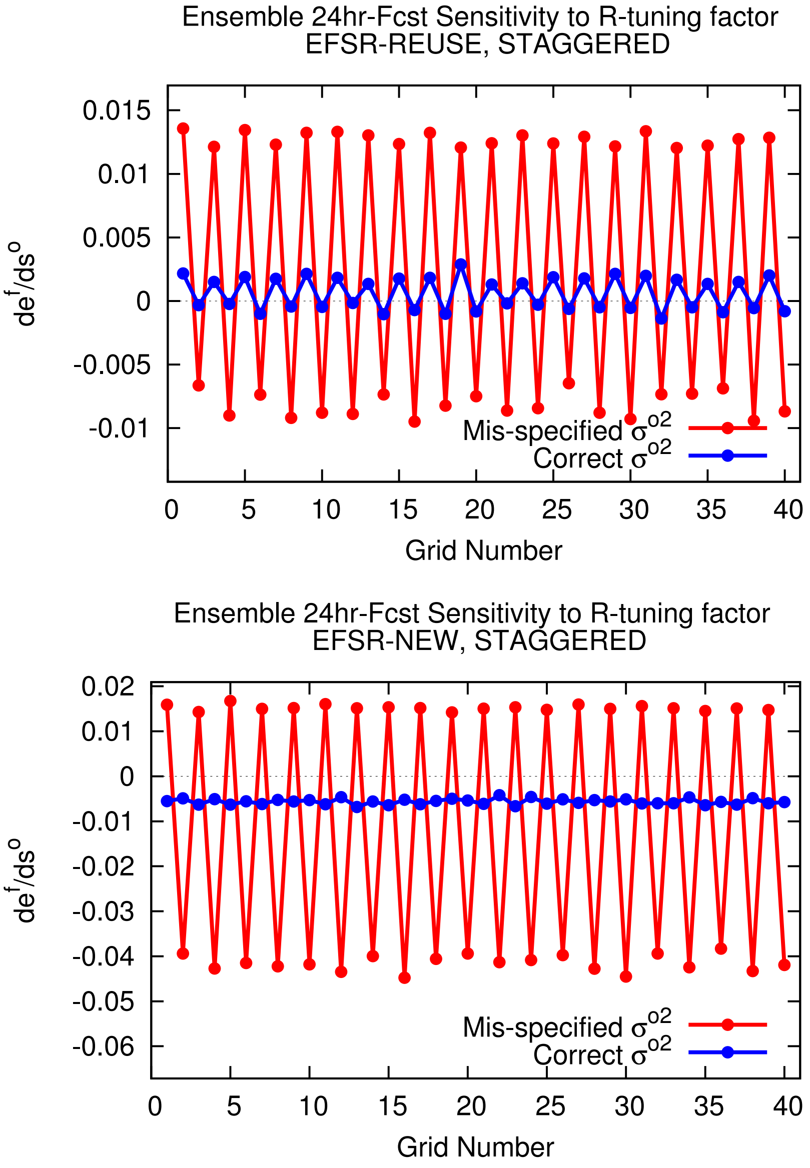


Figure 7.7: As in Figure 7.5, but for the forecast sensitivity to the tuning factor s^o_j of the observation error variances.

7.4.3.3 The LAND-OCEAN experiment

Figure 7.8 shows the analysis errors of the LAND-OCEAN experiment verified against the truth. Consistent with our expectation, for both the correct- and incorrect-**R** runs, the analysis is less accurate over the “land” ($1 \leq j \leq 20$) where the observations are less accurate (with error variance 0.3^2), than over the “ocean” ($21 \leq j \leq 40$) where the observations are more accurate (with variance 0.1^2). A somewhat unexpected, surprising feature here is that, unlike in the previous two experiments, there is not much difference in the accuracy of the analysis between the correct-**R** and incorrect-**R** run, except near the land-ocean boundaries (*i.e.*, between $j = 40$ and $j = 1$, and between $j = 20$ and $j = 21$). This means that, except near the boundaries, the analysis (and thus the forecast) of the incorrect-**R** run cannot be improved much by optimizing the observation error variances specified in the DA system. Thus, we cannot anticipate the FSR diagnostics to provide useful information on how to tune the observation error variances, except near the boundaries.

Consistent with our reasoning above, the ensemble FSR (Figure 7.9) and the adjoint FSR (Figure 7.10) are not as successful for this experiment as they are for the previous two experiments. For the incorrect-**R** run (red line), the “EFSR-REUSE” and “AFSR-REUSE” (top panels) both give correct guidance near the land-ocean boundaries (grid points $\#20 \sim \#23$, $\#40, \#1$ and $\#2$), giving positive/negative sensitivities to the “ocean”/”land” grid points. However, in the interior of the “land,” the sensitivities are rather random with the wrong sign; in the interior of the “ocean,” the sensitivities are very low. The “EFSR-NEW” and “AFSR-NEW” (bottom panels) are both more successful than the “EFSR-REUSE” and “AFSR-REUSE,” giving clear negative sensitivities (*i.e.*, we should increase the observation error variances) for the entire “land.” On the other hand, the results for the correct-**R** run (blue line) is not easy to interpret. Since the observation error variances are all correctly specified, we can expect that the sensitivity gradient should have more or less similar values, giving a flat profile. The “EFSR-NEW” and “AFSR-NEW” give a flatter profile than the “EFSR-REUSE” and “AFSR-REUSE,” so in this sense, the former seems to make more sense. From another aspect, however, “EFSR-REUSE” and “AFSR-REUSE” seem to be more reasonable: near the “land-ocean” boundaries, we can expect to improve the forecast by giving higher/lower credence to the more/less accurate ocean/land observations by assigning smaller/larger observation error variances. “EFSR-REUSE” and “AFSR-REUSE” do suggest us to decrease the observation error variances at the near-boundary “land” grid points (21st, 22nd, and 40th), while “EFSR-NEW” and “AFSR-NEW” suggest otherwise. It is thus not clear which of the four FSR formulations is the best.

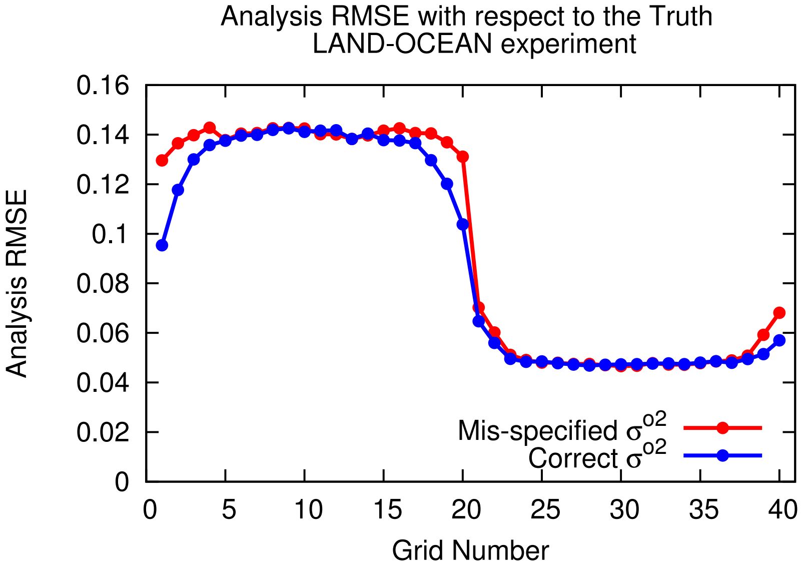


Figure 7.8: As in Figure 7.1, but for the LAND-OCEAN experiment.

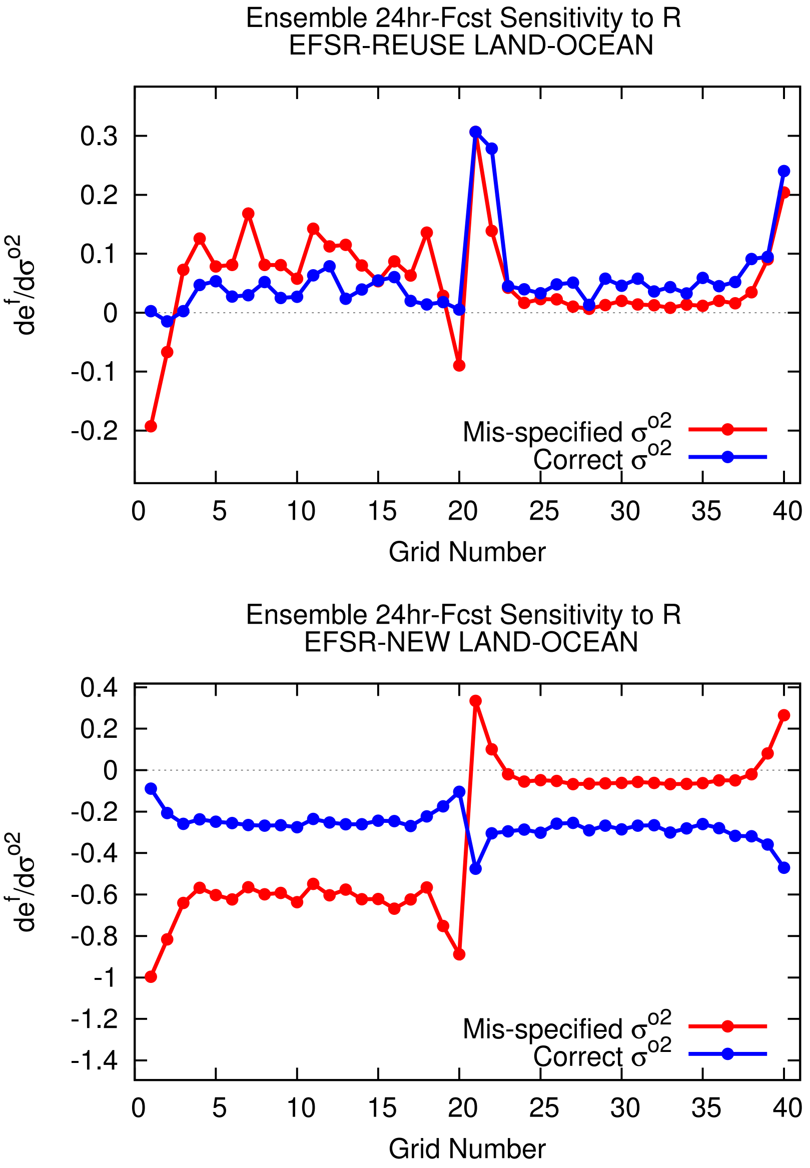
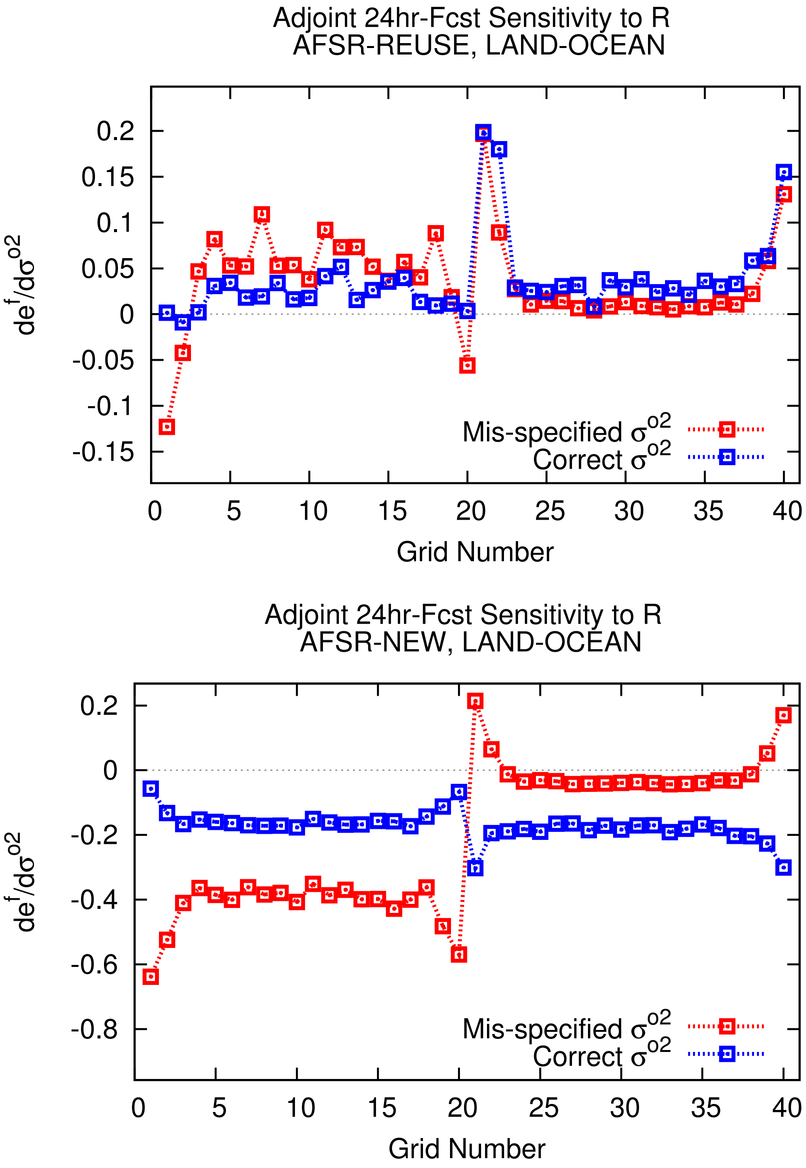


Figure 7.9: As in Figure 7.2, but for the LAND-OCEAN experiment.



Adjoint 24hr-Fcst Sensitivity to R
AFSR-REUSE, LAND-OCEAN

Adjoint 24hr-Fcst Sensitivity to R
AFSR-NEW, LAND-OCEAN

Figure 7.10: As in Figure 7.3, but for the LAND-OCEAN experiment.

7.5 Summary and discussion

In this chapter, we introduced the formulation of EFSR (ensemble-based forecast sensitivity to observation error covariance **R**) by combining an ensemble approximation similar to that of Kalnay et al. (2012) with the adjoint formulation of Daescu and Langland (2013), and examined its capability to detect whether the observation error variances prescribed to a DA system are larger or smaller than the actual values, using the 40-variable version of Lorenz '96 model within a framework of the"identical twin" experiment. Three different sets of experiment are performed, each consisting of two DA cycle runs, one with correctly specified observation error variances, the other with incorrectly specified observation error variances. In any of the experiments, adjoint and ensemble formulation gave consistent estimations, but the adjoint formulation resulted in a smaller sensitivity. Overall, the four tested formulations ("EFSR-REUSE," "EFSR-NEW," "AFSR-REUSE" and "AFSR-NEW") are all capable of detecting mis-matches in the actual and prescribed observation error variances. In particular, in one of the three experiments conducted, the SPIKE experiment in which only one observation at a specific grid point is less accurate than the other observations by a factor of 4, the mis-specification of the observation error variance was detected extremely well by all the four formulations tested, which corroborates the power of the FSR diagnostics.

Two caveats might need to be pointed out, however: first, FSR diagnostics could signal a "false alert" when, in fact, the observation error variances are perfectly specified, as we saw in the correct-**R** runs of the STAGGERED and LAND-OCEAN experiments. Second, in situations where an optimal specification of observation error variances does not lead to noticeable improvement in analysis or forecast, as is the case with the incorrect-**R** run of the LAND-OCEAN experiment, results of the FSR diagnostics can be less reliable.

These caveats being said, our results from the simple system, overall, do support the effectiveness of the FSR diagnostics. Encouraged with this success, we implemented and tested our EFSR formulation to the NCEP's lower-resolution version of the operational global NWP system, as we describe in the next chapter.

Chapter 8: Ensemble Forecast Sensitivity to observation error covariance (EFSR) II: Quasi-operational implementation with GFS LETKF/3D-Var hybrid GSI

8.1 Introduction

In the previous chapter we introduced an ensemble-based formulation of forecast sensitivity to observation error covariance matrix **R**, which we call EFSR, and showed with a simple toy system that the EFSR diagnostics is capable of detecting incorrect specification of the observation error covariance. In this chapter, we apply the EFSR diagnostics to a real NWP system, that is, the GFS model coupled with the LETKF/3D-Var hybrid GSI DA system which we used for our Proactive QC experiments. Unlike systems such as the one we used in the previous chapter, in a real NWP system we do not know the true values of the observation error covariance matrix **R**. Thus, we cannot directly tell if the results of EFSR diagnostics are right or not. In order to verify if the EFSR diagnostics yield meaningful results, we perform a tuning experiment in which the observation error variances for several observation types are modified (or tuned) based on the recommendations from the EFSR diagnostics and check if the tuning of **R** improves EFSO impacts from the tuned observation types. As we will show, our results do corroborate the effectiveness of the EFSR diagnostics.

8.2 Experimental setup

The experimental setup is identical to what we used for Proactive QC experiments (Chapter 3): we use a lower-resolution (T254L64 for control; T126L64 for ensemble) version the GFS EnKF/3D-Var GSI DA system with the EnKF part replaced by the LETKF. In this system, the observation error covariance matrix $\mathbf{R}$ is assumed to be diagonal and is shared between the GSI (3D-Var) and the LETKF. The observations assimilated in our experiments are identical to those in the operational system and the period is 31 days from 00 UTC of January 8th, 2012 to 18 UTC of February 7th with 6-hourly analysis. See Chapter 3 for more details.

We could simply apply EFSR diagnostics to the output we have from our Proactive QC experiments but we re-executed LETKF jobs because, when we performed the Proactive QC experiments, we did not store the ensemble mean analysis in the observation space $\bar{y}_0^a$ which is necessary for computing EFSR (Eq. (7.22)).

For each of the stattypes (see Section 3.5 and Tables 3.2 and 3.3) and the sensors (c.f. Table 3.4), we computed the forecast sensitivity to observation error variance scaling factors $s_i^o, i = 1, \cdots, I$ by Eq. (7.22) using the gradient vector $\frac{\partial e_{t|0}^f}{\partial \mathbf{y}_0^o}$ defined by Eq. (7.25) (the EFSR-REUSE formulation). Since we confirmed in the previous chapter with the toy system that the two formulations for gradient evaluation, EFSR-REUSE (Eq. (7.25)) and EFSR-NEW (Eq. (7.27)), are both capable of detecting mis-specification of observation error variances, we only examine the EFSR-REUSE formulation. The sensitivity to scaling factors is computed for each of the 124($= 31 \times 4$) analyses and are averaged over the whole one-month period.

8.3 Results

This section presents the forecast sensitivity to observation error variance scaling factors $\frac{\partial e^f_{t|o}}{\partial s^o_i}$ for different types of observations obtained from the EFSR diagnostics.

Figure 8.1 shows the forecast sensitivity to observation error variance scaling factors computed using the EFSR-REUSE formulation of the gradient vector (Eq. (7.25)). In the top panels ((a) and (b)) the forecast errors are evaluated with the moist total energy norm; in the bottom panels ((c) and (d)) they are evaluated with the dry total energy norm. Panels on the left ((a) and (c)) and those on the right ((b) and (d)) show, respectively, the results for 6-hour and 24-hour lead time. Positive values of the sensitivity mean that the forecast errors will increase by increasing the corresponding observation error variances and thus we should decrease them. Conversely, negative values indicate that we should increase the corresponding observation error variances.

From Figure 8.1, we see that, except for MODIS winds, all observation types exhibit positive sensitivities in all of the panels. From the discussion in Section 7.3.1, we can interpret this as indicating that the covariance inflation is insufficient

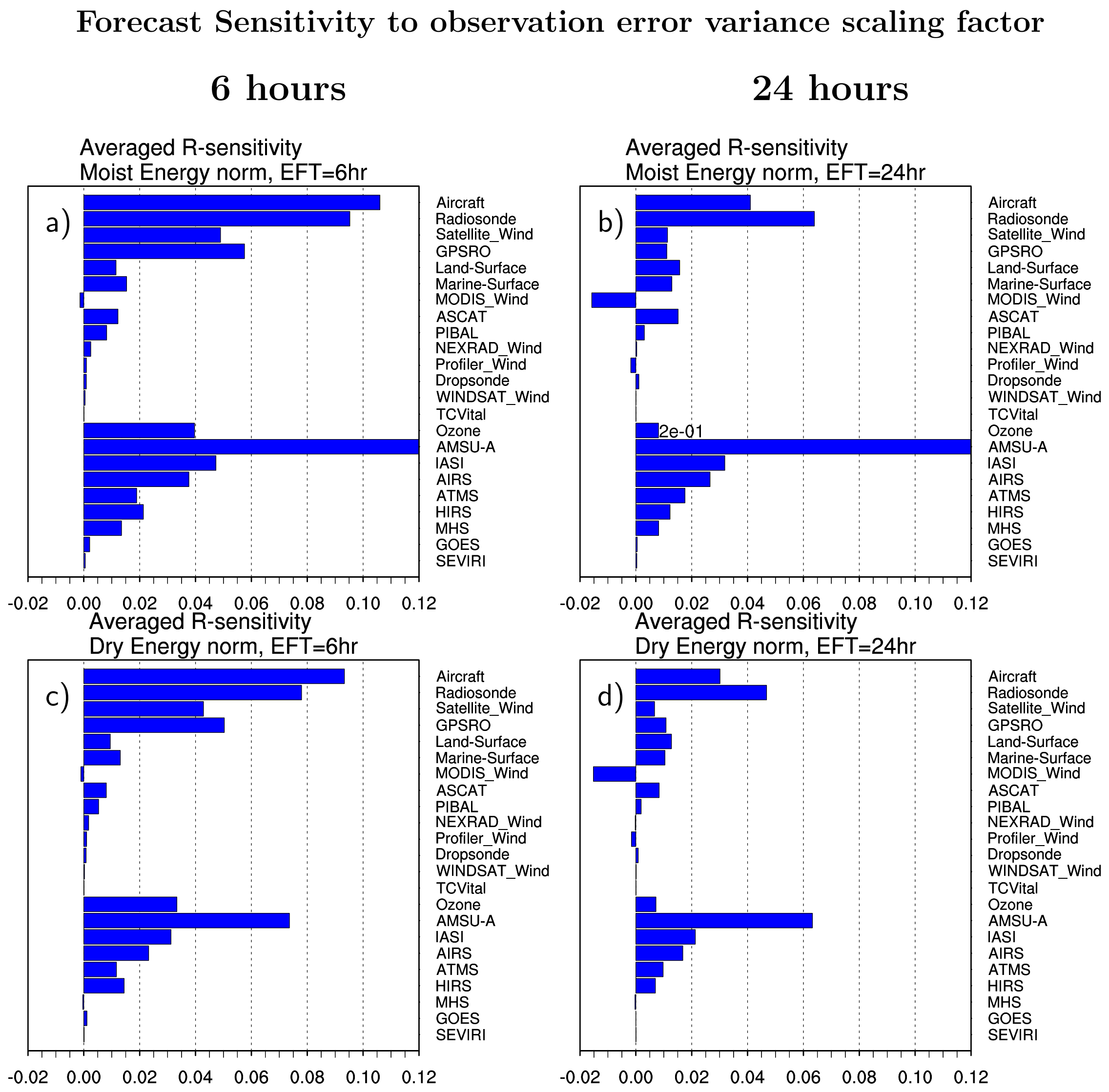


Figure 8.1: Forecast sensitivity to observation error variance scaling factors. Shown are (a) 6-hour forecast sensitivity measured with the moist total energy norm, (b) as in (a), but for 24-hour forecast, (c) as in (a), but measured with the dry total energy norm, and (d) as in (c), but for 24-hour forecast. All are verified against the control GSI analysis. The units are J kg^{-1}.

because, as we can see from Eq. (7.33), the forecast sensitivity to covariance inflation factor is negative (*i.e.*, a stronger inflation will decrease the forecast errors). This is understandable because, as we described in Section 3.4, the covariance inflation parameters we used are optimized for the higher resolution operational system; in EnKF or Extended Kalman Filter (EKF) methods, the background covariance inflation is introduced to compensate for the model errors that are not accounted for in the standard Kalman Filter algorithm (e.g., Jazwinski, 1970, Section 8.3). Thus, it is natural to assume that the covariance inflation that we used in our experiments is weaker than it optimally should be, since the model errors are likely to be larger for a lower-resolution model than for a higher-resolution model.

We can also observe from Figure 8.1 that, among all the observation types, Aircraft, Radiosonde and AMSU-A exhibit higher sensitivities than other types, and that MODIS winds show negative sensitivity. This feature is consistently seen in any combination of the lead times and the error norms. Thus, if our EFSR diagnostics is correct, we can expect to improve the forecast by decreasing the observation error variances for Aircraft, Radiosonde and AMSU-A, and by increasing the observation error variance for MODIS winds.

8.4 Tuning-of-$\mathbf{R}$ experiment

As we discussed in the previous section, we can expect to improve the forecast by decreasing the observation error variances for Aircraft, Radiosonde and AMSU-A, and by increasing the observation error variance for MODIS winds. This motivates us to perform a tuning experiment of the observation error variances; this also serves as a means to validate the results of EFSR diagnostics. This section describes the experimental setup and the results of our tuning experiment.

8.4.1 Experimental setup

The first thing we need to think of before conducting a tuning-of-**R** experiment is how much to tune it. As Daescu and Langland (2013) points out, however, the FSR diagnostics, be it adjoint-based or ensemble-based, does not provide information about how much we should tune; it only provides the direction for tuning, without any clue on the magnitude. Without better option, we decided to tune the observation error variances in a rather modest way; we multiplied the observation error variances of Aircraft, Radiosonde and AMSU-A by 0.9 and that of MODIS winds by 1.1.

Using the new, tuned observation error variances, we re-ran the cycling experiment for the entire period, including the first 7-day spin-up, and we used the last 31 days for verification (see Section 3.7).

8.4.2 Results

By this tuning, we expect to improve the forecast, and more specifically, to improve the EFSO impacts from the tuned observation types. Figure 8.2 compares the one-month averages of the EFSO impacts from each observation type before and after the tuning. The error bars shown in the figure represent the confidence interval at 95% level obtained by a standard t-test for the difference of paired data: Let us fix the observation type and let X_1 and X_2 denote, respectively, the EFSO impacts from that type. For each of the $31 \times 4 = 124$ cases, we compute the difference $d_i = X_{1,i} - X_{2,i}$, where i is an index for the cases (samples). Our null-hypothesis is that the population mean μ_d is zero. We define the test statistics t by $t = \bar{d}/\left(\frac{s_d}{\sqrt{n}}\right)$ where $\bar{d}$ and s_d represent, respectively, the sample mean and sample standard deviation of d, and $n = 124$ is the sample size. Under the assumption that d is normally distributed with zero mean, the test statistics t obeys the t-distribution with the degree of freedom $\nu = n - 1$. Thus, the confidence interval at 95% level can be obtained by $\pm t_\nu^{2.5\%} \times \frac{s_d}{\sqrt{n}}$ where $t_\nu^{2.5\%}$ is the 2.5 percentile of the t-distribution with the degree of freedom ν.

From Figure 8.2 we can observe the following features: (1) the EFSO impacts from Aircraft, Radiosonde and AMSU-A all had statistically significant improvement by the tuning, for any combination of the lead times and the error norms, (2) the EFSO impact from MODIS wind had no statistically significant improvement or degradation, (3) the EFSO impact from IASI is also improved, although its observation error variance is not tuned; however, the improvement is not statistically significant in panel (d) (the dry energy norm with 24-hour lead time), and (4) the EFSO impact from GPSRO is somewhat degraded, especially for 24-hour lead time.

The feature (1) is a desirable result that affirms our expectation that the EFSR diagnostics indeed yield meaningful information. The feature (2), on the other hand, seems puzzling at a first glance; we give an interpretation to this question in the next paragraph. The feature (3) is also difficult to interpret but perhaps we should not trust this result too much since the statistical significance is not very large compared to the differences for Aircraft, Radiosonde and AMSU-A. The feature (4) is also difficult to explain; here we only speculate that the impacts from GPSRO are somewhat obscured by the enhancement of impacts from Aircraft and AMSU-A; GPSRO data are mainly observed in the upper troposphere and in the stratosphere where Aircraft and AMSU-A data are also abundant. By giving more weight to Aircraft and AMSU-A observations by decreasing their observation error variances, GPSRO observations would receive relatively less weight, possibly resulting in weaker EFSO impacts.

Now, let us consider why the tuning of observation error variance for MODIS wind did not improve or degrade its EFSO impacts. As we saw in Chapter 6, during the period of our experiments, MODIS wind observations contained a lot of "flawed" observations with large negative impacts to the forecast. Figure 8.3 shows the time series of the EFSO impacts from MODIS wind observations evaluated for 6-hour lead time with moist total energy norm. As we can see, for most of the cases, the impact is positive (*i.e.*, negative values; negative values of EFSO decrease the forecast errors so their impacts are positive) but it sometimes becomes negative (positive values). In such negatively impacting cases, the MODIS wind are likely to have large observation errors. Thus, including such "outlier" cases would lead to non-Gaussianity of observation errors for MODIS wind, which violates the assumption we made in deriving EFSR formulation that all observation errors obey normal distributions with zero mean. Including such cases in the statistics thus would have rendered EFSR diagnostics less accurate, so we can expect to improve

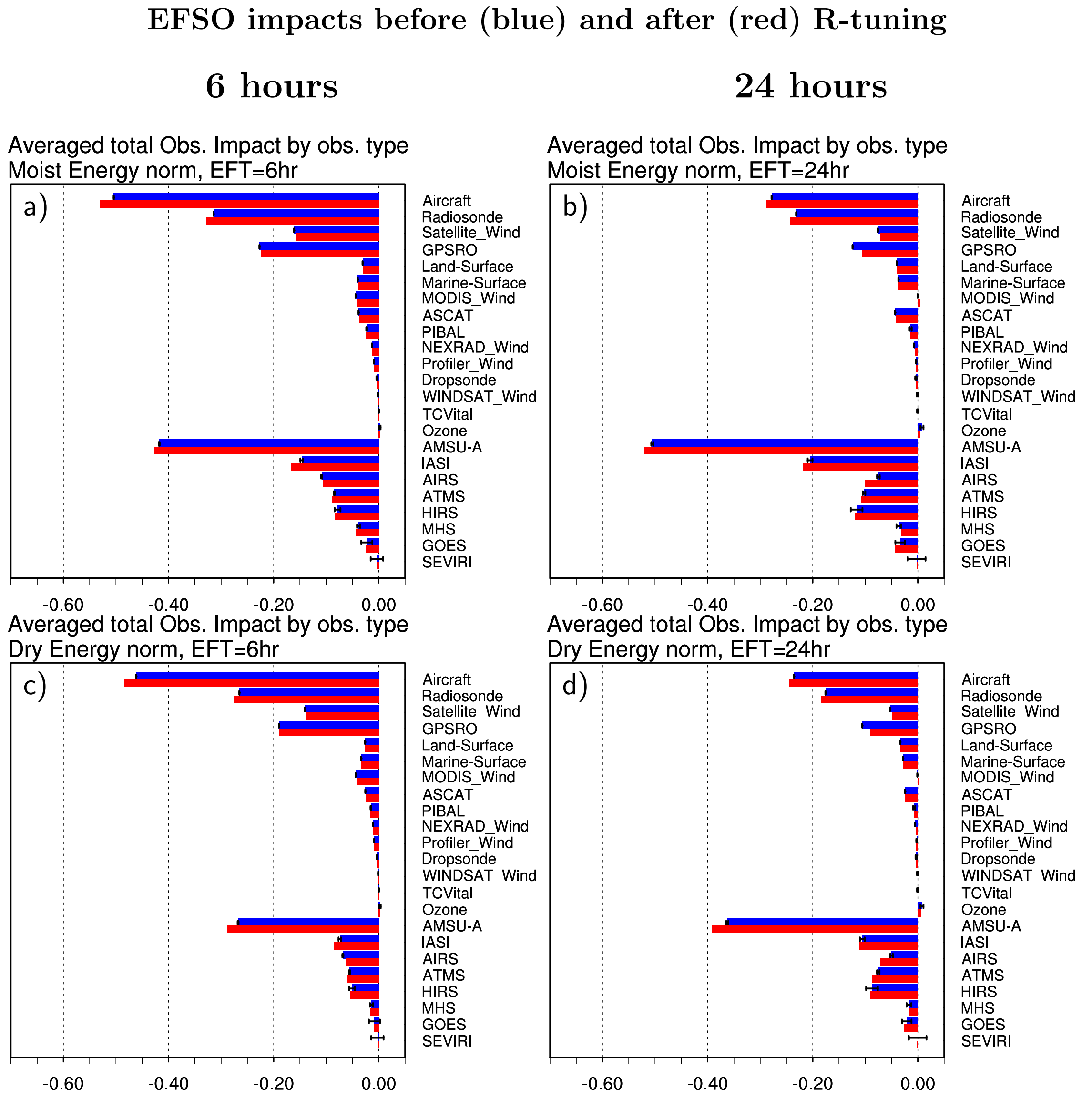


Figure 8.2: The one-month averages of the EFSO impacts from each observation type before and after the tuning of observation error variances. The EFSO impacts before the tuning are shown with blue bars; those after the tuning are shown with red bars. The error bars represent the confidence intervals at 95% level computed by a t-test for the difference of two paired data (see text for detail). Shown are (a) 6-hour EFSO impacts measured with the moist total energy norm, (b) as in (a), but for 24-hour EFSO, (c) as in (a), but measured with the dry total energy norm, and (d) as in (c), but for 24-hour EFSO. All are verified against the control GSI analysis. The units are J kg^{-1}.

the accuracy of the diagnostics by removing such “outlier” cases from the samples.

In Figure 8.4, we show the forecast sensitivity to observation error variance scaling factors averaged over the period but without using cases in which 6-hour EFSO impacts from MODIS wind, measured with the moist total energy norm verified against the control GSI analysis, were negative. The results are mostly the same with Figure 8.1. The only (and important) difference is that, in Figure 8.4, where the cases in which the observation errors of MODIS wind are likely to contain outliers are excluded, the forecast sensitivity to observation error variance for MODIS wind is nearly zero, for any combination of the lead times and the error norms. This means that a tuning of observation error variance for MODIS wind would have neutral impact on the forecast, which is consistent with the feature (2) we observed in Figure 8.2.

The important lesson we learned from this investigation is that, before applying EFSR diagnostics, we should remove “flawed” observations so that the observation errors would obey a normal distribution; otherwise, the EFSR diagnostics will be biased. Also, the flawed observations require a separate tuning from the good observations.

8.5 Summary and discussion

In this short chapter, we first applied the EFSR diagnostics, whose effectiveness is confirmed with a toy system in Chapter 7, to the lower-resolution version of the NCEP’s real operational DA system. The results suggest that the forecast will

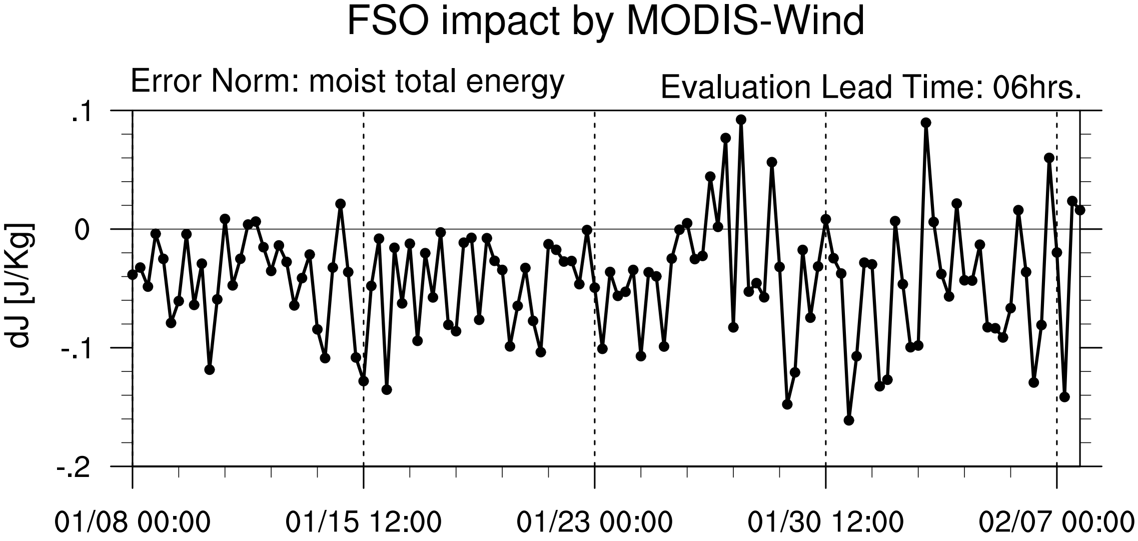


Figure 8.3: The EFSO impacts from MODIS wind observations for 6-hour lead time measured with the moist total energy norm shown as a time series. The verification is against the control GSI analysis. Note that positive values correspond to negative impacts because they increase the forecast errors.

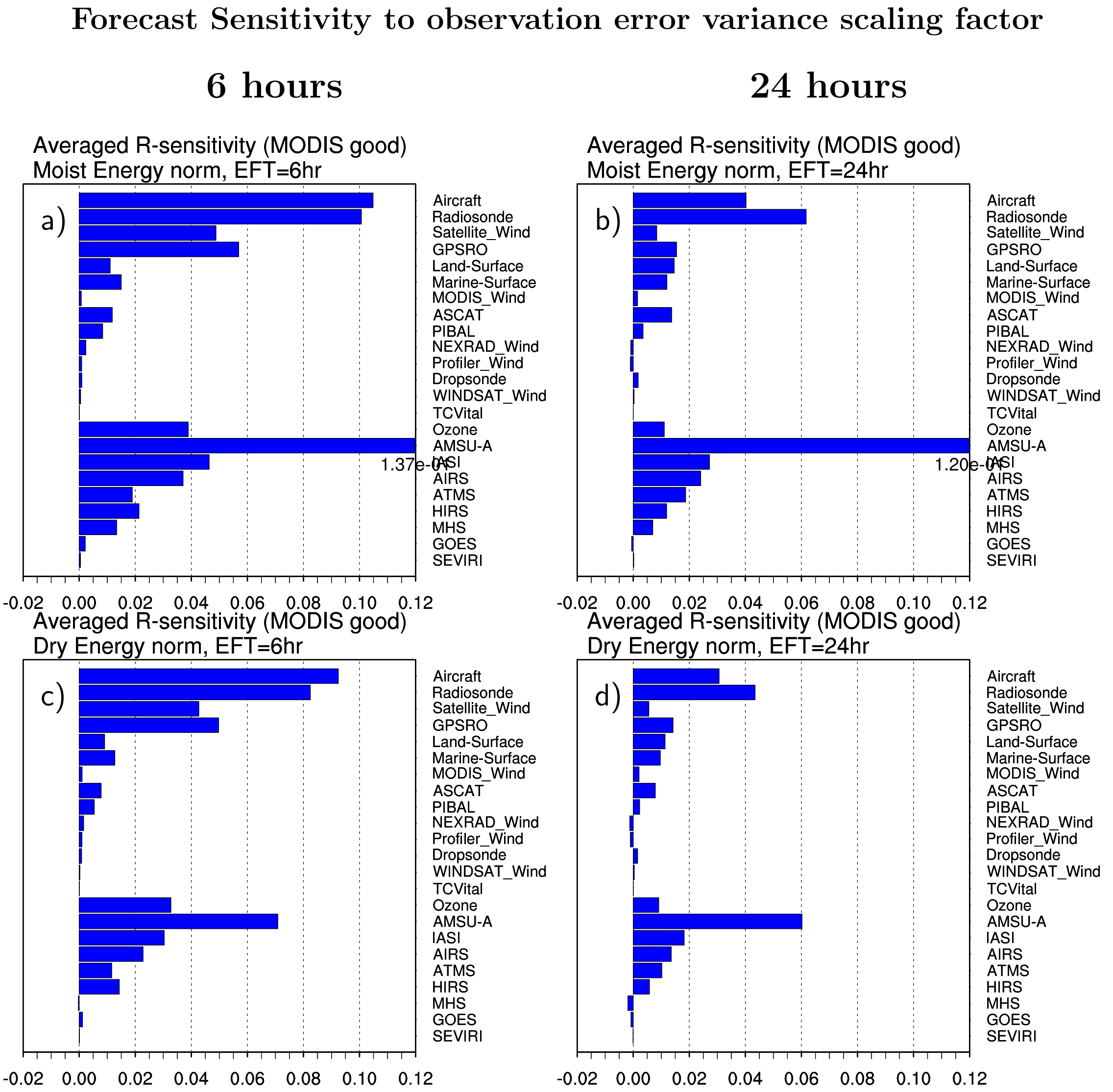


Figure 8.4: As in Figure 8.1, but the average is taken without using cases in which the 6-hour EFSO impacts from MODIS wind measured with the moist total energy norm was negative.

be improved by reducing the observation error variances for Aircraft, Radiosonde and AMSU-A and by increasing the observation error variance for MODIS wind. We then conducted a simple tuning experiment by increasing/decreasing the observation error variances for the the four observation types accordingly, and obtained results that do support our proposition that the EFSR diagnostics can provide useful guidance for an improvement of **R** matrix. Curiously, however, tuning of the observation error variance for MODIS wind did not lead to improved EFSO impact; by investigating the reason for this, we also obtained an important and practical lesson: in order for the EFSR diagnostics to work well, we need to first remove “flawed” observations which, if not removed, result in violation of the Gaussianity assumption for the observation errors. Failure to remove (or separately treat) the “flawed” observations biases the EFSR estimation for the “good” observations.

As we discussed in Section 1.5, the observation error covariance matrix **R** is one of the few parameters of DA systems that must be prescribed externally. Currently, operational NWP centers prescribe it in a more or less subjective manner and thus there should be plenty of room for it to be improved. Without a systematic method to tune it, however, up to present, operational centers have had to rely on a rather ad hoc, empirical tuning. Our EFSR-based tuning method is possibly innovative in this respect because it will allow NWP centers to optimize the **R** matrix in a systematic fashion. An advantage of our method that is important from a practical perspective is that, unlike the adjoint-based method of Daescu and Langland (2013), it does not require an execution of expensive diagnostics; it can “reuse” the sensitivity gradient that is readily available from EFSO, which, in turn, is also much more computationally efficient than the adjoint-based FSO.

As we described in Section 7.3.1, our EFSR can also be used to tune the background covariance inflation factor as well. Li et al. (2009) also proposed an algorithm that enables simultaneous estimation of the the background covariance inflation factor and the observation error covariance, based on the consistency diagnostics we briefly described in Section 1.5. Since both our method and that of Li et al. (2009) are applicable to any EnKF system, and they are based on different, complementary approaches, it would be interesting to compare them in a same system. One advantage of their approach that is missing in our method is that it can specify how observation error variances should be tuned; in our approach, as in the original adjoint-based method of Daescu and Langland (2013), we can only estimate whether we should increase or decrease them (we could iteratively optimize the scaling factors, as we do in variational DA methods, by using minimization techniques such as conjugate gradient or quasi-Newton methods, but that could be prohibitively expensive because each iteration requires a re-execution of DA cycles). The advantage of our approach over theirs is that, we can use information from forecast errors so that the tuning of $\mathbf{R}$ actually improves the forecast, whereas, the method of Li et al. (2009) only adjusts $\mathbf{R}$ and the inflation so that they satisfy the consistency condition as closely as possible. Due to this complementary nature of the two approaches, we can expect to establish a more robust tuning method by combining the two methods.

Chapter 9: Summary and Future Directions

9.1 Summary

9.1.1 Proactive QC

One of the major problems in the operational NWP systems is the so-called "forecast skill dropouts," that is, abrupt and large drops of the forecast skills. Recent studies have shown that some of the dropout cases are caused by the failure of operational QC system to detect and filter out flawed observations. Ota et al. (2013) showed that such flawed observations can be detected by 24-hour EFSO and that rejection of the detected observations from the analysis indeed significantly improves the analysis and forecast. Encouraged by their achievement, in this thesis, we proposed to exploit the EFSO's capacity to detect such flawed observations just after 6 hours from the analysis and then remove them and repeat the analysis and forecast. The goal of the first part of this thesis is to investigate if this technique, which we call Proactive QC, can improve the NWP forecast. Before implementing this technique into the operational system, however, several issues needed to be addressed.

First, we implemented the EFSO into a lower resolution version of the NCEP's operational global EnKF/3D-Var hybrid DA system whose EnKF part is replaced by the LETKF. Because we were the first to implement EFSO on a hybrid DA system, we carefully examined the consistency of our results with other previous FSO studies. We also examined the validity of using 6-hour EFSO by comparing it with the tried-and-true 24-hour EFSO using both control GSI analysis and ensemble mean LETKF analysis as truth.

Having confirmed that, somewhat to our surprise, EFSO results do not depend sensitively on the choice of the verifying truth and the forecast lead time, we investigated if we can detect occurrences of regional forecast dropouts after only 6 hours. Expanding the idea of Ota et al. (2013), we first introduced two different methods to divide the globe into smaller regions, the $30^\circ \times 30^\circ$ cells and spherical harmonics Y_{12}^{6} cells, and examined the statistics of two quantities, the normalized regional forecast errors and the regional forecast error reduction by the analysis, evaluated for both of the two methods to divide the globe. Based on the statistics we obtained, we proposed to use the "2σ" criterion for the $30^\circ \times 30^\circ$ cells, which selects the regions if their normalized regional forecast errors and the regional 6-hour forecast error divided by the regional 12-hour forecast error are both larger than their temporal averages by at least 2 standard deviations. This criterion picks up about ~ 2 regions per cycle, so operational systems should be able to afford to perform regional EFSO on the detected regions. Then, we performed 6-hour and 24-hour EFSO and confirmed that the "2σ" criterion can capture the cases for which, according to EFSO estimation, large ($\sim 25\%$ or higher) regional forecast improvements can be expected by the denial of the "flawed" observations.

We then performed data denial experiments to test whether the rejection of observations that are identified by EFSO as “flawed” really improves the forecast. In order to see how many of the observations we should reject given the information on the EFSO values of each observation, we tried four different criteria for setting the threshold of EFSO values above which the observations are rejected. We found that, in all of the 20 cases examined, we can in fact improve the forecast if we adopt strict thresholds for the rejection of observations. Forecast improves more dramatically if we reject all negatively impacting observations of the types identified by EFSO as “flawed,” but this approach tend to yield some forecast degradation. Thus, strict and loose thresholds for the rejection of data have their own pros and cons, and the threshold would have to be chosen somewhat subjectively depending on how much we can tolerate possibility of forecast degradation. In several of the examined cases, the forecast improvements attained by the rejection of the “flawed” observations was quite spectacular, with local relative improvement reaching as much as 30%–50%, indicating the power of Proactive QC.

In Section 1.7, we started out by posing four key questions. Here we review the accomplishments of our thesis by providing answers to the four key questions:

1. Are 6 hours long enough for the detection of “flawed” observations?

Our answer is “yes”: we confirmed, in Chapter 4, that 6-hour EFSO and 24-hour EFSO are highly consistent, both in terms of statistical properties and for individual cases. The consistency between 6-hour and 24-hour EFSO is also corroborated by the similarity of the forecast improvements of data denial experiments based on 6-

hour and 24-hour EFSO, for example, by checking the resemblance of Figure 6.2 (6 hours) and Figure 6.3 (24 hours). In Chapter 4, we also confirmed that the EFSO is not sensitive to the choice of the verifying truth by using either the control GSI analysis or the ensemble mean LETKF analysis.

2. How can we detect possible occurrences of "dropouts" after only 6 hours from analysis?
We found that the "2σ" criterion which we presented in Chapter 5 allows us to detect regional "dropouts" after only 6 hours from analysis. The 219 cases selected by the "2σ" criterion included 15 cases whose estimated forecast improvements are above 25%. Data denial experiments on the 20 cases extracted from the cases that passed the "2σ" criterion showed that, in seven cases, significant forecast improvements that locally exceed 30% can be achieved by rejecting the observations identified by 6-hour EFSO as "flawed."

3. What is the best threshold for rejection of "flawed" observations?
We would have to say that this question must be answered rather subjectively. If it is very important that Proactive QC never degrades the forecast at any location, then a stricter threshold will be favored; if some degradation of forecast can be tolerated, loose threshold such as the "allneg" criterion, in which all the observations whose EFSO impacts are negative are rejected, should be more favorable because it will allow more dramatic forecast improvement.

4. Does rejection of detected "flawed" observation really improve analysis and fore-

cast?

Our answer is yes. We confirmed in Chapter 6 that 24-hour forecasts can in fact be improved by the rejection of observations that are identified as “flawed” with 6-hour EFSO.

The novel findings of this work include the following: we confirmed that EFSO also works well with the EnKF within a hybrid DA system; we also confirmed that EFSO with forecast lead time as short as 6 hours is equally capable of estimating the impacts of each observation type and of detecting flawed observations to the one with the widely used 24-hour lead time; we demonstrated that Proactive QC in fact significantly improves the forecast. We would like to emphasize here that our proactive QC is a major innovation because it will allow us, for the first time ever, to carry out fully flow-dependent QC based on whether the observation actually degraded the forecast.

Our work does have some limitations. First, the resolution of the analysis and forecast adopted in this thesis (T126 for ensemble and T254 for control) is quite low compared to the current operational systems. It is not easy to infer whether the success of our results would be reproduced if we implement Proactive QC to the actual high-resolution operational system. This concern can be answered only by repeating the whole procedure of our study in the higher resolution system. Second, we did not examine if the forecast is also improved if we used the new 6-hour forecast after the denial of the observation as the background for the next cycle; the data denial experiments we conducted are “off-line” in the sense that the improvement of the forecast achieved by not assimilating the “flawed” observations is not taken over to the next cycle. However, if Proactive QC is implemented in the real-time operational NWP system, the improved forecast will be used as the background (first guess) at the next cycle. We do not see any reason by which this cycling should degrade the forecast or reduce the forecast improvement (on the contrary, it should reinforce the improvement), but this has to be confirmed before operational implementation.

9.1.2 EFSR

The observation error covariance matrix **R** is one of the few external parameters to a DA system and thus, in operational NWP systems, it is specified somewhat empirically and subjectively. In the second part of this thesis, we showed that the forecast sensitivity to observation error covariance matrix **R** (FSR) diagnostics derived with the adjoint technique by Daescu and Langland (2013), which can be used to optimize the **R** matrix, can be formulated for any EnKF DA system by using the approximations used in the derivation of EFSO formulation by Kalnay et al. (2012). We denote the ensemble based FSR diagnostics by "EFSR." We verified its validity with a series of experiments which use a toy system based on Lorenz '96 model. We then applied the EFSR diagnostics to the lower-resolution version of the NCEP's global NWP system that we used for testing Proactive QC and performed simple "tuning-of-**R**" experiment in which the observation error variances for the observation types with particularly high EFSR sensitivities (Aircraft, Radiosonde and AMSU-A) are reduced by 0.9, and the error variance for MODIS wind, which exhibited negative EFSR sensitivity, is inflated by 1.1. We found that, by this simple tuning of **R**, we can improve the EFSO impacts from the tuned observations except MODIS wind. After some investigation, we found that the EFSO impact from MODIS wind was not improved because the EFSR estimation of MODIS wind was not accurate due to the "flawed" MODIS wind observations; by taking samples excluding the cases where MODIS wind observations exhibited negative EFSO impacts, we found that the EFSR sensitivity for MODIS wind in fact became neutral, which is consistent with the neutral impact of observation error variance tuning.

Our work is the first to derive an ensemble-based FSR diagnostics. Also, it is the first to apply this technique to the quasi-operational global DA system. Our work is innovative in that we showed that the tuning of **R** based on FSR diagnostics can improve the FSO impact from the tuned observation types.

Although our EFSR work is innovative, we need to perform more detailed experiments. For example, in showing the EFSR sensitivity, we did not separately treat the different channels of satellite radiance observations. It is well acknowledged that radiance observations, especially those from hyperspectral sounders such as IASI and AIRS, have very different error characteristics for different channels. Thus, it should make more sense to treat observations from different channels separately in evaluating EFSR sensitivity.

Another topic that we did not cover in this dissertation is the EFSR's capacity to diagnose the impacts from off-diagonal elements of **R**. EFSR sensitivity can be computed not only for diagonal elements (*i.e.*, observation error variances) but also for off-diagonal elements of **R**. Thus, it can be used to address the correlation issues between the errors of different observations. Possible applications of this capacity include tuning of observation thinning (for dense observations such as AMVs) and channel selection strategy for hyperspectral sounders.

9.2 Future directions

First, we would like to repeat the entire experiment for Proactive QC with the operational resolution. In our data denial experiments, we obtained particularly high forecast improvement by the rejection of MODIS wind observations. Since MODIS wind observations are very dense, it is possible that a higher resolution DA system can assimilate these data more effectively than our lower-resolution system. If so, EFSO diagnostics would not identify MODIS wind as "flawed" observation type in the first place, or the denial of them do not lead to improved forecast. Thus, experiments with a higher resolution system will allow us to clarify if the MODIS wind observations that were identified as "flawed" in our experiment really had poor measurement error or they were detrimental because of a problem in the DA system's side.

In implementing Proactive QC to a real-time operational environment, having to wait 6 hours after the first analysis can be a bottleneck. It is thus important to minimize this delay. We have two ideas for mitigating this problem. The first idea is to exploit the time lag that exists between the so-called "early" analysis and the so-called "delayed" (or "cycled") analysis (this idea was suggested by Dr. John Derber of NCEP): some operational centers, including NCEP and JMA, maintain two different kinds of global DA jobs. One is executed solely to provide the initial conditions to the extended forecast. This DA job is called "early" analysis. The analysis from this DA job is used one time and is not taken over to the next cycle. The name comes from the fact that it has to finish earlier in time and thus has shorter cut-off time. The other DA job maintains the analysis-forecast cycle. Because it does not have to provide the initial condition for the extended forecast, it does not have to finish early, so the cut-off time for observation ingestion is longer, allowing it to assimilate observations that arrive late. The data dependency is shown in the schematic Figure 9.1 which is reprinted from JMA (2013). Suppose we would like to detect flawed observations at 12 UTC. In order to perform 6-hour EFSO for 12 UTC, we need an analysis at 18 UTC. If we only have a single cycled DA job, the analysis for 18 UTC cannot be computed without the analysis for 12 UTC. However, in the system shown in Figure 9.1, the early analysis at 18 UTC (labeled "GA18") is computed from the forecast from the early analysis at 12 UTC (labeled "GA12"). Thus, after the completion of GA18 and the cycled analysis at 12 UTC (CA12), we can compute 6-hour EFSO for the observations assimilated at CA12. We can then repeat[1] CA12 without using the observations at 12 UTC that are identified as "flawed" by 6-hour EFSO. Note that CA12 can finish later than 18 UTC of physical time because its output is not used for the extended forecast until GA00. In fact, in JMA's operational system, the finish time of GA18 (20:30 UTC) is earlier than the start time of CA12 (23:50 UTC). We can exploit this time difference for the computation of 6-hour EFSO after GA18 so that "flawed" observations are rejected

[1] As we show in the next paragraph, in fact, analysis does not have to be repeated.

at the second execution of CA12. The only new assumption in this approach is that, analysis from GA is more accurate than the background from CA, which is generally true.

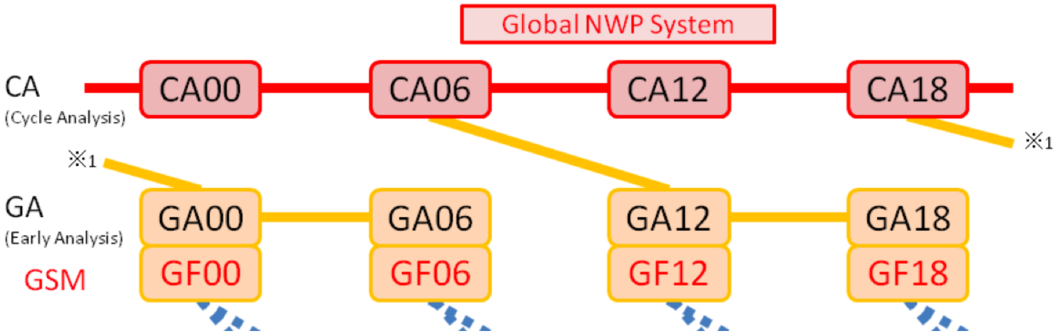


Figure 9.1: Schematic of JMA's global NWP system. The data dependency is depicted by the thick lines: for example, the oblique dark-orange line connecting CA06 and GA12 denotes that GA12 (early analysis for 12UTC) depends on the short-term forecast from CA06 (cycle analysis for 06UTC). Adapted from JMA (2013).

The second idea for the mitigation of the delay is not to repeat the analysis and forecast: we showed in Chapter 2 that the analysis that would be obtained by denying a subset of the assimilated observations can be approximated by EFSO (Eq. (2.24)) without doing the analysis. Thus, once we have the "flawed" observations that should be rejected, we can approximately obtain the improved analysis without even performing an analysis. Similarly, by using the approximate equation Eq. (2.27) for the forecast that would be obtained by denying a subset of the observations, we can approximately obtain the improved forecast, without repeating the forecast. As we discussed in Section 2.2.2.3, in Eq. (2.24) and Eq. (2.27), the Kalman gain $\mathbf{K}$ is assumed to be the same for the original analysis (with full set of observations including the "flawed" observations) and the new analysis (without the “flawed” analysis). This assumption should be reasonably good if the number of denied observations is much smaller than the total number of the assimilated observations. The number of observations denied in Proactive QC is at most $\sim 10^4$ and typically $\sim 10^2$, which is a tiny fraction of the total number of the assimilated observations ($\sim 3 \times 10^6$). We can thus expect this technique to work well. In fact, Ota et al. (2013) used the same approximation and obtained very good consistency between the nonlinear forecast change and its linear approximation (see their Figure 9).

In Section 6.5, we argued that the fact that the denial of MODIS wind resulted in particularly large forecast improvements suggests that MODIS wind might have had some technical problem and that developers of MODIS wind should look into any potential deficiencies in their algorithm. This motivates us to explore the possibility of another major advance from Proactive QC, beside the direct outcome of improving the forecasts: real-time operation of Proactive QC would enable us to build up a detailed database of failed observations by collecting all their occurrences along with relevant metadata. Such database can then be provided to algorithm developers to help them to identify the problem that produced the bad observations and avoid them in the future. For this application, close collaboration with the developers of instruments is indispensable in order to determine what type of information and metadata would be most helpful to them.

Finally, we would like to propose a very powerful application of EFSO, Proactive QC and EFSR diagnostics that would allow a more efficient and precise determination of the optimal way to assimilate new observing systems: if a new observing system becomes available to NWP systems, NWP developers typically evaluate the usefulness of the new data by conducting the OSE-type experiments: namely, the forecast started from “experiment” (or “test”) analysis, which is made by using the new observations, is compared to that started from “control” analysis, which is made without using the new observations. This current approach has difficulties in obtaining statistically significant results in the presence of the rest of the available observing systems. This is because the current NWP systems already produce very accurate analysis by assimilating a myriad of observations from a plethora of observation types. The Proactive QC should address this problem by finding the short-term impact of each observation and allowing the comparison of the impact of different observation algorithms. In fact, Lien (2014) showed, using the Tropical Rainfall Measurement Mission (TRMM)-retrieved global precipitation data as an example of a new observing system, that EFSO can be effectively used to systematically design a data-selection strategy without which the new observing system fails to improve the forecast; without EFSO, one would have to perform a huge number of expensive OSE experiments with trial and error before arriving at an appropriate data-selection strategy.

In designing an assimilation method for new observing systems, an optimal specification of the observation error variance for them is also a difficult problem. As we discussed in Section 1.5, the traditional approach based on the statistics of observation-minus-background (O-B) departure assumes the diagonality of **R** (Hollingsworth and Lönnberg, 1986). This assumption is becoming increasingly inappropriate, especially for new, remotely sensed observing systems, because of their high resolution both in the horizontal and the vertical. We believe our EFSR diagnostics will be useful for this purpose, because, as we can see from Eq. (7.23), it is capable of estimating not only the diagonal elements (*i.e.*, variances) of $\mathbf{R}$ but also its off-diagonal element.

The use of EFSO and EFSR will thus allow us to greatly accelerate the development of DA methods for new observing systems.

Appendix

Chapter A: A semi-implicit modification to Lorenz N-cycle scheme and its application to an AGCM

A.1 Introduction

A unique feature of the atmospheric and oceanic sciences is that, unlike other fields of natural sciences, controlled experiments are difficult to perform. Accordingly, numerical experimentation has become an increasingly important methodology in meteorology and physical oceanography. A key role in numerical experimentation is played by atmospheric or oceanic models which numerically integrate hydrodynamic partial differential equations (PDEs) that describe the governing laws of geophysical fluid flows. There is thus a high demand for improvements of accuracy of such models.

One of the major challenges in designing numerical integration schemes for atmospheric models, in particular the Atmospheric General Circulation Models (AGCMs), is the so-called "stiffness" problem: the equations solved by AGCMs contain, not only the slower waves that are relevant to the actual weather phenomena, but also the faster waves that are of little meteorological interest. The phase speeds of the faster waves are typically an order-of-magnitude faster than those of the slower waves. In order to satisfy the Courant-Friedrichs-Lewy (CFL) stability condition, an explicit temporal integration scheme necessitates use of an overly short time step just to maintain stability, making the integration significantly more expensive. Most AGCMs resolve this issue by adopting a semi-implicit scheme which treats the terms responsible for the fast waves implicitly and the other terms explicitly (Robert, 1969); this treatment clears the CFL condition for the fast waves and thus allows an efficient integration with much longer time steps. The availability of this semi-implicit treatment is thus a prerequisite for a temporal integration scheme to be used in AGCMs.

In AGCMs, despite recent advancements in computational fluid dynamics, a rather simple centered-differencing scheme, commonly known as the leapfrog scheme, remains in wider use than any other schemes. The leapfrog scheme has several desirable properties, which include: ease of implementation, availability of a stable semi-implicit version, low cost in computational time, low memory consumption, and conservation of energy for a non-dissipative system.

The above desirable properties are, however, tainted by the following undesirable features (Durran, 1991): First, the scheme is unstable when applied to a system with dissipation. Second, being a three-time-level scheme, it necessitates special treatment at the very first several steps. Lastly, and most importantly, the leapfrog scheme produces, when applied to a nonlinear system, a spurious computational mode, which, if left unattenuated, results in time-splitting instability. In AGCMs, the first issue is typically dealt with by applying the leapfrog only to non-dissipative dynamics part; dissipative processes such as physics and damping are treated with separate schemes, such as the explicit Forward-Euler or implicit Backward-Euler scheme. This treatment unfortunately comes with the side-effect of making the scheme only first-order accurate. The second issue is commonly dealt with by running a two-time-level scheme, such as Forward-Euler scheme, at the very first step. The third issue can be resolved by filtering-out the computational mode by applying Robert-Asselin (RA) filter (Asselin, 1972; Robert, 1966); this treatment also introduces the side effect of degrading the scheme to first-order accuracy by damping not only the computational mode but also the physical mode. Despite these disadvantages, the leapfrog scheme with semi-implicit modification (Robert, 1969), combined with RA filter and separate treatment of dissipative part, has long been remained the most widely-used scheme for AGCMs. A better scheme for AGCMs which is free from these limitations has thus been sought after.

One way to achieve this goal is to alleviate the limitations by improving the classical RA-filtered leapfrog scheme. Recently, Williams (2009) proposed an improvement to the RA filter. The new filter, called Robert-Asselin-Williams (RAW) filter, preserves the second order accuracy of the leapfrog scheme without any significant increase in computational cost. The advantage of the RAW filter over the RA filter is confirmed also for the semi-implicit leapfrog scheme (Amezucua et al., 2011; Williams, 2011). Williams (2013) devised further improvements in this line, leading to schemes with even higher accuracy in amplitude (up to 7th-order; the phase error remains second-order). While these improved schemes effectively eliminate the undesirable artificial damping of physical modes, other shortcomings of the filtered leapfrog scheme remain unresolved. The efficacy of rendering the RA-filter second-order is also diminished if a first-order scheme is used for dissipative terms to suppress instability.

Attempts have also been made to seek for alternative schemes that are better suited for atmospheric and oceanic models. Multi-step schemes such as the Adams-Bashforth family of schemes, for example, can have the order of accuracy that is higher than the leapfrog without increasing computational expenses. Durran (1991) found, however, that while the three-step third-order Adams-Bashforth scheme is a viable alternative to the RA-filtered explicit leapfrog scheme, this scheme cannot replace the semi-implicit leapfrog scheme because the Adams-Bashforth scheme becomes unstable if it is combined with a semi-implicit scheme for fast modes. The Runge-Kutta family of schemes can also be more accurate than the leapfrog. Kar (2006) and Whitaker and Kar (2013), for example, have successfully developed semi-implicit versions of Runge-Kutta-type schemes and showed their advantages over the RA-filtered semi-implicit leapfrog scheme. Their schemes, however, consume more memory space than the leapfrog.

In 1971, Edward Norton Lorenz devised an ingenious temporal integration scheme, now called Lorenz N-cycle scheme, for a system of first-order ordinary differential equations (ODEs) (Lorenz, 1971). Its ingenuity resides in high order of accuracy, low computational expenses, and the economy of memory usage. It is self-starting (i.e., it does not require model states at multiple steps for initiating integration), computationally as efficient as the leapfrog, both in terms of speed and memory usage, but yet, can be of N-th order accurate at every N steps for an integer $N \leq 4$. Despite these advantages, Lorenz N-cycle remained rather obscure in atmospheric and oceanic sciences. Although there is at least one oceanic model that uses this scheme (Gent and Cane, 1989), it seems not to have been applied to an atmospheric model. In particular, a semi-implicit modification to this scheme has not been developed.

The aim of this study is to present a new semi-implicit modification to Lorenz N-cycle and show that this scheme can improve the accuracy of an AGCM. In designing our semi-implicit version, we put particular emphasis on preserving the low memory consumption of the original explicit scheme. As we describe in Section A.2, Lorenz N-cycle schemes can be thought of as a special subfamily of Runge-Kutta schemes. In fact, treating Lorenz 3-cycle as a special case of Runge-Kutta scheme, Whitaker and Kar (2013) proposed a semi-implicit formulation. Their scheme, however, is designed from a different motivation than ours: while our priority is in minimizing memory footprint, their priority was in ensuring stability. Consequently, the two schemes have different pros and cons, which is discussed in section 3.

This Appendix is organized as follows: Section A.2 concisely summarizes the algorithm of Lorenz N-cycle and discusses its advantages, especially in comparison with the traditional leapfrog scheme. Section A.3 first describes the traditional semi-implicit modification to the leapfrog scheme and then presents the formulation of our semi-implicit modification to Lorenz N-cycle. It then gives the analysis of its accuracy and stability. Section A.4 briefly describes the AGCM, called SPEEDY model, whose dynamical core is used to test our schemes. It also describes our verification method which is based on a standardized baroclinic-wave dynamical core test, whose results are presented in Section A.5. Section A.6 concludes the Appendix with a summary and an outlook for our future research.

A.2 Explicit Lorenz N-cycle

This section describes the algorithm of explicit Lorenz N-cycle and discusses its advantages, especially in comparison to the traditional RA-filtered leapfrog scheme.

A.2.1 The algorithm

Let us consider a problem of numerically integrating the following system of ODEs:

$$\frac{dx}{dt} = F(x) \tag{A.1}$$

where $x = (x_1(t), \dots, x_M(t))$ is an M-dimensional vector function of t and $F(x) = (F_1(x_1, \dots, x_M), \dots, F_M(x_1, \dots, x_M))$ is a function from $\mathbb{R}^M \to \mathbb{R}^M$. In trying to find an economical scheme, Lorenz (1971) derived two "isomeric" versions of schemes for the above problem. Using the elegant notation introduced by Purser and Leslie (1997), the algorithm of one version of the schemes, which we refer to by "version A" hereafter, can be schematically written as following:

<u>N-cycle A</u>

```
$w^0 = 1,$                                   (A.2)
$w^k = \frac{N}{N-k}\ (k = 1, \ldots, N-1)$   (A.3)
**do** $k = 0, \ldots$
    $w \leftarrow w^{\bmod (k,N)}$             (A.4)
    $G \leftarrow wF(x) + (1-w)G$             (A.5)
    $x \leftarrow x + G\Delta t$              (A.6)
**end do**
```

Similarly, the other version, which we call "version B," can be schematically written as following:

<u>N-cycle B</u>

```
$w^0 = 1,$                                   (A.7)
$w^k = \frac{N}{k}\ (k = 1, \ldots, N-1)$     (A.8)
**do** $k = 0, \ldots$
    $w \leftarrow w^{\bmod (k,N)}$             (A.9)
    $G \leftarrow wF(x) + (1-w)G$             (A.10)
    $x \leftarrow x + G\Delta t$              (A.11)
**end do**
```

Note that the two versions differ only in the ordering of the "weight" coefficients w^k. In implementing these schemes, we only need to store the two arrays of M words, x and G. Also, they require only one evaluation of the tendency term $F(x)$ per time step.

The mathematical idea behind the above algorithms is simple: to "reuse" the previously computed tendencies ($F(x^{k-j}), j = 1, 2, \ldots, k-1$, where x^{k-j} denotes the value of x at the $(k-j)$-th step of each cycle) by forming a weighted average of them to produce a tendency which yields the highest order of accuracy after the completion of the N-th step, under the constraint that each intermediate step retains at least first order accuracy. In this sense, Lorenz N-cycle can be regarded as a special instance of the Runge-Kutta family. In fact, for example, Lorenz 4-cycles, both versions A and B, are equivalent to the classical 4-step 4th-order Runge-Kutta scheme with time step $4\Delta t$ if F is linear.

If F is linear, at every N steps, both versions A and B give the Taylor expansion with respect to Δt of the true solution truncated at the $N+1$-th order term. Thus, for a linear case, if we only look at results at every N steps, the two versions are identical N-th order schemes.

If F is nonlinear, the accuracy of the N-cycle schemes reduces to second order for $N \geq 3$. However, for $N = 3$ and for $N = 4$, the $N+1$-th order term in the truncation error of the versions A and B can be shown to be of the same magnitude with opposite signs. Thus, for these values of N, N-th order accuracy can be attained by running both A and B cycles simultaneously and then averaging the predictions, at the expense of doubling the computational cost in both speed and memory consumption.

In order to avoid doubling of computational cost, Lorenz (1971) proposed to use the versions A and B in a suitable alternating sequence, based on the intuition that the errors of the two versions should tend to cancel each other. In fact, Lorenz (1971) claims, without proof, that, for $N = 3$, true 3rd-order accuracy can be retained even for a nonlinear case by alternating versions A and B. Likewise, it is claimed that, for $N = 4$, full 4th-order accuracy can be achieved by forming a $4N(= 16)$-cycle of the sequence A,B,B,A. Numerical computations for a simple nonlinear ODE performed by Purser (2007) corroborates Lorenz's claim for 3-cycle, but there seems to have been no work that supports or denies the claim for $N = 4$. In Section A5.3, we present a result for an AGCM that partly corroborates this claim for $N = 4$.

The stability of Lorenz N-cycle schemes is investigated by Lorenz (1971) for the case of scalar linear F and by Israeli and Gottlieb (1974) for linear partial differential equations (PDEs) discretized in space by centered finite differencing. Unlike the leapfrog scheme which is stable for non-dissipative (hyperbolic) systems but unstable for dissipative (parabolic) systems, Lorenz N-cycle schemes are shown to be reasonably stable for both hyperbolic and parabolic systems. For example, as we stated above, Lorenz 4-cycle with a time step of Δt and the classic Runge-Kutta 4-th order scheme with the time step of $4\Delta t$ for a linear system yield mathematically equivalent results and thus their stability conditions are identical.

A.2.2 Advantages of Lorenz N-cycle

The principal advantage of Lorenz N-cycle, besides its high-order accuracy, is its computational efficiency in terms of both speed and memory consumption. It requires only one evaluation of $F(x)$ per time step which, in most cases, is the most computationally demanding part. Also, the scheme consumes only $2M$ words of memory. Thus, Lorenz N-cycle has the same computational cost as the widely-used leapfrog scheme. Compared to the 4th-order Runge-Kutta scheme which is accurate but too expensive for the purpose of AGCMs, the Lorenz N-cycle consumes less than half the memory and can run 4 times faster.

Another great advantage of Lorenz N-cycle is the absence of computational modes: the Lorenz N-cycle, being a two time-level method rather than a three time-level method like leapfrog scheme, does not suffer from the presence of computational modes. This feature proves to be particularly useful for nonlinear systems for which computational modes can grow exponentially, causing divergence of numerical solution from the actual solution. Being a two-time level (or single step) scheme also facilitates the initialization process. Unlike schemes with three or more time levels such as the leapfrog or Adams-Bashforth method, Lorenz N-cycle does not require any special treatment for the initial step(s).

A.3 Semi-Implicit Modification

As we discussed in the introduction, the equations solved by AGCMs are stiff: the external inertia-gravity waves (also known as Lamb waves), which are contained in the solutions of these equations but are of little meteorological importance, exhibit very fast phase speed (approximately ~ 300 m/s), whereas, waves which are relevant to the actual weather phenomena, such as internal inertia-gravity waves and Rossby waves, exhibit an order-of-magnitude slower phase speed. Due to the CFL restriction, the presence of these fast waves requires an order-of-magnitude shorter time-stepping than what is otherwise required to resolve the slower, meteorologically meaningful waves. In AGCMs, it is customary to circumvent this issue by using semi-implicit technique (Robert, 1969).

This section presents our new semi-implicit modification to Lorenz N-cycle scheme and discuss its accuracy and stability. In deriving our semi-implicit scheme, we utilize a somewhat non-conventional notation in which the tendency term (not the model state itself) is modified to account for semi-implicit treatment. To familiarize the readers with our tendency-based notation, the classical semi-implicit leapfrog scheme is presented in this notation in Section A.3.1. Section A.3.2 describes our semi-implicit formulation of Lorenz N-cycle. Section A.3.3 then discusses its accuracy and stability for illustrative linear cases.

A.3.1 Semi-implicit leapfrog scheme

Consider integration of an equation of the form:

$$\frac{dx}{dt} = F^E(x) + L^I x \tag{A.12}$$

where $F^E : \mathbb{R}^M \to \mathbb{R}^M$ is a nonlinear function and L^I is an $M \times M$ matrix. It is assumed that the term $L^I x$ is responsible for the fast external inertia-gravity waves. The semi-implicit modification to the leapfrog scheme which was originally introduced by Robert (1969) takes the following form:

$$\frac{x^{n+1} - x^{n-1}}{2\Delta t} = F^E(x^n) + L^I \left(\alpha x^{n+1} + (1-\alpha) x^{n-1} \right) \tag{A.13}$$

where x^n signifies the predicted state at the n-th step and $0 \leq \alpha \leq 1$ is a "centering factor." $\alpha = 1/2$, $\alpha = 1$ and $\alpha = 0$ correspond, respectively, to Crank-Nicolson, backward Euler, and forward Euler scheme.

To solve Eq.(A.13) for x^{n+1}, let us first define the "tendency" δx by

$$\delta x \quad = \quad \frac{x^{n+1} - x^{n-1}}{2\Delta t} \tag{A.14}$$

and express x^{n+1} on the right hand side as $x^{n+1} = x^{n-1} + 2\Delta t \delta x$. Then, substituting this expression to Eq.(A.13), we have:

$$\begin{aligned} \delta x \quad &= \quad F^E(x^n) + L^I x^{n-1} + 2\alpha \Delta t L^I \delta x && \text{(A.15)} \\ \Leftrightarrow \delta x \quad &= \quad (I - 2\alpha \Delta t L^I)^{-1} (F^E(x^n) + L^I x^{n-1}) && \text{(A.16)} \end{aligned}$$

where I is the identity matrix. Once δx is obtained, the integration can be completed by

$$x^{n+1} = x^{n-1} + 2\Delta t \delta x \tag{A.17}$$

In summary, to solve Eq.(A.13) for x^{n+1}, we first evaluate the nonlinear tendency $F^E(x^n)$ at the central step, and then evaluate and add the linear tendency $L^I x^{n-1}$ at the older step. We then multiply it by the inverse matrix $(I - 2\alpha\Delta t L^I)^{-1}$ and finally integrate the equation by Eq.(A.17). Note that the matrix $(I - 2\alpha\Delta t L^I)$ is constant as long as the time step Δt is unchanged, so that the matrix inversion needs to be carried out only once for the whole integrations.

A.3.2 Formulation of the semi-implicit Lorenz N-cycle

The important advantage of Lorenz N-cycle schemes over other schemes such as the leapfrog or the Runge-Kutta scheme is its economy in terms of memory consumption. Thus, in designing a semi-implicit modification to it, we sought to preserve this favorable property. Our proposed scheme achieves this goal by applying tendency adjustment similar to Eq.(A.16) on each step of the N-cycle:

```
do $k = 0, \ldots$
    $w \leftarrow w^{\mathrm{mod}(k,N)}$                                (A.18)
    $G \leftarrow wF^E(x) + (1 - w)G$                        (A.19)
    $\delta x = (I - \alpha\Delta t L^I)^{-1}(G + L^I x)$             (A.20)
    $x \leftarrow x + \Delta t \delta x$                            (A.21)
end do
```

Note that no additional variables are introduced in this semi-implicit formulation. Also, unlike other semi-implicit Runge-Kutta type schemes (e.g., Whitaker and Kar, 2013), this scheme involves only one matrix inversion, which allows simplicity in implementation.

A.3.3 Stability and accuracy analysis

Semi-implicit time-stepping schemes are traditionally examined by applying them to the following split-frequency linear oscillation equation (Durran, 1991; Whitaker and Kar, 2013; Williams, 2011):

$$\frac{dx}{dt} = i\omega_L x + i\omega_H x \tag{A.22}$$

where the first and second terms on the right hand side correspond, respectively, to $F^E(x)$ and $L^I x$ in Eq.(A.12).

By carrying out the algorithm, we can show, as in the case of explicit N-cycle, that, for a linear system, the versions A and B give identical expression at every N steps. The truncation errors are:

$$\frac{x^N - x^{\text{Exact}}}{x^0} = \frac{1}{2N}(1-2\alpha)\omega_H(\omega_H+\omega_L)(N\Delta t)^2 + O(\Delta t^3) \tag{A.23}$$

Thus, the semi-implicit Lorenz N-cycle can be of second order by taking $\alpha = 1/2$ (i.e., Crank-Nicolson scheme).

Stability of these schemes for each ω_L and ω_H can be visualized by plotting the modulus of the corresponding amplification factor $|A|$ as a function of $(\omega_L \Delta t, \omega_H \Delta t)$. The scheme is unstable if $|A|$ exceeds unity. In order for a fair comparison among different values of N, we define the average amplification factor per step A by

$$A \quad := \quad \sqrt[N]{x^N / x^0}. \tag{A.24}$$

Figure A.3.3 shows $|A|$ of Crank-Nicolson ($\alpha = 1/2$) semi-implicit Lorenz N-cycle schemes for values of N from $N = 1$ to $N = 6$. The areas of instability are filled with light gray ($1 < |A| < 1.01$) or with dark gray ($|A| > 1.01$). In general, semi-implicit techniques are applied when the oscillations produced by the implicitly-treated term are faster than those produced by the explicitly-treated term. Thus, in interpreting Figure A.3.3, we should focus on the area above the thick red lines ($|\omega_H| > |\omega_L|$). From panels (a) and (b), we see that the Lorenz 1- and 2-cycles are stable when the two frequencies ω_H and ω_L are of opposite signs and the magnitude is larger for ω_H. Unfortunately, however, these schemes are unconditionally unstable if ω_H and ω_L are of same sign. The 3-cycle scheme (Figure A.3.3c) exhibit better stability for cases where ω_H and ω_L are of same sign, but it also introduces weak instability for regions where the two frequencies are of opposite signs. The 4-cycle (Figure A.3.3d) has larger stability region than the 3-cycle, but increasing N further does not improve the situation; the 5- and 6-cycles (Figure A.3.3e and A.3.3f) show stability regions that are smaller than that of the 4-cycle. Values of N larger than 6 (from 7 to 12; not shown) result in even smaller stability regions. For these reasons, hereafter we focus only on Lorenz 3- and 4-cycles.

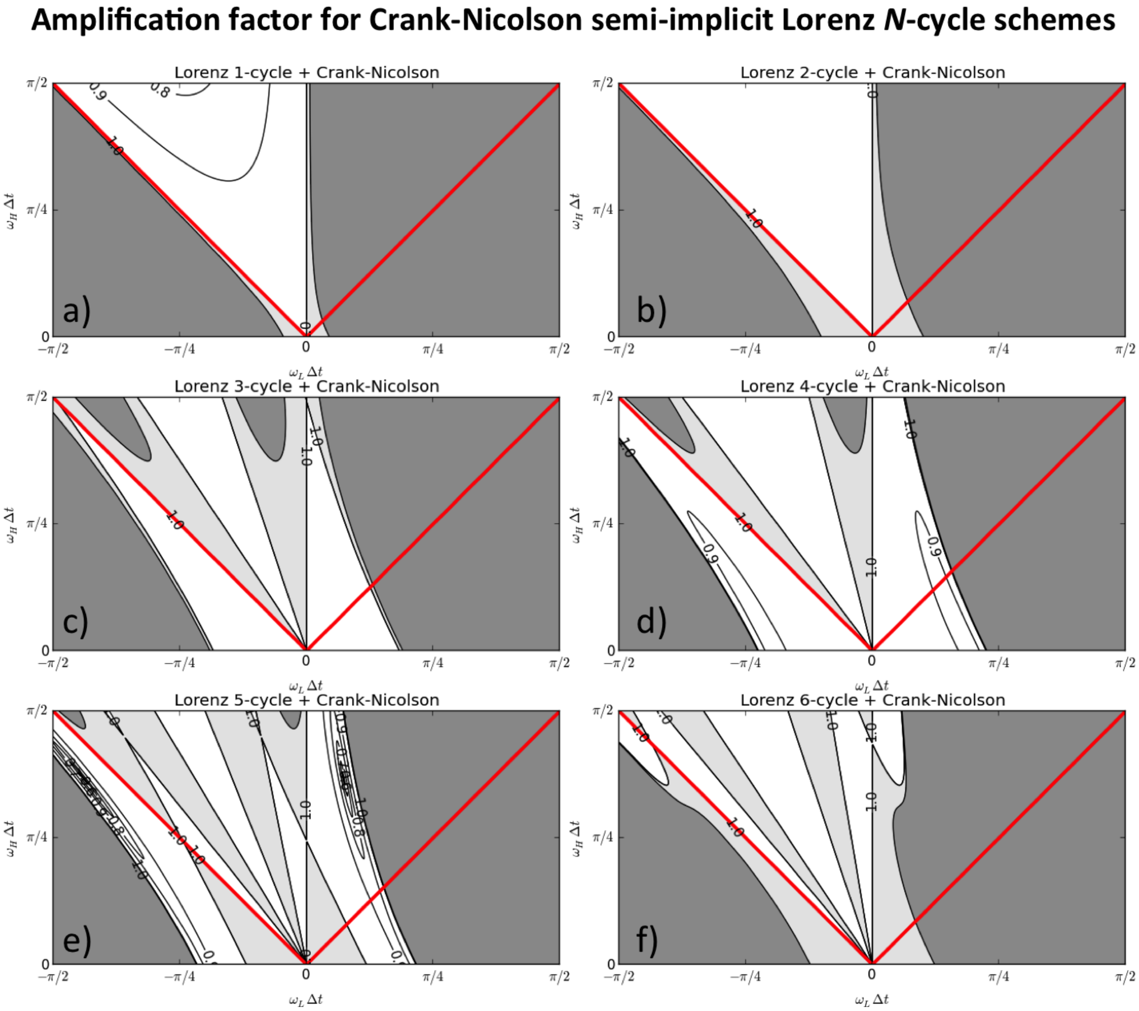


Figure A.1: The modulus of the average amplification factor per step $|A|$ defined by Eq. (A.24) for the Crank-Nicolson semi-implicit Lorenz N-cycle schemes applied to the scalar split-frequency problem Eq. (A.22). Panels (a)–(f) correspond, respectively, to $N = 1, 2, 3, 4, 5$ and 6. The contour intervals are 0.1. However, regions with $1 < |A| < 1.01$ are filled with light gray to show where the schemes are slightly unstable. Regions of instability with $|A|$ exceeding 1.01 are filled with dark gray. Red thick lines in each panel represent $|\omega_H| = |\omega_L|$; we are interested in the region above these lines.

Figure A.2 shows the relative phase errors (in percent) of Lorenz (a) 3- and (b) 4-cycles. Here, the relative phase errors are defined as:

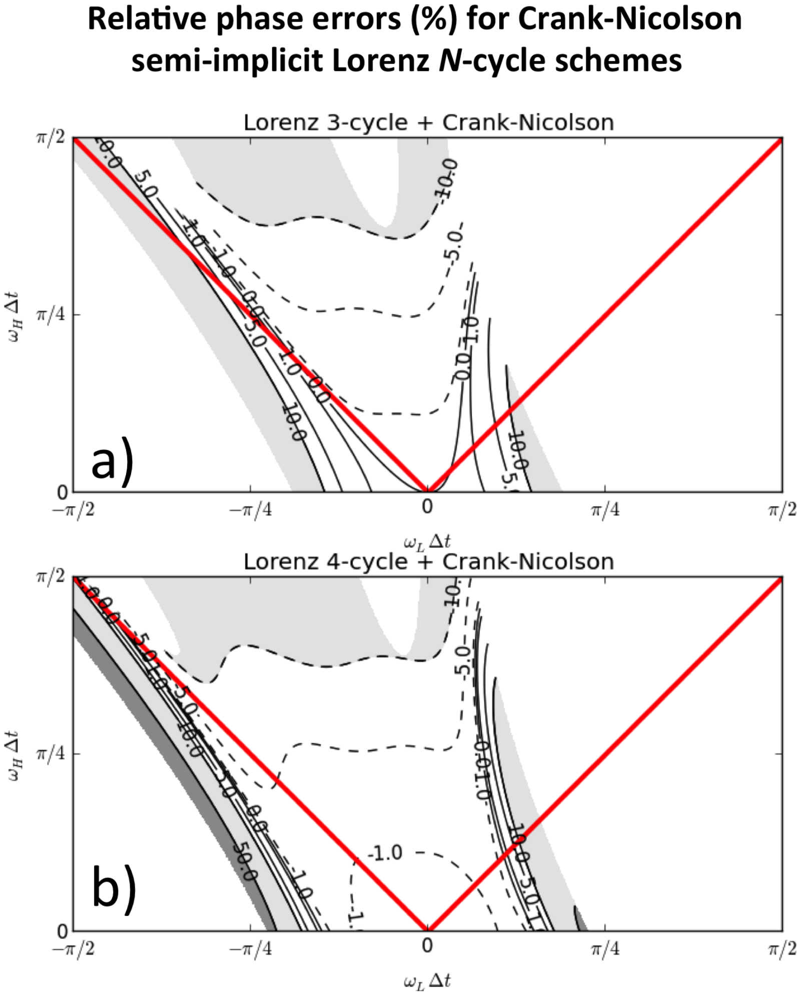
$$\frac{\arg A - (\omega_L + \omega_H)\Delta t}{(\omega_L + \omega_H)\Delta t} \times 100 \tag{A.25}$$

Lorenz 3- and 4-cycles show similar phase errors. The errors are small (less than 10%) in most of the stability regions except very near to the boundaries. Interestingly, the phase errors are predominantly negative, which means the oscillations in the numerical solutions tend to be slower than the exact solution. This is consistent with our intuition that semi-implicit method stabilizes the scheme by slowing down high-frequency waves (Kalnay, 2003).

Stability analysis based on the split-frequency equation Eq. (A.22) provides us with an insight on how the schemes would behave for a pure, non-dissipative oscillating system. However, it is also important to examine how the schemes behave for a system with dissipation because most geophysical fluid systems, including AGCMs, contain dissipative terms. Following Kar (personal communication, 2013), we account for dissipation in the stability analysis by introducing imaginary component to ω_L in Eq.(A.22):

$$\frac{dx}{dt} = i\omega_L x + i\omega_H x = \{iRe(\omega_L)x - Im(\omega_L)x\} + i\omega_H x \tag{A.26}$$

As in the case of Eq.(A.22), we integrate this equation with the semi-implicit Lorenz N-cycle schemes treating the ω_L-term explicitly and ω_H-term implicitly, and



Relative phase errors (%) for Crank-Nicolson semi-implicit Lorenz *N*-cycle schemes

Lorenz 3-cycle + Crank-Nicolson

Lorenz 4-cycle + Crank-Nicolson

Figure A.2: As in Figure A.3.3, but for the average phase error per step defined by Eq. (A.25) for the Crank-Nicolson semi-implicit Lorenz (a) 3-cycle and (b) 4-cycle schemes, applied to the scalar split-frequency problem Eq. (A.22). The contour levels are $\pm50\%, \pm10\%, \pm5\%, \pm1\%$ and 0. Non-negative and negative contours are drawn, respectively, with solid and dashed lines. Regions where the magnitude of the relative errors exceeds 1% and 10% are filled, respectively, with light and dark gray. Red thick lines in each panel represent $|\omega_H| = |\omega_L|$; we are interested in the region above these lines. Phase errors are drawn only for areas where the modulus of the amplification factor (shown in Figure A.3.3c,d) is smaller than 1.01.

examine the stability by looking at the modulus of the amplification factor $|A|$. We first fix the frequency of the implicitly treated term, $\omega_H \Delta t$ to a prescribed value and draw the contour of $|A|$ on a complex plane for $\omega_L \Delta t$. Figure A.3.3 shows the stability regions of the semi-implicit Lorenz (a) 3- and (b) 4-cycles. Each curve represents the boundary of stability region for the corresponding value of $\omega_H \Delta t$ shown in the legend. For both 3- and 4-cycles, the instability that is present in the non-dissipative case (Figure A.3.3c,d) for negative $Re(\omega_L)\Delta t$ and positive $\omega_H \Delta t$ can be suppressed by very small damping ($Im(\omega_L)\Delta t \lesssim 0.025$). If we focus on the stability of 3- and 4-cycles for small damping ($Im(\omega_L)\Delta t \lesssim 0.1$), we find that 4-cycle has broader range of stability than 3-cycle.

From the above stability analysis we conclude that, with our semi-implicit formulation, Lorenz 4-cycle is more stable than Lorenz 3-cycle. For this reason, in the experiments with an AGCM described in later sections we focus only on Lorenz 4-cycle.

A.4 Experimental setup

A.4.1 SPEEDY model

In this study, we implement and test the semi-implicit Lorenz N-cycle in a simplified low-resolution AGCM known as Simplified Parametrizations, Primitive Equation Dynamics (SPEEDY) model (Molteni, 2003). This model was originally designed as a climate model and thus only had outputs of one-month averages. Miyoshi (2005) modified it to produce 6-hourly output of snapshot values so that

Stability regions for semi-implicit Lorenz *N*-cycles schemes applied to damped oscillation equation

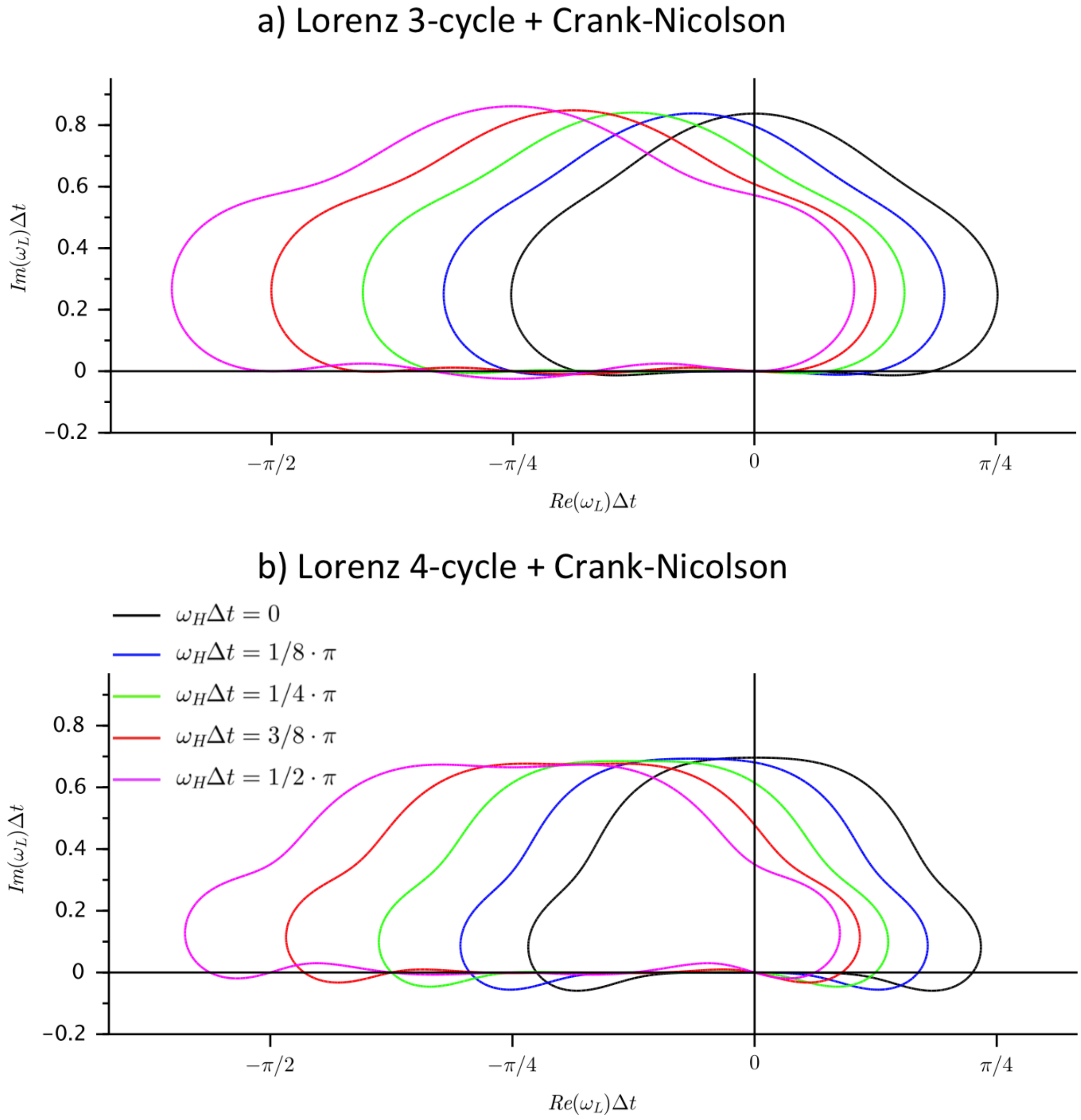


Figure A.3: Stability regions of semi-implicit Lorenz (a) 3- and (b) 4-cycles applied to the split-frequency damped oscillation problem Eq. (A.26) for different values of $\omega_H \Delta t$. Each curve represent the contour of $|A = 1|$ for $\omega_H \Delta t$ labeled in the legend. The scheme is stable in the regions encircled by theses curves.

it can be used in data assimilation researches. This model is a primitive equation model with T30L7 resolution; its horizontal discretization is spectral representation with respect to the spherical harmonics triangularly truncated at the total wavenumber of 30, and its vertical discretization is finite differencing on 7 layers in σ-coordinate system. In grid-point space, it has 96 longitudinal points and 48 latitudinal points. For temporal discretization, the SPEEDY model uses the standard RA-filtered semi-implicit leapfrog scheme. However, due to the intrinsic instability of leapfrog scheme against dissipative processes, SPEEDY model treats physical parametrizations with first-order Forward Euler scheme (with the time step of $2\Delta t$) to prohibit numerical instability. Moreover, it treats horizontal spectral bi-harmonic dampings for momentum and temperature equations with implicit Backward Euler scheme to achieve further stabilization. The scheme can be written in pseudo-code as:

```
do $k = 1, \ldots$
    $\delta x \leftarrow F_{\mathbf{Dyn}}^{\mathbf{NL}}(x^k) + F_{\mathbf{Phys}}(x^{k-1}) + L_{\mathbf{Dyn}} x^{k-1}$    (A.27)
    $\delta x \leftarrow (1 - 2\alpha\Delta t L_{\mathbf{Dyn}})^{k-1} \delta x$    (A.28)
    $\delta x \leftarrow \left(I + \kappa(2\Delta t)\nabla^4\right)^{k-1} \left(\delta x - \kappa(2\Delta t)\nabla^4 x^{k-1}\right)$    (A.29)
    $x^{k+1} \leftarrow x^{k-1} + 2\Delta t \delta x$    (A.30)
    $x^k \leftarrow x^k + \nu(x^{k+1} - 2x^k + x^{k-1})$    (A.31)
    $t \leftarrow t + \Delta t$    (A.32)
    $x^{k-1} \leftarrow x^k, \quad x^k \leftarrow x^{k+1}$    (A.33)
end do
```

where $F_{\mathbf{Dyn}}^{\mathbf{NL}}(x)$ represents the nonlinear part of the tendency terms associated with dynamical process, $L_{\mathbf{Dyn}} x$ the linear part, and $F_{\mathbf{Phys}}(x)$ the tendency terms from physical parametrizations. κ in Eq.(A.29) represents the diffusion coefficient for the bi-harmonic hyper-diffusion. Note that, since SPEEDY model is a spectral model based on spherical harmonics, in spectral space, horizontal bi-harmonic operator ∇^4 and its inverse both become simple scalar multiplication. The smoothing parameter ν in the RA filter Eq.(A.31) is set to 0.05. The centering parameter for the implicit component α is 1/2 for Crank-Nicolson and 1 for Backward-Euler. The use of implicit Backward-Euler scheme for harmonic damping in Eq.(A.29) reduces the formal accuracy of the scheme to only first order. This debasement might be justifiable for the SPEEDY's default dynamical core since its formal accuracy is only first order due to the use of RA filter. In our study, however, this is undesirable, because our goal is to achieve second-order accuracy by adopting our semi-implicit version of Lorenz N-cycle. Also, as we have shown in Figure A.3.3, unlike leapfrog scheme, Lorenz N-cycle does not suffer from instability when applied to a dissipative system. For this reason, in implementing our schemes to the SPEEDY's dynamical core, we modified it so that the harmonic dampings are included in the term $F^{\mathbf{NL}}_{\mathbf{Dyn}}(x)$ and thus are treated explicitly.

The SPEEDY model includes a simplified set of physical parametrizations whose descriptions can be found in the appendix of Molteni (2003). Simplified physical process and coarse resolution enable the SPEEDY model to be integrated very fast. Despite such simplification, this model is able to produce realistic simulations of a wide range of the atmospheric phenomena including precipitation, mid-latitude synoptic features and climatology. As we describe in the next section, however, we switch-off all physical parametrizations in the SPEEDY model to evaluate its performance under the framework of a dynamical core test.

A.4.2 Jablonowski-Williamson dynamical core test

In our study, we are interested in assessing how our semi-implicit Lorenz N-cycle schemes behave for various values of time step Δt, in particular, compared to the conventional RA-filtered leapfrog scheme. For this purpose, physical parametrizations are undesirable because some of them are designed to work well only for a specific time step. Adjustment processes such as large scale condensation violates the assumption that the tendency term $F(x)$ is smooth (*i.e.*, has continuous first derivatives), which also complicates the interpretation of the results. Thus, we test our schemes under the framework of a dynamical core test.

Several standardized benchmarks have been proposed for dynamical cores of AGCMs (e.g., Boer and Denis, 1997; Held and Suarez, 1994). Among those standardized test cases we adopt the baroclinic-wave test case proposed by Jablonowski and Williamson (2006). Unlike other previously proposed test cases which are primarily focused on quantifying long-term or climatological performance, this baroclinic-wave test case allows us to evaluate the performance of a dynamical core in an initial-value problem.

In the baroclinic-wave test case of Jablonowski and Williamson (2006), the initial and boundary conditions are designed so that a train of unstable baroclinic waves develops from a small disturbance superposed on a zonally uniform steady state. This specific configuration enables us to measure the performance of a model in terms of its ability to simulate baroclinic waves. Precise specification of the initial and boundary conditions for this test case is given in the next subsection.

A.4.3 The Initial and Boundary Conditions for the Baroclinic-Wave Test Case

The initial condition for the baroclinic-wave test case of Jablonowski and Williamson (2006) comprises of two parts: first, a zonally symmetric state which is an analytic steady-state solution of the primitive equations, and second, a horizontally localized disturbance to the steady state which triggers development of a baroclinic wave train. We first describe the steady state and then describe the disturbance, followed by the description of the boundary condition.

First, we define an intermediate vertical coordinate σ_v by

$$\sigma_v = (\sigma - \sigma_0) \times \frac{\pi}{2} \tag{A.34}$$

with $\sigma_0 = 0.252$. Here $\sigma \in [0, 1]$ represents the σ-vertical coordinate. The zonal wind of the steady state is defined as:

$$\bar{u}(\varphi, \sigma) = u_0 \cos^{3/2} \sigma_v \sin^2 (2\varphi) \tag{A.35}$$

with $u_0 = 35\mathrm{m\ s}^{-1}$. Here φ represents the latitude in radian. The meridional wind $\bar{v}$ is set to zero everywhere:

$$\bar{v}(\varphi, \sigma) = 0. \tag{A.36}$$

The temperature is defined by

$$\begin{aligned} \bar{T}(\varphi, \sigma) \quad = \quad & \langle T(\sigma) \rangle + \frac{3}{4} \frac{\sigma \pi u_0}{R_d} \sin \sigma_v \cos^{1/2} \sigma_0 \\ & \times \left[\left\{ -2 \sin^6 \varphi \left(\cos^2 \varphi + \frac{1}{3} \right) + \frac{10}{63} \right\} 2 u_0 \cos^{3/2} \sigma_0 \right. \\ & \left. + \left\{ \frac{8}{5} \cos^3 \varphi \left(\sin^2 \varphi + \frac{2}{3} \right) - \frac{\pi}{4} \right\} a \Omega \right] \end{aligned} \tag{A.37}$$

where $R_d = 289.0\text{J kg}^{-1}\text{ K}^{-1}$ is the ideal gas constant for dry air, $a = 6.371229 \times 10^6$ m is the mean radius of the Earth, with the horizontal average temperature $\langle T(\sigma)\rangle$ given by

$$\langle T(\sigma)\rangle = \begin{cases} T_0\sigma^{R_d\Gamma/g} & \text{for } 1 \geq \sigma \geq \sigma_t, \\ T_0\sigma^{R_d\Gamma/g} + \Delta T\left(\sigma_t - \sigma\right)^5 & \text{for } \sigma_t > \sigma, \end{cases} \tag{A.38}$$

where $g = 9.80616$ m s^{-2} is the gravitational acceleration, $\Gamma = 0.005$ K m^{-1} is the temperature lapse rate, $\Delta T = 4.8 \times 10^5$ K, and $\sigma_t = 0.2$ is the tropopause level. The surface pressure p_s is globally set to a constant value:

$$p_s = 10^5\text{Pa} \tag{A.39}$$

which completes the specification of the steady state.

On top of this steady state, a horizontally localized disturbance of the zonal wind $u'(\lambda, \varphi, \sigma)$ centered at $(\lambda_c, \varphi_c) = (\pi/9, 2\pi/9)$ (=(20 °E,40 °N)) is superposed to form the complete initial condition (other prognostic variables are not touched). The disturbance in the zonal wind u' is specified by:

$$u'(\lambda, \varphi, \sigma) \quad = \quad u_p \exp\left\{-\left(\frac{r}{R}\right)^2\right\} \tag{A.40}$$

with radius $R = a/10$ and $u_p = 1\text{m s}^{-1}$. The great circle distance r from the center

(λ_c, φ_c) is defined by

$$r \;=\; a\cos^{-1}\left\{\sin\varphi_c \sin\varphi + \cos\varphi_c \cos\lambda - \lambda_c\right\}. \tag{A.41}$$

Finally, we describe the boundary condition. The orography (or surface height) z_s is also zonally uniform and is specified by:

$$\begin{aligned} z_s(\lambda, \varphi) \;=\;& u_0 \cos^{3/2}\left(1 - \sigma_0\right) \\ & \times \left[\left\{-2\sin^6\varphi\left(cos^2\varphi + \frac{1}{3}\right) + \frac{10}{63}\right\} u_0 \cos^{3/2}(1 - \sigma_0) \right. \\ & \left. + \left\{\frac{8}{5}\cos^3\varphi\left(\sin^2\varphi + \frac{2}{3}\right) - \frac{\pi}{4}\right\} a\Omega\right] \end{aligned} \tag{A.42}$$

For the upper boundary condition, Jablonowski and Williamson (2006) requires that no Rayleigh friction near the model top be applied. Thus, in our experiment, we switched-off the Rayleigh friction which, in SPEEDY model, is applied to only the horizontal winds at the topmost layer.

A.4.4 Verification method

An analytic solution is not known to this problem. Jablonowski and Williamson (2006) thus provides a set of reference solutions that can be used by the users of this test case. Their reference solutions are produced by integrating multiple different high-resolution models. The differences among these reference solutions can be used to quantify their uncertainties. We do not use their reference solutions, however, because our main focus of this study is on temporal integration schemes:

comparison with solutions from a higher-resolution model would complicate the interpretation by introducing spatial discretization as another factor. In order to facilitate a fair comparison among different temporal integration schemes, we produce a reference solution by implementing the traditional explicit 4th-order Runge-Kutta (RK4) scheme to the SPEEDY model and then integrating it with a very small time step $\Delta t = 10$ s. The reference solution thus obtained is regarded and used as "truth" in our verification.

Following Jablonowski and Williamson (2006), for quantitative comparison, we use the difference of surface pressure p_s from the reference RK4 solution measured with l^2-norm:

$$l_2(p_s(t)) \quad := \quad \left[\frac{1}{4\pi} \int_0^{2\pi} \int_{-\pi/2}^{\pi/2} \left\{ p_s(\lambda, \varphi, t) - p_s^{\mathrm{RK4}}(\lambda, \varphi, t) \right\}^2 \cos\varphi d\varphi d\lambda \right]^{1/2} \tag{A.43}$$

where λ, φ and t denote, respectively, the longitude, latitude and forecast time. p_s^{RK4} denotes the reference solution produced from RK4 scheme described in the last paragraph. Jablonowski and Williamson (2006) reports that choice of verified variables (temperature or vorticity vertically interpolated to some specified pressure level, in addition to surface pressure) and error norms (l^1 and l^∞, in addition to l^2) do not sensitively affect results of verification.

A.4.5 Temporal Integration Schemes implemented to SPEEDY model

The temporal integration schemes implemented to SPEEDY model in our study are summarized in Table A.1. This table also defines the abbreviated names of each scheme that we use in the following sections. As we described in Section A.4.1, SPEEDY model, by default (ImLF+CN), adopts a somewhat complicated combination of different schemes. This is necessary because of the limited stability of leapfrog scheme when it is applied to dissipative terms. In contrast, our Lorenz N-cycle schemes (ImL4+CN, ImL4+BE and ExL4) treat all the terms consistently with a single scheme; this is possible because Lorenz N-cycle, being a variant of Runge-Kutta scheme, can better tolerate dissipation.

Since Lorenz N-cycle scheme has two versions (which we refer to by A and B), we can construct several variants of our Lorenz N-cycle schemes by using one of the two versions or by using them in alternating sequences. For each of our Lorenz N-cycle schemes (ImL4+CN, ImL4+BE and ExL4), we tried versions A, B, AB and ABBA. For example, the ImL4+CN scheme with Lorenz 4-cycle version ABBA is referred to, hereafter, by ImL4+CN-ABBA.

Table A.1: List of temporal integration schemes implemented to SPEEDY model in this study. Each scheme treats different terms of the right hand side of the governing equation differently. In the leftmost column, “Ext.Grav.Wave” signifies the terms that are responsible for the external gravity waves; “Diffusion” signifies spectral bi-harmonic diffusion; “Other Dynamics” signifies the rest of the terms that arise from dynamical processes; “Phys” signifies tendencies from physical parametrizations. “N/A” in each column of the last row (“Filter”) indicates that no filters are applied. (Im) or (Ex) in each item indicate that the scheme that follows them is, respectively, implicit or explicit.

Name	**ImL4+CN**	**ImL4+BE**	**ExL4**	**ImLF+CN**	**ImLF+BE**	**RK4**
Ext.Grav.Wave Diffusion Other Dynamics Phys. Param.	(Im) Crank-Nicolson (Ex) Lorenz 4-cycle	(Im) Backward-Euler	(Ex) Lorenz 4-cycle	(Im) Crank-Nicolson (Im) Backward-Euler (Ex) Leapfrog (Ex) Forward-Euler	(Im) Backward-Euler	(Ex) RK4
Filter	N/A			Robert-Asselin		N/A

A.5 Results

This section presents the results of Jablonowski-Williamson baroclinic-wave test case described in the previous section. Particular emphasis is placed on comparison between our semi-implicit Lorenz 4-cycle scheme and the traditional RA-filtered semi-implicit leapfrog scheme. To grasp the qualitative features of these schemes, we first show the snapshot pictures of the evolution of the baroclinic waves. We then show the results of the estimation of the orders of accuracy.

A.5.1 Snapshots

Figure A.4 shows the snapshots of surface pressure field at the 9th day of integration for different integration schemes. In the reference solution produced from RK4 scheme with a small time step $\Delta t = 10$ s (Figure A.4a), a deep low with a minimum of less than 960hPa develops on a grid (61.23^{o}N, 146.25^{o}W). The SPEEDY's default scheme ImLF+CN with the time step of $\Delta t = 1200$ s (Figure A.4b) also produces a low with its minimum at the same grid as the reference solution, but with a weaker intensity. With this traditional leapfrog scheme, successful simulation of the intensity of the low requires a much smaller time step, as shown in Figure A.4c (ImLF+CN with $\Delta t = 10$ s). On the other hand, our new ImL4+CN-A scheme successfully produces the deep low even with the longer time step of $\Delta t = 1200$ s (Figure A.4d). Other versions of ImL4+CN schemes (versions B, AB and ABBA; not shown) also produced solutions that are visually indistinguishable to Figure A.4d. This qualitative result suggests that ImL4+CN is the most advantageous scheme in that it alone successfully reproduces the deep low with the larger time step. Quantitative results shown in the following section strongly supports this supposition.

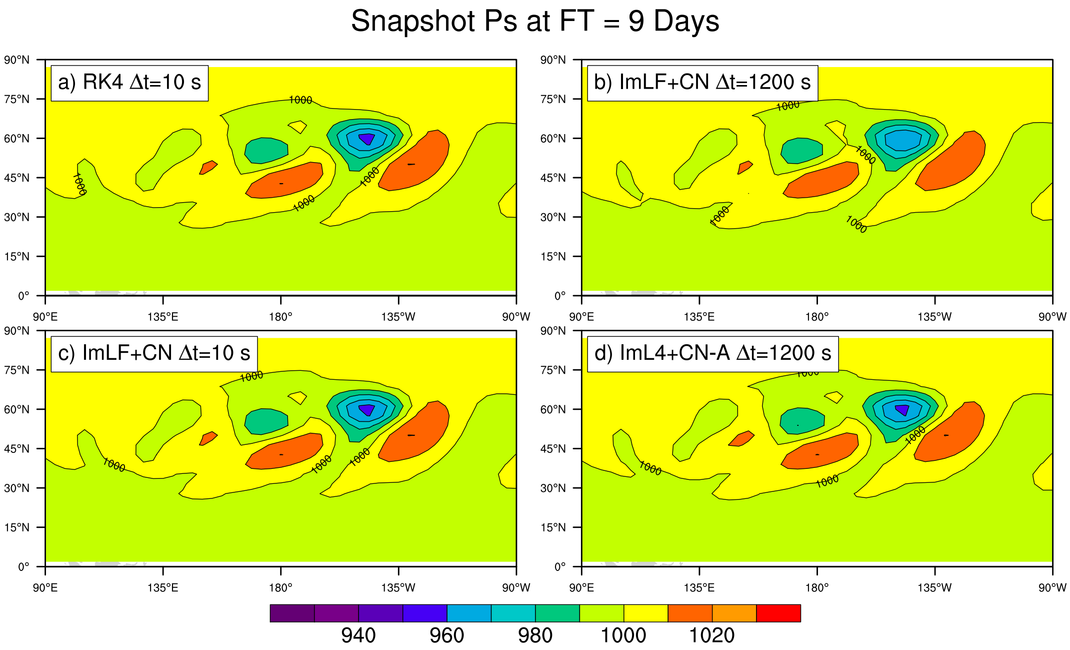


Figure A.4: Snapshots of surface pressure (in hPa) at the 9th day of integration. (a) Reference solution produced from RK4 scheme with time step $\Delta t = 10$ s, (b) SPEEDY's default ImLF+CN scheme with $\Delta t = 1200$ s, (c) as in (b), but with $\Delta t = 10$ s, and (d) ImL4+CN-A scheme with $\Delta t = 1200$ s. Contour intervals are 10 hPa.

A.5.2 Order estimation

As we discussed in Section A.3.3, theoretically, our Lorenz 4-cycle schemes combined with Crank-Nicolson scheme for the semi-implicit part (ImL4+CN-A, ImL4+CN-B, ImL4+CN-AB and ImL4+CN-ABBA schemes) should have second-order accuracy. Figure A.5 verifies this expectation. For 5-day forecasts (the left panel), if we focus on small values of time step ($\Delta t < 300$ s), all versions of ImL4+CN schemes collapse on a single line which has a slope of 2.0 (meaning that they are of second-order accuracy). The errors for these schemes are clearly better than that of ImLF+CN scheme whose slope is 1.0 (first-order accuracy). If we look at larger values of Δt, however, the curves begin to saturate as Δt becomes larger. Curiously, all the schemes with Backward-Euler scheme for the implicit part (ImLF+BE, ImL4+BE-A, ImL4+BE-B, ImL4+BE-AB and ImL4+BE-ABBA) exhibit slopes of 0.4 instead of the theoretical expectation 1.

The result becomes somewhat different if we look at 20-day forecasts (Figure A.5, right panel). All the schemes with Backward-Euler scheme for the implicit part now exhibit first-order accuracy, which is consistent with theoretical expectation. ImL4+CN-A and ImL4+CN-B schemes continue to be of second-order accuracy, with the latter being more accurate for $\Delta t > 300$ s. However, the alternating combinations of the two versions (ImL4+CN-AB and ImL4+CN-ABBA) exhibit larger errors for $\Delta t > 300$ s. In fact, ImL4+CN-ABBA scheme with $\Delta t = 1200$ s even blows up at the 30th days of integration. It is not clear why the alternation of versions A and B does not improve the accuracy of ImL4+CN schemes. Nonetheless, superiority of uncombined versions ImL4+CN-A or ImL4+CN-B over the traditional ImLF+CN is clear for all time steps. Interestingly, ImL4+CN-B is more accurate than ImL4+CN-A, although, for a linear system, the two schemes give identical predictions.

A.5.3 Explicit Lorenz 4-cycles

While the main focus of this study is on our new semi-implicit Lorenz N-cycle scheme, it would be interesting to see how the original explicit Lorenz N-cycle schemes perform when applied to the AGCM. As we mentioned in the penultimate paragraph of Section A.2.1, we show a result that supports the claim of Lorenz (1971) that explicit Lorenz 4-cycle can attain 4th-order accuracy by running it in the alternating A-B-B-A sequence.

Figure A.6 shows equivalent of Figure A.5 for explicit Lorenz 4-cycle schemes (ExL4-A, ExL4-B, ExL4-AB and ExL4-ABBA). Unlike Figure A.5, the left and right panels show, respectively, errors of 10-day and 30-day forecasts. At the 10th day, all versions exhibit 4th-order accuracy, while, at the 30th day, the versions A and B only have second-order accuracy. However, their alternating combinations AB and ABBA both retain 4th-order accuracy. These results are consistent with theoretical expectations that both versions A and B of Lorenz 4-cycle should be 4th-order accurate for a linear problem. In the Jablonowski-Williamson's baroclinic-wave dynamical core test, the integration starts from a zonally-uniform steady-state solution which is superposed with a weak and small localized perturbation. Thus, during the initial period of the integration, the system is only weakly nonlinear, until the breaking of the baroclinic wave occurs at $\sim$ day 10. By day 30, the system develops into a fully nonlinear regime, making versions A and B only second-order. As claimed by Lorenz (1971), the combination ABBA actually regains 4th-order accuracy. What is surprising is that, unlike what was suggested in Lorenz (1971), the simpler combination AB also regains 4th-order accuracy; in fact, for $\Delta t > 150$ s, it yields more accurate predictions than ABBA.

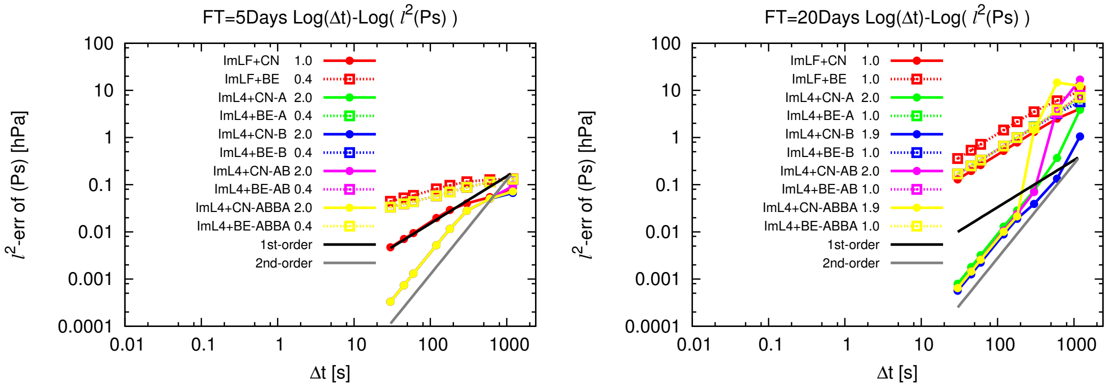


Figure A.5: Errors of surface pressure measured with the l^2-norm for various semi-implicit temporal integration schemes plotted against time steps Δt on a log-log plane. Shown are the errors for $\Delta t = 1200$, 600, 300, 180, 120, 90, 60, 45 and 30 s. The errors are computed with respect to the reference solution produced by running RK4 scheme with $\Delta t = 10$ s. The slopes of regression lines fitted using errors for 30 s $< \Delta t <$ 180 s are shown for each scheme in the legend. The names of the schemes are defined in Table 1 and Section A.4.4. The left and right panels show, respectively, the results for 5-day and 20-day forecasts.

A.6 Summary and discussion

Lorenz N-cycle scheme for numerical integration of ODEs proposed by Lorenz (1971) has remained less widely used in the atmospheric and oceanic sciences despite its major advantages, notably its economical use of memory, higher-order accuracy and the ease of implementation. Part of the reasons is perhaps lack of its semi-implicit formulation. Recently, Whitaker and Kar (2013) proposed a semi-implicit scheme based on Lorenz 3-cycle and reported promising results with both an idealized shallow-water system and an operational numerical weather prediction (NWP) model. Their focus, however, was on improving stability of their previously proposed scheme and thus the economical memory use of Lorenz N-cycle was lost in their formulation. In this study, we presented a new semi-implicit formulation of Lorenz N-cycle which can preserve the memory efficiency.

The accuracy and stability analysis conducted for a linear, univariate split-frequency oscillation equation shows that the accuracy of our new schemes all can be made second-order if we adopt Crank-Nicolson scheme for the implicit part. For a purely oscillatory equation with no damping, our semi-implicit Lorenz N-cycle combined with Crank-Nicolson scheme for the implicit part is found to be unstable for $N = 1$ and $N = 2$. The stability is improved by increasing N to 3 and 4, with the latter having larger stable region, but it becomes less stable again if N is further increased. A linear stability analysis which allows dissipation shows that the small instability regions found in the semi-implicit 4-cycle scheme disappears if a small damping is present in the system.

Numerical experiments performed using the dynamical core of the SPEEDY AGCM under the framework of the baroclinic wave test case of Jablonowski and Williamson (2006) confirmed that our semi-implicit Lorenz 4-cycle scheme combined with Crank-Nicolson scheme for the implicit part, both version A (ImL4+CN-A) and version B (ImL4+CN-B), exhibit second-order accuracy and are more accurate than the traditional semi-implicit leapfrog scheme (ImLF+CN) for any time step Δt. Intriguingly, however, contrary to our expectation that running the two versions in alternating sequence should improve the scheme because their truncation errors tend to cancel each other, the alternating combinations of the two (ImL4+CN-AB and ImL4+CN-ABBA) actually proved to be less accurate than the non-alternating versions. ImL4+CN-ABBA even proved to be unstable for $\Delta t = 1200$ s although other versions were stable for this Δt.

We have also confirmed that, for explicit Lorenz 4-cycle, running the two versions (ExL4-A and ExL4-B) in alternating sequences (ExL4-AB and ExL4-ABBA) in fact improves the accuracy to 4th order. In view of the critical comment posed by Purser (2001) that "the efficacy of this strategy [alternation of the two versions] in a highly nonlinear numerical weather prediction model is extremely doubtful," the fact that the alternation strategy suggested by Lorenz (1971) in fact worked for the primitive equation system is itself a surprising result. Unlike what is claimed by Lorenz (1971), however, for this particular problem, a simpler combination ExL4-AB turned out to be more accurate than the suggested ExL4-ABBA.

It remains unclear why, for our semi-implicit scheme, alternation of the two versions did not lead to improved accuracy. One possible reason is that, unlike in the original explicit scheme, truncation errors of the two versions that arise from nonlinearity of the explicit tendency ($F^E(x)$ of Eq.(A.12)) do not cancel each other.

In summary, the semi-implicit Lorenz 4-cycle schemes we propose in this work are computationally as efficient as the traditional Robert-Asselin-filtered semi-implicit leapfrog scheme, in terms of both the amount of computation and memory usage. Moreover, our scheme have second-order accuracy, which is higher than that of the traditional scheme. Another advantage of our scheme over the traditional leapfrog is that it is a two-time level scheme, meaning that it requires only the model state at a single time to initialize the integration. Being a two-time level scheme also means that it is free from troublesome computational modes.

Given our success with the primitive-equations model, it is tempting to hope that we might be able to improve operational weather forecasts by adopting our scheme to operational NWP systems. We conclude this work by suggesting some future work in this direction. Most advanced operational NWP centers, including European Centre for Medium-Range Weather Forecasts (ECMWF), Environment Canada, Japan Meteorological Agency (JMA) and UK MetOffice, all adopt semi-implicit semi-Lagrangian temporal integration schemes in their global models. National Centers for Environmental Prediction (NCEP) is also developing semi-Lagrangian version of their global model. Our next challenge is therefore to formulate a semi-implicit semi-Lagrangian version of Lorenz N-cycle.

The trend in regional NWP is to use non-hydrostatic models which include acoustic waves in their solutions. In fact, for example, JMA, NCEP, UK MetOffice and the German weather service (DWD) are already using non-hydrostatic regional models in their operation. In such models, the very fast acoustic waves are typically accommodated by using split-explicit techniques. Implementing Lorenz N-cycle within a split-explicit scheme should be straightforward. Thus, application of Lorenz N-cycle in regional non-hydrostatic models would also be an attractive research topic.

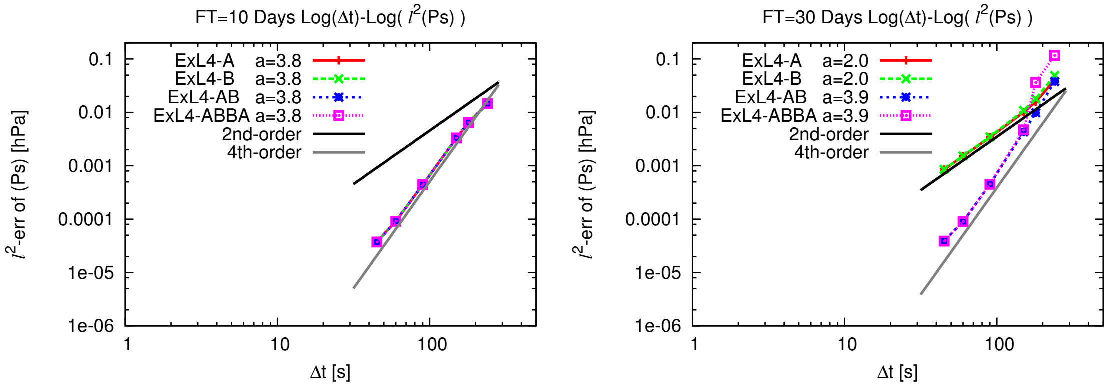


Figure A.6: As in Figure A.5, but for explicit Lorenz 4-cycle schemes. Shown are the errors for $\Delta t = 240, 180, 150, 90, 60,$ and 30 s. The slopes of regression lines fitted using errors for 30 s $< \Delta t <$ 120 s are shown for each scheme in the legend. The left and right panels show, respectively, the results for 10-day and 30-day forecasts.

Bibliography

Alpert, J. C., D. L. Carlis, B. A. Ballish, and V. K. Kumar, 2009: Using pseudo RAOB observations to study GFS skill score dropouts. *23rd Conference on Weather Analysis and Forecasting/19th Conference on Numerical Weather Prediction*, Omaha, NE, American Meteorological Society, Extended Abstract.

Amezucua, J., E. Kalnay, and P. D. Williams, 2011: The effects of the RAW filter on the climatology and forecast skill of the SPEEDY model. *Mon. Wea. Rev*, **139**, 608–619.

Anderson, E. and H. Järvinen, 1999: Variational quality control. *Q. J. R. Meterol. Soc*, **125 (554)**, 697–722.

Anderson, J. L., 2001: An ensemble adjustment Kalman filter for data assimilation. *Mon. Wea. Rev*, **129 (12)**, 2884–2903.

Anderson, J. L., 2007: Exploring the need for localization in ensemble data assimilation using a hierarchical ensemble filter. *Physica D: Nonlinear Phenomena*, **230 (1)**, 99–111.

Asselin, R., 1972: Frequency filter for time integrations. *Mon. Wea. Rev*, **100**, 487–490.

Balgovind, R., A. Dalcher, M. Ghil, and E. Kalnay, 1983: A stochastic-dynamic model for the spatial structure of forecast error statistics. *Mon. Wea. Rev*, **111 (4)**, 701–722.

Bergthorsson, P. and B. Döös, 1955: Numerical weather map analysis. *Tellus*, **7**, 329–340.

Bjerknes, V., 1904: Das Problem der Wettervorhersage, betrachtet vom Standpunkte der Mechanik und der Physik (The problem of weather prediction, considered from the viewpoints of mechanics and physics). *Meteorologische Zeitschrift*, **21**, 1–7, URL `http://www.schweizerbart.de/resources/downloads/paper_free/74383.pdf`, (originally in German; translated and edited by VOLKEN E. and S. BRÖNNIMANN – *Meteorologische Zeitschrift* **18** (2009), 663–667.

Boer, G. and B. Denis, 1997: Numerical convergence of the dynamics of a GCM. *Clim. Dyn.*, **13**, 359–374.

Buehner, M., 2005: Ensemble-derived stationary and flow-dependent background-error covariances: Evaluation in a quasi-operational NWP setting. *Q. J. R. Meterol. Soc*, **131 (607)**, 1013–1043.

Cardinali, C., 2009: Monitoring the observation impact on the short-range forecast. *Q. J. R. Meterol. Soc*, **135 (638)**, 239–250.

Charney, J. G., R. Fjørtoft, and J. von Neuman, 1950: Numerical integration of the barotropic vorticity equation. *Tellus*, **2**, 237–254.

Clayton, A., A. Lorenc, and D. Barker, 2013: Operational implementation of a hybrid ensemble/4D-Var global data assimilation system at the Met Office. *Q. J. R. Meterol. Soc*, **139 (675)**, 1445–1461.

Da Silva, A., J. Pfaendtner, J. Guo, M. Sienkiewicz, and S. Cohn, 1995: Assessing the effects of data selection with DAO's Physical-space Statistical Analysis System. *International Symposium on Assimilation of Observations*, Tokyo, Japan, WMO, 133–17.

Daescu, D. N., 2008: On the sensitivity equations of four-dimensional variational (4D-Var) data assimilation. *Monthly Weather Review*, **136 (8)**.

Daescu, D. N. and R. H. Langland, 2013: Error covariance sensitivity and impact estimation with adjoint 4D-Var: theoretical aspects and first applications to NAVDAS-AR. *Q. J. R. Meterol. Soc*, **139**, 226–241.

Daescu, D. N. and R. Todling, 2009: Adjoint estimation of the variation in model functional output due to the assimilation of data. *Mon. Wea. Rev*, **137 (5)**.

Dalcher, A. and E. Kalnay, 1987: Error growth and predictability in operational ECMWF forecasts. *Tellus A*, **39 (5)**, 474–491.

Desroziers, G. and S. Ivanov, 2001: Diagnosis and adaptive tuning of observation-error parameters in a variational assimilation. *Q. J. R. Meterol. Soc*, **127 (574)**, 1433–1452.

Durran, D. R., 1991: The third-order Adams-Bashforth method: An attractive alternative to leapfrog time differencing. *Mon. Wea. Rev*, **119**, 702–720.

Ehrendorfer, M., 2007: A review of issues in ensemble-based Kalman filtering. *Meteorologische Zeitschrift*, **16 (6)**, 795–818.

Ehrendorfer, M., R. M. Errico, and K. D. Raeder, 1999: Singular-vector perturbation growth in a primitive equation model with moist physics. *J. Atmos. Sci*, **56 (11)**, 1627–1648.

Gaspari, G. and S. E. Cohn, 1999: Construction of correlation functions in two and three dimensions. *Q. J. R. Meterol. Soc*, **125 (554)**, 723–757.

Gasperoni, N. and X. Wang, 2013: Improving an ensemble-based observation impact estimate using a group filter technique, WMO Sixth symposium on Data Assimilation, College Park, MD, USA.

Gelaro, R., R. H. Langland, S. Pellerin, and R. Todling, 2010: The THORPEX Observation Impact Intercomparison Experiment. *Mon. Wea. Rev*, **138 (11)**.

Gelaro, R. and Y. Zhu, 2009: Examination of observation impacts derived from observing system experiments (OSEs) and adjoint models. *Tellus A*, **61 (2)**, 179–193.

Gent, P. and M. Cane, 1989: A reduced gravity, primitive equation model of the upper equatorial ocean. *J. Comput. Phys.*, **81**, 444–480.

Gill, A., 1980: Some simple solutions for heat-induced tropical circulation. *Q. J. R. Meterol. Soc*, **106 (449)**, 447–462.

Gill, A. E., 1982: *Atmosphere-ocean dynamics*, Vol. 30. Academic press, 662 pp.

Held, I. M. and M. J. Suarez, 1994: A proposal for the intercomparison of the dynamical cores of atmospheric general circulation models. *Bull. Amer. Meteor. Soc*, **75**, 1825–1830.

Holdaway, D., R. Errico, R. Gelaro, and J. G. Kim, 2014: Inclusion of linearized moist physics in NASA's Goddard Earth Observing System data assimilation tools. *Mon. Wea. Rev*, **142 (1)**.

Hollingsworth, A. and P. Lönnberg, 1986: The statistical structure of short-range forecast errors as determined from radiosonde data. Part I: The wind field. *Tellus A*, **38 (2)**, 111–136.

Hotta, D. and Y. Ota, 2011: Learning Data Assmilation with the Lorenz '96 model. *Additional Volume to Report of Numerical Prediction Division (Suuchi Yohouka Houkoku Bessatsu)*, **57**, 144–158, (in Japanese).

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

Ishibashi, T., 2010: Optimization of error covariance matrices and estimation of observation data impact in the JMA global 4D-Var system. *CAS/JSC WGNE Research Activities in Atmospheric and Oceanic Modelling*, **40**, 1–11.

Israeli, M. and D. Gottlieb, 1974: On the stability of the N cycle scheme of Lorenz. *Mon. Wea. Rev*, **102**, 254–265.

Jablonowski, C. and D. L. Williamson, 2006: A baroclinic instability test case for atmospheric model dynamical cores. *Q. J. R. Meterol. Soc*, **132**, 2943–2975.

Jazwinski, A. H., 1970: *Stochastic Processes and Filtering Theory.* Academic Press, 376 pp.

JMA, 2013: *Outline of the operational numerical weather prediction of the Japan Meteorological Agency.* Appendix to WMO Technical Progress Report on the Global Data-processing and Forecasting System (GDPFS) and Numerical Weather Prediction (NWP) Research, Japan Meteorological Agency, 188 pp., URL `http://www.jma.go.jp/jma/jma-eng/jma-center/nwp/outline2013-nwp/index.htm`.

Kalnay, E., 2003: *Atmospheric modeling, Data Assimilation and Predictability.* Cambridge University Press, 341 pp.

Kalnay, E., Y. Ota, T. Miyoshi, and J. Liu, 2012: A simpler formulation of forecast sensitivity to observations: application to ensemble Kalman filters. *Tellus A*, **64**, 18 462.

Kar, S. K., 2006: A semi-implicit Runge-Kutta time-difference scheme for the two-dimensional shallow-water equations. *Mon. Wea. Rev*, **134**, 2916–2926.

Kleist, D. T., 2012: An evaluation of hybrid variational-ensemble data assimilation for the NCEP GFS. Ph.D. dissertation, University of Maryland.

Kleist, D. T., D. F. Parrish, J. C. Derber, R. Treadon, R. M. Errico, and R. Yang, 2009a: Improving incremental balance in the GSI 3DVAR analysis system. *Mon. Wea. Rev*, **137 (3)**.

Kleist, D. T., D. F. Parrish, J. C. Derber, R. Treadon, W.-S. Wu, and S. Lord, 2009b: Introduction of the GSI into the NCEP Global Data Assimilation System. *Weather & Forecasting*, **24 (6)**.

Kumar, K., J. C. Alpert, D. L. Carlis, and B. A. Ballish, 2009: Investigation of NCEP GFS Model forecast skill "dropout" characteristics using the EBI Index. *23rd Conference on Weather Analysis and Forecasting/19th Conference on Numerical Weather Prediction*, Omaha, NE, American Meteorological Society, Extended Abstract.

Langland, R. H. and N. L. Baker, 2004: Estimation of observation impact using the NRL atmospheric variational data assimilation adjoint system. *Tellus*, **56A**, 189–201.

Li, H., E. Kalnay, and T. Miyoshi, 2009: Simultaneous estimation of covariance inflation and observation errors within an ensemble Kalman filter. *Q. J. R. Meterol. Soc*, **135**, 523–533.

Li, H., J. Liu, and E. Kalnay, 2010: Correction of 'Estimating observation impact without adjoint model in an ensemble Kalman filter'. *Q. J. R. Meterol. Soc*, **136 (651)**, 1652–1654.

Lien, G.-Y., 2014: Ensemble assimilation of global large-scale precipitation. Ph.D. dissertation, University of Maryland.

Liu, J. and E. Kalnay, 2008: Estimating observation impact without adjoint model in an ensemble Kalman filter. *Q. J. R. Meterol. Soc*, **134**, 1327–1335.

Liu, J., E. Kalnay, T. Miyoshi, and C. Cardinali, 2009: Analysis sensitivity calculation in an ensemble Kalman filter. *Q. J. R. Meterol. Soc*, **135 (644)**, 1842–1851.

Liu, Z.-Q. and F. Rabier, 2002: The interaction between model resolution, observation resolution and observation density in data assimilation: A one-dimensional study. *Q. J. R. Meterol. Soc*, **128 (582)**, 1367–1386.

Lorenc, A., 1981: A global three-dimensional multivariate statistical interpolation scheme. *Mon. Wea. Rev*, **109 (4)**, 701–721.

Lorenc, A. C., 2003: The potential of the ensemble Kalman filter for NWP — a comparison with 4D-Var. *Q. J. R. Meterol. Soc*, **129 (595)**, 3183–3203.

Lorenc, A. C. and R. T. Marriott, 2013: Forecast sensitivity to observations in the Met Office Global numerical weather prediction system. *Q. J. R. Meterol. Soc*, **140**, 209–223.

Lorenz, E. N., 1971: An *N*-cycle time-differencing scheme for step-wise numerical integration. *Mon. Wea. Rev*, **119**, 1612–1623.

Lorenz, E. N., 1996: Predictability: A problem partly solved. *Proc. Seminar on predictability*, Reading, UK., ECMWF, Vol. 1, 1–18.

Lorenz, E. N. and K. A. Emanuel, 1998: Optimal sites for supplementary weather observations: Simulation with a small model. *J. Atmos. Sci*, **55 (3)**, 399–414.

Matsumoto, M. and T. Nishimura, 1998: Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation*, **8 (1)**, 3–30.

Matsuno, T., 1966: Quasi-geostrophic motions in the equatorial area. *J. Meterol. Soc. Japan*, **44 (1)**, 25–43.

Miyoshi, T., 2005: Ensemble Kalman filger experiments with a primitive-equation global model. Ph.D. dissertation, University of Maryland.

Miyoshi, T., 2006: Ensemble Kalman Filter – Fusing Ensemble Forecasting with Data Assimilation –. *Additional Volume to Report of Numerical Prediction Division (Suuchi Yohouka Houkoku Bessatsu)*, **52**, 88–99, (in Japanese).

Molteni, F., 2003: Atmospheric simulations using a GCM with simplified physical parametrizations. I: Model climatology and variability in multi-decadal experiment. *Climate Dyn.*, **20**, 175–191.

NCEP, 2013: Table 2. Code table for PREPBUFR report types used by Global GFS and GDAS GSI analyses. http://www.emc.ncep.noaa.gov/mmb/data_processing/prepbufr.doc/table_2.htm.

Onogi, K., 1998: A data quality control method using forecasted horizontal gradient and tendency in a NWP system: dynamic QC. *J. Meterol. Soc. Japan*, **76 (4)**, 497–516.

Ota, Y., J. C. Derber, T. Miyoshi, and E. Kalnay, 2013: Ensemble-based observation impact estimates using the NCEP GFS. *Tellus A*, **65**, 20 038.

Parrish, D. F. and J. C. Derber, 1992: The National Meteorological Center's spectral statistical-interpolation analysis system. *Mon. Wea. Rev*, **120 (8)**, 1747–1763.

Penny, S. G., 2014: The hybrid local ensemble transform Kalman filter. *Mon. Wea. Rev*, **142**, 2139–2149.

Persson, A., 2005a: Early Operational Numerical Weather Prediction outside the USA: an historical introduction. Part I: Internationalism and engineering, NWP in Sweden, 195269. *Meteorol. Appl.*, **12**, 135–159.

Persson, A., 2005b: Early Operational Numerical Weather Prediction outside the USA: an historical introduction. Part II: Twenty countries around the world. *Meteorol. Appl.*, **12**, 269–289.

Persson, A., 2005c: Early Operational Numerical Weather Prediction outside the USA: an historical introduction. Part III: Endurance and mathematics – British NWP, 19481965. *Meteorol. Appl.*, **12**, 381–413.

Purser, R. J., 2001: Proposed semi-implicit adaptations of two low-storage Runge-Kutta schemes. Part I: Theoretical formulation and stability analysis. *NCEP Office Note*, **435**, 40pp.

Purser, R. J., 2007: Accuracy considerations of time-splitting methods for models using two-time-level schemes. *Mon. Wea. Rev*, **135**, 1158–1164.

Purser, R. J. and L. M. Leslie, 1997: High-order generalized Lorenz N-cycle schemes for semi-lagrangian models employing second derivatives in time. *Mon. Wea. Rev*, **125**, 1261–1276.

Robert, A. J., 1966: The integration of a low order spectral form the primitive meteorological equations. *J. Meterol. Soc. Japan*, **44**, 237–245.

Robert, A. J., 1969: The integration of a spectral model of the atmosphere by the implicit method. *Proc. WMO-IUGG Symp. on Numerical Weather Prediction*, Tokyo, Japan, Japan Meteorological Agency, Vol. VII, 19–24.

Rodwell, M. J. and Coauthors, 2013: Characteristics of occasional poor medium-range weather forecasts for Europe. *Bull. Amer. Meteor. Soc*, **94**, 1393–1405.

Roh, S., M. G. Genton, M. Jun, I. Szunyogh, and I. Hoteit, 2013: Observation quality control with a robust ensemble Kalman filter. *Mon. Wea. Rev*, **141 (12)**.

Roulstone, I. and J. Norbury, 2013: *Invisible in the Storm*. Princeton University Press, 346 pp.

Sela, J. G., 1980: Spectral modeling at the National Meteorological Center. *Mon. Wea. Rev*, **108 (9)**, 1279–1292.

Simmons, A., 2011: From observations to service delivery: challenges and opportunities. *WMO Bulletin*, **60 (2)**, 96–107.

Sommer, M. and M. Weissmann, 2014: Observation impact in a convective-scale localized ensemble transform Kalman filter. *Q. J. R. Meterol. Soc*, doi: 10.1002/qj.2343.

Talagrand, O., 1999: A posteriori evaluation and verification of the analysis and assimilation algorithms. *Proceedings of Workshop on Diagnosis of Data Assimilation Systems*, Reading, UK., ECMWF, 17–28.

Tavolato, C. and L. Isaksen, 2010: Huber norm quality control in the IFS. *ECMWF Newsletter*, **122**, 27–31.

Todling, R., 2013: Comparing two approaches for assessing observation impact. *Mon. Wea. Rev*, **141**, 1484–1505.

Trémolet, Y., 2004: Diagnostics of linear and incremental approximations in 4D-Var. *Q. J. R. Meterol. Soc*, **130 (601)**, 2233–2251.

Trevisan, A., M. D'Isidoro, and O. Talagrand, 2010: Four-dimensional variational assimilation in the unstable subspace and the optimal subspace dimension. *Q. J. R. Meterol. Soc*, **136 (647)**, 487–496.

Wang, X., D. Parrish, D. Kleist, and J. Whitaker, 2013: GSI 3DVar-based ensemble-variational hybrid data assimilation for NCEP Global Forecast System: Single-resolution experiments. *Mon. Wea. Rev*, **141 (11)**.

Whitaker, J. S. and T. M. Hamill, 2002: Ensemble data assimila- tion without perturbed observations. *Mon. Wea. Rev*, **130**, 1913–1924.

Whitaker, J. S. and T. M. Hamill, 2012: Evaluating methods to account for system errors in ensemble data assimilation. *Monthly Weather Review*, **140 (9)**.

Whitaker, J. S. and S. K. Kar, 2013: Implicit-explicit Runge-Kutta methods for fast-slow wave problems. *Mon. Wea. Rev*, **141**, 426–3434.

Williams, P. D., 2009: A proposed modification to the Robert-Asselin time filter. *Mon. Wea. Rev*, **137**, 2538–2546.

Williams, P. D., 2011: The RAW filter: An improvement to the Robert-Asselin filter in semi-implicit integrations. *Mon. Wea. Rev*, **139**, 1996–2007.

Williams, P. D., 2013: Achieving seventh-order amplitude accuracy in leapfrog integrations. *Mon. Wea. Rev*, **141**, 3037–3051.

Wu, W.-S., R. J. Purser, and D. F. Parrish, 2002: Three-dimensional variational analysis with spatially inhomogeneous covariances. *Mon. Wea. Rev*, **130**, 2905–2916.

Yoon, Y.-N., E. Ott, and S. Istvan, 2010: On the propagation of information and the use of localization in ensemble Kalman filtering. *J. Atmos. Sci*, **67 (12)**, 3823–3834.

Zhang, F., C. Snyder, and J. Sun, 2004: Impacts of initial estimate and observation availability on convective-scale data assimilation with an ensemble Kalman filter. *Mon. Wea. Rev*, **132 (5)**, 1238–1253.