ABSTRACT

Title of dissertation: MULTIVARIATE CORRELATIONS: BALANCE OPERATORS AND VARIABLE LOCALIZATION IN ENSEMBLE DATA ASSIMILATION

Catherine Anne Thomas,
Doctor of Philosophy, 2017

Dissertation directed by: Professor Kayo Ide
Department of Atmospheric and
Oceanic Science

Localization is performed in ensemble data assimilation schemes to eliminate correlations that are contaminated by sampling error. This method is frequently necessary within numerical weather prediction (NWP) applications due to the computational constraints present, limiting the number of ensemble members to a size much smaller than the dimension of the system. The most common form of localization occurs in the spatial dimensions, reducing the correlations for points that are distant and likely dominated by sampling error. Spatial localization can introduce imbalance in the system due to the disruption of physical relationships that are dictated by gradients or column integrated quantities, which produce fast-moving gravity waves within NWP models and degrade the forecast.

The first part of this dissertation explores the impact of including a balance operator within ensemble data assimilation schemes and how the type of spatial localization interacts with it. The inclusion of a balance operator allows the localization to be performed on the unbalanced portion of the correlation, preserving the balanced correlation. Two data assimilation schemes are explored: a hybrid 4D ensemble-variational (4DEnVar) scheme and a Local Ensemble Transform Kalman Filter (LETKF). Observing system simulation experiments are performed using an intermediate complexity model, SPEEDY. It is shown that localizing on the background error as in the hybrid 4DEnVar is more effective than localizing on the observation error as in the LETKF. Within the LETKF, the balance operator can only propagate information one way, for example, from streamfunction to temperature, but not vice versa as in the hybrid 4DEnVar.

Many applications contain variables that are physically unrelated and should not be correlated, but contain nonzero correlations. The second part of this dissertation presents two forms of variable localization in a unified framework: observation space variable localization (VO) and model space variable localization (VM). VO restricts the impact that observations have to certain model variables. VM removes the cross-correlations during the computation of the background error covariance. VM is more computationally expensive, but it has the added advantages of not requiring knowledge of observation types and allowing a single observation to impact multiple model variables whose cross-correlations have been removed.

MULTIVARIATE CORRELATIONS: BALANCE OPERATORS AND VARIABLE LOCALIZATION IN ENSEMBLE DATA ASSIMILATION

by

Catherine Anne Thomas

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
2017

Advisory Committee:
Professor Kayo Ide, Chair/Advisor
Professor Jim Carton
Professor Brian Hunt
Professor Eugenia Kalnay
Dr. Daryl Kleist

Acknowledgments

I sincerely thank my advisor, Kayo Ide, for her unwavering guidance and support, both financially and academically. Her attention to detail and instilling of good habits in me has helped me immensely, not only in producing this thesis, but in my transition to my professional career. I am indebted to Eugenia Kalnay for bringing me into the program and supporting me for my first few years. I am also grateful for the mentorship of Daryl Kleist, both at the University of Maryland and at the Environmental Modeling Center. Along with Brian Hunt and Jim Carton, I am thankful for my committee for their time and valuable feedback for this thesis.

The experiments performed in this thesis would not have been possible without Takemasa Miyoshi, whose code from his thesis experiments provided a launching point for this work. Fred Kucharski also gratefully provided code and support for a recent version of the SPEEDY model.

I am extremely grateful for the support of I.M. Systems Group and the Environmental Modeling Center for allowing me to continue my studies while working for them. I am especially thankful to my supervisors (Andrew Collard, John Derber, and Mike Pecnick) for their patience and support while I completed my degree.

I have been blessed with an unbelievable support network of family and friends. My parents provided me with an environment that fostered a love for science and math as well as a role model for women in STEM. My husband's unyielding patience and support along with my friends' enthusiastic encouragement allowed me to continue to push though. Last but not least, Pearl was a steadfast presence, always at

my side through many late nights, my watchful guardian.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Background and Motivation

Data assimilation refers to the class of methods that initializes the model forecast by optimally combining information on the state from observations, previous forecasts, and dynamical balances. This optimization requires the characterization of various errors that exist within the system, including errors due to representativeness, observation instrumentation, and the forecast model, among others. The relative weights of these errors determine the influence of each of our information sources. The background error covariance matrix describes the errors associated with the a priori information, usually a previous short term forecast. It provides spatial correlations between every grid point as well as the correlations between the different variable types. The spatial correlations determine the structure of the observation impact and its weight relative to the observation error determines the magnitude.

With the advancement of scientific computing, ensemble methods (Evensen, 1994; Houtekamer and Mitchell, 1998; Whitaker and Hamill, 2002) emerged to compensate for some of the disadvantages of the original Kalman filter (Kalman, 1960), namely that the background error was impractical to calculate in modern numerical

weather prediction (NWP). Ensemble methods utilize a Monte Carlo approach to approximate the background error by sampling a large ensemble of forecasts, allowing the background covariance to vary in time and contain spatial correlations that are anisotropic. However, computational constraints do not allow for a full sampling of the NWP system since the dimension is so large. This under-sampling can create unphysical correlations in the estimated background error. Various methods, such as localization and inflation, have been devised to handle spurious correlations as well as typically under-spread ensembles, removing correlations that are likely unphysical and stabilizing model integration (Anderson and Anderson, 1999; Houtekamer and Mitchell, 2001).

Most frequently used in the spatial dimensions, localization eliminates correlations by modifying either the background error covariance or the observation error covariance. To apply spatial localization in model space, the background error is multiplied by a correlation function that decreases with distance, reducing the correlations for points that are distant and retaining the correlations for points that are nearby (Houtekamer and Mitchell, 2001). This form of localization is also applied within ensemble-variational schemes, such as the hybrid 4DEnVar (Kleist and Ide, 2015b). To apply spatial localization in observation space, the observation error is multiplied by a correlation function that increases with distance, increasing the observation error and therefore reducing the impact for observations that are distant (Hunt et al., 2007). Alternatively, localization in observation space can be implemented as a form of observation selection, where only observations within a certain region are considered in the calculation of the analysis (Houtekamer and

Mitchell, 1998). Both forms of observation space localization are applied within the local ensemble transform Kalman filter (LETKF, Hunt et al., 2007). The reduction of the background error and the increase of the observation error have an equivalent impact on observation influence through the Kalman gain.

By reducing all correlations to zero beyond a certain distance, large scale balances that are dictated by gradients or column integrated quantities, such as geostrophic balance, are disrupted (Cohn et al., 1998; Lorenc, 2003; Mitchell et al., 2002). Even though localization removes spurious correlations, it introduces imbalance to the system. Imbalances within NWP initial conditions result in the production of fast moving gravity waves that degrade the forecast. As computational resources continue to grow, the model resolution also continues to increase. The number of ensemble members will likely not be sufficient to fully estimate the background error in the foreseeable future. Therefore, localization will continue to be necessary within ensemble data assimilation schemes, despite the imbalances it causes. The particular implementation of the localization can have an impact on the degree that the localization affects the balance.

While localization is most commonly applied spatially, it can be applied to other portions of the ensemble correlations as well. Variable localization, or the removal of correlations between different variable types, has been implemented in a number of applications (Clayton et al., 2013; Kang et al., 2011), though the formulation is not consistent. With an increased interest in strongly coupled data assimilation (Han et al., 2013; Liu et al., 2013; Sluka et al., 2016) and the assimilation of chemistry components (Coman et al., 2012; Liu et al., 2012; Pagowski and

Grell, 2012; Schwartz et al., 2014), the need for variable localization will likely increase. A common formulation comparing the different types of variable localization would be advantageous when constructing a new system as well as an investigation into their strengths and weaknesses.

1.2 Thesis Objectives

The objectives of this thesis relate to the construction and balance of the background error covariance matrix. The aims of this thesis are to:

1. Apply a balance operator to the ensemble portion of a deterministic hybrid 4DEnVar scheme within the SPEEDY model.

2. Apply a balance operator to the ensemble mean and spread in an LETKF within the SPEEDY model.

3. Examine the effect of the balance operator on the analysis and forecast skill as well as the balance of the analysis within two types of spatial localization.

4. Present a unified framework for two forms of variable localization within three ensemble data assimilation schemes and identify their strengths and weaknesses.

By applying a balance operator within ensemble data assimilation schemes, the localization can act on the unbalanced part of the ensemble correlations, preserving the balanced correlations and reducing imbalance. While this method has been applied within ensemble schemes previously (Clayton et al., 2013), the form of spatial

localization was not considered. Whether the application of the spatial location is on the background error or observation error impacts the effectiveness of the balance operator. As previously stated, the method of variable localization is also not novel. However, discussions regarding the classification of different variable localization methods as well as their relative strengths and weaknesses have not been presented.

The first three objectives will be addressed by performing observing system simulation experiments (OSSEs). These experiments will take place within an intermediate complexity global atmospheric model, SPEEDY (Molteni, 2003). The SPEEDY model will be described in Chapter 2, along with validation of its climatology and variability. Model biases between different horizontal resolutions will be diagnosed and the observing network and data assimilation configuration will also be described in Chapter 2. Chapter 3 describes the application of a balance operator within the ensemble portion of a hybrid 4DEnVar (Objective 1), which utilizes spatial localization on the background error. Single observation impact tests as well as OSSEs with a full observing network will be evaluated. Chapter 4 examines how the balance operator implementation differs within an LETKF (Objective 2), which utilizes spatial localization upon the observation error. Similar impact tests and OSSEs are performed, demonstrating that the balance operator is unable to fully propagate balanced information within this form of spatial localization (Objective 3). Variable localization is explored in Chapter 5, where two forms of variable localization are identified: model space variable localization and observation space variable localization (Objective 4). These two forms are formulated within three data assimilation schemes: the ensemble square root filter (EnSRF, Whitaker and

Hamill, 2002), LETKF, and EnVar. The strengths and weaknesses of the two forms of variable localization are shown to be consistent across ensemble schemes.

Through these objectives, it will be shown that the construction of the background covariance is critical to model performance. Localization, either spatial or variable, can be applied in model space or observation space. While the background error and observation error are related through the Kalman gain, there are additional consequences of choosing to localize in either model space or observation space that need to be considered for the particular application.

Chapter 2: System Configuration

Chapters 3 and 4 contain experiments to determine the impact of a balance operator within two ensemble data assimilation schemes using an intermediate complexity model. This chapter begins with a description of the model's formulation (Section 2.1.1), then presents verification for its mean fields and variability (Section 2.1.2), and diagnoses model biases between different horizontal resolutions (Section 2.1.3). Section 2.2 contains a description of the observation network and the chapter concludes with the data assimilation configuration used in the experiments (Section 2.3).

2.1 SPEEDY Model

2.1.1 Model Description

The SPEEDY model (Simplified Parameterizations, primitivE-Equation DYnamics) is a global atmospheric general circulation model (AGCM) of intermediate complexity (Molteni, 2003). It was created for use in climate studies, being global in scope yet significantly faster computationally than state-of-the-art NWP models, approximately an order of magnitude for the same spatial resolution. These savings

come primarily from its simplified parameterization schemes. SPEEDY contains the same parameterization components as most complex AGCMs (convection, vertical diffusion, cloud cover/thickness, condensation, long- and short-wave radiation, and momentum and energy surface fluxes) though their formulation is more basic, making assumptions specifically for a model with very coarse vertical resolution. Model integration uses a leap-frog time scheme, where the spurious computational mode is damped through the Robert-Asselin-Williams (RAW) filter (Amezcua et al., 2011). SPEEDY was first adapted for data assimilation experiment within NWP by Miyoshi (2005) and has since been used to test many new data assimilation methodologies (Greybush et al., 2011; Harlim and Hunt, 2007; Kang et al., 2011, 2012; Li et al., 2009; Miyoshi, 2011; Sluka et al., 2016; Zhou, 2014) in a framework that is similar in structure to the models used in the NWP centers, but uses much less computational resources and contains less uncertainty.

SPEEDY contains a spectral dynamic core with a coarse vertical resolution of eight σ levels (σ =0.95, 0.835, 0.685, 0.51, 0.34, 0.2, 0.095, 0.02), where the bottom level represents the planetary boundary layer and the top two levels represent the stratosphere. Several horizontal resolution options are available (Kucharski, 2012). The following experiments utilize horizontal resolutions of T30 and T63, corresponding to a standard Gaussian grid of 96 by 48 grid points (approximately 3.75° at the equator) and 192 by 96 grid points (approximately 1.875° at the equator) respectively.

Climatological boundary conditions for both model resolutions are the mean from the ERA-Interim (Dee et al., 2011) for the years 1979-2008. Annual mean

values are used for the bare-surface albedo and the vegetation fractional coverage. Monthly mean values are stored for sea-surface temperature (SST), sea ice fraction, top soil layer temperature, top soil layer moisture, and snow depth, which are interpolated to compute daily values. For the SST field, an anomaly is added to the climatological value in order to have time varying SST to represent interannual variability. SST anomalies are provided by the NOAA_ERSST_V3 (Smith et al., 2008; Xue et al., 2003) and are available from 1854 to 2010. The incoming solar radiation at the top of the atmosphere is a daily mean value, resulting in no diurnal cycle. Topographic height and the land-sea mask are constant.

2.1.2 Model Climatology

The experiments presented use a fraternal twin configuration where the model itself provides the true state that the performance is evaluated against. As a result, model performance is not required to closely reflect reality. Nevertheless, if the results from these experiments are to be applied to a state-of-the-art system in the future, the model should provide a reasonable representation of the atmosphere. To evaluate how closely SPEEDY simulates the true atmosphere, a five year nature run from the T63 SPEEDY will be compared with the NCEP/NCAR Reanalysis (Kalnay et al., 1996) in both mean statistics as well as variance.

Figure 2.1 shows the zonal mean zonal wind for the December, January, and February months (DJF). Comparing the nature run to the reanalysis, they contain the same large scale features: westerly winds in the midlatitudes, weaker easterly

winds in the tropics, and high velocity jets in the upper atmosphere. The northern hemisphere jet is fairly accurate in location using SPEEDY, though its maximum is slightly weaker than in the reanalysis. The maximum in the southern hemisphere jet is displaced to the north and is higher in the atmosphere in the SPEEDY nature run. In the southern hemisphere, SPEEDY also has a hint of a double banded maximum in the upper troposphere, which is not seen in the reanalysis. When comparing the months of June, July, and August (JJA, Figure 2.2), the large scale similarities are present again. The jets in both hemispheres, however, are stronger in the nature run than in the reanalysis. More notable is the structure of the southern hemisphere jet in the model truth. The jet has a much stronger bimodal configuration than DJF, with two peaks in wind speed around 55°S and 25°S. There is only a slight indication of this in the reanalysis around 400 hPa. SPEEDY's jets in both seasons are elongated into the stratosphere compared to the reanalysis. This is likely due to the lack of vertical resolution within SPEEDY.

Figure 2.3 shows the DJF mean of mid-level geopotential heights, $\sigma = 0.5$ in SPEEDY and 500 hPa in the reanalysis, in the northern hemisphere. The large scale features are comparable with a trough extending through Kamchatka, though the trough is deeper in the reanalysis. Another trough exists in both datasets in Eastern Europe though SPEEDY contains standing waves around the Himalayas. The magnitudes of the heights are also lower around the North Pole in SPEEDY, likely due to SPEEDY treating the North Pole as a rigid boundary at 87°N. The comparison for JJA (Figure 2.4) is similar. Heights in the polar regions are too low in SPEEDY, with standing waves again present in the vicinity of the Himalayas.

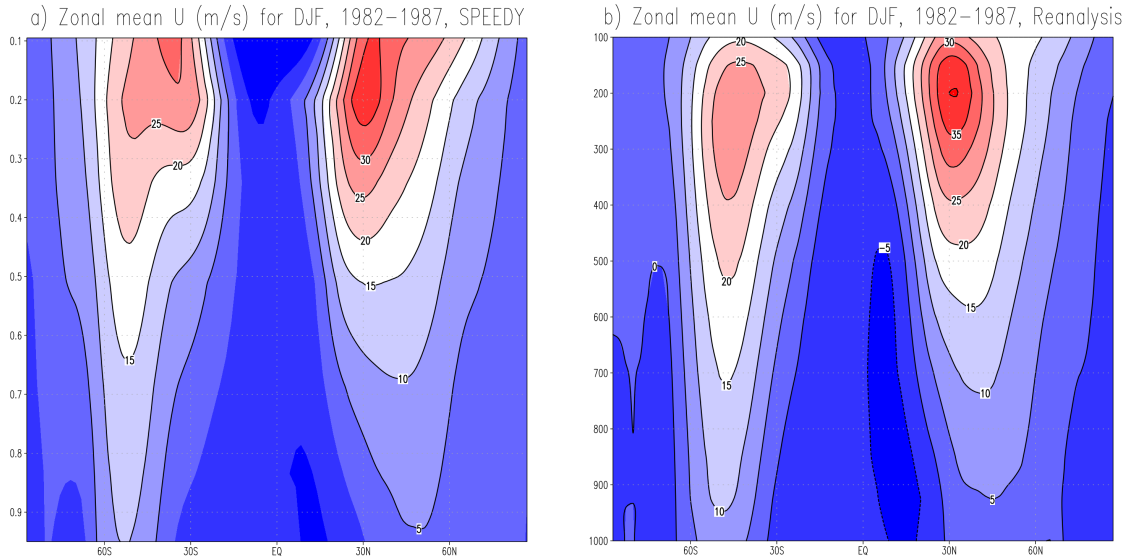

Figure 2.1: Zonal mean zonal wind in m/s for (a) the T63 nature run of the SPEEDY model and (b) the NCEP/NCAR Reanalysis, shown for DJF. The contouring interval is 5 m/s.

However, a trough exists by the Bering Strait and off the California coast in both figures.

The variability of a model can be equally as important as its mean state; therefore, the standard deviation of the nature run fields is compared against the NCEP/NCAR reanalysis. Figure 2.5 shows the zonal wind in the upper troposphere, $\sigma = 0.2$ in SPEEDY and 200 hPa in the reanalysis, for DJF. The values of the standard deviation in the nature run are lower than in the reanalysis on average. However, the spatial distribution is qualitatively similar. They both feature maxima in the North Pacific and Atlantic along the storm tracks. The double banded feature in the southern midlatitudes in both datasets mirrors the zonal mean zonal wind. It is much more prominent and zonally constrained in SPEEDY than the reanalysis. Comparing for the months of JJA (Figure 2.6), the spatial distribution is broadly similar: lower variability in the tropics and higher in the midlatitudes. Both

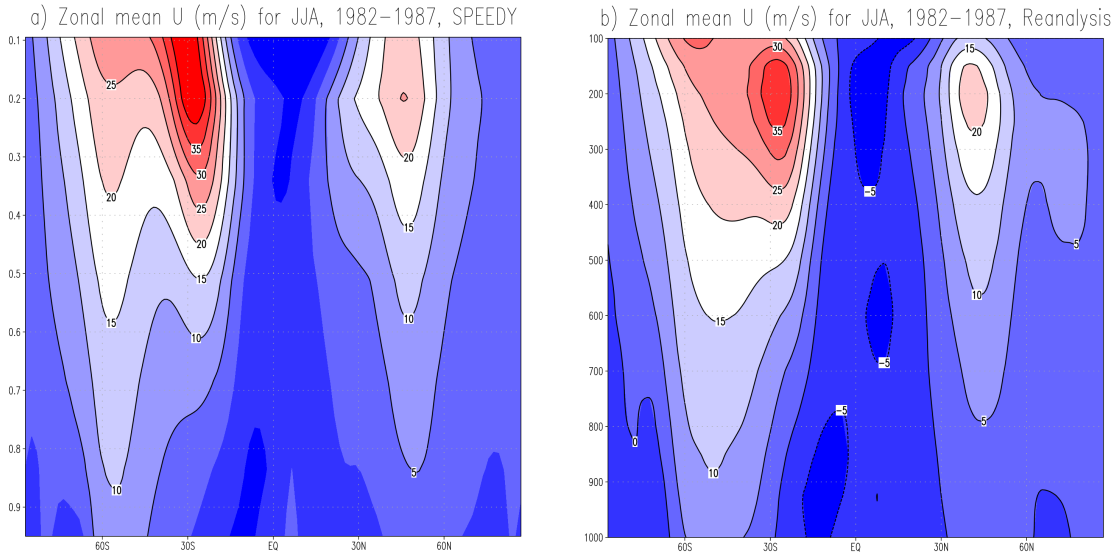

Figure 2.2: Zonal mean zonal wind in m/s for (a) the T63 nature run of the SPEEDY model and (b) the NCEP/NCAR Reanalysis, shown for JJA. The contouring interval is 5 m/s

datasets have a maximum in the eastern Pacific, though SPEEDY has a much larger magnitude. Similar to the zonal mean, the double banded feature is more distinct than DJF in SPEEDY with clear separation between the bands. The reanalysis has this feature to a lesser extent.

The variance of the global geopotential height, $\sigma = 0.5$ in SPEEDY and 500 hPa in the reanalysis, is also compared for DJF (Figure 2.7). The spatial pattern is quite similar with appropriate maxima locations in the North Pacific, North Atlantic, and South Pacific and low values throughout the tropics. Both datasets contain a wavenumber-3 pattern in the northern hemisphere. The JJA patterns (Figure 2.8) are not as closely represented, with the maximum around the Aleutian islands in the reanalysis missing in SPEEDY. There is also a large area of high variability in SPEEDY over northeast Asia and western Alaska that is not seen in the reanalysis. Since SPEEDY has a rigid boundary at 87°N and S, its representation

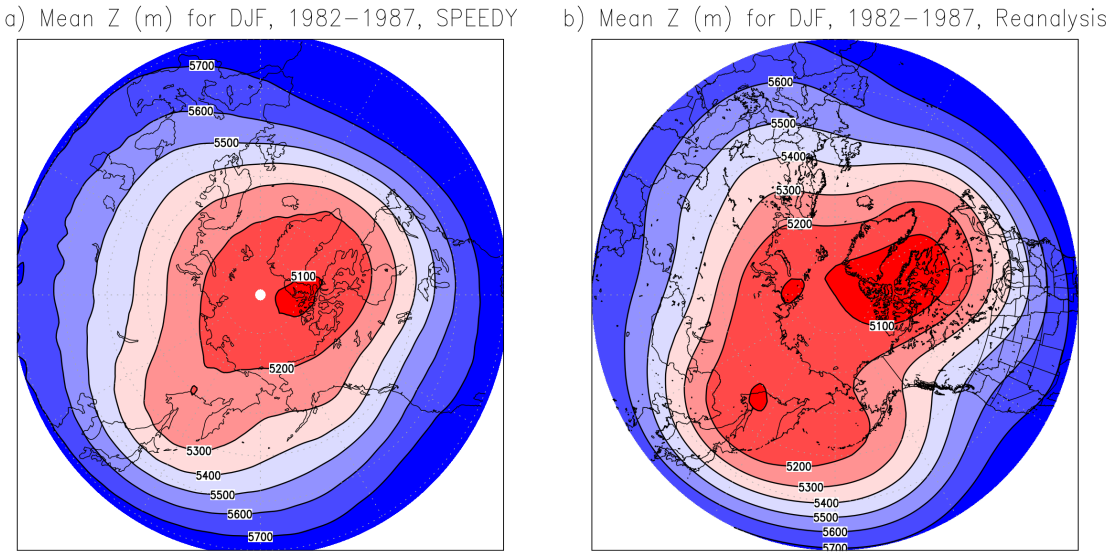

Figure 2.3: Mean northern hemisphere geopotential heights in meters for (a) the T63 nature run of the SPEEDY model at $\sigma = 0.5$ and (b) the NCEP/NCAR Reanalysis at 500 hPa, shown for DJF. The contouring interval is 100 m.

of the poleward-most regions are likely not well represented.

This series of comparisons with different reanalysis data sets demonstrate that the SPEEDY model has a reasonable large-scale circulation, suitable for intermediate level climate studies as it was designed. The finer scale structure and magnitude of some of the mean fields are lacking, but the key features are present, giving a realistic representation of the atmosphere. The variability of SPEEDY also does not match all of the features of the reanalysis, but the range of values and the large scale structure are similar. For the experiments presented, the model dynamics and physics appear accurate enough that any experiments dealing with large-scale balances should not be fundamentally different were they applied to a state-of-the-art AGCM.

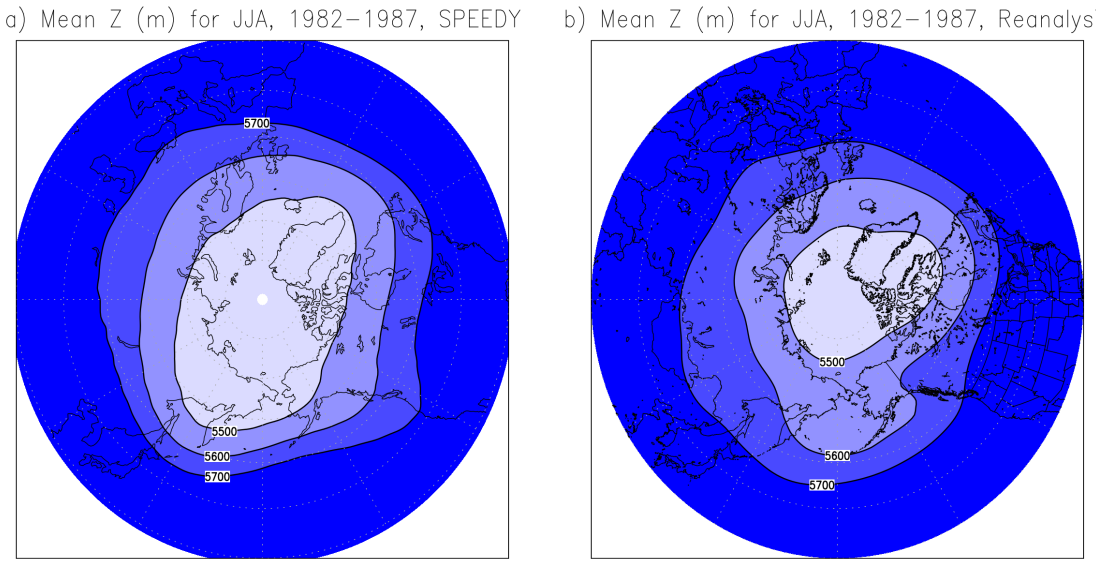

Figure 2.4: Mean northern hemisphere geopotential heights in meters for (a) the T63 nature run of the SPEEDY model at $\sigma = 0.5$ and (b) the NCEP/NCAR Reanalysis at 500 hPa, shown for JJA. The contouring interval is 100 m.

2.1.3 Model Bias

In the previous section, some of the deficiencies of the model were highlighted with comparison to the observed atmosphere, though when performing fraternal twin experiments, these model errors are not critical since our model provides the true state. Since the T63 resolution will be considered our true state, there will be model errors when the lower resolution T30 forecast is used with respect to the higher resolution that will be relevant to the experiment performance.

The method that will be used to diagnose initial model bias is from Danforth et al. (2007). The authors of that study calculated the initial model bias between SPEEDY and the NCEP/NCAR reanalysis. Their method is adopted to calculate the initial model bias between the high resolution and the low resolution SPEEDY configurations (Figure 2.9). A high resolution truth ($\mathbf{x}^t_{63}$) is integrated forward in

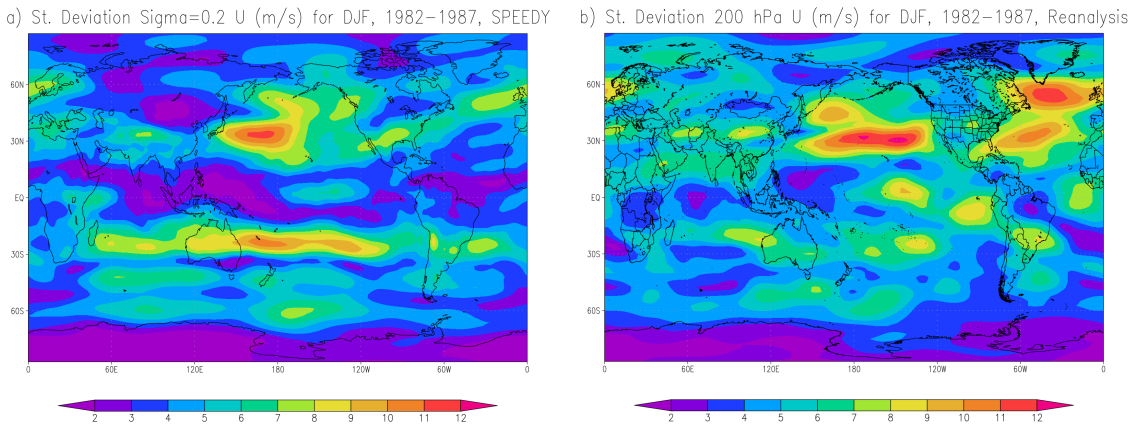

Figure 2.5: Standard deviation of monthly zonal wind in m/s for (a) the T63 nature run of the SPEEDY model at $\sigma = 0.2$ and (b) the NCEP/NCAR Reanalysis at 200 hPa, shown for DJF.

time with the high resolution model (M_{63}). The high resolution truth is spectrally truncated to make a low resolution equivalent ($\mathbf{x}_{30}^t$) with transformation $\mathbf{T}$. Then, the low resolution state is integrated forward six hours using the low resolution model (M_{30}) to get a low resolution forecast ($\mathbf{x}_{30}^f$). Over a period of five years, the difference between the T30 equivalent of the T63 truth and the T30 forecast ($\mathbf{x}_{30}^t - \mathbf{x}_{30}^f$) is calculated, producing the mean initial bias due to model resolution.

Figure 2.10 contains the initial model resolution bias of the zonal wind in the upper troposphere (σ=0.2) for (a) DJF and (b) JJA over a five year period. It is readily apparent that for both seasons the largest errors occur over the poles, with the stronger bias over the winter pole. There is also a region of small biases associated with the high topography of the Himalayas, which will be represented differently in the two resolutions.

The initial model resolution bias in the surface pressure fields (Figure 2.11) is almost exclusively associated with areas of sharp gradients in topography, such as the

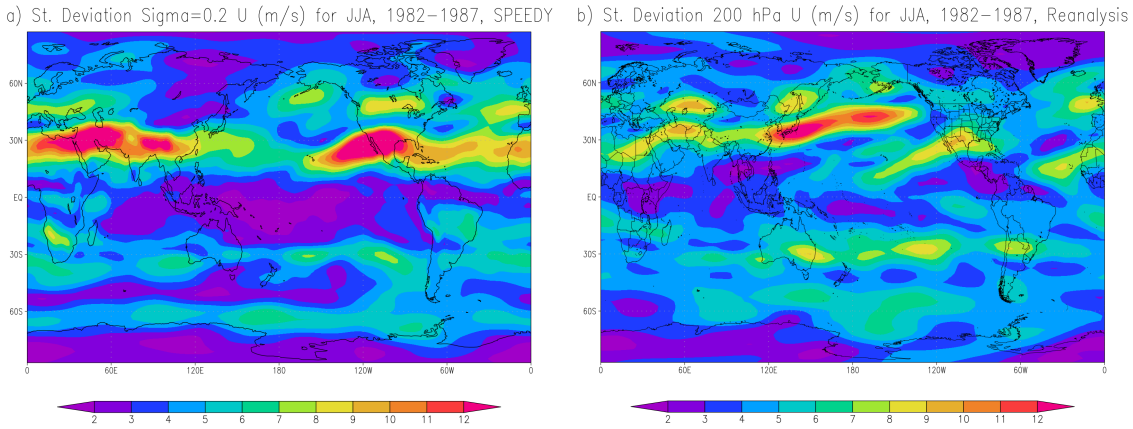

Figure 2.6: Standard deviation of monthly zonal wind in m/s for (a) the T63 nature run of the SPEEDY model at $\sigma = 0.2$ and (b) the NCEP/NCAR Reanalysis at 200 hPa, shown for JJA.

edges of the ice sheets in Antarctica and Greenland, the Himalayas, and the Andes, showing very little seasonal dependence. The lower resolution spectral model cannot resolve these topographical features to the extent that a higher resolution model can. The differences in surface pressure result predominantly from the definition of the surface. Comparing the surface orography from each resolution (not shown), there are areas that are several hundred meters different, which would lead to significant differences in surface pressure.

Figure 2.12 contains the initial model resolution bias for midtropospheric temperature (σ=0.51) for the DJF and JJA months. Large biases are again evident in the polar regions, though the seasonality of the winter pole is not apparent. There are some additional bias features closely associated with areas of high terrain, where the difference in the surface definition is reflected in the height of the σ levels.

The biases revealed using the method from Danforth et al. (2007) were due to the differing spectral model resolutions. Due to the nature of spectral models,

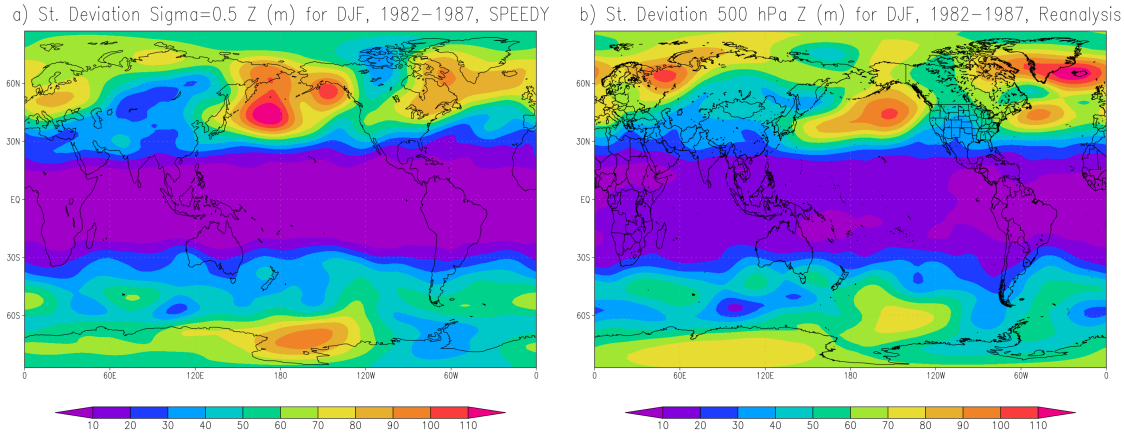

Figure 2.7: Standard deviation of geopotential height in m for (a) the T63 nature run of the SPEEDY model at $\sigma = 0.5$ and (b) the NCEP/NCAR Reanalysis at 500 hPa, shown for DJF.

this results in a noisy bias field rather than a bias that is more dynamics-based and therefore likely smoother. If bias correction were to be applied, these bias fields would be added into the analysis at every analysis cycle. The noise added by the bias correction could accumulate and be detrimental to the model forecast. The SPEEDY model, designed for long-term integrations, has been proven to be unstable in cases of high frequency instabilities and particularly sensitive to noise in the initial conditions. Appendix A contains results from previous experiments, demonstrating the nature of SPEEDY's instability and various steps that were taken to reduce it. For these reasons, we have chosen not to apply bias correction.

As presented in Danforth et al. (2007), the six hour forecast comparisons are supposed to reveal the initial biases that exist between the two models, which are likely linear in nature. Comparing the mean upper level zonal wind of a low resolution integration with the nature run for JJA (Figure 2.13), it is clear that there is significant systematic bias that the previous method did not capture. The

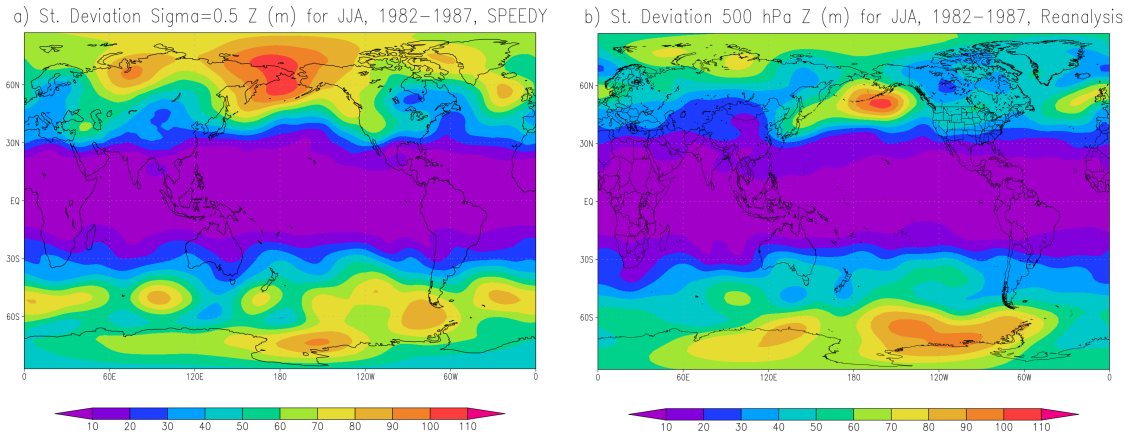

Figure 2.8: Standard deviation of geopotential height in m for (a) the T63 nature run of the SPEEDY model at $\sigma = 0.5$ and (b) the NCEP/NCAR Reanalysis at 500 hPa, shown for JJA.

stratospheric zonal wind speed is notably weaker globally. This damped wind speed is the result of inconsistent stratospheric damping timescales between resolutions. These timescales are longer than six hours, resulting in the previous method being insufficient to diagnose such biases as implemented here. If longer forecasts were created for the comparison, it is likely that this bias could have been diagnosed using Danforth et al's method.

2.2 Observation Network

The synthetic observation network used in the following experiments have two components: a synoptic radiosonde network and an asynoptic satellite network. All observations are created by adding a Gaussian random error to the T63 true state, scaled by the observation error associated with each observation type and instrument. All observation errors are assumed to be uncorrelated.

The radiosonde portion of the observation network consists of 416 station-

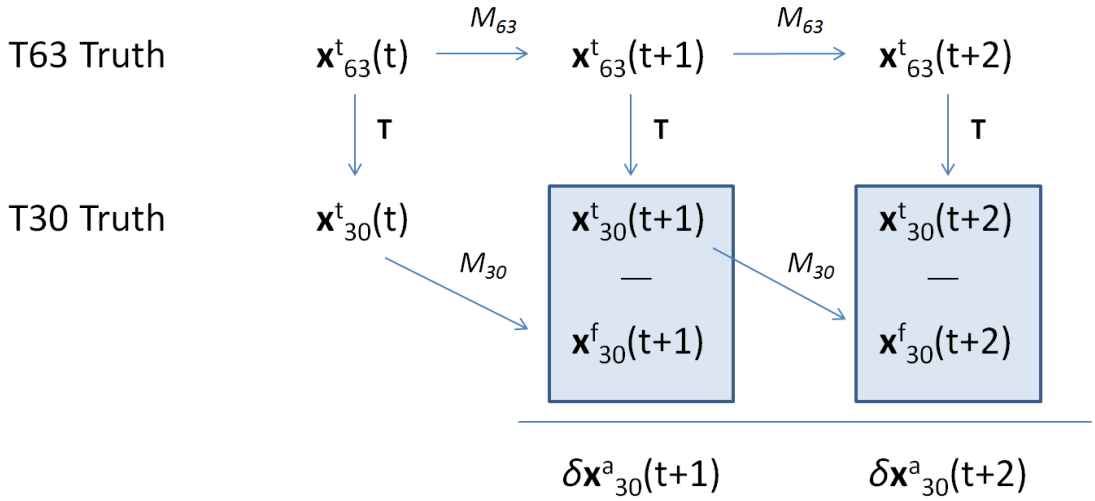

Figure 2.9: Schematic depicting the calculation of the initial model bias.

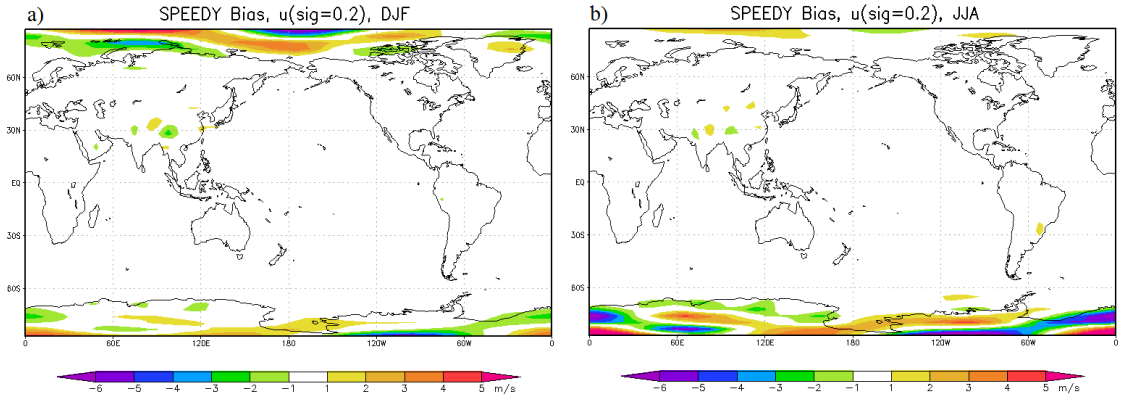

Figure 2.10: Initial bias between the T63 and T30 resolution from 1982-1987 for zonal wind at $\sigma = 0.2$ in m/s: a) DJF and b) JJA

ary sounding locations (Figure 2.14). The observation locations were constructed by Miyoshi (2005) to have a realistic distribution. The radiosondes are unevenly distributed, with more observations occurring over land than over ocean and more observations in the northern hemisphere than in the southern hemisphere. For each observation station, prognostic variables u, v, and T are present at all vertical levels, q at the bottom four levels, and P at the surface, with observations occurring every

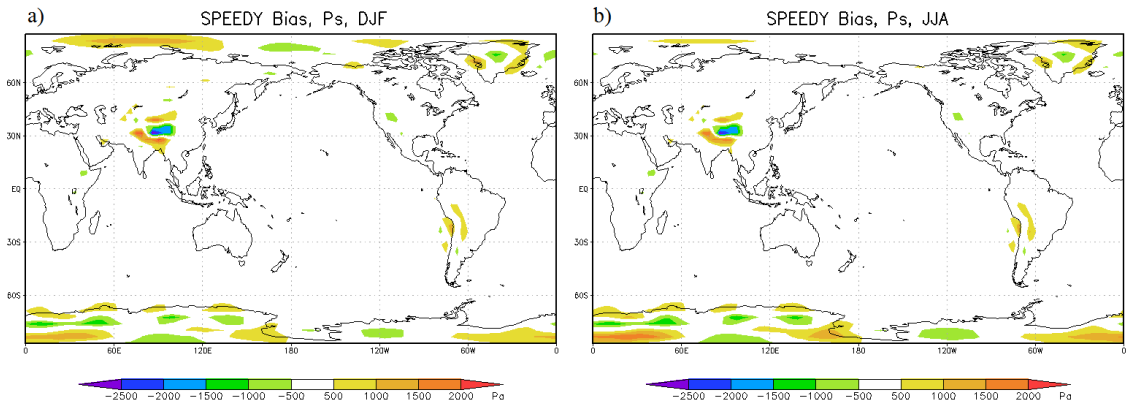

Figure 2.11: Initial bias between the T63 and T30 resolution from 1982-1987 for surface pressure in hPa: a) DJF and b) JJA

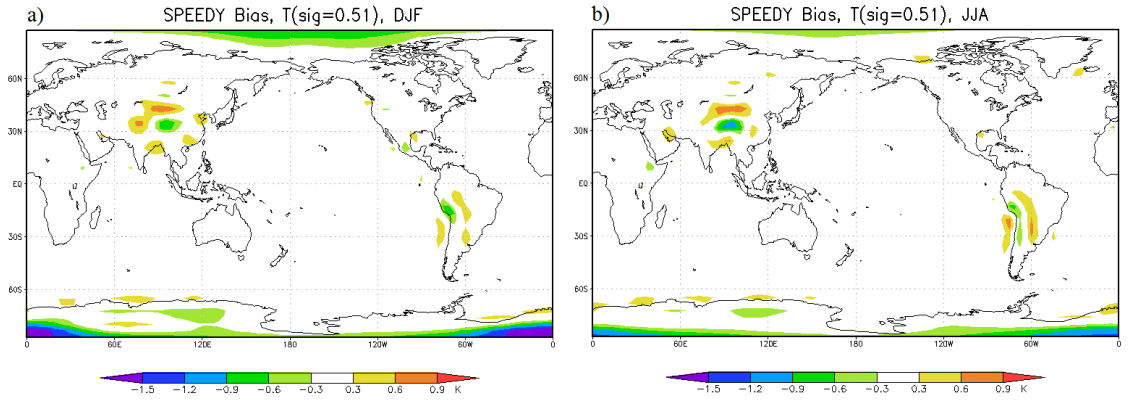

Figure 2.12: Initial bias between the T63 and T30 resolution from 1982-1987 for temperature at $\sigma = 0.51$ in K: a) DJF and b) JJA

six hours. The radiosonde observation errors for each variable type are shown in Table 2.1.

The asynoptic component of the observations comes from the simulated satellite observations. Observation values, in the form of retrievals at nadir only, are created in the same manner as the radiosonde data, using the truth and adding a random Gaussian error to it. The observation locations, error, and variable types were chosen to mimic the AIRS instrument on the Aqua satellite and SeaWinds on the QuikSCAT satellite. The simulated AIRS instrument provides retrievals for

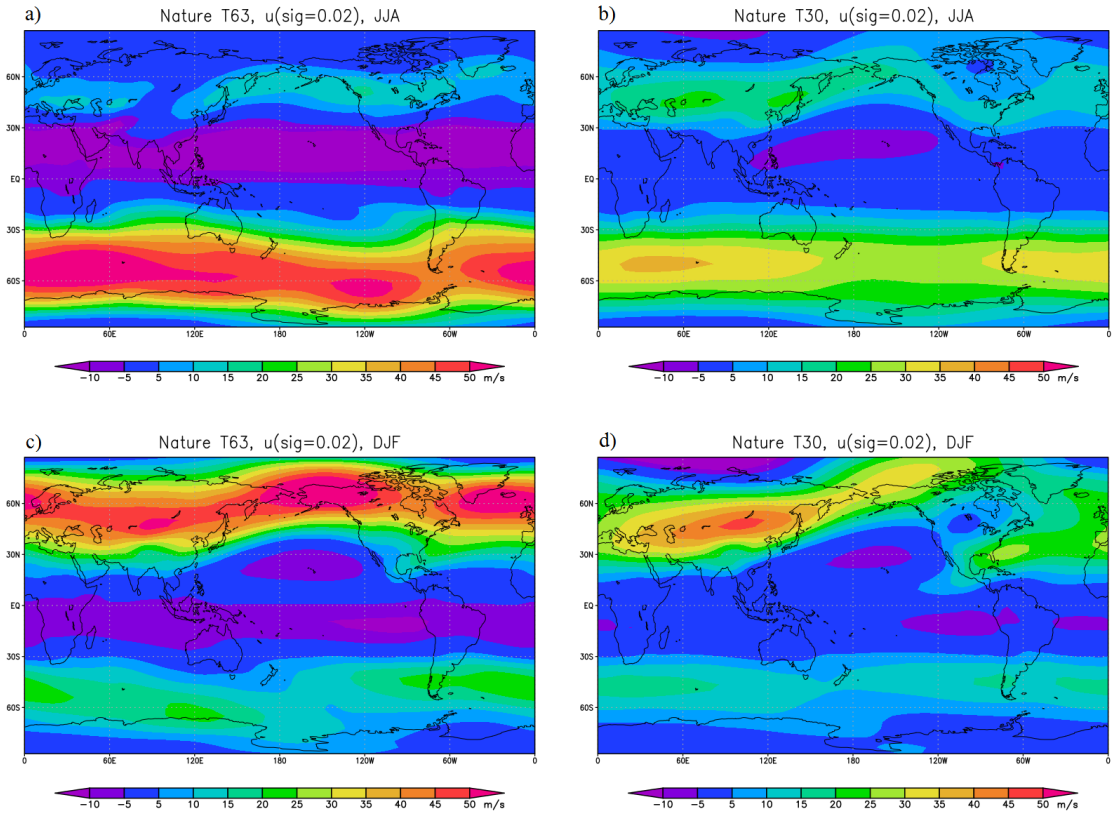

Figure 2.13: Mean zonal wind at the top model level for (top) June, July, August and (bottom) December, January, February for horizontal resolutions (left) T63 and (right) T30.

T (full column) and q (lowest four levels) while the simulated SeaWinds provides information about the surface winds over the ocean, represented as u and v observations at the lowest model level. Observation locations were chosen to mirror the satellite tracks as closely as possible, with the observations for AIRS (green) and SeaWinds (red) over a six hour period shown in Figure 2.14. From the simulated tracks, observations were taken at two minute intervals and binned hourly, giving approximately 2,300 observations every six hours. While observation locations vary by hour, they are identical for each day, i.e. 00z locations are the same for each day. The observation errors of the satellite data are higher than the radiosonde data,

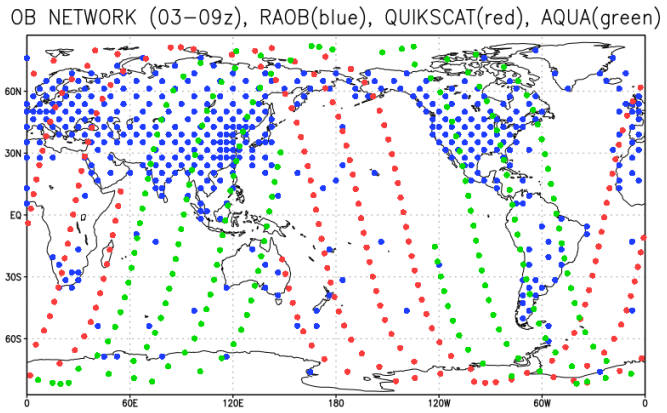

Figure 2.14: Observation distribution with radiosondes in blue, QuikSCAT-like observations in red and Aqua-like observations in green.

Table 2.1: Radiosonde observation error standard deviations for the prognostic variables in the SPEEDY model.

Observation Type	u	v	T	q	P
Observation Error	1 m s^{-1}	1 m s^{-1}	1 K	10^{-4} kg kg^{-1}	100 Pa

shown in Table 2.2.

Table 2.2: Satellite observation error standard deviations for the prognostic variables in the SPEEDY model.

Observation Type	u	v	T	q
Observation Error	1.5 m s^{-1}	1.5 m s^{-1}	2 K	2×10^{-4} kg kg^{-1}

Using both observing system components results in approximately 14,000 observations per cycle. While much less than the number of observations available for operational NWP systems, the dimension of SPEEDY is only $O(10^5)$. Assimilating $O(10^4)$ observations for a $O(10^5)$ system gives a ratio that is comparable to

operational systems.

2.3 Data Assimilation Configuration

A hybrid 4DEnVar was implemented within the SPEEDY system. The background ensemble and deterministic forecast are saved hourly for a six hour data assimilation window centered at synoptic times. The background covariances are composed of 10% static contribution and 90% ensemble contribution, $(\beta^f, \beta^e) = (\sqrt{0.1}, \sqrt{0.9})$, with $M = 20$ ensemble members used. Several values for the covariance weights and ensemble size were tested, tuning for analysis error reduction and long-term stability.

The static background error, $\mathbf{B}^f$, is calculated using the NMC method (Parrish and Derber, 1992). This method uses a series of lagged forecast pairs to calculate the climatological statistics of the background error. To generate the background error for these experiments, one year of 24 and 48 hour forecasts was generated. Statistics for the background error variance and length scales for the recursive filter (both contained within $\mathbf{U}^f$) and regression coefficients $(c, \mathbf{G}, \Omega)$ for the balance operator ($\mathbf{\Gamma}$, see Chapter 3 for details) were computed from pairs of forecasts that are valid at the same time. The coefficients used in the following experiments were calculated using forecasts generated by the SPEEDY model, though the structure of the coefficients are comparable to those generated by NCEP's Global Forecast System (GFS, Figure 2.15), demonstrating that the flow within SPEEDY contains realistic large scale balances.

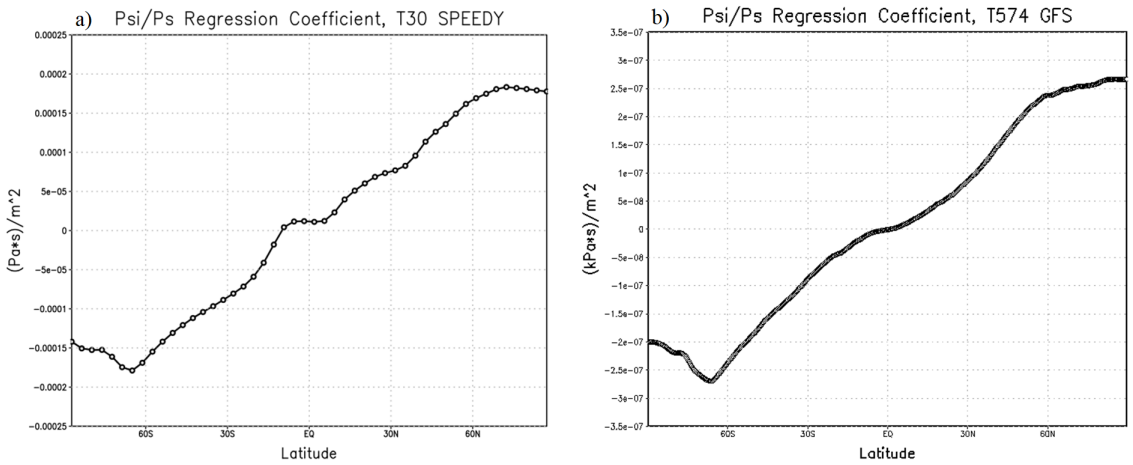

Figure 2.15: Regression coefficient Ω between variables ψ and P as calculated using the NMC method in (a) SPEEDY at horizontal resolution T30 and (b) GFS at horizontal resolution T574.

The localization of the ensemble perturbations is applied through a fourth order recursive filter, $\mathbf{F}$ (Purser et al., 2003). See Appendix B for the details of its formulation. Due to the extremely coarse vertical resolution of SPEEDY, the recursive filter is not applied in the vertical. The recursive filter is less effective at the boundaries, which for a fourth order filter is four grid points deep. Since SPEEDY only has eight vertical levels, every point would be a boundary. Instead, the ensemble correlations are fully localized in the vertical, with each level being independent of one another.

The analysis perturbations are provided by a local ensemble transform Kalman filter (LETKF, Hunt et al., 2007), which runs in parallel with the hybrid assimilation that generates the deterministic analysis. After the LETKF computes an analysis ensemble, the ensemble is recentered about the hybrid analysis rather than the calculated analysis ensemble mean. Using the same number of ensemble members as the hybrid system, the LETKF uses a fixed multiplicative inflation of 8%, horizontal

localization radius of 750 km, and vertical localization of 0.1 $\log(p)$ where p is the three-dimensional pressure. The multiplicative inflation and localization radii were also tuned with a series of experiments varying each parameter and evaluating the analysis root-mean-squared error (RMSE, Figure 2.16). The multiplicative inflation values tested ranged from 3% to 15% and the horizontal localization parameters tested ranged from 250 km to 1500 km. Since the localization in the LETKF is implemented through a distance dependent function rather than a recursive filter, this vertical localization does not have difficulties with the sparse vertical resolution of the model. The LETKF also uses a 4D configuration with hourly observation bins and model forecasts and a time localization length scale of three hours.

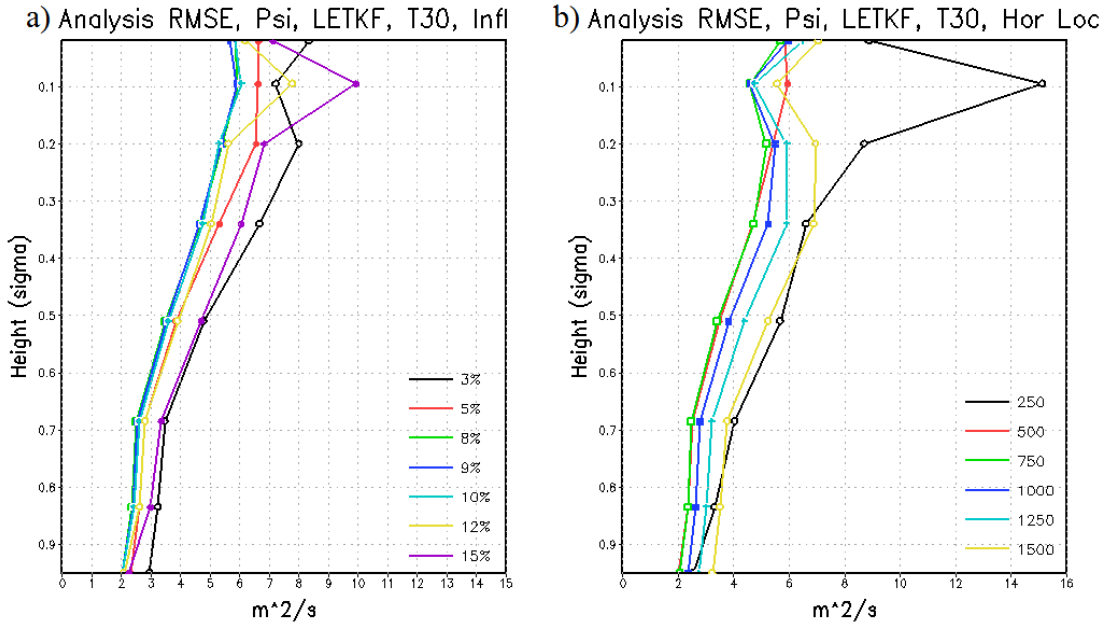

Figure 2.16: Analysis RMSE for ψ in LETKF for two years of cycling. a) Fixed multiplicative inflation parameters ranging from 3% to 15% and b) Horizontal localization radii ranging from 250 km to 1500 km.

The experiments that follow use a fraternal twin configuration where the truth is provided by the model itself at the higher resolution, T63, and the analysis and

forecasts are run at a lower resolution, T30. Having the truth and forecasts at different resolutions adds model error to the system (Section 2.1.3), making it more realistic and therefore allowing the proposed method to be more applicable to operational NWP. The true state was initialized from a state of rest on January 1, 1981 with the first year thrown out for spin up. Initial conditions for the T30 ensemble and deterministic forecasts were chosen from an interpolated true state at different times with the experiment period running from January 1, 1982 through January 1, 1984.

Chapter 3: Balance Operators in Ensemble Data Assimilation: Hybrid 4DEnVar

3.1 Introduction

Numerical weather prediction (NWP) is an initial value problem that needs both a representative numerical forecast model and appropriate initial conditions in order to make an effective forecast. With the chaotic nature of Earth's atmosphere, initial conditions must be as close to the truth as possible but also must be balanced to prevent the production of gravity waves, which propagate through the model and degrade the forecast. Imbalances in the initial conditions led to the failure of the first numerical weather prediction forecast by Lewis Fry Richardson in the early 20th century (Kalnay, 2003).

Many methods have been applied over the years to handle imbalances that data assimilation systems create. In variational schemes, these methods have been traditionally formulated as a strong constraint, a weak constraint, or initialization. The two types of constraints, initially formalized by Sasaki (1970), are incorporated within the data assimilation scheme itself, while initialization is treated as a post-processing step once the assimilation is complete. Strong constraints assume that

the balance in the model is perfect and strictly enforces it. Le Dimet and Talagrand (1986) formulated a strong constraint referred to as "a reduction in control variable", where the analysis is found for one variable and the other variables are found through the appropriate balance equations. In contrast, weak constraints assume that the model balance is approximate. Lorenc (1986) argues that since the balances that are typically represented in strong and weak constraints are not present at all scales, weak constraints are more appropriate than strong constraints in NWP. Many weak constraints formulations separate the model variables into balanced and unbalanced components through the use of a balance operator (Lorenc et al., 2003; Parrish and Derber, 1992; Wu et al., 2002) or add a penalty to the cost function (Courtier and Talagrand, 1990; Gauthier and Thépaut, 2001; Wee and Kuo, 2004). The addition of a term in the cost function is problematic since the cost function often becomes ill-conditioned and minimization is adversely impacted.

Balance operators were originally constructed for global models (Derber and Bouttier, 1999; Gauthier et al., 1999; Lorenc et al., 2003; Parrish and Derber, 1992), where geostrophic and hydrostatic balance are prevalent. As grid sizes continue to decrease as computational power increases and models relax the hydrostatic approximation, a significant portion of the flow no longer ascribes to these balances. Vetra-Carvalho et al. (2012) found that hydrostatic balance breaks down at 1.5 km horizontal resolution in the Met Office Unified Model. Non-geostrophic and non-hydrostatic cross-covariances can be dealt with in mesoscale models through the model itself within 4DVar (e.g. Kuo et al., 1996; Zou and Kuo, 1996; Zou et al., 1995) and ensemble-derived covariances. EnKFs have been used in mesoscale ap-

plications with much success (e.g. Caya et al., 2005; Snyder and Zhang, 2003; Tong and Xue, 2005). There have also been attempts at modifying the balance operator to allow for convective scale motions. Honda et al. (2005) separated the error into synoptic and mesoscale components, calculating the regression coefficients from forecasts that have had a low-pass filter applied to them. Barker et al. (2004) added an additional term to the calculation of balanced pressure to include cyclostrophic balance.

Initialization procedures are performed outside of the cost function after an analysis is obtained (Huang and Lynch, 1993; Machenhauer, 1977). In an effort to reduce the imbalance that the assimilation created, initialization often pulls the analysis back away from the observations, frequently undoing the work of the assimilation (Errico et al., 1993; Williamson et al., 1981). In order to reduce the degradation potentially imparted by the initialization, it is possible to include the initialization within the analysis calculation itself. Nonlinear normal-mode initialization (Baer and Tribbia, 1977; Machenhauer, 1977), where the time tendencies of the forecast model are projected onto the fast gravity wave modes and provide a correction to the original state, was included in a variational analysis by Courtier and Talagrand (1990) and Thepaut and Courtier (1991). Operationally, the National Centers for Environmental Prediction's (NCEP's) Global Data Assimilation System (GDAS) uses a tangent-linear normal-mode constraint (TLNMC, Kleist et al. (2009a)). It follows the theory of normal-mode initialization, but since the GDAS does not have a full nonlinear model available within the analysis cycle, a tangent-linear form of the tendency model is used instead, resulting in a reduction of imbalances in the

analysis without being computationally prohibitive.

In ensemble methods, the covariances are provided by the model itself, so imbalances within the analysis should be small provided that the covariances are not tampered with. However, in an attempt to handle shortcomings such as sampling error and an under-dispersive ensemble, the covariances that the model provides are modified. Sampling error due to small ensemble size is mitigated through the use of localization, eliminating covariances that are likely spurious and unphysical. This can be achieved by modifying the background error (decorrelating grid points that are at large distances from one another, e.g. Houtekamer and Mitchell (2001)) or modifying the observation error (increasing the error for observations that are located far from the grid point in question, e.g. Hunt et al. (2007)). These localization techniques create imbalances within the analysis (Buehner, 2005; Cohn et al., 1998; Greybush et al., 2011; Kepert, 2009; Lorenc, 2003; Mitchell et al., 2002). By bringing the correlations to zero at a certain distance from the grid point, the balance between one variable and another variable's gradient are greatly affected, particularly in height and wind relationships. Localizing in stream function and velocity potential space rather than zonal and meridional wind space has been shown to reduce imbalances due to the covariances of stream function and velocity potential being more isotropic than the covariances of zonal and meridional wind (Kepert, 2009).

This chapter evaluates a method of improving balance in the ensembles: applying a balance operator to the ensemble perturbations of a hybrid 4D ensemble-variational (4DEnVar) system and thereby localizing on the unbalanced variables

only, with the formulation being described in Section 3.2. Section 3.3 outlines the results of single observation impact tests as well as full network experiments. A summary and conclusions are presented in Section 3.4. The contents of this chapter are contained in Thomas and Ide (2017a).

3.2 Method

3.2.1 Balance Operator

In variational methods of data assimilation, analysis control variables are typically chosen such that their background errors are uncorrelated with one another. This allows for a much less challenging inversion of the background error covariance matrix (Parrish and Derber, 1992). In most cases, the multivariate correlations are provided by a dynamic constraint or balance operator, which represents prominent physical relationships between the variables. In the atmosphere, the strongest relationships that are typically represented are hydrostatic and geostrophic balance, which relate the horizontal wind with the temperature and surface pressure.

Following Wu et al. (2002), the increments of velocity potential χ, temperature T, and surface pressure P are broken down into balanced and unbalanced components. The balanced parts are formed by using a statistical linear regression with

the stream function ψ:

$$\delta\chi = \delta\chi^u + c\delta\psi, \tag{3.1a}$$

$$\delta T = \delta T^u + \mathbf{G}\delta\psi, \tag{3.1b}$$

$$\delta P = \delta P^u + \Omega\delta\psi, \tag{3.1c}$$

where δ indicates the increment, superscript u represents the unbalanced component, $c \in \mathbb{R}^{Q\times Q}$, $\mathbf{G} \in \mathbb{R}^{Q\times Q}$, and $\Omega \in \mathbb{R}^{S\times Q}$ are linear regression coefficients, S is the number of horizontal grid points per level, and Q is the total number of grid points. The balanced part of $\delta\chi$ is $c\delta\psi$, where $\delta\psi$ from a particular level only affects $\delta\chi$ at that same level, making c a diagonal matrix. The largest contribution from this correlation is from the planetary boundary layer, as ψ and χ are typically uncorrelated in the free atmosphere. Hollingsworth and Lonnberg (1986) confirmed the lack of correlation between the ψ and χ background errors using radiosonde data over North America. The balanced part of δT is $\mathbf{G}\delta\psi$, where each level of $\delta\psi$ contributes to δT at all levels, i.e. there is a vertical profile of balanced δT for each vertical level of $\delta\psi$. The balanced part of δP is $\Omega\delta\psi$, where all levels of $\delta\psi$ contribute to δP, with the largest contribution being from the lowest model level. The coefficients in c, $\mathbf{G}$, and Ω are estimated using the NMC method (Section 2.3) and assumed to vary with latitude and height only.

Figure 3.1a shows the $\mathbf{G}$ regression coefficient for a single level of $\delta\psi$ ($\sigma = 0.34$) as calculated from the SPEEDY model. Any adjustments made to $\delta\psi$ at that level will project onto a column of balanced δT that is latitude dependent. Since this

coefficient represents thermal wind balance, $\mathbf{G}$ and consequently the balanced part of T become very small close to the equator and minimal adjustments will be made in the tropics. Figure 3.1b contains the regression coefficients at each level for a certain latitude, 40°N. For example, the purple line shows the balanced part of δT that is correlated with the bottom level of $\delta\psi$ at $\sigma = 0.95$. The yellow line corresponds to the level in Figure 3.1a.

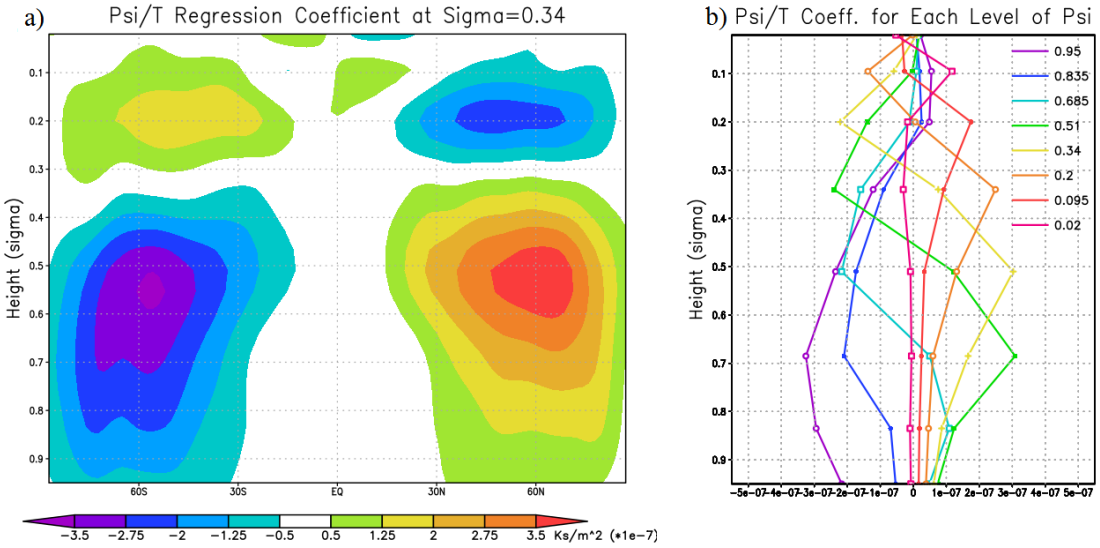

Figure 3.1: a) Linear regression coefficients, $\mathbf{G}$, for the SPEEDY model between $\delta\psi$ at $\sigma = 0.34$ and δT at all levels, dictating columns of δT^b which are latitudinally dependent. b) $\mathbf{G}$ at each level for 40°N. Each line represents a column of δT^b for a single level of $\delta\psi$.

The wind variables used in this balance operator construction, ψ and χ, are customarily preferred as control variables over the zonal and meridional wind, u and v respectively, in both variational (Daley, 1991) and ensemble (Kepert, 2009) applications. The two cartesian wind components of u and v are highly correlated with each other and their background errors are anisotropic. The nondivergent and irrotational wind components of ψ and χ have small cross covariances and contain isotropic self-correlations, making them more suitable for a symmetric and easily

invertible background error.

The increment of the model variables is represented as $\delta\mathbf{x} = (\delta\psi^T, \delta\chi^T, \delta T^T, \delta q^T, \delta P^T)^T \in \mathbb{R}^N$ where N is the dimension of the model. The increment for the variables in the unbalanced space, which are the control variables, are similarly represented as $\delta\mathbf{z} = (\delta\psi^T, (\delta\chi^u)^T, (\delta T^u)^T, \delta q^T, (\delta P^u)^T)^T \in \mathbb{R}^N$. The three balance equations, (3.1a) - (3.1c), are combined into a matrix, $\mathbf{\Gamma} \in \mathbb{R}^{N\times N}$, which will be referred to as the balance operator. When applied to $\delta\mathbf{z}$, it transforms the control variables to the total model variables:

$$\delta\mathbf{x} = \mathbf{\Gamma}\delta\mathbf{z}. \tag{3.2}$$

The balance operator transformation operates in the vertical only. While the variable transformation does not operate in the horizontal, impacts in the horizontal arise due to the choice of control variable. Changes to ψ and χ at a particular point result in changes to u and v away from that point, allowing the balance operator to represent horizontal correlations that are in close proximity. However, the balance operator is not able to represent long distance horizontal correlations, such as those associated with El Niño—Southern Oscillation.

This balance operator is applied within the data assimilation scheme itself, unlike initialization methods that occur after the analysis is found (e.g. Machenhauer, 1977), often pulling the state away from the observations to bring the analysis into balance (e.g. Williamson et al., 1981). Within the following sections, $\mathbf{\Gamma}$ is instead incorporated into the cost function minimization, taking balance into account during

the analysis calculation.

3.2.2 Variational application within a Hybrid 4DEnVar

Hybrid methods combine the advantages of both variational and ensemble methods (Buehner, 2005; Hamill and Snyder, 2000; Lorenc, 2003). The flow dependent errors of the ensemble system compensate for the climatological nature of the 3DVar background. Furthermore, adding a portion of the full rank 3DVar background covariance to the ensemble covariance compensates for the ensemble covariance being rank deficient. The hybrid method presented here is the extended control variable hybrid 4DEnVar as described in Kleist and Ide (2015b) and implemented at NCEP for the Global Forecast System (GFS) in May 2016. It uses a cost function framework, which can include other penalties for balance, bias, and quality control, and would be attractive to operational centers that already have a variational system in place. In this formulation, the ensemble enters the standard 3DVar cost function through an additional background term representing the weights of each ensemble member:

$$J(\mathbf{v}) = \frac{1}{2}\left(\mathbf{v}^f\right)^T \mathbf{v}^f + \frac{1}{2}\left(\mathbf{v}^e\right)^T \mathbf{v}^e + \frac{1}{2}\sum_{k=1}^{K}\left(\mathbf{d}_k - \mathbf{H}_k\delta\mathbf{x}_k\right)^T \mathbf{R}^{-1}\left(\mathbf{d}_k - \mathbf{H}_k\delta\mathbf{x}_k\right)^T \quad (3.3)$$

The control vector, $\mathbf{v} = ((\mathbf{v}^f)^T, (\mathbf{v}^e)^T)^T \in \mathbb{R}^{N+QM}$ is comprised of the control variables for the static and ensemble parts respectively where N is the dimension of the model and M is the number of ensemble members. The observation term of the cost function is constructed as in 3DVar and then summed over each time level, where

K is the total number of k time levels, $\mathbf{d}_k \in \mathbb{R}^L$ is the innovation, $\mathbf{H}_k \in \mathbb{R}^{L\times N}$ is the linear observation operator, $\delta\mathbf{x}_k \in \mathbb{R}^N$ is the increment, and $\mathbf{R} \in \mathbb{R}^{L\times L}$ is the observation error covariance.

The multivariate correlations for the ensemble portion of the background error covariance are provided by the ensemble itself, but a balance operator is needed to provide the multivariate correlations for the static 3DVar portion, where the control variable is chosen to be in the unbalanced variable space. By having a control vector whose multivariate correlations are small, the construction of the static background error, $\mathbf{B}^f \in \mathbb{R}^{N\times N}$, is simplified. Since the ensemble perturbations already contain multivariate correlations, $\mathbf{\Gamma}$ is conventionally only applied to the unbalanced static control variable.

The increment at time k, $\delta\mathbf{x}_k$, is constructed as a weighted sum of the static, or fixed (f), portion and the ensemble (e) portion:

$$\delta\mathbf{x}_k = \beta^f \mathbf{\Gamma}\mathbf{U}^f\mathbf{v}^f + \beta^e \sum_{m=1}^{M} (\mathbf{F}\mathbf{v}_m^e \circ (\mathbf{X}_m^e)_k), \tag{3.4}$$

where β^f and β^e are the scalar weighting coefficients for the static and ensemble covariance respectively, $\circ$ is the element-by-element multiplication operator or Schur product, and $(\mathbf{X}_m^e)_k \in \mathbb{R}^N$ is the m^{th} column of the background perturbations scaled by $\sqrt{M-1}$ at time k. Each part of the control vector is preconditioned on the square root of its own background error covariance to allow for a more rapid minimization of the cost function as it is typically ill conditioned. $\mathbf{U}^f \in \mathbb{R}^{N\times N}$ is the square root of the static background error, $\mathbf{B}^f = \mathbf{U}^f(\mathbf{U}^f)^T$, so the term $\mathbf{U}^f\mathbf{v}^f$ represents

the unbalanced static increment, $\delta\mathbf{z}^f$. The ensemble error covariance is a block diagonal matrix with each block comprised of a forward, ($\mathbf{F} \in \mathbb{R}^{Q\times Q}$), and backward, ($\mathbf{F}^T$), recursive filter (Purser et al., 2003). The recursive filter provides the spatial correlation for the ensemble error covariance matrix by spreading an impulse, or in this case an ensemble weight, spatially into a quasi-Gaussian shape. This dictates the geographical extent of the ensemble's influence, allowing the recursive filter to also function as the spatial localization of the ensemble perturbations. For more details on the recursive filter, see Appendix B.

The ensemble portion of the increment contains the ensemble perturbations of the total variables. The static portion of the control variable is in the unbalanced variable space. When $\mathbf{\Gamma}$ is applied to $\mathbf{U}^f\mathbf{v}^f$, the balanced part is incorporated into the static increment, resulting in a hybrid increment $\delta\mathbf{x}$ in the total variable space.

3.2.3 Ensemble application within a Hybrid 4DEnVar

In ensemble methods, the correlations within the ensemble members provide statistical information about the errors of the system. This includes information on the intervariable correlations, which is the same type of information provided by the balance operator (3.2). However, $\mathbf{\Gamma}$ is based on mean climatological statistics and does not represent the time dependent flow. This could lead to the belief that implementing a dynamic constraint into an ensemble data assimilation system would be redundant or even damaging to the numerical forecast. If the ensembles remained as the model provided, the covariances would be in balance, i.e. the variables would

be consistent with each other according to model dynamics and physics. Due to small ensemble size and sampling error, modifications are made to the ensemble to reduce filter divergence. When attempting to remove spurious correlations through spatial localization, some physical correlations are also inadvertently removed and add imbalance to the ensemble. By applying a balance operator to the ensemble portion of the covariance, the balanced part can be removed from the perturbations and the localization will work on the perturbations in the unbalanced space only. This will reduce the imbalance added by the localization.

Within a Hybrid EnVar, $\mathbf{\Gamma}$, is traditionally applied to the static control variable, $\mathbf{v}^f$ as in (3.4). If $\mathbf{\Gamma}$ is to be applied to the ensemble perturbations as well, instead of having an additional application to the ensemble portion of the increment, $\mathbf{\Gamma}$ is applied to the full increment:

$$\delta\mathbf{x}_k = \mathbf{\Gamma}\left(\beta^f \mathbf{U}^f \mathbf{v}^f + \beta^e \sum_{m=1}^{M} (\mathbf{F}\mathbf{v}_m^e \circ (\mathbf{Z}_m^e)_k)\right). \tag{3.5}$$

$(\mathbf{Z}_m^e)_k \in \mathbb{R}^N$ is the m^{th} column of the background perturbations in the unbalanced space scaled by $\sqrt{M-1}$ at time k. These perturbations are calculated by applying $\mathbf{\Gamma}^{-1}$ to the perturbations in the total variable space, $(\mathbf{X}_m^e)_k$, at the beginning of each analysis cycle. There is no formulation change to the cost function (3.3), only to the increment used within it.

This method allows for the ensemble localization, implemented as a recursive filter in the matrix $\mathbf{F}$, to work only on the control variables in the unbalanced space rather than the model variables. This preserves the balanced part and allows

for a propagation of information outside of the physical localization radius of the ensemble. This formulation was previously adopted by the UK Met Office in their operational hybrid ensemble-4DVar (Clayton et al., 2013) and is being investigated by the Chinese Meteorological Administration for a future implementation of the GRAPES Hybrid 3DEnVar (Chen et al., 2015).

3.3 Experiment Results

3.3.1 Impact Tests

To assess the impact of adding the balance operator to the ensemble portion of the increment, single observation experiments are performed. Within 3DVar, or alternatively $(\beta^f, \beta^e) = (1, 0)$ within a hybrid 4DEnVar, the multivariate correlations are determined solely the balance operator. Using this data assimilation scheme when assimilating a single observation isolates its effect. Figure 3.2a shows the analysis increments for T (contour) and ψ (shaded) at the lowest model level where a single T observation has been assimilated. The T increment is isotropic about the observation location as an effect of the climatological covariances and lack of flow dependence. Since there are no wind observations, $\delta\psi$ is determined solely by the multivariate covariances, which in this static case is $\mathbf{\Gamma}$. The ψ increment is negatively correlated with δT, which can be surmised by examining Figure 3.1b that contains a negative regression coefficient for the lowest model levels of both T and ψ. As also indicated by Figure 3.1, the smaller coefficients and therefore a weaker balance operator response occurs at matching levels of T and ψ. Figure 3.2b, a vertical cross

section through the observation location, shows the larger multivariate correlations and larger response higher in the troposphere. The ψ response to the T observation exhibits thermal wind balance, with an increase in ψ and therefore the heights above the increase in T.

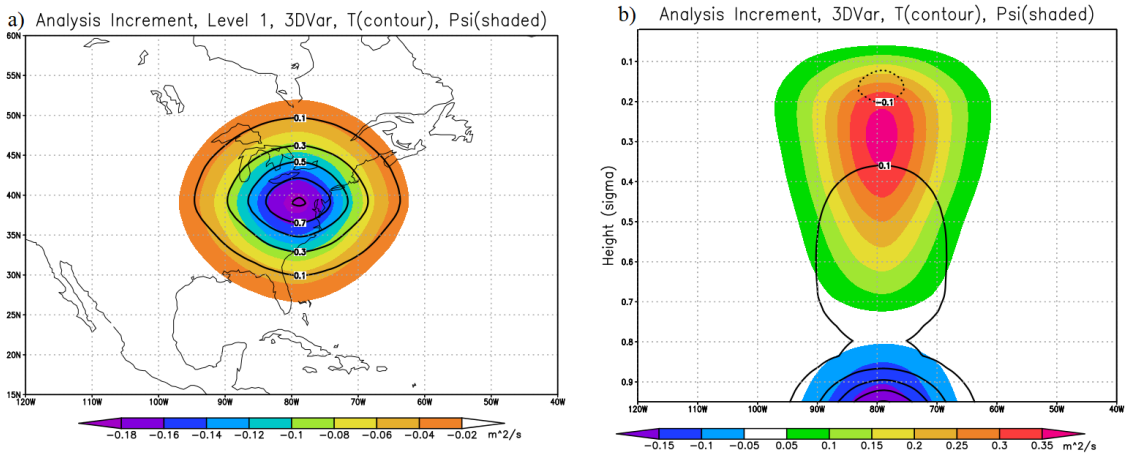

Figure 3.2: Analysis increments from the assimilation of a single T observation using 3DVar: (a) lowest model level and (b) vertical cross section at 39°N. T is contoured with 0.2 K interval and ψ is shaded.

To isolate the balance operator effect on the ensembles, the next set of tests use 100% ensemble contribution, $(\beta^f, \beta^e) = (0, 1)$. Figure 3.3a shows the analysis increments for T (contour) and ψ (shaded) at the level of the observation location using the standard EnVar configuration (CTL). This figure shows the standard case where the increment away from the observation is derived from the localized ensemble perturbations, demonstrating flow dependence. δT is roughly isotropic, but $\delta\psi$ exhibits a dipole about δT, increasing the height gradient and therefore the meridional wind at that location.

As in the static 3DVar case, the adjustment to both $\delta\psi$ and δT by $\mathbf{\Gamma}$ in the ensemble case should also be small at the level of the observation and larger at

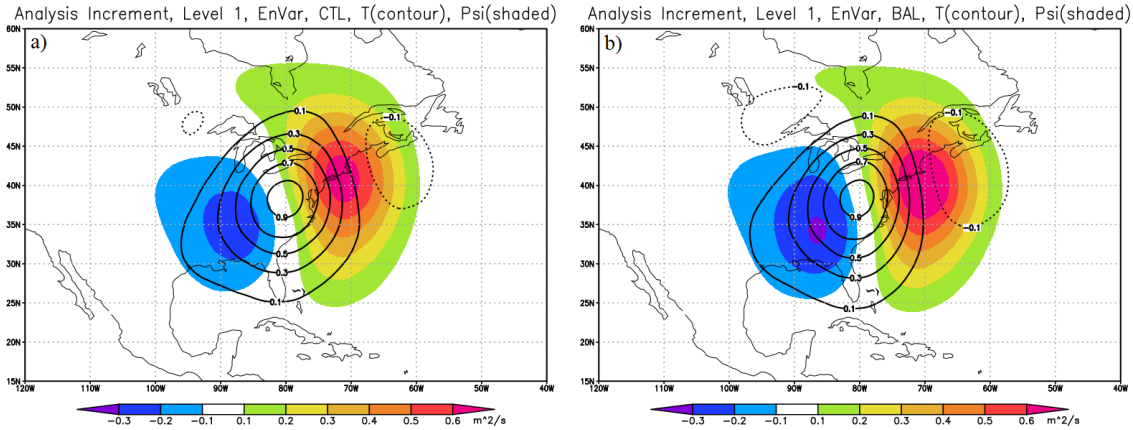

Figure 3.3: Analysis increments from the assimilation of a single T observation at the lowest model level using 4DEnVar: (a) CTL and (b) BAL. T is contoured with 0.2 K interval and ψ is shaded.

adjacent levels. Figure 3.3b shows the analysis increment for BAL at the level of the observation. As expected, the differences between BAL and CTL are small. There is a slight reduction (increase) in δT in areas with a positive (negative) $\delta\psi$ increment due to the negative value of $\mathbf{G}$ at this level. If the analysis increment is decomposed into its balanced and unbalanced parts (Figure 3.4), this adjustment is depicted as the balanced part of δT.

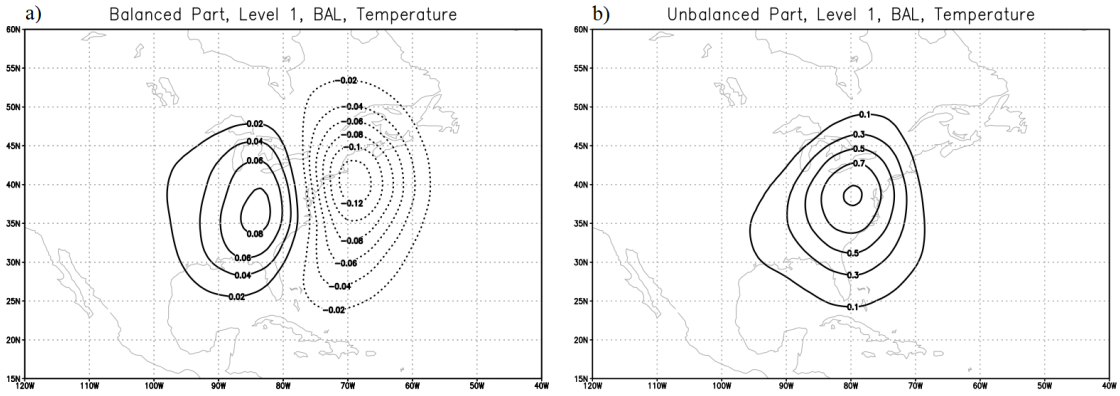

Figure 3.4: Decomposition of δT into its (a) balanced and (b) unbalanced components. T is contoured with 0.02 K interval for (a) and 0.2 K interval for (b).

Due to $\mathbf{G}$ having larger values at adjacent levels, the adjustment by $\mathbf{\Gamma}$ should

be larger in these areas as well. Figure 3.5 shows a latitudinal cross section of the analysis increments through the observation location. The localization of vertical ensemble correlations keeps the analysis increments confined to the observing level in CTL (Figure 3.5a). When $\mathbf{\Gamma}$ is added to the ensemble correlations in BAL, there is a much larger difference above the level of observation (Figure 3.5b). Information from the bottom level is now propagated vertically outside of the localization radius through $\mathbf{\Gamma}$. These correlations in the vertical exhibit thermal wind balance: there is an increase in $\delta\psi$, and therefore the heights, above the positive δT at the lowest model level. Also, above negative (positive) $\delta\psi$ in the lowest model level, there is now an increase (decrease) in δT throughout the column. The adjustment in the upper levels is qualitatively similar to the static case (Figure 3.2b), which was solely determined by the balance operator, though the effects of the flow dependence in the lower level ensemble covariances is apparent by the slight asymmetry of the vertical response. With a modification to both $\delta\psi$ and δT above the observing level, there is a two-way propagation of information between the variables; δT impacts $\delta\psi$ and $\delta\psi$ also impacts δT.

The configuration within SPEEDY uses strict vertical localization due to the recursive filter handling the coarse vertical resolution poorly. This allows the BAL configuration to show a maximum effect since $\mathbf{\Gamma}$ is the only means of propagating observation information vertically. In reality, NWP models rarely use vertical localization as strict as this system. In practice, this likely would reduce the impact of the BAL configuration if the vertical localization length scale was large. Figure 3.5c shows the single observation response when no vertical localization is used within

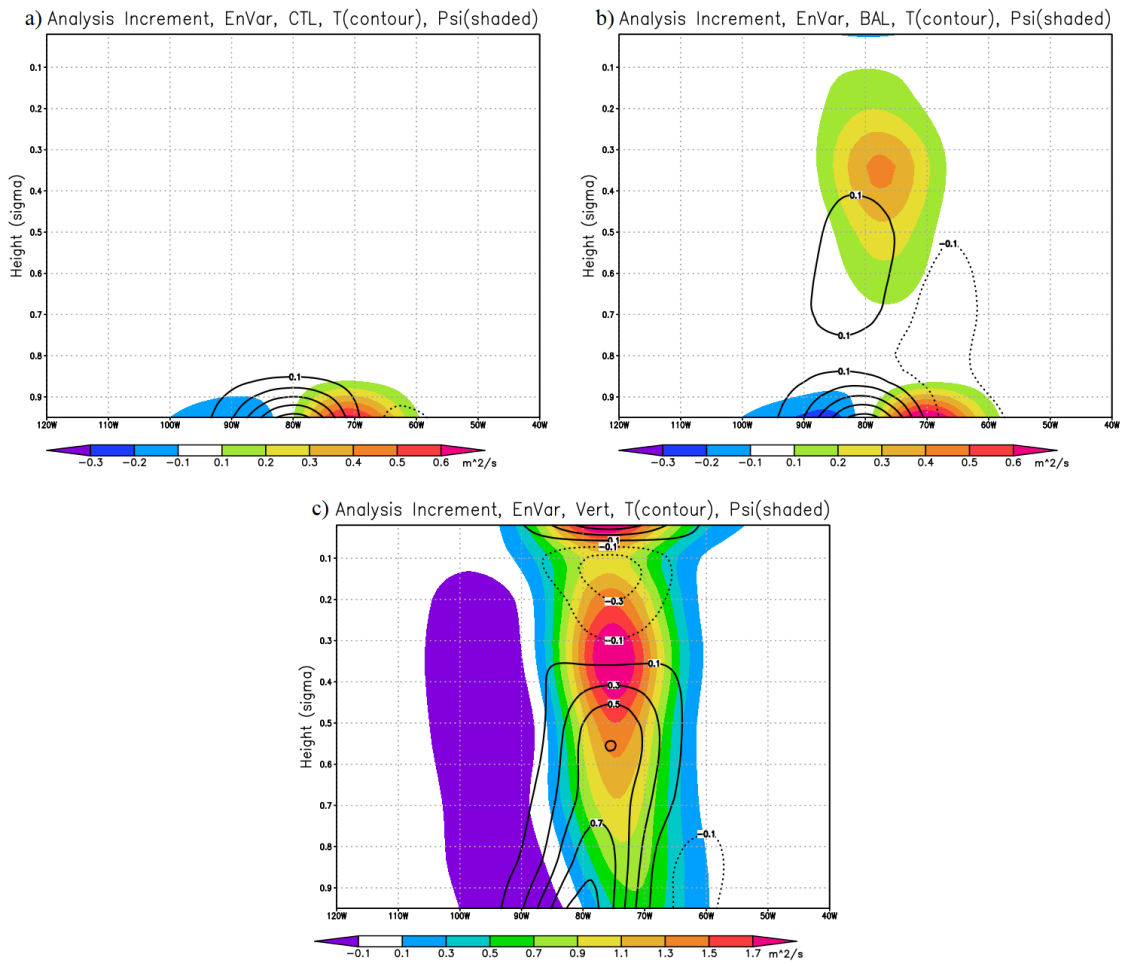

Figure 3.5: Vertical cross section of the analysis increments taken at 39°N when assimilating a single T observation at the lowest model level for (a) CTL, (b) BAL, and (c) CTL with no vertical localization. T is contoured with 0.2 K interval and ψ is shaded.

the EnVar. Only CTL is shown, as CTL and BAL with no vertical localization are nearly identical. While higher in magnitude, the response is qualitatively similar to BAL, exhibiting thermal wind balance. In this case, the thermal wind relationship is derived from the ensembles rather than from climatology, demonstrating that the lack of vertical information in CTL is due to the localization only and not a deficiency in the ensemble covariances. Figures 3.5b and c demonstrate the extremes of vertical localization and the associated BAL impact. In reality, vertical localization within ensemble data assimilation schemes falls in between and it is expected that

the relative impact of $\mathbf{\Gamma}$ would as well.

In addition to the single observation test, a single cycle analysis is performed using the full suite of observations (Figure 2.14) and the same background. Figure 3.6 shows the CTL and BAL T analyses at $\sigma = 0.835$ for the same background at 1982/06/01 00z as well as the difference between them. Comparing Figures 3.6a and 3.6b, several features are common, such as the negative increments in the Bering Sea and south of Hawaii, the dipole south of New Zealand, and the large increment on the Antarctic Peninsula. The magnitude of the BAL increments are larger than the CTL increments, which follows as a natural consequence of Figure 3.5b. In CTL, each observation primarily impacts the levels immediately adjacent to the observation since the ensemble correlations are fully localized in the vertical. Only the 10% static part can have an impact throughout the column in CTL. For BAL, the ensemble can also propagate information outside of the localization radius. This results in more information affecting each level, creating increments with larger magnitude.

Examining Figure 3.6c, the primary differences between these configurations occur in the southern hemisphere and over the ocean. These areas have few radiosonde observations compared to the northern hemisphere and over land. While there is satellite coverage in the southern hemisphere, the simulated satellites only observe T and q or lowest level winds. In contrast, the radiosondes observe all 3D variables at all levels, with the exception of upper level q. When a region is well observed, $\mathbf{\Gamma}$ would not typically need to make large adjustments since the observations, if they are close to the truth, should bring the analysis most of the way towards

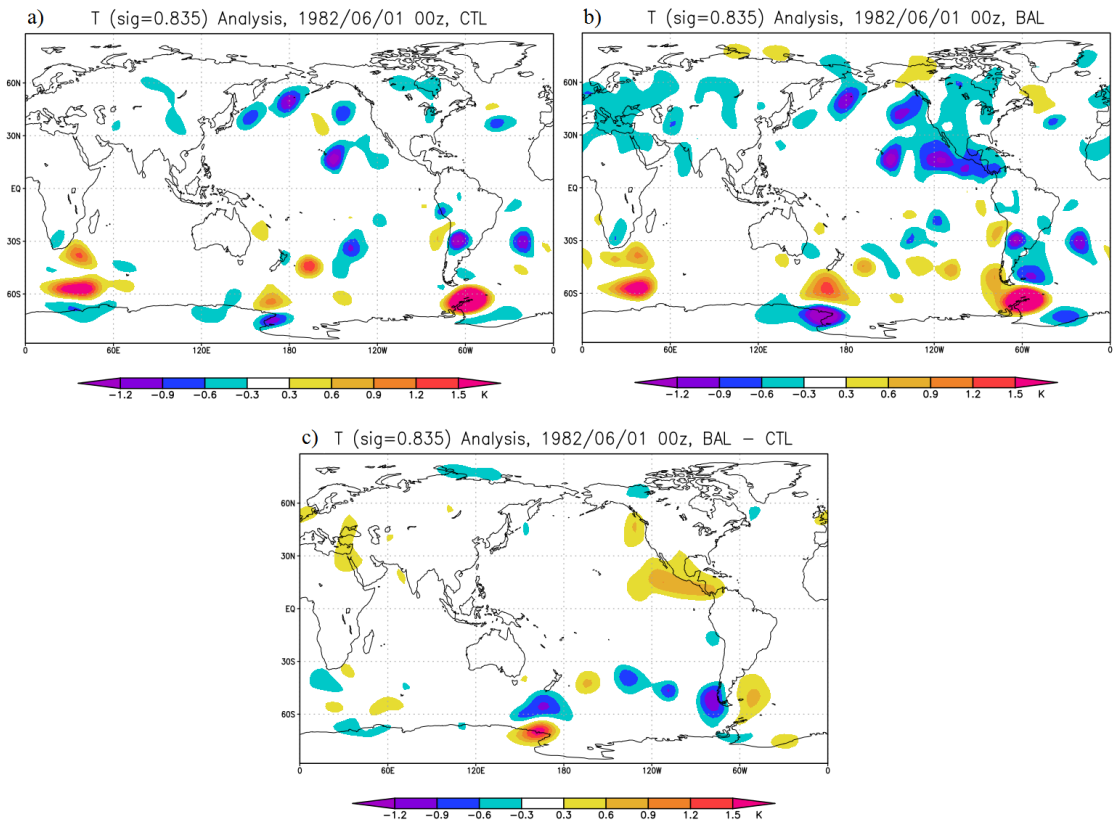

Figure 3.6: Single analyses for (a) CTL and (b) BAL at 1982/06/01 00z using the same background. (c) is the difference between (b) and (a). Shown for T at the second model level.

a balanced state. In a region that is not well observed, $\mathbf{\Gamma}$ would be expected to make larger corrections; if only one of the variables is observed, the other variables would need to be adjusted throughout the column to be brought into balance. This results in larger differences between BAL and CTL in regions with few radiosonde observations.

The single observation impact test demonstrates how $\mathbf{\Gamma}$ propagates the ensemble covariance information outside of the localization radius, with small adjustments at the level of the observation and larger adjustments at adjacent levels. The single analysis impact test shows that the propagation of information can lead to larger

magnitude increments in BAL than CTL. The uneven distribution of observations also leads to larger differences between BAL and CTL in regions in which all observations are not regularly observed.

3.3.2 Full Experiment Results

The full suite of simulated observations was assimilated for two parallel experiments: CTL and BAL. Two years of cycling were completed using the configuration described in Section 2.3 and the first month was rejected due to spin up. The proposed advantage of adding $\mathbf{\Gamma}$ to the ensemble is the improvement of balance. When imbalances exist within the system, gravity waves are produced, which degrade the forecast. A signature of this in the forecast fields would be an increase in surface pressure tendency. Figure 3.7a depicts the zonally averaged surface pressure tendency for CTL and BAL over the two year period, excluding the first month due to spin up, with the difference and 95% confidence intervals (Figure 3.7b). Little difference is seen in the northern hemisphere (Figure 3.7a), though there is some statistically significant degradation shown in Figure 3.7b in the areas surrounding the Himalayas. In the southern hemisphere, BAL exhibits a lower surface pressure tendency than CTL, where it was indicated previously that $\mathbf{\Gamma}$ was producing greater adjustments. This signifies that the inclusion of $\mathbf{\Gamma}$ in the ensemble perturbations has increased the balance in this region.

In many other initialization methods, to bring the state into balance, the analysis is moved away from the observations. In contrast, the method presented

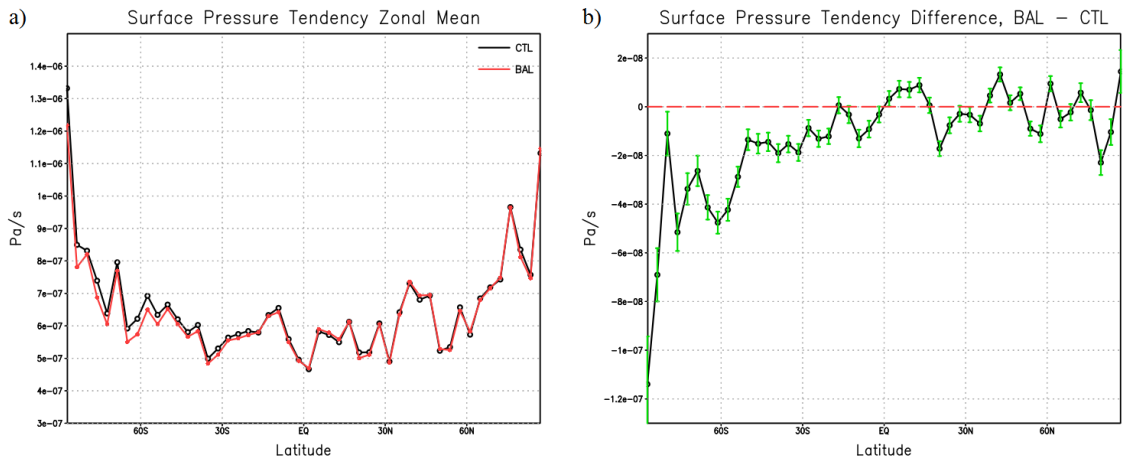

Figure 3.7: (a) Zonal mean surface pressure tendency for CTL and BAL and (b) the difference between the two experiments with 95% confidence intervals.

here attempts to correct for imbalance within the minimization itself, rather than after the assimilation. To assess the skill of the analysis, root-mean-square errors (RMSE) were calculated with respect to the high resolution true state. Figure 3.8 shows the results, calculated globally for ψ, χ, and T at each vertical level. The impact on q is neutral (not shown). This is expected since this particular formulation of $\mathbf{\Gamma}$ does not modify the moisture field; any improvement in humidity skill would be due to the improved wind field only. The largest improvement is seen in T. This is encouraging since T has the strongest relation to ψ. Tropospheric ψ and the lowest level χ show a marginally positive impact, also corresponding to regions that are affected by $\mathbf{\Gamma}$. There is a large decrease in skill in upper level ψ and to a smaller extent in χ. This is due to the model bias discussed in Section 2.1.3. The regression coefficients of (3.1a) - (3.1c) are derived using the low resolution model, which results in a statistical relationship between T and a damped stratosphere rather than the "true" stratosphere. When $\mathbf{\Gamma}$ is incorporated, it only exacerbates

the bias that already exists in this area.

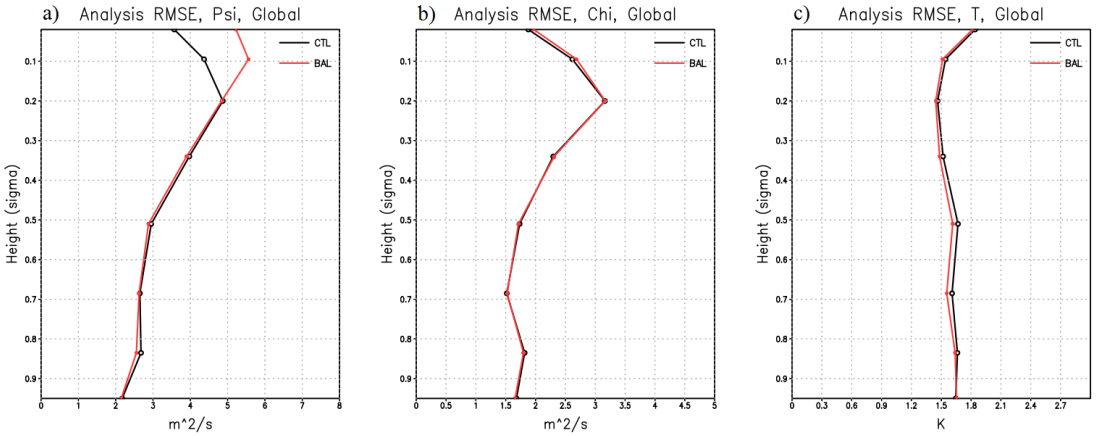

Figure 3.8: Analysis RMSE calculated globally with height for CTL (black) and BAL (red). Shown for variables (a) ψ, (b) χ, and (c) T.

When all of the variables are observed, $\mathbf{\Gamma}$ does not need to make as large of an adjustment as when only one variable is observed. When computing the analysis skill by hemisphere (Figure 3.9, shown for T only), $\mathbf{\Gamma}$ has a much larger positive impact in the southern hemisphere, where the observations are comprised mostly of satellite data, as discussed for the single cycle case. Figure 3.10a shows the zonally averaged RMS of difference in δT with height. The midlatitudes contain the largest differences, where geostrophic balance acts strongly, with a maximum in the midtroposphere. As in the RMSE of the analysis, the southern hemisphere shows greater differences between the increments of CTL and BAL.

Another conclusion of the single cycle case was that BAL has larger magnitude increments. To investigate whether this translates to the cycling system, rather than calculate the RMS of the increment difference, the difference of the RMS increment is calculated instead, shown in Figure 3.10b for T with latitude and height. There is no consideration of increment sign, just the size of the increment for each experiment.

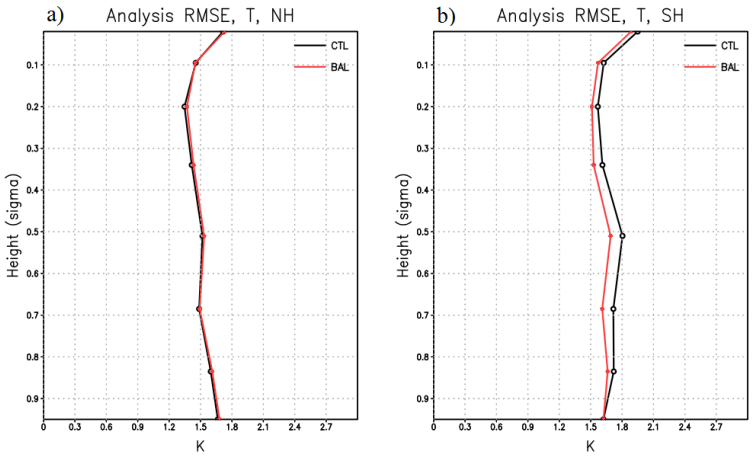

Figure 3.9: Temperature analysis RMSE for the CTL (black) and BAL (red) experiments over (a) the Northern and (b) Southern Hemispheres.

The size of the increment for BAL is larger than CTL for all latitudes and heights, with the largest increases occurring in the northern hemisphere midlatitudes. This is where the maximum amount of total observations is located. As stated in Section 3.3.1, $\mathbf{\Gamma}$ allows the ensemble covariances to impact levels away from the observation location, resulting in each level being affected by more observations and therefore producing larger increments. Both RMS difference figures (Figures 3.10a,b) have maximum differences in the midlatitude middle troposphere. The expected impact of $\mathbf{\Gamma}$ can be deduced by examining the structure of $\mathbf{G}$. Figure 3.10c shows the sum of the absolute value of $\mathbf{G}$, which represents the accumulated impact of $\mathbf{\Gamma}$ on δT summed over all of the levels of $\delta\psi$ (3.1b). In agreement with the RMS differences, the maximum accumulated impact is in the midlatitude middle troposphere.

An improvement in the balance of the initial conditions reduces the production of fast moving gravity waves, which should improve the long-term forecast skill. To assess the forecast skill in these experiments, anomaly correlation coefficients (AC)

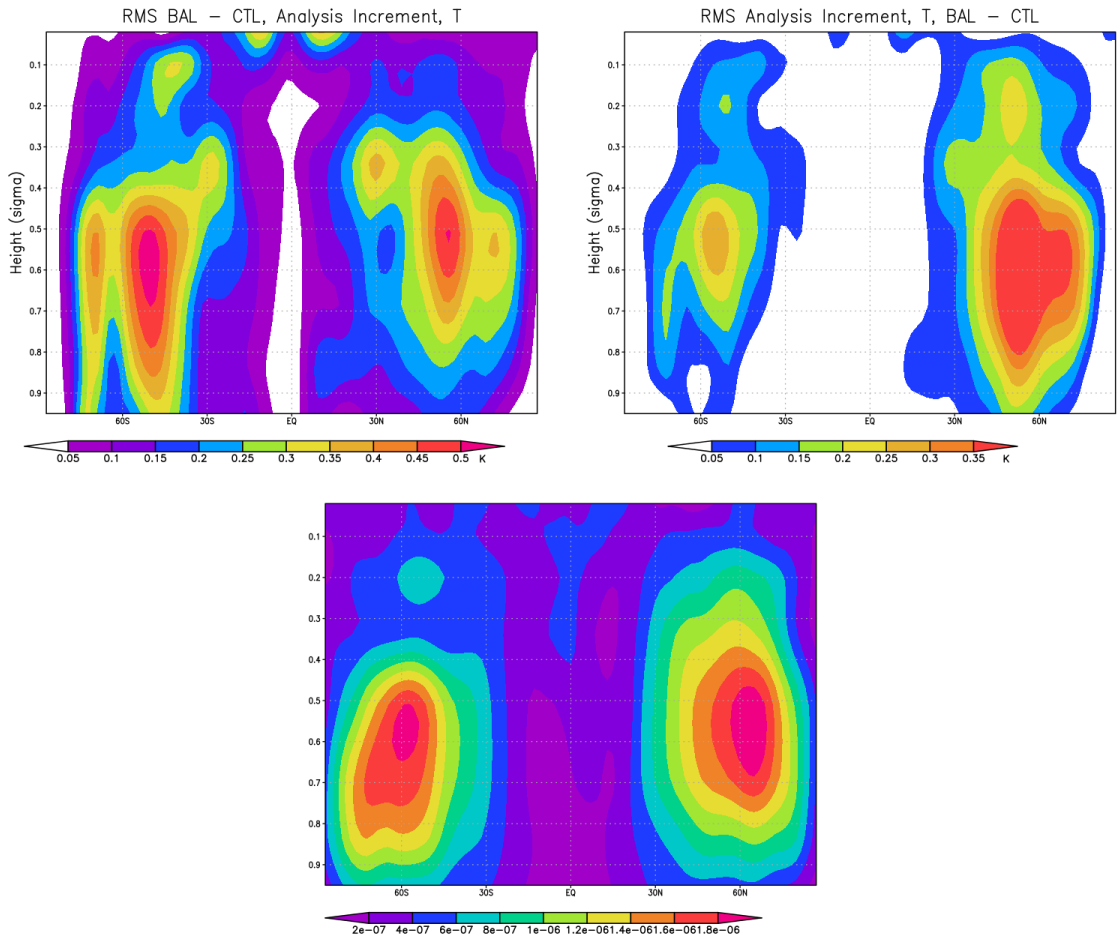

Figure 3.10: (a) RMS of the difference between the T analysis increments for the BAL and CTL configurations. (b) The difference between the RMS of the T analysis increments for the BAL and CTL configurations. (c) The sum of the absolute value of $\mathbf{G}$ over all vertical levels.

are computed in which 10 day forecasts initialized at 00z each day for the two year period are compared with the verifying truth. SPEEDY does not contain a diurnal cycle, so sampling at a single time each day is representative of the forecast behavior. Figure 3.11 shows the difference between the global AC for T and ψ with height for each experiment. The ψ AC show a similar degradation in the upper levels due to the bias in the stratosphere, though tropospheric ψ and all levels of T showed improvement for all forecast days calculated.

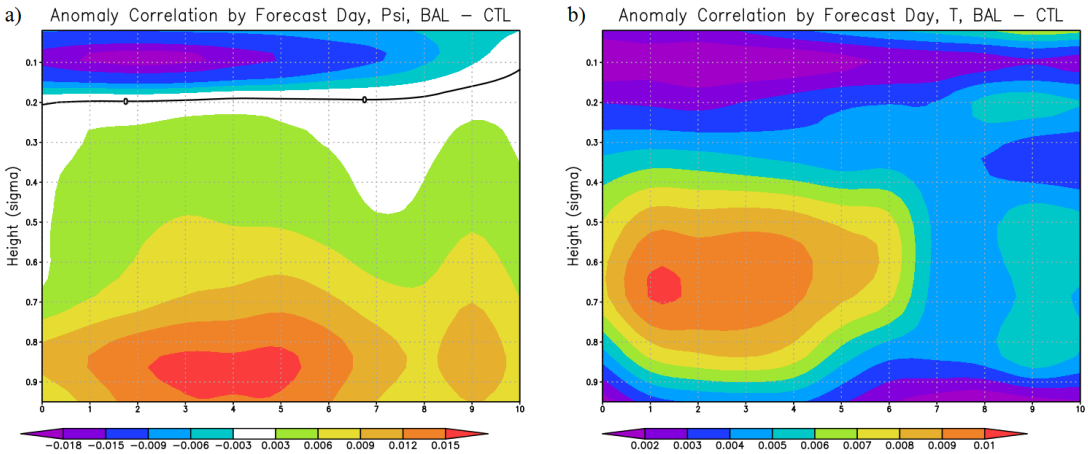

Figure 3.11: Difference in global AC between BAL and CTL by height and forecast day. Shown for (a) ψ and (b) T.

The statistical significance of these differences is shown in Figure 3.12, which contains the 95% confidence intervals for an upper and lower level of ψ and T. The forecast degradation for the upper level ψ is significant out to 8 days, but neutral by day 10. The T AC at the same level are improved over the 10 day forecast range. Though the improvement is small, it is statistically significant at all calculated forecast lengths. The lower levels show a highly significant improvement in the first five days for both T and ψ and remains significant through 10 days, though less so with time as would be expected. The impact of $\mathbf{\Gamma}$ on the forecast skill mirrors the impact on the analysis skill: degradation in the stratospheric wind fields where large model bias is present and significant improvement in T at all levels and the wind fields in the troposphere.

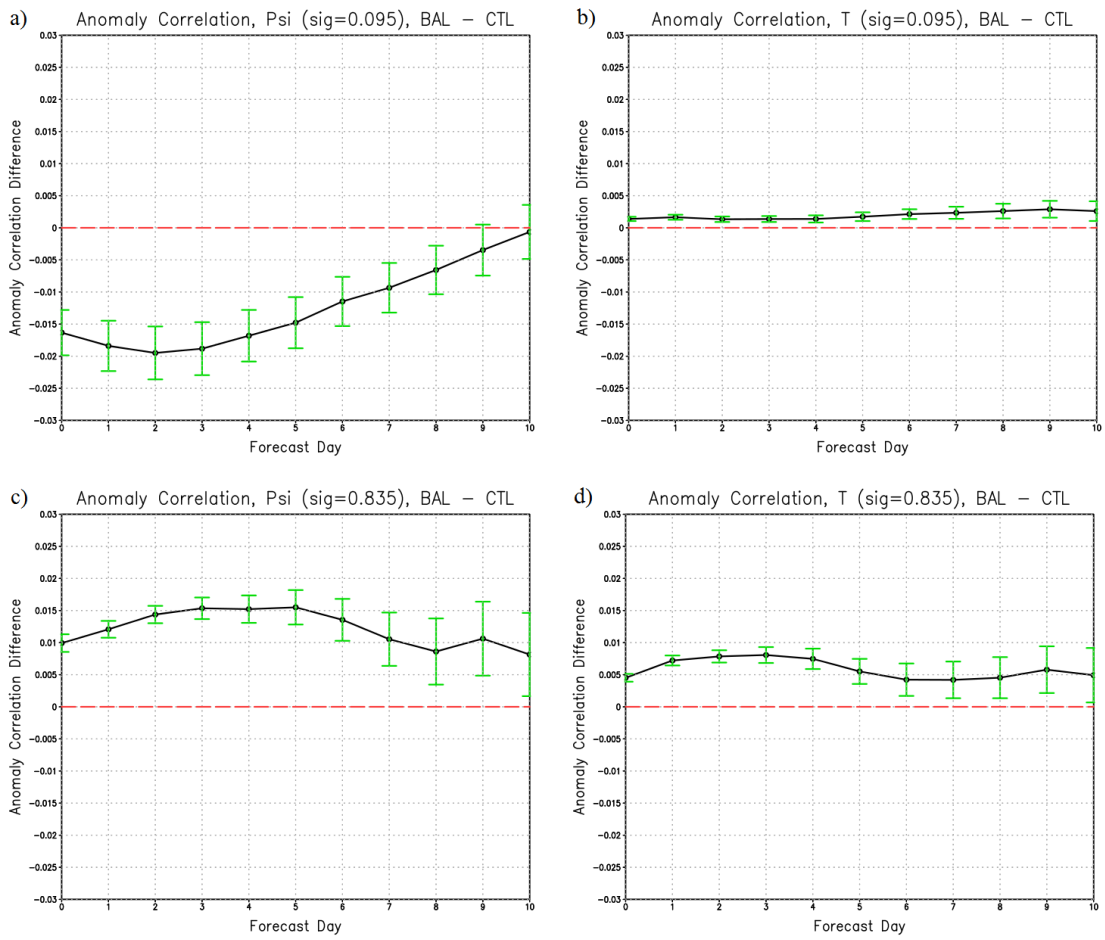

Figure 3.12: Difference in global AC between BAL and CTL with 95% confidence intervals for (left) ψ and (right) T at (top) an upper and (bottom) lower model level.

3.4 Summary and Discussion

By applying a balance operator, $\mathbf{\Gamma}$, to the ensemble portion of a hybrid 4DEnVar, the balance of the SPEEDY system is improved. Implemented within the intermediate complexity model, SPEEDY, this method allows the localization to act on the perturbations in the unbalanced space only and the balanced part remains unaffected. There is a reduction of the surface pressure tendency in the southern hemisphere where $\mathbf{\Gamma}$ has a larger impact. The forecast and analysis skill for T and

lower level ψ are improved, though the analysis and forecast skill of the stratospheric wind fields degrade due to the poor treatment of the zonal wind bias.

The added cost of including $\mathbf{\Gamma}$ for the ensemble perturbations is minimal. Rather than performing an extra balance operator step each time the increment is calculated, the preexisting balance operator step for the static part of the increment is merely being performed on the whole increment instead. The only added cost is the transformation of the ensemble perturbations to the unbalanced space, which only needs to be done once at the beginning of each analysis cycle. For the manner in which $\mathbf{\Gamma}$ is formulated here, it is highly parallelizable; it only works on the column, not on neighboring points. Each ensemble member can also be calculated independently.

SPEEDY is an intermediate complexity model and developed with climate applications in mind. It has a slower error growth than the state-of-the-art models have, so conclusions of improvements out to 10 days are not directly applicable to the more complex models in terms of forecast length. While many of the complexities of the major models are absent in SPEEDY, the key balances of hydrostatic and geostrophic balance are present and well represented. The configuration of these fraternal twin experiments also allowed for the clean evaluation of a known model bias. BAL did not perform well in these areas, which may draw concern for the application to a more realistic setting since the main NWP models have many biases, though this bias between resolutions was quite large (20 m s^{-1} in some locations). The potential advantages of localizing the perturbations in the unbalanced space could outweigh adverse impacts from model bias, particularly at such a low cost in

the computation of the analysis.

In the future, this method will be tested within NCEP's Gridpoint Statistical Interpolation (GSI, Kleist et al. (2009b)) for use in the GFS. The hybrid and balance operator formulations used in these experiments were chosen to mimic operations as closely as possible for greater ease of transition. One notable difference is that these experiments use an LETKF to create the ensemble perturbations compared to the EnSRF used operationally (Whitaker et al., 2008), but since the mean state estimation of the ensemble system is discarded and the ensemble is recentered about the hybrid estimation, the difference in ensemble method should have a minimal impact.

The GSI employs another method to increase balance, the tangent-linear normal-mode constraint (TLNMC, Kleist et al. (2009a)). Using 3DVar, Kleist et al. (2009a) found that the TLNMC compensates for the deficiencies within the prescription of the balance operator coefficients. On the other hand, in the attempt to fully replace the balance operator with the TLNMC, the balance operator proved to be of benefit. The experiments of this chapter use a hybrid 4DEnVar data assimilation scheme rather than 3DVar. With the TLNMCs application to the whole increment within the hybrid 4DEnVar formulation of the GSI rather than just the static, 3DVar-like portion, it is unknown what the interaction of the TLNMC would be with an ensemble application of a balance operator. The ensemble portion of the increment has 4-dimensional time information, which may not require the TLNMC as greatly as the static portion, if at all.

Both the ensemble balance operator and the TLNMC would have a similar ef-

fect outside of the vertical localization radius to increase balance, possibly reducing the impact of the BAL configuration in the presence of the TLNMC. Since the two operators function in a comparable manner, it would be desirable if the computationally inexpensive BAL could replace the costly TLNMC. However, BAL uses the same balance operator coefficients as the 3DVar, which the TLNMC provided an adjustment for within Kleist et al. (2009a). It is probable that the TLNMC would also adjust the increments of BAL, increasing the balance of the analysis and providing additional benefit. The interaction of the TLNMC and the ensemble balance operator will be explored in the future.

Chapter 4: Balance Operators in Ensemble Data Assimilation: Localization

4.1 Introduction

Ensemble methods of data assimilation use statistics calculated from a large number of model forecasts to approximate the background error. In numerical weather prediction (NWP) applications, the ensemble size is typically much smaller than the dimension of the problem due to computational cost, $\mathcal{O}(100)$ ensembles for an $\mathcal{O}(10^9)$ model, resulting in a large amount of sampling error. This can cause spurious correlations between two variables that are physically unrelated. For grid points that are close in distance, the correlations are likely dominated by signal, but for grid points that are at a large distance from one another where the true correlation is small, sampling error likely dominates (Hamill et al., 2001). Utilizing this assumption, the method of localization was devised to remove these long distance correlations, which are likely noisy and often degrade the analysis.

A common form of localization operates on the background error covariance, referred to as $\mathbf{B}$ localization. Detailed in Houtekamer and Mitchell (2001), this method calculates the Schur product of the background covariance and a correla-

tion function. The most commonly used correlation function is that of Gaspari and Cohn (1999), which ranges from one to zero as the distance increases, retaining the covariances for grid points that are physically close to one another and eliminating covariances for points that are distant. This method is regularly used in the perturbed observation form of the ensemble Kalman filter (Burgers et al., 1998; Houtekamer and Mitchell, 1998) and ensemble-variational hybrids (Buehner, 2005; Lorenc, 2003).

Localization is also frequently applied to the observation error covariance, referred to as $\mathbf{R}$ localization. A simplified version of this is implemented through observation selection where each grid point assimilates only a subset of local observations. Houtekamer and Mitchell (1998) noted that excluding remote observations from the calculation of a particular grid point improved the overall analysis. However, concerned with abrupt cutoffs between areas of observation influence, a more sophisticated version of $\mathbf{R}$ localization was implemented by Hunt et al. (2007) in the local ensemble transform Kalman filter (LETKF), where the method of observational selection was paired with a modification of the observation error covariance. In this method, the observation error was multiplied by a function that increases with distance from the observation to the analyzed grid point. This results in observations that are far away from a grid point having extremely large errors associated with them and subsequently having minimal impact on the analysis. The LETKF employs this method of localization rather than $\mathbf{B}$ localization, since the background error is never explicitly calculated within that data assimilation scheme. While these localization methods differ in their implementations, they achieve the same

goal. Through the Kalman gain, the background and observation error covariances work in opposing manners to determine the influence of each observation; decreasing the background error and increasing the observation error have the same qualitative impact. Greybush et al. (2011) highlight one notable difference: the optimal length scale for $\mathbf{R}$ localization schemes is shorter than the optimal length scale for an equivalent $\mathbf{B}$ localization scheme.

While localization greatly reduces undesirable noise in the ensemble correlations, it disrupts geostrophically and hydrostatically balanced correlations (Buehner, 2005; Cohn et al., 1998; Greybush et al., 2011; Kepert, 2009; Lorenc, 2003; Mitchell et al., 2002). By reducing analysis increments to zero at a certain distance, balances that are dependent on gradients or column quantities are disrupted. In Chapter 3, a method of preserving balance in localized ensembles was investigated. In a hybrid 4DEnVar formulation, the balance operator is applied to the full increment rather than the static portion only. This results in the localization being applied to the perturbations in the unbalanced variable space, thereby retaining the balanced part of the correlation. The impact is especially evident for temperature and streamfunction correlations, where rather than having a vertically localized increment, there is a balanced increment throughout the column. That formulation was implemented within an intermediate complexity global atmospheric model and observing system simulation experiments were performed. The change in the application of the balance operator reduced the surface pressure tendency, indicating improved balance, and forecast and analysis skill were generally increased. Regions that contained high amounts of model bias between the high resolution truth and the low reso-

lution forecast experienced analysis and forecast degradation due to the balance regression coefficients being based on the low resolution model.

The hybrid 4DEnVar of Chapter 3 utilizes $\mathbf{B}$ localization. The purpose of this chapter is to apply the balance operator methodology to an EnKF that uses $\mathbf{R}$ localization, LETKF, and explore the strengths and weaknesses in their application. The methods for both EnVar and LETKF are described in Section 4.2 as well as how the different forms of localization impact the effect of the balance operator. Results in a single observation setting and a full observing network are presented for LETKF in Section 4.3 and a summary and conclusions are discussed in Section 4.4. The contents of this chapter are contained in Thomas and Ide (2017b).

4.2 Balance Operators and Localization

Traditionally used in variational schemes, balance operators represent the multivariate correlations present in the model state. These correlations are based on known physical relationships such as geostrophic and hydrostatic balance. Following Wu et al. (2002), the variables of velocity potential χ, temperature T, and surface pressure P are broken down into two components: a "balanced" component that is correlated with streamfunction ψ and an "unbalanced" component, which becomes the analysis control variable (1) in Chapter 3. The matrix $\mathbf{\Gamma} \in \mathbb{R}^{N\times N}$, referred to as the balance operator, transforms ensemble perturbations from the unbalanced variable space, represented by $\mathbf{Z} = (\psi^T, (\chi^u)^T, (T^u)^T, q^T, (P^u)^T)^T \in \mathbb{R}^{N\times M}$, to the total variable space, represented by $\mathbf{X} = (\psi^T, \chi^T, T^T, q^T, P^T)^T \in \mathbb{R}^{N\times M}$, where N is

the dimension of the model, M is the number of ensemble members, and both sets of perturbations are normalized by $\sqrt{M-1}$. Application of $\mathbf{\Gamma}$ to $\delta\mathbf{z}$ incorporates the balanced component, $\mathbf{X} = \mathbf{\Gamma Z}$, where application of $\mathbf{\Gamma}^{-1}$ to $\mathbf{X}$ removes the balanced component. Application of $\mathbf{\Gamma}^T$ modifies ψ based on χ, T, and P. The regression coefficients with which the correlations are based are climatological, but model derived, i.e., they are static in time but consistent with the balance recognized by the model.

The balance operator is applied within the ensemble systems to separate the balance and unbalanced parts of the perturbations. Doing so ensures that the balanced portion of the analysis increment is not affected by spatial localization, as it is in conventional ensemble methods. The balance operator's interaction with localization is represented schematically for four cases in Figure 4.1. Figure 4.1a shows a depiction of the background error covariances for ψ and T without any localization, $\mathbf{XX}^T$. Correlations from the ensemble are depicted in red, with the primary feature being a strong spatial correlation along the diagonal. Because of inadequate ensemble size, sampling error exists in the calculation of the covariances and noise appears away from the diagonal, resulting in distant, unphysical correlations. The effect of the balance operator is shown in yellow, highlighting how ψ and T can impact each other at great distances, particularly in the vertical. Areas that are colored orange represent multivariate correlations that have contributions from the balance operator in addition to the ensemble forecasts.

Figure 4.1b depicts the conventional case when spatial localization based on the physical distance between grid points is applied to the background covariance,

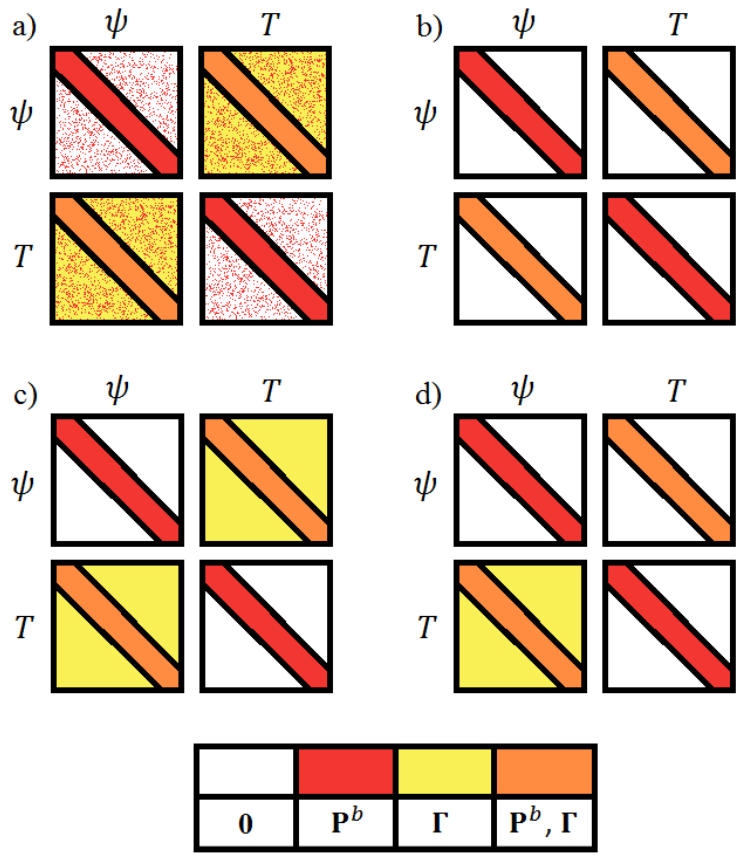

Figure 4.1: Illustration of a subset of the background covariance for variables ψ and T where red represents ensemble-derived correlations, yellow represents $\mathbf{\Gamma}$ derived correlations and orange represent correlations from both sources. a) No localization, b) conventional spatial localization, c) $\mathbf{\Gamma}$ in EnVar, d) $\mathbf{\Gamma}$ in LETKF

$\rho \circ (\mathbf{X}\mathbf{X}^T)$ where $\rho \in \mathbb{R}^{N \times N}$ represents the localization function and $\circ$ is the element-wise Schur product. Since $\mathbf{X} = \mathbf{\Gamma}\mathbf{Z}$, the impact of the balance operator is also localized. With the application of the spatial localization, this case only retains the correlations, whether univariate or multivariate, near the diagonal of each block. By reducing all of the correlations to zero beyond a certain distance, relationships that rely on gradients or vertically integrated quantities are disrupted. This has significant implications for correlations that dictate the large scale balance.

4.2.1 Application within a $\mathbf{B}$ localization method: EnVar

Chapter 3 demonstrates the benefit of the balance operator when applied to the ensemble portion of a Hybrid 4DEnVar in an intermediate complexity model. In that case, the balance operator allows for a propagation of the balanced multivariate correlations outside of the spatial localization radius. In this section, the formulation for its ensemble-only counterpart, EnVar, is presented.

In 3DVar, $\mathbf{\Gamma}$ is used to transform the control variable. In the EnVar, the control variable is the weight for each ensemble member ($\mathbf{v}^e \in \mathbb{R}^{QM}$ where Q is the number of grid points); therefore, $\mathbf{\Gamma}$ is not applied to the control variable but to the ensemble perturbations. The increment is written as:

$$\delta\mathbf{x} = \mathbf{\Gamma} \sum_{m=1}^{M} (\mathbf{F}\mathbf{v}_m^e \circ \mathbf{Z}_m^e) = \mathbf{\Gamma}\mathbf{E}_z\mathbf{U}\mathbf{v}^e. \tag{4.1}$$

The perturbations in the unbalanced space for each ensemble member, $\mathbf{Z}_m^e \in \mathbb{R}^N$, are created at the beginning of each analysis cycle by applying $\mathbf{\Gamma}^{-1}$ to the total model perturbations for each member, m. The matrix representation of the Schur product of the ensemble perturbations is $\mathbf{E}_z \in \mathbb{R}^{N\times QM}$. The control vector is preconditioned on the square root of its error covariance matrix, $\mathbf{L} = \mathbf{U}\mathbf{U}^T \in \mathbb{R}^{QM\times QM}$, which is a block diagonal matrix that provides the spatial correlation of the ensemble weights and localizes the perturbations. For this implementation, each block of $\mathbf{L}$ contains both a forward and backward recursive filter, $\mathbf{F} \in \mathbb{R}^{Q\times Q}$ and $\mathbf{F}^T$ respectively. The recursive filter, described in Purser et al. (2003), applies covariance localization in

the unbalanced space rather than model space, ensuring that the balanced covariances remain unaltered. Once the increment is found in the unbalanced variable space, $\mathbf{\Gamma}$ is applied to include the balanced variables once again to obtain the full increment.

The analytical solution for the ensemble weight is written as:

$$\mathbf{v}^e = (\mathbf{P}^a)^{-1}\,\mathbf{U}^T(\mathbf{E}_z)^T\mathbf{\Gamma}^T\mathbf{H}^T\mathbf{R}^{-1}\mathbf{d}, \tag{4.2}$$

where

$$\mathbf{P}^a = \mathbf{I} + \mathbf{U}^T(\mathbf{E}_z)^T\mathbf{\Gamma}^T\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\mathbf{\Gamma}\mathbf{E}_z\mathbf{U}, \tag{4.3}$$

is the analysis covariance, $\mathbf{H} \in \mathbb{R}^{L\times N}$ is the observation operator, L is the number of observations, $\mathbf{R} \in \mathbb{R}^{L\times L}$ is the observation error covariance and $\mathbf{d} \in \mathbb{R}^L$ is the innovation. (4.2) shows that when calculating $\mathbf{v}^e$, $\mathbf{\Gamma}^T$ is applied before the localization ($\mathbf{U}^T$). The localization matrix has no knowledge of where the observations are located, so it does not constrain the increment to be close to the observations. It propagates any information that is present spatially, whether the information is from the observations or the balance operator. Once $\mathbf{v}^e$ is found, the increment is calculated as in (4.1), where $\mathbf{\Gamma}$ is applied to the increment last. This permits the application of both $\mathbf{\Gamma}$ and $\mathbf{\Gamma}^T$ to be felt, allowing for a two-way propagation of information. Information from ψ is passed to the other variables and information from the other variables is also passed to ψ.

Figure 4.1c depicts this case, where the localization is applied to the pertur-

bations in the unbalanced space only, $\mathbf{\Gamma}(\rho \circ \mathbf{Z}\mathbf{Z}^T)\mathbf{\Gamma}^T$. The localization eliminates the noise for points that are distant for both the self-correlations and the cross-correlations, but the cross-correlations still have an impact from $\mathbf{\Gamma}$, including for points that are distant. The balanced correlations are able to be propagated outside of the radius of localization.

4.2.2 Application within an $\mathbf{R}$ localization method: LETKF

The method of localizing the covariances for the unbalanced variables only will now be applied to a different ensemble method, one utilizing observation error localization: the local ensemble transform Kalman filter (LETKF, Hunt et al., 2007). In this data assimilation scheme, the analysis is calculated locally at each grid point, only considering the observations that fall within a certain radius from that point. The observation error is also calculated locally with a distant-dependent localization function, ρ_O, applied to it, so that observations that are closer to the grid point in question have a smaller observation error associated with it. Ensemble weights are calculated locally and applied to the background perturbations to create the analysis mean and spread.

To make use of the perturbations in the unbalanced space, the background ensemble needs to be transformed to the unbalanced variables as in the EnVar: by applying $\mathbf{\Gamma}^{-1}$ globally to each ensemble member. The analysis weights are then calculated locally as in the standard LETKF. These calculations require the ensemble perturbations in observation space, which is equivalent for the full and unbalanced

perturbations since the unbalanced perturbations have to be transformed back to the full perturbations in order to apply the observation operator. This results in the analysis weights of each method also being equivalent. What differentiates this method from the standard LETKF is that the ensemble weights are locally applied to the ensemble mean and spread in the unbalanced space after they are found, producing an analysis ensemble in the unbalanced space:

$$\bar{\mathbf{z}}^a = \bar{\mathbf{z}}^b + \mathbf{Z}^b \bar{\mathbf{w}}_z^a, \tag{4.4a}$$

$$\mathbf{Z}^a = \mathbf{Z}^b \mathbf{W}_z^a, \tag{4.4b}$$

where $\bar{\mathbf{w}}_z^a \in \mathbb{R}^M$ is the analysis weights calculated using the perturbations in the unbalanced space and $\mathbf{W}_z^a \in \mathbb{R}^{M \times M}$ is the square root of the analysis covariance in ensemble space used to transform the background perturbations to the analysis perturbations:

$$\begin{aligned} \tilde{\mathbf{P}}^a &= \mathbf{W}_z^a \left(\mathbf{W}_z^a\right)^T \\ &= \mathbf{I} + \mathbf{Z}^T \boldsymbol{\Gamma}^T \mathbf{H}^T (\rho_O \circ \mathbf{R})^{-1} \mathbf{H} \boldsymbol{\Gamma} \mathbf{Z}, \end{aligned} \tag{4.5}$$

where $\rho_O \in \mathbb{R}^{L \times L}$ is the $\mathbf{R}$ localization function. It is an exponential function that increase with the distance from the observation to the analysis grid point, thereby increasing the observation error and reducing the impact of observations that are far away from the grid point being analyzed.

The analysis weights are calculated at each grid point using the following equation:

$$\bar{\mathbf{w}}_z^a = \left(\tilde{\mathbf{P}}^a\right)^{-1} \mathbf{Z}^T \boldsymbol{\Gamma}^T \mathbf{H}^T (\rho_O \circ \mathbf{R})^{-1} \mathbf{d}. \tag{4.6}$$

Once the weight is applied to the background perturbations, the balance operator is applied globally to each ensemble member, transforming from the unbalanced space back to the full variable space. This allows for a propagation of information from the anchor variable, in this case ψ, to the other variables and moves the ensemble towards a balanced state.

Acknowledging the equivalence between $\mathbf{E}_z$ and $\mathbf{Z}$, the similarities between the two ensemble formulations are apparent. The major difference between this equation and (4.2) is the spatial localization. The EnVar applies the localization on the background error, while the LETKF applies the localization on the observation error. When the analysis weight is calculated in the LETKF, ρ_O is applied to the observation error, which reduces the impacts of observations that are at a large distance from the grid point being analyzed. When an observation is sufficiently far, the inverse of $\rho_O \circ \mathbf{R}$ is zero. The application of the transpose of the balance operator in the weight calculation ($\mathbf{\Gamma}^T\mathbf{H}^T(\rho_O \circ \mathbf{R})^{-1}\mathbf{d}$) cannot force the weight to become nonzero. The local implementation of the LETKF equations further enforces this; when an observation is fully outside the radius of observation selection, there is no information that the balance operator can attempt to propagate during the computation of the analysis weights. This presents a major disadvantage when applying the balance operator within an observation space localization scheme. When the full analysis ensemble is calculated from the analysis ensemble in the unbalanced space, $\mathbf{\Gamma}$ is applied to the unbalanced variables which produces a one-way propagation of information from ψ to the other variables. However, the other variables have no avenue to impact ψ; there is not a two-way passing of information. Figure 4.1d

depicts the LETKF case, where $\mathbf{\Gamma}$ can only propagate information one way due to the $\mathbf{R}$ localization. The equivalent schematic covariance is written as $\mathbf{\Gamma}\left(\rho \circ \mathbf{Z}\mathbf{Z}^T\mathbf{\Gamma}^T\right)$. The noise of the distant correlations is eliminated as in cases (b) and (c), but distant balance correlations of case (c) only appear on one side of the matrix, resulting in a covariance matrix that is not symmetric.

The background covariances in the unbalanced space are transformed to the full variable space when used in the analysis weight calculation in order for the observation operator to be applied ($\mathbf{H\Gamma Z}$). This results in the analysis weights being equivalent for the unbalanced and the full variables ($\bar{\mathbf{w}}_z^a = \bar{\mathbf{w}}_x^a$). The analysis ensemble is brought into balance when the balance operator is applied to the analysis ensemble in the unbalanced space at the end of the cycle. The LETKF calculates its preferred analysis based on the observations and background ensemble and then the balance operator shifts χ, T, and P away from their analyzed states to be brought into balance with the analyzed ψ, unlike the EnVar that takes balance into account during the minimization and adjusts all of the variables in the analysis. While the LETKF adjustment is not a preferred means of bringing the analysis into balance, there is no unique balanced state for each unbalanced state (Daley, 1991).

If there was no localization applied within either of the data assimilation schemes presented, the application of the balance operator would have no impact. The formulations with and without the balance operator would be equivalent, emphasizing the importance of the details of the localization implementation.

4.3 Results

4.3.1 Single Observation Tests

To demonstrate the impact of the balance operator in the LETKF, a single observation impact test is performed. To showcase the effect of $\mathbf{\Gamma}$, two tests were conducted. A single temperature observation was assimilated for both tests, with 1 K innovation and observation error. The first case (Figures 4.2 and 4.3) assimilates an observation at the lowest model level. The second case (Figure 4.4) assimilates an observation in the upper troposphere.

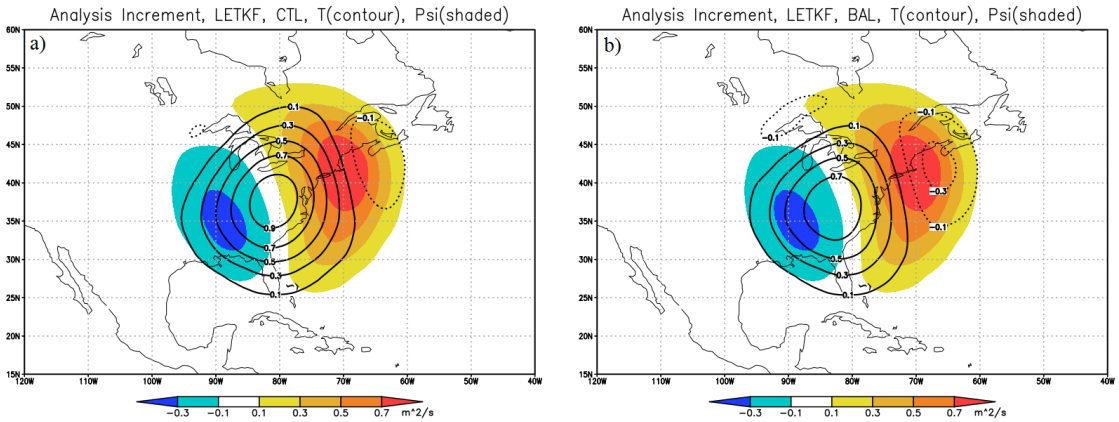

Figure 4.2: Analysis increment at the lowest model level with the assimilation of a single T observation at that level: (a) CTL and (b) BAL. T is contoured with 0.2 K interval and ψ is shaded.

Figure 4.2a shows the analysis increment for LETKF CTL at the lowest model level where a single T observation is assimilated. δT (contoured) is mostly isotropic and $\delta\psi$ (shaded) exhibits a dipole about δT, which is similar to EnVar CTL (Figure 3.3a). The T-ψ cross-covariances are defined solely by the ensemble perturbations in both LETKF and EnVar CTL. As expected, both CTL's behave similarly. However,

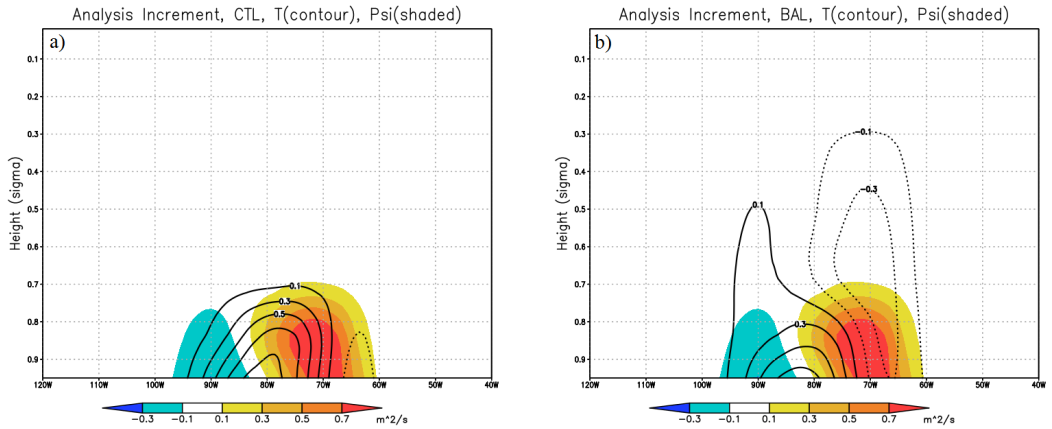

Figure 4.3: Analysis increment at 39°N with height for the assimilation of a single temperature observation at the lowest model level: (a) CTL and (b) BAL. T is contoured with 0.2 K interval and ψ is shaded.

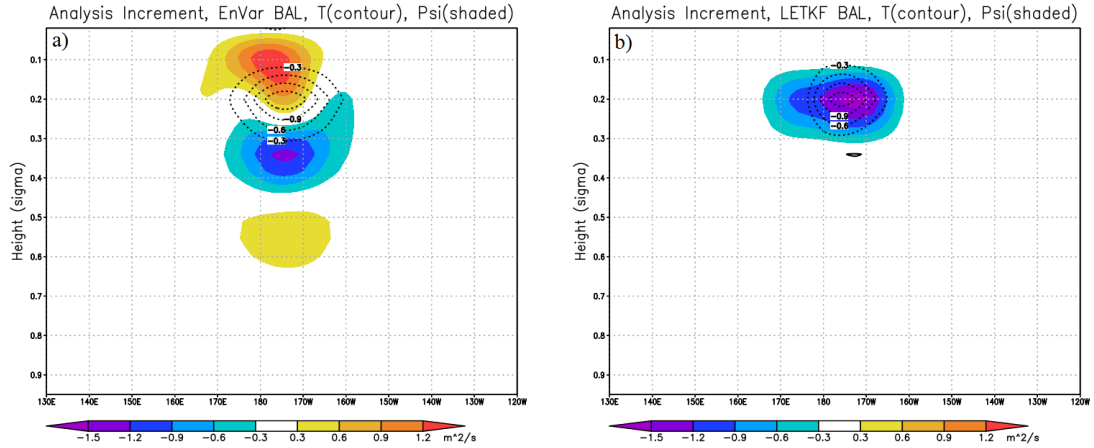

Figure 4.4: Analysis increment at 30°S with height for the assimilation of a single temperature observation at $\sigma = 0.2$: (a) EnVar BAL and (b) LETKF BAL. T is contoured with 0.3 K interval and ψ is shaded.

the BAL's exhibit much different behavior. The EnVar showed a slight adjustment to both $\delta\psi$ and δT. For LETKF BAL (Figure 4.2b), there is no change to $\delta\psi$ compared to CTL (4.6) but there is an adjustment to δT, a larger adjustment than the EnVar. There is a reduction (increase) in δT in the broad area of positive (negative) $\delta\psi$.

A vertical cross section through the observation location shows how ψ and T adjust hydrostatically. By comparing a vertical cross section of the LETKF CTL

increments (Figure 4.3a) with the EnVar CTL increments (Figure 3.5), the difference in the vertical localization is apparent. LETKF CTL and BAL has more relaxed vertical localization than EnVar CTL and BAL, allowing both δT and $\delta\psi$ to extend to the second model level, rather than being constrained to the bottom level only. This difference contributes to the larger adjustment seen in the Figure 4.2b. The regression coefficients between ψ and T in $\mathbf{\Gamma}$ are at a minimum at corresponding levels with the largest coefficient occurring at adjacent levels. With no increment above the bottom model level, the EnVar did not have a large adjustment at the level of the observation. For the LETKF, the increments in CTL extend above the bottom model level. When the balance operator is applied, the second model level's $\delta\psi$ is able to impact the lowest level δT, rather than just from $\delta\psi$ at the same level.

Figure 4.3b shows the vertical cross section for LETKF BAL. Similar to the increment at the level of the observation, $\delta\psi$ remains unchanged; this time it is evident throughout the column. This provides a stark difference from the EnVar BAL vertical cross section, where δT at a level hydrostatically leads to $\delta\psi$ throughout the depth of the troposphere. Instead of a slight adjustment to both $\delta\psi$ and δT, LETKF BAL shows a moderate adjustment to δT only, with cooling above regions with positive $\delta\psi$ and warming above regions with negative $\delta\psi$.

As mentioned in Section 4.2.2, the balance operator within the LETKF only imposes a one-way adjustment to bring the state into balance rather than the two-way adjustment of the EnVar. The lack of two-way communication has major implications for the assimilation system. In the EnVar BAL, there was a degradation in the stratospheric ψ compared to CTL (Figure 3.8a). This was due to the balance

operator being derived from a model that had a damped stratosphere; so when T is observed, it produces a vertical profile of $\delta\psi$ consistent with the damped model rather than the true model. In Figure 4.4a, a single T observation is assimilated in the upper troposphere with EnVar BAL. The impact on $\delta\psi$ extends through the troposphere and up into the stratosphere. Figure 4.4b shows the analysis increment when LETKF BAL is used instead. $\delta\psi$ is equivalent to the increment that was computed from CTL; there is no adjustment due to the balance operator. This should result in different stratospheric performance between the EnVar and LETKF, which is explored further in Section 4.3.2.

4.3.2 Full Observation Network

By construction, the addition of the balance constraint in the LETKF has a one-way impact: ψ alters the other variables, but remain unchanged itself (4.6). To evaluate whether the one-way effect is enough to have a positive impact on the forecast, two experiments with the full observing system were run: a standard 4D-LETKF as described in Section 2.3 (CTL) and a second experiment with the balance operator incorporated (BAL).

The global root-mean-square errors (RMSE) of the analysis is shown in Figure 4.5. There is a slight negative impact in the BAL analysis skill of ψ and χ (Figure 4.5a,b), though this is due to the cycling only and not the ψ and χ analyses themselves. The largest analysis impact is for T (Figure 4.5c), which is negative for the troposphere. As discussed in Section 4.2.2, full column T is adjusted after

the analysis weights are calculated, moving T away from what the LETKF calculates to be the optimal analysis in order to be brought it into balance with ψ. The improvement in the top level T is significant, but confined to the southern polar region, where BAL has a better representation of the Antarctic circumpolar vortex (not shown). Because q is not a part of the $\mathbf{\Gamma}$ formulation, it has no measurable change in its analysis skill (not shown) as expected. Analysis skill would only be impacted through a change in the flow and ψ and χ, as described in Section 4.2.2, do not change for individual analyses with BAL except for χ at the bottom model level.

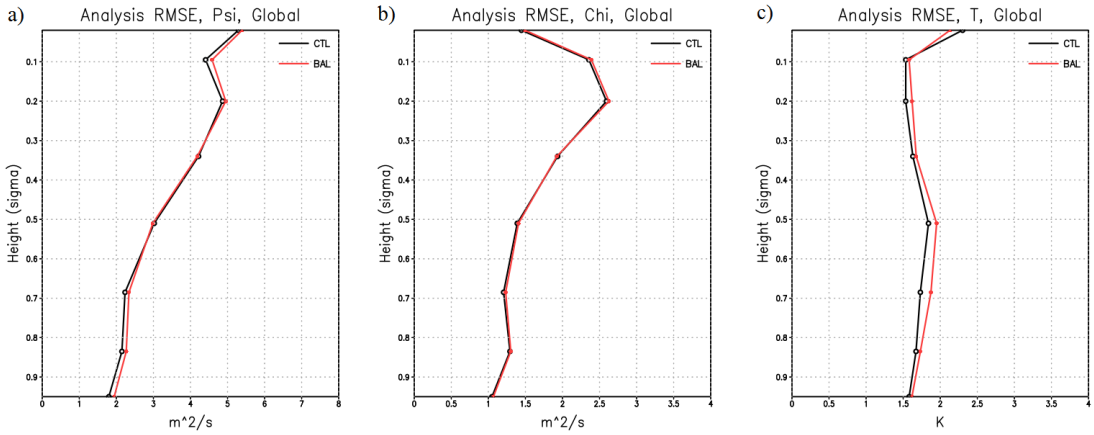

Figure 4.5: Global analysis RMSE with height for CTL (black) and BAL (red) for variables (a) ψ, (b) χ, and (c) T.

The balance operator has a larger impact in the southern hemisphere on the T RMSE (Figure 4.6), as in the EnVar (Figure 3.9). A notable difference between the EnVar and the LETKF is that the latter has a larger negative impact as opposed to a larger positive impact in the former. For the EnVar, it was conjectured that the larger impact in the southern hemisphere was due to the lack of observations (Section 3.3.1). The majority of observations in the southern hemisphere are T retrievals.

Due to the EnVar's two-way interaction, $\boldsymbol{\Gamma}$ is able to make larger adjustments since not all variables are observed. The LETKF's one-way interaction does not allow $\boldsymbol{\Gamma}$ to make adjustments based on T observations. The difference in the T analyses between the LETKF BAL and CTL are based solely on the magnitude of $\delta\psi$. Since the southern hemisphere is less observed and less skillful, the increments are larger than the northern hemisphere increments on average (Figure 4.7). This results in a larger adjustment to the temperature in the LETKF BAL in the southern hemisphere than in the northern hemisphere.

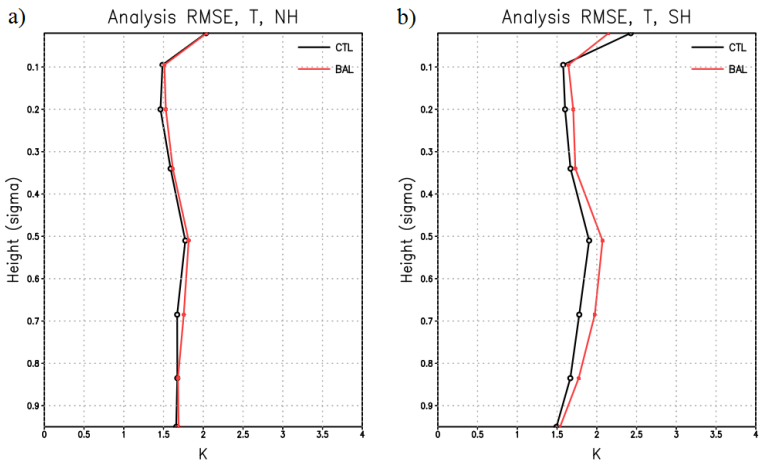

Figure 4.6: As in the right panel of Figure 4.5, for (a) the northern hemisphere and (b) the southern hemisphere.

In the EnVar, the main region that experienced a degradation by including the balance operator in the ensemble was the stratosphere. This was due to the large model bias between the T63 and T30 resolutions. The lower resolution was used to calculate the regression coefficients and therefore when the balance operator was included, it specified a relationship between the troposphere and a damped stratosphere. This resulted in significant degradation in skill of upper level ψ (Figure

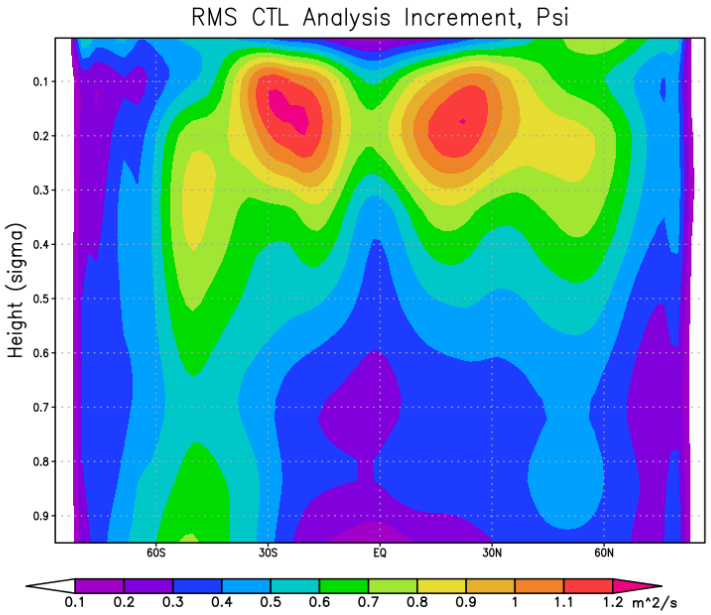

Figure 4.7: Zonally averaged RMS analysis increment for LETKF CTL with height for variable ψ.

3.8). LETKF does not have such a response: ψ shows only a slight degradation in the analysis RMSE. This is expected due to the effect of the localization on the balance operator. As shown in the single observation case in Section 4.3.1, $\mathbf{\Gamma}$ is not able to communicate between the stratosphere and the troposphere in LETKF BAL due to the $\mathbf{R}$ localization. This makes it unable to propagate the incorrect $T - \psi$ relationship either. In addition, the adjustments from $\mathbf{\Gamma}$ in LETKF BAL are primarily to T rather than ψ due to the one-way adjustment, reducing the ability of the improper regression coefficients to negatively impact ψ.

One of the motivations to improve balance in the analysis is to reduce the production of gravity waves which degrade the long term forecast (e.g. Baer and Tribbia, 1977). Adjustments to increase balance, unfortunately, often degrade the short term forecast since the previously calculated analysis is altered and moved towards a more balanced state (e.g. Williamson et al., 1981). To assess the forecast

skill in these experiments, the global anomaly correlation coefficients (AC) for the BAL and CTL experiments are examined. With little change to the analysis for ψ, χ, and q, the AC do not show significant differences between experiments either, as expected (not shown). Also in line with the analysis results (Figure 4.5c), T is the variable with the largest difference in the AC (Figure 4.8). Early in the forecast period, the detriment of the analysis adjustment is evident in the degraded AC. Figure 4.9a demonstrates that this is significant out to five days, shown for the bottom model level. At longer forecast times, however, there is an improvement in the AC for BAL over CTL in the troposphere. This demonstrates the benefit of balanced initial conditions for long term forecasts, with the improvements being marginally significant globally. When calculating the AC for the southern hemisphere only where $\mathbf{\Gamma}$ is acting more strongly (Figure 4.9b), the late forecast improvement is significant to at least the 95% confidence level, indicating that the improvement seen is likely due to the increased balance in the initial conditions.

As a measure of balance, the global surface pressure tendency was reduced significantly for EnVar BAL (Figure 3.8). For the LETKF, however, there is no significant difference in the surface pressure tendency between CTL and BAL (not shown). This further suggests that the one-way adjustment towards a balanced state is insufficient for NWP.

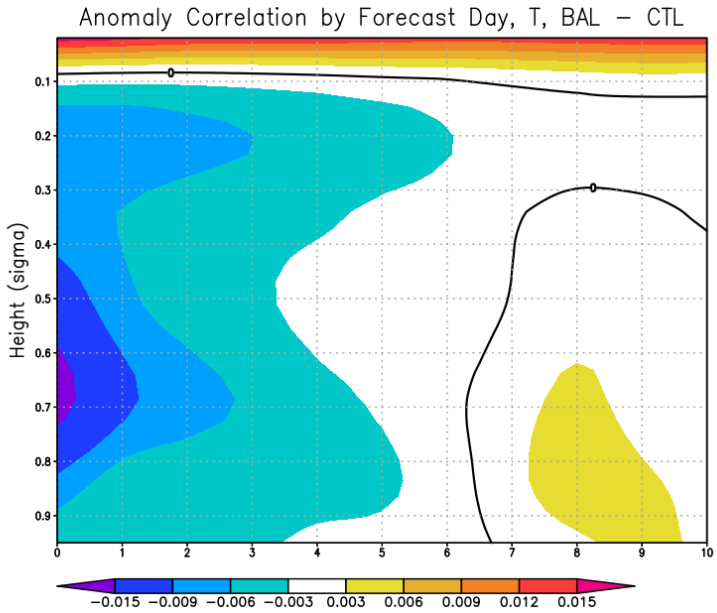

Figure 4.8: Global T anomaly correlation coefficient difference between BAL and CTL by forecast day with height.

4.4 Summary and Conclusions

In Chapter 3, a balance operator was implemented in the ensemble portion of a Hybrid 4DEnVar and showed significant improvements to the analysis and forecast skill. This chapter compares that implementation with one in a 4D-LETKF. These two ensemble data assimilation methods employ different forms of localization: the EnVar localizes on the background error and the LETKF localizes on the observation error. This chapter demonstrates that the observation error localization does not allow for a nonzero ensemble weight outside of the localization radius. Therefore, the only balanced information that gets transferred is from the anchor variable to the unbalanced variables and not vice versa. This results in the unbalanced variables being pushed away from the originally calculated analysis rather than all of the variables adjusting towards a balanced analysis.

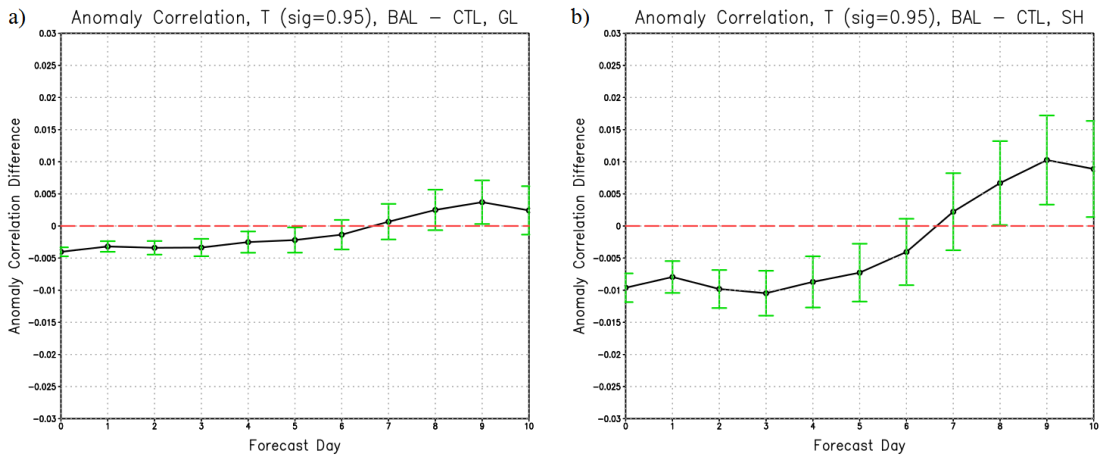

Figure 4.9: Lowest level T anomaly correlation coefficient difference between BAL and CTL (a) globally and (b) for the southern hemisphere.

The same set up as the Chapter 3 experiments is employed to allow for direct comparison. Using an intermediate complexity model, SPEEDY, a balance operator was included within a 4D-LETKF. There was a degradation in the analysis skill, particularly in T, which is consistent with the previously mentioned adjustment to the unbalanced variables. The forecast skill impact is neutral for most variables; however, the forecast skill for T is negative for short forecast lead times, but trends positive with forecast length, eventually becoming significantly positive by days 8 through 10. The effect is greater in regions where the balance operator made a larger impact, indicating that the increased balance in the T analysis leads to the increased forecast skill later in the time period. As with Chapter 3, caution must be taken when applying the results from these intermediate complexity model results with that of a state-of-the-art model. Developed as a climate model, SPEEDY's error growth is much slower than that of an NWP model, so conclusions about forecast length must be adjusted.

The method of the balance operator application is computationally efficient in terms of computational costs and is roughly equivalent to its application in the EnVar system; the ensemble members must be transformed to the unbalanced space at the beginning of each cycle, but the analysis weights are calculated as in the control case. While EnVar does not require the additional step of transforming each ensemble member into the full variables as in the LETKF, this task can be done in parallel. Even though the cost of this method is not great, the results of the method are not convincing enough to implement in a more complex system. The theoretical formulation of why the LETKF implementation should not be as effective as the EnVar implementation is supported by the underperforming results.

These results can be easily extended to other Ensemble Kalman filter algorithms. The Ensemble Square Root Filter (EnSRF; Whitaker and Hamill (2002)) is a commonly used EnKF and is operational at the National Centers for Environmental Prediction (Whitaker et al., 2008). The balance operator formulation is applied in a similar manner as in the LETKF: by computing the ensemble mean in the unbalanced variable space using the perturbations in the unbalanced space (see Appendix C for details). What sets it apart from the LETKF is its ability to have the spatial localization applied in either model space or observation space. When the spatial localization is applied in model space, the balance operator functions as it does in the EnVar: there is a two-way propagation of balanced information. However, since the full background error is localized, it is computationally expensive. More frequently, the observation space spatial localization is applied. This causes the balance operator in the EnSRF to function as in the LETKF: there is

only a one-way propagation of balanced information outside of the localization radius. While it is computationally less expensive, it will likely not have a positive impact on the forecast skill. Therefore, it is not recommended to include the balance operator within an EnSRF that is localized in observation space.

Chapter 5: Variable Localization in Ensemble Data Assimilation

5.1 Introduction

Ensemble data assimilation methods approximate the background error covariance using an ensemble of forecasts. The more ensemble members that are included in the estimation, the more accurately the sample covariance represents the true background error. Unfortunately, due to computational restraints, a limited number of ensembles must be used for numerical weather prediction (NWP) applications. For the state-of-the-art models, usually $\mathcal{O}$(10-100) ensembles are used to represent a 10^8 or greater system. The background covariance in this case contains a large amount of sampling error and is not a full rank matrix. These spurious correlations negatively impact the system, resulting in nonzero correlations for variables that are physically unrelated. Two grid points that are at a great distance from one another likely have a very small true correlation. For these points, sampling error likely dominates and contaminates the analysis (Hamill et al., 2001). These distant correlations are frequently eliminated through the method of localization.

Most commonly used in the spatial dimensions, localization can be applied to the background error covariance matrix or the observation error covariance matrix (Greybush et al., 2011). To localize on the background error covariance, it is

multiplied by a correlation function whose values range from one to zero as the distance between grid points increases, frequently the fifth order piecewise polynomial of (Gaspari and Cohn, 1999). This method, described in Houtekamer and Mitchell (2001), preserves the covariances for points that are in physical proximity and eliminates covariances for points that are distant. This requires the calculation of the full background error covariance, which is computationally prohibitive for NWP applications. Spatial localization can also be applied to the observation error covariance. A simple form of this is to only consider observations within a local region of an analysis grid point (Houtekamer and Mitchell, 1998). However, this could result in an abrupt cutoff of areas of observation influence. To curtail the negative impacts associated with these cutoffs, Hunt et al. (2007) proposed multiplying the observation error covariance matrix by a function that increases with distance in addition to the consideration of local observations, increasing the observation error and therefore reducing the impact of observations that are far from a grid point. Increasing the observation error and reducing the background error have a similar effect on the analysis increment through the Kalman gain. Greybush et al. (2011) compare these two formulations of spatial localization and found that their performance was comparable, though noting that the optimal length scale for localization on the background error is longer than the optimal length scale for localization on the observation error.

Typically, the length scales chosen for spatial localization are global in scope and applied to all variables equally. There is a great deal of evidence that this is far from optimal. Anderson (2007) notes that different localization functions are appro-

priate for different state variables. Chen and Oliver (2010) emphasize that the same localization function may not be suitable at all times. To address the lack of spatial homogeneity, Anderson and Lei (2013) developed an empirical localization function (ELF) that computes a non-Gaussian localization function between each observation type and state variable. Kang et al. (2011) also address the need for localization between the model variable types for certain applications. The authors implement a form of covariance localization that removes the spurious cross-covariances that arise between physically unrelated variables, which stabilized their carbon system. A strict form of variable localization was attempted by Clayton et al. (2013) in a Hybrid 4DVar scheme for global NWP by removing the cross-covariances between all variable types, but improvement was not seen in their case.

This chapter presents a unified view on the localization between variable types, called variable localization, presenting two different forms within three types of ensemble data assimilation schemes. Section 5.2 presents the two forms of variable localization qualitatively, while Section 5.3 describes their formulation within three ensemble data assimilation schemes. Section 5.4 presents a single observation case within one of the variable localization formulations. Discussion and conclusions are presented in Section 5.5. The contents of this chapter are contained in Thomas and Ide (2017c).

5.2 Variable Localization

In the earlier implementations of variational schemes, multivariate assimilation of observations required the construction of a balance operator or some other function to create cross-correlations between the analysis variables (Parrish and Derber, 1992). One considerable advantage of ensemble Kalman filters (EnKFs) is that the multivariate information can be naturally derived from the ensembles themselves rather than relying on a time invariant climatology. There are situations where portions of the correlation provided by the ensemble should be removed. For instance, grid points or observations that are located at a great distance from one another likely have small true correlations, which is difficult to represent in an ensemble whose size is much smaller than the dimension of the model. Spatial localization, a method commonly applied in EnKFs, removes these distant correlations.

Sampling error due to small ensemble size can also result in correlations between variable types that are not physically related. These multivariate correlations can be removed by a method called variable localization, which is less commonly applied within EnKFs. The most severe form of variable localization, removing all of the multivariate correlations, results in a univariate EnKF. This is typically undesirable as it loses the primary advantage that the ensemble methods have over the static covariance methods. However, there are situations that benefit from removing the correlation between certain variables.

Kang et al. (2011), hereafter referred to as K11, implemented a form of variable localization within a local ensemble transform Kalman filter (LETKF, Hunt et al.,

2007) to analyze carbon. Using an extension of the intermediate complexity model, SPEEDY (Molteni, 2003), carbon C and surface carbon fluxes CF were added to the standard set of prognostic variables of zonal wind u, meridional wind v, temperature T, specific humidity q, and surface pressure P. In previous applications of the SPEEDY system for data assimilation (Harlim and Hunt, 2007; Li et al., 2009; Miyoshi, 2005), synthetic observations of the meteorological variables were created and assimilated. For K11's implementation, C observations were also added to the system. While CF was not directly observed, the cross-correlations calculated by the ensembles were able to update CF in the analysis. The initial experiments performed poorly. By allowing for cross-covariances between all of the variables, the spurious correlations and poor initial conditions of CF negatively impacted the analysis of the meteorological variables; for example, q and CF should not be correlated. A form of variable localization was then implemented to eliminate the cross-correlations between variable types that do not have a direct physical relation.

When the true correlation between two distant grid points is small, sampling error dominates and degrades the analysis (Hamill et al., 2001). This line of reasoning also applies to different variables. When the true correlation between two variable types is small, sampling error dominates and the removal of these cross-correlations should improve the analysis. Variable localization refers to the set of methods that selectively remove cross-correlations between variables. Due to the differences in formulation among EnKFs, variable localization can be implemented in various ways. These methods can generally be divided into two categories: application in observation space or application in model space. Observation space

variable localization, which will be referred to as VO, removes the effect of the cross-correlations by allowing an observation to only impact certain control variables. Model space variable localization, which will be referred to as VM, removes the cross-correlations directly through the construction of the background error covariance matrix.

The VO form restricts an observation's impact and is straightforward to implement since it requires no change to the background error, which is not explicitly constructed in many NWP applications due to computational restraints. Only a subset of observations is used to calculate a subset of analysis variables, rather than all of the observations being used to calculate the analysis for all variables. For example, if T and u are to be uncorrelated, u observations are not assimilated when calculating the analysis for the variable T and vice versa.

The VM form of variable localization removes the correlations between model variables through the construction of the background error covariance. While this can be implemented as a Schur product between the background error and a correlation matrix as in the spatial localization, this is cost prohibitive within NWP applications since the background error covariance has dimensions of $N \times N$ where N is the dimension of the model. Instead, the background covariance matrix can be extended to create an additional ensemble, thus zeroing out the correlations between groups of variables. This application is advantageous compared to VO since it does not require knowledge of the observation types. There also does not need to be a direct correspondence between the model variable types and the observation types.

To illustrate the relationship between the two forms of variable localization,

consider the background ensemble perturbations, $\mathbf{X}^b \in \mathbb{R}^{N\times M}$, about the background ensemble mean, $\bar{\mathbf{x}}^b \in \mathbb{R}^N$, and normalized by $\sqrt{M-1}$ where M is the number of ensemble members. These perturbations can be partitioned into multiple groups of control variables. For two groups:

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \\ \mathbf{X}_2 \end{pmatrix}, \tag{5.1}$$

where $\mathbf{X}_1 \in \mathbb{R}^{N_1\times M}$ and $\mathbf{X}_2 \in \mathbb{R}^{N_2\times M}$. The two groups of control variables are mutually exclusive, i.e. $N = N_1 + N_2$. For the clarity of notation, the superscript b is dropped and the perturbations are assumed to refer to the background unless otherwise stated.

The background perturbations are frequently represented in observation space in EnKFs with $\mathbf{Y} = \mathbf{H}\mathbf{X} \in \mathbb{R}^{L\times M}$ where $\mathbf{H} \in \mathbb{R}^{L\times N}$ is the linearized observation operator transforming the perturbations from model space, $\mathbf{X}$, to observation space, $\mathbf{Y}$, and L is the number of observations. Similarly to $\mathbf{X}$, $\mathbf{H}$ can also be partitioned, though it can be partitioned along either dimension, creating submatrices of groups of control variables or observations:

$$\mathbf{H} = \begin{pmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} \\ \\ \mathbf{H}_{21} & \mathbf{H}_{22} \end{pmatrix} = \begin{pmatrix} \mathbf{H}_{M1} & \mathbf{H}_{M2} \end{pmatrix} = \begin{pmatrix} \mathbf{H}_{1O} \\ \\ \mathbf{H}_{2O} \end{pmatrix}, \tag{5.2}$$

where $\mathbf{H}_{Mj} = \begin{pmatrix} \mathbf{H}_{1j}^T & \mathbf{H}_{2j}^T \end{pmatrix}^T \in \mathbb{R}^{L\times N_j}$ is the observation operator partitioned in model space and $\mathbf{H}_{iO} = \begin{pmatrix} \mathbf{H}_{i1} & \mathbf{H}_{i2} \end{pmatrix} \in \mathbb{R}^{L_i\times N}$ is the observation operator partitioned in obser-

vation space. Unlike the groups of control variables, the groups of observations do not need to be mutually exclusive, i.e. $L_1 + L_2$ does not necessarily need to be equal to L and observation types can overlap between sets. The observation operator dictates which control variables each observation type projects on to. Partitioning $\mathbf{H}$ in observation space allows the separation of impact by observation type. Partitioning $\mathbf{H}$ in model space allows the separation of impact on particular control variables.

These two types of $\mathbf{H}$'s can also create two types of partitioned $\mathbf{Y}$'s. To partition $\mathbf{Y}$ in observation space, $\mathbf{H}$ in observation space is applied to the total background perturbations:

$$\mathbf{Y} = \begin{pmatrix} \mathbf{H}_{1O}\mathbf{X} \\ \\ \mathbf{H}_{2O}\mathbf{X} \end{pmatrix} = \begin{pmatrix} \mathbf{Y}_{1O} \\ \\ \mathbf{Y}_{2O} \end{pmatrix}, \tag{5.3}$$

where $\mathbf{Y}_{iO} \in \mathbb{R}^{L_i \times M}$. When $\mathbf{H}$ is partitioned in observation space, the observations are split into multiple groups. Therefore, each $\mathbf{Y}_{iO}$ is the background perturbations corresponding to the observations contained in the $\mathbf{H}_{iO}$ group only, which allows the control variables to be impacted by only certain observations.

To partition $\mathbf{Y}$ in model space, $\mathbf{H}$ in model space is applied to the background perturbations corresponding to that variable set:

$$\mathbf{Y} = (\mathbf{H}_{M1}\mathbf{X}_1 + \mathbf{H}_{M2}\mathbf{X}_2) = (\mathbf{Y}_{M1} + \mathbf{Y}_{M2}), \tag{5.4}$$

where $\mathbf{Y}_{Mj} \in \mathbb{R}^{L \times M}$. When $\mathbf{H}$ is partitioned in model space, the control variables are split into multiple groups. Correspondingly, each $\mathbf{Y}_{Mj}$ contains the projection from the observations onto certain control variables only. This permits only a portion of

the background perturbations to be considered, dictating which cross-correlations are included. (5.1) - (5.4) are formulated for two groups of control variables and observations, though more than two groups can be used.

In EnKF methods, the background error covariance matrix, $\mathbf{P}$, is approximated by the background perturbations ($\mathbf{P} = \mathbf{X}\mathbf{X}^T$). When the partitioning of control variables is included:

$$\mathbf{P} = \begin{pmatrix} \mathbf{X}_1\mathbf{X}_1^T & \mathbf{X}_1\mathbf{X}_2^T \\ \mathbf{X}_2\mathbf{X}_1^T & \mathbf{X}_2\mathbf{X}_2^T \end{pmatrix}. \tag{5.5}$$

The off-diagonal components of $\mathbf{P}$ represent the cross-correlations between the two groups of variables $\mathbf{X}_1$ and $\mathbf{X}_2$. Similarly, the background error can be represented in observation space as:

$$\mathbf{Y}\mathbf{Y}^T = \left(\mathbf{Y}_{M1}\mathbf{Y}_{M1}^T + \mathbf{Y}_{M2}\mathbf{Y}_{M1}^T + \mathbf{Y}_{M1}\mathbf{Y}_{M2}^T + \mathbf{Y}_{M2}\mathbf{Y}_{M2}^T\right). \tag{5.6}$$

VM removes the cross-correlation terms from (5.5), decorrelating the two variable groups:

$$\rho_{VM} \circ \mathbf{P} = \begin{pmatrix} \mathbf{X}_1\mathbf{X}_1^T & 0 \\ 0 & \mathbf{X}_2\mathbf{X}_2^T \end{pmatrix}, \tag{5.7}$$

where $\rho_{VM} \in \mathbb{R}^{N\times N}$ is a localization function comprised of blocks of zeros and ones, removing the cross-correlations between variable types from the background error directly. While intuitive, this form can be difficult to implement within NWP since the size of $\mathbf{P}$ is prohibitive to store the matrix explicitly.

An equivalent localization function can be constructed to localize the pertur-

bations in observation space. Rather than applying it to the background covariance directly, the localization function is applied to part of the Kalman gain, controlling the analysis increment. Shown for a single observation:

$$\begin{aligned} \rho_{VO} \circ \mathbf{X}\mathbf{Y}^T &= \rho_{VO} \circ \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \begin{pmatrix} \mathbf{X}_1^T & \mathbf{X}_2^T \end{pmatrix} \begin{pmatrix} \mathbf{H}_{M1}^T \\ \mathbf{H}_{M2}^T \end{pmatrix} \\ &= \rho_{VO} \circ \begin{pmatrix} \mathbf{X}_1\mathbf{X}_1^T\mathbf{H}_{M1}^T + \mathbf{X}_1\mathbf{X}_2^T\mathbf{H}_{M2}^T \\ \mathbf{X}_2\mathbf{X}_1^T\mathbf{H}_{M1}^T + \mathbf{X}_2\mathbf{X}_2^T\mathbf{H}_{M2}^T \end{pmatrix}, \end{aligned} \tag{5.8}$$

where $\rho_{VO} \in \mathbb{R}^N$ is a function that is also comprised of zeros and ones. For serial assimilation, a different ρ_{VO} needs to be constructed for each observation type that has the variable localization applied to it. When assimilating observations in batches, ρ_{VO} is applied to analysis increment itself, selecting which portion of the background to update. This form is more computationally efficient than (5.7), limiting the impact of the observations to certain control variables only. However, the cross-correlation terms will remain after the application of ρ_{VO} if an observation from one group impacts control variables from both groups.

An alternate method for localizing in model space can be formulated that does not require the full construction of $\mathbf{P}$. The control variable is split into two mutually exclusive groups as in 5.1. The perturbation matrix is then extended to separate

the perturbations from the two groups:

$$\tilde{\mathbf{X}} = \begin{pmatrix} \mathbf{X}_1 & 0 \\ 0 & \mathbf{X}_2 \end{pmatrix} \in \mathbb{R}^{N \times 2M}. \tag{5.9}$$

The tilde in this notation represents a matrix or vector that has been extended to include multiple groups of variables. Extending the perturbation matrix creates a second ensemble and removes the correlation between them. The background covariance constructed using this extended set of perturbations does not include the cross-correlation terms:

$$\mathbf{P} = \tilde{\mathbf{X}}\tilde{\mathbf{X}}^T = \begin{pmatrix} \mathbf{X}_1 & 0 \\ 0 & \mathbf{X}_2 \end{pmatrix} \begin{pmatrix} \mathbf{X}_1^T & 0 \\ 0 & \mathbf{X}_2^T \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1\mathbf{X}_1^T & 0 \\ 0 & \mathbf{X}_2\mathbf{X}_2^T \end{pmatrix}, \tag{5.10}$$

giving the same result as (5.7), but at a cheaper computational cost.

Combining the extended background covariance and the observation operator partitioned in model space, $\mathbf{Y}$ is extended as well:

$$\tilde{\mathbf{Y}} = \begin{pmatrix} \mathbf{Y}_{M1} & \mathbf{Y}_{M2} \end{pmatrix} = \begin{pmatrix} \mathbf{H}_{M1}\mathbf{X}_1 & \mathbf{H}_{M2}\mathbf{X}_2 \end{pmatrix} \in \mathbb{R}^{L \times 2M}. \tag{5.11}$$

Previously in (5.4), $\mathbf{Y}_{M1}$ and $\mathbf{Y}_{M2}$ combined into a single $\mathbf{Y}$ to represent the background perturbations projected onto an observation. In $\tilde{\mathbf{Y}}$, the contribution from each control variable group is separated. For example, if streamfunction ψ and velocity potential χ are control variables in each set and the zonal wind u is the observation being considered, $\mathbf{Y}_{M1}$ contains the rotational portion of the wind per-

turbations at the observation location while $\mathbf{Y}_{M2}$ contains the divergent portion. Forming the background error in observation space using $\tilde{\mathbf{Y}}$ also removes the cross-correlation terms from that matrix:

$$\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^T = \left(\mathbf{Y}_{M1}\mathbf{Y}_{M1}^T + \mathbf{Y}_{M2}\mathbf{Y}_{M2}^T\right). \tag{5.12}$$

The two types of partitioning for $\mathbf{H}$ and $\mathbf{Y}$ demonstrate how the two forms of variable localization, VO and VM, function. VO limits an observation's impact on control variables. The partitioning in observation space, $\mathbf{H}_{iO}$ and $\mathbf{Y}_{iO}$, only include L_i observations. When the analysis is calculated for a particular variable type, only a subset of observations will be contained in $\mathbf{Y}$. In contrast, VM modifies the background covariance directly and has no concern for observation groups. All observations are included within $\mathbf{H}_{Mj}$, but only a portion of the background error, $\mathbf{X}_j$, is considered within $\mathbf{Y}_{Mj}$, excluding the cross-covariances of (5.5). In VO, the total background perturbations are projected onto a subset of observations. In VM, a subset of the background perturbations is projected onto all of the observations. Both forms result in a portion of $\mathbf{P}$ being considered in the analysis.

The form of variable localization applied in K11 is VO where subsets of observations are used to calculate the analysis for different control variables in an attempt to remove the unphysical, spurious correlations between carbon and the other variables. Several configurations of VO were tested in K11. One of the configurations, referred to as C-univ, split the observations into two groups: (u, v, T, q, P) and (C). The meteorological observations are used to calculate $\mathbf{Y}_{1O}$ and the carbon observa-

tions are used to calculate $\mathbf{Y}_{2O}$. When the analysis for the meteorological variables is computed, $\mathbf{Y}_{1O}$ is used. Likewise, when the analysis for the carbon variables is computed, $\mathbf{Y}_{2O}$ is used. By also applying ρ_{VO}, the localization function only allows the meteorological observations to impact the meteorological analysis and the carbon observations to impact the carbon analysis.

VM considers all observations simultaneously, but it separates the control variables into groups to be decorrelated. In Chapter 3, a balance operator is applied to a Hybrid 4DEnVar. The control variables include streamfunction ψ and unbalanced velocity potential χ^u, which dictate the rotational and divergent components of the wind respectively. These two wind components are weakly correlated in the free atmosphere (Hollingsworth and Lonnberg, 1986), though because of sampling error, cross-correlations will likely exist in the ensemble derived $\mathbf{P}$. VO is unable to remove these cross-correlations. Wind observations should impact both of the wind components ψ and χ^u; thus, VO cannot remove the cross-correlations by restricting observation impact and is unable to accommodate this scenario. VM removes the cross-correlations independent of the observation types and would be able to be applied for these control variable types, either using ρ_{VM} or $\tilde{\mathbf{X}}$.

These two forms of variable localization, VM and VO, are analogous to the two forms of spatial localization: background error covariance, or $\mathbf{B}$, localization and observation error covariance, or $\mathbf{R}$, localization (Greybush et al., 2011). $\mathbf{B}$ localization removes the correlation between two grid points that are at a great distance from one another. It does so by directly modifying the background error, $\mathbf{P}$, which is also commonly referred to as $\mathbf{B}$. This form removes cross-correlations

in model space with no regard to the observations that are being assimilated and is analogous to VM. **R** localization removes the correlation between an observation and a grid point that are at a great distance from one another, restricting an observation's impact to grid points that are nearby. There frequently is also a modification to the observation error covariance matrix, **R**. This observation space spatial localization is analogous to VO.

Both of these spatial localization forms have an impact on the Kalman gain matrix. Defined as the ratio between the background error covariance and the total covariance in observation space, changes to either **P** or **R** alters the Kalman gain and therefore the analysis increment. If **P** increases, the observation provides a greater correction to the background. If **R** increases, the observation's impact is reduced. Each covariance matrix works in the opposite sense through the Kalman gain.

Greybush et al. (2011) compared the formulation of these two forms of spatial localization and applied them within a toy model as well as an intermediate complexity model. The authors found that the performance in skill and balance were comparable when using each localization scheme's optimal length scale, though the optimal length scale of the **R** localization was shorter than that for the **B** localization. The authors also found that while trying to find the optimal length scale for each form, the form of localization was more important than the data assimilation scheme used.

In the following section, the two forms of variable localization, VM and VO, are explored and applied to different data assimilation schemes. Similar to the findings of Greybush et al. (2011), the differences between these two forms of localization

schemes are greater than the differences in their application in various assimilation algorithms. Their relative strengths and weaknesses are consistent across the ensemble schemes presented.

5.3 Formulation

The two forms of variable localization discussed in the previous section, VO and VM, can be applied in many EnKF schemes. In this section, both of these forms are applied in three ensemble data assimilation schemes: the ensemble square root filter (EnSRF), the local ensemble transform Kalman filter (LETKF), and an ensemble-variational (EnVar) algorithm.

The variations of the EnKF solve a form of the original Kalman filter equations (Kalman, 1960). These equations update a prior state, or background, $\mathbf{x}^b \in \mathbb{R}^N$, and its error covariance, $\mathbf{P} \in \mathbb{R}^{N \times N}$, to include observation information. The resulting analysis state, $\mathbf{x}^a \in \mathbb{R}^N$, and its covariance, $\mathbf{P}^a \in \mathbb{R}^{N \times N}$, are computed based on the relative errors of the background and observations:

$$\mathbf{x}^a = \mathbf{x}^b + \delta\mathbf{x}^a, \tag{5.13a}$$

$$\mathbf{P}^a = (\mathbf{I} - \mathbf{KH})\mathbf{P}, \tag{5.13b}$$

where

$$\delta\mathbf{x}^{\mathbf{a}} = \mathbf{Kd}, \tag{5.14}$$

$$\mathbf{K} = \mathbf{PH}^T(\mathbf{HPH}^T + \mathbf{R})^{-1}, \tag{5.15}$$

$\delta\mathbf{x}^a \in \mathbb{R}^N$ is the analysis increment, $\mathbf{K} \in \mathbb{R}^{N\times L}$ is the Kalman gain, $\mathbf{d} \in \mathbb{R}^L$ is the innovation, or the difference between the observation and the background in observation space, and $\mathbf{R} \in \mathbb{R}^{L\times L}$ is the observation error covariance matrix. By acknowledging the duality between Optimal Interpolation (Daley, 1991) and variational methods, Courtier et al. (1994) present a variant formulation of (5.13b) and (5.15):

$$\mathbf{P}^a = [\mathbf{P}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}]^{-1}, \tag{5.16}$$

$$\mathbf{K} = \mathbf{P}^a\mathbf{H}^T\mathbf{R}^{-1}. \tag{5.17}$$

5.3.1 EnSRF

The early EnKF schemes relied on the addition of random perturbations to the observations in order to prevent an underestimation of the analysis error covariance (Burgers et al., 1998; Houtekamer and Mitchell, 1998). Whitaker and Hamill (2002) noted that the addition of perturbations to the observations degrades the accuracy of the analysis. They proposed an alternate formulation that does not require the observations to be perturbed, called the Ensemble Square Root Filter (EnSRF). Assimilating observations one at a time, the analysis increment $\delta\mathbf{x}^a \in \mathbb{R}^N$ is computed using the traditional Kalman gain and unperturbed observations:

$$\delta\mathbf{x}^a = \mathbf{K}d, \tag{5.18a}$$

$$\mathbf{K} = \mathbf{X}\mathbf{Y}^T(\mathbf{Y}\mathbf{Y}^T + R)^{-1}, \tag{5.18b}$$

where $\bar{\mathbf{x}}^a = \bar{\mathbf{x}}^b + \delta\mathbf{x}^a$. Once an observation is assimilated, the resulting analysis becomes the new background for the next observation, updating $\mathbf{X}, \mathbf{Y}$, and d. Assimilating observations serially reduces the size of the matrices greatly (d, R, $\mathbf{YY}^T \in \mathbb{R}^1$ and $\mathbf{XY}^T$, $\mathbf{K} \in \mathbb{R}^N$). This allows for matrix inversions that are much simpler than if all of the observations were assimilated simultaneously.

The covariance matrices, $\mathbf{P}$ and $\mathbf{P}^a$, are symmetric positive definite and can be broken down into their matrix square roots, which are not unique. In standard ensemble methods, $\mathbf{P}$ is approximated by the ensemble perturbations (5.5) and $\mathbf{YY}^T$ represents the background error covariance in observation space. The analysis covariance is similarly represented by the analysis spread, $\mathbf{P}^a = \mathbf{X}^a(\mathbf{X}^a)^T$. The analysis perturbations in the EnSRF are updated using a reduced Kalman gain rather than the traditional gain of the conventional Kalman filter equations:

$$\mathbf{X}^a = (\mathbf{I} - \gamma\mathbf{KH})\,\mathbf{X}, \tag{5.19}$$

where

$$\gamma = \left(1 + \sqrt{\frac{R}{\mathbf{YY}^T + R}}\right)^{-1}. \tag{5.20}$$

Another popular deterministic square root filter, the Ensemble Adjustment Kalman Filter (EAKF, Anderson (2001)), is theoretically similar to the EnSRF. It computes the same analysis ensemble mean and covariance, assimilating observations one at a time, though the membership of the analysis ensemble differs. Starting from the same analysis mean calculation, it then applies a rotation and scaling to

the background perturbations through an adjustment matrix ($\mathbf{X}^a = \mathbf{AX}$) in order to make the background error the identity matrix and compute the new analysis ensemble through the use of singular value decomposition. See Tippett et al. (2003) for a comparison of several square root filters.

Houtekamer and Mitchell (2001) perform spatial localization upon the background error covariance matrices in $\mathbf{K}$ through the use of a Schur or element-wise product between the covariance matrix and a correlation function:

$$\mathbf{K} = [(\rho_{SM} \circ \mathbf{X}\mathbf{X}^T)\mathbf{H}^T][\mathbf{H}(\rho_{SM} \circ \mathbf{X}\mathbf{X}^T)\mathbf{H}^T + R]^{-1}, \tag{5.21}$$

where $\rho_{SM} \in \mathbb{R}^{N \times N}$ is the covariance localization function which is based on the distance between each pair of grid points and SM stands for spatial localization in model space, e.g. a fifth-order piecewise Gaussian approximation with compact support of Gaspari and Cohn (1999). It retains the covariances for grid points that are in close proximity to one another, whose true correlation is likely large, and removes them for points that are distant, whose true correlation is likely small. This localized $\mathbf{K}$ is modified as in the traditional formulation in order to compute the analysis perturbations.

Alternatively, the background covariance can be localized in observation space, after the application of $\mathbf{H}$:

$$\mathbf{K} = (\rho_{SO} \circ \mathbf{X}\mathbf{Y}^T)(\mathbf{Y}\mathbf{Y}^T + R)^{-1}. \tag{5.22}$$

In this formulation, the localization function, $\rho_{SO} \in \mathbb{R}^N$ is based on the distance between the observation being assimilated and each grid point, and SO stands for spatial localization in observation space. This form of localization prohibits the observations from impacting grid points that are far from its location. As discussed in Greybush et al. (2011), the optimal length scale for the observation space localization is shorter than for the model space localization with the model space localization being more severe for the same length scale. Since only one observation is being assimilated at a time, $\mathbf{YY}^T$ is a scalar and does not need to be localized. Due to this removal of additional localization along with the reduction in size of ρ_{SO} and the elimination of the need to calculate $\mathbf{XX}^T$ explicitly, localizing in observation space is computationally preferred. This is also the preferred form of localization for the EAKF.

Like spatial localization, localization between variable types can also be applied in two ways as described in Section 5.2: in model space and in observation space. To apply these forms of variable localization within the EnSRF, the strategy of the spatial localization is followed. Beginning with VM, another covariance localization function is included in $\mathbf{K}$:

$$\mathbf{K} = [(\rho_{SM} \circ \rho_{VM} \circ \mathbf{XX}^T)\mathbf{H}^T][\mathbf{H}(\rho_{SM} \circ \rho_{VM} \circ \mathbf{XX}^T)\mathbf{H}^T + R]^{-1}. \tag{5.23}$$

The localization of the covariance terms function as (5.7) and the cross-correlations between $\mathbf{X}_1$ and $\mathbf{X}_2$ are removed from $\mathbf{K}$. Due to the commutative nature of the Schur product, these localization functions can be applied interchangeably. If all the cross-

correlations are to be retained and no variable localization is to be applied, $\rho_{VM} = 1$, (5.23) reduces to (5.21).

Continuing to follow the methodology of the spatial localization, an additional localization function is also added to the observation space form of $\mathbf{K}$:

$$\mathbf{K} = (\rho_{SO} \circ \rho_{VO} \circ \mathbf{X}\mathbf{Y}^T)(\mathbf{Y}\mathbf{Y}^T + R)^{-1}. \tag{5.24}$$

The numerator of $\mathbf{K}$ is localized as in (5.8). By assimilating a single observation at a time, no variable localization is necessary on $\mathbf{Y}\mathbf{Y}^T \in \mathbb{R}$, though as in (5.6), the effect of the cross-correlations remain in that term. If no variable localization is to be applied to an observation type, $\rho_{VO} = 1$, (5.24) reduces to (5.22).

One disadvantage of this formulation is that the cross-correlation terms, $\mathbf{X}_1\mathbf{X}_2^T$ and $\mathbf{X}_2\mathbf{X}_1^T$, remain (5.8). For many cases, this term becomes zero due to the construction of $\mathbf{H}$. The first cross-correlation term, $\mathbf{X}_1\mathbf{X}_2^T$, is multiplied by the $\mathbf{H}$ component for the second set of control variables, $\mathbf{H}_{M2}$. Likewise, the second cross-correlation term is multiplied by the $\mathbf{H}$ component of the first set of control variables. If the observation being assimilated only impacts variables from one control variable group, the cross-correlation terms vanish. However, if the observation impacts variables from both control variable groups, as a wind observation would for the aforementioned ψ and χ^u control variables, the cross-correlation terms would remain and the variable localization would be ineffective. The VO formulation cannot remove the cross-correlation between control variables that are impacted by the same observation.

The alternate form of VM described in Section 5.2 is applied to the EnSRF, where the background perturbations were extended, producing an additional ensemble and thereby removing the cross-correlations from the covariance matrices (5.10 and 5.12). A new $\mathbf{K}$ is constructed by using the extended background covariances:

$$\mathbf{K} = (\rho_{SO} \circ \tilde{\mathbf{X}}\tilde{\mathbf{Y}}^T)(\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^T + R)^{-1}, \tag{5.25}$$

and the analysis mean is computed as in the traditional EnSRF (5.18). The analysis perturbations are also computed as in the traditional EnSRF, (5.19) - (5.20), with the exception of the γ term where $\tilde{\mathbf{Y}}$ replaces $\mathbf{Y}$.

The spatial covariance localization in this alternate VM formulation is applied as in the standard form in observation space, through ρ_{SO}. The variable localization, however, is not applied through ρ_V; it is applied as a natural consequence of extending the background covariance. The cross-covariances between $\mathbf{X}_1$ and $\mathbf{X}_2$ have been eliminated through the extension of the matrix to $2M$, yet an observation can still impact all of the model variables if it is desired. This also allows the variable localization to be applied in model space without computing the full $N \times N$ background error covariance, granting significant computational savings when compared with the previous model space formulation.

To compare the alternate VM to VO (5.8), the numerator of $\mathbf{K}$ is constructed

for VM using the extended perturbations:

$$\begin{aligned}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T\mathbf{H}^T &= \begin{pmatrix} \mathbf{X}_1 & 0 \\ 0 & \mathbf{X}_2 \end{pmatrix} \begin{pmatrix} \mathbf{X}_1^T & 0 \\ 0 & \mathbf{X}_2^T \end{pmatrix} \begin{pmatrix} \mathbf{H}_{M1}^T \\ \mathbf{H}_{M2}^T \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{X}_1\mathbf{X}_1^T\mathbf{H}_{M1}^T \\ \mathbf{X}_2\mathbf{X}_2^T\mathbf{H}_{M2}^T \end{pmatrix}.\end{aligned} \tag{5.26}$$

The cross-correlation terms, $\mathbf{X}_1\mathbf{X}_2^T$ and $\mathbf{X}_2\mathbf{X}_1^T$, do not appear in this formulation. VM is able to remove the cross-correlations between the control variable groups without regard to what observations are assimilated. If an observation does not project onto the second variable set, $\mathbf{H}_{M2} = 0$ and the second row of the Kalman gain is also zero. This results in (5.26) being equivalent to (5.8). If $\mathbf{H}_{M2}$ is nonzero, these forms differ. The $\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^T$ term removes the cross-correlation terms as well, computed as in (5.12). Again, if an observation does not project onto the second control variable group, (5.12) simplifies to (5.6).

5.3.2 LETKF

Another deterministic square root EnKF is the local ensemble transform Kalman filter (LETKF, Hunt et al., 2007). This filter separates the solution for the analysis into several independent, local calculations in which the analysis for each grid point uses only the observations within the local region. The analysis ensemble mean is calculated by adding a local linear combination of the various background

perturbations to the background mean:

$$\bar{\mathbf{x}}^a = \bar{\mathbf{x}}^b + \delta\mathbf{x}^a = \bar{\mathbf{x}}^b + \mathbf{X}\bar{\mathbf{w}}^a, \tag{5.27a}$$

$$\mathbf{X}^a = \mathbf{X}^b\mathbf{W}^a, \tag{5.27b}$$

where $\bar{\mathbf{w}}^a \in \mathbb{R}^M$ controls each ensemble member's contribution to the analysis mean. Rather than pre-multiplying the background perturbations with a transformation matrix as in the EnSRF and the EAKF, the LETKF post-multiplies the background perturbations with the weight matrix, $\mathbf{W}^a \in \mathbb{R}^{M\times M}$. The analysis ensembles of the EnSRF, EAKF, and LETKF have different membership, but the ensembles span the same analysis covariance. Hunt et al. (2007) derived these weights by first constructing a cost function:

$$J(\bar{\mathbf{w}}) = \frac{1}{2}\bar{\mathbf{w}}^T\bar{\mathbf{w}} + \frac{1}{2}(\mathbf{d} - \mathbf{Y}\bar{\mathbf{w}})^T\mathbf{R}^{-1}(\mathbf{d} - \mathbf{Y}\bar{\mathbf{w}}), \tag{5.28}$$

and then setting the gradient to zero and solving for the weights analytically. Combining the weight with $\mathbf{X}$ as in (5.27a) produces the analysis increment:

$$\delta\mathbf{x}^a = (\mathbf{X}\mathbf{W}^a)\,(\mathbf{Y}\mathbf{W}^a)^T\mathbf{R}^{-1}\mathbf{d}, \tag{5.29}$$

where weights for the ensemble spread are the symmetric square root of the analysis covariance in ensemble space:

$$\mathbf{W}^a = (\mathbf{I} + \mathbf{Y}^T\mathbf{R}^{-1}\mathbf{Y})^{-\frac{1}{2}}. \tag{5.30}$$

For comparison with the EnSRF, the ensemble mean update can be written in terms of a Kalman gain:

$$\mathbf{K} = (\mathbf{X}\mathbf{W}^a)\,(\mathbf{Y}\mathbf{W}^a)^T\,\mathbf{R}^{-1}, \tag{5.31}$$

which is mathematically equivalent to (5.17).

Since the calculations are performed in ensemble space, the physical space representation of the background covariance, $\mathbf{X}\mathbf{X}^T$ (5.5), is not computed; therefore, the spatial localization is not easily implemented by applying a correlation function to the background perturbations. The LETKF, instead, applies two forms of spatial localization to $\mathbf{R}$. First, when the analysis is calculated locally at each grid point, it only incorporates the observations that fall within a specified radius. Separate weights are calculated for each grid point, but they are applied to all variable types, allowing the cross-covariances to work across the variables, but not at long distances. Second, to curtail the effect of a sudden cutoff at the edge of areas of observation influence (Hunt et al., 2007), $\mathbf{R}$ is also multiplied by a diagonal correlation matrix, $\rho_{SR} \in \mathbb{R}^{L\times L}$, that is based on distance from the observation to the grid point being analyzed ($\mathbf{R} \rightarrow \rho_{SR} \circ \mathbf{R}$). This alternate form of spatial localization gradually reduces an observation's influence on neighboring grid points as the distance increases.

K11 implemented a form of variable localization in the LETKF through the observation selection. Groups of analysis variables are chosen as well as groups of observations corresponding to each analysis group. The formulation here is demonstrated for two groups each, but theoretically, this formulation can also be extended to as many groups as variable types. The groups of analysis variables must be

mutually exclusive, but the groups of observations need not be.

To construct the background perturbations, $\mathbf{Y}$ is first computed using the full background perturbations. Then, the rows corresponding to the observations of interest are selected to create $\mathbf{Y}_{iO}$ (5.3) where i refers to the subset of observations. A different analysis weight is calculated for each subset of analysis variables using only the observations within the corresponding subset. These weights are derived by constructing separate cost functions, setting the gradient of each cost function to zero, and solving for each weight. Once the weights are found, the analysis increments can be created:

$$\delta \mathbf{x}_i^a = \mathbf{X}\bar{\mathbf{w}}_i^a = (\mathbf{X}\mathbf{W}_i^a)\left(\mathbf{Y}_{iO}\mathbf{W}_i^a\right)^T \left(\rho_{SRi} \circ \ \mathbf{R}_i\right)^{-1} \mathbf{d}_i, \tag{5.32}$$

where $\mathbf{R}_i^{-1} \in \mathbb{R}^{L_i \times L_i}$ and $\mathbf{d}_i \in \mathbb{R}^{L_i}$. The weight matrix, $\mathbf{W}_i^a \in \mathbb{R}^{M \times M}$, is similarly constructed using only the observations from each set:

$$\mathbf{W}_i^a = (\mathbf{I} + \mathbf{Y}_{iO}^T(\rho_{SRi} \circ \mathbf{R}_i)^{-1}\mathbf{Y}_{iO})^{-\frac{1}{2}}. \tag{5.33}$$

The analysis mean is updated using each increment and the perturbations are updated using each weight matrix:

$$\bar{\mathbf{x}}^a = \bar{\mathbf{x}}^b + \rho_{Vi} \circ \delta \mathbf{x}_i^a, \tag{5.34a}$$

$$\mathbf{X}_i^a = \mathbf{X}_i \mathbf{W}_i^a, \tag{5.34b}$$

where $\rho_{Vi} \in \mathbb{R}^N$ is variable localization function, comprised of zeros and ones, allowing the first weight to update the first set of analysis variables only and the second weight to update the second set of analysis variables only. Similarly to the serial assimilation of the EnSRF, the weights for each observation group need to be computed iteratively. Once the first weight is found and the analysis is updated, the analysis mean and spread become the new background for the calculation of the second weight. The serial assimilation of the groups is not significant if the control variables and observations from each group are independent. If they are not independent, the analysis for the second group will be different compared to if the analysis is not calculated iteratively. Consider the computation of $\mathbf{Y}_{2O}$. This term contains the background perturbations from the first control variable group:

$$\mathbf{Y}_{2O} = \mathbf{H}_{2O}\mathbf{X} = \begin{pmatrix} \mathbf{X}_1\mathbf{W}_1^a \\ \\ \mathbf{X}_2 \end{pmatrix}. \tag{5.35}$$

If $\mathbf{H}_{2O}$ does not project onto the first set of control variables, the change in $\mathbf{X}_1$ after the first analysis update to $\mathbf{X}_1\mathbf{W}_1^a$ is of no consequence. However, if the observations from the second group project onto the first group of control variables, the iterative calculation of weights becomes significant.

The standard LETKF formulation calculates one set of analysis weights that is applied to all of the variables. VO calculates multiple sets of analysis weights that are applied to different control variables, which restricts the impact of each observation group. This formulation is comparable to the EnSRF formulation in

Section 5.3.1, though there are subtle differences due to the serial nature of the EnSRF. In the EnSRF formulation of VO, an observation is restricted to impact only certain analysis variables. In the LETKF VO, an analysis variable is restricted to be impacted by certain observations. Table 5.1 provides a comparison of the analysis increments for each data assimilation scheme and variable localization form.

Even though the background error is not explicitly computed in the LETKF, a form of variable localization can be constructed that eliminates the cross-correlations between the analysis variables directly. In this form of VM, the observation types are not divided into multiple sets and are considered simultaneously. However, the control variables are again broken down into mutually exclusive groups of variables that are to be uncorrelated.

Each group of variables has its own set of ensemble weights, which are vertically concatenated into one vector, $\tilde{\mathbf{w}} = \left(\bar{\mathbf{w}}_1^T \; \bar{\mathbf{w}}_2^T \right)^T \in \mathbb{R}^{2M}$. A single background error matrix is also constructed containing the background perturbations from each set of variables and ensuring no correlation exists between the two groups, as in the EnSRF (5.9). The background perturbations in observation space $\tilde{\mathbf{Y}}$ are also computed as in the EnSRF (5.11), though all of the observations are considered instead of only one.

The concatenated matrices are substituted into the original LETKF formulation (5.29) to solve for both sets of weights simultaneously and compute a single increment:

$$\delta \mathbf{x}^a = \tilde{\mathbf{X}} \tilde{\mathbf{w}} = \left(\tilde{\mathbf{X}} \tilde{\mathbf{W}}^a \right) \left(\tilde{\mathbf{Y}} \tilde{\mathbf{W}}^a \right)^T \left(\rho_{SR} \circ \mathbf{R} \right)^{-1} \mathbf{d}, \tag{5.36}$$

where

$$\tilde{\mathbf{W}}^a = \left(\mathbf{I} + \tilde{\mathbf{Y}}^T \left(\rho_{SR} \circ \mathbf{R}\right)^{-1} \tilde{\mathbf{Y}}\right)^{-\frac{1}{2}} \in \mathbb{R}^{2M \times 2M}. \tag{5.37}$$

The weight matrix is also extended and used in conjunction with $\tilde{\mathbf{X}}$ to calculate $\mathbf{X}^a$:

$$\mathbf{X}^a = \tilde{\mathbf{X}} \tilde{\mathbf{W}}^a \mathbf{C}, \tag{5.38}$$

where $\mathbf{C} = \begin{pmatrix} \mathbf{I} & \mathbf{I} \end{pmatrix}^T \in \mathbb{R}^{2M \times M}$.

To compare with the VO formulation (5.32), the concatenated terms $\tilde{\mathbf{Y}}$ and $\tilde{\mathbf{w}}$ are expanded. This results in two weights that are dependent on each other:

$$\bar{\mathbf{w}}_1^a = \mathbf{W}_{M1}^a \left(\mathbf{Y}_{M1} \mathbf{W}_{M1}^a\right)^T \left(\rho_{SR} \circ \ \mathbf{R}\right)^{-1} (\mathbf{d} - \mathbf{Y}_{M2} \bar{\mathbf{w}}_2^a), \tag{5.39a}$$

$$\bar{\mathbf{w}}_2^a = \mathbf{W}_{M2}^a \left(\mathbf{Y}_{M2} \mathbf{W}_{M2}^a\right)^T \left(\rho_{SR} \circ \ \mathbf{R}\right)^{-1} (\mathbf{d} - \mathbf{Y}_{M1} \bar{\mathbf{w}}_1^a), \tag{5.39b}$$

where

$$\mathbf{W}_{Mj}^a = \left(\mathbf{I} + \mathbf{Y}_{Mj}^T (\rho_{SR} \circ \mathbf{R})^{-1} \mathbf{Y}_{Mj}\right)^{-1}. \tag{5.40}$$

One significant difference between the weights for each form of variable localization is that VM has an additional term. For $\bar{\mathbf{w}}_1^a$, this term is $\mathbf{Y}_{M1}^T \mathbf{R}^{-1} \mathbf{Y}_{M2} \bar{\mathbf{w}}_2^a$, which represents the influence of one observation on model variables from different groups. This is how the VM is able to address the ψ and χ^u scenario described in Section 5.2. If a single observation does not project onto analysis variables from both groups, this term is zero.

Another notable difference is the background perturbations in observation

space: VO uses $\mathbf{Y}_{iO}$ and VM uses $\mathbf{Y}_{Mj}$. Examining their structure highlights how the localization for each form operates. The background perturbations in VO are represented by $\mathbf{Y}_{iO} \in \mathbb{R}^{L_i \times M}$. Only certain observations are considered when computing a particular analysis grid point. In contrast, $\mathbf{Y}_{Mj} \in \mathbb{R}^{L \times M}$ in VM uses all of the observations when computing an analysis grid point. These perturbation matrices also consider different portions of the background covariance: $\mathbf{Y}_{iO}$ is calculated using the full background perturbations, $\mathbf{X}$, while $\mathbf{Y}_{Mj}$ only considers part of the background perturbations, $\mathbf{X}_j$. Combined, this illustrates how the two localization schemes work. VO does not localize the covariance itself, but selects rows of observation impact to include in the analysis. VM localizes the covariance directly and allows all observations to potentially impact all control variables.

The differences between the two forms of variable localization in the LETKF parallel the differences seen in the EnSRF. Both EnKF formulations face the same drawback of VO: its inability to remove background covariances from two variables affected by the same observation. It makes no difference that the EnSRF considers only one observation at a time while the LETKF splits the observation types into groups. Likewise, both formulations are benefited by the application of VM in this regard. It removes intervariable correlations without consideration of the observations. This conclusion mirrors one of the findings of Greybush et al. (2011) for spatial localization, where the authors determined that the choice of data assimilation algorithm was less impactful than the choice of spatial localization form when determining localization length scales.

5.3.3 EnVar

Ensemble-variational (EnVar) schemes use a cost function framework to incorporate ensemble perturbations into the background error covariance. The Hybrid 3DEnVar (Kleist and Ide, 2015a; Wang, 2010) draws parallels with the LETKF in its formulation, particularly evident through the cost function derivation of the LETKF. The primary difference between the formulations is the implementation of spatial localization: the LETKF applies it to $\mathbf{R}$ and the EnVar applies it to $\mathbf{P}$.

For simplicity and ease of comparison with the other EnKFs in the previous sections, the 100% ensemble covariance formulation of the EnVar is used rather than the hybrid form, removing the static parts of the increment and cost function. Following the notation of Wang (2010), the increment within an EnVar scheme is computed using a set of ensemble weights multiplied by the ensemble perturbations from each member normalized by $\sqrt{M-1}$:

$$\delta\mathbf{x} = \sum_{m=1}^{M} (\alpha^m \circ \mathbf{X}^m) = \mathbf{E}\boldsymbol{\alpha}, \tag{5.41}$$

where $\delta\mathbf{x} \in \mathbb{R}^N$ is the increment, $\alpha^m \in \mathbb{R}^Q$ represents the weights for each m ensemble member (analogous to $\bar{\mathbf{w}}$ in the LETKF), Q is the number of model grid points, $\mathbf{X}^m \in \mathbb{R}^N$ represents the normalized background perturbations for each member, $\mathbf{E} = \left(\mathbf{E}_1^T\ \mathbf{E}_2^T\right)^T \in \mathbb{R}^{N\times QM}$ represents the Schur product of ensemble perturbations, and $\boldsymbol{\alpha} \in \mathbb{R}^{QM}$ is a vertical concatenation of the weights for each member. A varia-

tional cost function is constructed to solve for the weights:

$$J(\boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{L}^{-1}\boldsymbol{\alpha} + \frac{1}{2}(\mathbf{d} - \mathbf{HE}\boldsymbol{\alpha})^T\mathbf{R}^{-1}(\mathbf{d} - \mathbf{HE}\boldsymbol{\alpha}), \tag{5.42}$$

where $\mathbf{L} \in \mathbb{R}^{QM\times QM}$ is the covariance for the ensemble weights. It is a block diagonal matrix with submatrices that define the spatial correlation of the control variable, which dictates the spatial localization. This cost function is of the same form as the LETKF cost function (5.28), where $\boldsymbol{\alpha}$ is equivalent to $\bar{\mathbf{w}}$ and $\mathbf{HE}$ is equivalent to $\mathbf{Y}$.

In practice, the cost function is minimized iteratively, but the control variable can be solved analytically by setting the gradient to zero and solving for $\boldsymbol{\alpha}$. Once found, the increment is formed:

$$\delta\mathbf{x} = (\mathbf{EV})\,(\mathbf{HEV})^T\,\mathbf{R}^{-1}\mathbf{d}, \tag{5.43}$$

where

$$\mathbf{V} = (\mathbf{L}^{-1} + \mathbf{E}^T\mathbf{H}^T\mathbf{R}^{-1}\mathbf{HE})^{-\frac{1}{2}} \in \mathbb{R}^{QM\times QM}. \tag{5.44}$$

This increment is of the same form as (5.29) in the LETKF, where $\mathbf{V}$ is analogous to $\mathbf{W}^a$. The primary difference is in the application of the spatial localization. In the LETKF, it is applied through $\mathbf{R}$ and a local implementation of the analysis equations. The EnVar applies the localization through $\mathbf{L}$ and uses a global implementation.

To apply VO within the EnVar framework, the same methodology of the

LETKF is used by computing multiple sets of weights:

$$\delta \mathbf{x}_i = \sum_{m=1}^{M} (\alpha_i^m \circ \mathbf{X}^m) = \mathbf{E}\boldsymbol{\alpha}_i. \tag{5.45}$$

The weights are calculated by solving separate cost functions iteratively using groups of observations as in (5.32), setting their gradients to zero, and solving for each $\boldsymbol{\alpha}_i$ to create separate increments:

$$\delta \mathbf{x}_i = \mathbf{E}\mathbf{V}_i \left(\mathbf{H}_{iO}\mathbf{E}\mathbf{V}_i\right)^T \mathbf{R}_i^{-1}\mathbf{d}_i, \tag{5.46}$$

where

$$\mathbf{V}_i = \left(\mathbf{L}_i^{-1} + \mathbf{E}^T\mathbf{H}_{iO}^T\mathbf{R}_i^{-1}\mathbf{H}_{iO}\mathbf{E}\right)^{-\frac{1}{2}}. \tag{5.47}$$

Once the increment is found, the analysis is computed in the same manner as LETKF (5.34a). As in LETKF VO, the increments must be computed serially, with the analysis for the first group of observations becoming the background for calculating the analysis for the second group of observations. Comparing with (5.32), the equivalence between the EnVar and the LETKF is apparent. The EnVar also suffers from the same drawback in its implementation of VO; the formulation cannot accommodate a single observation having impact across both control variable groups.

Similarly to the LETKF, VM can also be formulated for the EnVar. The model variables are broken down into mutually exclusive groups that are to be dissociated. The increment then includes multiple groups of weights multiplied by each group of

perturbations:

$$\delta \mathbf{x} = \sum_{m=1}^{M} (\alpha_1^m \circ \mathbf{X}_1^m) + \sum_{m=1}^{M} (\alpha_2^m \circ \mathbf{X}_2^m) = \tilde{\mathbf{E}} \tilde{\boldsymbol{\alpha}}, \tag{5.48}$$

where $\tilde{\mathbf{E}}$ is the extended Schur product for the multiple variable groups:

$$\tilde{\mathbf{E}} = \begin{pmatrix} \mathbf{E}_1 & 0 \\ 0 & \mathbf{E}_2 \end{pmatrix} \in \mathbb{R}^{N \times 2QM}. \tag{5.49}$$

Similarly to $\tilde{\mathbf{X}}$ in the VM form of the EnSRF and the LETKF, the extension of the $\mathbf{E}$ matrix removes the cross-correlation between the groups of variables. The weight error covariance, $\mathbf{L}$, is extended in the same manner to form $\tilde{\mathbf{L}}$. The control variable is also extended to include both sets of ensemble weights, $\tilde{\boldsymbol{\alpha}} = \begin{pmatrix} \boldsymbol{\alpha}_1^T & \boldsymbol{\alpha}_2^T \end{pmatrix}^T \in \mathbb{R}^{2QM}$.

As in the LETKF, a single cost function is constructed to solve for both sets of weights, considering all of the observations simultaneously. Even though in practice the cost function is solved numerically, the analytical solution is written for the increment to compare with the other EnKF schemes [(5.25) and (5.36)]:

$$\delta \mathbf{x} = \tilde{\mathbf{E}} \tilde{\boldsymbol{\alpha}} = \left(\tilde{\mathbf{E}} \tilde{\mathbf{V}} \right) \left(\mathbf{H} \tilde{\mathbf{E}} \tilde{\mathbf{V}} \right)^T \mathbf{R}^{-1} \mathbf{d}, \tag{5.50}$$

where

$$\tilde{\mathbf{V}} = \left(\tilde{\mathbf{L}}^{-1} + \tilde{\mathbf{E}}^T \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H} \tilde{\mathbf{E}} \right)^{-\frac{1}{2}} \in \mathbb{R}^{2QM \times 2QM}. \tag{5.51}$$

Refer to Table 5.1 for a comparison of the analysis increments for the three data assimilation schemes and each form of variable localization.

The concatenated matrices and vectors are then expanded in order to compare

the analysis weights with VO (5.46) and with the LETKF (5.39a,5.39b):

$$\boldsymbol{\alpha}_1 = \mathbf{V}_1 \left(\mathbf{HEV}_1\right)^T \mathbf{R}^{-1}(\mathbf{d} - \mathbf{HE}_2\boldsymbol{\alpha}_2), \tag{5.52a}$$

$$\boldsymbol{\alpha}_2 = \mathbf{V}_2 \left(\mathbf{HEV}_2\right)^T \mathbf{R}^{-1}(\mathbf{d} - \mathbf{HE}_1\boldsymbol{\alpha}_1). \tag{5.52b}$$

where

$$\mathbf{V}_i = \left(\mathbf{L}_i^{-1} + \mathbf{E}_i^T\mathbf{H}^T\mathbf{R}^{-1}\mathbf{HE}_i\right)^{-\frac{1}{2}}. \tag{5.53}$$

Similarly to the LETKF, when comparing the VM weights with its VO counterpart (5.46), an additional term is seen. This term, of the form $\mathbf{E}_1^T\mathbf{H}^T\mathbf{R}^{-1}\mathbf{HE}_2\boldsymbol{\alpha}_2$, is the term that allows one observation to impact both variable sets directly. There is also a similar difference in the observation operators ($\mathbf{H}_{iO}$ in VO compared to $\mathbf{H}$ in VM) and the background perturbations ($\mathbf{E}$ in VO compared to $\mathbf{E}_i$ in VM). Both VM implementations are capable of removing the cross-correlation terms while allowing observation impact on both control variable groups simultaneously.

5.4 Single Observation Demonstration

An example of the VM EnVar formulation using one of the scenarios discussed in Section 5.2 is now presented. Using a 4DEnVar scheme in the SPEEDY model, the control variables for this case are $(\psi, \chi^u, T^u, q, P^u)$, where superscript u represents the unbalanced portion of each variable. A balance operator provides the large scale, balanced correlations between ψ and (χ, T, P). Even though the balance operator prescribes the cross-covariances between the variables, there will be

additional correlations provided by the ensemble. To make ψ and (χ^u, T^u, P^u) fully uncorrelated, variable localization is applied. As discussed in Section 5.3.2, a u observation needs to be able to impact both ψ and χ^u; therefore, VO cannot accommodate this scenario. If an observation is able to impact more than one control variable, then those variables will retain their ensemble-derived correlations. VM, however, can handle this scenario. By separating the variables into two groups, ψ and (χ^u, T^u, q, P^u), two sets of weights are computed and the correlation between the two groups of variables is removed.

Part of $\mathbf{K}$ is constructed to show the impact of VM, considering only the two variables of interest, ψ and χ^u. First, the $\mathbf{XY}^T$ term without the variable localization is formulated in conjunction with the application of the balance operator, $\mathbf{\Gamma}$:

$$\begin{aligned}\mathbf{Z}\mathbf{Z}^T\mathbf{\Gamma}^T\mathbf{H}^T &= \begin{pmatrix} \mathbf{Z}_\psi \\ \mathbf{Z}_\chi \end{pmatrix} \begin{pmatrix} \mathbf{Z}_\psi^T & \mathbf{Z}_\chi^T \end{pmatrix} \begin{pmatrix} 1 & c \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mathbf{H}_\psi^T \\ \mathbf{H}_\chi^T \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{Z}_\psi\mathbf{Z}_\psi^T(\mathbf{H}_\psi^T + c\mathbf{H}_\chi^T) + \mathbf{Z}_\psi\mathbf{Z}_\chi^T\mathbf{H}_\chi^T \\ \mathbf{Z}_\chi\mathbf{Z}_\psi^T(\mathbf{H}_\psi^T + c\mathbf{H}_\chi^T) + \mathbf{Z}_\chi\mathbf{Z}_\chi^T\mathbf{H}_\chi^T \end{pmatrix},\end{aligned} \tag{5.54}$$

where $\mathbf{Z}$ represents the normalized ensemble perturbations in the unbalanced variable space and $\mathbf{X} = \mathbf{\Gamma}\mathbf{Z}$. The subscript ψ refers to the perturbations and observation operator for the ψ control variable and the subscript χ refers to the perturbations and observation operator for the χ^u control variable. Within $\mathbf{\Gamma}$, c is the part of the balance operator that represents the balanced correlations between ψ and χ. With no variable localization, observations with nonzero observation operators for both ψ

and χ^u can impact both variables. The cross-correlations between the two variables are present as well.

When VM is applied to this case, the background perturbations in the unbalanced space are extended as they were for the traditional background perturbations:

$$\begin{aligned}
&\tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^T\boldsymbol{\Gamma}^T\mathbf{H}^T \\
&= \begin{pmatrix} \mathbf{Z}_\psi & 0 \\ 0 & \mathbf{Z}_\chi \end{pmatrix} \begin{pmatrix} \mathbf{Z}_\psi^T & 0 \\ 0 & \mathbf{Z}_\chi^T \end{pmatrix} \begin{pmatrix} 1 & c \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mathbf{H}_\psi^T \\ \mathbf{H}_\chi^T \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{Z}_\psi\mathbf{Z}_\psi^T(\mathbf{H}_\psi^T + c\mathbf{H}_\chi^T) \\ \mathbf{Z}_\chi\mathbf{Z}_\chi^T\mathbf{H}_\chi^T \end{pmatrix}.
\end{aligned} \tag{5.55}$$

The cross-covariance terms, $\mathbf{Z}_\psi\mathbf{Z}_\chi^T$ and $\mathbf{Z}_\chi\mathbf{Z}_\psi^T$, have been removed, while the impact for both variables remains nonzero. This allows a single observation where the observation operators are nonzero for each variable to impact both of those variables, while simultaneously removing their cross-correlations.

A single observation impact test demonstrates this variable localization implementation. Assimilating a u observation without any variable localization, (Figure 5.1a), there is a response in ψ and χ that is consistent with an increase in the zonal wind at the point between the dipoles. There is also some anisotropy associated with the flow-dependent covariances. When applying VM, any cross-covariances for ψ and χ^u present in the ensemble correlations are removed (Figure 5.1b). The overall structure is similar between the two cases, with a majority of the information drawing from the direct impact of u on ψ and χ. However, slight differences are

present, as displayed in Figure 5.1c. The case with the cross-covariances removed is smoother than the case that includes the cross-covariances, suggesting that these covariances may have been subject to sampling error and should not be considered.

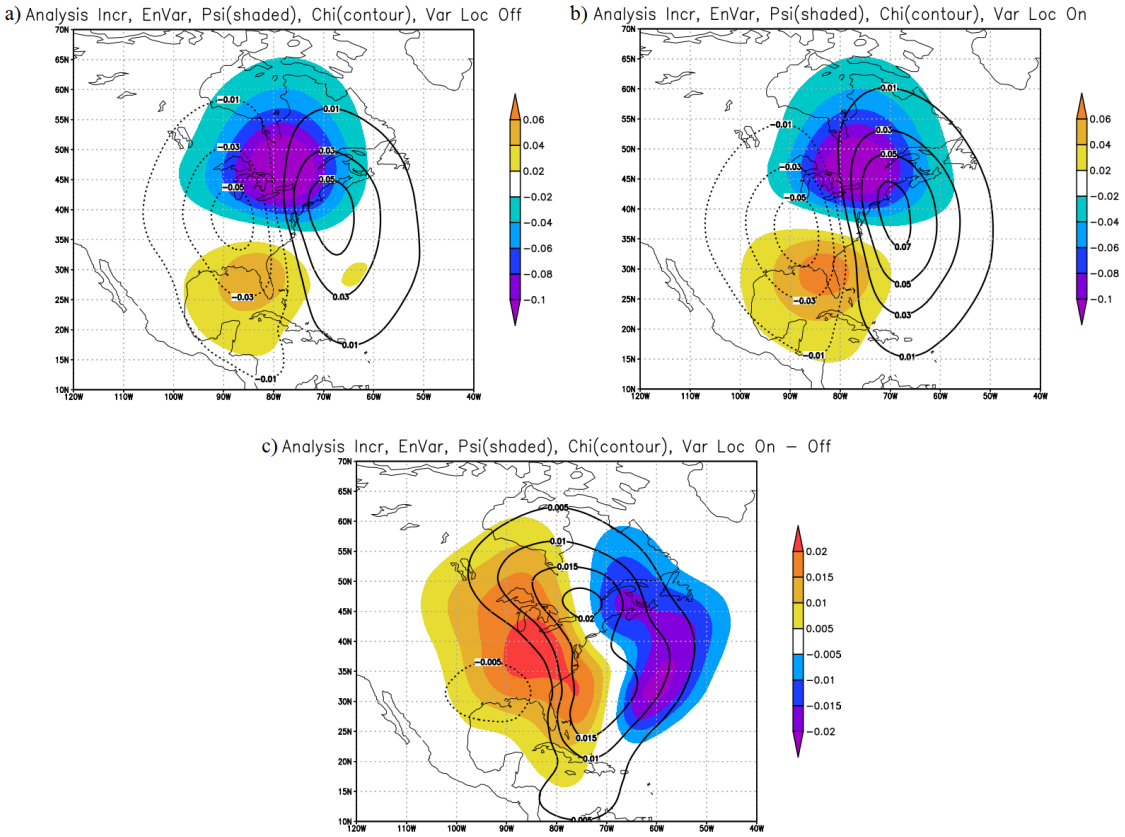

Figure 5.1: EnVar analysis increment for ψ (shaded) and χ (contoured) assimilating a single u observation at the lowest model level (a) without any variable localization and (b) with model space variable localization. (c) shows the difference with and without variable localization.

5.5 Summary and Discussion

Variable localization is a method by which to remove the spurious correlations that exist between variables that are physically unrelated. Two forms of variable localization were described: observation space variable localization (VO) and model

space variable localization (VM). Both forms were then implemented within three data assimilation schemes: the EnSRF, the LETKF, and the EnVar. Differences exist between the data assimilation schemes, but the forms of variable localization are implemented in similar ways. VO removes the correlation between model variables by not allowing an observation to impact certain analysis variables. VM extends the background error and forces the removal of the cross-correlations directly without regard to observation type.

For the LETKF and the EnVar, both forms of variable localization are implemented by calculating multiple sets of weights for different variable types. The method by which these weights are calculated differ for each form. VO computes each set of weights using its own group of observation types. VM computes two sets of weights simultaneously considering all of the observation types. While both formulations need the analysis variable groups to be mutually exclusive, the groups of observations for VO can overlap. In contrast, VM does not use two separate groups of observations; all observation types are considered at once. This is reflected in how the background error is constructed in each formulation. In VO, $\mathbf{Y}$ is calculated using the full background perturbations, $\mathbf{X}$, then rows corresponding to individual observations are selected. This gives control over which model variables feel the impact of a particular observation. In VM, the background perturbations $\mathbf{X}$ are localized by variable first, then $\mathbf{Y}$ is constructed for all observations. This form allows for a single observation to impact multiple analysis variables that also have their cross-correlations removed.

K11 explored several different configurations for removing cross-correlations

between different variable sets. One such configuration, referred to as C-univ in that paper, decorrelates the carbon variables with the meteorological variables in the analysis. While simplistic, the transport error covariances are not considered during the analysis of the carbon variables. In the K11 paper, VO is used within an LETKF to implement this scenario and it yields positive results, stabilizing the error and resulting in realistic CF fields, though the results are worse than the other configurations tested due to the lack of representation of the transport error. VM could also be used effectively in this scenario. The two sets of control variables for both formulations are (u, v, T, q, P) and (C, CF). VO also separates the meteorological observations from the carbon: (u, v, T, q, P) and (C).

Another configuration that was tested in K11 is the L-1way method, which allows for the cross-covariance between the carbon variables and the wind variables, but it does not allow for the carbon variables to negatively impact the wind, having a one-way feedback only. This configuration had the best performance of the configurations tested in that paper. Using VO, the analysis variables are again split into the meteorological variables and the carbon variables: (u, v, T, q, P) and (C, CF). The groups of observations, however, are split into overlapping sets: (u, v, T, q, P) and (u, v, C). The wind observations impact the carbon analysis, but the carbon observations do not impact the wind analysis. Unfortunately, VM cannot manage this scenario since it has no means to represent a one-way feedback; it does not split the observations into sets. In VO, by choosing to use a particular observation type in both sets, cross-covariances are being introduced among the two analysis variable sets. VM enforces a full decorrelation of the control variables and cannot implement

this configuration.

Section 5.4 presented a situation that was manageable by VM only. The control variables to be decorrelated are the rotational component of the wind ψ and the unbalanced divergent component of the wind χ^u. To assimilate a wind observation, the impact must be able to span both groups of control variables while also having the cross-correlations removed. VO is unable to accommodate such a scenario since it removes the cross-correlations by restricting the observation impact. VM removes the correlations through modification of the background error directly.

Both forms have their advantages and disadvantages and should be considered in different situations. A careful evaluation needs to be performed before applying any form of variable localization to ensure that beneficial, physical correlations do not exist between the variables that are to be separated that were not accounted for. The benefit of removing the spurious correlations should be sizable since variable localization comes with a computational cost. The VO forms are generally less expensive than their VM counterparts. For the EnSRF, VO has little additional cost since rows of $\mathbf{K}$ are being zeroed out. VM has a modest additional cost since the computation of $\mathbf{K}$ scales by the size of the ensemble (Tippett et al., 2003) and the size of the ensemble is effectively doubled. The LETKF, however, scales by the quadratic of the ensemble size (Hunt et al., 2007). The computation of two sets of weights approximately doubles the computational cost, but the extension of the matrices to double the size of the ensemble increases the computational expense four-fold. For VO in the EnVar, solving two separate cost functions doubles the computational cost. For VM, the control vector is being extended to include an

additional set of weights, increasing its size by the number of grid points, which constitutes a doubling. The background covariance matrix is being increased by a factor of four, though since the matrix is not explicitly calculated and contains identical sub-blocks, the increase in computation is not as large.

These forms of variable localization may not be appropriate for many current applications within NWP. However, for applications such as the one presented in K11, variable localization is necessary. There is an increasing amount of work done in the assimilation of chemistry components, such as trace gases (Coman et al., 2012; Liu et al., 2012) and aerosols (Pagowski and Grell, 2012; Schwartz et al., 2014), where variable localization may be needed. In coupled ocean-atmospheric modeling, there has been a shift away from weakly coupled data assimilation, where the individual components exchange information in the forecast but are analyzed separately (Doblas-Reyes et al., 2011; Ham et al., 2014; Hazeleger et al., 2013; Robson et al., 2012), towards strongly coupled data assimilation, where all observations from each system are considered in one analysis solution (Han et al., 2013; Liu et al., 2013; Sluka et al., 2016). With coupled systems, there will very likely be variables that are sensitive to the noisy correlations that will presumably be present across system boundaries and would benefit from variable localization. Along with the current prevalence of ensemble techniques, variable localization may become more valuable in the future than it has been.

Table 5.1: Analysis increments from each of the data assimilation schemes and variable localization forms.

Data Assimilation Scheme	Variable Localization Form	Analysis Increment
EnSRF	None	$\delta\mathbf{x} = [(\rho_{SM} \circ \mathbf{X}\mathbf{X}^T)\mathbf{H}^T][\mathbf{H}(\rho_{SM} \circ \mathbf{X}\mathbf{X}^T)\mathbf{H}^T + R]^{-1}d$ $\delta\mathbf{x} = (\rho_{SO} \circ \mathbf{X}\mathbf{Y}^T)(\mathbf{Y}\mathbf{Y}^T + R)^{-1}d$
EnSRF	VO	$\delta\mathbf{x} = (\rho_{SO} \circ \rho_{VO} \circ \mathbf{X}\mathbf{Y}^T)(\mathbf{Y}\mathbf{Y}^T + R)^{-1}d$
EnSRF	VM	$\delta\mathbf{x} = [(\rho_{SM} \circ \rho_{VM} \circ \mathbf{X}\mathbf{X}^T)\mathbf{H}^T][\mathbf{H}(\rho_{SM} \circ \rho_{VM} \circ \mathbf{X}\mathbf{X}^T)\mathbf{H}^T + R]^{-1}d$ $\delta\mathbf{x} = (\rho_{SO} \circ \tilde{\mathbf{X}}\tilde{\mathbf{Y}}^T)(\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^T + R)^{-1}d$
LETKF	None	$\delta\mathbf{x} = \mathbf{X}(\mathbf{I} + \mathbf{Y}^T(\rho_{SR} \circ \mathbf{R})^{-1}\mathbf{Y})^{-1}\mathbf{Y}^T(\rho_{SR} \circ \mathbf{R})^{-1}\mathbf{d}$
LETKF	VO	$\delta\mathbf{x}_i = \mathbf{X}(\mathbf{I} + \mathbf{Y}_{iO}^T(\rho_{SRi} \circ \mathbf{R}_i)^{-1}\mathbf{Y}_{iO})^{-1}\mathbf{Y}_{iO}^T(\rho_{SRi} \circ \ \mathbf{R}_i)^{-1}\mathbf{d}_i$
LETKF	VM	$\delta\mathbf{x} = \tilde{\mathbf{X}}(\mathbf{I} + \tilde{\mathbf{Y}}^T(\rho_{SR} \circ \mathbf{R})^{-1}\tilde{\mathbf{Y}})^{-1}\tilde{\mathbf{Y}}^T(\rho_{SR} \circ \mathbf{R})^{-1}\mathbf{d}$
EnVar	None	$\delta\mathbf{x} = \mathbf{E}(\mathbf{L}^{-1} + \mathbf{E}^T\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\mathbf{E})^{-1}\mathbf{E}^T\mathbf{H}^T\mathbf{R}^{-1}\mathbf{d}$
EnVar	VO	$\delta\mathbf{x}_i = \mathbf{E}(\mathbf{L}_i^{-1} + \mathbf{E}^T\mathbf{H}_{iO}^T\mathbf{R}_i^{-1}\mathbf{H}_{iO}\mathbf{E})^{-1}\mathbf{E}^T\mathbf{H}_{iO}^T\mathbf{R}_i^{-1}\mathbf{d}_i$
EnVar	VM	$\delta\mathbf{x} = \tilde{\mathbf{E}}(\tilde{\mathbf{L}}^{-1} + \tilde{\mathbf{E}}^T\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\tilde{\mathbf{E}})^{-1}\tilde{\mathbf{E}}^T\mathbf{H}^T\mathbf{R}^{-1}\mathbf{d}$

Chapter 6: Summary and Future Directions

6.1 Summary

Ensemble-derived background error covariance matrices are often undersampled for NWP applications since the ensemble size is limited by computing resources. The method of localization was devised in order to eliminate correlations that are dominated by sampling error, thereby improving the forecast. Localization is commonly applied in the spatial dimensions, removing correlations for points that are distant whose true correlations are likely small. Spatial localization can be applied to the background error or the observation error. However, both forms of spatial localization can result in imbalances by disrupting physical relationships based on gradients or column integrated quantities. Imbalances can degrade the forecast by producing fast moving gravity waves within the model.

A method to improve balance within ensemble data assimilation methods was explored in Chapters 3 and 4. By applying a balance operator to ensemble perturbations, the localization is performed on the unbalanced correlations only. This prevents the localization from acting on the balanced correlations. This method was implemented within two ensemble data assimilation schemes: a hybrid 4DEnVar (Chapter 3) and an LETKF (Chapter 4). It was found that the type of spatial

localization, whether applied to the background error or the observation error, impacts the effectiveness of the balance operator. The hybrid 4DEnVar employs the model space localization, which allows for a two-way propagation of balanced information outside of the localization radius, i.e. $\delta\psi$ can impact δT and vice versa. The LETKF applies the spatial localization to the observation error, which only allows a one-way propagation of balanced information, i.e. $\delta\psi$ can impact δT but δT cannot impact $\delta\psi$.

Observing system simulation experiments were performed in the SPEEDY model, whose climatology and biases were reviewed in Chapter 2. The lower resolution forecast model has a stratosphere that is significantly damped compared with the higher resolution truth, though the large scale climatological features matched well with reanalysis data. When the balance operator was added within the hybrid 4DEnVar, the analysis and forecast skill were improved in the troposphere but degraded for the stratospheric winds where the model bias dominates. This degradation is due to the regression coefficients of the balance operator being based on the low resolution model rather than the high resolution truth. When the balance operator was included in the LETKF, the analysis and short term forecasts were degraded for most variables and locations. Since the localization prevents a two-way communication of information, δT is adjusted to be brought into balance with $\delta\psi$ after the analysis is found, moving it away from the observations. The form of spatial localization used has a significant impact on the effectiveness of the balance operator.

Localization can also remove correlations between variable types, termed vari-

able localization. Chapter 5 presented a unified framework, describing two forms that mirror the two types of spatial localization: observation space variable localization (VO) or model space variable localization (VM). VO and VM were formulated for three data assimilation schemes: EnSRF, LETKF, and EnVar. The relative strengths and weaknesses of VO and VM were consistent throughout the three schemes. VO removes the multivariate correlations by restricting an observation's impact to only certain control variables. While computationally inexpensive, a single observation cannot impact multiple control variables whose correlations are to be removed. VM removes the cross-correlations from the background error directly. It is more computationally expensive but requires no knowledge of the observations. In VM, a single observation can impact multiple control variables while their cross-correlations are also removed.

6.2 Future Directions

In Chapter 3, the addition of the balance operator to the ensemble part of the hybrid 4DEnVar produced predominantly positive results using an intermediate complexity model, SPEEDY. To test whether these results translate to more complex models, this method will be applied within NCEP's GSI for use within the GFS. The system used in the OSSEs was designed to mimic the GSI to allow for a simpler transition. One difference between the systems is the computation of the ensemble perturbations. The experiments from Chapter 3 use an LETKF to generate the perturbations and the GSI uses an EnSRF operationally, though it has the option

to use LETKF. Since both systems recenter the ensemble perturbations about the EnVar deterministic analysis, this difference should be of little consequence.

NWP models continue to increase in horizontal and vertical resolution as computational resources also increase. It is unlikely that the ensemble will be large enough to eliminate the need of localization in a majority of applications. Therefore, the impact of the spatial localization on the balance should continue to be studied as well as ways to alter the implementation of localization to improve balance. Because of computational costs, many EnKFs within the NWP community use observation space spatial localization rather than model space (Anderson, 2001; Whitaker et al., 2008). There are additional disadvantages that have not been addressed here, for instance, how to define observation locations for radiance data (Campbell et al., 2010). Model space localization would be more desirable if the computational cost was reduced. Bishop (2017) explores a variant on the ETKF to increase the size of the ensemble while including model space localization as a consequence.

The use of EnKFs within storm scale models is increasing. As global models also approach convection allowing resolutions, the implications for large scale balance need to be addressed. The same control variables may no longer be appropriate. In the Met Office Unified Model, Vetra-Carvalho et al. (2012) found that hydrostatic balance breaks down at 1.5 km horizontal resolution. Within a 4DVar, Honda et al. (2005) separate the control variables into synoptic scale and mesoscale, computing the balanced component from the synoptic scale only. Li et al. (2015) construct a multiscale 3DVar, applying a form of variable localization between the

large scales and the small scales. Spatial localization can also be applied at different scales, as in Buehner (2012). This allows the larger scales, which are associated with geostrophic balance, to have a much larger localization radius than the smaller scales, which likely would improved balance throughout the system.

Bannister (2008) notes that streamfunction is not an ideal choice for the primary variable within a balance operator due to geostrophic adjustment. Other formulations of balance operators can be explored in the ensemble data assimilation context. The inclusion of an independent balance variable could allow for the background covariance to be more symmetric in the observation space spatial localization case. Variational methods have extensively explored different control variable options. The use of potential vorticity has been explored within 4DVar (Cullen, 2003) and would allow for an unbalanced streamfunction variable (Bannister, 2008). Also within 4DVar, Fisher (2003) computed a balanced vertical wind through use of the quasi-geostrophic omega equation. How different forms of balance operators interact with localization requires further exploration.

Appendix A: Instability in the SPEEDY Model

Prior to the experiments conducted in Chapters 3 and 4, several preliminary experiments were run using an older version of the SPEEDY model (Molteni, 2003) with different observation and data assimilation configurations. The older version of the model is single resolution and only has seven vertical levels. A simpler observation configuration of simulated radiosondes was assimilated within 3DVar, which contained a different formulation for the balance operator than (3.1a) - (3.1c). These experiments exhibited a large amount of instability, frequently terminating integration. This instability, caused by regularly spaced observations and noise in the background error, manifest in a standing wave checkered pattern in the tropospheric temperature. The results from these experiments guided the configuration for the experiments within Chapters 3 and 4. This appendix describes the details of the configuration of the previous experiments and their results, previously presented in Sabol (2011).

A.1 Formulation

Rather than the utilizing the Hybrid 4DEnVar scheme described in Chapter 3, the experiments that follow use 3DVar, the static covariance only predecessor. The

basis for the Hybrid 4DEnVar, this scheme iteratively minimizes a cost function to find the optimal state considering both the observations and the previous forecast:

$$J(\mathbf{v}) = \frac{1}{2}\mathbf{v}^T\mathbf{v} + \frac{1}{2}(\mathbf{d} - \mathbf{HUv})^T\mathbf{R}^{-1}(\mathbf{d} - \mathbf{HUv}), \tag{A.1a}$$

$$\nabla J(\mathbf{v}) = \mathbf{v} - \mathbf{U}^T\mathbf{H}^T\mathbf{R}^{-1}(\mathbf{d} - \mathbf{HUv}). \tag{A.1b}$$

In this formulation, the control variable ($\mathbf{v}$) is preconditioned on the square root of the background error covariance ($\mathbf{B} = \mathbf{UU}^T$). The control variable utilizes a difference variable set than the previous experiments, substituting the zonal and meridional wind (u and v respectively) for streamfunction and velocity potential.

In our current 3DVar formulation, different types of observations are assimilated independently of one another; the formulation does not recognize relationships among the prognostic variables (u, v, T, q, and p_s). Analysis increments can be dynamically inconsistent and result in imbalances in the next forecast cycle. Therefore, a dynamic constraint is added to our 3DVar system to represent the intervariable correlations, different in formulation from the balance operator described in Section 3.2.1. One of the strongest of the intervariable relationships in the atmosphere concerning our prognostic variables is geostrophic balance.

Using geostrophic balance, the horizontal velocity, $\mathbf{V}$, is expressed as:

$$\mathbf{V} = r\mathbf{V}^g + \mathbf{V}^u, \tag{A.2}$$

where $\mathbf{V}^g$ is the vector geostrophic wind that is in balance with T and p_s:

$$f\mathbf{k} \times \mathbf{V}^g = -RT\nabla ln(p_s) - \nabla\phi(T). \tag{A.3}$$

f is the Coriolis parameter, R is the gas constant for dry air, and ϕ is geopotential (related to T through hypsometric relations), $\mathbf{V}^u$ is the unbalanced, residual velocity, and r is the regression coefficient chosen so that $\mathbf{V}^g$ and $\mathbf{V}^u$ are statistically uncorrelated. Mathematically, r is represented as:

$$r = \frac{E[(\epsilon)(\epsilon^g)^T]}{E[(\epsilon^g)(\epsilon^g)^T]}, \tag{A.4}$$

where ϵ denotes the error of the total wind and ϵ^g denotes the error of the geostrophic wind. This coefficient represents the strength of geostrophic balance, where a value of $r = 1$ implies that the wind is fully geostrophic. In practice using the NMC method (Parrish and Derber, 1992), ϵ is the difference between the 18 and 24 hour forecasts of the full wind verifying at the same time and ϵ^g is the difference between the 18 and 24 hour forecasts of the geostrophic wind verifying at the same time. Computed for each latitude and vertical level using a year's worth of samples, it is an expression of how geostrophic the wind at a location is. High values of r occur in the midlatitudes, where geostrophic balance is strong, and low values occur in the tropics, where the Coriolis effect is weak (Figure A.1a). The correlation between $\mathbf{V}^g$ and $\mathbf{V}^u$ is extremely low as shown in Figure A.1b for the zonal average of the correlation with height.

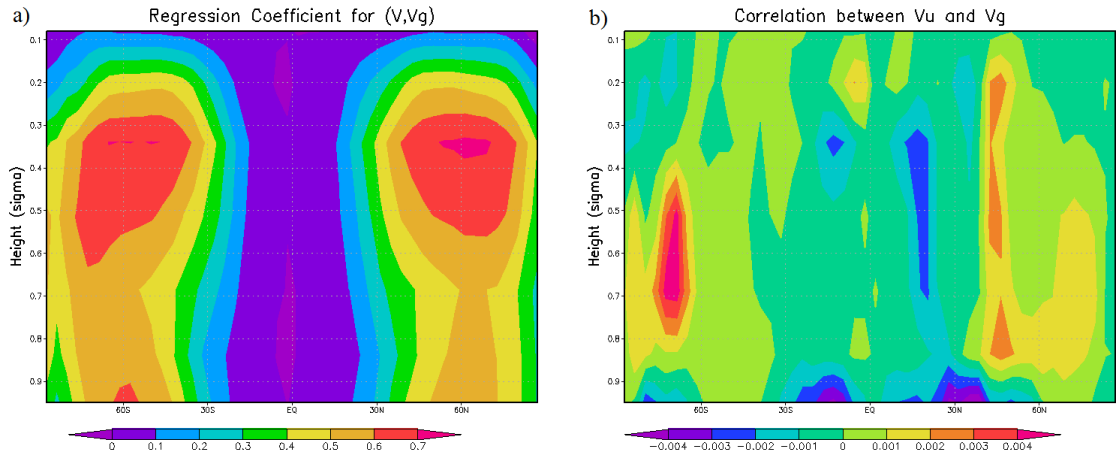

Figure A.1: (a) The linear regression coefficient, r, between the total wind and the geostrophic wind, shown with latitude along the x-axis and σ levels along the y-axis. (b) The correlation between the unbalanced wind and the geostrophic wind, by latitude and height.

To use the geostrophic constraint in the incremental 3DVar with preconditioning, the control variable is transformed to $\delta\mathbf{z}$:

$$\delta\mathbf{x} = \mathbf{G}\mathbf{U}_z\delta\mathbf{z}, \tag{A.5}$$

where $(\mathbf{U}_z\delta\mathbf{z})^T = [(\delta u^u)^T, (\delta v^u)^T, (\delta T)^T, (\delta q)^T, (\delta p_s)^T]$ and:

$$\mathbf{G} = \begin{bmatrix} 1 & 0 & -\frac{rR}{f}\mathbf{\Lambda}_y ln(p_s) - \frac{r}{f}\mathbf{\Lambda}_y\mathbf{C} & 0 & -\frac{rRT^b}{f}\mathbf{\Lambda}_y\left(\frac{1}{p_s}\right) \\ 0 & 1 & \frac{rR}{f}\mathbf{\Lambda}_x ln(p_s) + \frac{r}{f}\mathbf{\Lambda}_x\mathbf{C} & 0 & \frac{rRT^b}{f}\mathbf{\Lambda}_x\left(\frac{1}{p_s}\right) \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \tag{A.6}$$

$\mathbf{\Lambda}$ represents the centered-difference first order derivative in space and $\mathbf{C}$ represents the transformation from temperature to geopotential using the hypsometric equa-

tion.

Incorporating the geostrophy constraint into the cost function (A.1a) results in:

$$J(\delta\mathbf{z}) = \frac{1}{2}\delta\mathbf{z}^T\delta\mathbf{z} + \frac{1}{2}(\mathbf{d} - \mathbf{HGU}_z\delta\mathbf{z})^T\mathbf{R}^{-1}(\mathbf{d} - \mathbf{HGU}_z\delta\mathbf{z}), \tag{A.7a}$$

$$\nabla J(\delta\mathbf{z}) = \delta\mathbf{z} - \mathbf{U}_z^T\mathbf{G}^T\mathbf{H}^T\mathbf{R}^{-1}(\mathbf{d} - \mathbf{HGU}_z\delta\mathbf{z}). \tag{A.7b}$$

The background error variances based on the control variable becomes:

$$\mathbf{B} = \mathbf{GB}_z\mathbf{G}^T = \mathbf{GU}_z\mathbf{U}_z^T\mathbf{G}^T. \tag{A.8}$$

A.2 System Description

A.2.1 Model Description

The experiments that follow use an older version of the SPEEDY model compared to the version described in Chapter 2. One of the most significant differences is the vertical resolution. Rather than eight vertical levels, this version of SPEEDY has seven vertical levels, with the bottom six levels being identical and one less level in the stratosphere (σ=0.95, 0.835, 0.685, 0.51, 0.34, 0.2, 0.08). This version of the model also does not contain a high resolution configuration. There is only one resolution, T30, which corresponds to a standard Gaussian grid of 96 grid points zonally and 48 meridionally ($3.75^\circ \times 3.75^\circ$). This results in the truth being at the same resolution as the analysis and forecast steps. With no other model error, these

experiments are considered identical twin experiments rather than fraternal twin experiments.

A.2.2 Observation Network

The observation network configuration is different than described in Section 2.2, though the method of observation generation is the same: adding a Gaussian random error to the true state, scaled by the observation error associated with each observation type. There is no satellite component to the network. The network is comprised of simulated radiosondes only, which contain the observation types and errors as the experiments in Chapters 3 and 4 (Table 2.1).

There are three configurations for radiosonde-like observation locations used in these experiments, designed by Miyoshi (2005). First is the dense observation network (Figure A.2a). It consists of regular observational coverage at every other grid point, one out of every four surrounding grid points, for a total of 1056 stations. The second network is the sparse network (Figure A.2b). It is also regularly distributed, though the observations occur at every fourth grid point, or one out of 16 surrounding grid points. This makes a total of 264 observing stations. The third network is the realistic radiosonde network (Figure A.2c). The observations are unevenly distributed, with more observations occurring over the land than over the ocean, and more observations in the northern hemisphere than in the southern hemisphere. The highest concentration is over Europe, Asia, and the United States, with the lowest concentration in the Southern Pacific Ocean basin. There are 415

stations in this configuration.

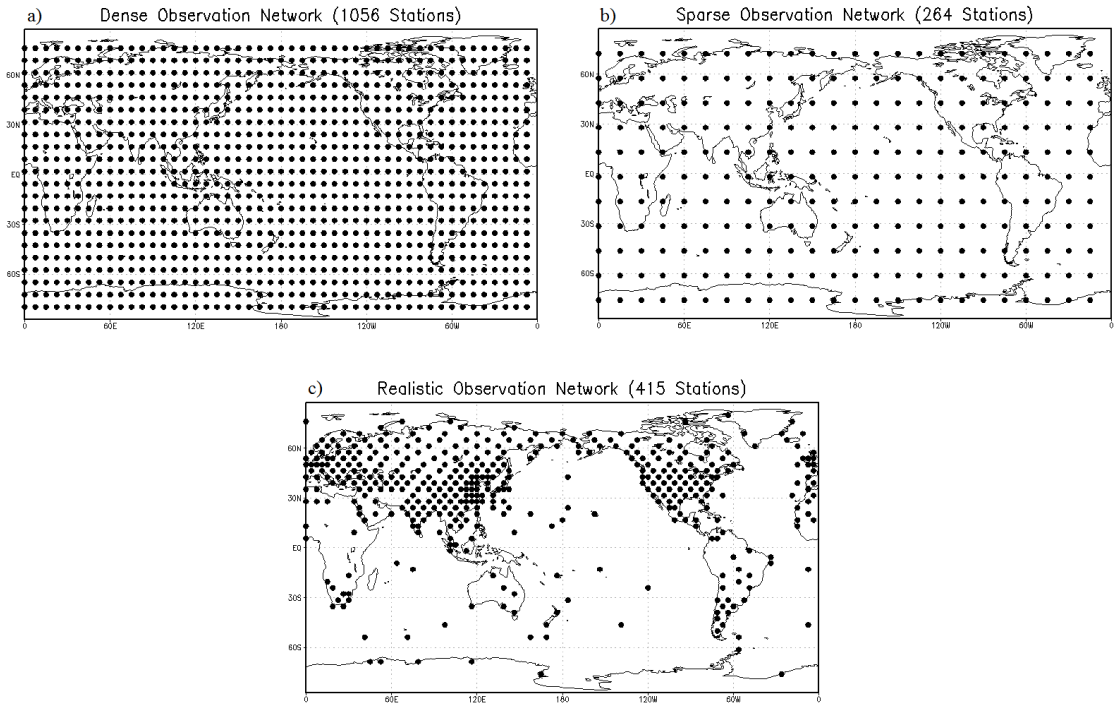

Figure A.2: Observation networks used in the SPEEDY experiments. (a) Dense network, (b) Sparse network, (c) Realistic network

A.2.3 Experimental Setup

Several experiments were run with different observation configurations, different background error variances, and with and without the geostrophic constraint. All experiments begin on January 1, 1982 with observations assimilated every six hours. The initial background is chosen from the truth at a different time in the integration: February 1, 1982. All experiments also use the same observation error covariance matrix, $\mathbf{R}$, which is a diagonal matrix with no correlation between observations. A summary of all experiments is provided in Table A.1.

The first set of experiments have a configuration that closely matches the ex-

periments of Miyoshi (2005), where each experiment uses a different observation network, but all three use the same background error variance and do not use the geostrophic constraint. The background error variance was derived as in those experiments: using the NMC method on two months of experiments with the dense observation network. Finding that the background error contained noise due to the small sample size, an additional experiment was run for the dense network case with the background variances zonally averaged. Another set of experiment was then conducted, increasing the sample size of lagged forecast pairs to one year for each observing network.

Three more experiments were conducted, this time with the geostrophic constraint incorporated. The background error for each observing network used a year's worth of samples and no horizontal smoothing was applied. Two additional experiments were conducted for the dense network case only with zonal and horizontal smoothing.

Table A.1 contains a summary of the described experiments, including whether it used the geostrophic constraint, which observing network was used, how $\mathbf{B}$ was computed, the length of the experiment, and root-mean-squared error (RMSE) statistics. These statistics were computed from the analysis compared to the nature run for temperature at the middle model level, $\sigma = 0.51$. RMSEs were calculated globally, north of 20°N, and south of 20°S.

A.3 Results

A.3.1 Without the Geostrophic Constraint

The analysis RMSEs for the three experiments using the $\mathbf{B}$ calculated with a two month sample are shown in Figure A.3. After approximately six years, the dense network case increases in error rapidly, eventually causing the model to terminate. The exact nature of these errors will be addressed in greater detail in Section A.3.2. To summarize, they are related to the noisiness of the background error variance and the regularity of the observations. The relation to observation regularity can be seen in the success of the realistic observation case, which completed a 15-year integration. This case exhibits an annual cycle in the analysis RMSE, with higher accuracy in January (boreal winter) and lower accuracy in July (austral winter). This is due to the higher amount of observations in the northern hemisphere, allowing the dynamic instabilities associated with the winter season to be captured in the northern hemisphere more than in the southern hemisphere.

Using a small sample size of only two months, the background error contains noise that could contaminate the analysis and build over time. Zonally averaging the background error for use in the dense network case remedies the situation of instability in Figure A.3 in the long-term. In fact, a 50-year integration from 1950-1999 was completed and the analysis remains stable throughout the time period (not shown).

Figure A.4 shows successful 15-year integrations for the three experiments us-

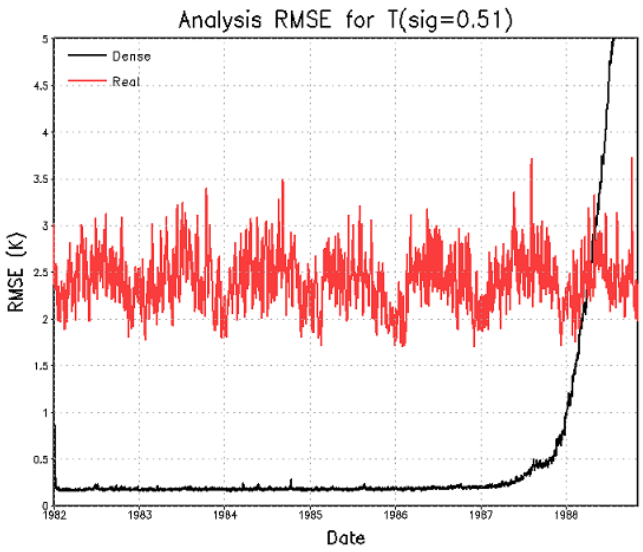

Figure A.3: Analysis RMSE for the midlevel T (in K) for seven-year integrations without the constraint for the dense network (black) and the realistic network (red). Both use the background error from the dense NMC case with two months of samples.

ing background errors calculated from one year of samples from their own observing networks. The realistic and sparse network cases use the unsmoothed background errors calculated from a year of samples from their respective networks. The dense network case uses the zonally smoothed background error used for the previous experiment for stability. There is improvement in the realistic and the sparse network cases over the initial integrations using the dense network background error (not shown). The dense network still greatly outperforms the other cases, as expected, with RMSEs of about 10% of the other two networks.

A.3.2 With the Geostrophic Constraint

Three experiments were performed, one for each observing network, with the geostrophic constraint included. Examining Figure A.5, the system performs well early in the experiment period, reducing the error over the first couple of weeks.

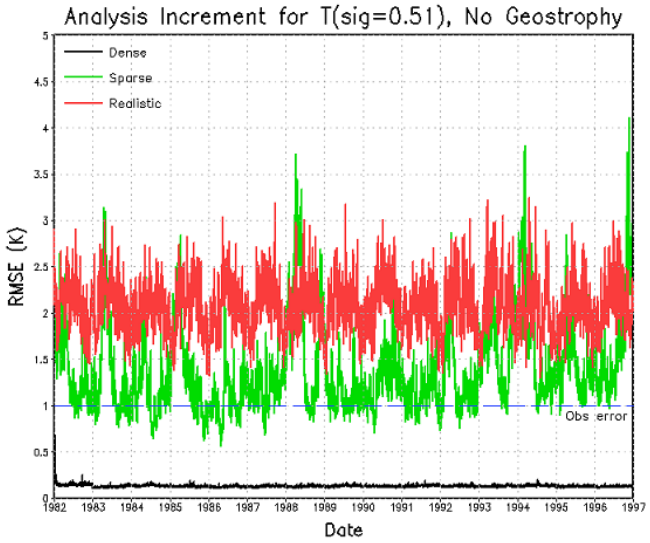

Figure A.4: The analysis RMSE for midlevel T (in K) for 15-year integrations without the constraint. The dense network (black), the sparse network (green), and the realistic network (red) use their own $\mathbf{B}$ with additional smoothing for the dense network.

Shortly after, the errors in the dense and sparse cases begin to increase as they did in the dense non-geostrophic case, except this time it occurs much sooner in the integration, on the order of months instead of years. As in the non-geostrophic case, the realistic network case is stable and is able to complete a 15-year integration (not shown).

For the remainder of this section, we identify the cause and nature of the increase in errors in the dense network case. When no geostrophic constraint is used, a smoothing of the variances resolved the problem of long term instability (Figure A.4). Zonal and horizontal smoothing are applied to the background error variances for the dense network, but they do not keep the errors from increasing; they only delay the response (Figure A.6).

Smoothing the errors spatially does not keep the model integration stable. If the magnitude of $\mathbf{B}$ is too small (Figure A.7a), the background error becomes much

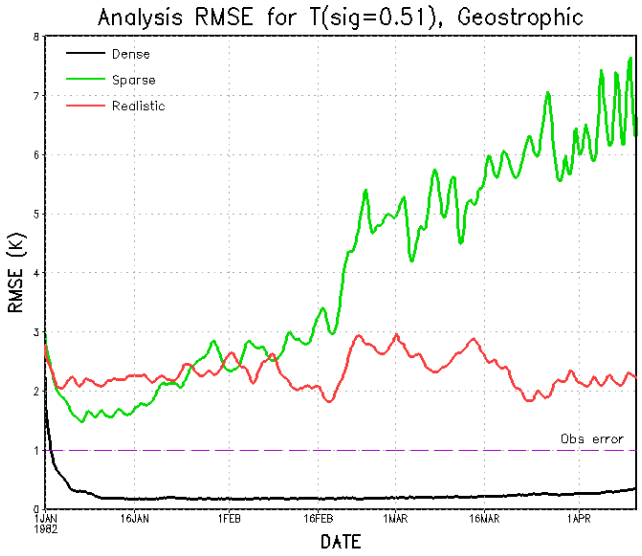

Figure A.5: Analysis RMSE for midlevel T (in K) using the geostrophic constraint. Results are shown for the dense (black), sparse (green), and realistic (red) networks.

smaller than the observation error and the background does not draw closely to the observations. As a result, the analysis can diverge from the true state, or not reach a state close to the truth at all. If the magnitude of $\mathbf{B}$ is too large, the background draws extremely close to the observations, possibly making the analysis noisier and contributing to the previous long-term instability. Figure A.7b shows that a very small change in the scaling parameter can alter the stability and accuracy of the assimilation. With a scaling factor of 0.48 being too large and 0.45 being too small, it seems that the analysis is very sensitive to the magnitude of $\mathbf{B}$.

In addition to the impact of the errors in the T at the middle model level, the behavior of the RMSE at other model levels for other variables is investigated as well as to ensure that these errors are not limited to the particular variable of choice. For the dense network case using the one year, unsmoothed background error, the RMSE for T and the zonal wind, u, are shown in Figure A.8 with height over time.

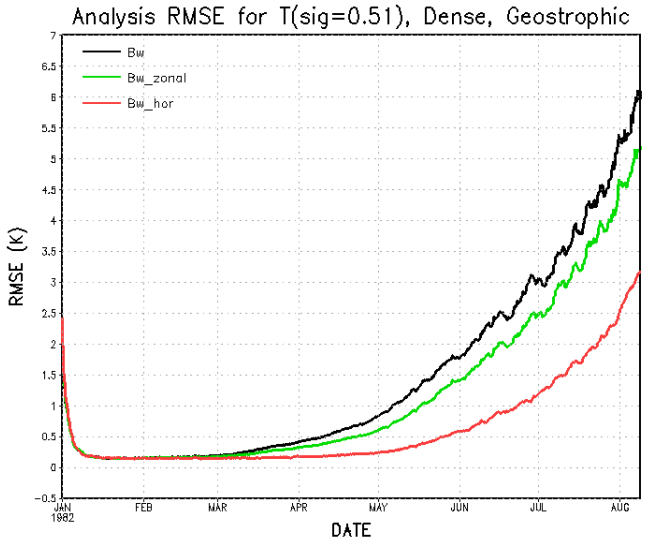

Figure A.6: Analysis RMSE for midlevel T (in K) using the geostrophic constraint for the dense network. Background errors are from the NMC method without smoothing (black), using a zonal mean (green), and using a horizontal mean (red).

The errors for both T and u increase over time for all vertical levels with the highest errors occurring in the upper troposphere.

For our standard middle model level of σ=0.51, the analysis of the last time step before model failure (Figure A.9a) is compared with the truth (Figure A.9b). The analysis for temperature at the middle level is extremely noisy. In fact, there are large amplitude standing waves in the analysis producing a lattice-like, checkered pattern. The crests and troughs of each wave occur between the observation locations. Also, the analysis is warmer than the true state, which shows the horizontal mean for the temperature bias at the same level over time (Figure A.10).

The extra energy for the warming can come from either of two places: the analysis increment or the forecast. The bias is examined in each of these steps, analysis minus background for the former and background minus analysis at the previous time step for the latter (Figure A.11). The analysis increment has zero

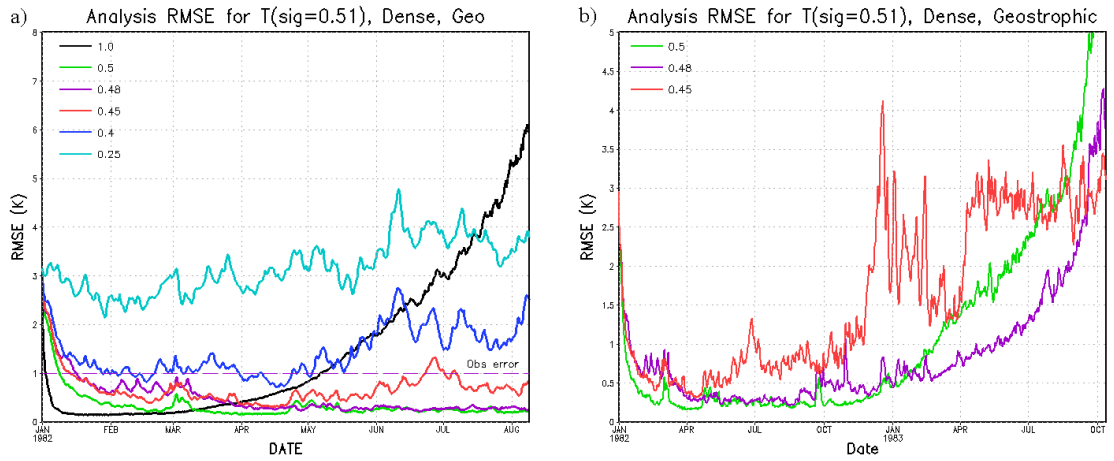

Figure A.7: Analysis RMSE for midlevel T (in K) using the geostrophic constraint for the dense network. Background errors are from the NMC method scaled by the factor indicated. (a) 7-month integration and (b) 22-month integration.

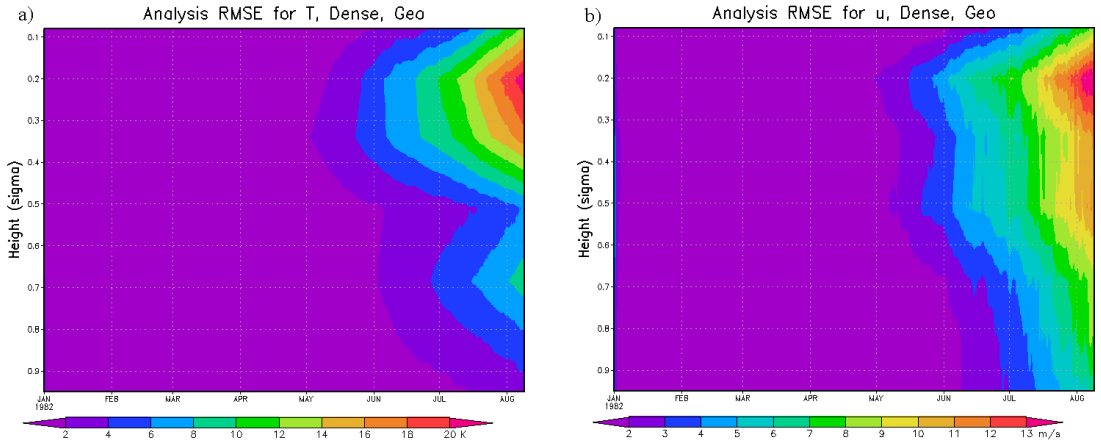

Figure A.8: Analysis RMSE for (a) T and (b) u with height over time. This analysis uses the dense observing network and an unsmoothed background error.

bias for the level and variable of interest for the first few months (Figure A.11a). The mean increment then begins to decrease rapidly. A mean negative increment implies that the background is too warm and that the analysis is trying to cool it down. This means that the extra energy is not coming from the analysis increment. When examining Figure A.11b, the bias in the forecast begins to increase after approximately six months, meaning that the mean temperature increases during

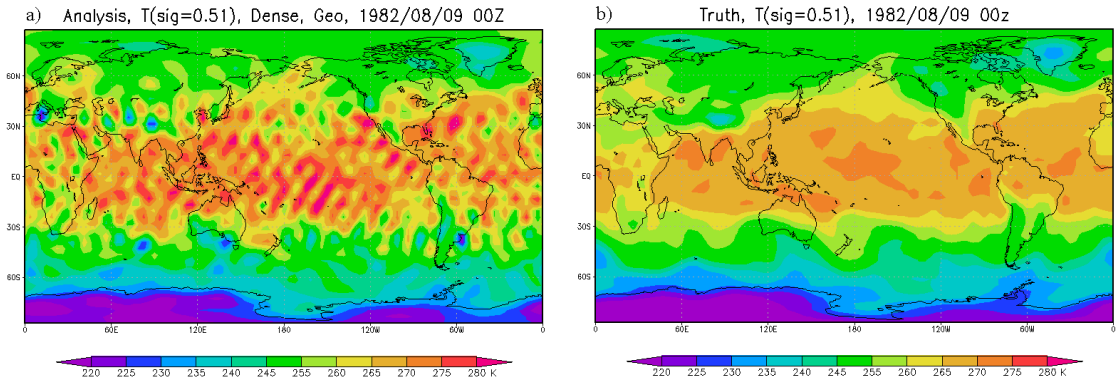

Figure A.9: The (a) analysis and (b) truth for midlevel T (in K) on 1982/08/09 00z, the last analysis cycle before the model failure. The analysis is computed using the dense network and geostrophic constraint, with no smoothing for the background error.

the 6-hour forecast. The extra energy, therefore, comes from the forecast step.

Although there is not a clear bias in the first few months and the errors do not start to increase until around May 1982, the checkered pattern is present in the mean analysis bias (analysis - truth) from January 15th to February 15th, 1982 (Figure A.12a). Similar to Figure A.9, the peaks are occurring off of the observation location, which is logical since the points that have observations should see the highest accuracy. The mean bias of the checkered pattern is zero, which is why there is no indication in the horizontal mean statistics. This implies that the increase in temperature is not the source of the problem, but the checkered pattern precedes the temperature increase. When we average from July 1st to August 1st, 1982, a higher amplitude bias has developed, though the pattern remains the same (not shown).

The observations being located exactly in between the troughs and crest of the waves, along with the success of the realistic observation network, implies that

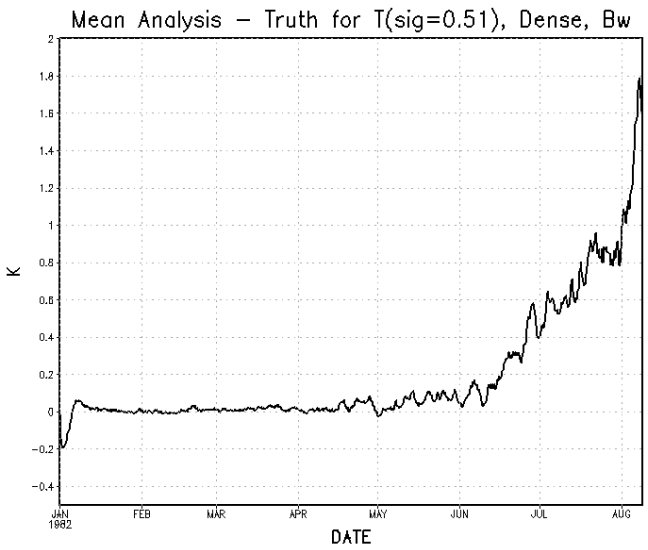

Figure A.10: The mean analysis bias for T at $\sigma = 0.51$ using the same configuration as Figure A.9

the regularity of the observations is involved in the long-term instability, perhaps through the generation of a resonating standing wave. There is no mean bias in either the analysis increment or the forecast for the realistic case (Figure A.13). However, if we examine the spatial distribution of the mean analysis error bias for the beginning of the assimilation as we did in Figure A.12, a weak checkered pattern is present in certain locations, particularly those with high density, regularly spaced observations (Europe, Asia, and North America; Figure A.14). This checkered pattern does not seem to be of a high enough amplitude or widespread enough to dominate the analysis field, allowing it to remain stable over much longer time scales.

While the choice of observation network clearly plays a role in the long term stability, the choice of constraint does as well. As presented in Section A.3.1, the dense network case without the geostrophic constraint and with an unsmoothed

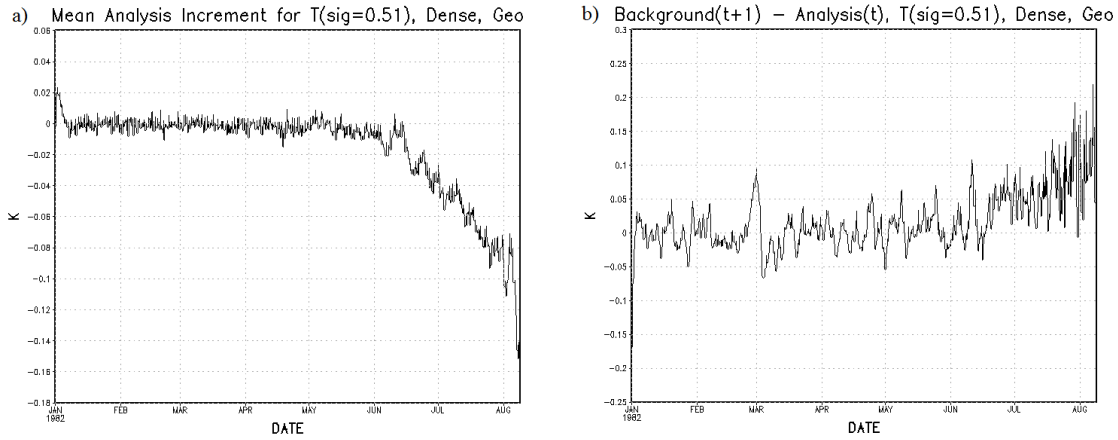

Figure A.11: The mean increment in temperature at the middle model level (in K) using the dense network and the geostrophic constraint with no smoothing in the background error. (a) The mean increment for analysis minus background and (b) the increment for the 6-hour forecast minus the analysis.

B experiences similar increases in error to what has been outlined in this section. However, without the constraint the analysis is stable over much longer time periods than when the constraint is used. This case can also be stabilized by smoothing the background error, unlike in the geostrophic constraint case. For the geostrophic constraint, not only are there increments at the observation point, there are analysis increments off of the observation point as well, producing a dipole about the observation. This difference in spatial distribution of increments could be responsible for exciting the standing waves in the analysis at a much quicker rate than in the non-geostrophic case.

A.4 Summary

A suite of 3DVar experiments were run with an intermediate complexity model, SPEEDY. Three different observation configurations were tested: a regularly spaced

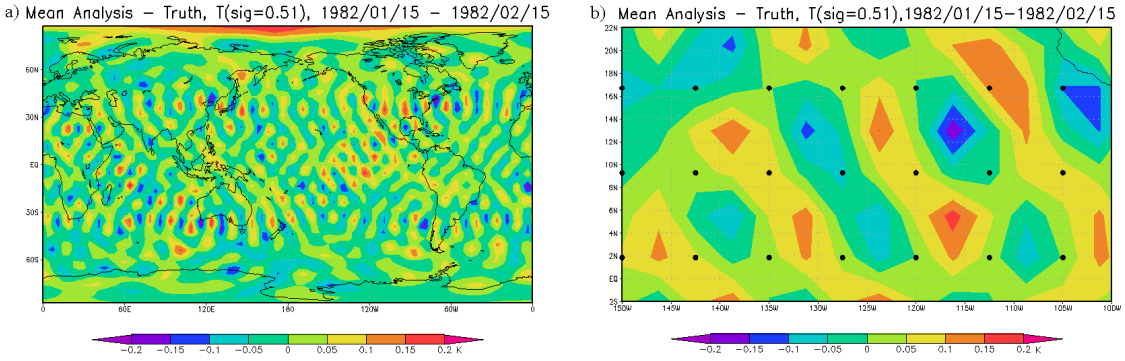

Figure A.12: (a) The mean temperature bias in space for the middle model level, calculated from January 15th to February 15th, 1982. (b) A close-up of (a) with the observation locations indicated.

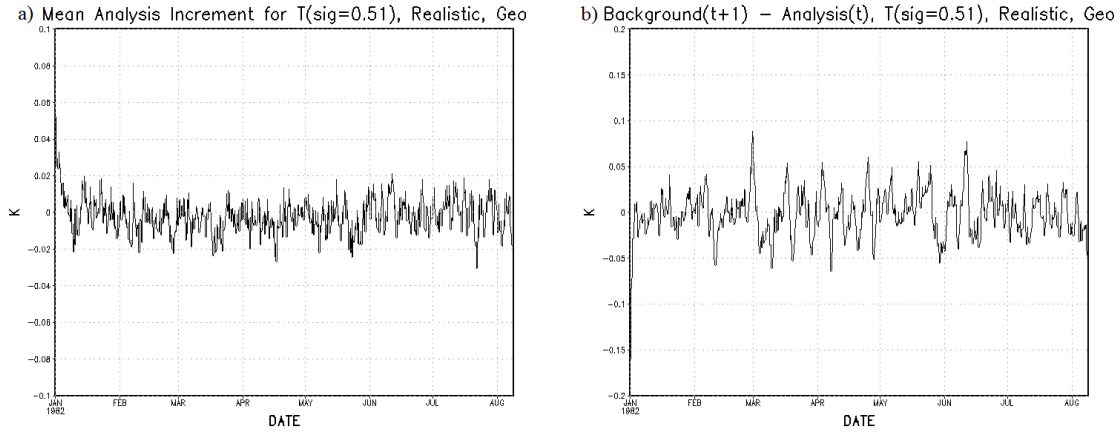

Figure A.13: Same as Figure A.11, but with the realistic observation network.

dense network, a regularly spaced sparse network, and an irregularly spaced realistic network. Both of the regularly spaced networks were unstable and frequently terminated integration. Without the geostrophic constraint, the dense network blew up after six years of cycling, though when the background error variances were smoothed, the integration was stable for 50 years. When the geostrophic constraint was applied, the instability accelerated and the model terminated after only eight months. The errors of the sparse network began to increase even more rapidly, within the first two months. It was determined that it was the forecast portion of

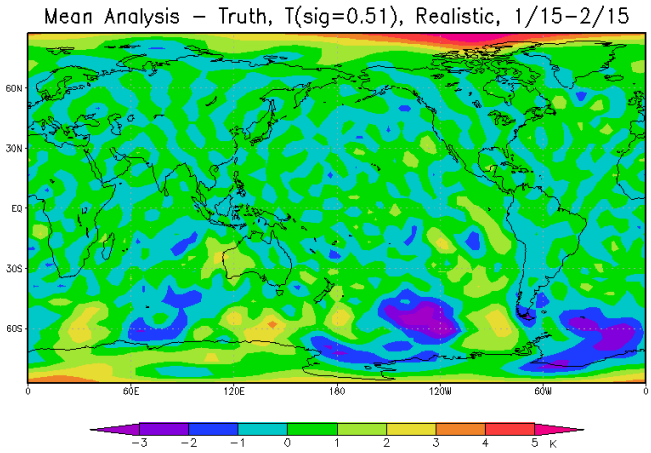

Figure A.14: Same as Figure A.12a, but with the realistic observation network.

the analysis cycle that was contributing to these errors. The model was increasing the temperature at all levels. A standing wave was created in the temperature field, with an amplitude exceeding 20 K. The previous method of smoothing the background error variances did not stabilize the integration in this case, only delayed it. The realistic network, however, showed no signs of instability throughout any of the configurations tested.

This set of SPEEDY experiments demonstrated several aspects of the system that allowed for more stable experiments moving forward. The experiments of this section were run by an older version of the SPEEDY model. Updating to a newer version of the model allowed for an increase in the number of vertical levels as well as additional options for higher resolution. It was determined that the regularly observing networks were unstable and should not be used in future experiments. The realistic network, however, was stable over 15 years of cycling, and this observation configuration was retained. The type of geostrophic constraint used in

these experiments was also abandoned. The balance operator used by NCEP was adopted (3.1a) - (3.1c), changing the control variables to be ψ and χ rather than u and v. Once these changes were in place, the T30 resolution experiments no longer exhibited the standing wave pattern or instability of the older experiments.

Table A.1: Summary of 3DVar-SPEEDY Experiments. RMSE calculations are for temperature at $\sigma = 0.51$ for 1982/01/10 - 1982/04/01. *Integration began on 1950/01/01.

Observation Network	Geostrophic Constraint	**B** Obs.	**B** Time	**B** Smoothing	Complete	End Date	Global RMSE	SH RMSE	NH RMSE
Dense	No	Dense	Two months	None	No	1988/11/27	0.174	0.176	0.171
Dense	No	Dense	Two months	Zonal	Yes	1997/01/01	0.174	0.176	0.173
Dense	No	Dense	Two months	Zonal	Yes	1999/01/01*	0.174	0.176	0.171
Dense	No	Dense	One year	Zonal	Yes	1983/01/01	0.176	0.154	0.192
Dense	Yes	Dense	One year	None	No	1982/08/09	0.164	0.163	0.166
Dense	Yes	Dense	One year	Zonal	No	1982/08/21	0.160	0.158	0.161
Dense	Yes	Dense	One year	Horizontal	No	1982/09/10	0.153	0.154	0.151
Sparse	No	Dense	Two months	None	Yes	1985/01/01	1.881	1.628	2.067
Sparse	No	Sparse	One year	None	Yes	1997/01/01	1.711	1.609	1.791
Sparse	Yes	Sparse	One year	None	No	1982/04/13	2.853	3.332	2.216
Realistic	No	Dense	Two months	None	Yes	1997/01/01	2.192	2.911	1.056
Realistic	No	Realistic	One year	None	Yes	1997/01/01	1.996	2.615	1.034
Realistic	Yes	Realistic	One year	None	Yes	1997/01/01	2.140	2.598	1.529

Appendix B: The Recursive Filter

The recursive filter F in (3.4) represents the spatial correlations on the model grid. Described by Purser et al. (2003), it smoothes an impulse into a quasi-Gaussian shape. A three-dimensional recursive filter is split up into vertical and horizontal correlations, where the horizontal correlations are further split up into the x-direction and the y-direction, i.e. $\mathsf{F} = \mathsf{F}_x\mathsf{F}_y\mathsf{F}_z$. For each spatial direction, there is a forward and backward component to the recursive filter:

$$\mathbf{b}_i = \beta\mathbf{a}_i + \sum_{j=1}^{n} \alpha_j \mathbf{b}_{i-j}, \tag{B.1a}$$

$$\mathbf{c}_i = \beta\mathbf{b}_i + \sum_{j=1}^{n} \alpha_j \mathbf{c}_{i+j}, \tag{B.1b}$$

where n is the order of the recursive filter, i is the grid point, treated in ascending order for the forward pass and in descending order for the backward pass, and α_j and β are coefficients that satisfy:

$$\beta = 1 - \sum_{j=1}^{n} \alpha_j. \tag{B.2}$$

The forward pass, (B.1a), smoothes in the positive direction, with $\mathbf{a}_i$'s as the input and $\mathbf{b}_i$'s as the output, and the backward pass, (B.1b), smoothes in the negative direction, with $\mathbf{b}_i$'s as the input and $\mathbf{c}$'s as the output. Thus, these equations can be represented as $\mathbf{b} = \mathsf{F}\mathbf{a}$ and $\mathbf{c} = \mathsf{F}^T\mathbf{b}$ with F and F^T as the forward and backward recursive filter. To determine the coefficients for the recursive filter, α_j and β, we follow Purser et al. (2003).

Let D represent a differential operator of order n:

$$D_{(n)} = 1 - \frac{\sigma^2\delta x^2}{2}\frac{\mathrm{d}^2}{\mathrm{d}x^2} + \frac{1}{2!}\left(\frac{\sigma^2\delta x^2}{2}\frac{\mathrm{d}^2}{\mathrm{d}x^2}\right)^2 + \cdots + \frac{1}{n!}\left(-\frac{\sigma^2\delta x^2}{2}\frac{\mathrm{d}^2}{\mathrm{d}x^2}\right)^n, \tag{B.3}$$

where δx is the uniform grid spacing and σ is the length scale with units of number of grid points. It has a spectral representation of:

$$\hat{D}_{(n)} = 1 - \sigma^2\left(\frac{k\delta x^2}{2}\right) + \frac{\sigma^4}{2!}\left(\frac{k^2\delta x^2}{2}\right)^2 + \cdots + \frac{\sigma^{2n}}{n!}\left(\frac{k^2\delta x^2}{2}\right)^n, \tag{B.4}$$

where k is the wavenumber. Let K represent a finite difference operator, which approximates a second order derivative of variable ψ:

$$\frac{\mathrm{d}^2}{\mathrm{d}x^2}\psi_i \approx \frac{\psi_{i-1} - 2\psi_i + \psi_{i+1}}{\delta x^2} = -\frac{K(\psi_i)}{\delta x^2}, \tag{B.5}$$

where i is the grid point index in the direction of x. This operator can be written in spectral form with respect to k:

$$\hat{K}(k) = \left[2\sin\left(\frac{k\delta x}{2}\right)\right]^2. \tag{B.6}$$

This relationship is inverted and expressed k in terms of the finite difference operator, using an expression for $\sin^{-1}$ as a series:

$$\sin^{-1} x = \sum_{i=0}^{\infty} \gamma_i x^{2i+1}, \tag{B.7}$$

with $\gamma_i = \frac{1}{(2i+1)} \frac{(2i-1)!!}{(2i)!!}$. Now, $(k^2 \delta x^2)$ can be expressed as a power series:

$$(k^2 \delta x^2) = \sum_{j \geq i} g_{i,j} \hat{K}^j. \tag{B.8}$$

The coefficients of $g_{i,j}$ are found in Table 1 of Purser et al. (2003), defined as $b_{i,j}$, for coefficients i and j up to 6. Substituting (B.8) into (B.4), expanding the series, and grouping the terms according to the power of K, the differential operator becomes:

$$\begin{aligned} D^{\star}_{(n)} = 1 + g_{1,1} \frac{\sigma^2}{2} K + \left[g_{1,2} \left(\frac{\sigma^2}{2} \right) + \frac{g_{2,2}}{2!} \left(\frac{\sigma^2}{2} \right)^2 \right] K^2 + \cdots \\ + \left[\sum_{j=1}^{n} \frac{g_{j,n}}{j!} \left(\frac{\sigma^2}{2} \right)^j \right] K^n. \end{aligned} \tag{B.9}$$

The 4DEnVar system in our experiments uses a fourth-order recursive filter. The differential operator of the fourth order is written as:

$$\begin{aligned} D_{(4)} = 1 + \frac{\sigma^2}{2} K + \left[\frac{\sigma^4}{8} + \frac{\sigma^2}{24} \right] K^2 + \left[\frac{\sigma^6}{48} + \frac{\sigma^4}{48} + \frac{\sigma^2}{180} \right] K^3 \\ + \left[\frac{\sigma^8}{384} + \frac{\sigma^6}{192} + \frac{7\sigma^4}{1920} + \frac{\sigma^2}{1120} \right] K^4. \end{aligned} \tag{B.10}$$

Since the differential operator can be written as a polynomial in K, its factorized

form is:

$$D_{(4)} = \prod_{j=1}^{4} \left(1 - \frac{K}{\kappa_j}\right). \tag{B.11}$$

By matching the coefficients in (B.10) and (B.11), the roots to the following fourth degree equation are found, which can also be factored as:

$$\begin{aligned} &ax^4 + bx^3 + cx^2 + dx + e \\ &\qquad = \left(x - \frac{1}{\kappa_1}\right)\left(x - \frac{1}{\kappa_2}\right)\left(x - \frac{1}{\kappa_3}\right)\left(x - \frac{1}{\kappa_4}\right) = 0, \end{aligned} \tag{B.12}$$

with the solutions being $(1/\kappa_j)$'s. By defining a shift operator, $Z\psi_i = \psi_{i+1}$ and $Z^{-1}\psi_i = \psi_{i-1}$, K is rewritten as $K = -Z^{-1}+2-Z$. Substituting into the polynomial factors of (B.11):

$$1 - \frac{K}{\kappa_j} = \frac{Z^{-1} - 2 + \kappa_j + Z}{\kappa_j} = \left(\frac{1 - \zeta_j Z^{-1}}{1 - \zeta_j}\right)\left(\frac{1 - \zeta_j Z}{1 - \zeta_j}\right), \tag{B.13}$$

where ζ is the smaller of the two solutions to the quadratic equation:

$$\zeta^2 - (2 - \kappa_j)\,\zeta + 1 = 0. \tag{B.14}$$

In the limit of the order of the differential operator to infinity, the inverse of the operator is a Gaussian function. The input signal, $\mathbf{a}$, is related to our output signal, $\mathbf{c}$, through this differential operator, $D_{(n)}$, with the objective of making the output signal Gaussian in shape:

$$D_{(n)}\mathbf{c} = \mathbf{a}. \tag{B.15}$$

$D_{(n)}$ is broken down using an LU decomposition into matrices $\mathbf{P}$ and $\mathbf{Q}$, which is used to separate the operation of (B.15) into two parts, including an intermediate signal, $\mathbf{b}$:

$$\mathbf{Pb} = \mathbf{a}, \tag{B.16a}$$

$$\mathbf{Qc} = \mathbf{b}. \tag{B.16b}$$

These two relations are the backward and forward iterations of the recursive filter, previously expressed in (B.1a) and (B.1b), respectively, where $\mathsf{P} = \mathsf{F}^{-1}$ and $\mathsf{Q} = \mathsf{F}^{-T}$. Using the expression from (B.11), the forward and backward recursive filter iterations are rewritten using the inverse of the differential operator:

$$\mathbf{b}_i = \prod_{j=1}^{4} \left(\frac{1-\zeta_j}{1-\zeta_j Z^{-1}} \right) \mathbf{a}_i, \tag{B.17a}$$

$$\mathbf{c}_i = \prod_{j=1}^{4} \left(\frac{1-\zeta_j}{1-\zeta_j Z} \right) \mathbf{b}_i. \tag{B.17b}$$

The coefficients α_j and β from (B.1a) and (B.1b) can also be determined:

$$\alpha_1 = \zeta_1 + \zeta_2 + \zeta_3 + \zeta_4, \tag{B.18a}$$

$$\alpha_2 = -\left(\zeta_1\zeta_2 + \zeta_3\zeta_4 + \zeta_1\zeta_3 + \zeta_1\zeta_4 + \zeta_2\zeta_3 + \zeta_3\zeta_4\right), \tag{B.18b}$$

$$\alpha_3 = \zeta_1\zeta_2\zeta_3 + \zeta_2\zeta_3\zeta_4 + \zeta_3\zeta_4\zeta_1 + \zeta_4\zeta_1\zeta_2, \tag{B.18c}$$

$$\alpha_4 = -\zeta_1\zeta_2\zeta_3\zeta_4, \tag{B.18d}$$

$$\beta = 1 - \alpha_1 - \alpha_2 - \alpha_3 - \alpha_4. \tag{B.18e}$$

Appendix C: Balance Operator in the Ensemble Square Root Filter

Spatial localization within the EnSRF can be applied in either model space:

$$\mathbf{K} = [(\rho_M \circ \mathbf{X}\mathbf{X}^T)\mathbf{H}^T][\mathbf{H}(\rho_M \circ \mathbf{X}\mathbf{X}^T)\mathbf{H}^T + \mathbf{R}]^{-1}, \tag{C.1}$$

or observation space:

$$\mathbf{K} = (\rho_O \circ \mathbf{X}\mathbf{Y}^T)(\mathbf{Y}\mathbf{Y}^T + \mathbf{R})^{-1}, \tag{C.2}$$

where $\mathbf{K} \in \Re^{N \times L}$ is the Kalman gain, $\mathbf{Y} = \mathbf{H}\mathbf{X} \in \Re^{L \times M}$ is the ensemble spread in observation space, and $\rho_M \in \Re^{N \times N}$ and $\rho_O \in \Re^N$ are covariance localization functions in model space and observation space respectively.

To apply the balance operator in either of these localization algorithms, the background perturbations are transformed to the unbalanced variable space as in both the EnVar and the LETKF. Then, a new Kalman gain is constructed to compute an analysis in the unbalanced space, shown for both the model space and observation space localization:

$$\mathbf{K}_z = [(\rho_M \circ \mathbf{Z}\mathbf{Z}^T)\mathbf{\Gamma}^T\mathbf{H}^T][\mathbf{H}\mathbf{\Gamma}(\rho_M \circ \mathbf{Z}\mathbf{Z}^T)\mathbf{\Gamma}^T\mathbf{H}^T + \mathbf{R}]^{-1}, \tag{C.3}$$

$$\mathbf{K}_z = (\rho_O \circ \mathbf{Z}\mathbf{Y}^T)(\mathbf{Y}\mathbf{Y}^T + \mathbf{R})^{-1}, \tag{C.4}$$

where $\mathbf{Y} = \mathbf{H}\boldsymbol{\Gamma}\mathbf{Z}$. For the model space localization Kalman gain, $\boldsymbol{\Gamma}^T$ appears outside of the localization in the first term, allowing it to vertically propagate information outside of the localization radius. Along with the transformation from the unbalanced analysis back to the full variables with the application of $\boldsymbol{\Gamma}$, there is a two-way propagation of information in this method and it functions similarly to the EnVar formulation. For the observation space localization Kalman gain, $\boldsymbol{\Gamma}^T$ occurs within the spatial localization in the first term, not allowing a spreading of balanced information outside of the localization radius. $\boldsymbol{\Gamma}$ is still applied once the unbalanced analysis is found to transform to the full variables, but it only allows a one-way propagation of information among the variables, analogous to the LETKF.

Bibliography

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

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

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

Anderson, J. L. and Anderson, S. L. (1999). A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Mon. Wea. Rev.*, 127:2741–2758.

Anderson, J. L. and Lei, L. (2013). Empirical localization of observation impact in ensemble Kalman filters. *Mon. Wea. Rev.*, 141:4140–4153.

Baer, F. and Tribbia, J. (1977). On complete filtering of gravity waves through nonlinear initialization. *Mon. Wea. Rev.*, 105:1536–1539.

Bannister, R. N. (2008). A review of forecast error covariance statistics in atmospheric variational data assimilation. II: Modelling the forecast error covariance statistics. *Quart. J. Roy. Meteor. Soc.*, 134:1971–1996.

Barker, D. M., Huang, W., Guo, Y.-R., Bourgeois, A. J., and Xiao, Q. N. (2004). A three-dimensional variational data assimilation system for MM5: Implementation and initial results. *Mon. Wea. Rev.*, 132:897–914.

Bishop, C. (2017). Gain form of the Ensemble Transform Kalman Filter and its relevance to satellite data assimilation with model space ensemble covariance localization. College Park, MD.

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

Buehner, M. (2012). Evaluation of a spatial/spectral covariance localization approach for atmospheric data assimilation. *Mon. Wea. Rev.*, 140(2):617–636.

Burgers, G. P., van Leeuwen, P. J., and Evensen, G. (1998). Analysis scheme in the ensemble Kalman filter. *Mon. Wea. Rev.*, 126:1719–1724.

Campbell, W. F., Bishop, C. H., and Hodyss, C. (2010). Vertical covariance localization for satellite radiances in ensemble Kalman filters. *Mon. Wea. Rev.*, 138:282–290.

Caya, A., Sun, J., and Snyder, C. (2005). A comparison between the 4DVAR and the ensemble Kalman filter techniques for radar data assimilation. *Mon. Wea. Rev.*, 133:3081–3094.

Chen, L., Chen, J., Xue, J., and Xia, Y. (2015). Development and testing of the GRAPES regional ensemble-3DVAR hybrid data assimilation system. *Journal of Meteorological Research*, 29:981–996.

Chen, Y. and Oliver, D. S. (2010). Cross-covariances and localization for EnKF in multiphase flow data assimilation. *Computational Geosciences*, 14:579–601.

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

Cohn, S. E., da Silva, A., Guo, J., Sienkiewicz, M., and Lamich, D. (1998). Assessing the effects of data selection with the DAO physical-space statistical analysis system. *Mon. Wea. Rev.*, 126:2913–2926.

Coman, A., Foret, G., Beekmann, M., Eremenko, M., Dufour, G., Gaubert, B., Ung, A., Schmechtig, C., Flaud, J.-M., and Bergametti, G. (2012). Assimilation of IASI partial tropospheric columns with an ensemble Kalman filter over Europe. *Atmospheric Chemistry and Physics*, 12:2513–2532.

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

Courtier, P., Thépaut, J.-N., and Holingsworth, A. (1994). A strategy for operational implementation of 4D-Var, using an incremental approach. *Quart. J. Roy. Meteor. Soc.*, 120:1367–1387.

Cullen, M. J. P. (2003). Four-dimensional variational data assimilation: A new formulation of the background-error covariance matrix based on a potential-vorticity representation. *Quart. J. Roy. Meteor. Soc.*, 129:2777–2796.

Daley, R. (1991). *Atmospheric data analysis*. Cambridge University Press.

Danforth, C. M., Kalnay, E., and Miyoshi, T. (2007). Estimating and correcting global weather model error. *Mon. Wea. Rev.*, 135:281–299.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Delsol, N. B. C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Holm, E. V., Isaksen, L., Kallberg, P., Kohler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J.-J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N., and Vitart, F. (2011). The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Quart. J. Roy. Meteor. Soc.*, 137:553–597.

Derber, J. and Bouttier, F. (1999). A reformulation of the background error covariance in the ECMWF global data assimilation system. *Tellus A*, 51(2):195–221.

Doblas-Reyes, F. J., Balmaseda, M. A., Weisheimer, A., and Palmer, T. A. (2011). Decadal climate prediction with the European Centre for Medium-Range Weather Forecasts coupled forecast system: impact of ocean observations. *J. Geophys. Res.*, 116.

Errico, R. M., Rosmond, T., and Goerss, J. S. (1993). A comparison of forecast analysis and initialization increments in an operational data assimilation system. *Mon. Wea. Rev.*, 121:579–588.

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

Fisher, M. (2003). Background error covariance modelling. In *ECMWF seminar of recent developments in data assimilation for atmopshere and ocean*, pages 45–64, Reading, UK. ECMWF.

Gaspari, G. and Cohn, S. E. (1999). Construction of correlation functions in two and three dimensions. *Quart. J. Roy. Meteor. Soc.*, 125:723–757.

Gauthier, P., Charrette, L., Koclas, P., and Laroche, S. (1999). Implementation of a 3D variational data assimilation at the Canadian Meteorological Centre. Part I: The global analysis. *Atmos.–Ocean*, 37(2):103–156.

Gauthier, P. and Thépaut, J.-N. (2001). Impact of the digital filter as a weak constraint in the preoperational 4DVAR assimilation system of the Météo-France. *Mon. Wea. Rev.*, 129:2089–2102.

Greybush, S. J., Kalnay, E., Miyoshi, T., and Ide, K. (2011). Balance and ensemble Kalman filter localization techniques. *Mon. Wea. Rev.*, 139:511–522.

Ham, Y. G., Rienecker, M. M., Suarez, M. J., Vikhliaev, Y., Zhao, B., Marchak, M., Vernirres, G., and Schubert, S. D. (2014). Decadal prediction skill in the GEOS-5 forecast system. *Climate Dyn.*, 42:1–20.

Hamill, T. M. and Snyder, C. (2000). A hybrid ensemble Kalman fi lter-3D variational analysis scheme. *Mon. Wea. Rev.*, 128:2905–2919.

Hamill, T. M., Whitaker, J. S., and Snyder, C. (2001). Distance dependent filtering of background error covariance estimates in an ensemble Kalman filter. *Mon. Wea. Rev.*, 129:2776–2790.

Han, G., Wu, X., Zhang, S., Liu, Z., and Li, W. (2013). Error covariance estimation for coupled data assimilation using a Lorenz atmosphere and a simple pycnocline ocean model. *J. Climate*, 26:10218–10231.

Harlim, J. and Hunt, B. R. (2007). Four-dimensional local ensemble transform Kalman filter: Numerical experiments with a global circulation model. *Tellus*, 59A:731–748.

Hazeleger, W., Wouters, B., van Oldenborgh, G. J., Corti, S., Palmer, T., Smith, D., Dunstone, N., Krger, J., Pohlmann, H., and von Storch, J. S. (2013). Predicting multiyear North Atlantic Ocean variability. *J. Geophys. Res.*, 118:1087–1098.

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

Honda, Y., Nishijima, M., Koizumi, K., Ohta, Y., Tamiya, K., Kawabata, T., and Tsuyuki, T. (2005). A pre-operational variational data assimilation system for a non-hydrostatic model at the Japan Meteorology Agency: Formulation and preliminary results. *Quart. J. Roy. Meteor. Soc.*, 131:3465–3475.

Houtekamer, P. L. and Mitchell, H. L. (1998). Data assimilation using an ensemble Kalman filter technique. *Mon. Wea. Rev.*, 126:796–811.

Houtekamer, P. L. and Mitchell, H. L. (2001). A sequential ensemble Kalman filter for atmospheric data assimilation. *Mon. Wea. Rev.*, 129:123–137.

Huang, X.-Y. and Lynch, P. (1993). Diabatic digital-filtering initialization: Application to the HIRLAM model. *Mon. Wea. Rev.*, 121:589–603.

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

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *J. Basic Eng., Trans. ASME*, 82:35–45.

Kalnay, E. (2003). *Atmospheric modeling, data assimilation and predictability*. Cambridge University Press.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Leetmaa, A., Reynolds, R., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K., Ropelewski, C., Wang, J., Jenne, R., and Joseph, D. (1996). The NCEP/NCAR 40-year reanalysis project. *Bull. Amer. Meteor. Soc.*, 77:437–471.

Kang, J.-S., Kalnay, E., Liu, J., Fung, I., Miyoshi, T., and Ide, K. (2011). variable localization in an ensemble Kalman filter: Application to the carbon cycle data assimilation. *J. Geophys. Res.*, 116.

Kang, J.-S., Kalnay, E., Miyoshi, T., Liu, J., and Fung, I. (2012). Estimation of surface carbon fluxes with an advanced data assimilation methodology. *J. Geophys. Res.*, 117.

Kepert, J. (2009). Covariance localisation and balance in an ensemble Kalman filter. *Quart. J. Roy. Meteor. Soc.*, 135:1157–1176.

Kleist, D. T. and Ide, K. (2015a). An OSSE-based evaluation of hybrid variational-ensemble data assimilation for the NCEP GFS. Part I: System description and 3D-hybrid results. *Mon. Wea. Rev.*, 143:433–451.

Kleist, D. T. and Ide, K. (2015b). An OSSE-based evaluation of hybrid variational-ensemble data assimilation for the NCEP GFS. Part II: 4DEnVar and hybrid variants. *Mon. Wea. Rev.*, 143:452–470.

Kleist, D. T., Parrish, D. F., Derber, J. C., Treadon, R., Errico, R. M., and Yang, R. (2009a). Improving incremental balance in the GSI 3DVAR analysis system. *Mon. Wea. Rev.*, 137:1046–1060.

Kleist, D. T., Parrish, D. F., Derber, J. C., Treadon, R., Wu, W.-S., and Lord, S. (2009b). Introduction of the GSI into the NCEP Global Data Assimilation System. *Wea. Forecasting*, 24:1691–1705.

Kucharski, F. (2012). personal communication.

Kuo, Y.-H., Zou, X., and Guo, Y.-R. (1996). Variational assimilation of precipitable water using a nonhydrostatic mesoscale adjoint model. Part I: Moisture retrieval and sensitivity experiments. *Mon. Wea. Rev.*, 124:122–147.

Le Dimet, F.-X. and Talagrand, O. (1986). Variational algorithms for analysis and assimilation of meteorological observations. *Tellus*, 37A:97–110.

Li, H., Kalnay, E., and Miyoshi, T. (2009). Simultaneous estimation of covariance inflation and observation errors within an ensemble Kalman filter. *Quart. J. Roy. Meteor. Soc.*, 135:523–533.

Li, Z., McWilliams, J. C., Ide, K., and Farrara, J. D. (2015). A multiscale variational data assimilation scheme: Formulation and illustration. *Mon. Wea. Rev.*, 143(9):3804–3822.

Liu, J., Fung, I., Kalnay, E., Kang, J.-S., Olsen, E. T., and Chenh, L. (2012). Simultaneous assimilation of AIRS Xco2 and meteorological observations in a carbon climate model with an ensemble Kalman filter. *J. Geophys. Res.*, 117.

Liu, Z., Wu, S., Zhang, S., Liu, Y., and Rong, X. Y. (2013). Ensemble data assimilation in a simple coupled climate model: The role of ocean-atmosphere interaction. *Advances in Atmopsheric Sciences*, 30:1235–1248.

Lorenc, A. (1986). Analysis methods for numerical weather prediction. *Quart. J. Roy. Meteor. Soc.*, 112:1177–1194.

Lorenc, A., Ballard, S., Bell, R., Ingelby, N., Andrews, P., Barker, D., Bray, J., Clayton, A., Dalby, T., Li, D., Payne, T., and Saunders, F. (2003). The Met Office global three-dimensional variational data assimilation scheme. *Quart. J. Roy. Meteor. Soc.*, 126(270):2991–3012.

Lorenc, A. C. (2003). The potential of the ensemble Kalman filter for NWP–A comparison with 4D-Var. *Quart. J. Roy. Meteor. Soc.*, 129:3183–3203.

Machenhauer, B. (1977). On the dynamics of gravity oscillations in a shallow water model, with application to normal mode initialization. *Contrib. Atmos. Phys.*, 50:253–271.

Mitchell, H. L., Houtekamer, P. L., and Pellerin, G. (2002). Ensemble size, balance, and model-error representation in an ensemble Kalman filter. *Mon. Wea. Rev.*, 130:2791–2808.

Miyoshi, T. (2005). *Ensemble Kalman filter experiments with a primitive-equation global model*. PhD thesis, University of Maryland, College Park.

Miyoshi, T. (2011). The Gaussian approach to adaptive covariance inflation and its implementation with the local ensemble transform Kalman filter. *Mon. Wea. Rev.*, 139:1519–1535.

Molteni, F. (2003). Atmospheric simulations using a GCM with simplified physical parameterizations. I: model climatology and variability in multi-decadal experiments. *Climate Dyn.*, 20:175–191.

Pagowski, M. and Grell, G. A. (2012). Experiments with the assimilation of fine aerosols using an ensemble Kalman filter. *J. Geophys. Res.*, 117.

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

Purser, R. J., Wu, W.-S., Parrish, D. F., and Roberts, N. M. (2003). Numerical aspects of the application of recursive filters to variational statistical analysis. Part I: Spatially homogeneous and isotropic Gaussian covariances. *Mon. Wea. Rev.*, 131:1524–1535.

Robson, J. I., Sutton, R. T., and Smith, D. M. (2012). Initialized decadal predictions of the rapid warming of the North Atlantic ocean in the mid 1990s. *Geophys. Res. Lett.*, 39.

Sabol, C. A. (2011). Three-dimensional variational data assimilation experiments with the SPEEDY atmospheric general circulation model. Master's thesis, University of Maryland, College Park.

Sasaki, Y. (1970). Some basic formalisms on numerical variational analysis. *Mon. Wea. Rev.*, 98:875–883.

Schwartz, C. S., Liu, Z., Lin, H.-C., and Cetola, J. D. (2014). Assimilating aerosol observations with a "hybrid" variational-ensemble data assimilation system. *Journal of Geophysical Research Atmospheres*, 119:4043–4069.

Sluka, T. C., Penny, S. G., Kalnay, E., and Miyoshi, T. (2016). Assimilating atmospheric observations into the ocean using strongly coupled ensemble data assimilation. *Geophys. Res. Lett.*, 43:752–759.

Smith, T. M., Reynolds, R. W., Peterson, T. C., and Lawrimore, J. (2008). Improvements to NOAA's historical merged landocean surface temperature analysis (1880-2006). *J. Climate*, 21:2283–2296.

Snyder, C. and Zhang, F. (2003). Assimilation of simulated Doppler radar observations with an ensemble Kalman filter. *Mon. Wea. Rev.*, 131:1663–1677.

Thepaut, J.-N. and Courtier, P. (1991). Four-dimensional variational data assimilation using the adjoint of a multilevel primitive-equation model. *Quart. J. Roy. Meteor. Soc.*, 117:1225–1254.

Thomas, C. A. and Ide, K. (2017a). Balance operators in ensemble data assimilation, Part I: Hybrid 4DEnVar.

Thomas, C. A. and Ide, K. (2017b). Balance operators in ensemble data assimilation, Part II: Localization.

Thomas, C. A. and Ide, K. (2017c). An overview of variable localization methods within ensemble data assimilation schemes.

Tippett, M. K., Anderson, J. L., Bishop, C. H., Hamill, T. M., and Whitaker, J. S. (2003). Ensemble square root filters. *Mon. Wea. Rev.*, 131:1485–1490.

Tong, M. and Xue, M. (2005). Ensemble Kalman filter assimilation of Doppler radar data with a compressible nonhydrostatic model: OSS experiments. *Mon. Wea. Rev.*, 133:1789–1807.

Vetra-Carvalho, S., Dixon, M., Migliorini, S., Nichols, N. K., and Ballard, S. P. (2012). Breakdown of hydrostatic balance at convective scales in the forecast errors in the Met Office Unified Model. *Quart. J. Roy. Meteor. Soc.*, 138:1709–1720.

Wang, X. (2010). Incorporating ensemble covariance in the gridpoint statistical interpolation (GSI) variational minimization: a mathematical framework. *Mon. Wea. Rev.*, 138:2990–2995.

Wee, T.-K. and Kuo, Y.-H. (2004). Impact of a digital filter as a weak constraint in MM5 4DVAR: An observing system simulation experiment. *Mon. Wea. Rev.*, 132:543–559.

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

Whitaker, J. S., Hamill, T. M., Wei, X., Song, Y., and Toth, Z. (2008). Ensemble data assimilation with the NCEP Global Forecast System. *Mon. Wea. Rev.*, 136:463–482.

Williamson, D. L., Daley, R., and Schlatter, T. (1981). The balance between mass and wind fields resulting from multivariate optimal interpolation. *Mon. Wea. Rev.*, 109:2357–2376.

Wu, W.-S., Parrish, D. F., and Purser, R. J. (2002). Three-dimensional variational analysis with spatially inhomogeneous covariances. *Mon. Wea. Rev.*, 130:2905–2916.

Xue, Y., Smith, T. M., and Reynolds, R. W. (2003). Interdecadal changes of 30-yr SST normals during 1871-2000. *J. Climate*, 16:1601–1612.

Zhou, Y. (2014). *Minimizing reanalysis jumps due to new observing systems*. PhD thesis, University of Maryland, College Park.

Zou, X. and Kuo, Y.-H. (1996). Rainfall assimilation through an optimal control of initial and boundary conditions in a limited-area mesoscale model. *Mon. Wea. Rev.*, 124:2859–2882.

Zou, X., Kuo, Y.-H., and Guo, Y.-R. (1995). Assimilation of atmospheric radio refractivity using a nonhydrostatic adjoint model. *Mon. Wea. Rev.*, 123:2229–2250.