TECHNIQUES FOR THE IMPROVEMENT OF NUMERICAL WEATHER PREDICTION: INVESTIGATING THE DYNAMICS OF FORECAST UNCERTAINTY AND THE IMPLEMENTATION OF REGIONALLY ENHANCED GLOBAL DATA ASSIMILATION

A Dissertation

by

MICHAEL AARON HERRERA

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Istvan Szunyogh
Committee Members,	Ramalingam Saravanan
	Robert L. Korty
	Robert D. Heltand
Head of Department,	Ping Yang

December 2016

Major Subject: Atmospheric Sciences

ABSTRACT

This dissertation describes and shows results from two projects which focused on investigating and improving current methods of numerical weather prediction. First, we show a new data assimilation (DA) approach that produces a global analysis that is enhanced by information from multiple regions of a limited area model (LAM). The enhancement is introduced by blending the model states from the global and regional models and using the blended state to compute the innovations (the difference between the observations and their predicted value). The approach is tested by an implementation on the US Navy's operational global data assimilation system and global and limited area numerical weather prediction models. The resulting system is evaluated by month long analysis-forecast experiments. It is found that REG DA can provide improved global forecasts with the largest improvements observed for Hurricane Sandy and frontal passages over the central plains. This result suggests that operational NWP centers that produce both global and regional forecast products could potentially improve their global forecasts without a significant increase in computational resources by implementing the proposed approach.

Next, we employ local linear, spatial spectral, and Lorenz-curve based diagnostics to investigate the dynamics of uncertainty in global numerical weather forecasts in the NH extratropics. The diagnostics are applied to ensembles in the THORPEX Interactive Grand Global Ensemble (TIGGE). The initial growth of uncertainty is found to be the fastest at the synoptic scales (zonal wave numbers 7-9) most sensitive to baroclinic instability. At later forecast times, the saturation of uncertainties at the synoptic scales and the longer sustainable growth of uncertainty at the large scales lead to a gradual shift of the wave number of the dominant uncertainty towards zonal wave number 5. At the sub-synoptic scales, errors saturate as predicted by Lorenz's classic theory. While the ensembles cap-

ture the general characteristics of the uncertainty dynamics efficiently, there are locations where the predicted magnitude and structure of uncertainty have considerable time-mean errors. In addition, the magnitude of systematic errors in the prediction of the uncertainty increases with increasing forecast time. These growing systematic errors are dominated by errors in the prediction of low frequency changes in the large scale flow.

DEDICATION

I dedicate this dissertation to my wife of almost eight years, Samantha Phelps-Herrera. I can absolutely say that without her love and support, I would not be where I am now.

ACKNOWLEDGMENTS

I first want to thank my committee chair and advisor, Dr. Istvan Szunyogh for his immense help over the course of my graduate career. He's afforded me the opportunity to travel across the U.S. visiting national labs and conferences to learn from other scientists and present my work. Without his guidance I would not be where I am today. The other members of my research group, both past and present, have also played a significant role in my graduate career. Thank you to Dr. Gyorgyi Gyarmati who manages all of our IT needs in the group, Dr. Christina Holt who acted as a mentor when I first entered graduate school, Michael Battalio for being a good friend and invaluable resource in both the classroom and in my research, my mentee Carlee Loeser who helped me better understand and defend my research, and the remainder of our group, Fan Han, Adam Brainard, and Max Gawryla.

I would also like to thank my family and friends. My parents help instill in me a curiosity about our world that has led me to pursue a scientific career. I've also been lucky enough to have two very close cohorts over my time in graduate school and they have both been an important source of support, without which my time in graduate school would have been much more difficult.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a dissertation committee consisting of Professors Istvan Szunyogh, Ramalingam Saravanan, and Robert Korty of the Department of Atmospheric Sciences and Professor Robert Hetland of the Department of Oceanography.

The study shown in Section 2 was conducted in collaboration with scientists from two of the NRL institutions. From NRL Monterey, Craig Bishop, Allen Zhao, and Teddy Holt, and from NRL DC, Dave Kuhl and Karl Hoppel.

The study shown in Section 3 was conducted in collaboration with Joseph Tribbia from the National Center for Atmospheric Research (NCAR).

All other work conducted for the dissertation was completed by the student independently.

Funding Sources

Graduate study was supported by a fellowship from Texas A&M University and the following grants:

- National Science Foundation Grant ATM-AGS-1237613
- Office of Naval Research Grant N00014-12-1-0785
- Office of Naval Research Grant N00014-16-1-3011
- Naval Research Lab Award N00173-16-1-3011

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION*

This dissertation focuses on improving the current state of numerical weather prediction (NWP) by developing and implementing a new data assimilation (DA) method and investigating the dynamics of forecast uncertainty in state of the art global ensemble forecast systems from the leading NWP centers around the world. Improvements in NWP are incremental and this new DA technique called Regionally Enhanced Global (REG) DA would allow for global weather forecast models to see an incremental improvement without a significant increase in computational resources. The second half of this dissertation focuses on how well the leading ensemble forecast systems are able to capture the magnitude and space of forecast uncertainty.

Most major numerical weather prediction centers of the world produce both global and limited area model (LAM) forecast products. Focusing on smaller domains than their global counterparts, LAMs can operate at higher resolutions for the available finite computational resources. The assumption is that the higher resolution LAM can provide accurate forecasts of the atmospheric state at the scales that are smaller than those that the coarser resolution global model can resolve for a limited time (up to about 2-3 day) in the interior of the LAM domain. The goal of this project is to formulate and test a global DA approach, REG DA, that can take advantage of this presumed useful forecast information about the smaller scales.

Previous studies with highly idealized Lorenz models of the atmospheric dynamics demonstrated the potential for improvement of both the global and limited area analyses and forecasts by preparing the global and limited area analyses concurrently by a unified

*Parts of Section 1 are reprinted with permission from "Forecast Uncertainty Dynamics in the THORPEX Interactive Grand Global Ensemble (TIGGE)" by M. Herrera, I. Szunyogh, and J. Tribbia, 2016. Monthly Weather Review, 144, 2739-2766, Copyright 2016 by the American Meteorological Society.

state estimation (data assimilation) process (Yoon et al., 2012; Kretschmer et al., 2015). These idealized studies were motivated by the mixed results of the first attempt to incorporate higher resolution limited area model information into the global data assimilation process (Merkova et al., 2011). That study used operational models of the National Centers for Environmental Prediction (NCEP) and the research data assimilation system of Szunyogh et al. (2008). The results showed that further research was necessary to refine the technical approach for the introduction of the limited area forecast information into the global data assimilation process. The present study is the first attempt since that of Merkova et al. (2011) to test the concept on operational models. The technical approach used is similar but not completely identical to that of Kretschmer et al. (2015). In essence, the limited area forecast information is utilized in the computation of the innovations (the differences between the observations and their predicted values), but does not affect the global background state estimate updated by the assimilation of the observations.

To test the new approach, we use the U.S. Navy's operational global and limited area models, and global data assimilation system. The global model is the Navy Global Environment Model (NAVGEM) (Hogan et al., 2014), a spectral transform model operationally run at a resolution of T425. The regional model is the Coupled Ocean/Atmosphere Mesoscale Prediction System (COAMPS) (Hodur, 1997), using the Navy's standard configuration without ocean/atmosphere coupling. COAMPS is implemented operationally on more than 80 limited area domains over the world, with a varying horizontal resolution from 25km to 500m. The global data assimilation system of NAVGEM, the Naval Research Laboratory Atmospheric Variational Data Assimilation System - Accelerated Representer (NAVDAS-AR), is a four-dimensional variational (4D Var) data assimilation scheme (Xu et al., 2005; Rosmond and Xu, 2006). Implementation of the REG DA system would allow for the Navy to combine the DA systems of NAVGEM and COAMPS, reducing the run time by eliminating the need for DA for each individual limited area domain,

and improving both the global and regional analyses.

Data assimilation produces the initial conditions of numerical weather forecasts by a statistical interpolation of the atmospheric observations (e.g. Szunyogh, 2014). Thus the initial conditions have an inherently random error component, which we will call the *analysis uncertainty*. The amplification of the analysis uncertainty by the chaotic dynamics of the atmosphere would lead to an inevitable growth of the magnitude of the *forecast uncertainty*, even if models were perfect replica of the atmosphere. Because models are not perfect and use statistical considerations to account for processes at the unresolved scales, *model errors* and *model uncertainty* also contribute to the forecast uncertainty.

In perhaps the most influential paper ever written on the dynamics of forecast error (uncertainty) growth, Lorenz (1969b) investigated the role of scale interactions in the error growth process. He argued that forecast errors saturated, predictability was lost, at increasingly larger scales as forecast time increased. Lorenz's results were most recently revisited by Tribbia and Baumhefner (2004), Rotunno and Snyder (2008), and Durran and Gingrich (2014). Tribbia and Baumhefner (2004) augmented Lorenz's description of the process by adding that in the extratropics the dominant errors asymptoted to the baroclinically active scales, where they then grew exponentially. Rotunno and Snyder (2008) replaced the original two-dimensional vorticity equation in Lorenz's model by the surface geostrophic equation. They pointed out that the rapid downscale propagation of errors at the mesoscales played an important role in the rapid saturation of the forecast errors at the smaller scales. Durran and Gingrich (2014) extended this argument to emphasize that synoptic scale errors, even if they had small magnitude, led to a rapid saturation of the errors at the smaller scales.

In their ensemble based predictability studies, Kuhl et al. (2007), and Satterfield and Szunyogh (2010, 2011) (hereafter reffered to collectively as KSS) observed the same spectral evolution of the forecast uncertainty as Tribbia and Baumhefner (2004). They also

found that a 40-80 member ensemble of forecasts was efficient in capturing the dominant synoptic scale patterns of forecast uncertainty. The present study extends the investigations of KSS to the THORPEX Interactive Grand Global Ensemble (TIGGE), which is comprised of operational global ensemble forecast data from the major operational centers. The primary objective of the study is to verify that the findings of KSS also hold for the operational ensemble forecast systems. The results shown later strongly suggest the affirmative. The secondary objective is to provide information about the performance of the operational ensemble forecast systems.

2. REGIONALLY ENHANCED GLOBAL DATA ASSIMILATION

2.1 The Concept Of REG DA

Baek et al. (2006) argued that an efficient approach to account for model errors due to differences between the model attractor and the true attractor in data assimilation was to modify the observation function $\mathbf{h}(\mathbf{x})$ as

$$\mathbf{h}(\mathbf{x}) \rightarrow \mathbf{h}(\mathbf{x} - \mathbf{b}), \tag{2.1}$$

where $\mathbf{b}$ was a time dependent estimate of the model error in the model state $\mathbf{x}$. The general idea was that in cases where the model error was due to difference between the attractors, it was more advantageous to account for the effects of model errors in the computation of the innovations than to push the background from the model attractor towards the true attractor. Baek et al. (2006) estimated the unknown $\mathbf{b}$ by the method of state augmentation. Baek et al. (2009) pointed out that the approach was closely related to the techniques used for the correction of the state dependent observation bias in the practice of data assimilation. Baek et al. (2009) made this argument by considering the effects of differences between the model orography and the true orography in the data assimilation process. The differences between the orographies lead, for instance, to state dependent errors in the prediction of the surface pressure, which are obvious examples of errors due to differences between the model attractor and the true attractor. These errors can be efficiently accounted for by modifying the observation function as

$$\mathbf{h}(\mathbf{x}) \rightarrow \mathbf{h}(\mathbf{x}) + \mathbf{c}, \tag{2.2}$$

where $\mathbf{h}(\mathbf{x})$ is now the observation function for the surface pressure observations and $\mathbf{c}$ is

the state dependent surface pressure bias. That is the correction of the entire state by b can be replaced by the correction of the predicted value of the observations, $\mathbf{h}(\mathbf{x})$, by $\mathbf{c}$, which can formally be written as

$$\mathbf{h}(\mathbf{x} - \mathbf{b}) \approx \mathbf{h}(\mathbf{x}) + \mathbf{c}. \tag{2.3}$$

The state dependent bias $\mathbf{c}$ can also be estimated by the method of state augmentation.

A similar argument can be made in connection with the correction of the bias of radiance observations. In that case, $\mathbf{h}(\mathbf{x})$ is a radiative transfer model and experience shows that the observations cannot be gainfully assimilated without a bias correction in the form of Eq. 2.2. There are strong indications (personal communication, Massimo Bonavita of ECMWF), however, that the observation bias correction algorithms correct as much for the effects of errors in $\mathbf{x}$ (the model prediction of the state) as for biases in the radiance observations or the radiation transfer models.

In the approach proposed here, the high resolution limited area forecast information plays the role of $\mathbf{b}$. For instance, due to the higher resolution, more realistic orography of the limited area model, the use of LAM information is expected to lead to a better interpretation of the surface pressure and radiance observations.

2.2 Formulation For An Incremental 4D-Var

Our goal is to obtain an estimate (analysis) $\mathbf{x}_g^a$ of the global atmospheric state, where the components of the discrete state vector $\mathbf{x}_g^a$ are defined by a model representation of the atmospheric state. We assume, that in addition to the global model, we also have access to a limited area model that can add resolution and accuracy to the global model forecast in a limited area domain. To simplify notation, we assume in this section that limited area model information is available for the entire globe.

We define the regionally enhanced global state vector $\mathbf{x}_e$ by

$$\mathbf{x}_e = (1-\alpha)\mathbf{I}(\mathbf{x}_g) + \alpha\mathbf{x}_\ell \tag{2.4}$$

where $\mathbf{I}$ is the linear operator that maps the lower resolution global model state $\mathbf{x}_g$ to the higher resolution of the limited area model, $\mathbf{x}_\ell$ is the high resolution state vector whose components are provided by the limited area models, and $0 \leq \alpha \leq 1$ is a (possibly location dependent) *blending coefficient*. Notice that the dimension of $\mathbf{x}_e$ is higher than that of $\mathbf{x}_g$. Let $\mathbf{h}_g(\cdot)$ and $\mathbf{h}_e(\cdot)$ be the observation functions operating on $\mathbf{x}_g$ and $\mathbf{x}_e$, respectively. Under the assumption that $\mathbf{x}_e$ is a more accurate representation of the atmospheric state than $\mathbf{x}_g$, $\mathbf{h}_e(\mathbf{x}_e)$ is expected to be a more accurate prediction of the observations than $\mathbf{h}_g(\mathbf{x}_g)$. Thus the effect of the change

$$\mathbf{h}_g(\mathbf{x}_g) \rightarrow \mathbf{h}_e(\mathbf{x}_e) \tag{2.5}$$

of the observation function of a global data assimilation scheme is analogous to that of Eq. (2.1).

In the incremental formulation of 4D-Var, the increment $\delta\mathbf{x}_g$ is defined by

$$\delta\mathbf{x}_g(t_j) = \mathbf{x}_g(t_j) - \mathbf{x}_g^b(t_j), \tag{2.6}$$

where $t_j, j = 0, \ldots, N$, are the times at which observations are available for the assimilation time window, and $\mathbf{x}_g^b$ is the global background state. The analysis

$$\mathbf{x}_g^a = \mathbf{x}_g^b + \delta\mathbf{x}_g^a \tag{2.7}$$

at time t_0 is obtained by computing the analysis increment $\delta\mathbf{x}_g^a = \delta\mathbf{x}_g(t_0)$ that minimizes

the cost function (e.g. Szunyogh, 2014)

$$\begin{aligned}\mathrm{J}\left[\delta\mathbf{x}_g(t_0)\right] = &\left[\delta\mathbf{x}_g(t_0)\right]^{\mathbf{T}}\mathbf{Pb}^{-1}\delta\mathbf{x}_g(t_0)+ \\ &\sum_{j=0}^{N}\left[\delta\mathbf{y}^{\mathrm{o}}(t_j)+\mathbf{H}_g(t_j)\mathbf{M}_g(t_0,t_j)\delta\mathbf{x}_g(t_0)\right]^{\mathbf{T}}\times \\ &\mathbf{R}_{t_j}^{-1}\left[\delta\mathbf{y}^{\mathrm{o}}(t_j)+\mathbf{H}_g(t_j)\mathbf{M}_g(t_0,t_j)\delta\mathbf{x}_g(t_0)\right],\end{aligned} \tag{2.8}$$

and adding the result to the background $\mathbf{x}_g^b$. In Eq. (2.8), the expression in the square brackets of the second(observation) term stands for $\mathbf{y}^{\mathrm{o}}(t_j) - \mathbf{h}\left[\mathbf{x}_g(t_j)\right]$. The equality of the two expressions,

$$\mathbf{y}^{\mathrm{o}}(t_j) - \mathbf{h}_g\left[\mathbf{x}_g(t_j)\right] = \delta\mathbf{y}^{\mathrm{o}}(t_j) + \mathbf{H}_g(t_j)\mathbf{M}_g(t_0,t_j)\delta\mathbf{x}_g(t_0), \tag{2.9}$$

follows from

$$\mathbf{x}_g(t_j) = \mathbf{M}_g(t_0,t_j)\delta\mathbf{x}_g(t_0), \tag{2.10}$$

$$\delta\mathbf{y}^{\mathrm{o}}(t_j) = \mathbf{y}^{\mathrm{o}}(t_j) - \mathbf{h}_g\left[\mathbf{x}_g^b(t_j)\right], \tag{2.11}$$

and

$$\mathbf{h}_g\left[\mathbf{x}_g(t_j)\right] \approx \mathbf{h}_g\left[\mathbf{x}_g^b(t_j)\right] + \mathbf{H}_g(t_j)\delta\mathbf{x}_g(t_j), \tag{2.12}$$

where $\mathbf{H}_g(t_j)$ is the linearization of $\mathbf{h}_g\left[\mathbf{x}_g(t_j)\right]$ about $\mathbf{x}_g^b(t_j)$. $\mathbf{M}_g(t_0,t_j)$ represents the linearized model dynamics between times t_0 and t_j for the nonlinear state space trajectory associated with the initial condition $\mathbf{x}_g^b(t_0)$.

Equation (2.8) describes the cost function minimized by the inner loop of the algorithm. When multiple steps of the outer loop are taken, the nonlinear trajectory is recomputed at the beginning of each step of the outer loop, using the updated background from the previous step as initial condition. This involves recomputing $\mathbf{H}_g(t_j)$, for the updated

values of $\mathbf{x}_g(t_j)$.

2.2.1 The Effect Of Replacing $\mathbf{h}_g(\mathbf{x}_g)$ By $\mathbf{h}_e(\mathbf{x}_e)$

The analogue of Eq. (2.9) for $\mathbf{h}_e(\mathbf{x}_e^b)$ is

$$\mathbf{y}^{\mathrm{o}}(t_j) - \mathbf{h}_e\left[\mathbf{x}_e(t_j)\right] = \delta_e\mathbf{y}^{\mathrm{o}}(t_j) + \mathbf{H}_e(t_j)\mathbf{M}_e(t_0, t_j)\delta\mathbf{x}_e, \tag{2.13}$$

where

$$\delta_e\mathbf{y}^{\mathrm{o}}(t_j) = \mathbf{y}^{\mathrm{o}}(t_j) - \mathbf{h}_e\left[\mathbf{x}_e^b(t_j)\right], \tag{2.14}$$

and the linearization $\mathbf{H}_e(t_j)$ of $\mathbf{h}_e\left[\mathbf{x}_e(t_j)\right]$ about $\mathbf{x}_e^b(t_j)$ satisfies

$$\mathbf{h}_e\left[\mathbf{x}_e(t_j)\right] \approx \mathbf{h}_e\left[\mathbf{x}_e^b(t_j)\right] + \mathbf{H}_e(t_j)\delta\mathbf{x}_e(t_j), \tag{2.15}$$

where

$$\begin{aligned}\delta\mathbf{x}_e(t_j) &= \mathbf{x}_e(t_j) - \mathbf{x}_e^b(t_j) \\ &= (1-\alpha)\mathbf{I}\left[\mathbf{x}_g(t_j) - \mathbf{x}_g^b(t_j)\right] + \alpha\left[\mathbf{x}_\ell(t_j) - \mathbf{x}_\ell^b(t_j)\right].\end{aligned} \tag{2.16}$$

Introducing the notation

$$\delta\mathbf{x}_\ell(t_j) = \mathbf{x}_\ell(t_j) - \mathbf{x}_\ell^b(t_j), \tag{2.17}$$

Eq. (2.16) can be written as

$$\delta\mathbf{x}_e(t_j) = (1-\alpha)\mathbf{I}\delta\mathbf{x}_g(t_j) + \alpha\delta\mathbf{x}_\ell(t_j). \tag{2.18}$$

With the help of $\mathbf{M}_g(t_0, t_j)$ and $\mathbf{M}_\ell(t_0, t_j)$, where the latter represents the linearization of the limited area dynamics about the nonlinear trajectory $\mathbf{x}_\ell(t_j)$, Eq. (2.18) can be written

as

$$\delta\mathbf{x}_e(t_j) = (1-\alpha)\mathbf{I}\mathbf{M}_g(t_0,t_j)\delta\mathbf{x}_g(t_0) + \alpha\mathbf{M}_\ell(t_0,t_j)\delta\mathbf{x}_\ell(t_0). \tag{2.19}$$

Because the right-hand side of Eq. (2.19) is a linear mapping of $\delta\mathbf{x}_e(t_0)$, Eq. (2.19) can also be written as

$$\delta\mathbf{x}_e(t_j) = \mathbf{M}_e(t_0,t_j)\delta\mathbf{x}_e(t_j), \tag{2.20}$$

where $\mathbf{M}_e(t_0,t_j)$ is the operator that represents the linear mapping. This result shows that the proposed model error correction can be implemented by replacing $\mathbf{M}_g(t_0,t_j)\delta\mathbf{x}_g$ by Eq. (2.19) and $\mathbf{H}_g$ by $\mathbf{H}_e$ in the cost function by Eq. (2.8). That is, the new cost function is

$$\begin{aligned}\mathrm{J}\left[\delta\mathbf{x}_g(t_0)\right] = &\left[\delta\mathbf{x}_g(t_0)\right]^{\mathbf{T}}\mathbf{P}\mathbf{b}^{-1}\delta\mathbf{x}_g(t_0)+ \\ &\sum_{j=0}^{N}\left[\delta_e\mathbf{y}^{\mathrm{o}}(t_j)+\mathbf{H}_e(t_j)\mathbf{M}_e(t_0,t_j)\delta\mathbf{x}_e(t_0)\right]^{\mathbf{T}}\times \\ &\mathbf{R}_{t_j}^{-1}\left[\delta_e\mathbf{y}^{\mathrm{o}}(t_j)+\mathbf{H}_e(t_j)\mathbf{M}_e(t_0,t_j)\delta\mathbf{x}_e(t_0)\right].\end{aligned} \tag{2.21}$$

The second (observation) term of the cost function depends on the control variable $\delta\mathbf{x}_g$ through Eq. (2.18), which shows that $\delta\mathbf{x}_e(t_0)$ is a function of $\delta\mathbf{x}_g$. Notice that the resolution of the analysis increment produced by this cost function is the same as that produced by the original cost function.

2.2.2 Implementation Without Using The Regional TLM $\mathbf{M}_l(t_0,t_j)$

According to Eq. (2.19), the implementation of the model correction algorithm requires the integration of the TLM for both the global and the limited area model. Next we show that in practice the algorithm can be implemented such that it requires the integration of only the global TLM. Notice that the evaluation of $\mathbf{I}\mathbf{M}_g(t_0,t_j)\delta\mathbf{x}_g(t_0)$ in Eq. (2.19) requires an integration of the global TLM and the subsequent interpolation of the result to

a higher resolution. While the operators $\mathbf{I}$ and $\mathbf{M}_g(t_0, t_j)$ do not commute formally, the order of the two operations they represent can be reversed by first interpolating $\delta\mathbf{x}_g(t_0)$ to a higher resolution and then integrating the TLM at that higher resolution. For the time being, we keep the notation $\mathbf{M}_g(t_0, t_j)$ for the higher resolution TLM dynamics and rewrite Eq. (2.19) as

$$\delta\mathbf{x}_e(t_j) = (1-\alpha)\mathbf{M}_g(t_0, t_j)\mathbf{I}\delta\mathbf{x}_g(t_0) + \alpha\mathbf{M}_l(t_0, t_j)\delta\mathbf{x}_l(t_0). \tag{2.22}$$

Then, we assume that the higher resolution global TLM also provides an acceptable approximation to $\mathbf{M}_l(t_0, t_j)$, which leads to

$$\delta\mathbf{x}_e(t_j) \approx (1-\alpha)\mathbf{M}_g(t_0, t_j)\mathbf{I}\delta\mathbf{x}_g(t_0) + \alpha\mathbf{M}_g(t_0, t_j)\delta\mathbf{x}_l(t_0) = \mathbf{M}_g(t_0, t_j)\delta\mathbf{x}_e(t_0) \tag{2.23}$$

by making use of Eq. (2.18). Comparing Eqs. (2.20) and (2.23) yields

$$\mathbf{M}_e(t_0, t_j) \approx \mathbf{M}_g(t_0, t_j). \tag{2.24}$$

In other words, the model error correction can be implemented by propagating the increment of the regionally enhanced global state vector by the higher resolution global TLM and apply the observation function to the result in the inner loop of the data assimilation.

2.3 Implementation On NAVGEM And COAMPS

The implementation of REG DA on the Navy's operational global and limited area model, and global data assimilation system has the following top level steps:

1. preparation of short-term global (NAVGEM) forecast,

2. preparation of short-term regional (COAMPS) forecast.

3. interpolation of both the NAVGEM and the COAMPS forecast fields onto a common high-resolution grid,

4. formation of the composite state,

5. 4D-Var computation of the analysis increment,

6. preparation of the analysis.

The key steps of REG DA are steps 3 and 4, which blend $\mathbf{x}_g$ from the global model and $\mathbf{x}_\ell$ from the limited area model into a single state $\mathbf{x}_e$. To form the innovations according to Eq. (2.13), a composite state is needed for each time t_j of the observation time window. Because the observation time window of NAVDAS-AR is six hours, centered around the analysis time, the data assimilation time window spans the period from forecast time 3 h to forecast time 9 h from the previous analysis time. NAVGEM and COAMPS both provide hourly output, so a composite state is created for each of the six hours of the observation time window.

The common grid used for the blending of the global state and the limited area states has to have a resolution similar to that of the limited area model, because otherwise the data assimilation system could not take advantage of the higher resolution forecast information provided by the limited area model. Both the global and the limited area model fields have to be interpolated onto this common global grid. Because NAVGEM is a spectral transform model, there are two options for the interpolation of the NAVGEM fields onto the common grid. The first option is to transform the spectral fields onto the Gaussian grid of NAVGEM, then use an interpolation technique for the gridded fields, such as a cubic spline interpolation. The second option is to first prepare a higher resolution spectral representation of the fields by zero padding, that is, by adding higher wave number components with zero spectral coefficients to the original spectral representation of the fields,

then transforming the fields onto a higher resolution Gaussian grid. We found that while this second method did not introduce any noticeable noise into the interpolated fields, the first method introduced noise that had a magnitude similar to that of the COAMPS fields in the blended state when a weight of $\alpha = 0.1$ was used in Eq. (2.4) (Figure B.1). This experience motivated the use of the second method in all of our analysis-forecast experiments.

Because COAMPS is a finite-difference model, which represents the fields on a grid in all phases of the model computations, the interpolation of the COAMPS fields require a grid-space interpolation. There are two main parts of this interpolation: the horizontal interpolation and the vertical interpolation. We use a nearest neighbor approach for the horizontal interpolation. The vertical interpolation is more involved, because the two models have different vertical coordinates. In particular, NAVGEM uses a hybrid sigma-pressure coordinate with a pressure-based definition of sigma; $\sigma = \frac{p}{p_s}$, where p_s is the surface pressure. The coordinate transitions gradually from pure sigma at the Earth's surface to pure pressure at the top of the model atmosphere (Hogan et al., 2014). COAMPS, in contrast, uses a height-based sigma coordinate, $\sigma = H\left(z - z_s\right) / \left(H - z_s\right)$, where z_s is the terrain height and H is the depth of the model atmosphere. The vertical interpolation is performed after completion of the horizontal interpolation of the COAMPS fields on the height-based sigma surfaces of the model to the horizontal location of the grid points of the common grid. The hybrid sigma-pressure coordinate of each grid point is computed based on the horizontally interpolated COAMPS pressure and surface pressure fields. For this calculation, the horizontally interpolated surface pressure field is corrected for the difference between the orography of NAVGEM and the orography of COAMPS at each horizontal grid point location. The procedure follows that of Baek et al. (2009); that is, the corrected surface pressure $\hat{p}_s$ is obtained by

$$\widehat{p}_s = p_s \times e^{g\Delta z/R_d\overline{T}}, \tag{2.25}$$

where p_s is the original surface pressure, g is standard gravity (9.8 m s^{-1}), Δz is the terrain difference between the two models, R_d is the gas constant for dry air (287 J kg^{-1} K^{-1}), and $\overline{T}$ is the mean temperature of the $|\Delta z|$ deep (hypothetical) atmospheric layer, which we estimate by

$$\overline{T} = T_s + \frac{1}{2}\gamma\Delta z. \tag{2.26}$$

In Eq. (2.26), γ is the standard atmosphere lapse rate (6.5×10^{-3}K m^{-1}), and T_s is the surface temperature. Because the temperature is available only for the model levels, we estimate the surface temperature by

$$T_s = T_1 \left(1 - \frac{\gamma R \ln\sigma_1}{g + \frac{1}{2}\gamma R \ln\sigma_1}\right) \tag{2.27}$$

where σ_1 and T_1 are the values of sigma and the temperature, respectively, at the lowest NAVGEM model level. Fig. B.2 illustrates the effect of the correction for the orography difference on the COAMPS surface pressure field. The figure shows that the correction efficiently removes the direct effect of the orography difference on the surface pressure. The remaining differences reflect differences in the evolution of the atmospheric flow between the two models, which, at least in part, may be the result of the indirect effects of the orography differences..

After the computation of the hybrid sigma-pressure coordinate of each grid point for the horizontally interpolated COAMPS fields, the COAMPS fields are interpolated linearly onto the hybrid sigma-pressure surfaces of NAVGEM. A seemingly problematic aspect of this interpolation is that it does not allow for the computation of the COAMPS fields at locations where the NAVGEM vertical coordinate is associated with a pressures that is

higher than the pressure at the lowest COAMPS vertical level. (This is the situation at locations where the orography is lower in NAVGEM than in COAMPS.) At such locations the interpolation introduces discontinuities into the COAMPS fields. In REG DA, these discontinuities do not pose a problem, because the interpolated fields are used only for the computation of the value of observation function and the vertical position of an observation cannot fall below the COAMPS orography.

The final step is the computation of the blended fields. In each experiment described in this thesis, we use a single value of the blending coefficient for all COAMPS domains. Approaching the lateral boundaries and the top of the COAMPS model atmosphere, α is gradually tapered to zero to avoid introducing discontinuities into the analysis of the atmospheric fields at the boundaries.

An example of the virtual potential temperature analysis increment produced by our system is shown in Figure B.3 for the model level nearest to the surface. The top panel is the increment from a control run of NAVGEM without using information from the regional model (equivalent to using $\alpha = 0.0$). The middle panel is the increment for $\alpha = 0.3$ (30% blend) and the bottom panel is the difference between the two fields. The difference between the two increments is constrained to the limited area domains.

2.4 Experiment Design

Three types of experiments are discussed in this paper: a control experiment, a “blend skip" control experiments, and the experiments that use nonzero blending coefficients (Table A.1). The control experiment is carried out with the standard configuration of the global forecast system, but at a reduced resolution of T119, with the computations of the inner loop of NAVDAS-AR done at resolution T47. The implementation of the REG DA experiments is summarized by Figure B.4: NAVGEM is run at a horizontal resolution of T119, COAMPS at 32 km, and the resolution of the common grid is T319. The blend

skip experiment is a control experiment in which the REG DA system is used, but without creating a blended state (Figure B.5). This experiment is equivalent to a REG DA experiment that uses $\alpha = 0$ for the blending coefficient. The purpose of this experiment is twofold. First, it allows for the evaluation of the effects of the interpolation schemes of the REG DA implementation on the analyses and forecasts: had the analyses and forecasts of the "blend skip" experiment been less accurate than those of the control experiment, the negative effects of the interpolation errors would reduce or even overwhelm, the potential positive effects of REG DA. Second, it allows for a clearer assessment of the effects of the COAMPS forecast information on the analyses and forecasts than the control experiment, because unlike the control experiment, it carries out the calculations of the inner loop of NAVDAS-AR at the same T119 resolution as the REG DA experiments.

The same three limited area domains are used in all experiments (Figure B.6): a North American (CON) domain, a Northeast Pacific (NEPAC) domain, and a European (EUR) domain. We chose these particular regions, because they provide a representative sample of limited area domains routinely used by the U.S. Navy. The analyses of all experiments are obtained by continuously cycling NAVDAS-AR for the period from 00 UTC October 1, 2012 to 00 UTC November 1, 2012. An analysis is prepared every 6 h and we assimilate all observations that were assimilated by the U.S. Navy in real time FNMOC.

2.5 Results

This section consists of three parts, each describing a different sets of verification diagnostics. The first set of diagnostics measure the stability of the REG DA system, the second set evaluates the general forecast performance of the approach for the 120 forecasts hours, while the third set focuses on its performance for tropical cyclones.

2.5.1 Cycling Diagnostics

These diagnostics are used to verify that the implementation of REG DA has no adverse effects on the qualitative behavior of NAVDAS-AR. We first examine the speed of convergence of the algorithm to find the minimizer of the cost function, as a significant increase of the number of inner loop iterations required for convergence or a complete lack of convergence would indicate a failure of REG DA. The number of iterations required to achieve the prescribed level of reduction of the cost function is shown in Figure B.7 for all three types of experiments. The blend skip and 30% REG DA configurations of NAVDAS-AR require only 10 additional iterations, which can be explained by the increased resolution of the TLM. We expect this small difference in the number of iterations to completely disappear once the operational resolutions are used. Another way to diagnose the speed of the convergence is to plot the residual after each iteration of the inner loop. The results shown for this diagnostic in Fig. B.8 confirm that REG DA has minimal effect on the speed of convergence.

The next set of diagnostics investigate the effects of REG DA on atmospheric balance between the mass and wind fields in the analyses. These diagnostics are particularly important, because intuition suggests that blending forecast fields from two different models may weaken the balance between the mass and wind fields. The standard approach of numerical weather prediction to quantify the degree of the lack of balance in the atmospheric initial conditions is to compute the surface pressure tendencies in the forecasts for the first few forecast hours: large surface pressure tendencies at short forecast times indicate a rapid adjustments of the mass field, that is, a lack of balance in the initial conditions. The surface pressure tendency tends to be slightly larger for the first few time steps of model integration even for a state-of-the-art data assimilation system. While this is also the case for NAVDAS-AR (left panel of Fig. B.9), neither the interpolation errors and the

increased resolution of the inner loop (middle panel), nor the blending of COAMPS forecast information (right panel) leads to a further increase of the magnitude of the surface pressure tendencies. In short, REG DA does not weaken the atmospheric balance in the initial conditions.

2.5.2 Forecast Diagnostics

The following diagnostics focus on the forecast performance for the first 120 forecast hours (first five forecast days). Each forecast is verified against the ECMWF analysis valid at the particular forecast time. One of the main diagnostics discussed is the mean-square-error (MSE) that we calculate in two different ways: we either average the estimates of the errors over all the forecasts to produce a map of the spatial distribution of the typical magnitude of the errors, using the formula

$$\epsilon_T = \frac{1}{T}\sum_{t=1}^{T}(x_t^f - x_t^v)^2, \tag{2.28}$$

where T is the number of forecasts, x_t^f is the forecast at time t, and x_t^v is the verifying analysis; or average over both space and time to describe the typical magnitude of the error by a single number for each forecast lead time, using the formula,

$$\epsilon = \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}(x_{ti}^f - x_{ti}^v)^2, \tag{2.29}$$

where N is the number of grid points. The MSE can be decomposed into a variance and a square mean (bias) component. We also use this decomposition to gain further insight into the behavior of the forecast errors in the different experiments.

We will refer to the first type of diagnostic that we show as *impact diagnostics*. These diagnostics show the MSE difference between two experiments in terms of percentages. For instance, when a REG DA experiment is compared to the control experiment, the

impact diagnostics are computed by

$$\frac{MSE_{control} - MSE_{REG\ DA}}{MSE_{control}} \times 100, \tag{2.30}$$

where $MSE_{control}$ is the MSE for a particular forecast variable of the control experiment, while MSE_{REGDA} is the MSE for the same forecast variable of the REG DA experiment. The difference is always taken such, that a positive value of the diagnostic indicates forecast improvement, while a negative value indicates forecast degradation due to REG DA.

Figure B.10 is the first of the series of figures that show impact diagnostics. It shows the diagnostics for the geopotential height forecasts of the control experiment and the blend skip control experiment. The different panels show the results for different regions: the left panel for the entire Northern Hemisphere (NH), while the next three panels for the three limited area domains. The results indicate that the forecasts of the blend skip control experiment are clearly more accurate than the forecasts of the control experiment. That is, the improvements due to the higher resolution of the calculations of the inner loop in the blend-skip experiment definitely outweigh the degradation that may arise due to interpolation errors.

The format of Fig. B.11 is the same as that of Fig. B.10, but it compares geopotential height forecasts of the control and the 30% REG DA experiment. Similar to the forecasts of the blend-skip control experiment, the forecasts of the 30% REG DA experiment are clearly more accurate in general than those from the control experiment. This result shows that the introduction of COAMPS model information into the data assimilation process does not lead to forecast degradations that would outweigh the improvements due to the higher resolution inner loop, but it does not prove that COAMPS forecast information added to the positive effects. That question can be addressed by comparing the forecasts of the REG DA experiments to those from the blend-skip experiment. The next four figures

(Figs. B.12-B.15) show the impact diagnostics for that comparison.

The format of Figs. B.12-B.15 is similar to that of Figs. B.10 and B.11, except that they have two rows of panels: the top shows the results for the 30% blend experiment and the bottom row for the 100% blend experiment. Fig. B.12 is directly comparable to Figs. B.10 and B.11, as it also shows results for the 500-hPa geopotential height. The results shown in this figure suggest that the introduction of COAMPS information has a generally positive impact on the analyses and ensuing forecasts in the lower troposphere (below 500 hPa), and a more mixed impact in the upper troposphere and above. The results are also more positive at the longer forecast times (beyond about 36-48 hours). The contrast between the short and longer term forecast performance is particularly striking for the 100% blend experiment. In addition, while the 30% blend leads to similarly large improvements as the 100% blend for certain levels and forecast times, it leads to much more modest degradations for the others. There are also noticeable differences between the results for the different domains: the results are more positive for the North American and Northeast Pacific domains than for the European domain.

The impact diagnostics for the temperature (Figure B.13) and the two components of the horizontal wind vectors (Figures B.14 and B.15), show both some similarities and some differences compared with the results for the geopotential height. The most important similarity is the higher ratio of the number of improved diagnostics to the number of degraded diagnostics for the 30% blend experiment than for the 100% blend experiment. Most importantly, the forecasts of the two components of the wind for the entire NH have smaller errors in the 30% blend experiment than in the blend skip experiment at almost all vertical levels and forecast times. In addition, while the patterns of improvements and degradations are less clustered in the diagrams for the temperature than for the geopotential height, the improvements for the temperature and the wind components are less confined to the lower troposphere than for the geopotential height.

Figures B.16 and B.17 show the difference between the spatiotemporally averaged diagnostics for the blend-blend skip control experiment and the REG DA experiments. These two figures show results for the geopotential height at the 850 hPa and 500 hPa level, respectively. Each curve shows results for a different verification (forecast) domain and positive values indicate improvements due to REG DA. The results suggest a general picture in which the analysis impact of REG DA is nearly neutral (slightly positive or negative), but as forecast time increases, its forecast impact becomes increasingly more positive. The forecast improvements peak at about 72-96 h forecast time. The one exception to the hitherto described general behavior is the behavior of the curves for the European region at both 850 hPa and 500 hPa for the 100% blend and at 500 hPa for the 30% blend. These curves indicate degradations due to REG DA. These degradations are strong enough to lead to a degradation for the entire NH domain at 120 h forecast lead time.

Should the overall positive results shown so far be considered a strong evidence that REG DA works as we hoped? Scientists usually answer such questions by trying to estimate the probability that the dominantly positive results were produced by random chance alone. The usual approach to estimate this probability is to do a statistical significance test. Unfortunately, as it was pointed out in a recently released statement (Wasserstein and Lazar, 2016) of the American Statistical Association (ASA), p-values computed by a statistical significance test do not measure the probability that the results were produced by random chance alone. Instead, the p-values can indicate how incompatible the data are with the statistical model specified in the test. In the same statement, ASA suggested using methods that emphasize estimation over testing. This motivates us to follow the approach of Roh et al. (2013, 2015) by using box plots rather than significance testing to investigate the strength of our evidence for the usefulness of REG DA. Figures B.18 and B.19 show box plots that correspond to the MSE curves of Fig. B.16 and B.17. The bottom and top

of each box represents the first and third quartiles respectively, the line in between is the median value, and the marker is the mean. The top and bottom whiskers are the maximum and minimum values within 1.5 times the inner quartile range, and any values that fall outside this range are considered outliers and shown as plus (+) symbols. The mean values shown in Figs. B.18 and B.19 at the different forecast lead times are the same as the MSE values shown in Figs. B.16 and B.17.

A positive value of the median in a box plots indicates that more forecasts were improved than degraded by REG DA, while a negative value indicates that more forecasts were degraded than improved. Thus, when both the mean and the median are positive, the MSE was reduced, in part because more forecasts were improved than degraded. A pair of a positive mean and a negative median indicates that large forecast improvements at a fewer times offset the effects of the more frequent but smaller magnitude degradations at the other times. For the 30% blend experiments (the top panels of Figs. B.18 and B.19), both the mean and the median are positive for the NH at forecast times equal to or longer than 36 h. Except for the European region, the statistics behave similarly in the limited area domainsas in the NH. For the 100% blend experiments (the bottom panels of Figs. B.18 and B.19), the results are somewhat less positive. For instance, for the entire NH in the 36 h-120 h forecast range, while typically both the mean and the median indicate forecast improvements, the median is negative at the 72 h lead time and both the median and the mean are negative at the 120 h lead time for the 500 hPa geopotential height. The other statistics depicted by the box plots also indicate that REG DA performed better for the 30% blend than the 100% blend, as the distribution of changes in the forecast errors is more skewed in the direction of positive values (improvements due to REG DA).

Next, we show results (Figures B.20 and B.21) for the decomposition of MSE into error variance and square mean error (bias),

$$MSE = Var + Bias^2. \tag{2.31}$$

The square bias is a measure of the magnitude of systematic errors in the forecasts, but it should be kept in mind that in a relatively small sample of forecasts a very few forecasts with large errors can also contribute to the bias. The error variance is the average magnitude of the transient component of the errors. The two figures show that while differences in the error variance drive the differences in MSE at the 850 hPa level, differences in both the error variance and the bias contribute the MSE difference at the 500 hPa level. That is, the MSE reduction due to REG DA below the 500 hPa level (Fig. B.12) is the result of a reduction of the magnitude of the transient component of the errors. The two figures also show that the degradation of forecast accuracy due to REG DA in the European region, which is particularly well pronounced for the 100% blend experiment, is the result of an increase of the magnitude of systematic errors.

Hitherto the focus has been on investigating averages over specific regions. Next, we turn our attention to the examination of the spatial structure of the errors for selected forecast variables at a specific level and forecast time. For example, Figure B.22 shows the MSE for the 1000 hPa geopotential height field at analysis time: ECMWF analyses rather than a depiction of the differences between the errors of our analyses: at the locations where the values are positive, on average, the 30% blend REG DA analyses are more similar to the ECMWF analyses than the blend-skip analyses, where the values are negative, the blend-skip analyses are more similar to the ECMWF analyses. The top left panel of the figure is the MSE for the blend skip control experiment. It shows that the DA has difficulties near to the surface in areas of high orography. The figure also indicates higher errors along the tropical cyclone paths over the Atlantic in the North American domain. The top right panel is the MSE difference for the forecasts of the blend skip control and the

30% REG DA experiment, in which positive values indicate improvements and the negative values indicate degradations due to REG DA. REG DA has the largest forecast impact in regions of high orography and most of the changes in forecast accuracy are confined to the North American and Northeast Pacific domains. In addition to the improvement in mountainous regions, there are clear improvements along the paths of the tropical cyclones in the North American domain. The decomposition of MSE into a variance and the square bias component is shown by the bottom two panels. These panels show that the reduction of MSE due to REG DA is dominantly the result of a combination of bias reduction in the Rocky Mountain region, and a reduction of the magnitude of the transient errors in the mountainous regions of western North America and along the paths of the two Atlantic tropical cyclones.

To illustrate the vertical structure of the diagnostics shown in Fig. B.22, Fig. B.23 shows them at several levels for the North American domain. The figure shows that the bias reduction over the Rocky Mountains extends well into the upper troposphere, the reduction of the error variance in the mountainous regions is confined to the atmospheric layer below 925 hPa, and the reduction of the error variance for the tropical cyclones is restricted to the layer below 500 hPa.

Figure B.24 shows the evolution of the 500 hPa diagnostics from Fig. B.23 with forecast time. Except for the improvements in the regions along the paths of the tropical cyclones, the MSE reduction due to REG DA disappears at forecast time 12 h, just to reappear at 24 h and further amplify at the later times. This behavior is the result of a temporary increase of the bias at 12 h lead time that offsets the reduction in error variance. It is unclear whether the temporary increase of bias is an artifact of the verification method due to errors in the verifying data, or the result of a transient adjustment process in the model. The fact that a similar increase of the bias is not observed at analysis time, at which the same verifying analyses are used, suggests that the latter explanation is more likely.

The same series of diagnostics that have just been shown for the geopotential height (Figs. B.22, B.23, and B.24) are also shown for the wind components (Figs. B.25, B.26, and B.27), and the air temperature (Figs. B.28, B.29, and B.30). The first set of figures shows that the improvements of the wind forecasts are overwhelmingly due to the reduction of transient forecast errors. The second set of figures suggests that the changes in the accuracy of the temperature forecasts above the 700 hPa level are negligible (Fig. B.29), the changes of forecast accuracy are the largest, but somewhat mixed near the surface, and clearly positive at 850 hPa.

One particularly interesting feature in Fig. B.30 is the large reduction of bias over the central plains that rapidly grows with increasing forecast time. An inspection of the individual forecasts that contributed to this result revealed that the large bias correction was due to the systematic improvement of the prediction of frontal passages in the region. An example for such an improved forecast is shown in Fig. B.31.

2.5.3 Tropical Cyclone Verification

The results shown earlier suggested that REG DA significantly improved the forecasts of tropical cyclones that passed through the North American domain. This motived a more careful examination of the forecasts of the tropical cyclones. It turned out that the forecast improvements due to REG DA were all associated with forecasts of Hurricane Sandy, as the effects of REG DA on the forecasts of Hurricane Rafael, the other tropical cyclone in our sample of forecasts, were essentially neutral. To determine the location of Hurricane Sandy within the global model, we use NAVGEM's internal vortex tracker. This information, as well as the verification is passed to the Developmental Testbed Center's (DTC) Model Evaluation Tools - Tropical Cyclone (MET-TC) program to produce the forecast error statistics of this section. MET-TC uses Best Track data from the National Hurricane Center for the estimation of the position and intensity errors. The results for

the track and intensity errors are presented by box plots. Unlike the box plots shown earlier (Figs. B.18 and B.19), which showed the distribution of the difference between the errors of the different experiments, these box plots show the distribution of the errors for each experiment next to each other. We note that a similar format to show the statistical distribution of the track and intensity forecast errors was used before by Holt et al. (2013, 2015).

Figure B.32 shows the box plots for the track errors. In this figure the results for the forecasts started at different times are grouped together by forecast lead time. While the track error generally increases with lead time, the errors remain nearly constant between lead times 48 h and 96 h. While the errors of the three experiments behave similarly at short lead times, the two REG DA experiments develop a small but clear advantage over the blend-skip experiment by 72 h: the mean and median errors are smaller and the entire distribution of the errors is shifted toward the smaller values for the REG DA experiments. The results are somewhat mixed at 84 h lead time (the mean and the largest errors are smaller for the REG DA experiments, but the median and the smallest errors are smaller for the blend-skip experiment) the clear advantage of the REG DA experiments return once the errors start to grow again: at lead time 120 h, the mean, the median, and the first and third quartiles are all shifted toward the lower errors for the REG DA experiments. As for the comparison of the 30% and 100% blend experiments, the latter leads to generally larger track error reduction.

Next, we decompose the track errors into an *along track error* component, and an orthogonal component called the *cross track error*, as shown in Fig. B.33. Notice that the value of either of the two error components can be negative. A forecast is better than another with respect to these measures, if their absolute value is smaller. The box plots for the along track error and the cross track error are shown in Fig. B.34 and B.35, respectively. While the along track analysis errors tend to be negative for the blend-skip

experiment, they are more evenly distributed between the negative and positive errors for the REG DA experiments, leading to a reduction of both the median and the mean error. By lead times 84 h and 96 h, the two REG DA experiments develop a clear advantage over the blend-skip experiment. Beyond that time the along track movement of Sandy becomes too fast in the model, turning the advantage of the REG DA experiments at 96 h lead time into a disadvantage by 108 h. The cross track errors are very similar for the three experiments up to lead time 84 h, but at 96 h and 120 h the two REG DA experiments have a clear advantage over the blend skip experiment.

REG DA greatly improves the accuracy of the analyses of the intensity of Sandy by shifting the entire distribution of the analysis errors for the mean sea level pressure towards smaller values. While most of the error reductions due to REG DA gradually diminish by 48 h, the REG DA forecasts maintain a small advantage at the later lead times.

3. FORECAST UNCERTAINTY IN THE THORPEX INTERACTIVE GRAND GLOBAL ENSEMBLE (TIGGE)*

3.1 The TIGGE Data Set

We provide a general description of the TIGGE data set and briefly discuss the techniques that the centers use to represent the effects of initial condition and model uncertainties.

3.1.1 The Data Set

The TIGGE data set is a collection of global ensemble forecasts from the major NWP centers, (Bougeault and Coauthors, 2010; Swinbank and Coauthors, 2016). The goal of TIGGE is to provide ensemble data to support both academic research and operational product development. The forecasts are collected in real time and made available to the scientific community by the archiving centers in an easily accessible uniform format.

We analyze data from the following forecast centers:

- European Centre for Medium-Range Weather Forecasts (ECMWF)
- US National Centers for Environmental Prediction (NCEP)
- UK Met Office (UKMO)
- China Meteorological Administration (CMA)
- Japan Meteorological Agency (JMA)
- Korean Meteorological Administration (KMA)

*Section 3 is reprinted with permission from "Forecast Uncertainty Dynamics in the THORPEX Interactive Grand Global Ensemble (TIGGE)" by M. Herrera, I. Szunyogh, and J. Tribbia, 2016. Monthly Weather Review, 144, 2739-2766, Copyright 2016 by the American Meteorological Society.

- Meteorological Service of Canada (CMC)
- Météo-France

We do not consider data from two of the NWP centers, the Australian Bureau of Meteorology (BoM) and the Centro de Previsão de Tempo e Estudos Climáticos (CPTEC), that provide data to the TIGGE data set: data was not available from the BoM ensemble for the time period of this study, while CPTEC discovered an error in their ensemble and was planning to regenerate the data.

3.1.2 Initial Condition Perturbations

The degrees of freedom of the dynamics of an operational model is orders of magnitude larger than the operationally attainable number of ensemble members. The forecast centers have addressed this challenging aspect of ensemble forecasting by developing techniques for the generation of initial condition perturbations that efficiently represent the growing part of the analysis (initial condition) errors. Table A.2 shows a list of the ensemble generation techniques of the different centers.

The bred vector method, (Toth and Kalnay, 1993, 1997), which was originally developed and implemented at NCEP, is currently used by CMA and KMA. To create the bred vectors, the analysis is randomly perturbed and the full non-linear model is run for a short period (eg. six hours) for both the control (unperturbed) and perturbed analyses. The control forecast is subtracted from the perturbed forecasts and the resulting perturbations are rescaled to the magnitude of the initial perturbations. The cycle is repeated by adding the rescaled perturbations to the next analysis. After several days of "breeding", growing patterns dominate the spatio-temporal evolution of the perturbations.

Another type of initial condition perturbations, which is used by ECMWF, JMA, and Météo-France are known as (right) singular vectors (Buizza et al., 1993; Molteni and Palmer, 1993; Mureau et al., 1993); these vectors are the initial perturbations that grow

fastest with respect to a preselected norm and optimization (forecast) time. For the ensembles included in this study, the norm is a quadratic norm with energy dimension (e.g., Buizza et al., 1993) and the optimization time is 48 h forecast hour. The Météo-France ensemble uses a combination of singular vectors and evolved singular vectors, where the evolved singular vectors are created such that the analysis time, for which the initial perturbations are created, coincides with the end of the optimization period. The evolved singular vectors are hoped to represent analysis uncertainties that were likely to grow in the analysis cycles of the immediate past. ECMWF also used evolved singular vectors in the past, but by the time of the present study they have switched to using an ensemble of data assimilations (EDA) to account for the error growth during the previous data assimilation cycles (Buizza et al., 2008). To create these perturbations, observations are perturbed randomly in accordance with their presumed error statistics in the data assimilation system; each set of perturbed observations is assimilated into a different ensemble member.

The method currently used for the generation of ensemble perturbations at NCEP is similar to the generation of bred vectors, but it uses information from the data assimilation system to determine a spatio-temporally varying rescaling factor. This method is called Ensemble Transform with Rescaling (ETR) and was developed by Wei et al. (2008); ensemble perturbations valid for the analysis time are obtained through an ensemble transform of a previous set of forecast perturbations, taking into account the observation error statistics and centering the perturbations on the analysis.

The UK Met Office uses a local Ensemble Transform Kalman Filter (ETKF) to generate their perturbations, (Bishop et al., 2001; Wang and Bishop, 2003; Bowler and Mylne, 2009). The largest difference between the ETKF and the ETR methods is that the ETKF produces an ensemble of full analyses rather than an ensemble of rescaled perturbations. The analysis perturbations, which are obtained by taking the difference between the mem-

bers of the analysis ensemble and the mean of the analysis ensemble, are added to the operational 4D-Var analysis to obtain the ensemble of perturbed initial conditions. Lastly, CMC uses an Ensemble Kalman Filter (EnKF) to generate the analysis ensemble. Unlike at UKMO, their ensemble is centered on the mean analysis produced by the EnKF.

3.1.3 Model Error Parameterization Techniques

In addition to chaotic model dynamics acting on uncertain initial conditions, model errors also contribute to the forecast error growth. Model errors affect the forecasts continuously during the entire forecast period. They also contribute to the initial conditions uncertainty through the forecast phase of the analysis cycles.

The main sources of model errors are thought to be the parameterization schemes for the sub-grid processes. One technique to account for these sources is the method of Stochastically Perturbed Parameterization Tendencies, SPPT (Buizza et al., 1999; Palmer et al., 2009). This technique perturbs the total contribution of the parameterized processes to the tendency of the state variables in the model. Another technique is to use different parameterization schemes for the same processes, or to use different values of the prescribed parameters of the parameterization schemes. This approach is known as the multi-physics technique (Berner et al., 2011; Houtekamer, 2002).

The effects of uncertainties injected at the smallest resolved scales cannot be directly simulated by the models, because the interactions between those scales and the larger scales are distorted by the models: some scale interactions are explicitly eliminated by the truncation strategies, while others are eliminated by dampening the smaller scale motions. Time integration schemes also contribute to the diffusiveness of the models at scales where nature is not diffusive. An approach to make the representation of the effect of upscale propagating uncertainties by the ensemble more realistic, called Stochastic Energy Backscattering (SKEB), was introduced and described by Shutts (2005, 2013); Berner

et al. (2009); Bowler et al. (2009); Charron et al. (2010); Tennant et al. (2011). One other method used to account for this uncertainty is called Stochastic Time Tendency Perturbations (STTP), Hou et al. (2008). Whereas SKEB focuses on subgrid scale error, STTP adds stochastic forcing at all scales. Since SPPT/multi-physics and SKEB/STTP simulate different aspects of model error dynamics, they can be used in conjunction. This practice is followed at both ECMWF and CMC.

3.2 Local Diagnostics

An ensemble forecast provides a flow (synoptic situation) dependent prediction of the probability distribution of the forecast uncertainty. We focus on examining the mean and the covariance matrix of the predicted probability distribution. The significance of the covariance matrix of the distribution is that it describes both the structure and the magnitude of the predicted uncertain flow features. In addition, under the assumption that the probability distribution of the uncertainty is Gaussian, the mean and the covariance matrix together provide a complete description of the predicted probability distribution.

3.2.1 Local Vectors And Their Covariance

Following the approach of KSS, we define a *local state vector* $\mathbf{x}_{V_\ell}$ to describe the state in a local atmospheric volume V_ℓ centered at model grid point ℓ. The components of $\mathbf{x}_{V_\ell}$ are the grid point variables of the model in V_ℓ. We assume the availability of a K-member forecast ensemble and define the K-member *ensemble of local state vectors* $\mathbf{x}^k_{V_\ell}$, k=1,...,K, by the relevant components of the state vectors that represent the ensemble. Then, a K-member *ensemble of local perturbations*, $\mathbf{X}^k_{V_\ell}$, k=1,...,K, can be defined by

$$\mathbf{X}^k_{V_\ell} = \mathbf{x}^k_{V_\ell} - \bar{\mathbf{x}}_{V_\ell}, \tag{3.1}$$

where

$$\bar{\mathbf{x}}_{V_\ell} = \frac{1}{K} \sum_{k=1}^{K} \mathbf{x}_{V_\ell}^k \tag{3.2}$$

is the *local ensemble mean*, which is the prediction of the mean of the probability distribution of the local forecast uncertainty. We note that the same local framework is employed in the widely used Local Ensemble Transform Kalman Filter (LETKF) data assimilation scheme (Ott et al., 2004; Hunt et al., 2007; Szunyogh et al., 2005, 2008).

In what follows, we treat all local vectors as column vectors. The prediction of the local covariance matrix of the forecast uncertainty is

$$\mathbf{P}_{V_\ell} = \frac{1}{K-1} \sum_{k=1}^{K} \mathbf{X}_{V_\ell}^k [\mathbf{X}_{V_\ell}^k]^T. \tag{3.3}$$

The local state vector, the local ensemble perturbations, and the local covariance matrix can be defined for all locations ℓ, local volumes V_ℓ, and forecast times t_f, (including the analysis time, $t_f = 0$). We assume that the spaces spanned by the local ensemble perturbations are linear spaces. To be precise, we define the *space of local ensemble perturbations* $\mathbb{S}_{V_\ell}^K(t_f)$ by the range of $\mathbf{P}_{V_\ell}(t_f)$, and assume that any linear combination of the local ensemble perturbations is a plausible local perturbation of the atmospheric state. Because we compute the diagnostics for all locations ℓ and use a location independent definition of the local volumes V_ℓ, we drop the subscripts from the notation. To further simplify notation, we also drop the argument t_f. For instance, we replace the notation $\mathbf{P}_{V_\ell}(t_f)$ by $\mathbf{P}$.

3.2.2 Diagnostics For The Predicted Magnitude Of The Uncertainty

3.2.2.1 Optimality Conditions

Our error diagnostics are based on investigating the statistical properties of the difference

$$\delta \mathbf{x}_t = \mathbf{x}_t - \bar{\mathbf{x}} \tag{3.4}$$

between a proxy $\mathbf{x}_t$ of the true state and the ensemble mean $\bar{\mathbf{x}}$. We treat the values of $\delta \mathbf{x}_t$ computed for the different forecasts, forecast times, and locations as realizations of the random vector variable $\delta \mathbf{x}_t$. In ensemble forecasting, the difference $\delta \mathbf{x}_t$ is usually interpreted as an estimate of the error in the ensemble mean forecast. This terminology is fully justified when the ensemble mean is used as a deterministic forecast of the atmospheric state. In our interpretation, $\delta \mathbf{x}_t$ is an estimate of the difference

$$\varepsilon_r = \mathbf{x}_T - \overline{\mathbf{x}}_T \tag{3.5}$$

between the (unknown) true state $\mathbf{x}_T$ and the (unknown) true mean $\overline{\mathbf{x}}_T$ of the probability distribution of the state given all sources of uncertainty.

The vector $\boldsymbol{\varepsilon}_r$ is a representation of the forecast uncertainty, because if there were no uncertainties, $\bar{\mathbf{x}}_T$ would be identical to $\mathbf{x}_T$, leading to $\boldsymbol{\varepsilon}_r = \mathbf{0}$. As the magnitude of the forecast uncertainty increases with increasing forecast time, $\bar{\mathbf{x}}_T$ becomes increasingly different from $\mathbf{x}_T$, leading to an increase of the magnitude of $\boldsymbol{\varepsilon}_r$. At the forecast time at which predictability is completely lost, $\bar{\mathbf{x}}_T$ becomes identical to the climatological mean state of the atmosphere and the magnitude of $\boldsymbol{\varepsilon}_r$ converges to the magnitude of the climatological atmospheric variability (Epstein, 1969; Leith, 1974).

Because

$$\mathbf{x}_t = \mathbf{x}_T + \varepsilon_t, \tag{3.6}$$

where ε_t is the error in the proxy for the true state and

$$\overline{\mathbf{x}} = \overline{\mathbf{x}}_T + \varepsilon_m, \tag{3.7}$$

where ε_m is the error in the prediction of the mean, Eq. (5) can be also written as

$$\varepsilon_r = (\mathbf{x}_t - \varepsilon_t) - (\overline{\mathbf{x}} - \varepsilon_m) = \delta\mathbf{x}_t + (\varepsilon_m - \varepsilon_t). \tag{3.8}$$

Eq. (8) shows that the estimate $\delta\mathbf{x}_t$ of ε_r has an error of $\boldsymbol{\varepsilon}_e = (\varepsilon_m - \varepsilon_t)$. Similar to $\mathbf{x}$, the random variables $\boldsymbol{\varepsilon}_m$, $\boldsymbol{\varepsilon}_r$ and $\boldsymbol{\varepsilon}_t$ depend on the location, the initial time of the forecast, and the forecast time.

According to its definition, the random variable $\boldsymbol{\varepsilon}_r$ satisfies the condition that

$$E\left(\boldsymbol{\varepsilon}_r\right) = \mathbf{0}, \tag{3.9}$$

where $E(\cdot)$ is the expected value function. Hence, the estimate $\delta\mathbf{x}_t$ of $\boldsymbol{\varepsilon}_r$ should satisfy the condition

$$E\left(\delta\mathbf{x}_t\right) = \mathbf{0}. \tag{3.10}$$

In addition, because $\mathbf{P}$ is considered the prediction of the covariance matrix of $\boldsymbol{\varepsilon}_r$ in ensemble prediction, the two covariance matrices should also be equal. Because these two conditions cannot be verified for a single realization of $\delta\mathbf{x}_t$, ensemble forecast verification techniques investigate whether the behavior of the ensemble forecast system is consistent or not with these two conditions over a large number of realizations of $\delta\mathbf{x}_t$.

3.2.2.2 The Magnitude Of The Forecast Uncertainty

The mean square magnitude of the ensemble based estimate of the difference between the true state and the true mean of the probability distribution of the state is,

$$TV = E\left[(\delta \mathbf{x}_t)^2\right] = E\left((\delta \mathbf{x}_t)^T \delta \mathbf{x}_t\right). \tag{3.11}$$

A standard approach for the evaluation of an ensemble forecast system is to verify whether the mean of the ensemble variance

$$VS = E\left(\text{trace}(\mathbf{P})\right) \tag{3.12}$$

satisfies the approximate equality

$$VS \approx TV. \tag{3.13}$$

This criteria is a necessary condition for trace$(\mathbf{P})$ being an accurate prediction of $E\left[\boldsymbol{\varepsilon}_r^2\right]$ (the trace of the covariance matrix of $\boldsymbol{\varepsilon}_r$) under the assumption that $\boldsymbol{\varepsilon}_e = \mathbf{0}$.

3.2.2.3 The Uncertainty In The Proxy For The True State

The presence of $\boldsymbol{\varepsilon}_t$ is a limitation of any forecast verification technique. In the particular case of TV, the contribution of $\boldsymbol{\varepsilon}_t$ is not negligible at short (e.g., shorter than 12-48 h) forecast times, at which its magnitude can be comparable to the magnitude of $\boldsymbol{\varepsilon}_r$. The correlations between $\boldsymbol{\varepsilon}_t$ and the error components $\boldsymbol{\varepsilon}_m$ and $\boldsymbol{\varepsilon}_r$ can be reduced to near zero, even at short forecast times, by the proper choice of $\mathbf{x}_t$. For instance, when a time series of $\mathbf{x}_t$ is defined by analyses, those analyses should be other than those used for the production of the verified forecasts.

3.2.2.4 The Error In The Prediction Of The Mean

A nonzero value of $\boldsymbol{\varepsilon}_m$ is usually expected due to the inevitable presence of systematic model errors that can lead to a drift of the predicted probability distributions in state space. In addition, flaws in the design of the ensemble system can also contribute to the error in the ensemble mean. A testable sufficient condition for $\boldsymbol{\varepsilon}_m \neq \mathbf{0}$ at the longer forecast times is $M^2 = E^2(\delta \mathbf{x}_t) \gg 0$. The sufficient nature of this condition can be seen by first making use of Eqs. (8) and (3.9), which lead to

$$M = E(\delta \mathbf{x}_t) = E(\boldsymbol{\varepsilon}_m) - E(\boldsymbol{\varepsilon}_t) \approx E(\boldsymbol{\varepsilon}_m) \tag{3.14}$$

by taking into account that the magnitude of $E(\boldsymbol{\varepsilon}_t)$ is small and its contribution to $E(\delta \mathbf{x}_t)$ can be neglected if $t_f > 12 - 48$ h.

Because

$$TV = \text{Var}(\delta \mathbf{x}_t) + M^2, \tag{3.15}$$

where Var$(\cdot)$ is the variance function for the time period of verification, $E(\boldsymbol{\varepsilon}_m)$ can contribute to TV. If the climate and the model representation of the climate were both stationary, for an infinitely long forecast, M would be equal to the mean error in the climatological mean state of the model. However, because forecast verification is usually done only for a season of a single year and numerical predictions are for finite forecast times, M typically asymptotes to a value that includes systematic errors in the prediction of the lower frequency variability of the atmosphere.

3.2.2.5 Lorenz Curves

The evolution of VS, TV, and M for the different TIGGE ensembles will be compared by figures whose format will be mostly familiar to the reader. The only unusual comparison we do is based on fitting Lorenz-curves (Lorenz, 1969a, 1982) to both $\sqrt{TV}$ and $\sqrt{VS}$,

and comparing the parameters of the fitted curves. We parameterize the Lorenz-curves by the function

$$\frac{dF}{dt} = (\alpha F + \beta)\left(1 - \frac{F}{F_\infty}\right), \tag{3.16}$$

which was proposed by Dalcher and Kalnay (1987). Our choice of this particular parameterization is motivated by the paper Magnusson and Kallen (2013), which used it successfully to separate the factors that contributed to the improvement of the operational ECMWF forecasts from 1979 to 2011.

In our application of Eq. (3.16), F is either $\sqrt{TV}$ or $\sqrt{VS}$, α is the parameter that describes the (exponential) growth of F for the linear phase of uncertainty dynamics, and F_∞ is the saturation (asymptotic) value of F at long forecast times. For $F = \sqrt{TV}$, the parameter β can be considered a static (time-independent) estimate of the contribution of model errors to the forecast uncertainty tendency, while for $F = \sqrt{VS}$, it can be considered an estimate of the contribution of the technique used for the representation of the effect of model errors to the tendency of $\sqrt{VS}$. If β is smaller for $\sqrt{TV}$ than $\sqrt{VS}$, the chosen technique(s) overestimates the contribution of model errors, while if it is smaller for $\sqrt{VS}$ than $\sqrt{TV}$, it underestimates the contribution of model errors.

We emphasize that Eq. (16) is a crude parameterization of the function that describes the error growth process. It is based on the assumption that the initial error growth process is linear and nonlinear effects become important only later, once the magnitude of the errors becomes sufficiently large. While this is a reasonable assumption for the error growth at the synoptic scales, it is clearly violated by the rapidly saturating errors at the smaller scale. In addition, earlier studies (e.g. Orrell et al., 2001; Vannitsem and Toth, 2002; Nicolis et al., 2009) also demonstrated that model errors tended to lead to a nonlinear short term error growth. Hence, Eq. (16) is expected to provide a better description of the error growth process from the forecast times at which synoptic scale errors become

dominant; and the estimates of β should be considered a particularly crude estimate.

3.2.3 Diagnostics For The Predicted Structure Of The Uncertainties

The local vector $\delta\mathbf{x}_t$ can be decomposed as

$$\delta\mathbf{x}_t = \delta\mathbf{x}_t^{(\|)} + \mathbf{x}_t^{(\perp)}, \tag{3.17}$$

where $\delta\mathbf{x}_t^{(\|)}$ is the component of $\delta\mathbf{x}_t$ that projects onto $\mathbb{S}^K$ and $\mathbf{x}_t^{(\perp)}$ is the component that projects onto the null space of $\mathbf{P}$. Heuristically, the vector $\delta\mathbf{x}_t^{(\|)}$ represents the collection of uncertain local forecast features that the ensemble is able to capture. Likewise, $\mathbf{x}_t^{(\perp)}$ represents the collection of uncertain local forecast features that the ensemble is unable to capture. (This interpretation assumes that $\boldsymbol{\varepsilon}_e \approx \mathbf{0}$.)

The set of normalized eigenvectors $\{\mathbf{u}_k : k = 1, \ldots, K-1\}$ associated with the largest $K-1$ eigenvalues of $\mathbf{P}$ provide a convenient orthonormal basis to compute $\delta\mathbf{x}_t^{(\|)}$ by

$$\delta\mathbf{x}_t^{(\|)} = \sum_{k=1}^{K-1} \left([\delta\mathbf{x}_t]^T \mathbf{u}_k \right) \mathbf{u}_k. \tag{3.18}$$

The origin of the local orthogonal coordinate system defined by the basis vectors $\{\mathbf{u}_k : k = 1, \ldots, K-1\}$ is the local ensemble mean $\bar{\mathbf{x}}$. Heuristically, these basis vectors describe the structure of the local uncertain forecast features.

The efficiency of $\mathbb{S}^K$ in capturing $\delta\mathbf{x}_t$ can be assessed by comparing

$$TVS = E\left(\|\delta\mathbf{x}_t^{(\|)}\|^2 \right) = E\left(\left(\delta\mathbf{x}_t^{(\|)} \right)^T \delta\mathbf{x}_t^{(\|)} \right) \tag{3.19}$$

to TV. TVS always satisfies the relation $TVS \leq TV$, with the equality indicating the ideal situation, in which $\delta\mathbf{x}_t$ lies entirely in $\mathbb{S}^K$. There are three reasons why TVS can be smaller than its optimal value of TV. First and most importantly, $\mathbb{S}^K$ may not provide a

perfect representation of the space in which $\boldsymbol{\varepsilon}_r$ evolves. Second, the origin may be shifted, that is, $\boldsymbol{\varepsilon}_m \neq \mathbf{0}$. Third, $\boldsymbol{\varepsilon}_t$ typically has no significant projection on $\mathbb{S}^K$, which reduces TVS at short forecast lead times.

Finally, it should be noted that there are approaches different than ours to define an orthogonal basis for the investigation of the evolution of ensemble perturbations. For instance, Leutbecher and Lang (2014) defined an orthogonal basis by the leading right singular vectors of the tangent linear version of the model, while Zagar et al. (2015) by the normal mode functions consisting of vertical structure functions, each associated with a set of horizontal Hough functions.

3.2.4 Estimation Of The Expected Value

We estimate the expected value by either an average over all forecasts of the same lead time and all locations ℓ in the verification region, or an average over all forecasts of the same lead time. In the former case, the result is a scalar that depends only on the forecast lead time, while in the latter case, the result is a field of grid point values that depends on the forecast lead time.

We compute diagnostics for forecasts that were started between 1 January, 2012 0000 UTC and 29 February, 2012 1800 UTC. The diagnostics are computed for the entire forecast range of each ensemble system. Diagnostics that require the estimation of temporal means are computed by taking averages over all forecasts of equal forecast time. Spatiotemporal means for the NH extratropics are computed by averaging the temporal means over all locations between 30°N and 75°N.

3.3 The Atmospheric Flow

For the time period of our investigation, we describe the synoptic scale transients with the help of the eddy kinetic energy equation (EKE) and the low frequency transients by the zonal anomalies of the seasonal mean flow.

3.3.1 High-Frequency (Synoptic Scale) Transient Components Of The Flow

The EKE equation of Orlanski and Katzfey (1991), also see Orlanski and Chang (1993) and Chang (2000), is

$$\frac{\partial}{\partial t}\langle K_e\rangle = \overbrace{-\langle\nabla\cdot\mathbf{v}K_e\rangle}^{1}\overbrace{-\langle\nabla\cdot\mathbf{v}'\phi'\rangle}^{2}\overbrace{-\langle\omega'\alpha'\rangle}^{3}\overbrace{-\langle\mathbf{v}'\cdot(\mathbf{v}'_{\mathbf{3}}\cdot\nabla_3)\mathbf{v}_{\mathbf{m}}-\mathbf{v}'\cdot\overline{(\mathbf{v}'_{\mathbf{3}}\cdot\nabla_3)\mathbf{v}'}\rangle}^{4}$$
$$\overbrace{-[\omega' K_e]_s+[\omega' K_e]_t}^{5}\overbrace{-[\omega'\phi']_s+[\omega'\phi']_t}^{6}\overbrace{+\langle(Residue)\rangle}^{7}. \tag{3.20}$$

In this equation, the prime indicates the eddy component of the state variables and K_e is the eddy kinetic energy given by

$$K_e = \frac{1}{2}\mathbf{v}'\cdot\mathbf{v}', \tag{3.21}$$

where $\mathbf{v}'$ is the eddy component of the horizontal wind vector. The symbols ∇, ∇_3, $\mathbf{v_m}$ and $\mathbf{v}'_{\mathbf{3}}$ denote the horizontal nabla operator, the three dimensional nabla operator, the mean component of the horizontal wind vector and the eddy component of the three-dimensional wind vector for pressure vertical coordinate, respectively. Otherwise, the conventional notation is used for the state variables. The bar denotes a seasonal mean, while the symbol $\langle\cdot\rangle$ indicates a vertical average in pressure coordinate system, and $[\cdot]$ indicates a surface integral across the surface (s) or top (t) of the model atmosphere.

The first term of the right-hand side describes the horizontal eddy kinetic energy transport, the second term is the geopotential flux convergence, the third term is the baroclinic energy conversion, and the fourth term is the barotropic energy conversion. Term five describes the vertical eddy kinetic energy transport through the bottom and the top surfaces, while term six represents the transport of eddy potential energy through the same

surfaces. Finally, the last term is the residue term that represents the bulk effect of the errors of the numerical calculations and all processes unaccounted for by the other terms. The most important such process is dissipation, which usually makes the residue term negative. Because not all variables necessary for the computation of the terms of the eddy kinetic energy equation are available in the TIGGE data set, we use data from the ERA Interim reanalysis for the description of the flow. Unlike the previous diagnostics which are calculated on local volumes, the eddy kinetic energy equation is calculated on the global grid.

The computation of the terms of the eddy kinetic energy equation starts with a decomposition of the spatiotemporally evolving atmospheric state variables into a spatially varying time-mean component and a spatiotemporally evolving eddy component. We compute the time-mean for January-February-March, because even though all forecasts start in January and February, some of them end in March. The time-mean component of the geopotential height field at the 500 Pa pressure level is shown in Fig. B.37 for both the ECMWF and NCEP analyses: the time-mean flow has a dominantly zonal wavenumber two structure, with negative zonal anomalies in the Pacific and the Atlantic storm track regions and positive zonal anomalies in the exit regions of the storm tracks.

Figure B.38 shows the time-mean eddy kinetic energy (top left panel) and the time-mean of the three terms of the eddy kinetic energy that dominate the changes in the time-mean eddy kinetic energy (other three panels). These terms represent baroclinic energy conversion (top right panel), barotropic energy conversion (bottom left panel) and the horizontal transport of the eddy kinetic energy (bottom right panel). The largest local maxima of the eddy kinetic energy are located in the eastern sector of the Pacific storm track. These maxima are due to the local generation of kinetic energy by baroclinic energy conversion and the transport of eddy kinetic energy generated upstream by baroclinic energy conversion.

Barotropic energy conversion is the (nonlinear) transfer of kinetic energy between the synoptic scale transients and the seasonal mean flow. Where it is positive, kinetic energy is transferred to the synoptic scale eddies, while where it is negative, kinetic energy is transferred to the seasonal mean flow. While kinetic energy is transferred to the seasonal mean flow in the exit regions of the storm tracks, kinetic energy is transferred to the synoptic scale eddies over North America and Western Europe.

A comparison of Figs. B.38 and B.37 show the close relationship between the high- and low-frequency transients: baroclinic energy conversion at the synoptic scales is the most intense in regions of the negative anomalies of the time-mean flow, while (negative) barotropic energy conversion from the synoptic to the large scales plays a direct role in the slow changes of the large scale flow in regions of positive anomalies. These nonlinear barotropic energy conversion processes control the energy transfer between the high-frequency transients and the slowly varying large scale flow.

3.4 Results On The Predictions Of The Magnitude Of The Uncertainty

In this section, our attention is focused on studying the relationship between the evolutions of VS, TV, and M in the forecasts. We examine the behavior of both the spatiotemporally and the temporally averaged forms of the three diagnostics.

3.4.1 The Evolution Of VS, TV And M^2 In The Forecasts

3.4.1.1 Diagnostics Based On Averages Over All Forecasts And Locations

We first examine the evolutions of VS, TV, and M^2 in the forecasts qualitatively, with the help of Figs. B.39 and B.40, which show the evolution of the spatiotemporally averaged version of the three quantities. A common feature of the behavior of the different ensembles at analysis time is that VS (green curve) tends to be smaller than TV (black curve). That is, the ensembles have a tendency to underestimate the analysis uncertainty. This feature is the most pronounced for the JMA ensemble and barely noticeable for the

CMC ensemble. Because the match between VS and TV for the latter ensemble is essentially perfect at all forecast times between 12 h and 132 h, the slight difference at analysis time is likely to be due to uncertainty in the proxy for the true state rather than to an underestimation of the magnitude of the initial uncertainty.

For most ensembles, VS quickly (in about 48-72 forecast hours) asymptotes to TV. The rapid recovery of the ensemble variance is particularly notable for the JMA ensemble. The unique short term behavior of this ensemble can be explained by the fact that this is the only ensemble in TIGGE that is purely based on right singular vector initial condition perturbations: because the right singular vectors grow very rapidly during the optimization period, which is 48 h for the JMA ensemble, the magnitude of the analysis perturbations must be small to avoid over-shooting TV at 48 h forecast time. The ensemble that shows a somewhat similar behavior, but with a much less severe underestimation of the analysis uncertainty, is the ECMWF ensemble. The similarity is not by accident; some of the initial condition perturbations in the ECMWF ensemble are right singular vectors. The underestimation of the uncertainty in the ECMWF ensemble is much less severe, because it mixes the right singular vectors with perturbations produced by an ensemble of data assimilations. The latter perturbations grow much slower than the right singular vectors, but their initial magnitude is larger, leading to an overall larger magnitude of the analysis perturbations.

The ensemble for which the gap between VS and TV remains relatively large at all forecast times is the CMA ensemble. This behavior is most likely due to the feature of the CMA ensemble that it is one of only two ensembles in TIGGE that does not use any "parameterization" scheme to continuously increase the magnitude of the evolving forecast perturbations. The only other TIGGE member that does not "parameterize" the effects of model uncertainty is the KMA ensemble, but for that ensemble the gap between VS and TV is smaller than for the CMA ensemble at initial time, which helps at the longer

forecast times as well.

While there are more pronounced differences in the evolution of M^2 (purple curve) than in the evolution of TV and VS between the different ensembles, there are also some important similarities: the relative contribution of M^2 to TV is the largest at analysis time, while M^2 is typically an order of magnitude smaller than TV. The shape of the curves that describe the evolution of M^2 suggests that M^2 is growing due to systematic errors in the prediction of the low-frequency transients.

3.4.1.2 Sensitivity Of The Results To The Choice Of The Proxy For The True State

To test the robustness of the diagnostic results shown in Figs. B.39 and B.40 to the choice of the proxy $\mathbf{x}_t$ for the true state, we computed some of the diagnostics using analyses from different centers for the definition of $\mathbf{x}_t$. An example for the results of these calculations is shown in Figure B.41, which shows the diagnostics for the UKMO ensemble using ECMWF or NCEP analyses as $\mathbf{x}_t$. While the results slightly change quantitatively,* the choice of $\mathbf{x}_t$ has no effect on our qualitative observations about the relationships between the evolution of the diagnostics.

3.4.1.3 Diagnostics Based On Averages Over All Forecasts

To save space, we show temporally averaged forms of the diagnostics only for selected ensembles. Figures B.42 and B.43 show TV at four different forecast times for the ECMWF and the CMC ensembles, respectively. We choose these two ensembles, because while they are among the better performing ensembles, they are generated by using different techniques for the generation of the initial conditions and the representation of the effects of model errors. As discussed later, the results of our investigation also suggest that ECMWF and CMC use different tuning conditions for their systems. The two figures show

*For instance, the underestimation of TV by VS at analysis time for the ensemble is even smaller when the ECMWF analyses are used as proxy for the true states.

that the differences between the two systems are the largest at analysis time and rapidly diminishing with increasing forecast time. These diminishing differences are the result of the fast growth of TV in the storm track regions in both ensembles. By forecast time 360-h, the forecast uncertainty becomes the largest in the exit region of the Pacific and the Atlantic storm tracks, where the magnitude of the (positive) zonal anomaly of the time-mean flow is the largest (see Fig. B.37). This behavior suggests that uncertainties in the prediction of both the high- and low-frequency transients contribute to the large forecast uncertainty. As for the differences at analysis time, the large differences in the region of the Tibetan Plateau and the Himalayas are most likely due to differences in the orography of the verified and verification data sets. The related local maximum quickly disappears with increasing forecast time for both ensembles. The more important initial difference is the markedly lower magnitude of the uncertainty in the storm track regions, especially over the Pacific, for the CMC ensemble. This initial difference is most likely due to the larger magnitude of the CMC analysis perturbations.

Figures B.44 and B.45 show VS at four different forecast times for the ECMWF and the CMC ensembles, respectively. A striking feature of the two figures is their similarity. In particular, the growth of the ensemble is the fastest in the storm track regions. Both ensembles clearly underestimate the analysis uncertainty, but the underestimation is more severe for the ECMWF ensemble. For the CMC ensemble, VS slightly underestimates TV in the storm track region at all forecast times, but, in general, it correctly captures the main patterns of uncertainty. For the ECMWF analysis, VS slightly underestimates TV at all locations at 120 h forecast time, but later there are an increasing number of locations where it overestimates TV. The most important shortcoming of VS for the ECMWF ensemble is that at 360-h lead time, its maximum of the Atlantic is shifted westward (from Iceland to Newfoundland) compared to the related maximum of TV.

To shed some light on the origin of the aforementioned shift, we also plot the evolu-

tion of M for the ECMWF ensemble (Fig. B.46). The growing component of M, which becomes dominant by forecast time 360 h, is associated with errors in the prediction of the zonal anomalies of the time mean-flow, which are shown in Fig. B.37. In particular, the dipole pattern in the Atlantic region in the lower right panel of Fig. B.46 indicates a significant error in the prediction of the large scale flow in that region. The fact that the growing component of M is due to growing systematic errors in the prediction of the large scale flow is further illustrated by Fig. B.47. This figure takes advantage of the property of M that it can be written as

$$M = E\left(\mathbf{x}_t\right) - \frac{1}{K}\sum_{k=1}^{K} E\left(\mathbf{x}^k\right). \tag{3.22}$$

The figure shows $E\left(\mathbf{x}_t\right)$ (black lines) and the K-member ensemble of $E\left(\mathbf{x}^k\right)$ (grey lines). M would be zero, if the mean of the grey curves was identical with the black curve. This ideal situation cannot occur, because the ensemble members do not capture the zonal anomalies of the large scale flow. Figure B.48, which shows the same type of spaghetti diagram for four different ensembles at 360 h forecast time, illustrates that the behavior we have just described is a common shortcoming of all ensemble systems.

3.4.2 Spectral Evolution Of VS And TV In The Forecasts

We illustrate the spectral evolution of the forecast uncertainty and the ensemble spread with the example of the ECMWF and the CMC ensembles. Figure B.49 shows the spectral evolution of the two quantities for the meridional component of the wind vector at 500 hPa. The left panels show the evolution of the spectral distribution of VS, and the right panels show the evolution of the spectral distribution of TV for the meridional wind vector component of the state vector. This figure was obtained by first computing the zonal power spectra at each latitude within the verification region, then computing the

meridional average of the zonal spectra.

The evolution of the spectra of TV (right panels) is very similar for the two ensembles. At analysis time, the two spectra are white, except for a slow drop of the power at the highest wave numbers. Initially, the growth is the fastest at wave number 9, but with increasing forecast time, the maximum power gradually shifts to wave number 4 by forecast time day 14. This shift of the maximum power toward the lower wave numbers is the result of the earlier saturation of the power at the wave numbers at which it grows faster initially. At the sub-synoptic scales (wave numbers larger than about 12-14) the errors saturate as Lorenz (1969b) described: the steepness of the saturation spectra is k^{-3}, along which the uncertainty and the spread saturate at increasingly larger scales as forecast time increases.

The difference in the spectra of the ensemble spread between the two ensembles (left panels) is the largest at analysis time: the shape of the spectra of VS for the ECMWF ensemble is similar to that of the spectra of TV, but it has significantly less power; while the shape of the spectra of VS for the CMC ensemble is different from that of the spectra of TV, but the average power of the VS and TV spectra are more similar than for the ECMWF ensemble. The evolution of VS captures the generic characteristics of the evolution of TV for both ensembles. An interesting difference, however, is that at forecast time day 14, the power has its maximum at wave number 5 rather than wavenumber 4 for both ensembles.

3.4.3 Qualitative Description Of The Forecast Uncertainty Growth Process

The information provided by Figs B.49, B.42, B.43, and B.46 suggests the following general description of the forecast uncertainty growth process in the NH extratropics:

1. the initial growth is the fastest at the synoptic scales that are most sensitive to baroclinic instability;

2. the growing, and later saturating, synoptic scale features of uncertainty fill the region

that extends from the entrance region of the Pacific storm track to the exit region of the Atlantic storm track;

3. because the position and the spatial structure of the storm track regions are controlled by the slowly varying large-scale flow, the low-frequency transients have a major influence on the spatiotemporal distribution of the forecast uncertainty;

4. as uncertainties start saturating at the scales most sensitive to baroclinic instability, the wave number of dominant instability gradually shifts towards the larger scales (lower wave numbers),

5. at the sub-synoptic scales (zonal wave numbers larger than about 12-14), the uncertainty saturates as predicted by Lorenz's theory.

The diagnostic results for the ensemble spread (Figs B.44, B.45, B.46, and B.49) suggest that all ensemble forecast systems can capture the main characteristics of the error growth process. Figures B.39 and B.40 show, however, that the differences between the models, analysis systems, and ensemble generation techniques have important effects on the accuracy of the quantitative prediction of the uncertainty.

3.4.4 Lorenz-Curve Based Analysis Of The Evolution Of VS And TV With Increasing Forecast Time

3.4.4.1 Estimation Of The Parameters

Obtaining estimates of the parameters of the Lorenz curve requires the availability of dF/dt at each forecast time. We compute an approximate value of $dF/dt\,(t_f)$ by the centered-difference scheme

$$\frac{dF}{dt}(t_f) = \frac{F\,(t_f + \Delta t) - F\,(t_f - \Delta t)}{2\Delta t}, \tag{3.23}$$

where Δt =12 h. We compute $dF/dt(t_f)$ for all forecast times $t_f \geq 12$ h, but we ignore the first ($t_f = 12$ h) data point when fitting the Lorenz curve for $F = \sqrt{TV}$. The reason to exclude this data point is that the estimates of both dF/dt and F have a large relative error due to the error ε_t in the verification data.

The parameters α, β, and F_∞ can be estimated by fitting a second-order polynomial to the pairs of $(F, dF/dt)$ data using the standard least-square approach for function fitting. The estimates of α and β are

$$\alpha = -p_2 F_\infty \tag{3.24}$$

$$\beta = p_0, \tag{3.25}$$

where p_2 is the coefficient of the second order term of the fitted polynomial and p_0 is the coefficient of the zeroth order term, while F_∞ is the positive root of the fitted polynomial.

3.4.4.2 Comparison Of The Estimated Parameters Of The Lorenz-Curves

The estimated parameters of the Lorenz-curves for the forecast uncertainty ($F = \sqrt{TV}$) and the ensemble spread ($F = \sqrt{VS}$) are summarized by Table A.3. In addition, the pairs of $(F, dF/dt)$ values and the fitted curves for four selected ensembles (ECMWF, NCEP, CMC, JMA) are shown by Fig. B.50 for the forecast uncertainty, and by Fig. B.51 for the ensemble spread. The two figures show that there are some outliers in terms of the quality of the fit of the curves to the data. In particular, in the upper left panel of Fig. B.50 the forecast error is significantly overestimated at both 12 h and 24 h forecast times. Hence, in addition to the data point for 12 h forecast time, the data point for 24 h forecast time is also excluded from the estimation of the parameters of the Lorenz-curve. The bottom right panel of Fig. B.51 shows that the curve fitting for the ensemble spread has also failed. This failure suggests that the rapid initial growth of the SV perturbations

is inconsistent with the growth process that a Lorenz-curve can describe. The results for the CMC ensemble (bottom left panel) are also somewhat suspect.

To make the interpretation of the values of α in Table A.3 more transparent, the table also includes the values of $e^{\alpha t}$ for $t = 1$ day. This quantity is the daily growth rate of F for the time range in which the uncertainty dynamics is linear to a good approximation. For the ECMWF, NCEP, UKMO, and KMA ensembles, the daily linear growth rate for the forecast uncertainty is 1.5. The same growth rate for the ensemble spread is slightly lower (1.4) for the ECMWF, UKMO, and KMA ensembles, and slightly higher (1.6) for the NCEP ensemble. The latter result suggests that the underestimation of TV by VS in the NCEP ensemble (top right panel of Fig. B.39) is not due to an underestimation of the linear error growth. The CMC and the CMA ensembles underestimate the linear error growth to various degrees. (As indicated earlier, the estimation of the parameters for $\sqrt{VS}$ has failed for the JMA ensemble.)

We recall from section 3.4.4.1 that we consider the parameter β a measure of the contribution of model errors to the growth of the forecast uncertainty $\sqrt{TV}$, and a measure of the contribution of the parameterization of model errors to the ensemble spread $\sqrt{VS}$. (The model error parameterization techniques used by the ensembles of the different centers are listed in Table A.2.) Examining the results for β, it should always be kept in mind that the estimates are particularly sensitive to the errors of curve fitting at the short forecast times.

The most interesting conclusions that can be drawn about β are the following. An inefficient representation of the model error forcing by the ensemble is the most likely main source of the underestimation of the forecast uncertainty by the NCEP ensemble (top right panel of Fig. B.39). The good balance between TV and VS in the CMC ensemble (bottom left panel of Fig. B.39) is the result of a compensation of the underestimation of the linear error growth by an overestimation of the model error forcing. We note that CMC made

major changes to their ensemble system, including the representation of model error forcing, in 2013 (recall that our study is for 2012 data). The results of Reynolds et al. (2015) suggest that those changes led to significant changes in the behavior of the CMC ensemble. It is highly likely, therefore, that the Lorenz curves would behave differently than reported here for the current operational configuration of the CMC ensemble. The estimate of $\sqrt{TV_\infty}$ is an estimate of the saturation level of the forecast uncertainty. All ensembles, except for the ECMWF ensemble, underestimate this saturation level ($\sqrt{VS_\infty} < \sqrt{TV_\infty}$).

For comparison, we also computed the parameters of Lorenz-curves for the more conventional choice of the 500 hPa geopotential height rather than a combination of the virtual temperature and the two horizontal wind components in the layer between 1000 hPa and the 200 hPa levels. In these calculations, we computed $\sqrt{TV}$ and $\sqrt{VS}$ without localization. (The expected value was estimated by temporal averaging only.) Because a norm based on the geopotential height gives much less weight to errors at the smaller scales than a norm based on energy, the parameters of these Lorenz curves are far less sensitive to error growth at the smaller scales than those that we described earlier. As expected, we found the curve fitting more robust for the geopotential height. In particular, we did not have to exclude data points at the short forecast times and the curve fitting never failed. The results are summarized in Table 3. The linear error growth is lower than before (1.2 rather than 1.5-1.6) and uniform for the different ensembles. All ensembles do a good job with capturing this linear growth rate, which suggests that they are all tuned to perform well for diagnostics based on the 500 hPa geopotential height. The ECMWF ensemble also correctly simulates the contribution of model errors to the forecast uncertainty (β), while the NCEP ensemble still underestimates the contribution of model errors to the forecast uncertainty. Interestingly, the new Lorenz curves suggest that the CMC ensemble greatly underestimates the contribution of model errors, while the earlier curves indicated that it greatly overestimated the contribution of model errors. This discrepancy suggests that the

CMC ensemble introduces the effect of model errors at the wrong scales and/or model levels. We suspect that the changes made to the CMC system in 2013 greatly reduced this discrepancy.

3.4.4.3 The Relationship Between $d|M|/dt$ And $|M|$

We also prepared Lorenz-curve style figures for $|M|$ (Fig. B.52), but without fitting curves to the pairs of data points. These figures are evidently more dissimilar for the four ensembles than those for $\sqrt{VS}$ and $\sqrt{TV}$. This difference is most likely due to the fact that model errors can be diverse, while the evolution of $\sqrt{VS}$ and $\sqrt{TV}$ is primarily driven by the universal sensitivity of the synoptic scale transients to random perturbations. In general, the shape of the curves is consistent with our earlier conclusion that the growth of $|M|$ is dominated by errors in the prediction of low frequency transients.

3.5 Results On Predictions Of The Structure Of Uncertainty

We examine the evolution of TVS in the forecasts by comparing it to the evolution of TV and VS.

3.5.1 Diagnostics Based On Averages Over All Forecasts And Locations

For the examination of the spatiotemporally averaged form of TVS, we return to Figs. B.39 and B.40. While comparing the evolution of TVS and TV, it should be kept in mind, that the two diagnostics must satisfy the relation $TVS \leq TV$ at all forecast times. In addition, smaller differences between TVS and TV indicate a better performance of the ensemble in capturing the local forecast uncertainty.

The generally small differences betweens TVS (red curves) and TV (black curves) beyond the 48-72 h forecast times for the TIGGE ensembles is in a good agreement with the behavior that was reported by KSS for a research ensemble. This result confirms that an ensemble which is operationally attainable in size can efficiently span the local linear

space of the forecast uncertainty beyond forecast times 48-72 h.

For most ensembles, the asymptotic value of TVS tends to a level that is lower than the saturation level of TV. The only ensemble that is virtually unaffected by this problem is the ECMWF ensemble, while the ensemble that it affects most severely is the CMA ensemble. This result for the ECMWF analysis suggests that it captures all important forecast uncertainties that develop during the investigated time period.

A small difference between TVS and TV indicates that the ensemble captures the important uncertain forecast features, but it does not guarantee that the ensemble correctly captures the magnitude of those uncertainties. That ideal situation is indicated by a small difference between TVS and VS in addition to the small difference between TVS and TV. The ECMWF ensemble satisfies this requirement at the forecast times where the difference between TVS and TV is small. As for the relationship between VS and TVS in the other ensembles, with the exception of the JMA ensemble, VS tends to overestimate TVS at the analysis and the short forecast times. In other words, the ensembles compensate for part of the loss of the magnitude that results from not capturing all uncertain analysis and forecast features by over-inflating the magnitude of the correctly captured features. For most ensembles, this strategy pays off at later forecast times in the form of a good match between VS and TVS (NCEP, UKMO and CMA), or between VS and TV (CMC and KMA). The fact that for the JMA ensemble VS remains smaller than TVS at all forecast times suggests that the magnitude of the initial perturbations in that ensemble could be increased somewhat without negative effects on the performance of the ensemble.The CMC ensemble tends to underestimate the magnitude of the uncertainties that it captures correctly.

A comparison of the two left panels of Fig. B.39 suggests that the outstanding performance of the ECMWF ensemble is due to a combination of a faster convergence of TVS to TV in the first 48-72 forecast hour and a continued convergence beyond those forecast

times, leading to an almost perfect fit of the two curves beyond forecast time 192 h .

3.5.2 Diagnostics Based On Averages Over All Forecasts

Figures B.53 and B.54 show the spatiotemporal evolution of TVS in the forecasts for the ECMWF and the CMC ensembles, respectively. The main spatial patterns in these figures are very similar to those of TV in Figs. B.42 and B.54, but the magnitudes of the patterns are typically smaller for TVS. This result indicates that the ensembles correctly predict the regions of main forecast uncertainty, but they do not capture all uncertain forecast features in those regions. The ratio between TVS and TV tends to be smaller for the CMC than the ECMWF ensemble, which indicates that the CMC ensemble is less efficient in capturing uncertain forecast features.

A particularly good example for the ensembles capturing a large part of the structure of the most important local uncertainty is the matching pair of local maxima in TVS and TV over Iceland for the ECMWF ensemble at forecast time 360 h. This shows that the shift of the related maximum in the spread (Fig. B.44) is the result of not capturing all uncertain forecast features over Iceland and overestimating the spread over Newfoundland. Another example is the local maximum of TV over the northeast Pacific and the related maxima of TVS and VS. In that case, the maxima of VS is at the right location, but its magnitude is overinflated, most likely by the representation of the effect of model uncertainty, to compensate for the loss of spread due to the ensemble not capturing all uncertain forecast features.

An overinflation of VS in large regions is a general property of the ECMWF ensemble (Fig. B.55). This result suggests that while the parameters of the algorithms for the representation of the effect of model uncertainty can be efficiently tuned to achieve the good match shown between TV, VS, and TVS (Fig. B.39), they do not guarantee a near optimal representation of the local structure of the forecast uncertainty. We note that the

CMC ensemble does not suffer from a similar problem.

4. CONCLUSIONS*

4.1 Regionally Enhanced Global Data Assimilation

Our results demonstrate that REG DA has the potential to improve global forecasts without a significant increase in computation cost by using already available limited area forecast information. For an NWP center that prepares operational global and regional model products, the only extra computations associated with the approach are the interpolation of the model field onto a common grid and the creation of the composite state.

We found that by including COAMPS information in the formation of the innovations, we were able to improve global forecasts, especially over the North American region, where the approach led to better prediction of both tropical cyclones and frontal passages. While the results are more mixed for the Northeast Pacific and European regions, the overall results for the NH are clearly positive. For the experiments shown in this paper, the blending weight was kept constant across the different regions, but different weights could be potentially used to achieve a near optimal performance in all regions.

The inclusion of regional model information in the global data assimilation system led to improved track forecasts for Hurricane Sandy for both tested values of the blending coefficients. The speed of the movement of Sandy along its track was better forecast between 48 h and 96 h, while the cross track error was reduced the most at lead times 96 h and 120 h.

The next step for this project is to evaluate the system at the operational resolution of the Navy models. In these future experiments, NAVGEM and COAMPS will be run at horizontal resolution T425 and 10 km, respectively, the blending will take place on a

*Parts of Section 4 are reprinted with permission from "Forecast Uncertainty Dynamics in the THORPEX Interactive Grand Global Ensemble (TIGGE)" by M. Herrera, I. Szunyogh, and J. Tribbia, 2016. Monthly Weather Review, 144, 2739-2766, Copyright 2016 by the American Meteorological Society.

Gaussian grid that corresponds to a spectral resolution T1023, and the inner loop iterations of NAVDAS-AR will be carried out at resolution of T119.

Since this project started, the Navy has transitioned to using a hybrid 4D-Var system, which uses an ensemble-based approach to enhance the estimate of the background error covariance matrix. The REG DA system can easily be implemented within a hybrid framework, with each global ensemble member having a corresponding regional ensemble member. A composite state can then be created for each ensemble member to find an ensemble based covariance matrix that can be combined with the static covariance.

In this study, we focused on the investigation of the effects of REG DA on the the global model forecasts. The analyses produced by the approach can also provide initial conditions for the limited area model. Our investigation into the effects of REG DA on the COAMPS forecasts is underway and its results will be reported in a future study.

4.2 Forecast Uncertainty Dynamics In TIGGE

We proposed a description of the forecast uncertainty growth process summarized by an itemized list in Section 3.4.3, which emphasizes the earlier rapid growth of uncertainties at the synoptic-scales and the later shift of the dominant error growth towards the large scales in the spatiotemporal evolution of forecast uncertainty. We found that the TIGGE ensembles were able to capture the main characteristics of the error growth process. The results also showed, however, that the accuracy of the quantitative prediction of the forecast uncertainty was strongly system dependent. While the best performing ensembles did well with respect to the spatiotemporally averaged diagnostics, the location dependent temporally averaged diagnostics revealed that even the best performing ensembles had large errors in the representation of the local properties of the uncertainty.

A result that was particularly interesting from both a theoretical and a practical point of view was the typical growth of the error in the prediction of the mean state with increasing

forecast time. Such a drift of the predicted probability distribution of the state in state space greatly reduces the utility of the longer range (week 2 and beyond) ensemble predictions. If the growth of the mean forecast uncertainty is primarily due to shortcomings of the models and/or the ensemble generation techniques, there is hope that it can be greatly reduced by refining the models and the ensemble generation techniques. However, if it is due to some fundamental properties of the atmospheric dynamics, it may turn out to be a major barrier to the extension of numerical weather forecasts into the sub-seasonal to seasonal forecast range.

Our analysis showed that the growing errors in the prediction of the mean state were dominantly due to errors in the prediction of low-frequency changes in the large scale flow. One school of thoughts suggests that low frequency variability is a manifestation of the internal (chaotic) variability of the atmospheric dynamics at the large scales (e.g., Legras and Ghil, 1985). If this was true and the models fully captured the internal variability of the atmosphere, the systematic errors in the prediction of the mean state would not increase with forecast time and the ensembles were able to capture the related forecast uncertainty. Hence, one potential explanation for our result is that the models cannot fully capture the internal variability of the atmosphere; for instance, due to poor representation of the atmospheric dynamics in the tropics and/or atmosphere-ocean interactions. The results of Reynolds et al. (2015), for example, suggest that ensemble forecasts tend to lose temporal variability with increasing forecast time. Another potential explanation is that the low frequency variability is not dominated by internal variability of the atmospheric dynamics. In particular, some authors (e.g., Sura et al., 2005) have argued that low frequency variability may be due to state-dependent variations of stochastic feedbacks. Such state-dependent stochastic feedback in a long range prediction is provided by the high-frequency (synoptic-scale) transients.

REFERENCES

Baek, S. J., B. R. Hunt, E. Kalnay, E. Oott, and I. Szunyogh, 2006: Local ensemble Kalman filtering in the presence of model bias. *Tellus, Series A: Dynamic Meteorology and Oceanography*, **58 (3)**, 293–306.

Baek, S.-J., I. Szunyogh, B. R. Hunt, and E. Ott, 2009: Correcting for Surface Pressure Background Bias in Ensemble-Based Analyses. *Mon. Wea. Rev.*, **137 (7)**, 2349–2364.

Berner, J., S. Y. Ha, J. P. Hacker, A. Fournier, and C. Snyder, 2011: Model uncertainty in a mesoscale ensemble prediction system: stochastic versus multi physics representations. *Mon. Wea. Rev.*, **139**, 1972–1995.

Berner, J., G. J. Shutts, M. Leutbecher, and T. N. Palmer, 2009: A spectral stochastic kinetic energy backscatter scheme and its impact on flow-dependent predictability in the ECMWF ensemble prediction system. *J. Atmos. Sci.*, **66**, 603–626.

Bishop, C., B. Etherton, and S. Majumdar, 2001: Adaptive sampling with the ensemble transform kalman filter. part i: Theoretical aspects. *Mon. Wea. Rev.*, **129**, 420–436.

Bougeault, P., and Coauthors, 2010: The THORPEX Iinteractive Grand Gglobal Eensemble (TIGGE). *Bull. Amer. Meteor. Soc.*, **91**, 1059–1072.

Bowler, N., and K. Mylne, 2009: Ensemble transform kalman filter perturbations for a regional ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **135**, 757–766.

Bowler, N. E., A. Arribas, S. E. Beare, K. R. Mylne, and G. J. Shutts, 2009: The local ETKF and SKEB: Uparades to the MOGREPS short-range ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **135**, 767–776.

Buizza, R., M. Leutbecher, and L. Isaksen, 2008: Potential use of an ensemble of analyses in the ecmwf ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **134**, 2051–2066.

Buizza, R., M. Miller, and T. N. Palmer, 1999: Stochastic representation of model uncer-

tainties in the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **125**, 2887–2908.

Buizza, R., J. Tribbia, F. Molteni, and T. Palmer, 1993: Computation of optimal unstable structures for a numerical weather prediction model. *Tellus*, **45**, 388–407.

Chang, E., 2000: Wave packets and life cycles of troughs in the upper troposphere. part ii: Seasonal and hemispheric variations. *Mon. Wea. Rev.*, **128**, 25 – 50.

Charron, M., G. Pellerin, L. Spacek, P. L. Houtekamer, N. Gagnon, H. L. Mitchell, and L. Michelin, 2010: Toward random sampling of model error in the Canadian ensemble prediction system. *Mon. Wea. Rev.*, **138**, 1877–1901.

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

Durran, D., and M. Gingrich, 2014: Atmospheric predictability: Why butterflies are not of practical importance. *J. Atmos. Sci.*, **71**, 2476–2488.

Epstein, E., 1969: Stochastic dynamic prediction. *Tellus*, **6**, 739–759.

Hodur, R. M., 1997: The Naval Research Laboratory's Coupled Ocean/Atmosphere Mesoscale Prediction System (COAMPS). *Mon. Wea. Rev.*, **125 (7)**, 1414–1430.

Hogan, T. F., and Coauthors, 2014: The Navy global environmental model. *Oceanography*, **27**.

Holt, C., I. Szunyogh, G. Gyarmati, S. M. Leidner, and R. N. Hoffman, 2015: Assimilation of Tropical Cyclone Observations: Improving the Assimilation of TCVitals, Scatterometer Winds, and Dropwindsonde Observations. *Mon. Wea. Rev.*, **143 (10)**, 3956–3980.

Holt, C. R., I. Szunyogh, and G. Gyarmati, 2013: Can a moderate-resolution limited-area data assimilation system add value to the global analysis of tropical cyclones? *Mon. Wea. Rev.*, **141 (6)**, 1866–1883.

Hou, D., Z. Toth, Y. Zhu, and W. Yang, 2008: Impact of a stochastic perturbation scheme on ncep global ensemble forecast system. *Proceedings of the 19th AMS Conference on*

Probability and Statistics, 21-24 January 2008, New Orleans, Louisiana.

Houtekamer, P. L., 2002: The use of multiple parameterizations in ensembles. *ECMWF Workshop on Model Uncertainty, 20–24 June 2011*, Reading, UK, ECMWF, 163–174.

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

Kretschmer, M., B. R. Hunt, E. Ott, C. H. Bishop, S. Rainwater, and I. Szunyogh, 2015: A composite state method for ensemble data assimilation with multiple limited-area models. *Tellus*, **67**.

Kuhl, D., and Coauthors, 2007: Assessing predictability with a local ensemble kalman filter. *J. Atmos. Sci.*, **64**, 1116–1140.

Legras, B., and M. Ghil, 1985: Persistent anomalies, blocking and variations in atmospheric predictability. *J. Atmos. Sci.*, **42**, 433–471.

Leith, C., 1974: Theoretical skill of monte carlo forecasts. *Mon. Wea. Rev.*, **102**, 409–418.

Leutbecher, M., and S. T. K. Lang, 2014: On the reliability of ensemble variance in subspaces defined by singular vectors. *Quart. J. Roy. Meteor. Soc.*, **140**, 1453–1466.

Lorenz, E. N., 1969a: Atmospheric predictability as revealed by naturally occurring analogues. *J. Atmos. Sci.*, **26**, 636–646.

Lorenz, E. N., 1969b: Predictability of a flow which possesses many scales of motion. *Tellus*, **21**, 289–307.

Lorenz, E. N., 1982: Atmospheric predictability as revealed by naturally occurring analogues. *Tellus*, **34A**, 505–513.

Magnusson, L., and E. Kallen, 2013: Factors influencing the skill improvements in the ecmwf forecasting system. *Mon. Wea. Rev.*, **141**, 3142–3153.

Merkova, D., I. Szunyogh, and E. Ott, 2011: Strategies for coupling global and limited-area ensemble Kalman filter assimilation. *Nonlinear Processes in Geophysics*, **18 (3)**, 415–430.

Molteni, F., and T. Palmer, 1993: Predictability and finite-time instability of the northern winter circulation. *Quart. J. Roy. Meteor. Soc.*, **119**, 269–298.

Mureau, R., F. Molteni, and T. Palmer, 1993: Ensemble prediction using dynamically conditioned perturbations. *Quart. J. Roy. Meteor. Soc.*, **119**, 299–323.

Nicolis, C., P. Perdigao, and S. Vannitsem, 2009: Dynamics of prediction errors under the combined effect of initial condition and model errors. *J. Atmos. Sci.*, **66**, 766–778.

Orlanski, I., and E. Chang, 1993: Ageostrophic geopotential fluxes and upstream development of baroclinic waves. *J. Atmos. Sci.*, **50**, 212–225.

Orlanski, I., and J. Katzfey, 1991: The life cycle of a cyclone wave in the southern hemisphere. part i: Eddy energy budget. *J. Atmos. Sci.*, **48**, 1972–1998.

Orrell, D., L. Smith, J. Barkmeijer, and T. Palmer, 2001: Model error in weather forecasting. *Nonlin. Proc. Geophys.*, **8**, 357–371.

Ott, E., and Coauthors, 2004: A local ensemble kalman filter for atmospheric data assimilation. *Tellus*, **56**, 263–277.

Palmer, T. N., R. Buizza, F. Doblas-Reyes, T. Jung, M. Leutbecher, G. J. Shutts, M. Steinheimer, and A. Weisheimer, 2009: Stochastic parameterization and model uncertainty. ECMWF Technical Memorandum 598, European Centre for Medium-range Weather Forecasts, 42 pp.

Reynolds, C., E. A. Satterfield, and C. H. Bishop, 2015: Using forecast temporal variability to evaluate model behavior. *Mon. Wea. Rev.*, under review.

Roh, S., M. C. Genton, M. Jun, I. Szunyogh, and I. Hoteit, 2013: Observation quality control with a robust ensemble kalman filter. *Mon. Wea. Rev.*, **141**, 4414–4428.

Roh, S., M. Jun, I. Szunyogh, and M. C. Genton, 2015: Multivariate localization methods for ensemble kalman filtering. *Mon. Wea. Rev.*, **22**, 723–735.

Rosmond, T., and L. Xu, 2006: Development of NAVDAS-AR: Non-linear formulation and outer loop tests. *Tellus*, **58A**, 45–58.

Rotunno, R., and C. Snyder, 2008: A generalization of lorenz's model for the predictability of flows with many scales of motion. *J. Atmos. Sci.*, **65**, 1063–1076.

Satterfield, E., and I. Szunyogh, 2010: Predictability of the performance of an ensemble forecast system: Predictability of the space of uncertainties. *Mon. Wea. Rev.*, **138**, 962–981.

Satterfield, E., and I. Szunyogh, 2011: Assessing the performance of an ensemble forecast system in predicting the magnitude and the spectrum of analysis and forecast uncertainties. *Mon. Wea. Rev.*, **139**, 1207–1223.

Shutts, G. J., 2005: A kinetic energy backscatter algorithm for use in ensemble prediction systems. *Quart. J. Roy. Meteor. Soc.*, **131**, 3079–3102.

Shutts, G. J., 2013: Coarse graining the vorticity equation in the ECMWF integrated forecasting system: the search for kinetic energy backscatter. *J. Atmos. Sci.*, **70**, 1233–1241.

Sura, P., M. Newman, C. Penland, and P. Sardeshmukh, 2005: Multiplicative noise and non-gaussianity: A paradigm for atmospheric regimes? *J. Atmos. Sci.*, **62**, 1391–1409.

Swinbank, R., and Coauthors, 2016: The TIGGE project and its achievements. *Bull. Amer. Meteor. Soc.*, **97**, 49–67.

Szunyogh, I., 2014: *Applicable Atmospheric Dynamics*. 1st ed., World Scientific.

Szunyogh, I., E. Kostelich, G. Gyarmati, D. Patil, B. Hunt, E. Kalnay, E. Ott, and J. Yorke, 2005: Assessing a local ensemble kalman filter: Perfect model experiments with the ncep global model. *Tellus*, **57A**, 528–545.

Szunyogh, I., E. J. Kostelich, G. Gyarmati, E. Kalnay, B. R. Hunt, E. Ott, E. Satterfield, and J. A. Yorke, 2008: A local ensemble transform Kalman filter data assimilation system for the NCEP global model. *Tellus A*, **60 (1)**, 113–130.

Tennant, W. J., G. J. Shutts, A. Arribas, and S. A. Thompson, 2011: Using a stochastic kinetic energy backscatter scheme to improve MOGREPS probabilistic forecast skill. *Mon. Wea. Rev.*, **139**, 1190–1206.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at nmc: The generation of perturbations. *Bull. Amer. Meteor. Soc.*, **74**, 2317–2330.

Toth, Z., and E. Kalnay, 1997: Ensemble forecasting at ncep and the breeding method. *Mon. Wea. Rev.*, **125**, 3297–3319.

Tribbia, J. J., and D. P. Baumhefner, 2004: Scale interactions and atmospheric predictability: an updated perspective. *Mon. Wea. Rev.*, **132**, 703–713.

Vannitsem, S., and Z. Toth, 2002: Short-term dynamics of model errors. *J. Atmos. Sci.*, **59**, 2594–2604.

Wang, X., and C. Bishop, 2003: A comparison of breeding and ensemble transform kalman filter ensemble forecast schemes. *J. Atmos. Sci.*, **60**, 1140–1158.

Wasserstein, R., and N. Lazar, 2016: The ASA's statement on p-values: Context, process, and purpose. *The American Statistician*, **70:2**, 129–133.

Wei, M., Z. Toth, R. Wobus, and Y. Zhu, 2008: Initial perturbations based on the ensemble transform (et) technique in the ncep global operational forecast system. *Tellus*, **60**, 62–79.

Xu, L., T. Rosmond, and R. Daley, 2005: Development of NAVDAS-AR: Formulation and initial tests of the linear problem. *Tellus*, **57A**, 546–559.

Yoon, Y., B. Hunt, E. Ott, and I. Szunyogh, 2012: Simultaneous global and limited-area ensemble data assimilation using joint states. *Tellus A*, **64 (0)**.

Zagar, N., R. Buizza, and J. Tribbia, 2015: Modal analysis of atmospheric predictability based on the ECMWF ensemble. *J. Atmos. Sci.*, **72**, under review.

APPENDIX A

TABLES

Table A.1: Summary of the model parameters of the experiments.

	Global Resolution	Inner Loop Resolution	Regional Resolution	Intermediate Resolution	Blending Weight
Control	T119L50	T47	N/A	N/A	N/A
Blend Skip	T119L50	T119	N/A	T319	N/A
REG DA 30%	T119L50	T119	32 km	T319	0.30
REG DA 100%	T119L50	T119	32 km	T319	1.00

Table A.2: Ensemble Forecast Systems Included From TIGGE

NWP Center	Representation of Model Error and Uncertainty	Initial Perturbation Strategy	Max Forecast Lead Time	Ensemble Size
ECMWF	SKEB/SPPT	Singular Vectors & EDA	360 Hr	50 Members
NCEP	STTP	Ensemble Transform & Rescaling	384 Hr	20 Members
UKMO	SKEB	ETKF	360 Hr	14 Members
CMA	None	Bred Vectors	240 Hr	14 Members
CMC	SKEB/SPPT	EnKF	384 Hr	20 Members
KMA	None	Bred Vectors	240 Hr	23 Members
JMA	SPPT	Singular Vectors	216 Hr	50 Members
Météo-France	Multi-Physics	Singular Vectors & Evolved Singular Vectors	72 Hr	34 Members

		α $[day^{-1}]$	β $[J^{\frac{1}{2}}day^{-1}]$	$e^{\alpha t}$ (t=1 day)	F_∞ $[J^{\frac{1}{2}}]$
ECMWF	$\sqrt{TV}$	0.39	0.0	1.5	86
	$\sqrt{VS}$	0.32	3.7	1.4	86
NCEP	$\sqrt{TV}$	0.39	1.1	1.5	90
	$\sqrt{VS}$	0.46	0.0	1.6	80
CMC	$\sqrt{TV}$	0.32	3.0	1.4	91
	$\sqrt{VS}$	0.17*	8.5*	1.2*	88
JMA	$\sqrt{TV}$	0.45	0.0	1.6	85
	$\sqrt{VS}$	X	X	X	X
UKMO	$\sqrt{TV}$	0.38	1.3	1.5	88
	$\sqrt{VS}$	0.32	2.9	1.4	79
CMA	$\sqrt{TV}$	0.51	1.3	1.7	91
	$\sqrt{VS}$	0.28	2.6	1.3	80
KMA	$\sqrt{TV}$	0.41	0.2	1.5	86
	$\sqrt{VS}$	0.34	2.9	1.4	81

Table A.3: Estimates of the parameters of the Lorenz curves for the different ensembles. The symbol X indicates parameters for which the estimation process failed, while * indicates estimates that most likely have unusually large errors.

		α $[day^{-1}]$	β $[m\ day^{-1}]$	$e^{\alpha t}$ (t=1 day)	F_∞ $[m]$
ECMWF	$\sqrt{TV}$	0.16	1.5	1.2	126
	$\sqrt{VS}$	0.18	1.5	1.2	120
NCEP	$\sqrt{TV}$	0.15	3.0	1.2	129
	$\sqrt{VS}$	0.20	0.7	1.2	111
CMC	$\sqrt{TV}$	0.13	3.1	1.1	134
	$\sqrt{VS}$	0.18	0.7	1.2	120
JMA	$\sqrt{TV}$	0.17	2.7	1.2	118
	$\sqrt{VS}$	0.18	2.3	1.2	106
UKMO	$\sqrt{TV}$	0.16	1.5	1.2	126
	$\sqrt{VS}$	0.16	1.7	1.2	108
CMA	$\sqrt{TV}$	0.15	3.5	1.2	132
	$\sqrt{VS}$	0.19	0.5	1.2	104
KMA	$\sqrt{TV}$	0.18	1.6	1.2	122
	$\sqrt{VS}$	0.18	1.6	1.2	109

Table A.4: Estimates of the parameters of the Lorenz curves for the different ensembles using 500 hPa geopotential height without localization in the calculation of $\sqrt{TV}$ and $\sqrt{VS}$.

APPENDIX B

FIGURES

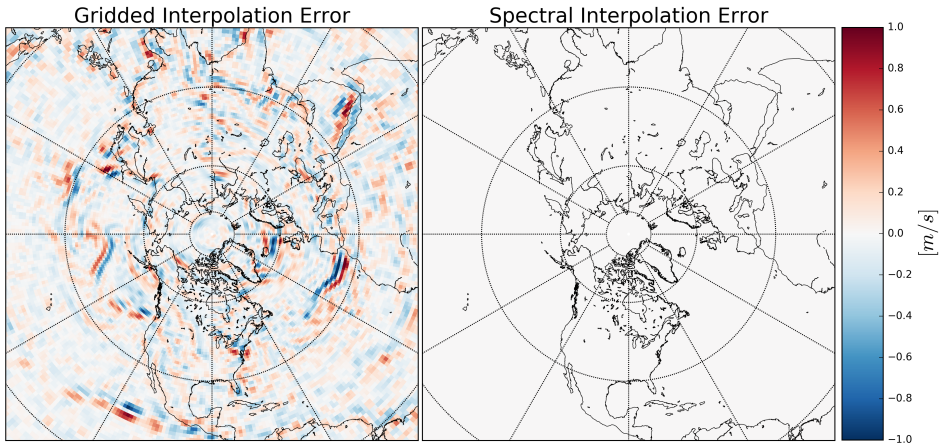


Figure B.1: Comparison of the errors of the two options for the interpolations. Shown is the near surface zonal wind field after an interpolation from T119 to T319 and then back to T119. The left panel shows the result for option 1, while the right panel shows the result for option 2 (the method used in all experiments.)

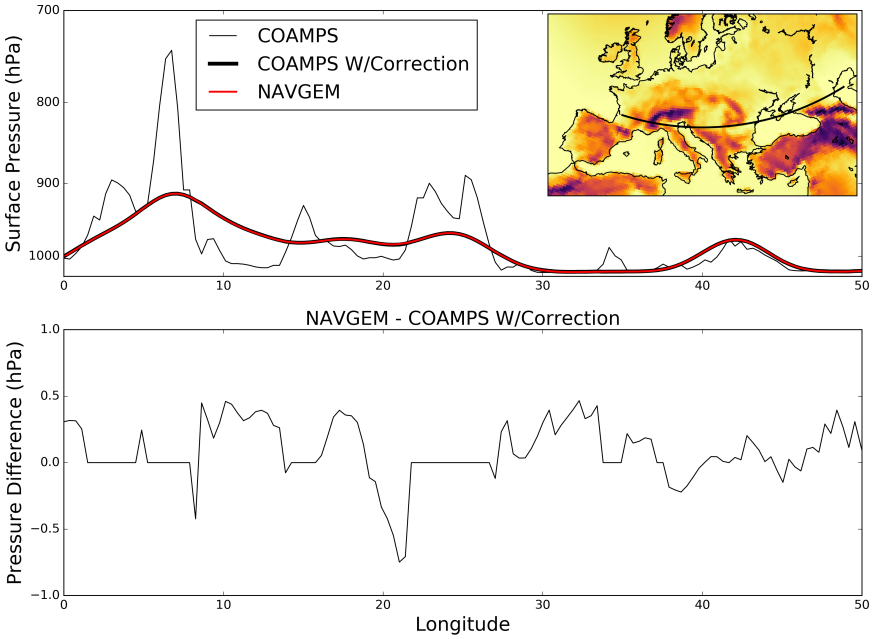


Figure B.2: Illustration of the difference between the COAMPS and the NAVGEM surface pressure before and after the correction of the COAMPS surface pressure for the orography difference. The pressure values are shown along the latitude segment indicated by a black line in the insert of the top panel. The top panel shows the surface pressure values for NAVGEM and for COAMPS before and after the correction for the orography difference. The bottom panel shows the difference between the NAVGEM surface pressure and the corrected COAMPS surface pressure.

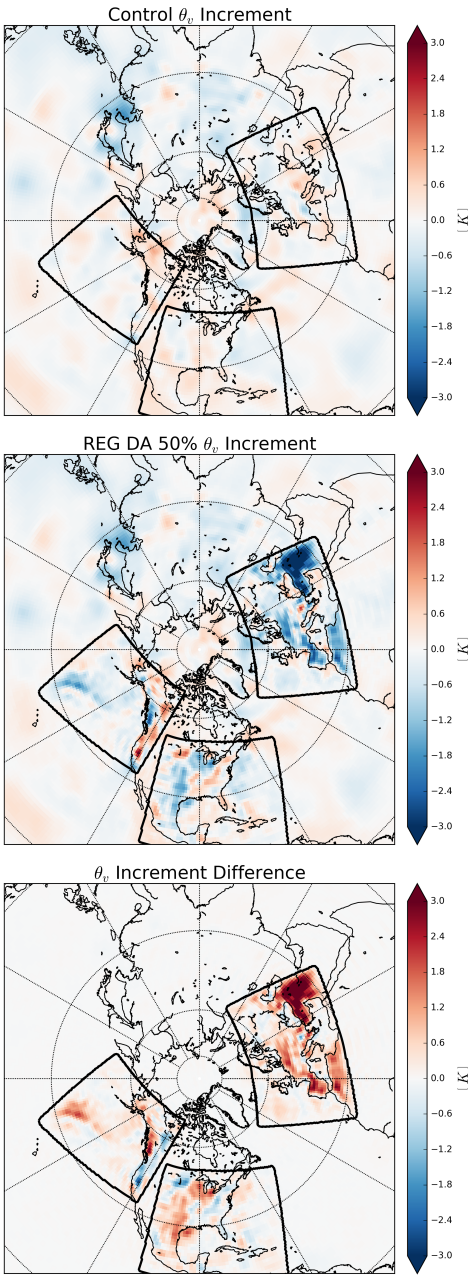


Figure B.3: Virtual potential temperature analysis increment for the (top) blend skip control and the (middle) 50% REG DA experiment at the model level nearest to the surface. The bottom panel is the difference between the two increments.

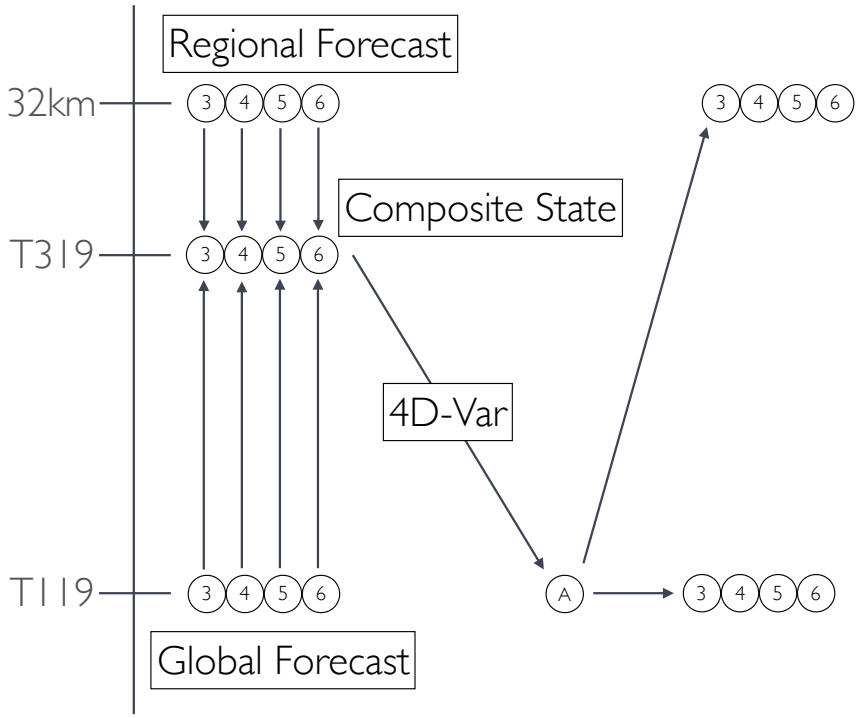


Figure B.4: Schematic illustration of the implementation of REG DA on the U.S. Navy's model. Each filled circle represents a state vector for a given lead time. For simplicity, only four time steps are shown here, but in reality the process is continuously cycled for the entire time period of the experiments.

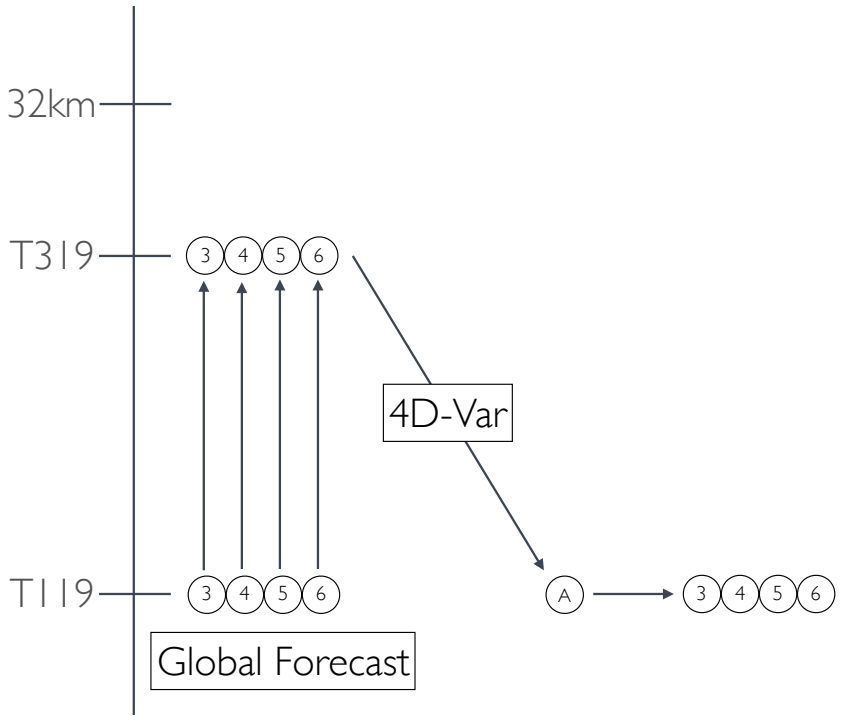


Figure B.5: Same as Fig. B.4, except for the blend skip control experiment.

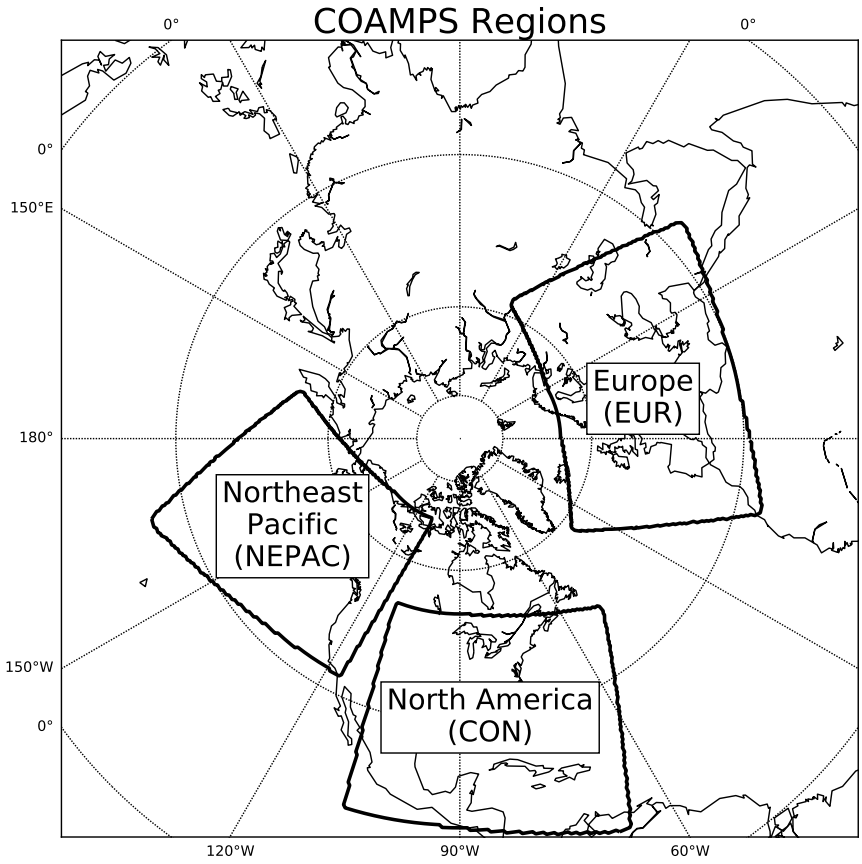


Figure B.6: The three COAMPS regions used in the experiments.

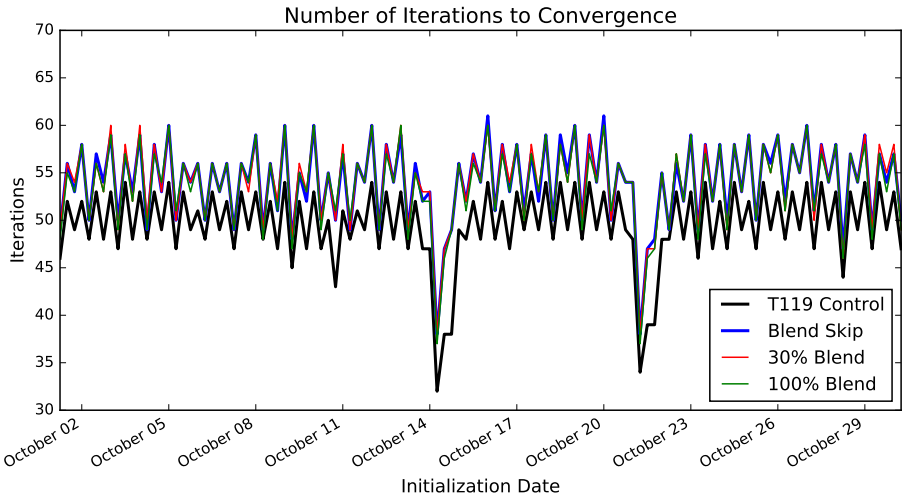


Figure B.7: The number of iterations of the inner loop required to achieve the prescribed reduction of the cost function. Shown are the number of iterations for the control, the blend skip control, and the 30% and 100% REG DA experiments.

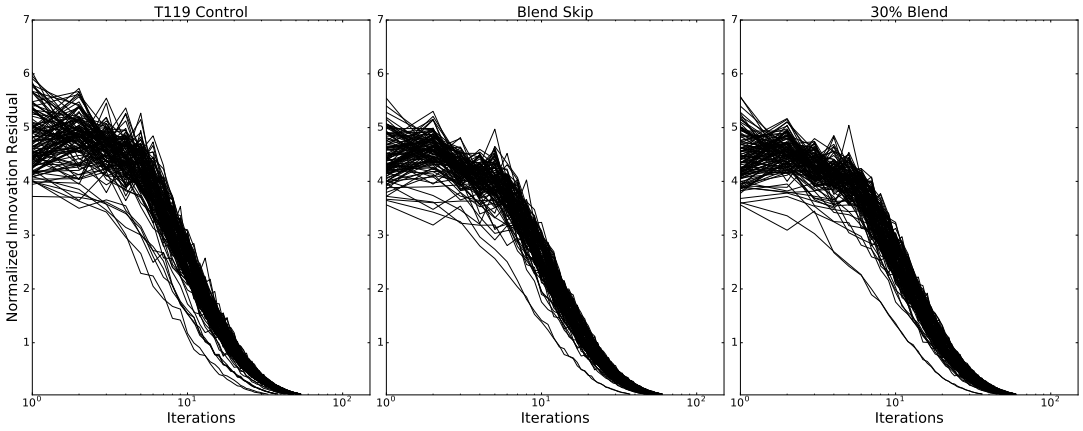


Figure B.8: The evolution of the residual with the iteration steps. Results are shown (left) control, (middle) blend skip control, and (right) 30% REG DA experiment. Each line represents different DA cycles.

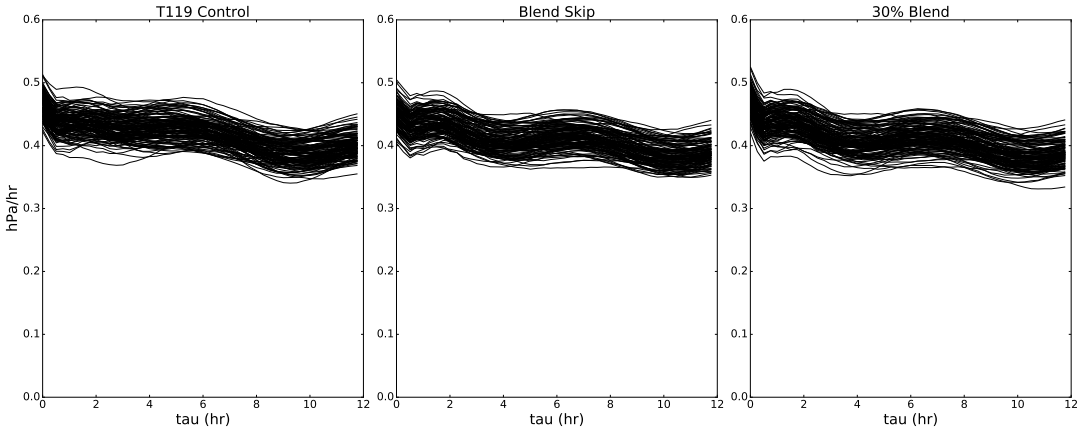


Figure B.9: Global root-mean-square of the surface pressure tendency for the first 12 forecast hours. Shown are the results for (left) control, (middle) blend skip control, and (right) 30% REG DA experiment.

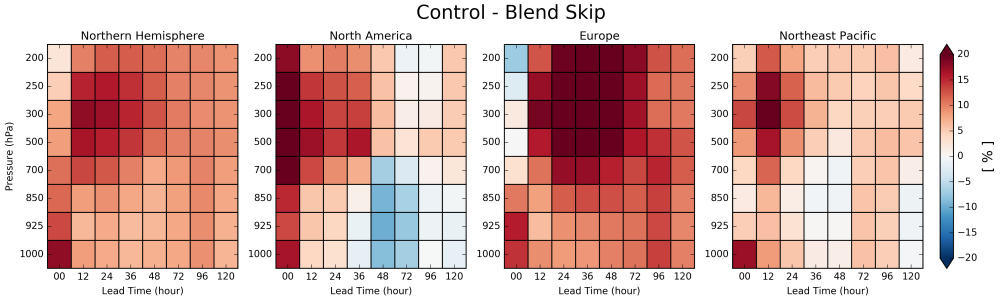


Figure B.10: Impact diagnostics for the control and the blend skip control experiments for the geopotential height. Red shades indicate forecasts that are more accurate for the blend skip control experiment, while blue shades indicate forecasts that are more accurate for the control experiment.

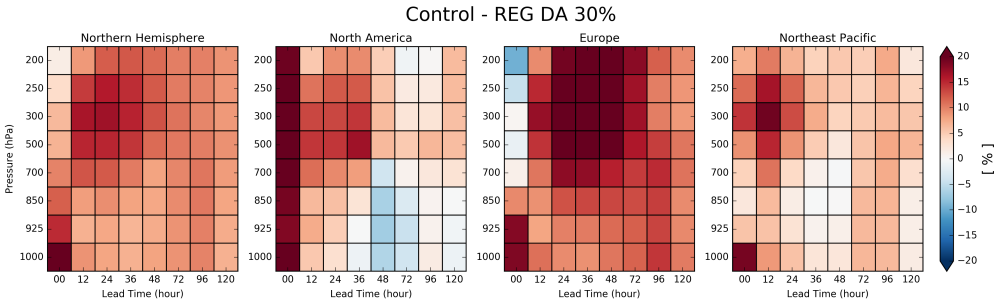


Figure B.11: Impact diagnostics for the blend skip control and 30% REG DA experiments for the geopotential height. Red shades indicate forecasts that are more accurate for REG DA, while blue shades indicate forecasts that are more accurate for the blend skip control.

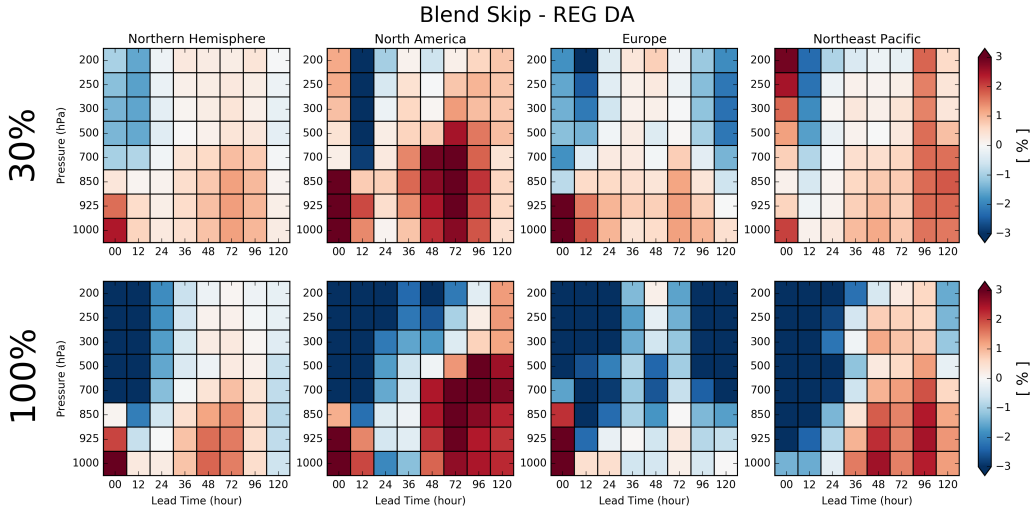


Figure B.12: Impact diagnostics for the blend skip control and the (top row) 30% and (bottom row) 100% REG DA experiments for the geopotential height. Red shades indicate forecasts that are more accurate for REG DA, while blue shades indicate forecasts that are more accurate for the blend skip control.

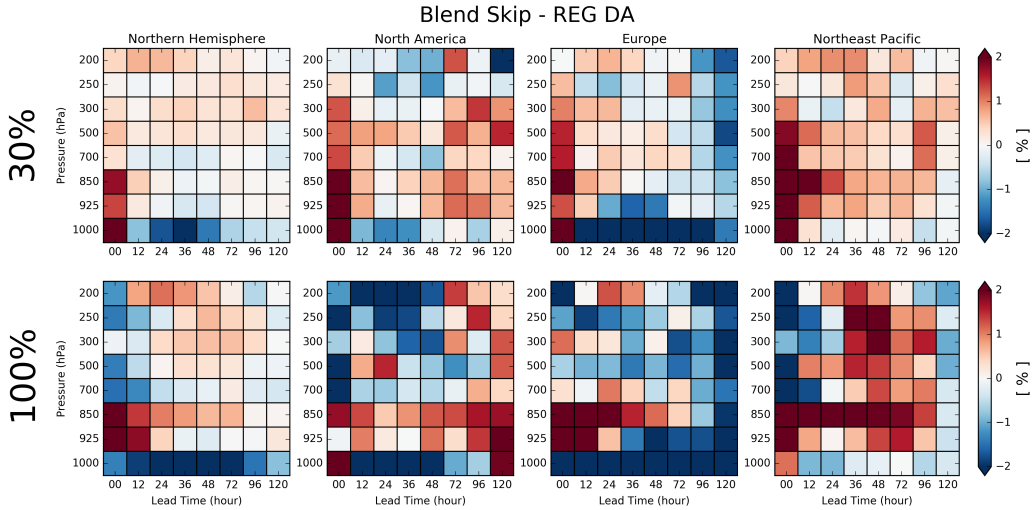


Figure B.13: Same as Fig B.12, except for temperature.

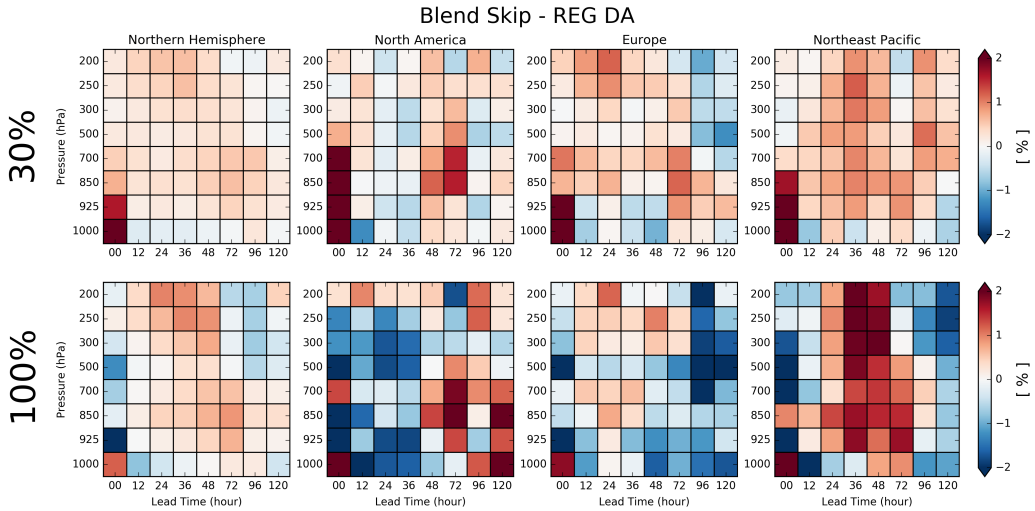


Figure B.14: Same as Fig B.12, except for the zonal component of the wind.

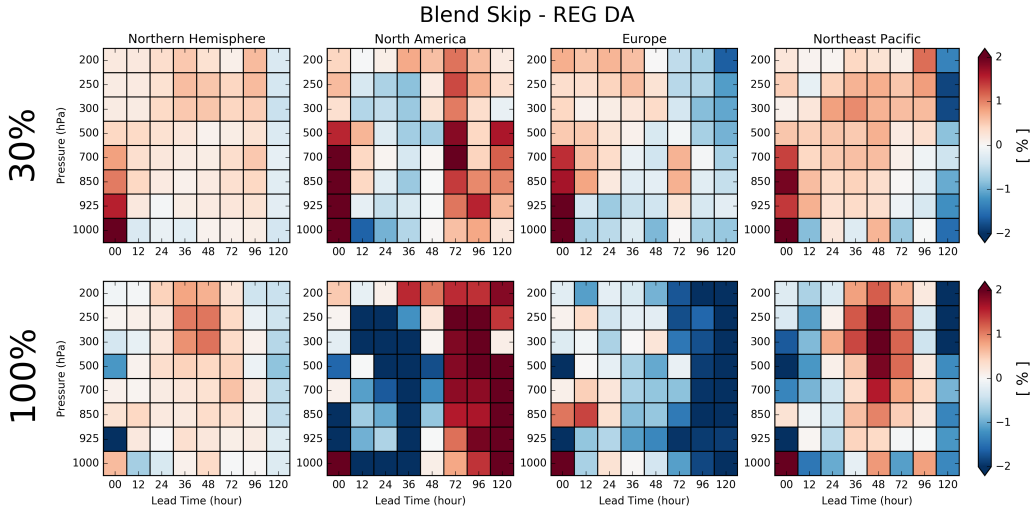


Figure B.15: Same as Fig B.12, except for the meridonal component of the wind.

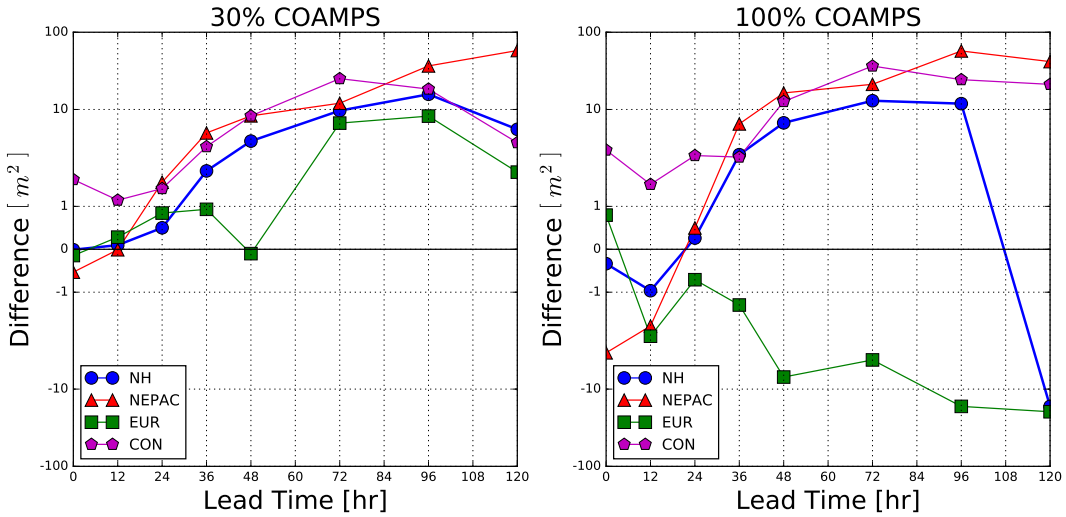


Figure B.16: Evolution of the MSE difference between the experiments for the 850 hPa geopotential height with forecast lead time. Shown are the results for the comparison between the blend skip and the (left) 30% blend REG DA experiment and (right) 100% blend REG DA experiment. Positive values indicate improvement and negative values indicate degradation due to REG DA.

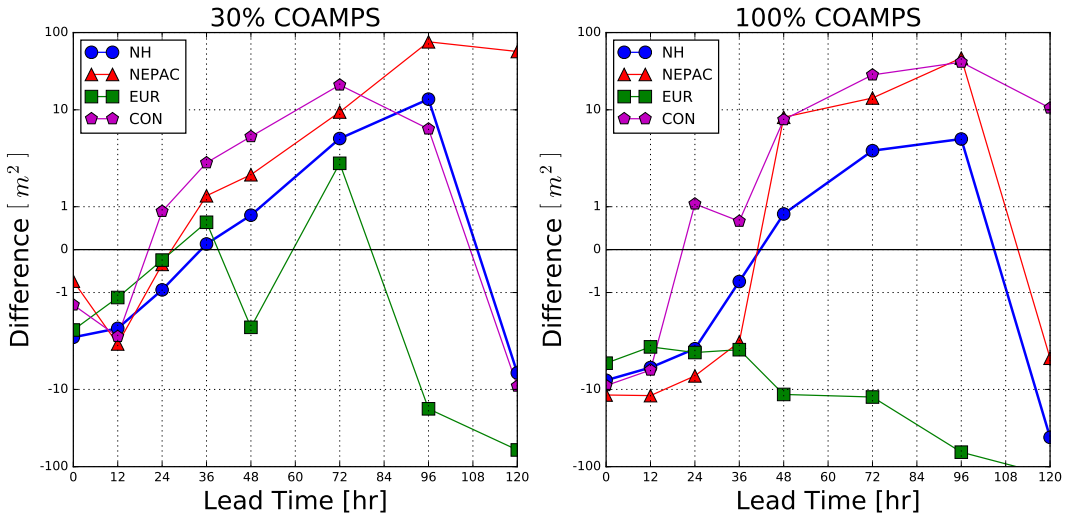


Figure B.17: Same as Fig. B.16, except now for the 500 hPa level.

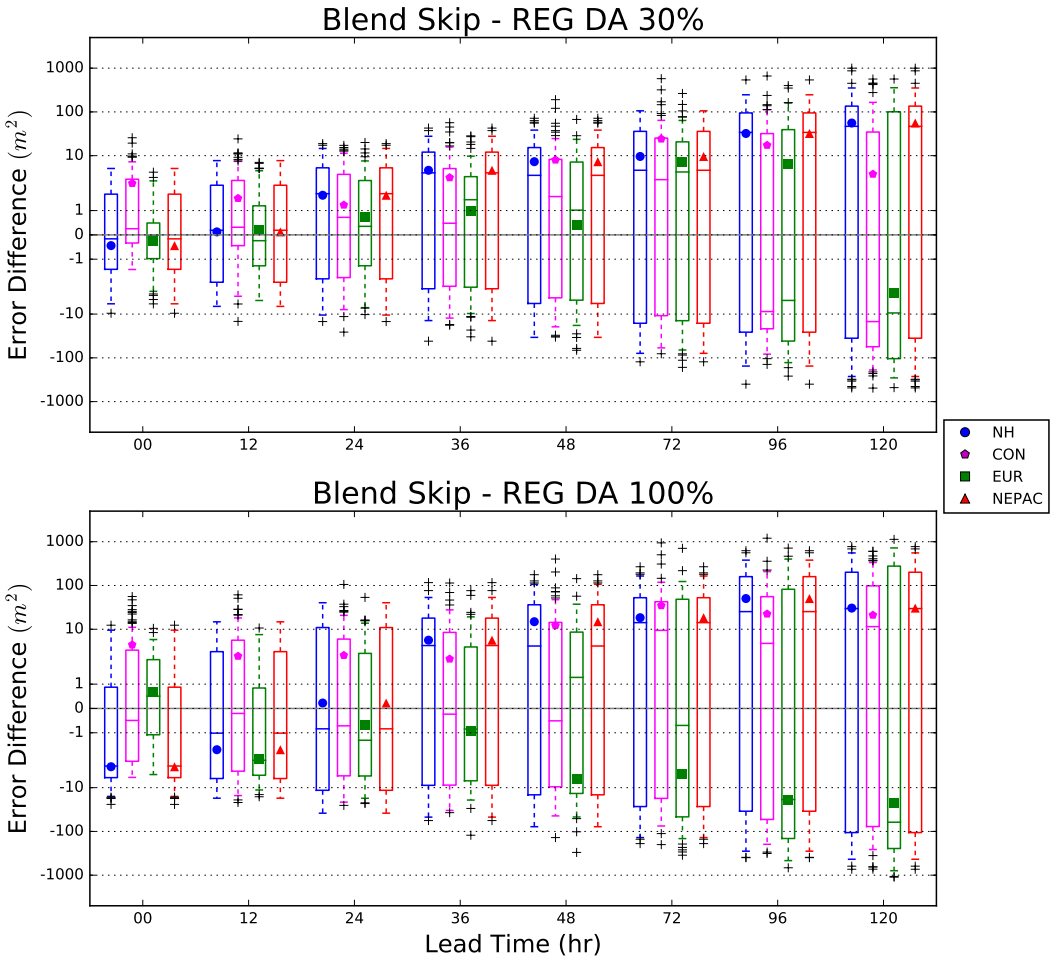


Figure B.18: Box and whisker diagram for the MSE difference between the blend skip control and the REG DA experiments for geopotential height at 850 hPa. Each marker represents the mean of the distribution, the line is the median value, and crosses indicate outliers.

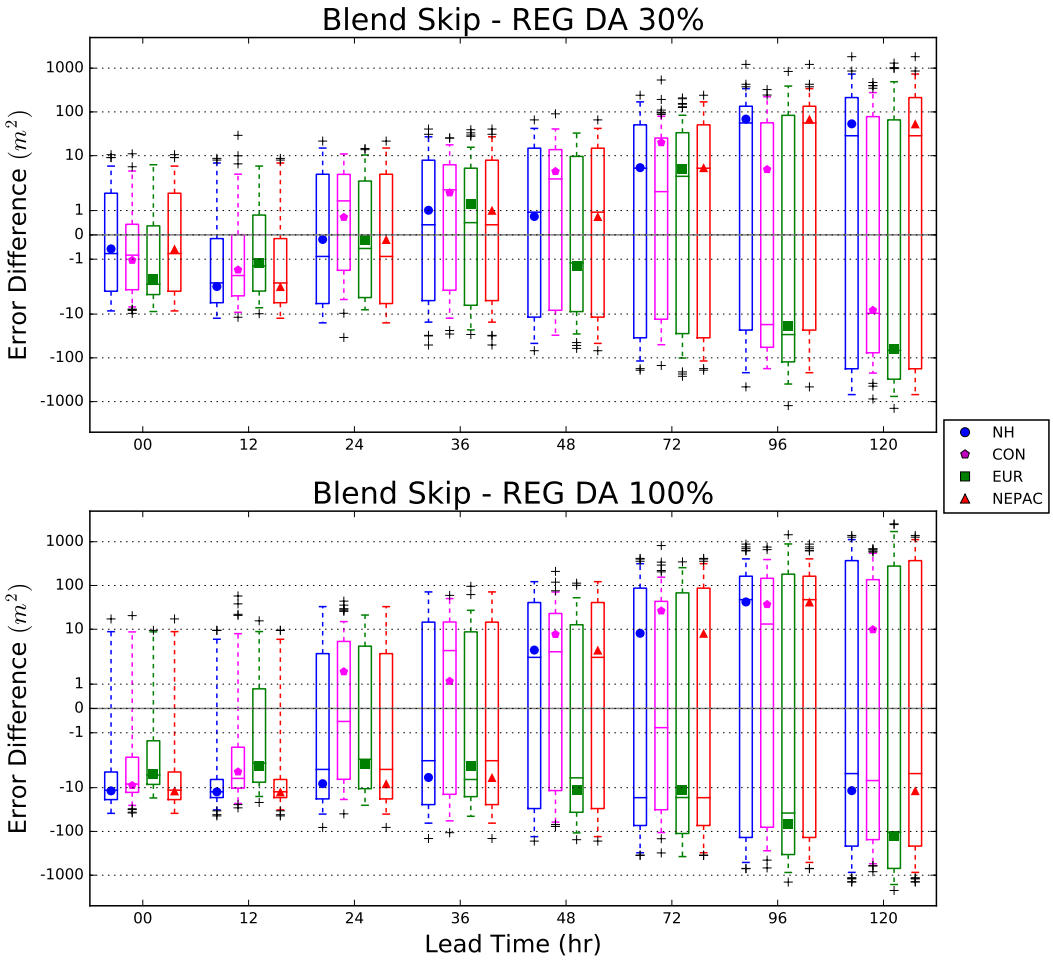


Figure B.19: Box and whisker diagram for the MSE difference between the blend skip control and the REG DA experiments for geopotential height at 500 hPa. Each marker represents the mean of the distribution, the line is the median value, and crosses indicate outliers.

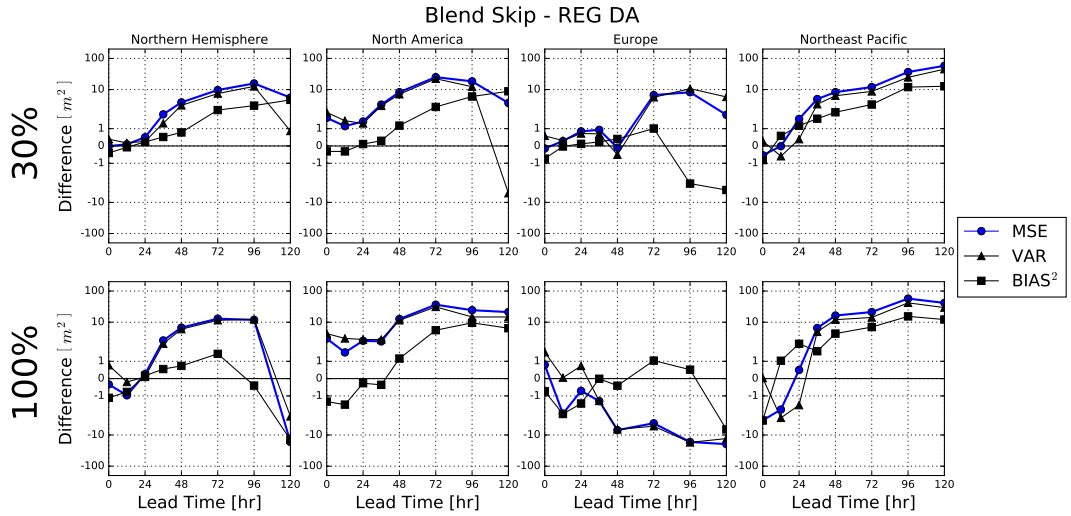


Figure B.20: Evolution of the decomposition of the MSE difference for 850 hPa geopotential height with forecast lead time. Shown are the results for the difference between the blend skip control and the (top row) 30% blend experiment and (bottom row) 100% blend experiments for the (left) NH, (second from left) North American domain (second from right) European domain, (right) Northeast Pacific domain.

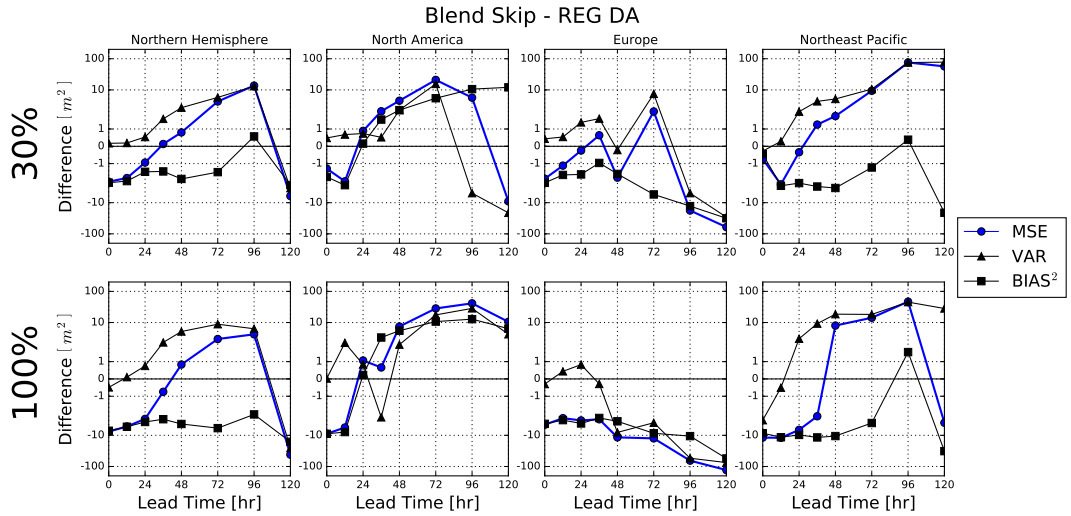


Figure B.21: Same as Fig. B.21, except now for the 500 hPa level.

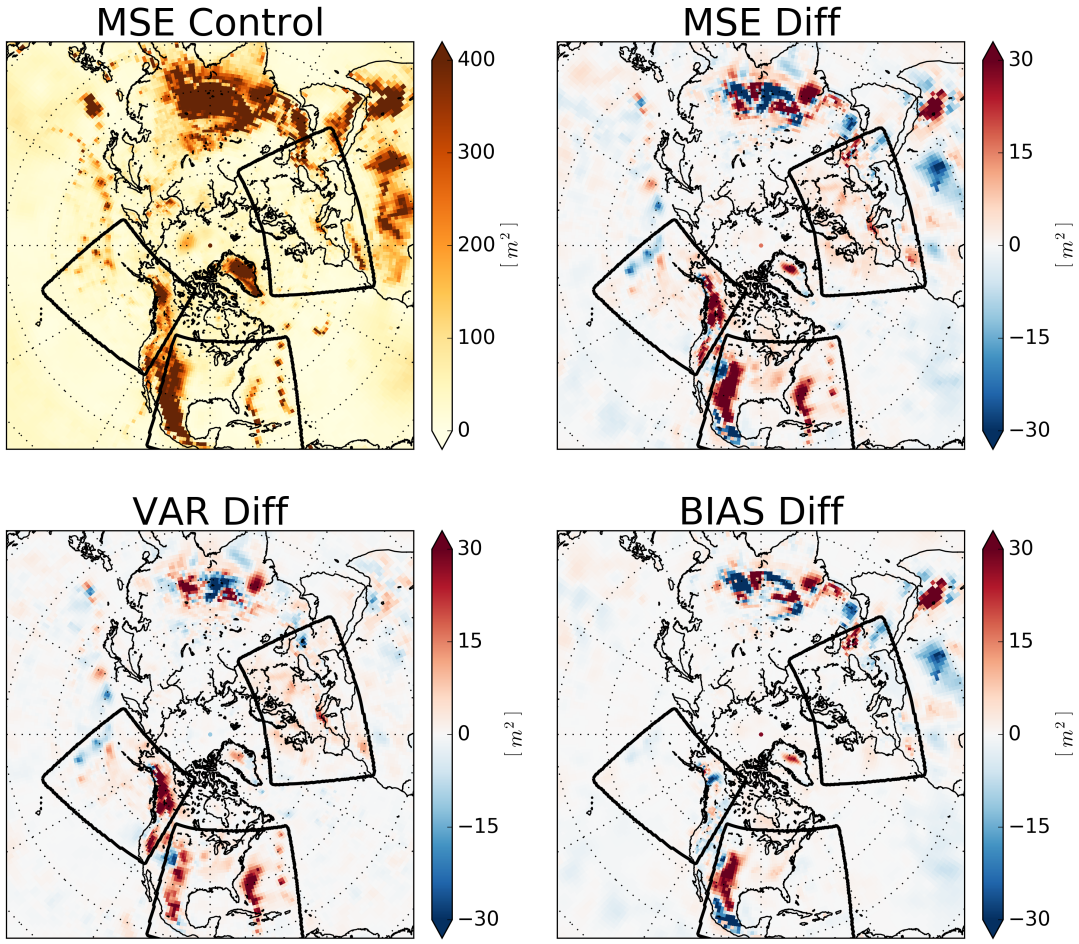


Figure B.22: Decomposition of the MSE for the 1000 hPa geopotential height field at analysis time. Shown are (top left) the MSE for the blend skip control experiment (top right) the MSE differences for the blend skip control and the 30% REG DA experiment, (bottom left) the difference between the error variance, and (bottom right) the difference between square bias. In the three difference figures, red shades indicate improvement and blue shades indicate degradation due to REG DA.

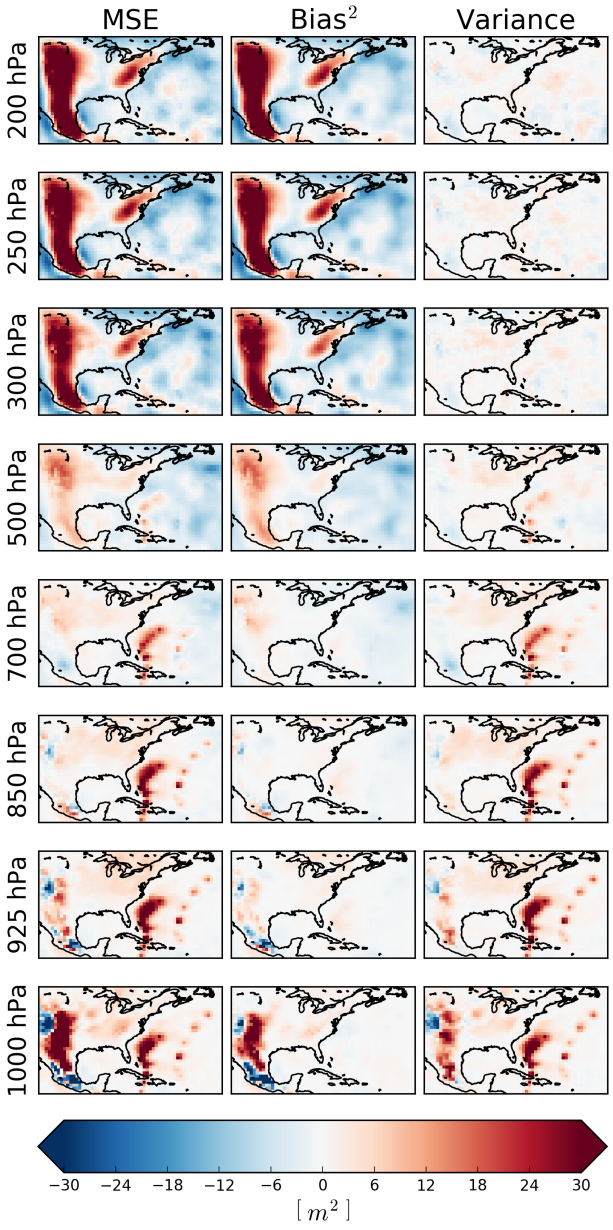


Figure B.23: Vertical structure of the decomposition of the MSE difference between the blend skip control and 30% blend REG DA experiment for geopotential height. The different rows show results for different pressure levels, with pressure deceasing from bottom to top. Shown are (left) the MSE difference, (middle) difference in the square bias and (right) difference in the variance. Red shades indicate improvement and blue shades indicate degradation due to REG DA.

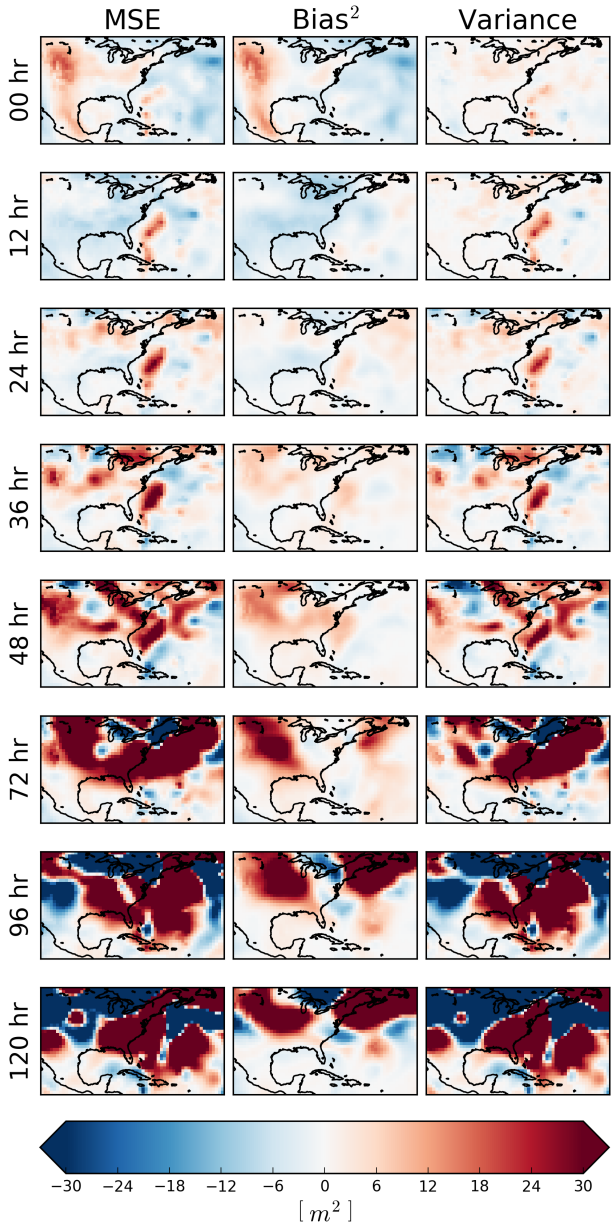


Figure B.24: Temporal evolution of the MSE difference between the blend skip control and 30% blend REG DA experiment for geopotential height. The different rows show results for different lead times. Shown are (left) the MSE difference, (middle) difference in the square bias and (right) difference in the variance. Red shades indicate improvement and blue shades indicate degradation due to REG DA.

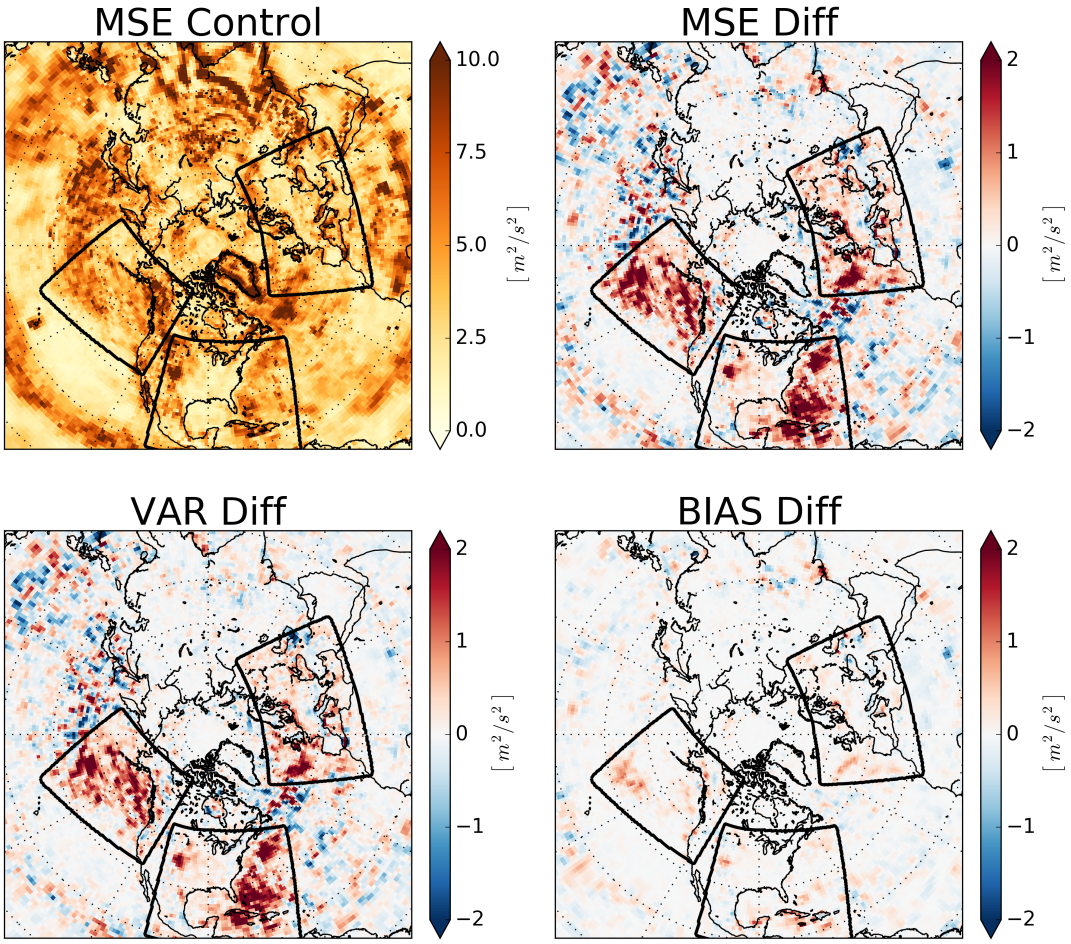
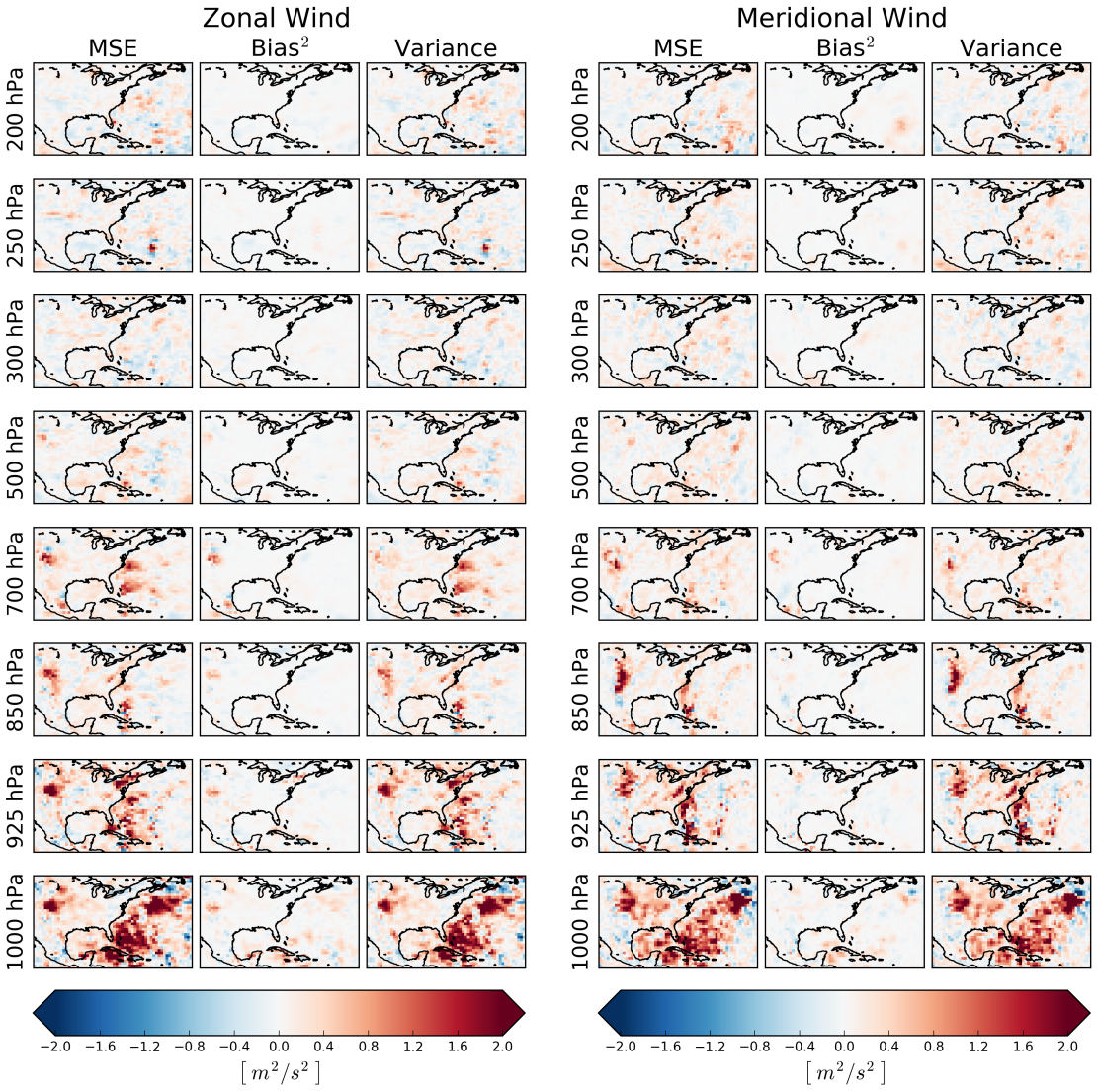


Figure B.25: Decomposition of the MSE for the 1000 hPa zonal wind field at analysis time. Shown are (top left) the MSE for the blend skip control experiment (top right) the MSE differences for the blend skip control and the 30% REG DA experiment, (bottom left) the difference between the error variance, and (bottom right) the difference between square bias. In the three difference figures, red shades indicate improvement and blue shades indicate degradation due to REG DA.



(a) (b)

Figure B.26: Vertical structure of the decomposition of the MSE difference between the blend skip control and 30% blend REG DA experiment for the two horizontal components of the wind. The different rows show results for different pressure levels, with pressure deceasing from bottom to top. Shown are (left) the MSE difference, (middle) difference in the square bias and (right) difference in the variance. Red shades indicate improvement and blue shades indicate degradation due to REG DA.

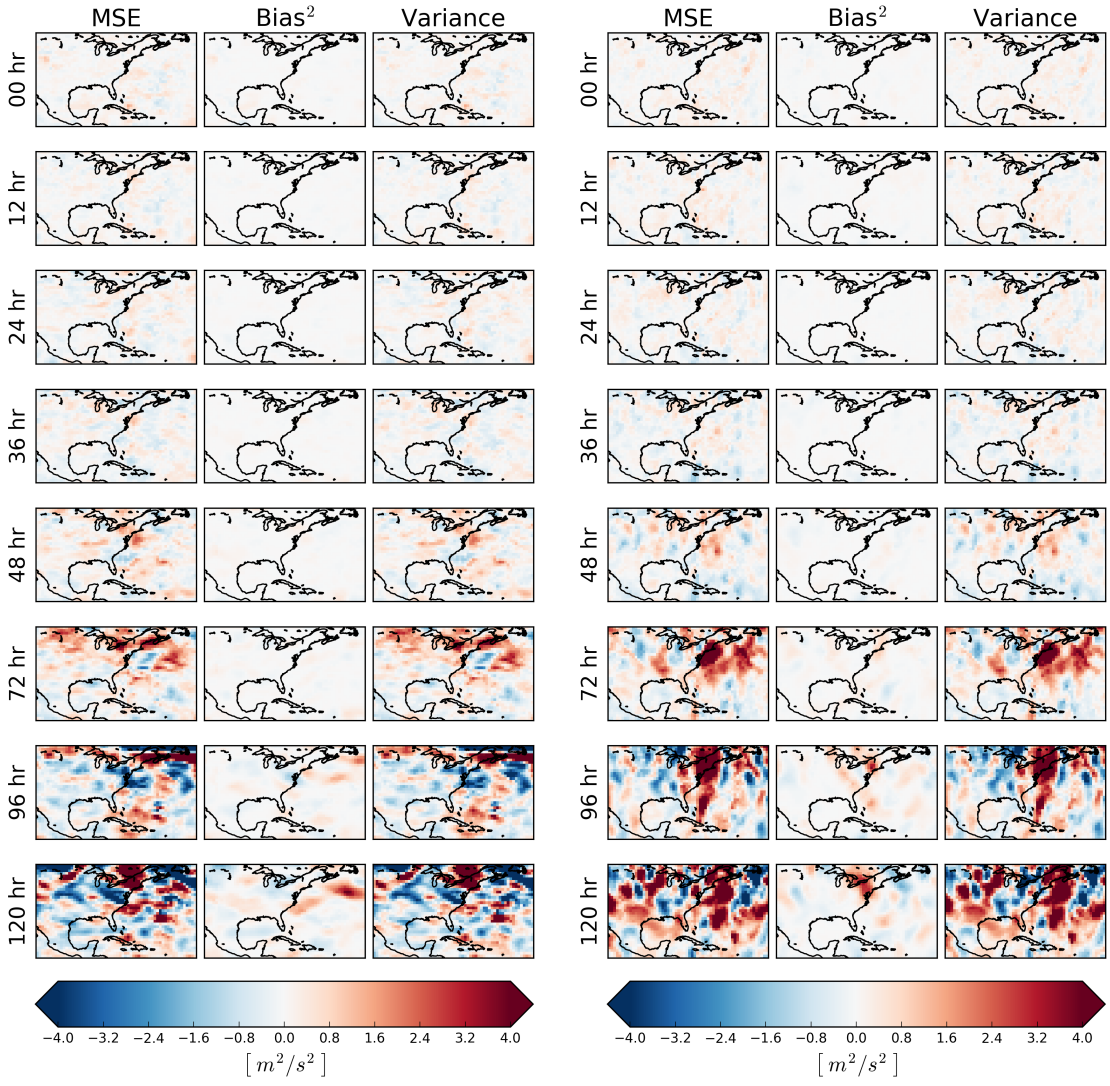


Figure B.27: Temporal evolution of the MSE difference between the blend skip control and 30% blend REG DA experiment for the two horizontal components of the wind. The different rows show results for different lead times. Shown are (left) the MSE difference, (middle) difference in the square bias and (right) difference in the variance. Red shades indicate improvement and blue shades indicate degradation due to REG DA.

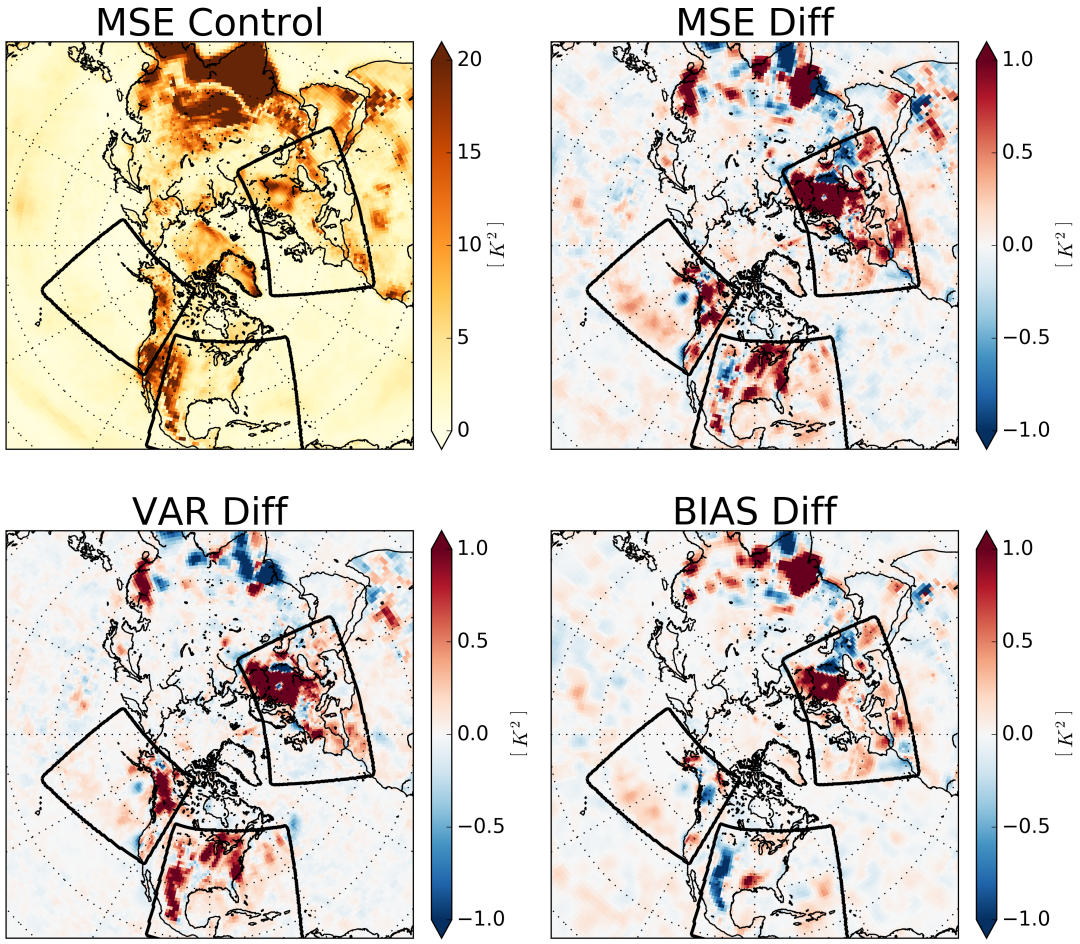


Figure B.28: Decomposition of the MSE for the 1000 hPa air temperature field at analysis time. Shown are (top left) the MSE for the blend skip control experiment (top right) the MSE differences for the blend skip control and the 30% REG DA experiment, (bottom left) the difference between the error variance, and (bottom right) the difference between square bias. In the three difference figures, red shades indicate improvement and blue shades indicate degradation due to REG DA.

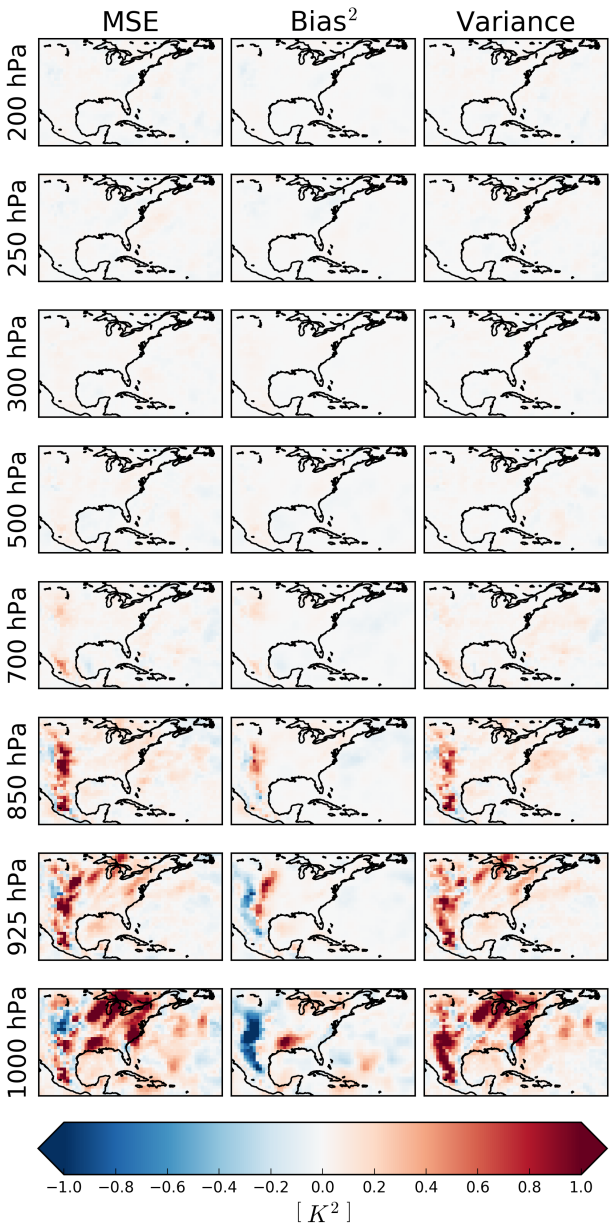


Figure B.29: Vertical structure of the decomposition of the MSE difference between the blend skip control and 30% blend REG DA experiment for air temperature. The different rows show results for different pressure levels, with pressure deceasing from bottom to top. Shown are (left) the MSE difference, (middle) difference in the square bias and (right) difference in the variance. Red shades indicate improvement and blue shades indicate degradation due to REG DA.

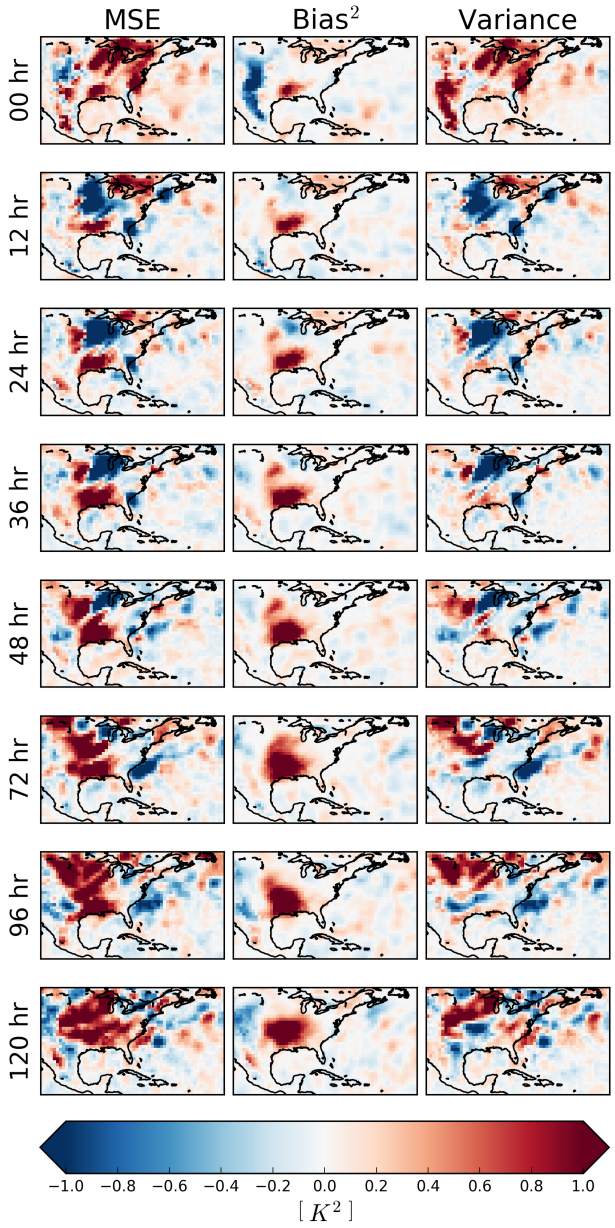


Figure B.30: Temporal evolution of the MSE difference between the blend skip control and 30% blend REG DA experiment for air temperature. The different rows show results for different lead times. Shown are (left) the MSE difference, (middle) difference in the square bias and (right) difference in the variance. Red shades indicate improvement and blue shades indicate degradation due to REG DA.

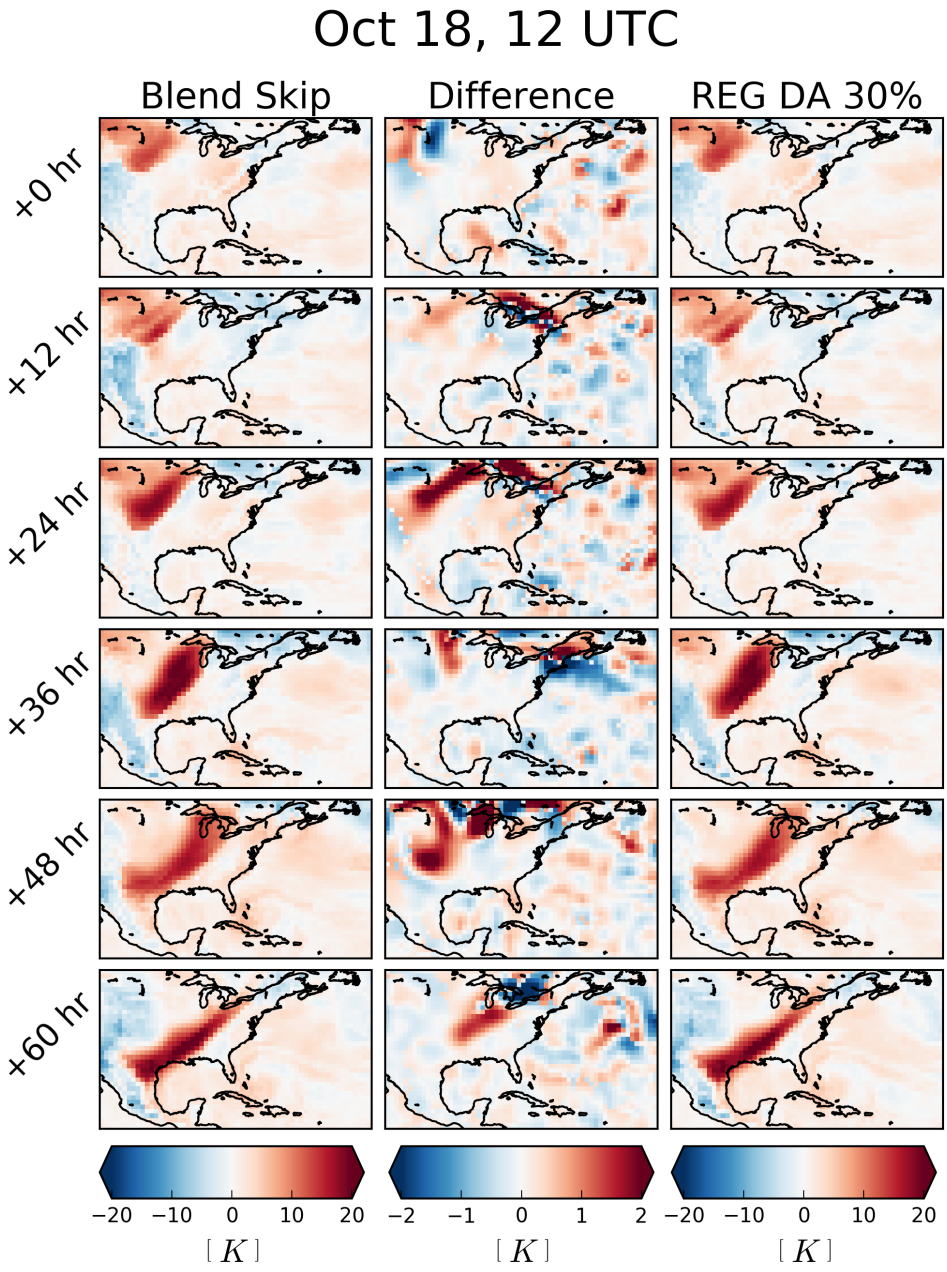


Figure B.31: A case study of the evolution of the 120 h forecast error. The different rows show results for individual forecasts initialized starting on (top) October 18th at 12 UTC and increasing in initialization time towards the bottom. Shown are forecast errors for (left) the blend skip control (right) REG DA 30% experiment and (middle) the difference in the absolute value of the error between the blend skip and REG DA 30%.

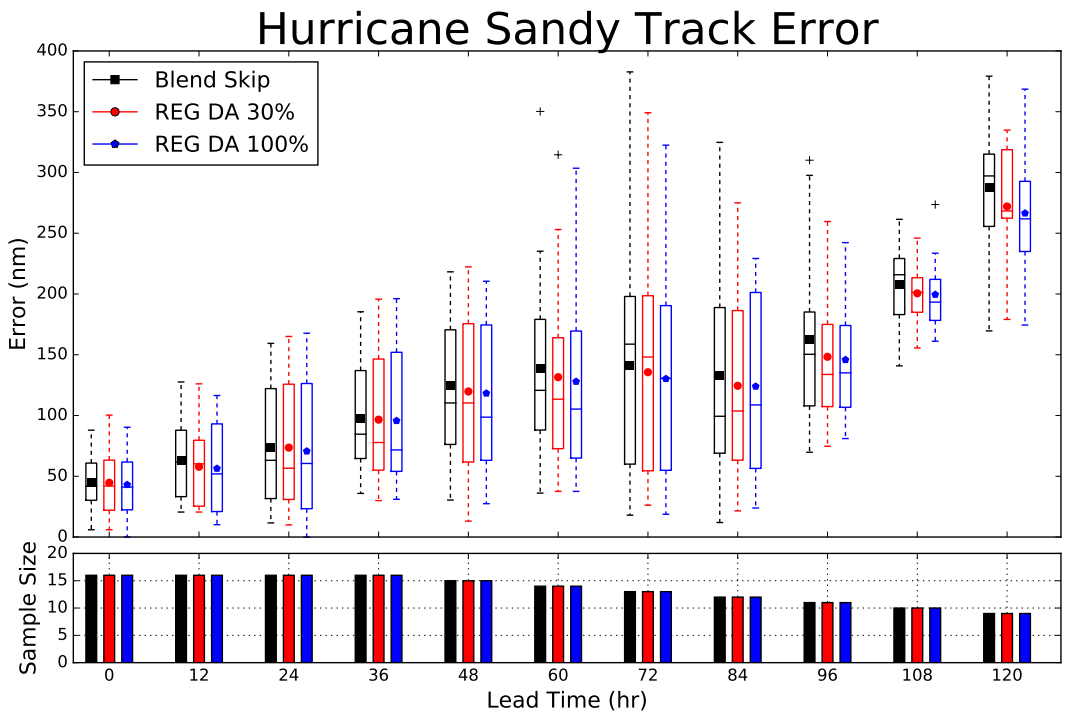


Figure B.32: Box and whisker diagram for the Hurricane Sandy track forecast errors. Each marker represent the mean of the distribution and the line is the median value. The error for the three experiments is shown as a function of lead time, with the sample size shown in the bottom panel.

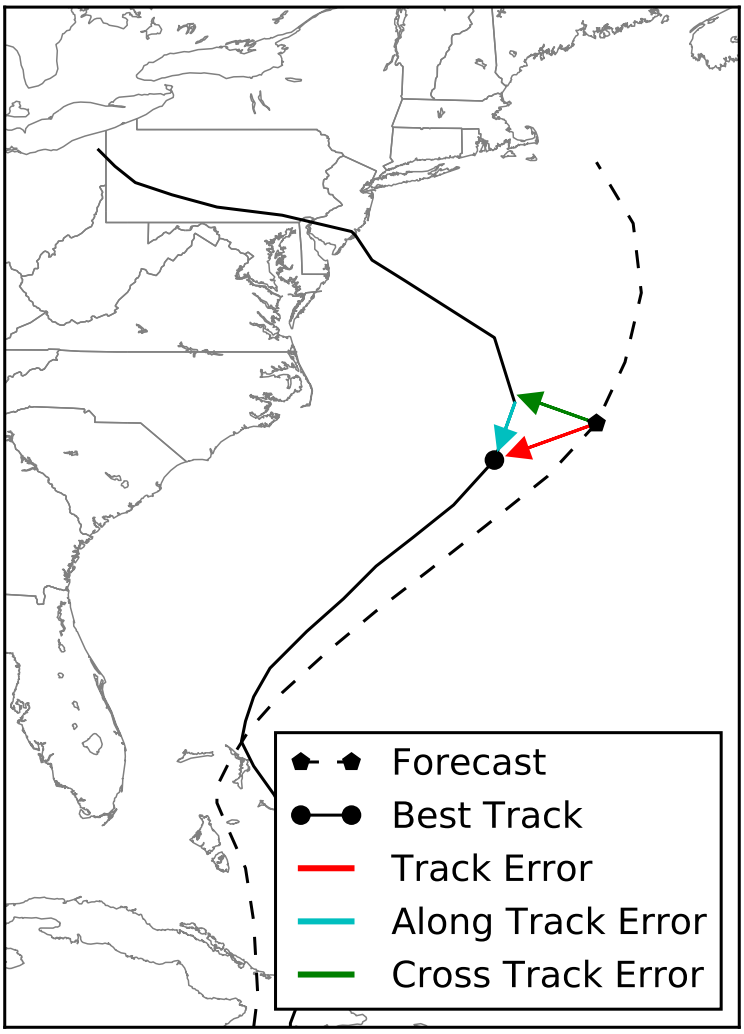


Figure B.33: Schematic illustration of the decomposition of the track error into an along track and a cross track component.

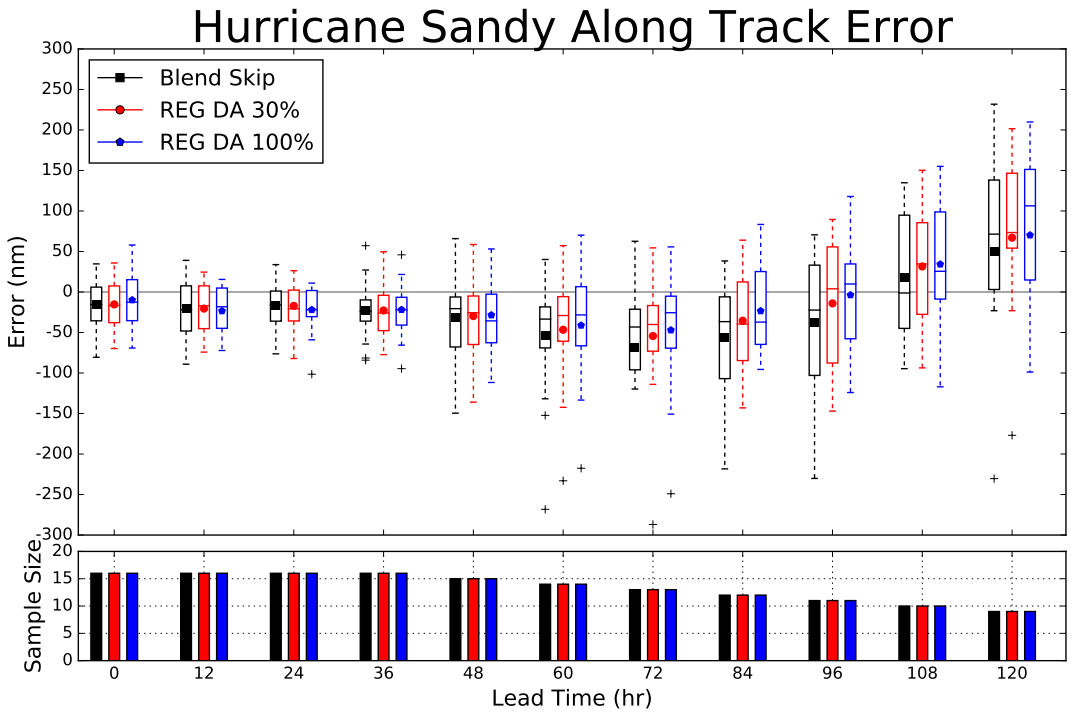


Figure B.34: Box and whisker diagram for the Hurricane Sandy *along* track forecast error component. Each marker represent the mean of the distribution and the line is the median value. The error for the three experiments is shown as a function of lead time, with the sample size shown in the bottom panel.

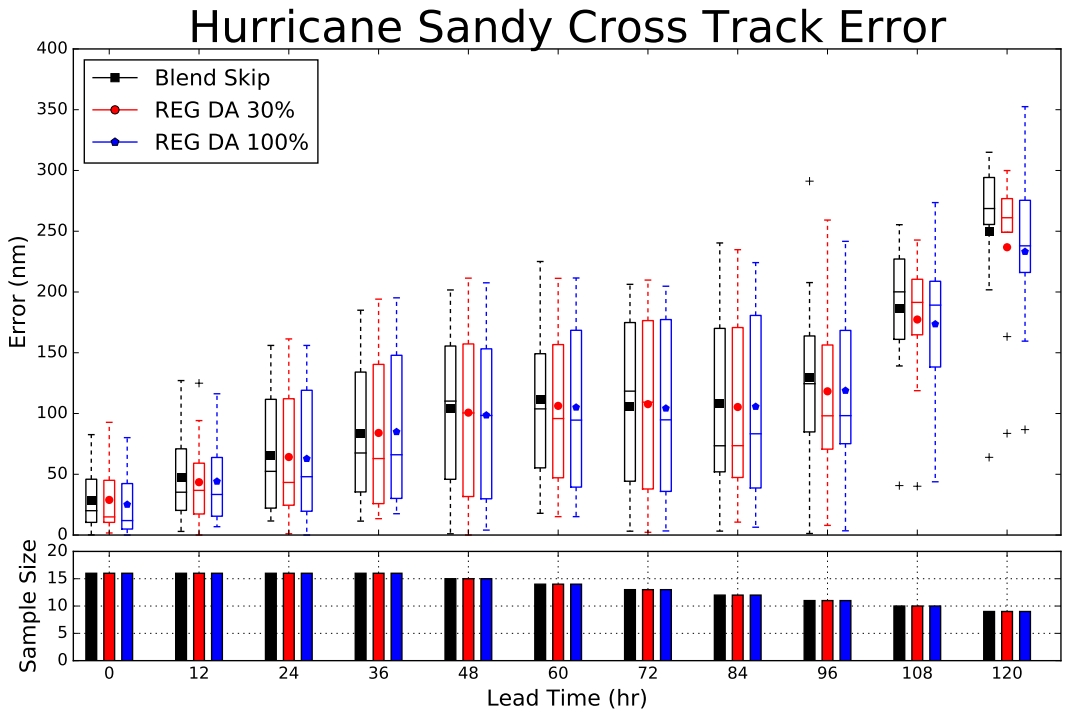


Figure B.35: Box and whisker diagram for the Hurricane Sandy *cross* track forecast error component. Each marker represent the mean of the distribution and the line is the median value. The error for the three experiments is shown as a function of lead time, with the sample size shown in the bottom panel.

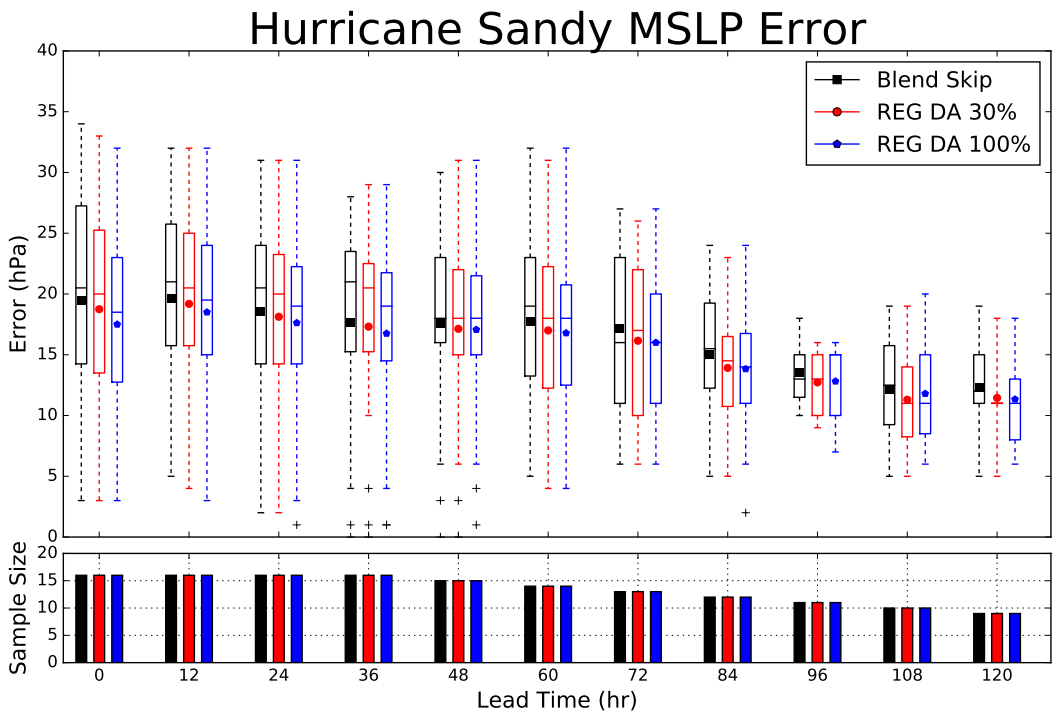


Figure B.36: Box and whisker diagram for the Hurricane Sandy minimum mean sea level pressure forecast errors. Each marker represent the mean of the distribution and the line is the median value. The error for the three experiments is shown as a function of lead time, with the sample size shown in the bottom panel.

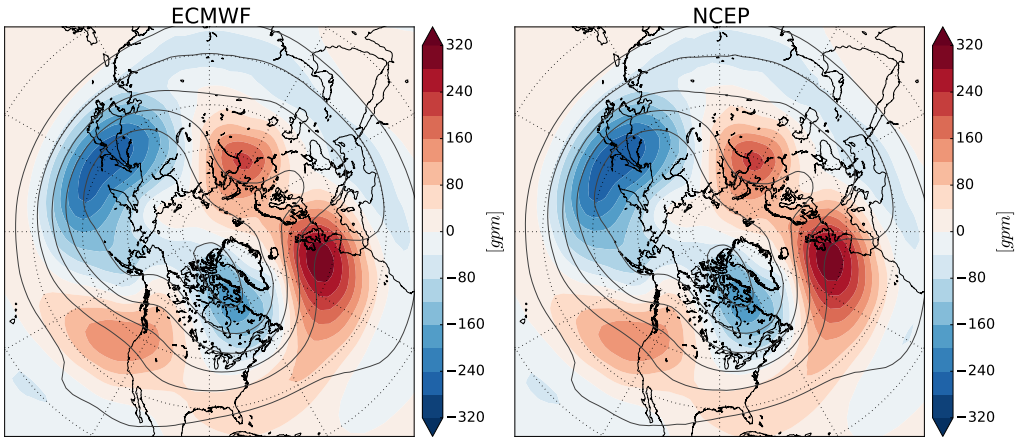


Figure B.37: Zonal anomalies of the time-mean flow. Color shades indicate the zonal anomalies for the investigated time period based on the ECMWF (left) analyses and NCEP (right) analyses, while contours show the time-mean flow (geopotential height) at the 500 hPa level based on the same analyses.

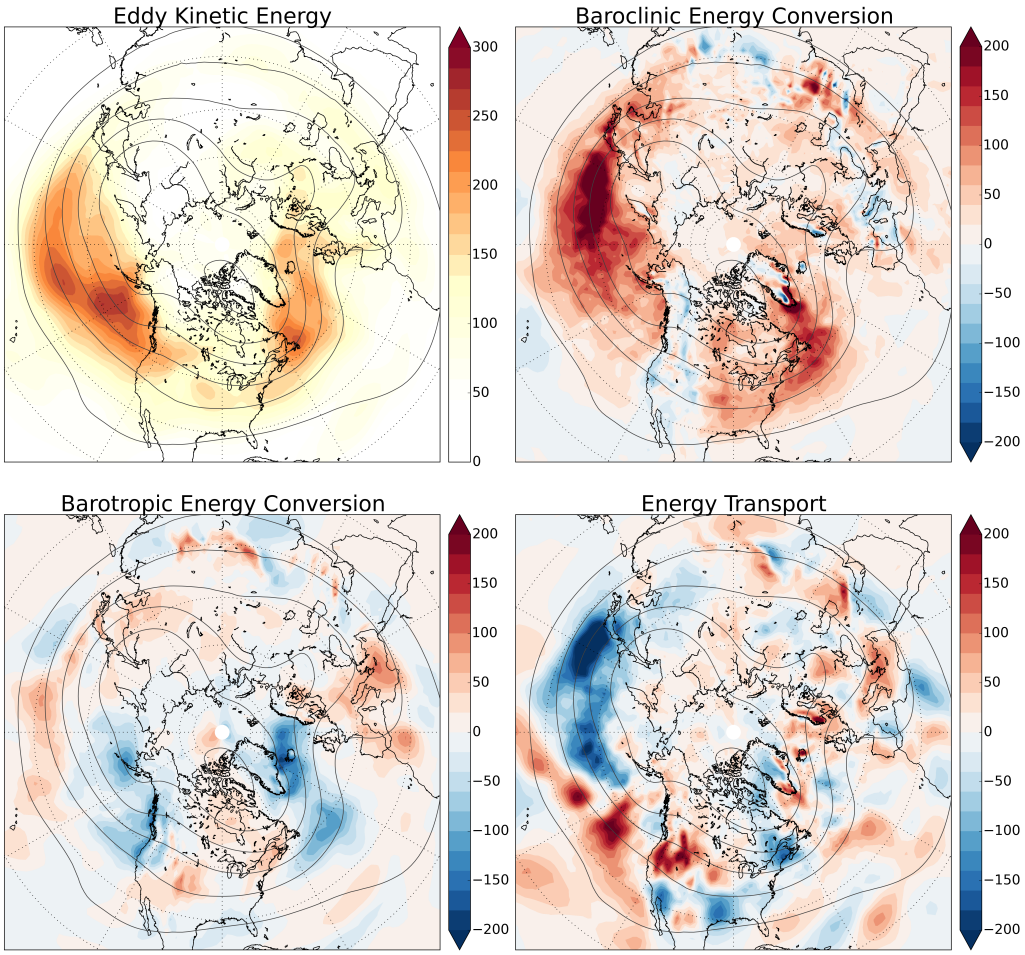


Figure B.38: The time-mean of the eddy kinetic energy conversion processes for the investigated time period. Color shades show the time mean of the eddy kinetic energy [J] (top left), baroclinic energy conversion [J/day] (top right), barotropic energy conversion [J/day] (bottom left), and horizontal transport of the eddy kinetic energy [J/day] (bottom right).

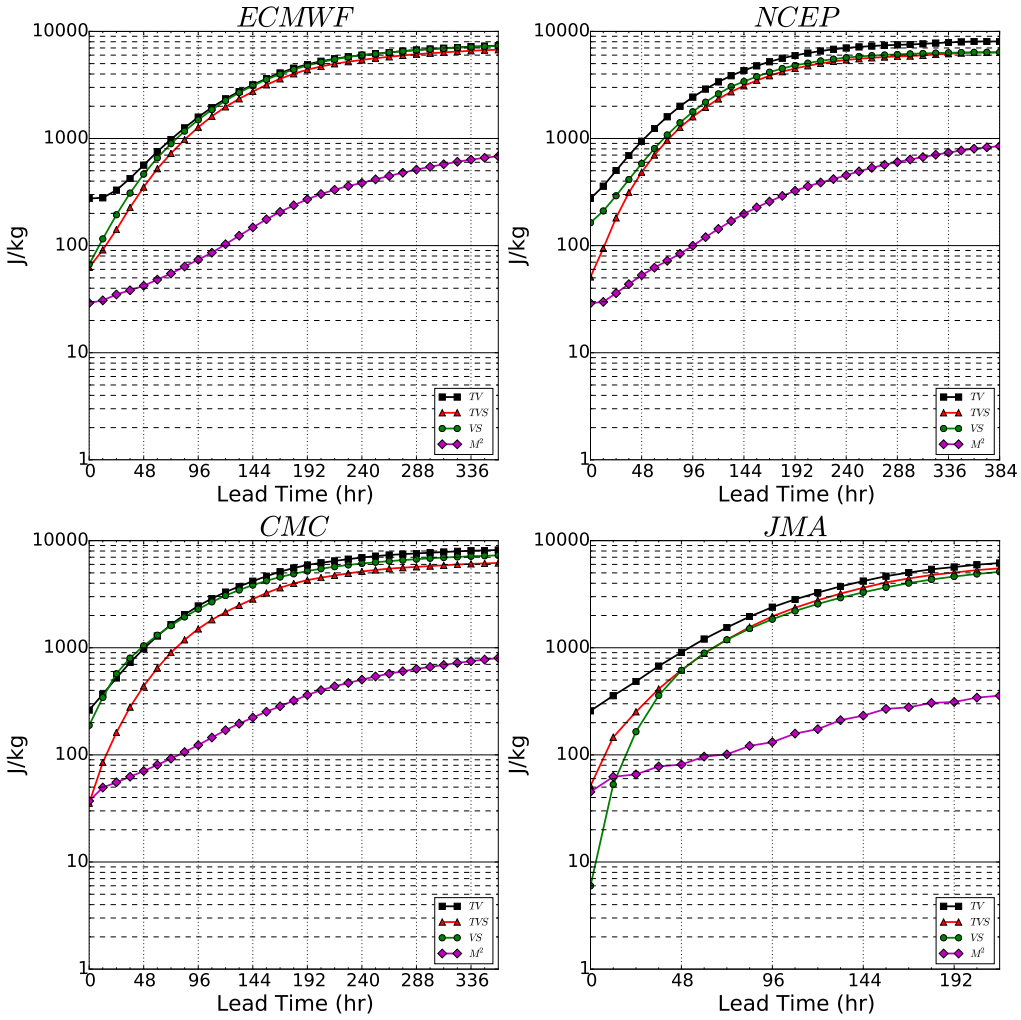


Figure B.39: The evolution of the diagnostics, VS, TV, TVS, and M^2 in the forecasts for four of the ensembles, averaged over the NH extra-tropics and all forecasts started between January 1, 2012 and February 29, 2012. It should be noted that the models have different max forecast lead times.

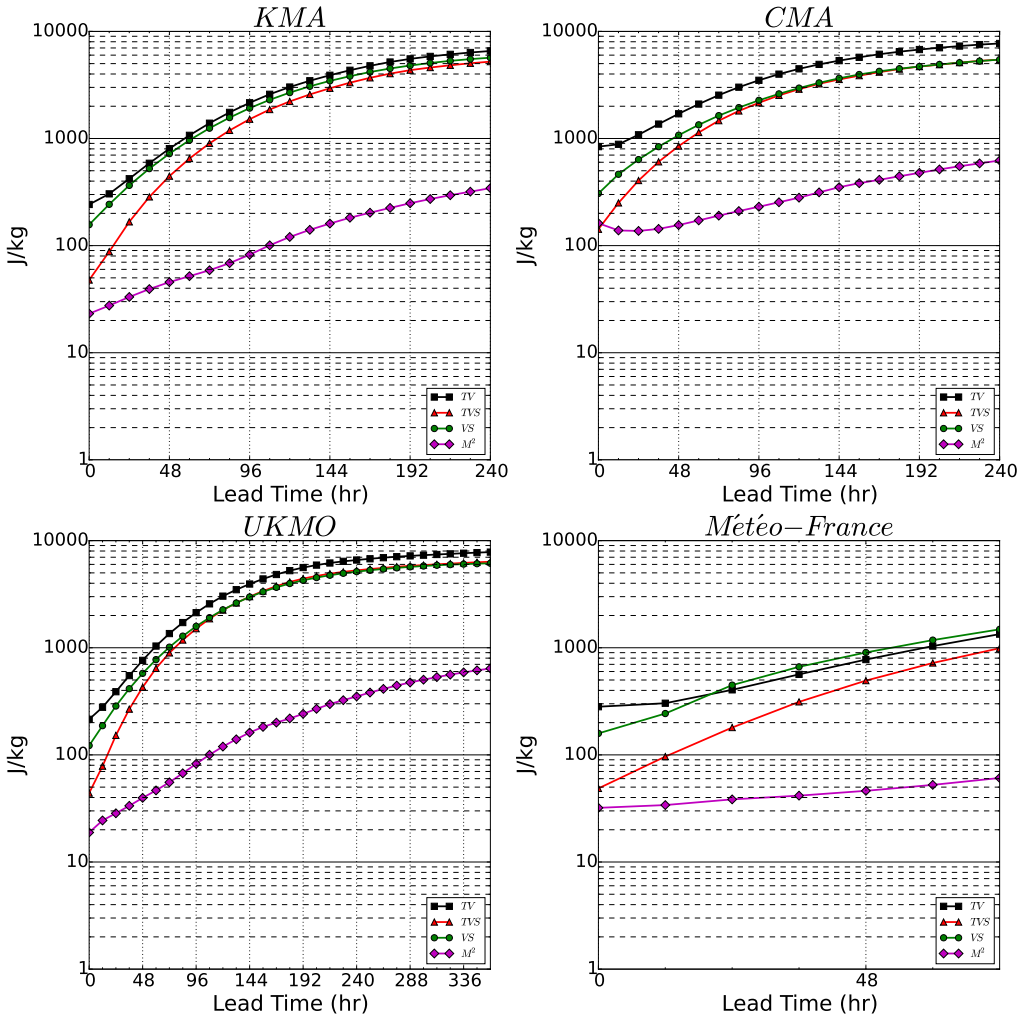


Figure B.40: The evolution of the diagnostics VS, TV, TVS and M^2 in the forecasts for the remaining four ensembles, averaged over the northern hemisphere extra-tropics and all forecasts started between between January 1, 2012 and February 29, 2012. It should be noted that the models have different max forecast lead times.

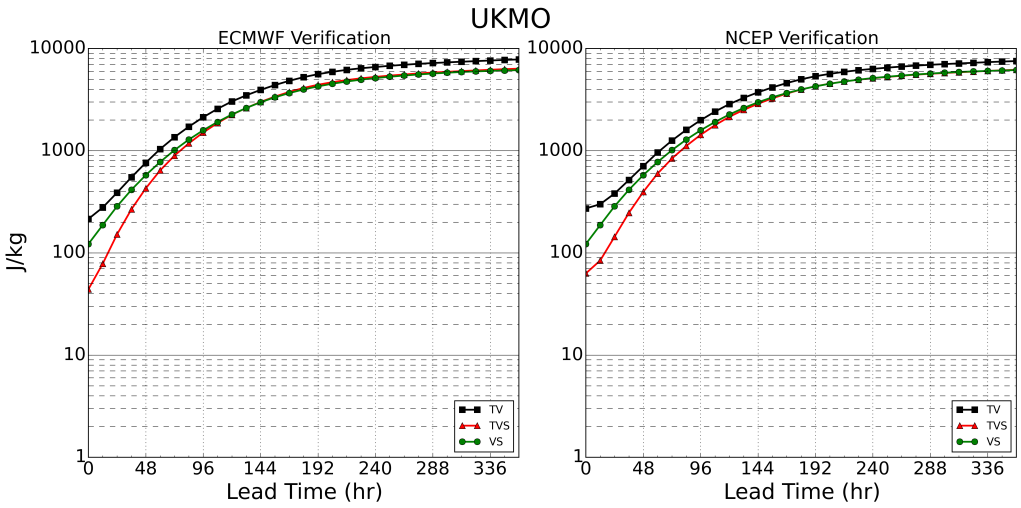


Figure B.41: Illustration of the robustness of the results of Figs. B.39 and B.40 to the choice of the proxy for the true state. The diagnostics are shown for the UKMO ensemble for the cases, in which the proxy for the true state is defined by (left) the ECMWF analyses and (right) the NCEP analyses.

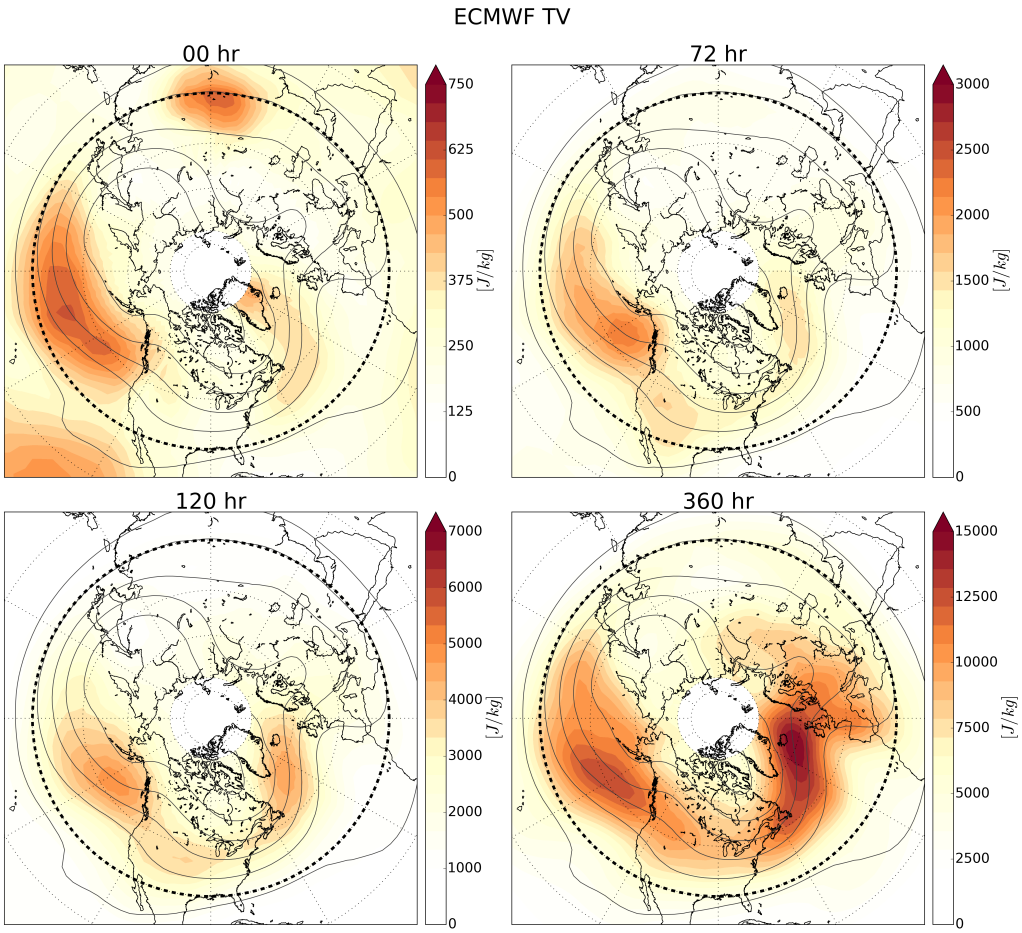


Figure B.42: Spatial distribution of the average forecast uncertainty for the ECMWF ensemble. Shown are TV (color shades) and the time mean of the geopotential analyses at 500 hPa (contours) at analysis and three different forecast times. Dashes indicate the southern boundary of the region used for the computation of the spatial averages of Figs. B.39 and B.40.

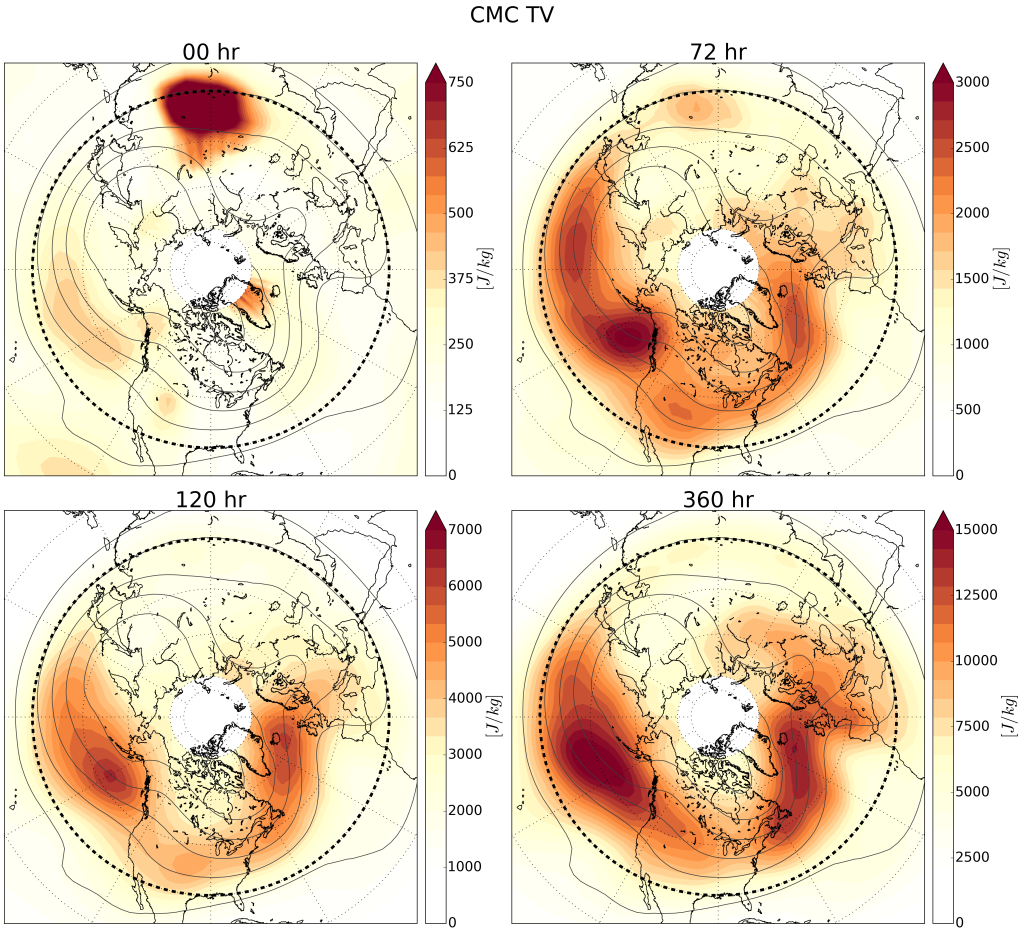


Figure B.43: Same as Fig. B.42, except for the CMC ensemble

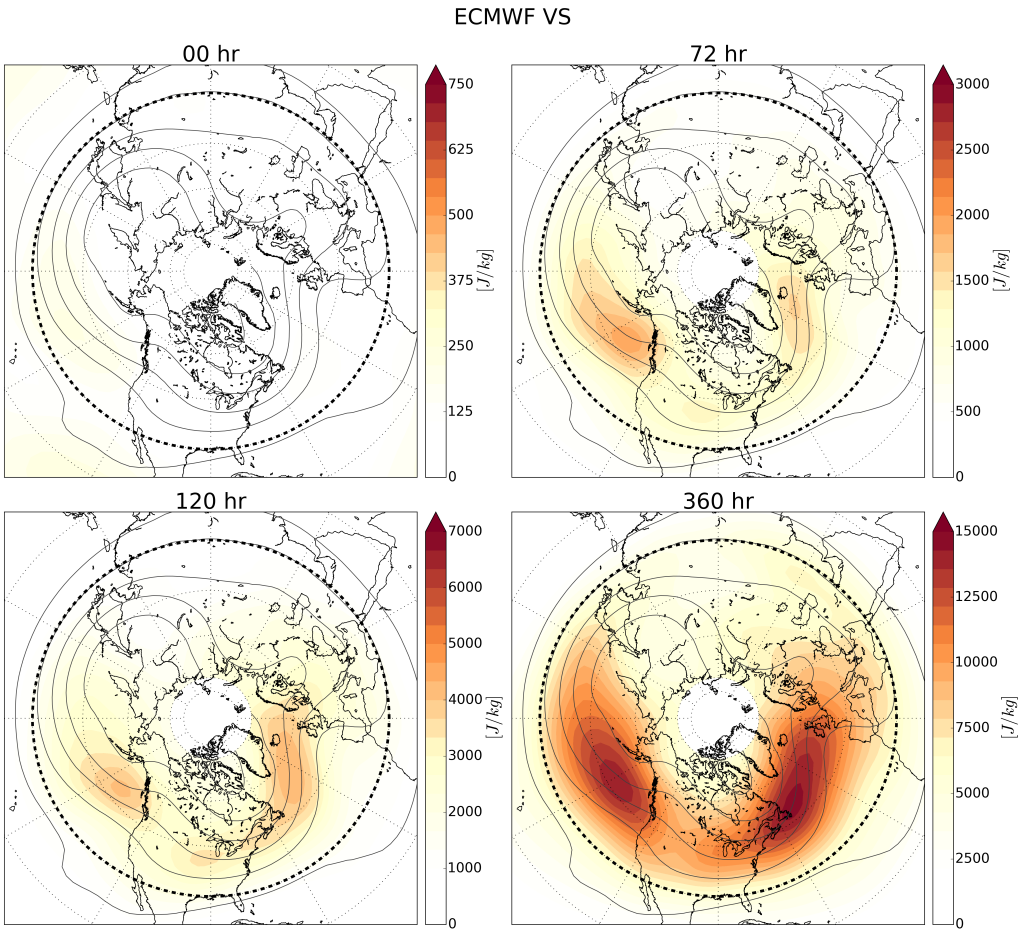


Figure B.44: Spatial distribution of the average ensemble spread for the ECMWF ensemble. Shown are VS (color shades) and the time mean of the geopotential analyses at 500 hPa (contours) at analysis and three different forecast times. Dashes indicate the southern boundary of the region used for the computation of the spatial averages of Figs. B.39 and B.40.

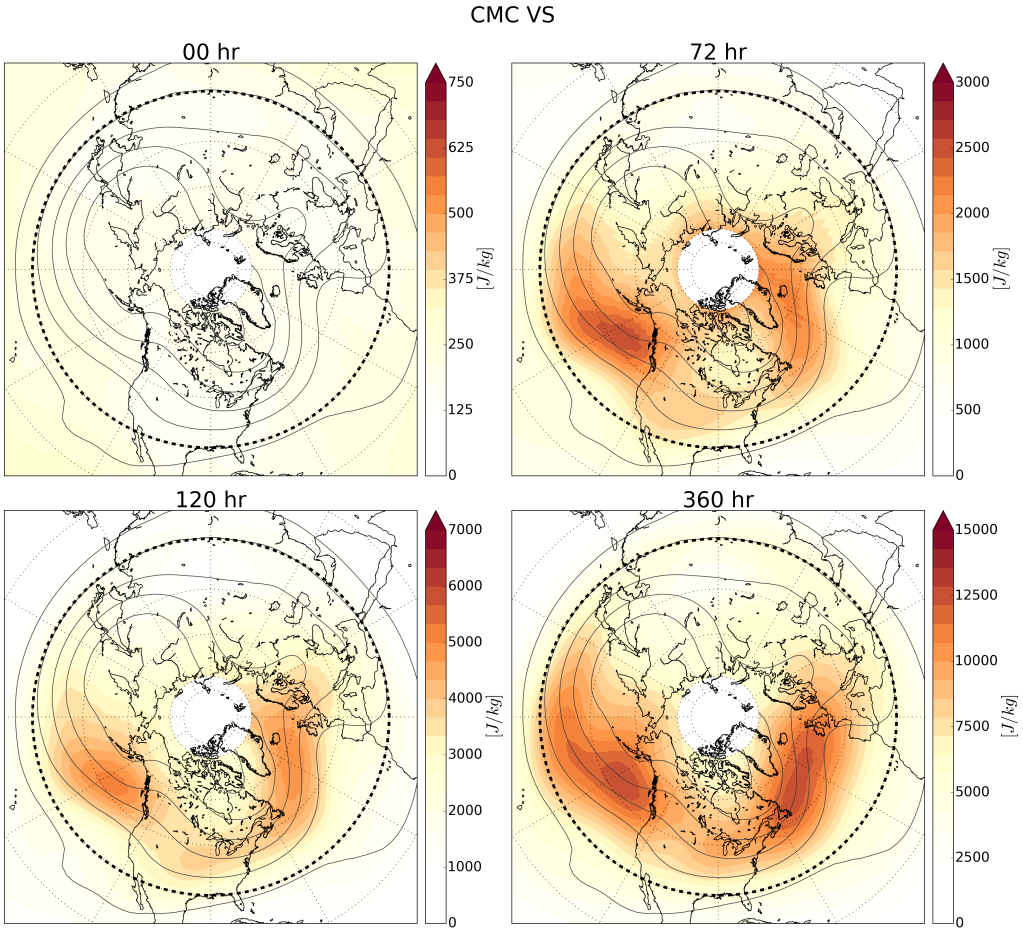


Figure B.45: Same as Fig. B.42, except for the CMC ensemble

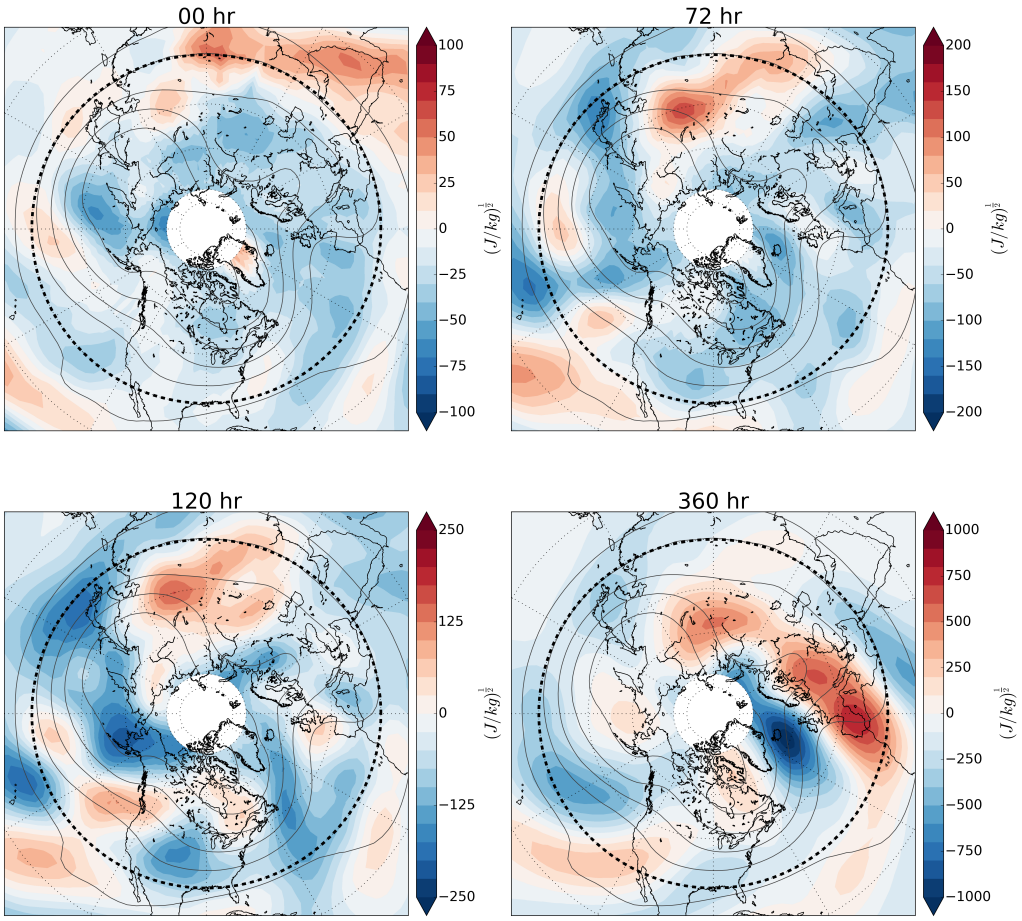


Figure B.46: Spatial distribution of the mean uncertainty for the ECMWF ensemble. Shown are M (color shades) and the time mean of the geopotential analyses at 500 hPa (contours) at analysis and three different forecast times. Dashes indicate the southern boundary of the region used for the computation of the spatial averages of Fig. B.39 and B.40.

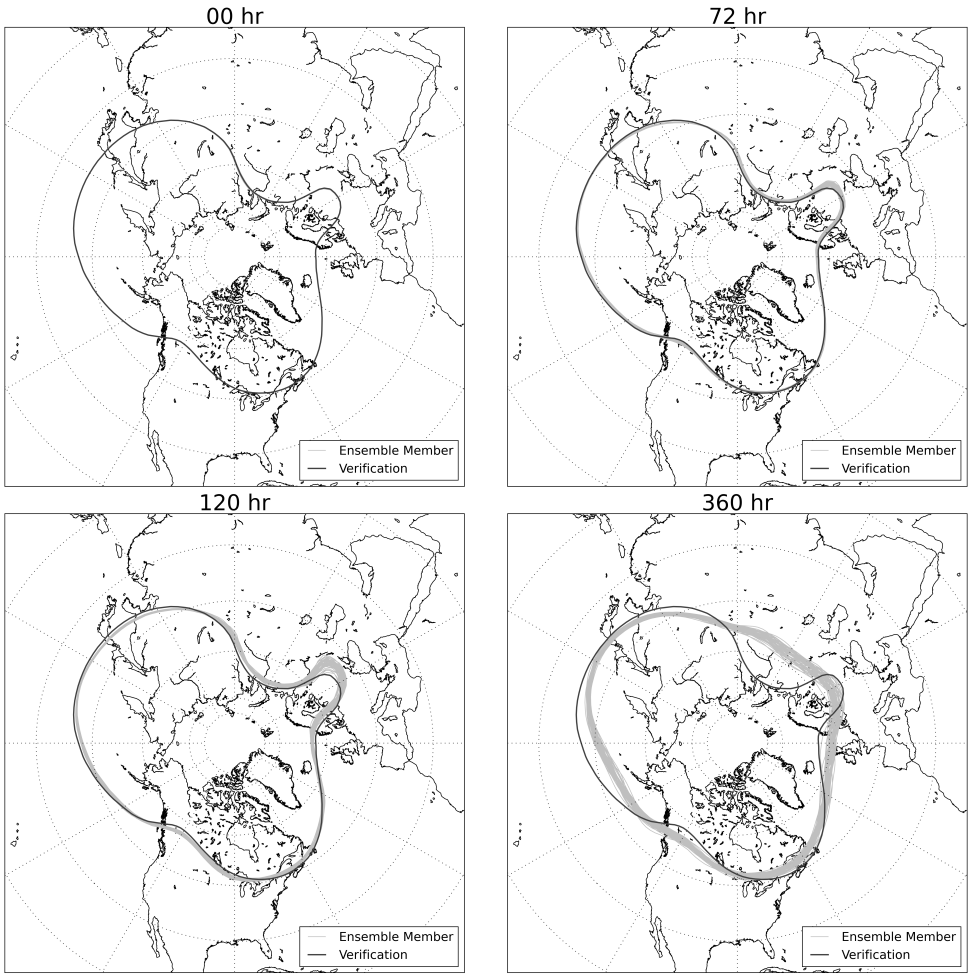


Figure B.47: Spaghetti diagram for the ensemble of mean forecasts. The mean forecasts were obtained by averaging each member of the ECMWF ensemble over the investigated time period. Shown by grey contour lines are the 5350 gpm isohypses for the ensemble members. The black contour line shows the time-mean of the ECMWF analyses for the investigated time period.

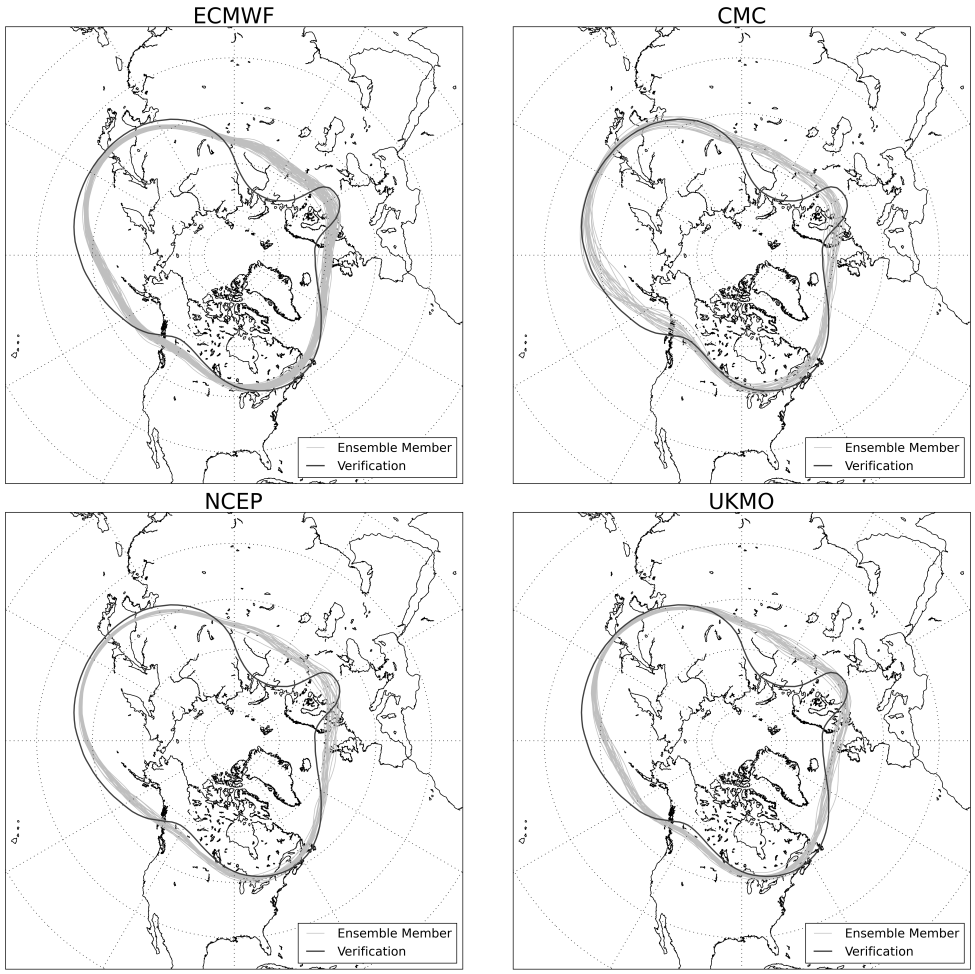


Figure B.48: Spaghetti diagrams for select ensemble systems at 360 hr lead time. The mean forecasts were obtained by averaging each member of the ensemble over the investigated time period. Shown by grey contour lines are the 5350 gpm isohypses for the ensemble members. The black contour line shows the time-mean of the ECMWF analyses for the investigated time period.

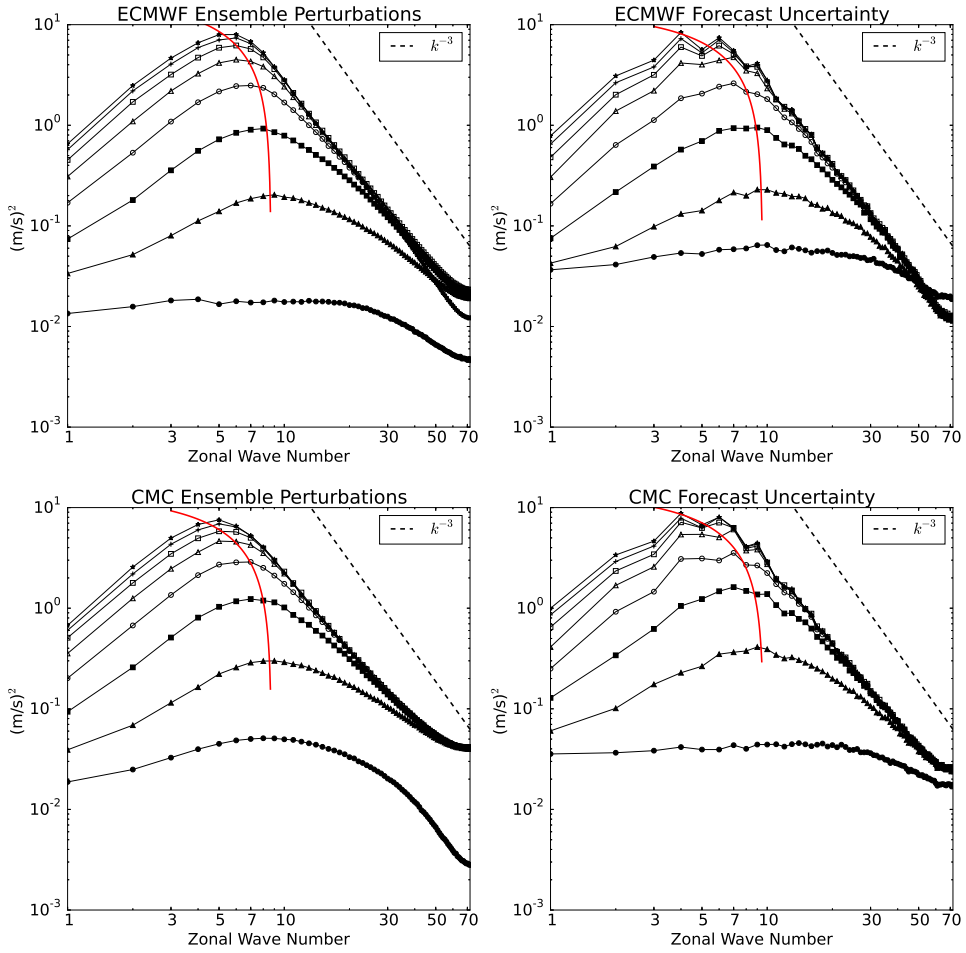


Figure B.49: Spectral evolution of the forecast uncertainty and the ensemble spread for the meridional component of the wind vector at 500 hPa for the ECMWF and CMC ensembles. Shown are the meridional averages of the zonal power spectra of the meridional wind associated with (left) the ensemble spread and (right) the forecast uncertainty. The lowest curves show the spectra at analysis time, while the other curves show the spectra with 2-day increments of the forecast time (the top curves are for day 14 forecast time). The red curve shows the linear regression of the maximum power and the associated zonal wave number for all of the forecast times.

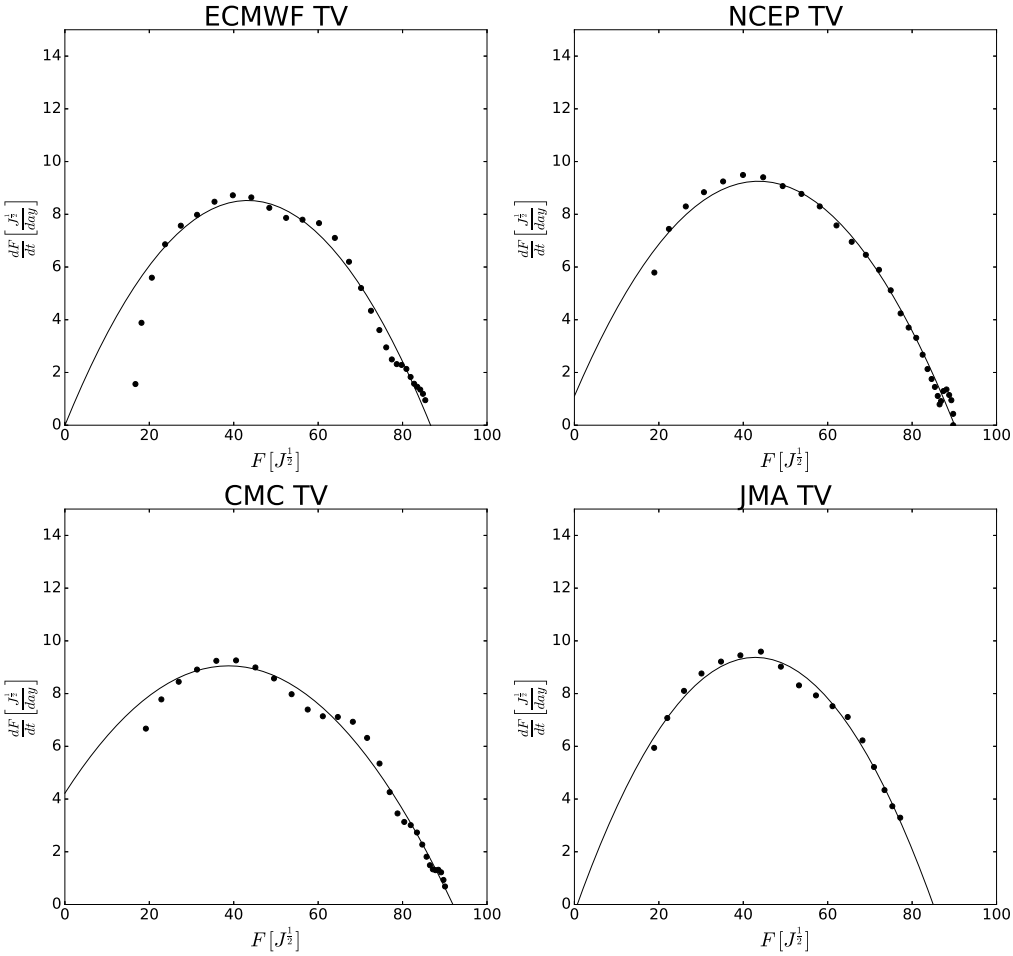


Figure B.50: Lorenz curves for the forecast uncertainty ($F = \sqrt{TV}$) for the ECMWF, NCEP, CMC, and JMA ensembles.

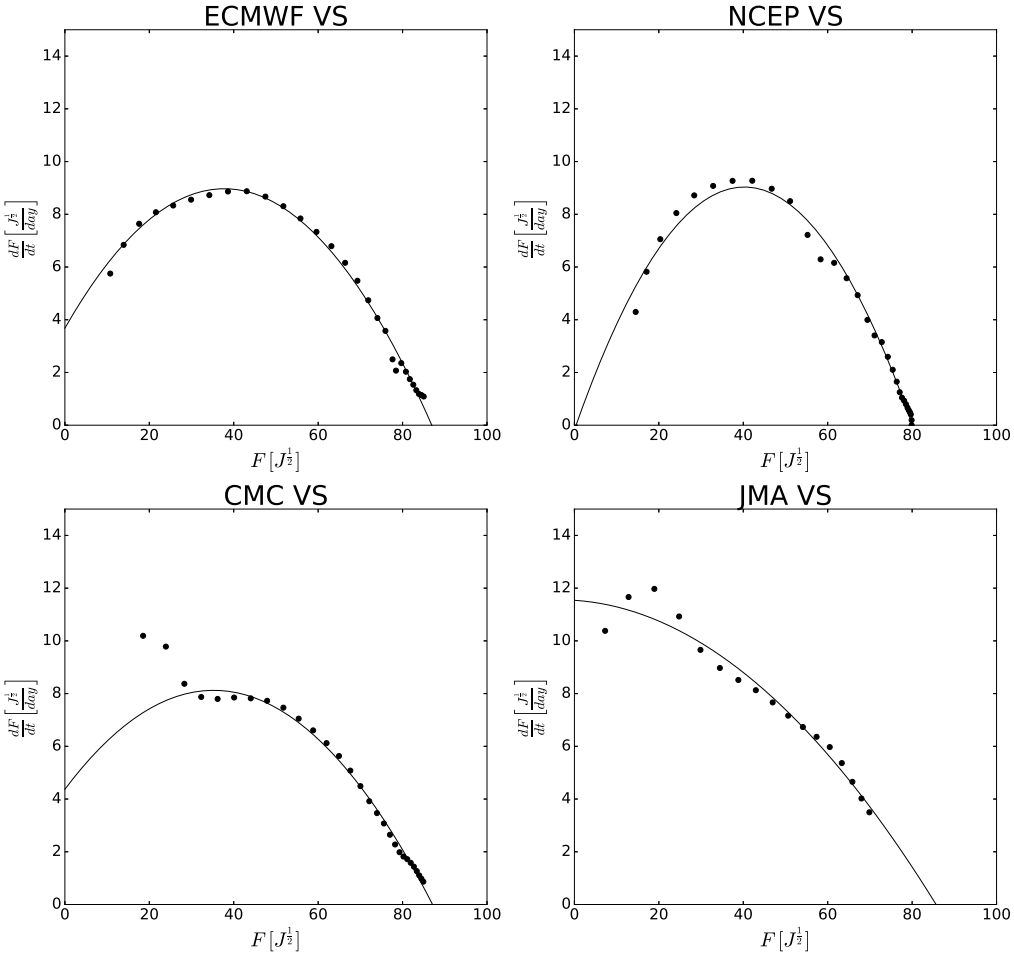


Figure B.51: Lorenz curves for the ensemble spread ($F = \sqrt{VS}$) for the ECMWF, NCEP, CMC, and JMA ensembles.

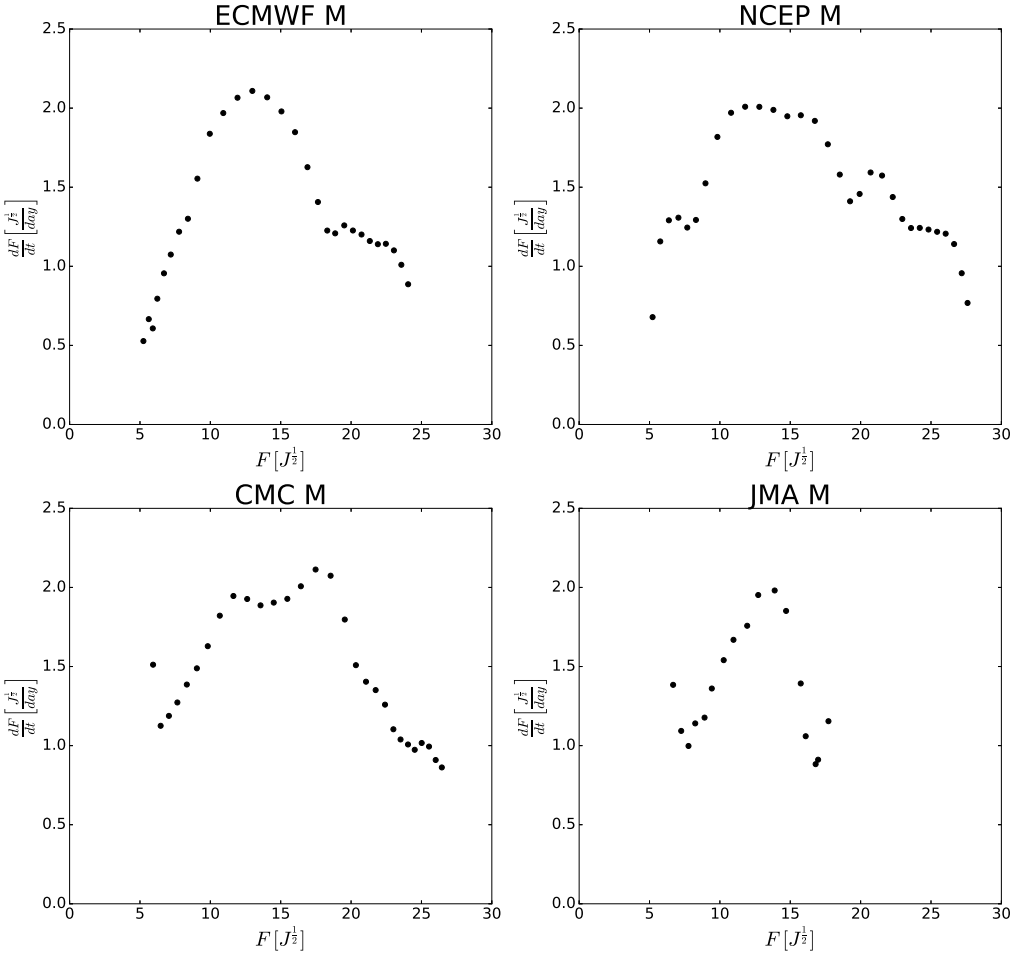


Figure B.52: Estimates of $d\left|M\right|/dt$ as function of $\left|M\right|$ ($F = \left|M\right|$) for the ECMWF, NCEP, CMC, and JMA ensembles.

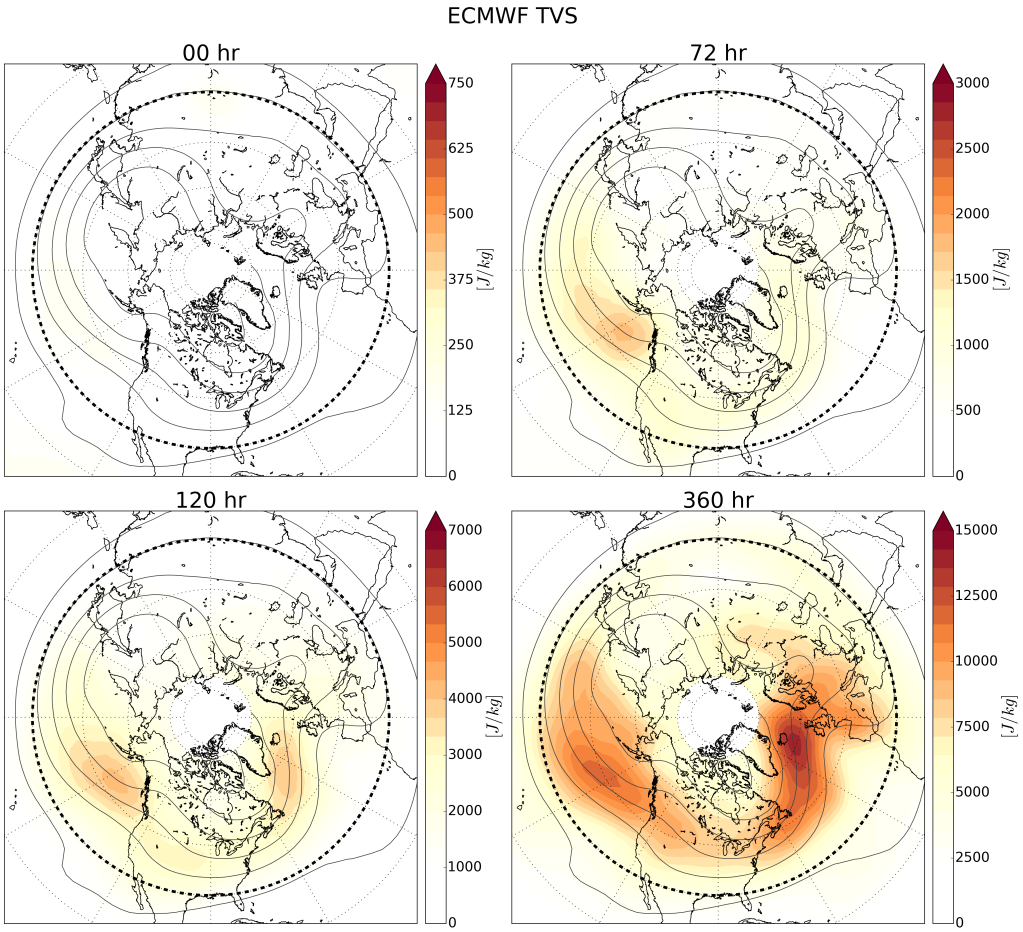


Figure B.53: Spatiotemporal evolution of TVS (color shades) in the ECMWF ensemble forecasts. Also shown is the time mean of the geopotential analyses at 500 hPa (contours). Dashes indicate the southern boundary of the region used in the computation of the spatial averages shown in Figs. B.39.

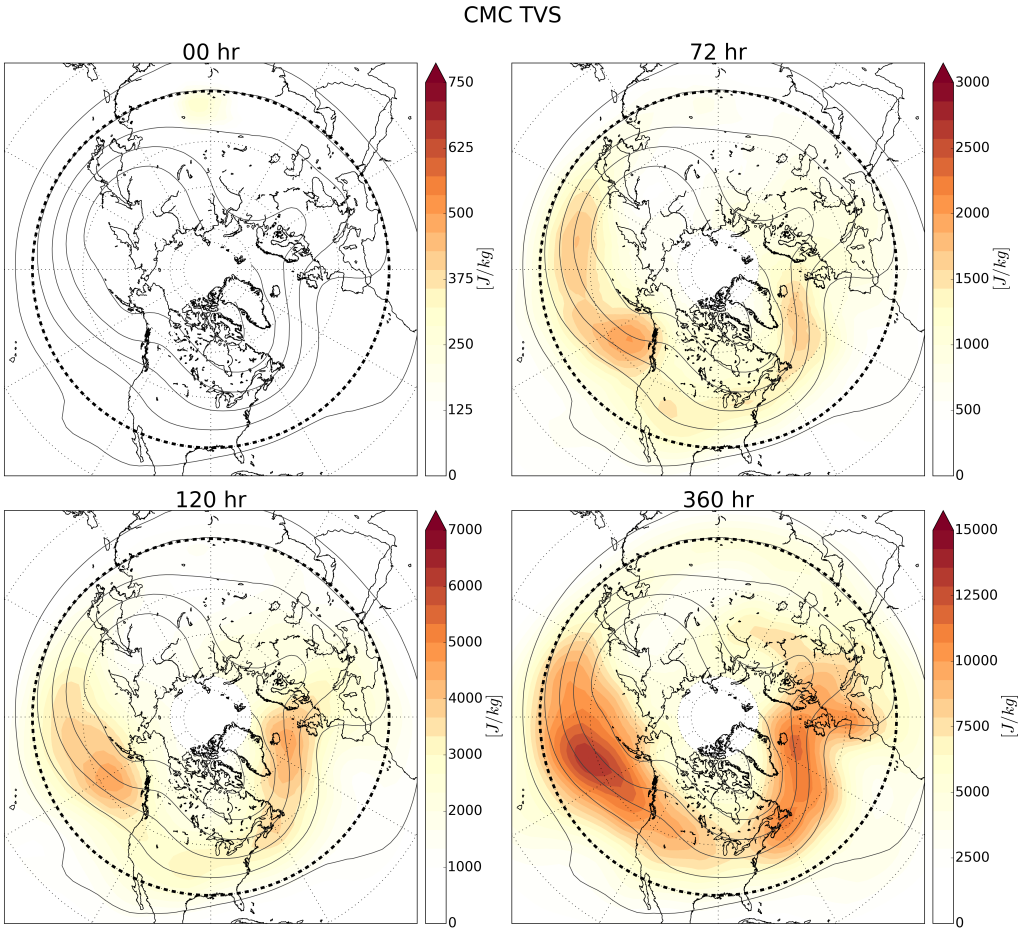


Figure B.54: The same as Fig. B.53, except for the CMC ensemble.

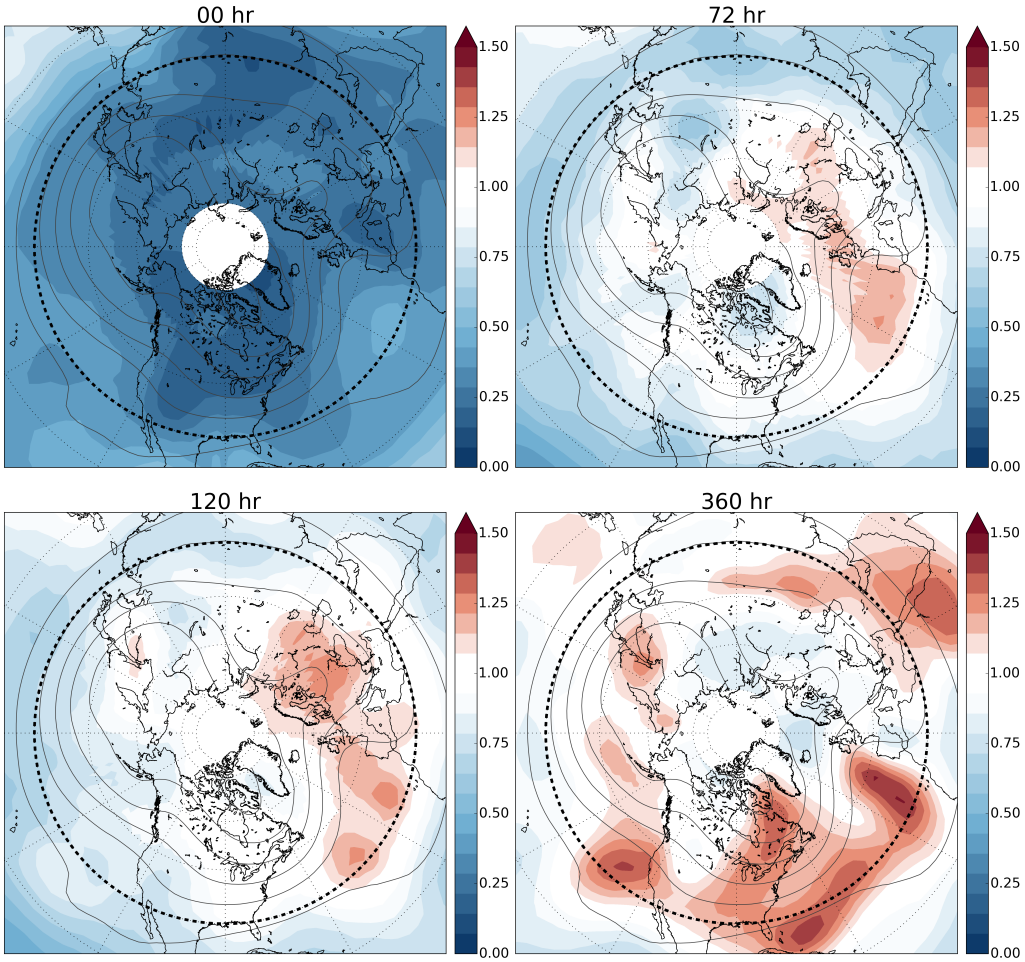


Figure B.55: The spatiotemporal evolution of the local ratio between VS and TV (color shades) in the ECMWF ensemble forecasts. Also shown is the time mean of the geopotential analyses at 500 hPa (contours). Dashes indicate the southern boundary of the region used in the computation of the spatial averages shown in Figs. B.39.