NUMERICAL WEATHER PREDICTION ERROR OVER THE NORTH PACIFIC AND WESTERN NORTH AMERICA: AN INVESTIGATION WITH SHORT-RANGE ENSEMBLE TECHNIQUES

JOSHUA P. HACKER

B.S. The University of California Los Angeles, 1994
M.Sc. The University of British Columbia, 1997

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES
(Atmospheric Sciences Programme, Department of Earth and Ocean Sciences)

We accept this thesis as conforming to the required standard
In presenting this thesis in partial fulfillment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Earth and Ocean Sciences
The University of British Columbia
Vancouver, Canada

Date 6 Nov. 2001
Abstract

Numerical weather prediction models, which are discretized approximations to the physical equations of the atmosphere, are a critical part of weather forecasting. They can project an observed state of the atmosphere into the future, but forecasts have many error sources. To counter this, several forecasts made from slightly different initial states or models can be made over the same domain and time period. This approach, called ensemble forecasting, can forecast uncertainty, and produce a better average forecast.

In this study, ensemble forecasts are generated and analyzed to understand the nature of initial condition (IC) and model error over the North Pacific. A poorly-predicted storm event (a bust) in Feb. 1999 is a useful case study period. To approximate different aspects of IC uncertainty, three IC-perturbation methods are used: (1) ranked perturbations that target coherent structures in the analyses; (2) perturbations that simulate differences between operational analyses from major forecast centers; and (3) random perturbations. An ensemble of different model configurations approximates model uncertainty.

Ensembles are verified several ways to separate model and IC uncertainty, and evaluate ensemble performance. It is found that during the period surrounding the bust, IC error is greater than model error. But for one critical forecast, model error is unusually high while error from ICs is unusually small.

Comparison with rawinsonde observations shows that differences between operational analyses cannot account for analysis error. Ensembles generated with this information show that analysis differences contain some spatial information about analysis uncertainty, but the magnitude is too small. To account for total forecast error, model uncertainty must be included with IC uncertainty.

Comparing different ensembles reveals that a scaled ranked-perturbation ensembles has the best characteristics, including perturbation magnitude, uncertainty growth vs. error growth, spread-error correlation, and shape of the variance spectrum. Its properties are verified by running an independent case study, and only minor differences are found.

Contributions to the field of weather prediction include a new ranked-perturbation method that results in improved short-range ensemble-mean forecasts, an understanding of the relationship between analysis error and analysis differences, and the confirmation that model uncertainty is necessary to account for short-range forecast error.
# Contents

Abstract ii  
List of Figures vi  
List of Tables xii  

1 Introduction and literature review 1  
   1.1 The role and effectiveness of numerical weather prediction 1  
   1.2 Chaotic systems 3  
   1.3 Probabilistic forecasts 5  
   1.4 IC ensemble methods 6  
      1.4.1 Monte Carlo methods 7  
      1.4.2 Perturbation of sensitive modes 9  
      1.4.3 Relation to data assimilation 12  
   1.5 Model error 13  
   1.6 Goals of this research 15  

2 Experiment Environment 16  
   2.1 Statistical definitions 16  
   2.2 Model and domain 17  
   2.3 The W transform 21  
   2.4 Generation of IC ensembles using perturbations in W space 24  
      2.4.1 Ranked perturbations 24  
      2.4.2 Perturbations correlated with analysis differences 27  
   2.5 Generation of model-error ensembles using physics parameterization options 28
## 3 Description of a case study

<table>
<thead>
<tr>
<th>3.1 Introduction</th>
<th>29</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.2 Storm description</td>
<td>33</td>
</tr>
</tbody>
</table>

## 4 NWP Error and Uncertainty Around a North Pacific Bust

<table>
<thead>
<tr>
<th>4.1 Experiment design</th>
<th>43</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.2 IC and model error with ranked perturbations</td>
<td>45</td>
</tr>
<tr>
<td>4.2.1 Spread-error correlation</td>
<td>45</td>
</tr>
<tr>
<td>4.2.2 Model and IC error</td>
<td>47</td>
</tr>
<tr>
<td>4.2.3 Qualitative aspects</td>
<td>50</td>
</tr>
<tr>
<td>4.3 Sympathetic data denial</td>
<td>51</td>
</tr>
<tr>
<td>4.4 Comparison with perfect ensemble</td>
<td>53</td>
</tr>
</tbody>
</table>

## 5 Perturbations correlated with analysis differences

<table>
<thead>
<tr>
<th>5.1 Experiment and analyses</th>
<th>58</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1.1 Ensemble generation</td>
<td>60</td>
</tr>
<tr>
<td>5.1.2 MRF and GEM analyses</td>
<td>61</td>
</tr>
<tr>
<td>5.2 Perfect ensemble verification</td>
<td>66</td>
</tr>
<tr>
<td>5.3 Verification against analyses</td>
<td>67</td>
</tr>
<tr>
<td>5.4 Conclusions</td>
<td>72</td>
</tr>
</tbody>
</table>

## 6 Comparison of ensemble methods

<table>
<thead>
<tr>
<th>6.1 Perfect ensemble comparison</th>
<th>74</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1.1 Error and spread</td>
<td>75</td>
</tr>
<tr>
<td>6.1.2 Spread-error correlation</td>
<td>77</td>
</tr>
<tr>
<td>6.2 Verification against analyses</td>
<td>78</td>
</tr>
<tr>
<td>6.2.1 Ensemble mean errors</td>
<td>78</td>
</tr>
<tr>
<td>6.2.2 Spread-error correlation</td>
<td>81</td>
</tr>
<tr>
<td>6.2.3 Spectral dependency</td>
<td>83</td>
</tr>
<tr>
<td>6.2.4 Vertical error and spread structure</td>
<td>87</td>
</tr>
<tr>
<td>6.3 Further variations on ensemble experiments</td>
<td>96</td>
</tr>
<tr>
<td>6.3.1 Doubling initial spread</td>
<td>97</td>
</tr>
<tr>
<td>6.3.2 Reducing the number of ensemble members</td>
<td>97</td>
</tr>
<tr>
<td>6.4 Summary of findings and recommendations</td>
<td>101</td>
</tr>
</tbody>
</table>

## 7 Independent Performance Tests

<table>
<thead>
<tr>
<th>7.1 Description of data set</th>
<th>102</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.2 Ensemble verification and comparison with the 1999 case study</td>
<td>105</td>
</tr>
</tbody>
</table>
8 Summary, Conclusions, and Implications 112
  8.1 Summary of the approach 112
  8.2 A North Pacific bust 113
     8.2.1 Summary of findings related to the case-study storm 113
     8.2.2 Implications 114
  8.3 Broader implications about short-range numerical predictability 115
     8.3.1 Analysis error and forecast uncertainty 115
     8.3.2 Ensemble performance 116
  8.4 Prediction in western North America — some recommendations 118

A Acronym definitions 120

Bibliography 122
List of Figures

2.1 Power spectra of a 24 h forecast 50 kPa geopotential height field on the computational domain........................................ 19
2.2 Complete MC2 domain with verification region for the MC2, MRF, and GEM models................................................. 20
2.3 Zonal mean 50.0 kPa geopotential height spread versus longitude. Each curve represents a different forecast time (h), given in the legend. The mean is determined by averaging all values between 30° and 60°N... 21
2.4 Power spectra of the 50 kPa geopotential height field on the computational domain, calculated with both the 2D FFT and the W transform. 24
2.5 Dynamically-initialized perturbations of the 50.0 kPa geopotential height, valid 00 UTC 10 Feb. 1999. The left panel (a) shows the categorical initialization and (solid) and ensemble member one (dashed). The right panel (b) shows the ensemble mean (solid lines) and the rms spread (shaded, 10 m contour intervals) ........................................ 26

3.1 Map of southern British Columbia (BC), Canada, and northern Washington (WA), USA, showing Vancouver international airport (CYVR) in the Lower Mainland, Tahta Lake West, and Vancouver Island .......................... 30
3.2 CYVR Precipitation and wind speed observations for the period 9–16 Feb. 1999. The frontal passage occurred near 12 UTC 13 Feb., and precipitation prior to that was convective air-mass. Precipitation values are 24 h accumulations from 12 UTC to 12 UTC around the 00 UTC date of each mark ......................... 31
3.3 Analyzed (near 160°W longitude) and 48 h forecast (near 140°W longitude) central pressure (kPa) and location, forecast from 00 UTC 10 Feb. 1999. The Medium-Range Forecast (MRF) model, Global Environmental Multiscale (GEM) model, Mesoscale Compressible Community (MC2) model, and Pacific Weather Centre (PWC) hand analysis are shown. Two low-pressure centers (LPC1 and LPC2) are shown in the PWC analysis at 48 h, whereas the model forecasts only evolved one low-pressure center (LPC2).

3.4 GOES 10 visible image of the North Pacific, valid 00 UTC 10 Feb. 1999.

3.5 GOES 10 visible image of the North Pacific, valid 00 UTC 11 Feb. 1999.

3.6 GOES 10 visible image of the North Pacific, valid 23 UTC 11 Feb. 1999.

3.7 (a) GEM and (b) MRF 50.0 kPa geopotential height analyses and 24 h forecast errors (analysis—forecast), valid 00 UTC 10 Feb. 1999. Contour intervals are 60 m for the forecasts (thick solid contours) and 30 m for the errors (thin dashed contours). Rawinsonde observations are also plotted as station models with observed geopotential heights.

3.8 (a) GEM and (b) MRF 48 h 50.0 kPa geopotential height forecast (contours) and forecast errors (forecast—observed) in meters, valid 00 UTC 12 Feb. 1999. Contour intervals are 60 m.

3.9 GOES 10 visible image of the North Pacific, valid 00 UTC 13 Feb. 1999.

4.1 50.0 kPa geopotential height correlation coefficients r for forecasts (a) 8, and (b) 10 Feb. 1999. The categorical (triangles), ensemble mean (circles), and every ensemble member forecast are shown (other eight lines).

4.2 Normalized 50.0 kPa geopotential height error variance (a) calculated for the entire period, and (b) for the forecast initialized 00 UTC 10 Feb. 1999. Growth rates $\frac{\partial^2 \epsilon}{\partial t^2}$ with units m (6 h)$^{-1}$ are shown in (c) for the entire period, and (d) for the 10 Feb forecast. Results from all the experiments except DENY are shown (see text for explanation of experiments).

4.3 Normalized 50.0 kPa geopotential height error variance of each ensemble forecast at 24, 48, and 72 h. (a) shows results from experiment PERT, and (b) shows results from experiment PHYS+MRF. The date on the abcissa refers to the date of each initialization, at 00 UTC.

4.4 (a) PERT ensemble and (b) PHYS ensemble storm tracks and 48 h central pressures from the forecast initialized 00 UTC 10 Feb. 1999. The MRF forecast is also shown in (b) as open circles.
4.5 Normalized 50.0 kPa geopotential height error variances of DENY, showing (a) the experimental period mean (solid) and for the forecast initialized 10 Feb. 1999 (dashed). Also shown is (b) the daily variability of the error variances at 24, 48, and 72 h. 

4.6 Number of aircraft observations within the box bounded by 35° to 60°N latitude, and 120° to 160°W longitude. The data assimilation system at NCEP did not modify or throw the observations deemed “good”. 

4.7 Comparison of perfect ensemble configuration and an ensemble verified against the GEM analysis. (a) shows the evolution of the perfect ensemble error variance $E$ (solid) and spread $S^2$ (long dash). The bias$^2$ is determined against GEM. (b) compares $S - E$ correlation coefficients ($r$) from both verifications. All units are m$^2$. 

4.8 Comparison of the error variance $E$ averaged over the whole period with that for the forecast initialized 10 Feb: (a) results from the perfect configuration, and (b) results when the ensemble is verified against GEM analyses. All units are m$^2$. 

4.9 Daily variability of error measures for the (a) perfect ensemble configuration, and (b), (c), and (d) verification against GEM analyses valid at 24, 48, and 72 h forecasts. (a) and (b) are plots of $E$, (c) shows bias$^2$, and (d) shows the $mse$. The curves in (b) and (c) sum to the curves in (d). All units are m$^2$. 

5.1 Comparison of 50.0 kPa geopotential height perturbations for ensembles (a) DIFF-MRF and (b) RND-MRF, valid 00 UTC 5 Feb. 1999. All contour intervals are 20 m. Dashed lines show MRF–GEM differences and solid lines show initial spread. 

5.2 Period-averaged 00 UTC 50.0 kPa geopotential heights at 120 m intervals (solid lines), and period averaged absolute MRF–GEM differences at 10 m intervals (dashed lines). 

5.3 Period-averaged 2D FFT spectra of 00 UTC analyzed geopotential heights ($Z$) at 100.0, 85.0, 50.0, and 25.0 kPa. Panel (a) shows MRF results, and panel (b) shows GEM results. The curves are normalized by the variances shown in Table 5.1. The wavenumbers $k$ are grid-relative. 

5.4 Period-averaged 00 UTC MRF–GEM analysis geopotential height 2D FFT spectra at (a) 100.0, (b) 85.0, (c) 50.0, and (d) 25.0 kPa. The wavenumbers $k$ are grid-relative.
5.5 DIFF-MRF and DIFF-GEM ensemble mean \( rmse \) against rawinsonde observations, and \( rms \) GEM–MRF differences for the 50.0 kPa geopotential height fields valid at 00 UTC on each day 5-12 Feb. 1999.

5.6 Period-averaged 50.0 kPa geopotential height \( rmse \) and \( S \) vs. forecast hour for the perfect ensemble configuration.

5.7 Period-averaged 50.0 kPa geopotential height \( rmse \) and \( S \) vs. forecast hour. DIFF-GEM is verified against MRF analyses, and DIFF-MRF is verified against GEM analyses.

5.8 Period-averaged 50.0 kPa geopotential height \( rmse \) and \( S \) vs. forecast hour. Ensemble FULL combines DIFF-MRF and DIFF-GEM runs, ensemble CNTL is the two control runs, and ensemble mean \( bias \) is added to the \( S \) curves in panel (b). The \( rmse \) curves for FULL and CNTL are nearly indistinguishable.

5.9 Period-averaged 50.0 kPa geopotential height \( S-rmse \) (spread-error) correlation coefficients vs. forecast hour.

5.10 Period-averaged 50.0 kPa geopotential height \( S \) and \( rmse \). \( S \) is a combination of the FULL \( S \), its \( bias \), and the PHYS+MRF \( S \).

6.1 Comparison of experiments (a) PERT, (b) DIFF-MRF, (c) DIFF-GEM, and (d) FULL 50.0 kPa geopotential height \( S \) and \( rmse \) in the perfect ensemble configuration.

6.2 Spread growth rates \([m \,(12 \,h)^{-1}]\) for experiments PERT, DIFF-MRF, DIFF-GEM, and FULL 50.0 kPa geopotential height in the perfect ensemble configuration.

6.3 Spread-error \((S-E)\) correlation coefficients for the perfect ensemble configuration.

6.4 Comparison of experiments (a) PERT, (b) DIFF-MRF, (c) DIFF-GEM, and (d) FULL 50.0 kPa geopotential height \( rmse \) verified against GEM analyses. Thin lines show categorical forecasts and thick lines show ensemble-mean forecasts.

6.5 Comparison of experiments (a) PERT, (b) DIFF-MRF, (c) DIFF-GEM, and (d) FULL 50.0 kPa geopotential height \( rmse \) verified against rawinsonde observations. Thin lines show categorical forecasts and thick lines show ensemble-mean forecasts.

6.6 Spread-error \((S-E)\) correlation coefficients, verified against (a) GEM and (b) MRF analyses.
6.7 Experiment PERT period-averaged 2D FFT spectra of 50.0 kPa geopotential height spread $S$, ensemble mean $rmse$, and categorical forecast $rmse$ at (a) initialization and (b) 72 h. Also shown are the time evolution of the (c) spread, and (d) ensemble mean error. The legends in the left column are valid in the adjacent frames. Wavenumbers $k = 2, 10$, and 60 correspond roughly with wavelengths 6000, 1200, and 200 km, respectively.

6.8 Experiment DIFF-MRF period-averaged 2D FFT spectra of 50.0 kPa geopotential height $S$, ensemble mean $rmse$, and categorical forecast $rmse$ at (a) initialization and (b) 72 h. Also shown are the time evolution of the (c) spread, and (d) ensemble mean error. The legends in the left column are valid in the adjacent frames.

6.9 Experiment DIFF-GEM period-averaged 2D FFT spectra of 50.0 kPa geopotential height $S$, ensemble mean $rmse$, and categorical forecast $rmse$ at (a) initialization and (b) 72 h. Also shown are the time evolution of the (c) spread, and (d) ensemble mean error. The legends in the left column are valid in the adjacent frames.

6.10 Experiment FULL period-averaged 2D FFT spectra of 50.0 kPa geopotential height $S$, ensemble mean $rmse$, and categorical forecast $rmse$ at (a) initialization and (b) 72 h. Also shown are the time evolution of the (c) spread, and (d) ensemble mean error. The legends in the left column are valid in the adjacent frames.

6.11 Vertical profiles of geopotential height ensemble-mean $rmse$ (against GEM analyses) for each of the ensembles, valid at forecast hour (a) 0, (b) 24, (c) 48, and (d) 72.

6.12 Vertical profiles of geopotential height ensemble spread for each of the ensembles, valid at forecast hour (a) 0, (b) 24, (c) 48, and (d) 72.

6.13 Vertical profiles of ensemble-mean geopotential height $rmse$ (against GEM analyses) components $E_\beta^1$ and $bias$.

6.14 Vertical profiles of geopotential height ensemble spread every 24 h, for each ensemble (a) PERT, (b) DIFF-MRF, (c) DIFF-GEM, and (d) FULL.

6.15 Vertical profiles of geopotential height ensemble-mean $rmse$ (against GEM analyses) every 24 h, for each ensemble (a) PERT, (b) DIFF-MRF, (c) DIFF-GEM, and (d) FULL.
6.16 Summary of ensemble PERTx2 50.0 kPa geopotential height performance. The perfect ensemble configuration spread and skill is shown in (a). The other panels show (b) rmse, (c) S, and (d) S-E correlation of ensembles PERT and PERTx2.

6.17 Summary of ensemble FULL8 50.0 kPa geopotential height performance. The perfect ensemble configuration spread and skill is shown in (a). The other panels show (b) rmse, (c) S, and (d) S-E correlation of ensembles FULL and FULL8.

7.1 Averaged 2D FFT spectra of Aviation model 00 UTC 50.0 kPa geopotential heights (Z). A total of 37 days between March and July 2001 are included in the average.

7.2 Average ensemble AVNPERT performance. Panel (a) shows the perfect ensemble configuration, and the spread and error are nearly indistinguishable. Panel (b) shows comparison against Eta model analyses.

7.3 Comparison of ensemble error and the corrected spread that includes the AVNPERT S, its bias, and the AVNPHYS S.

7.4 Comparison of spread-error correlations for ensembles AVNPERT and AVNPHYS in the (a) perfect-ensemble configuration, and (b) verification against Eta analyses.

7.5 Experiment AVNPERT period-averaged 2D spectra of 50.0 kPa geopotential height S, ensemble mean rmse, and categorical forecast rmse at (a) initialization and (b) 72 h. Also shown are the time evolution of the (c) spread, and (d) ensemble mean error. The legends in the left column are valid in the adjacent frames.

7.6 Experiment AVNPHYS period-averaged 2D spectra of 50.0 kPa geopotential height S, ensemble mean rmse, and categorical forecast rmse at (a) initialization and (b) 72 h. Also shown are the time evolution of the (c) spread, and (d) ensemble mean error. The legends in the left column are valid in the adjacent frames.
List of Tables

2.1 The MC2 model dynamics and discretization. .......... 18
2.2 The MC2 model physical parameterization schemes. .. 18

4.1 Contingency tables of ranked forecast spread (rows) and ranked error (columns) for 00, 24, 48, and 72 hour 50.0 kPa geopotential height forecasts over the entire experiment period. Each element in the table is based on the average of about 4000 values. .......... 46

5.1 Total period-averaged 00 UTC analysis geopotential height variances (m^2). ........................................ 62
5.2 Contributions to the total period-averaged 50.0 kPa geopotential height spread (m) shown in Figure 5.10. ..................... 71

7.1 List of dates included in ensembles AVNPERT and AVNPHYS. ... 103
7.2 Contributions to the total period-averaged 50.0 geopotential height spread (m) shown in Figure 7.3. ..................... 106

A.1 Miscellaneous acronym definitions. ....................... 120
A.2 Operational forecasting centers (country location). ....... 120
A.3 Model acronyms. ........................................... 121
A.4 Ensemble experiments. .................................... 121
Chapter 1

Introduction and literature review

1.1 The role and effectiveness of numerical weather prediction

Though the concept of numerical weather prediction (NWP) was developed decades earlier, it was not until the 1960's that scientists were able to document the connection between human forecast skill and NWP skill. In 1966, Charney et al. (1966) described the problems that would face NWP and forecasting during subsequent decades. He estimated a two-week limit of predictability with imperfect observations, and suggested that under the hypothesis that the atmosphere is deterministic or nearly deterministic, we can approach this limit by accounting for more physics in the models (including condensation, which was largely ignored up to that point) and addressing truncation error by increasing resolution. He also recognized that the atmosphere must be observed as "a single entity", and that a useful forecast depends on implementing a timely and coherent global observation system. By stating that beyond two weeks, averages and statistical distributions can be forecast, he built on then-current work by Lorenz and Epstein, and provided impetus to pursue probabilistic forecasts.

For the two decades following Charney's paper, national forecast centers in the developed world attacked the NWP problem vigorously, staying near computational limits and getting closer to estimated limits of predictability. In 1989, Shuman (1989) reviewed the progress of U.S. National Meteorological Center (NMC) forecasts, and linked improvements directly to improvements in modeling systems. Examining the S1 score (Wilks 1995) of 50.0 kPa geopotential heights, he documented a steady improvement from 1955, when the forecasts were purely subjective. Each major jump in the S1 score was linked to the implementation of a new, more sophisticated
model, which was often allowed by a new computer acquisition. Shuman (1989) suggested that by 1960, with only a barotropic (one-layer) model in use at NMC, the quality of some forecast products had already surpassed what could be produced manually without numerical guidance. He also estimated that about 95% of the NMC forecast products were automated, and the remaining 5% consisted primarily of sensible weather products such as precipitation and warnings.

For most of the first few decades of NWP, limitations in computer power traded with limitations in understanding of atmospheric dynamics and numerical technique, and the two progressed steadily hand-in-hand. But Shuman (1989) suggested that the observation network was as limiting as computer power. He also noted that the scores over western North America were consistently worse than further east. By acknowledging that the observation system limits predictability, he implied that improved forecasts with greater grid resolution was nearing a point of diminishing return. Up to that point, his S1 analysis led him to conclude that an order of magnitude increase in computer power allowed a doubling of resolution, resulting in a maximum increase of 15% in skill.

Since 1989, computer power and model resolution have continued to progress rapidly, and the limitations of the observing network have likely become more acute. Similar to Shuman (1989), Kalnay et al. (1998) more closely examined operational NWP, and compared it to human forecasts, stating that “beyond the first 6-12 h, the forecasts are based mostly on numerical guidance.” They noted that prior to 1976, three-day forecast skill was largely constant, but had steadily improved since, owing to the ability of NWP to show skill beyond three days. Kalnay et al. (1998) also reported that for the three-day sea-level pressure anomaly correlation, the human forecast scores tracked the NWP scores from 1968 onward, but human forecasters’ scores were always slightly better than NWP alone.

The studies discussed above focused on predicting mid- or upper-tropospheric forecast parameters, but the true value of a weather forecast is normally measured by how it affects a decision-making process that relies on sensible weather products. Several studies have attempted to quantify the contribution and economic value of NWP. Olson et al. (1995) found that the slight quantitative precipitation forecast (QPF) advantage of human forecasters over NWP alone had been nearly constant for 30 years. Roebber and Bosart (1996) examined several years of precipitation forecasts from a selection of systems that included both NWP and human forecasts using NWP. Their cost-loss analysis showed only a slight gain over NWP alone, reflecting what Kalnay et al. (1998) found from sea-level pressure. They noted that while improvements in forecasts aloft have been steady, improvement in surface sensible-weather forecasts were not as large as during the decade prior to their study. They attributed
the declining rate of improvement to a saturation in model ability to forecast precipitation, and the advantage is realized only by forecasters associating model behavior with specific synoptic situations.

Consensus in the literature indicates that since its inception, NWP has been overwhelmingly responsible for forecast improvements of all weather parameters. Thus it behooves researchers to continue their work on NWP systems. Improving NWP necessarily means understanding the sources of error, which intuitively arise from error in initial atmospheric state conditions and error in the NWP model itself. Observations are imperfect, and initial condition (IC) error may include instrument and representativeness errors in the observations (Daley 1996), mistaken assumptions about the error characteristics, and misunderstanding of the relationship between an observation and state variables. Models are imperfect approximations to real atmosphere evolution, and model error may include errors of representativeness, errors in estimating physical constants, truncation and other numerical errors, and misunderstanding of physical processes. Each of these error sources defines its own body of research, and some are discussed below.

1.2 Chaotic systems

Lorenz (1963), without using the term “chaos”, described the conditions under which the atmosphere would behave chaotically. He truncated a model of convection, initially described by Saltzman (1962), to arrive at a nonlinear system of three ordinary differential equations with three time-dependent variables. Still used in predictability studies today (Kalnay et al. 2000, Krishnamurti et al. 2000), it is often called the “Lorenz system”, and represents a forced, dissipative (nonconservative), hydrodynamical system. Integrated forward in time, the state of the system in phase space (a sub-space of which can be represented by plots of dependent variables against other dependent variables, rather than against independent variables) is bounded, but it never repeats itself. It follows that a trajectory in phase space is never duplicated, but it may come arbitrarily close to an earlier trajectory segment for an arbitrary length of time. Thus it is nonperiodic (also called quasi-periodic), even under constant forcing and constant dissipation. Lorenz explains that this behavior is different than other studies showing that linear systems of equations, subject to random forcing, will produce random solutions.

His pioneering study has important implications for the atmosphere, where periodicity has not been observed, so it is assumed to be nonperiodic. First, a limit on predictability exists unless the present state of the atmosphere is known exactly, which
is impossible. Second, we may explore the predictability of idealized systems, which may include NWP models, by comparing solutions obtained from slightly different ICs. His study laid the foundation for all atmospheric predictability and ensemble forecasting studies that followed.

Extending his own ideas, Lorenz (1965) examined the predictability of a 28-variable model, concluding that errors tend to project along non-random directions in phase space. Assuming that observation error can be made arbitrarily small, but not zero, and that the equations describing the evolution of the atmosphere are exactly known (a perfect model), he showed that seemingly indeterministic solutions can be obtained from a deterministic system. A solution from some arbitrary ICs may then be regarded as an exact solution of the deterministic system, or the “truth”. The difference between the exact solution, and the solution from ICs that are slightly different from the ICs of the exact solution, is the error.

Using a long integration as control and truth, Lorenz made small (linear) random perturbations and analyzed the error growth over forecast periods of several durations. In phase space at the initial time, the random perturbations all lie on a sphere surrounding the truth. Integrating the model from the perturbed states leads to a distortion of the sphere, with the largest semi-axis corresponding to the direction of the largest error growth. The 28-variable model can be linearized, and a new system of equations results when a small error term is added. The solution of this system contains coefficients for the error in each variable. The system can be written in matrix form, where the error coefficients define an error-propagator matrix ($A$). Lorenz showed that the eigenvalues of $AA^t$, where $t$ denotes transpose, are in fact related to the axes of the error ellipse in phase space. The members of $A$ are a function of the linear model and the initial state. Thus, the direction of maximum error growth can be predicted as the direction of the largest eigenvalue of $AA^t$.

The conclusions of Lorenz (1965) had critical implications for further ensemble and predictability studies. First, the growth of errors is highly dependent upon the flow of the day. They are not random, but grow primarily in directions that are dynamically sensitive in the linearized model. Other components of random error actually shrink. Second, the first few eigenvalues dominate the error growth after only a short time, and at longer ranges the error is further skewed toward the largest eigenvalues, rendering lower-ranked eigenvalues less important. Finally, error growth can be adequately characterized by considering only the first few eigenvalues, thus a few ensemble members with appropriate perturbations can account for most of the distribution of random errors. Operational numerical forecasting groups at the U.S. National Centers for Environmental Prediction (NCEP) and the European Centre for Medium Range Weather Forecasting (ECMWF) adhere to these conclusions in their
ensemble systems, focusing on only the errors that will grow. Scientists at ECMWF go one step further, and explicitly determine the eigenvalues similarly to Lorenz (1965) to construct their perturbations.

Lorenz (1969) again followed his own work by studying a system that more closely resembled the earth's atmosphere. Again assuming a perfect model, he examined the chaotic nature of the forecasts. If error at some future time cannot be reduced below some limit, even when the difference between ICs is as small as possible above zero, the system appears indeterministic. Defining "predictable" as the condition that the error is less than the difference between randomly chosen states of the system, the atmosphere has a finite range of predictability with all but the unattainable perfect ICs. Also, reduction in IC error will likely not result in a matching forecast error reduction.

When many scales of motion are considered, uncertainty in one scale affects the uncertainty in other scales, further reducing predictability. In his experiments with realistic atmospheric scaling, predictability changed when the energy spectrum (a prescribed parameter in his experiments) had a different shape, indicating that the predictability of certain systems can be arbitrarily increased by improving ICs. Unfortunately, the observed shape of the atmospheric energy spectrum is not within this class, suggesting that atmospheric predictability is in fact limited. Lorenz found that cumulus scales have an inherent predictability of just under an hour, while synoptic scales may be predictable to a few days. Those results were consistent even when the initial error was confined to the largest scales; and when the initial errors were halved, the predictability did not increase appreciably. Thus the atmosphere behaves indeterministically except when the error is identically zero.

Lorenz allowed that these estimates are likely not correct because of the many simplifications and assumptions that he employed, and that the error growth rates were likely overestimated, but the evidence of a finite limit to predictability is difficult to ignore. The concept of probabilistic forecasts was formalized to maximize the practical predictability within the limits of theoretical predictability.

1.3 Probabilistic forecasts

The work of Gleeson (1966, 1967), and Epstein (1969a,b) formalized the fact that any estimate of the unknown true state of the atmosphere is probabilistic, and therefore any prediction is also probabilistic. Traditionally, ICs have been defined as some form of maximum-likelihood estimate. In reality, no single estimate of the state of the atmosphere, given unavoidable observation error, can be known as the optimal or
best estimate. Rather, an infinite number of states are possible, bound by the physics of the earth-atmosphere system and the capabilities of the model. All possible states define a probability density function (pdf) for a given time. In the context of NWP, all possible ICs for a forecast period define an initial pdf, which evolves through a model to define a forecast pdf.

Gleeson (1966, 1967) used simple prognostic systems, both linear and nonlinear, to show that IC uncertainty results in forecast uncertainty, and demonstrated some methods for examining the relationship between IC and forecast uncertainty. Epstein (1969a) solidified the concept by re-examining Gleeson’s examples with the goal of understanding the behavior of individual members of an ensemble versus the behavior of the first moment of the pdf — the ensemble mean. He concluded that the evolution of the ensemble mean will not be identical to the evolution of any one member of the ensemble. His result implies that it is impossible to forecast the state of the atmosphere with a deterministic system, as long as the ICs are imperfect, as they always are. Because the ICs are really probabilistic, the forecast must be probabilistic as well.

In the same year, Epstein (1969b) demonstrated a stochastic method for producing a probabilistic forecast by explicitly forecasting the first and second moments of the pdf of an extremely simplified form the Navier-Stokes equations (Lorenz 1969). He showed how to derive extra prognostic equations for these moments, and used assumptions about the statistical nature of the higher-order moments to close the system. For comparison, he also performed Monte Carlo forecasts. He concluded that stochastic prediction provides a better forecast, and useful information about the uncertainty. He also discusses what a stochastic prediction can be used for, including long-range forecasting, analysis of observing systems, and evaluating models. By introducing all these topics, Epstein essentially set the stage for modern ensemble applications. Computational demands are the major drawback of the stochastic method.

1.4 IC ensemble methods

Ensemble forecasting methods using IC perturbations have traditionally fallen into three groups, each attempting to exploit different aspects of the analysis procedure. The first, classical, approach is generally a Monte Carlo method, where ensemble members are chosen randomly, and it is assumed that all are equally likely. Another approach identifies sensitive dynamic properties of the flow, and perturbs those. Some of the world’s leading operational NWP centers, such as NCEP, and the ECMWF in
England, take this approach. The third approach, developed more recently, uses an ensemble as a filter that helps to define the actual shape of the initial pdf, given a set of observations, and attempts to approximate it with a Monte Carlo ensemble. This approach is implemented as an essential part of the data assimilation cycle. Most of this work has been done with simplified models, but experimentation with more complex sets of equations is ongoing. The next three sections describe each approach, and discuss some advantages and disadvantages of each.

1.4.1 Monte Carlo methods

Monte Carlo approaches are truest to the approximate stochastic solution, as described by Epstein (1969b). Leith (1974) examined the theoretical skill of Monte Carlo forecasts, assuming that the categorical analysis defines the first moment of the pdf. It is optimal in the least-squares sense if the nature of the observational error is known. He ignored model error, assuming that one of the forecasts in the ensemble could actually follow the true evolution of the atmosphere. He also applied a linear-regression correction to the Monte Carlo ensembles, using known error statistics — a technique preceding current nonlinear filter approaches. Using random perturbations in phase space, he concluded that Monte Carlo forecasts are potentially as useful as stochastic forecasts, but are much cheaper. He noted that by averaging the ensemble, smaller, less-predictable scales are filtered out, and the larger, more predictable scales remain, thereby improving skill. Ensemble forecasters still have to grapple with that problem of loss of small-scale information.

Following Leith (1974), the literature is full of applications of Monte Carlo forecasts. Errico and Baumhefner (1987), Tribbia and Baumhefner (1988), and Mullen and Baumhefner (1989, 1994) all used random perturbations to generate ensembles on limited-area domains. They constructed perturbations with spectra that approximate the spectra of the actual two-dimensional error as estimated by Daley and Mayer (1986). More specifically, Errico and Baumhefner (1987) examined factors that limit ensemble spread when a limited area model (LAM) is used. They found that using common boundary conditions appears to limit spread between ensemble members, as differences are swept out of the domain at mean advective speeds. While local spread can double rapidly, domain-averaged spread (error) growth is much slower and shows the effects of common boundaries. They also restricted perturbations to small scales to examine the scale-evolution of the spread. Their results indicate that large-scale forecasts control small scales, and that if the large scales do not diverge the spread is limited across the spectrum. They also noted that spread reduced when an initialization (balancing) procedure was used, and during a geostrophic adjustment.
Mullen and Baumhefner (1989) used Monte Carlo simulations to examine the impact of IC uncertainty versus model uncertainty in forecasting explosive cyclogenesis. They found large case-to-case variability in sensitivity among explosively deepening cyclones, leading them to speculate that both storm dynamics and differences in imposed random errors contribute to total forecast error. This implies that a specific random perturbation may not adequately represent IC error where it would interact with sensitive storm dynamics. The variability also shows why studies should include ensemble runs over as many cases as possible before generalizing conclusions. Using the analysis of significant improvement in Tribbia and Baumhefner (1988), they showed that explosively deepening storms have a much greater sensitivity to IC error than the flow around them. But because all of the forecasts from perturbed ICs still deepened explosively, they concluded that the dynamics of the storms are strong enough to overcome random perturbations, heuristically demonstrating that a reduction in model error would still result in an improvement for short-range forecasts of explosive cyclogenesis.

The papers reviewed above demonstrate disadvantages with the Monte Carlo approach, including the necessity for assumptions about the structure of the IC errors, no explicit relationship to the flow dynamics, and a perceived lack of spread between the perturbed forecasts, especially at short ranges. Many of the studies reference Daley and Mayer (1986), and construct perturbations that have similar spectral and magnitude characteristics as errors estimated in that study.

Daley and Mayer (1986) estimated global analysis error for objectively analyzed fields with an observation system simulation experiment (OSSE). Judging by the number of times it has been referenced, it has been regarded as one of the best estimates of IC error, but the authors admit flaws in the study. Assumed random and systematic observation and representativeness errors were added to the pseudo-observations in their experiments. Though not addressed, the results could be sensitive to those assumptions. The general circulation model (GCM) that generates the “truth” in their study is run for only one month, and is known to drift relative to the real atmosphere.

The short sample used by Daley and Mayer (1986) provides results that are valid for only the particular regime that dominated during that period, because observational errors may be regime dependent. The drift, which is likely different from the model that serves as background for the objective analysis scheme, means that results are suspect in data-sparse regions. Finally, they used an observing network that was valid in 1979. They acknowledged flaws in the error estimate, and the fact that the observing network has changed since 1979, suggest that their quantitative results should not be strictly followed when creating perturbations. Daley and Mayer (1986)
also argue that their error estimates represent a lower bound on actual analysis errors. They found that above wavenumber 20, the error spectrum is saturated (the error is as strong as the signal of the field itself), but this particular wavenumber may depend on flow regime, the observation network, the assimilating model, and assumptions about the observation error. Still, some of their more general results are intuitive and may be useful.

1.4.2 Perturbation of sensitive modes

The second class of ensemble forecasts that are used extensively are based on the assertion that random errors are not important in an analysis cycle, and that perturbations should be concentrated on the fastest-growing modes in the atmosphere. NWP researchers at NCEP have pursued a method called the "breeding method", while at ECMWF researchers developed a singular-vector approach with a tangent linear and adjoint version of their medium-range operational model. Both of these methods are designed to project perturbations on the fastest-growing modes in the model atmosphere. The perturbations are similar, but contain important differences.

Toth and Kalnay (1993) described the breeding method, arguing that random perturbations should not be included in an ensemble at all, because they do not adequately project upon the growing modes that account for most of the forecast error. The breeding method attempts to extract the most important realistic perturbations from the analysis cycle, making it an important step toward linking data assimilation and ensemble forecasting.

Perturbations are constructed in the analysis (assimilation) cycle, by scaling the difference between a short-range control forecast and a short-range perturbed forecast to create a new perturbation. Toth and Kalnay (1993) argued that the differences represent the fastest growing forecast errors, which are only partially removed in the analysis cycle by the addition of new observations, and are thus "naturally selected" in the cycle. They wrote that perturbations should be constructed along only those modes because they dominate the forecast error, and because of the large number of Monte Carlo forecasts that it takes to adequately sample the initial pdf. Randomly selected perturbations will project mostly on non-growing modes, and will therefore have little effect on the forecast, they argued. Their results from a few case study days show an improved anomaly correlation over the control forecast, with as little as two bred members of the ensemble. They were unable to achieve the same improvement with Monte Carlo ensembles of 16 members.

The linear error-propagator concepts developed by Lorenz (1965) form the basis for the singular-vector (SV) approach, described in Molteni and Palmer (1993), Mureau
et al. (1993), and Buizza (1994a,b, 1995). A model is linearized, and an error term is added. The solution of this system of equations defines an error-propagator matrix. The eigenvalues of the error-propagator matrix squared (singular vectors) reveal the directions of the greatest error growth. They are dependent upon the time over which the operator is applied. The adjoint of the linear model projects the errors back onto the initial state. The errors are used to generate perturbations for an ensemble forecast. Molteni and Palmer (1993) described the initial implementation of the SV approach at the ECMWF. They used a linearized quasi-geostrophic (QG) model, and its adjoint, to determine errors that could be projected onto the categorical analysis that results from the ECMWF data assimilation system.

Because the SV approach does not determine the sign of the error vector, but only the axes of the error ellipse (Lorenz 1965), each of the error vectors is both added and subtracted to the initial conditions. Molteni and Palmer (1993) also scaled the initial perturbations to agree with analysis error estimates. They reported that SV growth far exceeded growth of normal modes, which do not explain observed forecast error growth.

In a companion paper, Mureau et al. (1993) applied the approach to four cases, finding both positive and negative aspects. They looked at the correlation between ensemble spread and skill. That is, one utility of an ensemble of forecasts is to forecast the current predictability of the atmosphere. A large spread between ensemble members should indicate an unpredictable forecast that would likely display low skill, and vice versa. The four forecasts showed evidence of a spread-skill correlation. They also found that ensemble skill improved with larger initial perturbations. But in general, they reported that the ensemble spread is lower than desired, or the ensemble was under-dispersive. They attributed this feature to the difference between the QG model (likely used for computational efficiency) and the actual forecast model, and suggested that a linear version of the full forecast model (using the primitive equations) would be more successful. They also noted that some consideration must be given to actual analysis error, which implies that information about both the most rapidly-growing modes and the likely analysis errors should be considered when constructing perturbations to most accurately represent forecast uncertainty.

Many of the follow-up papers by scientists at ECMWF describe refinements to the SV approach, focusing on computed vectors that are more appropriate for the real medium-range forecasting problem over Europe. Buizza (1994a) described implementation of a simple PBL turbulence scheme in the linearized model to reduce the appearance of non-physical SVs near the earth’s surface. He also examined the choice of time interval over which the error-propagator is applied (called the optimization time interval or OTI), concluding that 36 hours is most appropriate given
computational limitations. It is long enough to allow the errors to organize into desirable dynamic structures, but short enough to be computationally feasible. The linear approximation is still assumed valid with an OTI of 36 h, meaning that the linearized version of the model should not diverge substantially from the fully-nonlinear model.

Buizza (1994b) continued to describe perturbation localization using a projection operator, arguing that perturbations that affect the medium-range forecasts over Europe should be chosen to address operational predictability there. This amounts to discarding SVs that may be in other parts of the globe where the forecast is not relevant to Europe. Thus, the ensemble contains perturbations along vectors that would not have been otherwise perturbed, because they were ranked too low. The ensemble spread over the European verification area was increased with the application of the projection operator. Still, Buizza (1994b) considered the ensemble to be under-dispersive. He speculated on a possible connection to model error. That is, the errors in physical parameterizations are too large for an ensemble to account for the true evolution of the atmosphere. The lack of some physical processes in the linear and adjoint versions of the model may also contribute. In 1995, Buizza further examined the evolution of the perturbation amplitude. He concluded that nonlinear processes become important at 2-2.5 days lead time. Increasing the amplitude of the initial perturbations had the effect of increasing the ensemble spread while decreasing the skill. Decreasing perturbation amplitude had the opposite effect.

The ensemble method used at ECMWF focuses on medium-range forecasts, and many of the papers examining ensemble performance lament that the ensemble spread is too narrow for the true evolution of the atmosphere to always lie within the envelope of ensemble members. Consequently, many of the improvements to the system have been designed to increase ensemble spread, despite Mureau et al. (1993) recognizing that some information about actual analysis error should be included in the perturbation strategy.

The goals of correctly sampling a pdf of initial conditions, and of maximizing the ensemble spread for a specific verification domain and time, are to a large extent mutually exclusive. Large spread could be generated even when true analysis error is small, if perturbations are along sensitive vectors that are correctly analyzed. Buizza (1995) reported a reduction in skill when the perturbation amplitude was increased. This result is not really a surprise when any spread-skill correlation exists. Larger perturbations to the most rapidly growing modes should produce a larger spread. If the true evolution of the atmosphere lies outside the envelope of ensemble forecasts, either model error is dominating, or the analysis is sufficiently poor that no model realization can ever follow its evolution, and may even tend toward a completely different attractor.
For all these reasons, it may not be appropriate to intentionally (and unrealistically) maximize the spread. Rather, an ensemble should be designed to adequately sample the pdf defined by realistic analysis error. This is especially true for short-range forecasts, when the nonlinearity may not yet dominate. To ensure that the true evolution of the atmosphere lies within the ensemble, the models and the analysis procedures must be improved.

1.4.3 Relation to data assimilation

A natural link between data assimilation (DA) and ensemble forecasting exists because DA systems require a statistical characterization of forecast error, and ensemble forecasting systems are designed to produce it. DA systems need background (forecast) error covariances to properly weight new observations when updating an analysis. In traditional analysis systems (such as optimal interpolation), covariances are determined by explicitly comparing forecasts and observations over many forecast cycles, or are assumed homogeneous, isotropic, and independent of the flow. But covariances can also come from an ensemble of forecasts with the correct properties.

One advantage of this approach is that the covariances are time dependent, and can thus be related to the “flow of the day”. Another advantage is that the DA system yields information about the uncertainty in the analysis, which can then be used to generate ensembles that capture this uncertainty. Because of the adaptive nature of a DA/ensemble system sharing information, it is often referred to as a Kalman filter when the error statistics are assumed Gaussian.

One way to write a minimum-variance update step in a data assimilation system is

\[ \psi^a = \psi^f + K (O - H \psi^f) \] (1.1)

where \( \psi^a \) is the resulting analysis state vector, \( \psi^f \) is the background forecast model state vector, \( K \) is a Kalman gain matrix. The Kalman gain matrix operates on the observation increment, where \( O \) is the observation vector, and \( H \) is relates model variables to observations. \( K \) is a combination of background error covariances and observation error covariances that determine the relative weighting of the background and observations. Equation 1.1 determines a categorical analysis from which a new forecast can be launched.

Evensen (1994) introduced an Ensemble Kalman Filter (EnKF), where each member of an ensemble coexists with its own updating step such as equation 1.1. The Kalman gain matrix is redefined to include information on error covariances derived...
from the forecast ensemble, and the relationship between the analyzed and forecasted ensemble mean is the same as the categorical analysis and forecast. Evensen (1994) showed some problems with using the full EnKF as originally formulated, related to using common observations within each ensemble member DA cycle. Burgers et al. (1998) solved this problem by adding different random errors to the observations within each cycle, and treating them as random variables. This approach allows the ensemble covariance to appropriately model the error covariances.

Houtekamer and Mitchell (1998) further investigated the EnKF approach with small ensembles, and applied it to a different model. To maintain ensemble quality even with a small ensemble, they used two independent EnKFs, and checked them against each other. Their results show that using a pair of filters helps the ensemble maintain a spread that is representative of error due to initial conditions. Hamill and Snyder (2000) compared a similar approach in a simplified environment and found that the resulting ensemble had better characteristics than either bred or optimal perturbation ensembles.

The ability of the EnKF to adapt to current flow regimes, for daily variability in ensemble spread to exist in the analyses, and for the ensemble mean to improve on the control (categorical) analysis makes this method an attractive alternative for ensemble prediction and DA schemes. But several hurdles still exist. The studies to date have used atmospheric models that are much simpler and contain many fewer degrees of freedom than current operational models. It is not yet clear how these techniques would extend to a full-dimensional model. Practical difficulties include the expense of performing matrix operations on huge matrices (with rank on the order of $10^6$ or larger) for each ensemble member. Because of the expense of running an ensemble with a full-dimensional model and a coupled DA system, research is limited outside of major operational NWP centers. Still, it has shown enough promise to spur ongoing work.

1.5 Model error

Sources of model error include discretization (truncation) and other numerical errors, and errors in representing sub-grid scale physical processes such as turbulence or cloud microphysics. It is extremely difficult to quantify, and literature on its characteristics is sparse, but it affects all possible ensemble experiments. To avoid the issue, most of the literature reviewed above assumes a "perfect model". To assume a model is perfect for an ensemble experiment, the model does not need to produce a perfect forecast given a perfect analysis (neither of which is achievable anyway), but it must
produce a statistically realistic forecast.

A climatology of model forecasts does not even need to accurately reproduce climate means (a bias is easily removed), but the model climatological variance must match the climatological variance of the atmosphere. This restriction ensures that an ensemble with a climatological spread that is the same as the climatological forecast error of the control can be used to generate a climatological estimate of uncertainty. It is tantamount to saying that a model can reproduce an energy spectrum of all resolvable scales that is the same as the real atmosphere. For example, if a model cannot produce the same energy in its representation of many mid-latitude storms as actually might exist in a set of such storms of the same magnitude, then any information about uncertainty or error gleaned from an ensemble experiment is unreliable. This says nothing about the daily variability of spread or error, and an ideal ensemble experiment includes many cases to generate meaningful estimates of climatology.

Tribbia and Baumhefner (1988) used ensembles to estimate the relative contribution of model and IC errors, but explicitly characterizing model error has been much more elusive. Furthermore, because model error in such experiments is estimated by comparing different models, the common assumptions and errors in the models mean that those estimates represent a lower bound on model error (Orrell et al. 2001). Mitchell and Houtekamer (2000) address model error within an EnKF framework. They parameterize model error, and allow the EnKF system to incorporate it into the ensemble. Model error covariances can then be estimated from the filter. Orrell et al. (2001) take a more formal approach, developing a method for evaluating model error. By looking specifically for model drift away from a target orbit that lies on a model attractor, they find that model error dominates forecast error in the very short-term (a few hours). Their results exist as counter-examples to the notion that model error is small relative to IC error, but they neither weaken the perfect model assumption nor diminish its utility. Rather, they emphasize that much work on improving models lies ahead.

Another way to estimate the importance of model error is by assuming a "perfect ensemble" (Ziehmann 2000). The perfect ensemble assumption is an extension of the perfect model assumption, where the true evolution of the weather is assumed to always lie within the ensemble forecast distribution. This is achieved by using a randomly-chosen member of the ensemble as the verifying analysis, and any member of the ensemble is equally as likely to be the truth. Because ensemble and model bias are both removed, such an ensemble can easily be evaluated for its relation to IC error. Model error can be estimated by comparing the perfect ensemble with the ensemble verified against an independent analysis.
1.6 Goals of this research

The literature review above is not exhaustive, but it is a cross-section of many of the most important contributions in the field, and defines the paradigm within which most scientists are studying predictability and prediction. If nothing else, it leads to the conclusion that very little is understood.

This research focuses on regional NWP error characteristics over the North Pacific and western North America, and explores ways to account for the error. Short-range (0-3 days) prediction and predictability is examined for a specific event, largely (but not entirely) within the perfect model framework. More specifically, the goals of this research are threefold:

- Characterize model and IC uncertainty for an event affecting the North Pacific and western North America.

- Determine whether any useful information regarding IC error or uncertainty exists in operational analyses.

- Develop an ensemble prediction system that exploits any knowledge gained about the errors in short-range forecasting over the North Pacific and western North America, and test it on an independent data set.

The final goal is an evaluation of how the results may be generalized. None of these goals is independent, and the ensemble system is developed as a tool for understanding the error characteristics. The relationship between forecast error and the behavior of the ensemble system is a measure for the effectiveness of the ensemble system. At the same time, the ensemble system itself provides clues about important uncertainties in the ICs.

More details about construction of the ensembles are contained in Chapter 2. For the revealing case of a busted forecast (Chapter 3), NWP error and uncertainty are examined in Chapter 4. In Chapter 5, a description is provided of a method to generate IC ensembles, with perturbations based on differences between analyses produced by operational forecast centers. Both Chapter 4 and Chapter 5 discuss the relative contributions of model and IC error to the ensemble prediction of the period around the forecast bust. Chapter 6 compares the IC perturbation methods, and concludes with the recommendation of a new ensemble procedure. This is tested with independent data in Chapter 7, and overall conclusions and recommendations are summarized in Chapter 8. Acronym tables are in Appendix A.
Chapter 2

Experiment Environment

2.1 Statistical definitions

$I$ denotes the number of ensemble members, and the unweighted ensemble average is defined as:

\[ \bar{f} = \frac{1}{I} \sum_{i=1}^{I} f_i \]  

(2.1)

Here, $\bar{f}$ is the average of $f_i$ ensemble members, defined at each grid point.

For verification, the mean square error (mse) of the ensemble-mean forecast is simply

\[ mse = \frac{1}{NM} \sum_{n=1}^{N} \sum_{m=1}^{M} \left( \bar{f}_{nm} - a_{nm} \right)^2 \]  

(2.2)

where $n$ and $m$ are grid indices of an $N \times M$ grid, $\bar{f}_{nm}$ denotes the ensemble average at one grid point, and $a$ is the verifying analysis. The $mse$ is related to the error variance $E$ as

\[ E = mse - bias^2 \]  

(2.3)

where $bias$ is defined linearly, and is not to be confused with the bias score used in a distributed verification approach (Wilks 1995).

Under the perfect model and ensemble assumptions, $bias$ from the model is zero, and $E = mse$. To implement the perfect ensemble assumption, $a$ is one of the $f_i$ in
equation 2.1, chosen randomly. When verifying, the $f_i$ that is chosen is removed from the calculation of the ensemble mean.

If the ensemble is verified against an analysis that is not part of the ensemble, $bias \neq 0$, and the difference between $E$ and $mse$ measures it. Systematic differences between this verification and the perfect ensemble verification also indicate model error.

Ensemble spread is defined quadratically and grid point-wise as

$$S = \left[ \frac{1}{I} \sum_{i=1}^{I} (f_i - \bar{f})^2 \right]^{1/2}$$

which is just the standard deviation of the ensemble distribution. In an ensemble that is supposed to provide information on forecast confidence, $S \approx rmse$, where $rmse = (mse)^{1/2}$, in the perfect ensemble environment.

### 2.2 Model and domain

To perform ensemble experiments that simulate real NWP error, a realistic NWP model is the minimum requirement. It must satisfy the requirements of a perfect model, as outlined in section 1.5. Here, the experimental model and data are described, and the degree to which it realistically approximates the atmosphere is evaluated.

All the experiments reported here use 72 h forecasts with the Canadian Mesoscale Compressible Community Model (MC2). The MC2 is a LAM designed for mesoscale studies, but the dynamics are valid at nearly all scales, and many of the physical parameterization schemes are also used in the global models that the Canadian Meteorological Centre (CMC) uses operationally. In fact, its limitations are probably greater at finer scales because of limitations in the physical parameterization package. For all experiments, the model is run on a polar-stereographic grid projection with $128 \times 128$ horizontal grid points and 19 vertical levels. The horizontal grid spacing is $\Delta X = 100$ km, true at 60°N. Vertical levels are logarithmically-distributed and terrain following, from Gal-Chen and Sommerville (1975).

A thorough description of both the dynamics and the physics of the MC2 is given in Benoit et al. (1997). That paper was accurate for version 4.1 of the MC2, but a few changes have been made since, including using a more implicit time stepping scheme (Thomas et al. 1998). These experiments use version 4.8, and relevant information and example references are given in Tables 2.1 and 2.2. As described in section 2.5,
### Table 2.1: The MC2 model dynamics and discretization.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Approach</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Approximation</td>
<td>Hydrostatic, fully compressible</td>
<td>Benoit et al. 1997</td>
</tr>
<tr>
<td>Time-stepping</td>
<td>Semi-implicit</td>
<td>Thomas et al. 1998</td>
</tr>
<tr>
<td>Elliptic solver</td>
<td>Conjugate residual</td>
<td>Skamarock et al. 1997</td>
</tr>
<tr>
<td>Boundaries</td>
<td>“Sponge” zone</td>
<td>Yakimiw and Robert 1990</td>
</tr>
<tr>
<td>Horizontal diffusion</td>
<td>Fully implicit in time</td>
<td>Jakimow et al. 1992</td>
</tr>
</tbody>
</table>

### Table 2.2: The MC2 model physical parameterization schemes.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short-wave radiation</td>
<td>Fouquart and Bonnel 1980</td>
</tr>
<tr>
<td>Long-wave radiation</td>
<td>Garand 1983</td>
</tr>
<tr>
<td>Land surface</td>
<td>Deardorff 1978</td>
</tr>
<tr>
<td>PBL/turbulence</td>
<td>Benoit et al. 1989</td>
</tr>
<tr>
<td>Deep convection</td>
<td>Fritsch and Chappell 1980</td>
</tr>
<tr>
<td>Cloud microphysics</td>
<td>Sundqvist et al. 1989</td>
</tr>
<tr>
<td>Shallow convection</td>
<td>None</td>
</tr>
<tr>
<td>Gravity wave drag</td>
<td>None</td>
</tr>
</tbody>
</table>

The deep convection and cloud microphysics parameterization schemes are changed for one ensemble experiment that tries to approximate model error. The MC2 must be realistic to consider it in a perfect-model experiment. One way to test this assumption is to look at the energy spectrum of MC2 forecasts or simulations. Figure 2.1 shows a spectrum of 50.0 kPa geopotential height, calculated with a 2D FFT (Press et al. 1986). Because these experiments use a limited domain, the wavenumbers $k$ shown in Figure 2.1 are not true global wave numbers, and the computation was done on the polar-stereographic projection of the MC2. Thus, $k = 1$ corresponds to one wave across the domain. Still, the slope follows the $-3$ relationship that characterizes the global energy cascade at scales above the inertial subrange (Daley 1996). The peak at baroclinic wave scales is obscured by the meridional gradient ($k = 1$), but a bump can be seen near $k = 4$. A rough estimate with the value of $\Delta X = 100$ km places the bump at a wavelength of approximately 3000–4000 km, which does correspond to the variance peak at baroclinic wavelengths.
is persistent over all the cases presented in this work.

Both Medium-Range Forecast (MRF) model and Global Environmental Multiscale (GEM) model analyses and forecasts are used for initial and boundary conditions, and sometimes verification, depending on the experiment. The fact that boundary conditions need to be specified is one limitation of using the MC2 (or any LAM) for ensemble experiments. When ICs are perturbed, only the boundary conditions for the initialization time are perturbed. Throughout the rest of the integration, every ensemble member is prescribed the same boundary conditions. To minimize potential boundary contamination, the domain is chosen to be large, and verification is done on a smaller 50 x 50 gridpoint subdomain (Fig. 2.2).

The effects of the boundaries can be evaluated by looking at the average ensemble spread along the direction of the mean flow in the verification subdomain. Results with the perturbation strategy described in section 2.4.1 are taken as an example. Results of 50.0 kPa geopotential height spread averaged over all the forecast cycles in the experiment are shown in Figure 2.3, where the mean flow is from left to right in the plot.

At initialization (solid line), the longitudinal variation of mean spread is small, indicated by the flat curve. As the forecasts progresses, the spread at all longitudes increases, but the change is not constant across the domain. A peak in the spread develops near 140°W, and by forecast hour 72 the mean spread is approximately 1.5 times the spread upstream (left) at 180°. The values also drop rapidly toward
Figure 2.2: Complete MC2 domain with verification region for the MC2, MRF, and GEM models.
Figure 2.3: Zonal mean 50.0 kPa geopotential height spread versus longitude. Each curve represents a different forecast time (h), given in the legend. The mean is determined by averaging all values between 30° and 60°N.

the downstream (right) boundary. There is no persistent feature to which maximum spread is tied, such as a longwave trough, at 140°W that might maintain a spread maximum, and yet the characteristic shape of these curves is persistent through the experiment period.

Residence time of an air parcel in the domain is important to the growth of spread in the ensemble, as would be expected. An air parcel travelling along 45°N at an average zonal velocity of 50 m s\(^{-1}\) would travel approximately 165° of longitude in 72 h, or completely through the domain. Such a high average zonal speed is not often observed, but it illustrates the potential problems. Though upstream development allows spread to increase near the upstream boundary, the peak is well into the domain. The peak is not significantly greater than the values near the boundaries until the very end of the forecast period. If a longer forecast shows this continuing, it would suggest that the limit of spread has been reached upstream.

### 2.3 The W transform

The results of Whitaker and Loughe (1998) show that perfect-model errors in short-range forecasts are dominated by the statistical properties of IC error, while the errors in medium-range forecasts are dominated by nonlinear error growth. For medium-
range forecasts, methods that perturb along a singular vector (Palmer 1993), or along a bred vector (Toth and Kalnay 1993) are intended to exploit nonlinear error growth, and may not adequately sample the probability density function that describes all possible sets of initial conditions (Smith et al. 1999). While such perturbations lead to ensemble-average forecasts that are better than categorical forecasts for days 4 through 10, the ensemble-average skill can be worse for the 1-3 day short-range forecasts. Hence, singular-vector or bred-vector methods are not appropriate for short-range ensemble forecast (SREF) experiments. For SREF systems, correctly sampling the IC error is more important than identifying regions sensitive to the ICs, although both effects could operate synergistically to reduce forecast skill. This chapter describes a mathematical framework that is flexible and efficient enough to allow investigation of many aspects of IC error.

Perturbations are made in \( W \) space, constructed with a wavelet-like transform of analysis fields. Following Nychka et al. (1999), it is called the \( W \) transform. A 2D discrete transform first decomposes the field, providing spectral information from grid-point scales \((N/2)\Delta X\) to \(8\Delta X\), recursively stepping by a factor of 2:

\[
W = T_{8x8} \ldots T_{MxN/2} (\Psi_{MxN}) T_{MxN/2} \ldots T_{8x8}^t
\]  

(2.5)

Matrix \( W \) stores the transform coefficients, \( T \) contains both the scale and basis of the wavelet, \( \Psi \) is an analysis field, \( N = M = 128 \) is the dimension of the grid here, and \( t \) denotes transpose. The basis function and scale functions are the same as that used in Nychka et al. (1999) to model climatological precipitation distributions. The transform is similar to a discrete wavelet transform, where the basis function of Nychka et al. (1999) appears to pick out the appropriate features in the physical fields, including high and low pressure centers, regions of warm and cold, and wind direction and speed maxima and minima.

As an example, the structure of the \( 8 \times 8 \) transform operator is:

\[
T_{8x8} = \begin{bmatrix}
2 & 3 & -1 & 0 & 0 & 0 & 0 & 0 \\
0 & -1 & 3 & 3 & -1 & 0 & 0 & 0 \\
0 & 0 & 0 & -1 & 3 & 3 & -1 & 0 \\
0 & 0 & 0 & 0 & 0 & -1 & 3 & 2 \\
2 & -3 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & -1 & 3 & -3 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & -1 & 3 & -3 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & -1 & 3 & -2 \\
\end{bmatrix}
\]  

(2.6)

where the first four rows of \( T_{8x8} \) impose the scale and the last four rows impose the
basis. This transform is not orthogonal, so true spectral variance cannot be obtained.

To determine true spectral power an orthogonal wavelet transform must be used. In the work here, a Haar wavelet has been chosen. The Haar wavelet is more compact than the W transform above, so spectral information at scales as small as two grid points can be extracted with

$$T^{2\times2} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

(2.7)

where division by $\sqrt{2}$ ensures orthonormality.

Daley and Mayer (1986) provide a quantitative estimate of the analysis error spectrum finding a peak for some variables at wavenumbers coincident with baroclinic waves, indicating a maximum in analysis accuracy. But they suggest that no useful information exists at global wavenumbers greater than 30. Their results are from a 19-day period, and the spectrum is constructed by calculating variances over all the analyses of the period. It does not examine the daily variability of the analysis error, which leaves the possibility that the errors may exist at different wavenumbers for each analysis. Mullen and Baumhefner (1989) exploit the results of Daley and Mayer (1986), employing a white-noise perturbation strategy at wavenumbers below 30, and random phase errors above. They apply random perturbations with spectral characteristics that do not change for each case. While the overall magnitude of the error may be smaller today compared to 1986, the shape of the spectrum is likely to be similar, except that the white-noise portion should extend to higher wavenumbers because of improved observation and data-assimilation systems.

The transform in equation 2.7 yields a power spectrum in the same way as a 2D Fast Fourier Transform (FFT) does, with the total variance of the field the sum of the spectral powers across all resolved wavenumbers. An example of the power spectra determined with both the FFT and the W transform is shown in Figure 2.4. The discrete wavelet transform of equation 2.7 does not strictly or naturally correspond with a wavenumber in the same way as an FFT. The wavenumber $k$ of the W spectrum shown in Figure 2.4 is actually $2 \times \sqrt{k}$, with $k$ relative to the grid for the FFT calculation.

For the work presented here, whether the W transform precisely reproduces atmospheric spectral values is less important that the fact that it can represent a spectrum and provide additional spatial information. The experiments described below rely on matching spectral information, not on specific values. As long as the transform coefficients are consistent, the spectral information will be consistent. More comparisons between FFT and W transforms are made later.
Section 2.4 proposes three different strategies for using the W transform, each designed to approach a different possible aspect of initial condition uncertainty. Because of its flexibility, the W space is useful when generating an ensemble of ICs. Global or local features of the flow can easily be targeted. The performance of the ensembles indicates the success of each perturbation strategy.

2.4 Generation of IC ensembles using perturbations in W space

2.4.1 Ranked perturbations

A control analysis is assumed to be the mean of some IC probability distribution with an unknown shape. Perturbations are designed to sample from the distribution by perturbing different features across the analysis spectrum, where each perturbed analysis is assumed to be equally as likely. An attempt is made to account for the possibility that daily variability may exist in the scale and location of analysis error. These perturbation experiments use a model grid with $\Delta X = 100$ km, resolving waves that may be within the flat part of the error spectrum today. The perturbations target all resolvable scales similarly. The goal is to perturb different features at all scales in each ensemble member.
For each scale contained in \( W \), the coefficients are ranked by value from the largest to the smallest (largest negative). Each ranked coefficient is assigned to a quantile, where the number of desired ensemble members determines the quantile size. This experiment uses \( I = 8 \) perturbed ensemble members for each forecast (plus the categorical forecast), so \( I/2 = 4 \) quantiles are created (quartiles), each of which are used to generate a positive and negative perturbation.

Each ensemble member is generated by perturbing the coefficients in one quartile. Multiplying the coefficients in the first quartile by a number larger than one increases those coefficients, to create ensemble member one. Multiplying them by a number less than one decreases them, creating ensemble member two. Thus \( w_1' = w \times x \) and \( w_2' = w/x \), where the prime and index denote the perturbed wavelet coefficients for a particular ensemble member, and \( x \) is the number close to one that defines the magnitude of perturbation to the wavelet coefficients \( w \). This process is continued for the three other quartiles, resulting in eight perturbed members. Thus, the first two members of the ensemble contain perturbations to the strongest positive features at each scale in the analyses (e.g. ridges, warm pools, or U-wind maxima), and the last two contain perturbations to the strongest negative features at each scale (e.g. troughs, cold pools, or U-wind minima). The middle ensemble members contain perturbations where strong features are not present at a particular scale. An inverse transform builds the perturbed physical fields using the perturbed coefficients.

A constraint on the variance determines the magnitude of the perturbations. Perturbed fields are required to maintain a variance within 3% of the original field variance, a somewhat arbitrary choice ensuring that the perturbation size is within reasonable analysis error. A simple bisection algorithm iterates the transform-perturbation-inverse transform sequence described above until a value of \( x \) is found that satisfies the constraint.

The wavelet coefficients contain both amplitude and phase information. Perturbing the magnitude of the coefficients perturbs the amplitude of features in the analyses. The phase is not explicitly perturbed here, but after a few hours of forecast time, phase differences develop as baroclinic waves develop differently. Ranking the coefficients does not directly perturb different parts of the analysis spectrum, but the spectral location of strong or weak features in the analysis determine which scales are perturbed more for each ensemble member. Each perturbed analysis contains changes in the spectrum that are consistent across it.

This experiment, called PERT, uses the MRF model analyses, generated at NCEP, as a control. Ranked perturbations are made on each analyzed pressure level independently. After generating the perturbations, a dynamic initialization algorithm allows dissipation of gravity waves without generating irreversible phenomena such
as condensation and precipitation, forces agreement between the mass and momentum fields, and reduces the magnitude of the perturbations. Figure 2.5a shows an example of perturbed and unperturbed MRF 50.0 kPa geopotential height field after dynamic initialization. At this point, the \( \text{rms} \) of the perturbations is smaller than before the initialization. For example, the 50.0 kPa height \( \text{rms} \) ranges from zero to 30 m in Figure 2.5b.

The perturbation in this example slightly amplifies the dominant wave in the image. Transform coefficients are similarly ranked and perturbed through the depth of coherent atmospheric structures. This maintains the vertical structure of features in the analysis, and does little to alter the vertical error statistics. Mullen and Baumhefner (1989) use an extra correction step in their perturbation algorithm to ensure that the vertical error statistics are properly enforced. Vertical errors will differ between modeling systems, and the need to explicitly define them is a limitation. Such a step is unnecessary here, and knowledge of the vertical error structure is irrelevant.
2.4.2 Perturbations correlated with analysis differences

Despite the fact that major operational NWP centers have access to the same weather observations, large differences between analyses at synoptic times remain. These differences arise naturally within DA systems that use different subsets of available observations, different NWP models, and different error minimization algorithms.

In regions where in-situ observations are sparse, such as the North Pacific, DA systems rely heavily on satellite data to generate an analysis. The quality of an analysis depends on how these data are related to model background atmospheres. Because the data sets generated by satellite observations can be too large for current bandwidth and inversion by current computers, they are often thinned. How they are thinned will also affect the analysis.

As discussed in Chapter 1, evidence suggests that model error dominates in the first few hours of a forecast (Orrell et al. 2001). Different models contain different assumptions about the atmosphere, and thus contain attractors that are different from the true atmosphere and from each other. Even if two DA systems with different NWP models were started from the same analysis, different model errors would lead to differences in the next analysis. Over several cycles, the DA system is victim to model drift that causes it to diverge from the real atmosphere.

Every DA system is also subject to analysis errors arising from its error minimization algorithm. Those errors, combined with model errors, lead to forecast errors. DA systems minimize the analysis error subject to some constraint, but can never remove the error because of background and observational errors, and errors of representativeness (Daley 1996). Therefore, analyses also contain uncertainty.

An ensemble may be generated with perturbations based on the analysis differences to understand if the differences are important. Experiments with this strategy use MRF and GEM model analyses. If \( \Psi = \Psi_d = \Psi_M - \Psi_G \) in equation 2.5, where subscripts \( M \) and \( G \) indicate MRF and GEM analysis fields, a power spectrum of the difference between the MRF and GEM analyses can be constructed. Spectral information on the difference between analyses is used to generate two type of ensembles reflecting similar uncertainty spectra: (1) the difference field is randomly perturbed, then re-normalized according to the spectrum of \( \Psi_d \), and (2) purely random perturbations are normalized to have the same spectral characteristics as \( \Psi_d \). Case (1) will locally and globally reflect the differences, while case (2) will only globally contain similar information as \( \Psi_d \). The differences between the ensemble forecasts arising from the two perturbation strategies will measure the importance of spatial information in the analysis differences.

It is important to recognize that \( \Psi_d \) may not be representative of true analysis er-
ror. Its magnitude, statistical properties, and spectral properties may not be correct. The extent to which \( \Psi_d \) represents analysis error is discussed in chapter 5.

2.5 Generation of model-error ensembles using physics parameterization options

One way to simulate model error is by running the same dynamical core while changing subgrid-scale physical parameterization schemes for each run, generating a simple ensemble (Mullen and Baumhefner 1988, Stensrud et al. 2000, Hou et al. 2001). For these experiments, the perfect model assumption is obviously dropped, and the feedback between the different physical parameterization schemes and the common dynamical core generates spread between the ensemble members. The spread of this ensemble is not interpreted as an absolute measure of model error, but the day-to-day variation in spread indicates when model error might be most significant, just as the daily variability of IC-based ensembles indicates when IC error might be most significant.

Ensemble PHYS is created by using different microphysical and convective parameterization schemes for different ensemble members. Additional PHYS members are generated by including the MRF forecasts in the PHYS ensemble (PHYS+MRF), and including a run where the sub-grid parameterization schemes (i.e. microphysics, cumulus, turbulence, radiation, and gravity wave drag parameterizations) are turned off entirely (EXT). PHYS+MRF is perhaps the best measurement of model error, because it includes many other factors besides convective and microphysical parameterization schemes. EXT is an example of extreme model uncertainty — beyond what is realistic today — but is provided here as a benchmark.
Chapter 3

Description of a case study

3.1 Introduction

In early February 1999, residents of the Lower Mainland (Vancouver) area of south-western British Columbia (BC), Canada (Fig. 3.1), began preparing for an intense maritime cyclone that was forecast to bring heavy snow and rain, and strong winds to the area during the evening of 10 Feb., local time. On the morning of 10 Feb., the local forecast office issued both a wind warning, with winds forecast to reach 60 to 80 km/h later that day, and a snowfall warning with accumulations of 4 to 8 cm by evening.

Numerical model guidance indicated that the storm would continue into 11 Feb., leading the local forecast office to modify their afternoon forecast update issued on 10 Feb. to include wind and snowfall warnings, with winds increasing to 70 to 90 km h\(^{-1}\) and liquid-equivalent accumulations of 50 mm by 11 Feb. A low pressure center (LPC) was forecast to be a few hundred km off the central BC coast by the evening of 10 Feb., with an associated occluded front forecast to cross the Lower Mainland overnight.

The storm did not strike the Lower Mainland on the 10th or the 11th. Winds of 20 km h\(^{-1}\), and negligible precipitation, were recorded at Vancouver International Airport (CYVR) at 1600 PST (UTC = PST + 8 h) on 11 Feb. (Fig. 3.2). What made this storm remarkable was not just the forecast “bust” on the 10th and 11th, but also that subsequent numerical model runs initialized on 11 and 12 Feb. each produced revised guidance that brought the front with heavy precipitation and strong winds into the Lower Mainland quickly, and that each of these forecasts was also wrong. The region did not receive significant precipitation until after 0400 PST (12 UTC)
Figure 3.1: Map of southern British Columbia (BC), Canada, and northern Washington (WA), USA, showing Vancouver international airport (CYVR) in the Lower Mainland, Tahtsa Lake West, and Vancouver Island

Given the reliance on model guidance in modern operational forecasting, one can examine the role of the numerical models in contributing to the forecast bust. The 48 h storm tracks forecast by various operational and research numerical centers from 00 UTC 10 Feb. 1999 demonstrate a wide range of solutions for a single, strong LPC (Fig. 3.3). Only the Pacific Weather Centre (PWC; Pacific and Yukon Region, Meteorological Services of Canada) hand analyses show the cyclone splitting into two LPCs (Fig. 3.3). Hand analyses have the benefit of current satellite imagery and additional observations that may be outside model data assimilation cut-off times.

The storm was indeed as vigorous as forecast, but the forecast track and translation speed of the cyclone, and the frontal progression, were wrong. The front and strongest pressure gradient stalled and weakened near the north end of Vancouver Island for more than a day before crossing the Lower Mainland. The brunt of the storm was felt much further north along the west coast of BC. For example, 500 km north of Vancouver, the village of Tahtsa Lake West, in the coast mountains near the Alaska panhandle, recorded a BC one-day snowfall record of 145.0 cm on 11 Feb. But for the main urban and economic center of southwestern BC, the forecast was a repeated bust.
Figure 3.2: CYVR Precipitation and wind speed observations for the period 9-16 Feb. 1999. The frontal passage occurred near 12 UTC 13 Feb., and precipitation prior to that was convective air-mass. Precipitation values are 24 h accumulations from 12 UTC to 12 UTC around the 00 UTC date of each mark.
Figure 3.3: Analyzed (near 160°W longitude) and 48 h forecast (near 140°W longitude) central pressure (kPa) and location, forecast from 00 UTC 10 Feb. 1999. The Medium-Range Forecast (MRF) model, Global Environmental Multiscale (GEM) model, Mesoscale Compressible Community (MC2) model, and Pacific Weather Centre (PWC) hand analysis are shown. Two low-pressure centers (LPC1 and LPC2) are shown in the PWC analysis at 48 h, whereas the model forecasts only evolved one low-pressure center (LPC2).
The U.S. Weather Research Program (USWRP), Canadian Weather Research Program (CWRP), and the World Weather Research Program (WWRP), have recognized that data paucity over the mid-latitude oceans leads to numerical forecast degeneration downstream over continents. The lack of data exacerbates the underlying nonlinearity of the atmosphere, limits the range of predictability, and highlights sensitivity of the real atmosphere (and model atmosphere) to disturbances or errors in initial conditions (Lorenz 1963). Shapiro et al. (2001) noted unusually high NWP forecast error from 26 Jan.–10 Feb., 1999, and suggested that the errors may be related to ENSO. For this study, the forecast bust provides motivation to examine the causes of some NWP errors.

Three possible factors contributing to the failure of this forecast are: (1) initial condition (IC) error, (2) model errors for a particularly nonlinear or sensitive event, and (3) sympathetic data denial (when merchant ships and commercial aircraft, which make voluntary weather observations, avoid stormy areas). Determining the contribution of each factor to the bust can give us an idea of the state of NWP in the North Pacific and western North America, and how much improvement may be gained by improving observation networks in the North Pacific “data void”. Short-range ensemble forecasting (SREF) techniques that perturb ICs may provide insight to factor (1), for situations where the perturbation method adequately samples the range of possible ICs. If an estimate of model error is available, factor (2) can be investigated and the contribution to total forecast error can be separated (Mullen and Baumhefner 1989). Factor (3) is really one possible contribution to (1). It can be qualitatively investigated by comparing the amount and locations of in-situ data throughout the case study period, and by comparing forecasts made with different analyses of that sparse data. The next section describes the synoptic situation, and provides a qualitative assessment of analysis and forecast errors leading up to one critical forecast, initialized 00 UTC 10 Feb.

3.2 Storm description

This section describes the synoptic development of the storm (Hacker et al. 2001), and qualitatively compares model analyses and forecasts with satellite imagery and rawinsonde observations. GOES-derived winds and aircraft data (not shown here) were also used to verify the conclusions. It will be shown that easily-identifiable errors existed during this period. Observations over the Pacific will be shown to have a positive impact on subsequent analyses, as should be expected from the data assimilation systems. Model errors obviously play a role here, as any drift of the
model atmosphere away from the real atmosphere (or projection thereof) will carry through the data assimilation cycles, affecting future forecasts. Given the poor NWP guidance for this case, it is likely that the observations did not have enough of an effect on the data assimilation cycle to fully correct the model atmospheres that had drifted far from observations.

On 00 UTC 8 Feb. 1999, satellite imagery showed a leaf cloud in the central North Pacific, indicating a baroclinic disturbance there. Starting at roughly 00 UTC 09 Feb., the associated low pressured center (LPC1) deepened at a rate of 1.0 Bergerons [0.1 kPa h⁻¹ for 24 h adjusted for latitude (Sanders and Gyakum 1980)], and by 00 UTC 10 Feb. evolved into a mature cyclone near 45°N, 165°W, with a central pressure of 98.0 kPa (Fig. 3.4). During 10 Feb, a strong westerly jet streak of over 95 m s⁻¹ drove upper-level development to the southwest of the cyclone, suggesting ample zonal available potential energy (ZAPE). The upper-level flow split as a short-wave trough developed south of a persistent Omega block. Such blocking patterns have been linked with intense cyclogenesis (Bluestein 1993).

By 00 UTC 11 Feb., the upper-level wave had amplified considerably, with contributions from both a strong jet upstream and slow-moving continental air downstream. LPC1 had become vertically aligned with the upper-level trough by this time, but a new, active LPC2 had formed downstream near the frontal triple point at approximately 50°N, 140°W (Fig. 3.5). LPC2 had deepened 1.4 kPa to 97.0 kPa over 12 hours, and it deepened another 0.6 kPa to reach its minimum in the following 6 hours. Positive thickness and vorticity advection contributed to the rapid deepening. LPC1, with a slowly-evolving pressure of 97.2 kPa, was propagating east with the upper trough. Downstream, LPC2 was propagating rapidly northeast, causing cyclonic rotation of the two LPCs about each other. The BC coast, including Vancouver, experienced warm-frontal passage during the following 12 hours, but station CYVR reported no precipitation during this period.

By 00 UTC 12 Feb., the two co-existing LPCs reached meridional alignment, while the upper-level wave had further amplified (Fig. 3.6). The amplification coincided with a merger of a decaying Gulf of Alaska low with the developing trough as the Omega block weakened. The cold front had effectively stalled at the BC coast and was visibly weakening. Northward propagation of several vorticity maxima along the front, along with occlusion and orographic effects, may have contributed to the weakening. One such vorticity maximum can be seen near 51°N, 135°W in Fig. 3.6. According to surface observations, a high-pressure center (HPC) of 104.4 kPa had established itself further east, near 42°N, 114°W (roughly the corners of Utah, Idaho, and Nevada) with ridging ahead of the cyclone. The HPC was another factor that slowed eastward progression of the system.
Figure 3.4: GOES 10 visible image of the North Pacific, valid 00 UTC 10 Feb. 1999.
Figure 3.5: GOES 10 visible image of the North Pacific, valid 00 UTC 11 Feb. 1999.
Figure 3.6: GOES 10 visible image of the North Pacific, valid 23 UTC 11 Feb. 1999.
Figure 3.7: (a) GEM and (b) MRF 50.0 kPa geopotential height analyses and 24 h forecast errors (analysis—forecast), valid 00 UTC 10 Feb. 1999. Contour intervals are 60 m for the forecasts (thick solid contours) and 30 m for the errors (thin dashed contours). Rawinsonde observations are also plotted as station models with observed geopotential heights.

Analysis errors are difficult to quantify due to the paucity of observations over the North Pacific Ocean, but some deficiencies in MRF model and GEM model analyses, are obvious. First, the 00 UTC 09 Feb. 24 h forecast field, valid at 00 UTC 10 Feb., and the verifying analysis field are compared for both models. In both cases, the amplitude of the 50 kPa wave was greater in the analyses than in the forecast (Fig. 3.7). The differences were even larger at 25 kPa — up to 100 m geopotential height errors in the shortwave trough (not shown). Jet-stream winds were more meridional and extended further south in the analyses, suggesting a greater trough amplitude. Several data assimilation cycles occurred between the 00 UTC 09 Feb. analyses and the 00 UTC 10 Feb. analyses. The fact that analysis—forecast differences are greater than the differences between the rawinsonde observations and the current analysis suggest that the new observations positively influenced the 00 UTC 10 Feb. analyses, and partially corrected errant forecasts that served as background forecasts.

But significant errors (up to 40 m) at three isolated mid-North Pacific rawinsonde locations near 43°N, 155°W still exist in the 00 UTC 10 Feb. analyses (Fig. 3.7). If those observations were important in the analysis, even larger errors in surrounding regions devoid of observations may be expected. Those observations were used by the
data assimilation system at NCEP. Compared to these observations, the southerly wind component was under-analyzed by 8 m s\(^{-1}\) in the CMC analysis (Fig. 3.7a), and 6.5 m s\(^{-1}\) in the NCEP analysis (Fig. 3.7b). Westerly components were either comparable or too high in the analyses, which partially disqualifies the argument that underanalyzed wind magnitude was the cause.

Rather, this suggests an underanalysis of the 50 kPa wave amplitude by both 00 UTC 10 Feb. model analyses. GOES winds at 85.0 and 70.0 kPa also show that the trough was under-developed in both the analysis (not shown). Similarly, the downstream blocking ridge was underanalyzed in both model analyses. A CMC discussion of the analysis, issued on 10 Feb., notes 50 kPa geopotential height errors of 30-40 m in the ridge, a fact borne out by observations. The discussion also states that the jet stream over the Pacific was underestimated by 2.5-5 m s\(^{-1}\). Though anecdotal, this suggests that operational modelers were aware of potential errors in the numerical forecasts.

Large errors in the 48 h forecasts initialized at 00 UTC 10 Feb. are also apparent. The models moved the LPC2 too far east and neglected its northward movement. Because of this, the models predicted a lower surface pressure and a stronger surface pressure gradient at CYVR than was observed. The erroneous southeast displacement of the forecast LPC2 may have been related to the under-amplification of the upper-level steering flow.

The upper-level ridge was positioned over western North America after 48 hours. Geopotential height errors (model—rawinsonde) at rawinsonde locations in the ridge were on the order of −50 m at 50.0 kPa in the MRF and GEM models (Fig. 3.8), and −80 m 25.0 kPa. These errors suggest under-amplification of the forecast upper-level wave. Maximum errors were worse than −80 m at 50 kPa and −120 m at 25 kPa, and were located just west of the forecast ridge axis, suggesting that the ridge was forecast to propagate too rapidly. These errors might be linked to the two analysis errors discussed above: under-analyzed wave amplitude and weak meridional jet-stream speeds, both of which help determine upper-level development.

Near-surface differences between 48 h forecasts from 00 UTC 10 Feb., and 00 UTC 12 Feb. analyses show similar patterns to the 50.0 kPa forecast errors. Namely, positive forecast errors northwest of the modeled HPC, and negative errors west-northwest of the forecast LPC2, indicate that both were situated too far southeast in the forecasts.

The speed of the front was over-forecast as a result of under-amplification aloft, misrepresentation of rapidly propagating vorticity maxima, and perhaps poorly-handled interaction with coastal orography. Since frontal movement is best correlated with low-level/surface winds normal to the front in the cold air (Bluestein 1993), the
development of high-amplitude patterns generally corresponds to slower eastward frontal movement as southwest winds become southerly and southeasterly. Propagating mesoscale vorticity maxima aloft can induce negative surface-pressure perturbations along the front, which in turn create cyclonic mesoscale circulations that inhibit frontal movement (called frontal waves). The poor handling of frontal interaction with orography might be influenced by the coarse resolution of topography in the NWP models, which lessens the impact of topographic blocking.

Regardless of the cause, the actual front was considerably delayed and weakened, compared with the forecast, when it finally made its way inland (Fig. 3.9). Significant weather did not strike until 13 Feb., and it was a fraction of what had been expected over the previous three days. Frontal passage on 13 Feb. brought a 50° veering of surface wind, a slight pressure drop to 101.5 kPa, and a total of 4.4 mm of precipitation for the 24 h period ending 12 UTC 14 Feb. The surface temperature drop associated with the cold air mass behind the front occurred in the early hours of Feb. 14, and may have been delayed by interactions with orography and the formation of a trapped marine boundary layer.

In summary, persistent under-amplification of the geopotential height pattern is evident at all levels in both analyses and forecasts. GEM and MRF 00 UTC 10
Figure 3.9: GOES 10 visible image of the North Pacific, valid 00 UTC 13 Feb. 1999.
Feb. analysis errors were significant, particularly with respect to meridional jet wind speeds and height field amplitude in the North Pacific. After 48 h, both models had under-forecast the amplitude of the height field considerably. This led to overly zonally-progressive surface features in the models. Observations tended to push the model states toward reality, but the pattern was never adequately amplified. In the data assimilation cycle, model error, a lack of observations, and observations that were not given enough weight by the error-minimization algorithms, probably all contributed to the forecast errors. A qualitative assessment such as this suggests that both analysis and model errors are to blame, but choosing a dominating factor is nearly impossible without a quantitative approach, as described next.
Chapter 4

NWP Error and Uncertainty
Around a North Pacific Bust

4.1 Experiment design

To study the contributions of NWP analysis and model error for the case described in Chapter 3, different ensembles are run with the MC2 model for each forecast period. Experiment PERT uses ranked perturbations (section 2.4.1) to the initial conditions, experiment PHYS uses different physical parameterization schemes (section 2.5), experiment PHYS+MRF includes the MRF forecast, and experiment DENY uses identical runs with the MC2 started from the different NCEP and CMC analyses. Categorical (unperturbed) ICs and common boundary conditions for the ensemble experiments (except DENY) are taken from MRF forecasts. The difference between the MRF runs and the MC2 categorical runs results primarily from model differences, including boundary-condition implementation, with an expected small contribution from interpolation to the MC2 run-time grid.

The PHYS experiment contains differences from model error alone, because no perturbations are introduced to the ICs in this experiment. When perturbations are introduced (PERT), IC differences appear. The perfect model assumption is used to isolate the roles of IC and model uncertainty. Statistical verification against GEM analyses valid every 6 h for the period 5–15 Feb. helps to demonstrate the spread-error characteristics of the PERT experiment.

To study analysis errors such as could occur due to sympathetic data denial, two analyses produced independently by CMC and NCEP are used as initial conditions for the MC2 forecast runs. While these analyses had access to similar observation
data, the different government centers set different data cutoff times, and use different quality control systems, first-guess analysis fields, and error-minimization algorithms, all of which lead to different weights when assimilating the observations.

The results of Whitaker and Loughe (1998) show that if the probability distribution of ICs is Gaussian-shaped, and the ICs are adequately sampled in an ensemble system, the spread and error should be correlated. That is to say that when an ensemble forecast displays a large amount of spread, the ensemble-mean forecast is likely to be less accurate. Likewise, a small amount of spread indicates a more accurate forecast. A forecast with an average amount of spread, compared to a climatological spread, will yield little information on forecast confidence because it will not be related to improved or diminished accuracy. Climatological spread is the spread of a static ensemble system averaged over many cases.

Many factors, including model physics and an IC distribution, contribute to the relationship between spread and error. A positive spread-error correlation suggests that an ensemble accounts for them. This implies that among other things, the distribution of ICs is adequately sampled. The actual shape of the initial pdf is unknown, but a correlated spread and error suggests that it is partially sampled by the ensemble of perturbed ICs. This approach is valid during the short-range, when error growth is presumed to be nearly linear.

One way to measure the spread-error correlation is to construct a contingency table of ranked spread and error. Both spread and error are calculated at every verifying grid point at every forecast time over each 72 h forecast. They are ranked, and divided into quintiles. The frequency of both spread and error falling in the same quintile is an estimate of the joint spread-error probability. Results can be normalized by the total number of spread-error pairs in each quintile, to give a percentage of verification points where the spread and error fall into the same quintile. If there were no correlation between spread and error, each value in the table would be 0.2. A contingency table that indicated perfect spread-error correlation would have ones in the diagonal and zeroes everywhere else.

If ensemble PERT shows some spread-error correlation, suggesting it samples some of the IC pdf, it can be used to evaluate IC error compared to model error. If significant IC error is suggested, it is worthwhile to search for causes, including possible sympathetic data denial. Also, further ensemble verification can be carried out with a perfect ensemble framework. When ensemble bias is removed, comparison is simpler and can be used to support or refute any conclusions drawn from other comparisons of PERT and PHYS.

The next section discusses the spread-error correlation of ensemble PERT. Section 4.2.2 discusses the relative contributions of model and IC error, and section 4.2.3
qualitatively looks at the ensemble-forecast storm tracks. Later, section 4.3 uses aircraft data and ensemble DENY to assess sympathetic data denial, and section 4.4 constructs a perfect ensemble verification as another check on the results.

4.2 IC and model error with ranked perturbations

4.2.1 Spread-error correlation

A contingency table verifying 50.0 kPa geopotential height (Table 4.1), shows numbers greater than 0.2 along the diagonal, indicating a positive correlation between spread and error. This correlation improves for the extreme values in the first and last quintiles. That is, very large spread indicates very large error, and vice versa. This correlation also grows at longer forecast lead times, suggesting that more confidence information may be gained later in the forecast. For example, at 72 h a spread value ranked in the top quintile is three times more likely to show error in the top quintile than in the bottom quintile. In the middle quintiles, the diagonal numbers are only marginally greater than the surrounding numbers, indicating that near-median spread does not provide much information about forecast confidence. These results, using real forecasts, are similar to those obtained by Whitaker and Loughe (1998) for an idealized ensemble.

The spread-error correlation can also be seen by examining individual forecasts. Fig. 4.1 shows correlation coefficients between forecasts and the verifying analysis for the forecasts initialized 00 UTC on 8 and 10 Feb. 1999. They are chosen because they provide a contrast between one forecast characterized by high spread (8 Feb., Fig. 4.1a) and one characterized by low spread (10 Feb., Fig. 4.1b). The correlation coefficient is a generous measure of skill (Murphy 1988), but it is effective at showing the differences between the forecasts. Comparing Fig. 4.1a and b, one finds that all of the curves drop off faster in Fig. 4.1a, indicating greater error (less skill) at 72 h for the 8 Feb. forecast, and they also spread out more. Thus, high spread is associated with high error.

The correlation between ensemble spread and ensemble error suggests that this perturbation method is approximating at least part of the IC pdf. Its behavior is not perfect, and this sample of eight forecast cycles is too small to construct robust statistics. But within the “climatology” of these particular eight days, we can use the ensemble spread as an indicator of IC uncertainty by determining the forecast error variances. Error variance can be normalized by an estimate of the climatological variance and averaged over all verification locations. Mullen and Baumhefner
Table 4.1: Contingency tables of ranked forecast spread (rows) and ranked error (columns) for 00, 24, 48, and 72 hour 50.0 kPa geopotential height forecasts over the entire experiment period. Each element in the table is based on the average of about 4000 values.

<table>
<thead>
<tr>
<th></th>
<th>0-20%</th>
<th>20-40%</th>
<th>40-60%</th>
<th>60-80%</th>
<th>80-100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>00h</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0-20%</td>
<td>0.26</td>
<td>0.20</td>
<td>0.18</td>
<td>0.22</td>
<td>0.14</td>
</tr>
<tr>
<td>20-40%</td>
<td>0.22</td>
<td>0.21</td>
<td>0.22</td>
<td>0.19</td>
<td>0.15</td>
</tr>
<tr>
<td>40-60%</td>
<td>0.19</td>
<td>0.22</td>
<td>0.22</td>
<td>0.19</td>
<td>0.19</td>
</tr>
<tr>
<td>60-80%</td>
<td>0.18</td>
<td>0.19</td>
<td>0.19</td>
<td>0.20</td>
<td>0.24</td>
</tr>
<tr>
<td>80-100%</td>
<td>0.15</td>
<td>0.19</td>
<td>0.19</td>
<td>0.20</td>
<td>0.28</td>
</tr>
<tr>
<td>24h</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0-20%</td>
<td>0.38</td>
<td>0.24</td>
<td>0.19</td>
<td>0.12</td>
<td>0.07</td>
</tr>
<tr>
<td>20-40%</td>
<td>0.20</td>
<td>0.23</td>
<td>0.19</td>
<td>0.19</td>
<td>0.18</td>
</tr>
<tr>
<td>40-60%</td>
<td>0.17</td>
<td>0.19</td>
<td>0.20</td>
<td>0.20</td>
<td>0.24</td>
</tr>
<tr>
<td>60-80%</td>
<td>0.14</td>
<td>0.18</td>
<td>0.21</td>
<td>0.22</td>
<td>0.25</td>
</tr>
<tr>
<td>80-100%</td>
<td>0.11</td>
<td>0.16</td>
<td>0.20</td>
<td>0.27</td>
<td>0.26</td>
</tr>
<tr>
<td>48h</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0-20%</td>
<td>0.32</td>
<td>0.32</td>
<td>0.22</td>
<td>0.09</td>
<td>0.05</td>
</tr>
<tr>
<td>20-40%</td>
<td>0.25</td>
<td>0.22</td>
<td>0.22</td>
<td>0.18</td>
<td>0.13</td>
</tr>
<tr>
<td>40-60%</td>
<td>0.16</td>
<td>0.16</td>
<td>0.19</td>
<td>0.24</td>
<td>0.26</td>
</tr>
<tr>
<td>60-80%</td>
<td>0.13</td>
<td>0.14</td>
<td>0.18</td>
<td>0.25</td>
<td>0.29</td>
</tr>
<tr>
<td>80-100%</td>
<td>0.14</td>
<td>0.16</td>
<td>0.18</td>
<td>0.24</td>
<td>0.27</td>
</tr>
<tr>
<td>72h</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0-20%</td>
<td>0.37</td>
<td>0.32</td>
<td>0.19</td>
<td>0.08</td>
<td>0.04</td>
</tr>
<tr>
<td>20-40%</td>
<td>0.21</td>
<td>0.25</td>
<td>0.27</td>
<td>0.18</td>
<td>0.08</td>
</tr>
<tr>
<td>40-60%</td>
<td>0.17</td>
<td>0.19</td>
<td>0.23</td>
<td>0.21</td>
<td>0.20</td>
</tr>
<tr>
<td>60-80%</td>
<td>0.13</td>
<td>0.13</td>
<td>0.17</td>
<td>0.24</td>
<td>0.33</td>
</tr>
<tr>
<td>80-100%</td>
<td>0.12</td>
<td>0.11</td>
<td>0.13</td>
<td>0.29</td>
<td>0.35</td>
</tr>
</tbody>
</table>
Figure 4.1: 50.0 kPa geopotential height correlation coefficients $r$ for forecasts (a) 8, and (b) 10 Feb. 1999. The categorical (triangles), ensemble mean (circles), and every ensemble member forecast are shown (other eight lines).

(1989) argue that a comparison of normalized forecast error variance yields information about the importance of IC uncertainty versus model uncertainty. Suppose that IC uncertainty is approximated by the ensembles discussed above (PERT), and that model uncertainty can be approximated by a physics-based ensemble (PHYS). We can invoke the perfect model assumption and include exclusively IC or model uncertainty for each ensemble.

4.2.2 Model and IC error

Normalized error variances averaged over the entire period are shown in Fig. 4.2a and c. The PERT error variance decreases between forecast hour 0 and 12 during a period of further dynamic adjustment. Later, it shows an error growth rate that is greater than the PHYS error growth rate. The error values (Fig. 4.2a) are also much higher than those for PHYS. The PHYS+MRF error growth rate is larger than the PHYS throughout the forecast period, the error growth rates of PERT and PHYS+MRF are nearly identical at forecast hour 72.

The day-to-day variability in error variances determines a distribution. In this context, the differences between error variance curves are significant if the standard deviation of error variances is less than the difference between them (Tribbia...
and Baumhefner 1988). The PERT error variance is significantly greater than the PHYS+MRF error variance until the last few hours of the forecast period, and the difference between the PHYS and PHYS+MRF curves is significant after about 40 hours. If the PERT initialization is a realistic representation of analysis error, and the MC2 is a realistic representation of the atmosphere, one can not claim that the differences are an artifact of the ensemble strategy.

The results suggest that through the first 66 h of model forecasts during the entire eight-day period, initial condition uncertainty is more important than model uncertainty as measured by forecasts with both the MC2 and the MRF. In the last 6 hours, where the difference is not significant with respect to the standard deviation of PERT, one cannot reliably say that the initial condition uncertainty is greater. These results are valid only for this particular synoptic regime and period, within the ability of the MC2 to realistically reproduce atmospheric phenomena.

The IC uncertainty and model uncertainty for the single forecast initialized 00 UTC 10 Feb. are shown in Fig. 4.2b and d. This was one of the forecasts that provided poor short-range guidance to forecasters while the storm was developing offshore. The PERT curve shows lower error variances than the results from the entire period (Fig. 4.2a), and the lowest of any one forecast day, reflecting the fact that this forecast demonstrated small spread (Fig. 4.1b). The growth rate is also smaller (Fig. 4.2d). The PHYS and PHYS+MRF curves for 10 Feb. demonstrate the greatest values and growth rate of any one forecast day, dominating by hour 54 with values much greater than the period averages.

Viewed this way, it appears that model uncertainty is more important than IC uncertainty for the forecast initialized 00 UTC 10 Feb. That is, the uncertainty represented by the PERT experiment for this one day is much smaller than the average from the entire period, while the opposite is true for PHYS and PHYS+MRF. Considering the results of the previous section and of Whitaker and Loughe (1998), a forecaster would have erroneously placed high confidence in the forecast if the models were perfect because of the small amount of spread between ensemble members. But the large errors in PHYS and PHYS+MRF suggest this is actually not the case, illustrating a dilemma for operational forecasters using ensemble forecasts, which often include both runs with different ICs and different models.

The large reduction in PERT error variance over the first 12 hours suggests that the perturbations did not project on dynamically-sensitive modes for this forecast. Because the perturbations were selected with the same algorithm each day, and they are not truly random, it is likely that the verification sub-domain was not subject to dynamically-sensitive modes, despite the fact that the LPCs were inside the sub-domain at initialization. This could occur if the storm had already reached maturity,
Figure 4.2: Normalized 50.0 kPa geopotential height error variance (a) calculated for the entire period, and (b) for the forecast initialized 00 UTC 10 Feb. 1999. Growth rates $\frac{dE}{dt}$ with units m (6 h)$^{-1}$ are shown in (c) for the entire period, and (d) for the 10 Feb forecast. Results from all the experiments except DENY are shown (see text for explanation of experiments).
Figure 4.3: Normalized 50.0 kPa geopotential height error variance of each ensemble forecast at 24, 48, and 72 h. (a) shows results from experiment PERT, and (b) shows results from experiment PHYS+MRF. The date on the abcissa refers to the date of each initialization, at 00 UTC.

...and baroclinic instability was already diminished.

Looking at the daily variability of error variances throughout the experiment period also supports this conclusion (Fig. 4.3). At 72 h, the IC error is by far the greatest for the forecast initialized 00 UTC 8 Feb., while the model error is the greatest for the forecast initialized 00 UTC 10 Feb. Error growth for PERT is fairly uniform through 48 h, with a slight reduction in uncertainty later in the period as error variance values decrease. The spike on 8 Feb. does not become obvious until 72 h. A particularly sensitive feature dominating the verification domain at this time may also contribute, but the lack of a spike at 48 h for the forecast initialized 9 Feb. weakens that argument. Experiment PHYS displays a tendency for relatively high error variances as early as 24 h for the forecast initialized 00 UTC 10 Feb. For completeness, we also note that Fig. 4.2 and Fig. 4.3 suggest error doubling times of approximately 1.5–2 days through the 72 h period of the forecasts.

4.2.3 Qualitative aspects

Qualitatively, the IC and model uncertainty can be seen by tracking the low-pressure center through the forecast from 00 UTC 10 Feb. Fig. 4.4a shows the PERT forecast...
Figure 4.4: (a) PERT ensemble and (b) PHYS ensemble storm tracks and 48 h central pressures from the forecast initialized 00 UTC 10 Feb. 1999. The MRF forecast is also shown in (b) as open circles.

and Fig. 4.4b shows the PHYS forecast for the same case. Some tracks are close enough that they overlap for much of the period. The location and depth of the low varies much more in the PERT ensemble than in the PHYS ensemble. But when the MRF forecast is included, the spread at 48 h covers a similar geographic distance. Viewing only the cyclone-center location errors in Fig. 4.4, the differences between the PERT and PHYS experiments are not as visible as when many locations in the verification sub-domain are included, but the errors look similar. At 48 h in Fig. 4.2b, the errors are similar as well. One factor limiting the spread of the PHYS ensemble is the common PBL scheme and other diffusion properties. Another experiment would have to be run to determine the impact of those parameterizations.

4.3 Sympathetic data denial

When major cyclones are observed or forecast, ships and aircraft often avoid those regions. As a result, any automatic and manual weather observations made from those ships and aircraft do not sample the cyclone, where they are most needed to better predict cyclone evolution. This undesired loss of critical observations is known as “sympathetic data denial”, and can cause positive feedback in analysis and forecast error growth near the cyclone.
Although the behavior of the forecast initialized 00 UTC 10 Feb. is noticeably different from forecasts for the remaining days of the experimental period for both PERT and PHYS+MRF, the DENY results from that day are similar to the whole experiment period (Fig. 4.5). If DENY is an appropriate estimate of data denial, a comparison between error growth of the forecast initialized 10 Feb. and the period-mean error growth shows that particular forecast was not unusual. This assumption has several weaknesses because many other factors affect the analyses, including first-guess fields and algorithms for error minimization in the data assimilation systems. Systematic errors caused by minimization algorithms would tend to reduce daily variability in the spread, because it should mostly determine scale response and total error. Substantial day-to-day variability in spread and spread growth did exist in DENY (Fig. 4.5b), suggesting that these factors combine to produce significant differences between analyses. But the growth for the forecast initialized 10 Feb. was nearly average. Only part of the period is shown in Fig. 4.5b for comparison with Fig. 4.6 below.

A look at NCEP observations that were ingested by the data assimilation system does not provide any further insight. Rawinsonde and ship-borne observations over the ocean are too sparse to effectively determine whether changes in the network
were coincidental or the result of intentional circumnavigation of any developing storm system. Aircraft data are a bit more dense, and NCEP quality control flags give information about whether an observation was appropriately used. A plot of the number of temperature observations from aircraft, flying at altitudes near the tropopause, within the box bounded by 35° to 60°N latitude, and 120° to 160°W longitude, shows substantial variation in observations (Fig. 4.6). Observations were deemed "good" when the data assimilation system did not explicitly reduce the weights or throw the data out. Within the period shown, three intervals having very few observations are evident — all between 06 and 12 UTC — in the middle of the night local time. Presumably few aircraft are flying during these hours on any day.

Comparing Fig. 4.6 with Fig. 4.5b does not suggest any relationship between aircraft observations and the differences that result in experiment DENY. Relatively large 24 h spread for the forecast initialized 00 UTC 11 Feb. does not correspond with a period of few aircraft observations. The same is true for the large 72-h spread for the forecast initialized 00 UTC 9 Feb. Thus either experiment DENY is not an appropriate surrogate for data denial experiments, and/or variation in aircraft observations near the tropopause does not have a major impact on data assimilation systems. The latter is likely for three reasons: (1) increased reliance on satellite observations over the ocean, (2) the fact that most aircraft observations are made at a cruising altitude near the tropopause, and (3) a data assimilation system would not spread them over a large area without several concurrent observations. Thus, whether or not DENY is a reasonable surrogate for sympathetic data denial is unknown. Results of experiment DENY and an examination of observations used by the data assimilation system at NCEP are inconclusive.

4.4 Comparison with perfect ensemble

The perfect ensemble assumption is invoked by verifying the ensemble against one member of the ensemble. Often, this is accomplished by randomly choosing one verification member for each comparison (Ziehmann 2000). With only eight forecast periods, a random distribution would be highly unlikely if chosen independently. Thus, the perfect ensemble verification is carried out once for every member of the ensemble, and the results are averaged. When a particular member is chosen, it is removed from the calculation of the ensemble mean and the spread. This approach effectively increases the number of cases by a factor of \( I \), the number of ensemble members, though they are not independent and therefore do not actually improve the statistical robustness. Rather, it simply ensures a random distribution of verifying
Figure 4.6: Number of aircraft observations within the box bounded by 35° to 60°N latitude, and 120° to 160°W longitude. The data assimilation system at NCEP did not modify or throw the observations deemed “good”.

Because $bias = 0$ in a perfect ensemble, $E = mse$. In an ensemble that successfully predicts error, $E \approx S^2$. Figure 4.7a shows that when averaged over all forecasts, the evolution of $E$ and $S^2$ is similar, but the ensemble is over-dispersive. When verifying against the GEM analyses, the perfect ensemble assumption is dropped, and the $bias$ of the ensemble mean indicates model error (Fig. 4.7a). The average $bias^2$ is less than $E$ and $S$ for the perfect model. Thus, over the eight verifying forecasts, $E$ from the IC uncertainty makes a larger contribution to the total forecast error than $bias^2$; namely, the error variance is greater than systematic error. Linear spread-error correlation coefficients are better for the perfect ensemble, as expected (Fig. 4.7)b.

To examine the relative importance of model and IC error, Figure 4.8 shows the average error variances determined for the forecast initialized 10 Feb. By the end of the 72 h forecast period, the error variance for the 10 Feb. forecast is about half that of the period mean in the perfect-ensemble configuration (similar to Fig. 4.2). But when the GEM analyses are used for verification, $E$ for 10 Feb. actually surpasses the period mean, showing again that model error may have been unusually large for that forecast. Note that these values of $E$ have not been normalized by the climatological variance, as in Figure 4.2, which are really estimates of spread.

Model $bias$ and error variance can be separated, and the $bias$ can be removed from a forecast. The daily variability of the error terms provides evidence that an unusually
Figure 4.7: Comparison of perfect ensemble configuration and an ensemble verified against the GEM analysis. (a) shows the evolution of the perfect ensemble error variance $E$ (solid) and spread $S^2$ (long dash). The $bias^2$ is determined against GEM. (b) compares $S - E$ correlation coefficients ($r$) from both verifications. All units are m$^2$.

Figure 4.8: Comparison of the error variance $E$ averaged over the whole period with that for the forecast initialized 10 Feb: (a) results from the perfect configuration, and (b) results when the ensemble is verified against GEM analyses. All units are m$^2$. 

55
large contribution from model bias helped cause forecast failure on 10 Feb. Figure 4.9 compares the variation of several error parameters over the study period. The first panel (a) shows $E$ in the perfect model configuration. The forecast initialized 7 Feb. displays much greater values than any of the other days, and combined with 8 Feb. is largely responsible for the mean error variances being so much greater than those on 10 Feb. But when the ensemble is verified against GEM analyses (Fig. 4.9b), the forecast initialized 8 Feb. contains the greatest error at 72 h. Panel (b) may be compared to Figure 4.3b. A different order of averaging accounts for some of the differences, while a slightly different verification subdomain accounts for the rest\(^1\). The gross features are still preserved. The difference between panels (a) and (b) result from model error, and the relative increase in $E$ is actually greater for 10 Feb. than it is for 7 or 8 Feb. Panel (c), showing $bias^2$, also shows anomalously high model bias on 10 Feb. Panel (d) is the sum of (b) and (c), showing that despite unusually large bias, $E$ still dominates the total forecast error.

The results of this section support the results of section 4.2, which concluded that for the forecast initialized on 10 Feb., model error was unusually large. Experiment PHYS captures the anomalously high model error on 10 Feb., but the magnitude is underestimated, and the smaller peaks in Figures 4.3b and 4.9c do not match. Perfect ensemble experiments suggest that model bias was an unusually large contributor to total model error on 10 Feb.

\(^1\)The verification subdomain was slightly changed to better match up with the computational domain and avoid any interpotation errors that may be present in previous sections. The results and conclusions are not affected.
Figure 4.9: Daily variability of error measures for the (a) perfect ensemble configuration, and (b), (c), and (d) verification against GEM analyses valid at 24, 48, and 72 h forecasts. (a) and (b) are plots of $E$, (c) shows bias$^2$, and (d) shows the mse. The curves in (b) and (c) sum to the curves in (d). All units are m$^2$. 
Chapter 5
Perturbations correlated with analysis differences

Major operational numerical weather prediction (NWP) centers have access to similar weather observations, but large differences between analyses at synoptic times remain. These differences arise naturally within data assimilation (DA) systems that use different subsets of available observations, different NWP models, and different error minimization algorithms.

In regions where in-situ observations are sparse, such as the North Pacific, DA systems rely heavily on satellite data to generate an analysis. Analysis quality in these regions depends on how satellite data are related to model background atmospheres. Because the data sets generated by satellite observations can be too large for current communication bandwidth and fast inversion by current computers, these data are often thinned. How they are thinned will also affect the analysis.

Evidence suggests that model error dominates in the first few hours of a forecast (Orrell et al. 2001). Different models contain different assumptions about the atmosphere, and thus have attractors that are different from the true atmosphere and from each other. Even if two DA systems with different NWP models were started from the same analysis, different model errors would lead to differences in the next analysis. Over several cycles, the DA system is victim to model drift that causes it to diverge from the real atmosphere.

Every DA system is also subject to analysis errors arising from its error minimization algorithm. DA systems minimize the analysis error subject to some constraint, but can never totally remove the error because of background and observational errors, and errors of representativeness (Daley 1996).

Because analysis error cannot be perfectly defined without the unattainable true
state of the atmosphere, analysis uncertainty is a tractable replacement for error. The uncertainty contains both known and unknown errors.

One way to examine the effect of uncertainty in an analysis is with ensemble forecasting techniques. Some Monte Carlo ensemble techniques perturb a categorical analysis with noise representative of estimated analysis error characteristics (e.g. Errico and Baumhefner 1987, Mullen and Baumhefner 1989 and 1994). The ensemble Kalman filter (EnKF) approach perturbs observations within the DA system, and uses an ensemble of forecasts to propagate error statistics through the next cycle, thereby targeting uncertainty specific to that day (Evensen 1994, Burgers et al. 1998, Houtekamer and Mitchell 1998). Instead of looking explicitly for uncertainty, dynamic approaches, such as the singular vector (Molteni and Palmer 1993, Mureau et al. 1993, Buizza 1995) or bred vector (Toth and Kalnay 1993) schemes, perturb the most sensitive structures in the atmosphere, whether or not uncertainty exists there. The justification for these approaches is that nearly all forecast error will project upon the growing modes early in a forecast period (Mureau et al. 1993, Toth and Kalnay 1993), which are propagated in the first few eigenvectors of the system.

Monte Carlo and EnKF systems account for analysis uncertainty by introducing perturbations that simulate estimated realistic analysis errors. Within a perfect model environment, a large forecast spread suggests that analysis uncertainty has a large effect on the accuracy of a forecast. Given two analyses from different operational NWP centers, it seems intuitive that in regions where they differ, uncertainty may exist. Error may be large in one and small in the other, or both may have large error. Conversely, uncertainty in an analysis may be low where the analyses agree.

This idea is supported by studies showing improved skill with ensembles generated simply by combining forecasts from several operational centers (Ziehmann 2000) or research centers (Hou et al. 2001). Both of these studies found that a small ensemble of categorical forecasts showed desirable spread and ensemble-mean forecast properties. Because those categorical forecasts use different analyses and different forecast models, the relative contribution of the analysis uncertainty and model uncertainty is unknown. It remains to determine how much of the spread in multi-center ensembles may result from the analysis differences, versus how much is the result of model differences.

This part of the study addresses the first part of the problem – namely, are operational analysis differences a good representative of analysis uncertainty, and will they generate realistic spread in an ensemble? An ensemble based on operational analysis differences can be constructed by perturbing with structures that are representative of those differences. If the ensemble is integrated with one (perfect) model, ensemble spread is an indicator of uncertainty accounted for by analysis differences. To ap-
proach this, differences between the Canadian Meteorological Centre (CMC) Global Environmental Multiscale (GEM), and the National Centers for Environmental Prediction (NCEP) Medium Range Forecast (MRF) analyses are exploited to determine whether those differences are representative of analysis uncertainty that generates forecast uncertainty.

5.1 Experiment and analyses

5.1.1 Ensemble generation

Each DA system produces an analysis considered "optimal" by its own error minimization procedures. Though it is impossible, each analysis is considered equally as likely to be true because the best one depends on specific selection criteria. Different model errors and treatment of observations cause differences in the analyses that may contain some information about analysis uncertainty. Error is not the same as uncertainty, as an uncertain component in an analysis or forecast may be fortuitously close to the truth.

Perturbations are constructed as described in Section 2.4.2. After perturbation, each forecast is dynamically initialized to remove fast modes, slightly reducing the spread. Examples of 50.0 kPa spreads for DIFF-MRF and RND-MRF after initialization, valid 00 UTC 5 Feb. 1999, are shown in Figure 5.1. In each panel, the dashed lines show the difference between the control initialization from MRF and GEM analyses, or \( \Psi_d \). For this example, the correlation coefficients between the spread and the difference fields, averaged over all levels, are approximately 0.8 for DIFF-MRF and 0.05 for RND-MRF. In the correlated case (panel a), the difference and spread contours can cross because of the random component of the perturbations.

As an estimate of model error, ensemble PHYS+MRF is created by using different microphysical and convective parameterization schemes for different ensemble members. It also includes the MRF run from NCEP. Sensitivity to these physics has been shown in other studies (Mullen and Baumhefner 1988, Stensrud et al. 2000, Hou et al. 2001). All of these ensemble members are initialized with the same MRF analysis for each forecast period. Because model error is difficult to quantify, it is unknown whether ensemble PHYS+MRF is a realistic representation of model error. But the perfect-ensemble analysis of section 4.4 suggests that it behaves like model error.

Each ensemble includes a forecast from a categorical analysis, and eight perturbed members. Ensemble performance statistics are calculated for the 50.0 kPa geopotential height and compared for the four ensembles.
Figure 5.1: Comparison of 50.0 kPa geopotential height perturbations for ensembles (a) DIFF-MRF and (b) RND-MRF, valid 00 UTC 5 Feb. 1999. All contour intervals are 20 m. Dashed lines show MRF—GEM differences and solid lines show initial spread.

Skill of the ensemble mean is measured with $rmse = \sqrt{mse}$, where $mse$ (mean-squared error) is the sum of the error variance $E$ and the bias$^2$. Depending on what is being evaluated, the MRF analyses, GEM analysis, or available observations, may be the reference for verification.

### 5.1.2 MRF and GEM analyses

A map of the MRF-analyzed 00 UTC period-averaged 50.0 kPa geopotential height is shown in Figure 5.2. The trough over the North Pacific is flat because several short-waves propagated through the domain during the study period. Figure 5.2 also shows the average absolute difference between the 00 UTC MRF and GEM analyses. Over North America, the differences are small, but over the Pacific differences reach 20 m on average. Rawinsonde observations are virtually nonexistent over the Pacific, and the differences demonstrate the different methods of combining satellite data with model backgrounds.

A difference field $\Psi_d$ can be useful only if $\Psi_G$ and $\Psi_M$ contain a similar distribution of variance. Figure 5.3 shows both MRF (a) and GEM (b) 2D geopotential height
Figure 5.2: Period-averaged 00 UTC 50.0 kPa geopotential heights at 120 m intervals (solid lines), and period averaged absolute MRF—GEM differences at 10 m intervals (dashed lines).

Table 5.1: Total period-averaged 00 UTC analysis geopotential height variances (m$^2$).

<table>
<thead>
<tr>
<th>Level (kPa)</th>
<th>MRF</th>
<th>GEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>100.0</td>
<td>$4.28 \times 10^3$</td>
<td>$5.05 \times 10^3$</td>
</tr>
<tr>
<td>85.0</td>
<td>$1.12 \times 10^4$</td>
<td>$1.13 \times 10^4$</td>
</tr>
<tr>
<td>50.0</td>
<td>$9.07 \times 10^4$</td>
<td>$9.00 \times 10^4$</td>
</tr>
<tr>
<td>25.0</td>
<td>$2.80 \times 10^5$</td>
<td>$2.74 \times 10^5$</td>
</tr>
</tbody>
</table>

FFT spectra, averaged over all 00 UTC analyses from 5–12 Feb. 1999. They represent "climatological" spectra for the period studied here. Each curve is normalized by its total variance, which are shown in Table 5.1.

The greatest difference between the MRF and GEM analysis variances are near the surface at 100.0 kPa. During this period, the GEM analyses contained more low-level variance than the MRF analyses. Comparing the curves in Figure 5.3a and b, the 100.0 kPa GEM variance is slightly higher than the 100.0 kPa MRF curve, relative to its own variance aloft. This is most noticeable at higher wavenumber $k$.

The wavenumbers $k$ in Figure 5.3 are grid-relative, and are not true global wave numbers. But the spectra display the expected $-3$ relationship for the large scale,
Figure 5.3: Period-averaged 2D FFT spectra of 00 UTC analyzed geopotential heights (Z) at 100.0, 85.0, 50.0, and 25.0 kPa. Panel (a) shows MRF results, and panel (b) shows GEM results. The curves are normalized by the variances shown in Table 5.1. The wavenumbers \( k \) are grid-relative.

Except for the differences at low levels, the MRF and GEM geopotential height spectra are similar through most of the troposphere at all scales. Calculating \( \Psi_d \) will not introduce spurious differences arising from features that may be represented in one analysis but not the other, and thus \( \Psi_d \) is representative of true analysis differences.

MRF \( (\Psi_M) \) and MRF–GEM \( (\Psi_d) \) geopotential height spectra are shown in Figure 5.4. The difference field is closest to saturating the analysis variance in the lower troposphere (panel a). Higher in the troposphere, the differences are smaller relative to the analysis, and the wave number where saturation is closest gets smaller. Unlike the error estimates of Daley and Mayer (1986), \( \text{var}(\Psi_d) \) is not greater than the analyzed \( \text{var}(\Psi_M) \). But the shape of the spectrum is similar to Daley and Mayer (1986). The \( \text{var}(\Psi_d) \) curves in Figure 5.4 tend to flatten at lower wave numbers, suggesting the characteristics of the analysis differences is representative of analysis error even if the magnitude is not. Still, either analysis errors have reduced substantially since 1986, or \( \Psi_d \) is not a realistic estimate of analysis error.

Comparing the initialization 50.0 kPa ensemble mean geopotential height with rawinsonde observations confirms that \( \Psi_d \) does not account for all of the analysis
Figure 5.4: Period-averaged 00 UTC MRF–GEM analysis geopotential height 2D FFT spectra at (a) 100.0, (b) 85.0, (c) 50.0, and (d) 25.0 kPa. The wavenumbers $k$ are grid-relative.
An ensemble system based on $\Psi_d$ will not include enough uncertainty to account for realistic analysis error. Keeping this in mind, the extent to which $\Psi_d$ accounts for forecast error can still be estimated by verifying ensemble forecasts.

Figure 5.5 also shows that the true state of the atmosphere likely lies outside the envelope defined by operational analyses. This might result from inadequate NWP models and/or DA systems, or large observational error. Whatever the reason, two potential consequences are: (1) only pure luck or an NWP model with an unrealistic trajectory will cause a categorical model forecast to follow the trajectory of the real atmosphere; and (2) an ensemble that approximates analysis error is likely to contain the truth near the edge of the probability distribution, rather than in the middle, and any model bias will cause the ensemble to drift away from the truth.

These consequences imply that it is necessary to include an estimate of model uncertainty in any ensemble approach, even when realistic perturbations are introduced. Mitchell and Houtekamer (2000) pointed this out in the context of their ensemble Kalman filter. If a model is considered perfect in an ensemble prediction system, the ensemble spread must be unrealistically inflated to accommodate the truth.
5.2 Perfect ensemble verification

An ensemble forecast that is representative of IC uncertainty will eventually project onto growing modes if it is going to be useful in forecasting. Furthermore, it should project on modes that arise from analysis uncertainty, and not just on modes that are guaranteed to grow. Because analysis uncertainty is not necessarily correlated with growing modes, the perturbations do not need to be designed to project on growing modes. Rather, perturbations that are representative of analysis uncertainty should organize to affect modes with uncertain growth. In this section, the extent to which analysis differences project onto growing modes is assessed in a perfect-ensemble environment.

A perfect ensemble assumes that the true evolution of the atmosphere lies within the phase space bounded by the ensemble. Any one member of the ensemble can represent the verifying analysis. Over a large sample size, the verification may be chosen randomly (e.g. Ziehmann 2000). Because of the small ensembles and small sample size in these experiments, each member is chosen for verification once, ensuring a random-like distribution of verification statistics. When a member serves as verification, it is removed from all performance calculations. Each verification is not independent, so this procedure does not improve statistical robustness.

For the perfect configuration, bias = 0, making the error variance equal to the mean squared error (E = mse). Also, S is related to mse by nature of the verification. A comparison of perfect configurations of DIFF-MRF and RND-MRF is shown in Figure 5.6a, and Figure 5.6b compares DIFF-GEM and RND-GEM. The relationship between the rmse and S is clear, with the spread appearing slightly too small compared to the error, from initialization through the entire forecast period. This suggests that the ensembles are under-dispersive.

Because the perturbations of each ensemble are statistically similar, and are controlled by $\Psi_d$, the differences in initial rmse and S are determined by the dynamic initialization of the MC2. The RND ensembles show smaller S and rmse compared to the DIFF ensembles. Both the MRF and GEM analyses should be well-balanced with respect to their set of model equations. Perturbed fields based on the differences between them might have a better chance of being balanced compared to random perturbations. The dynamic initialization procedure removes some of the variance introduced by the unbalanced random perturbations. Thus S and rmse of the RND are reduced at initialization by about 20%. If the resulting perturbations are relevant to the forecast error, the growth rates of the DIFF and RND ensemble should still be similar.

The DIFF ensembles show faster growth in S than the RND ensembles. Ensemble
RND-MRF shows little $S$ growth until 48 h into the forecast, while the spread in DIFF-MRF grows from 12 h. The DIFF-GEM and RND-GEM behave similarly, with the DIFF-GEM growth rates slightly greater than the DIFF-MRF growth, and the RND ensemble spread growing almost the same for each control.

Because the initial differences are small and the dynamic initialization removes spurious modes, the difference in behavior results mostly from the perturbation strategies. Figure 5.6 shows that the DIFF ensembles project onto growing modes more than RND ensembles. Analysis differences are clearly not random, and may contain information that can be exploited in ensemble generation. The next section verifies the ensembles against separate analyses to determine whether the analysis differences can produce the proper ensemble spread — spread that accounts for forecast error.

## 5.3 Verification against analyses

Figure 5.5 shows that analysis differences are not as large as the $rms e$ of the analyses when compared with observations, but Figure 5.6 shows that perturbations that simulate analysis differences will grow in a forecast. The growth may be too slow to account for forecast error. This section verifies the ensembles against GEM and MRF analyses to compare forecast uncertainty resulting from $\Psi_d$ with ensemble-mean error.
Ensembles constructed around the MRF ensemble are verified against GEM analyses, and vice versa. Because the ensemble mean at initialization is closely related to the categorical forecast around which perturbations were applied (with differences arising from the dynamic initialization), the \( rmse \) at \( t = 0 \) should be nearly equal to the ensemble spread.

Figure 5.7 shows that initial spreads of the DIFF ensembles approximates initial ensemble mean error, but the growth in spread does not follow the growth in error. While the verifying analyses will follow a phase-space orbit that lies in the vicinity of the observations, an ensemble that does not adequately take into account analysis error at initialization (among other uncertainty sources) will not. Figure 5.5 shows that \( \Psi_d \) does is not large enough to accommodate analysis error, and thus the ensemble cannot be expected to envelop the truth.

A larger ensemble can be created by combining all the runs in DIFF-MRF and DIFF-GEM, called FULL. A two-member ensemble consisting of the two control runs initialized with unperturbed MRF and GEM analyses (called CNTL) can be generated for comparison. Figure 5.8a compares \( S \) for ensembles DIFF-MRF, DIFF-GEM, FULL, and CNTL. Any bias is easily removed from the \( rmse \), or similarly, added to \( S \) for comparison. \( S + bias \) can be considered a corrected spread. Figure 5.8b shows the results when the ensemble-mean bias is added to \( S \) for ensembles FULL and CNTL. The \( rmse \) and bias values used in Figure 5.8 are calculated with GEM.
verifying analyses. Verification against MRF analysis results in rmse values that are nearly identical, and slightly higher bias values. This suggests that the ensemble mean of FULL does not lie exactly in between the GEM and MRF analyses, and is a reflection of weak nonlinearity in the error-growth curves.

First, adding bias to the $S$ curves does not bring them up to the level of the ensemble mean error. Thus, the corrected $S$ from ensemble FULL cannot account for the total forecast error. Second, the rmse of CNTL and FULL is nearly indistinguishable, showing that the perturbations around $\Psi_d$ do not improve the ensemble mean forecast.

By combining DIFF-MRF and DIFF-GEM into FULL, the initial corrected $S$ is twice as large as the initial rmse, but the corrected $S$ for CNTL is almost identical to the initial rmse. This is a consequence of the perturbation strategy. The initial rmse of an ensemble of only the control forecasts is by design equal to the difference between the MRF and GEM control analyses.

Figure 5.8a shows that the initial rmse of the DIFF ensembles is slightly higher than rms(MRF--GEM) in Figure 5.5, but well below the rmse of the analyses compared to observations. When the bias is added (Fig. 5.8b), $S$ of ensemble FULL is much larger, but still below the analysis error shown in Figure 5.5.
Though the bias increases during the forecast period, its growth is nearly identical for the FULL and CNTL ensembles. Adding it simply shifts the FULL $S$ and CNTL $S$ curves upward, increasing their growth by the same amounts. The FULL $S$ curve does not display any increase in growth rate compared to the CNTL $S$ curve, suggesting that error growth is nearly linear through the 72 h period. This result does not conflict with the results of Tribbia and Baumhefner (1988), Houtekamer and Derome (1995), and Ziehmann (2000), for example. But the increase in $S$ does lead to improved spread-error correlations (Fig. 5.9).

As in the perfect ensemble verification in section 5.2, these results show that ensembles DIFF-MRF and DIFF-GEM are slightly under-dispersive relative to ensemble-mean error. Given that the initial spreads are less than analysis error, and the error growth is nearly linear over the 72 h forecast period, this is not a surprise. Thus, when operational analysis differences are the basis for an ensemble, model uncertainty must be included to approach an ensemble spread that is commensurate with ensemble-mean error. To check this, spreads from ensemble PHYS+MRF are added to the FULL spread curve in Figure 5.8b. The combined growth in $S$ closely follows the error growth after about 36 h (Fig. 5.10).

Table 5.2 shows the contributions from FULL, PHYS+MRF, and bias to the $S$ curve in Figure 5.10. Through the entire 72 h period, FULL $S$ accounts for more than half of the total corrected spread. As the model uncertainty grows (PHYS+MRF $S$), the relative contribution of FULL $S$ decreases because its growth is more linear.
Figure 5.10: Period-averaged 50.0 kPa geopotential height $S$ and $rmse$. $S$ is a combination of the FULL $S$, its $bias$, and the PHYS+MRF $S$.

Table 5.2: Contributions to the total period-averaged 50.0 kPa geopotential height spread (m) shown in Figure 5.10.

<table>
<thead>
<tr>
<th>Forecast h</th>
<th>FULL $S$</th>
<th>PHYS+MRF $S$</th>
<th>$bias$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>14.43</td>
<td>1.15</td>
<td>5.11</td>
</tr>
<tr>
<td>24</td>
<td>15.25</td>
<td>3.75</td>
<td>6.09</td>
</tr>
<tr>
<td>48</td>
<td>22.02</td>
<td>7.83</td>
<td>7.08</td>
</tr>
<tr>
<td>72</td>
<td>27.16</td>
<td>12.78</td>
<td>7.94</td>
</tr>
</tbody>
</table>
5.4 Conclusions

Other studies have shown that ensembles consisting of categorical runs from different operational or research centers can perform desirably (Ziehmann 2000, Hou et al. 2001). This study uses five different ensembles to estimate the importance of different operational analyses in generating an effective ensemble. Perturbations are made about MRF and GEM analyses, with statistics and spatial structure that are similar to the difference between the analyses. For comparison, random perturbations with similar statistics and an ensemble designed to estimate model uncertainty are also run. The ensembles are integrated for eight cases surrounding a storm event that was notable for its poor NWP guidance.

A comparison of the \( \text{rmse} \) of the MRF and GEM analyses with the \( \text{rms} \) differences between the MRF and GEM analyses show that analysis error is much larger than the differences throughout the entire resolved spectrum. Differences are larger over the Pacific where rawinsonde data are sparse, compared to over North America.

Ensembles generated with perturbations that are spatially correlated to analysis differences show greater spread than ensembles with similar statistics but no spatial correlation. This result suggests that some useful information regarding analysis uncertainty can be extracted from analysis differences. This result is satisfying because one expects analyses to agree more where in-situ observations are denser. It also confirms that the Pacific “data void” is detrimental to forecast skill over North America, and it exists despite the wealth of satellite information.

The ensemble-mean error growth is greater than the growth in spread for all of the ensembles generated with perturbations. Both the ensemble-mean error and the spread grow linearly over the 72 h forecast period. When spreads are corrected for ensemble bias and to account for model uncertainty, the spread grows at the same rate as the ensemble-mean error.

These results suggest that spread induced by perturbations representative of analysis differences can account for most of the spread shown in multi-center ensemble studies such as Ziehmann 2000 and Hou et al. (2001), at least for the first 72 h of forecast time. For an ensemble with the desirable property of spread equal to ensemble mean error, the spread must be increased, and model error and uncertainty can accomplish this. These estimates are consistent with nearly-linear error growth from analysis differences that are smaller than the observed analysis error.

The results here should be verified against a larger, independent data set, and preferably with a global model. Boundaries likely limit the spread of the ensembles. If true, model uncertainty may be less important than these results suggest.

As models will never be perfect, the only way to accommodate the truth in an
ensemble is to perfectly account for both initial condition and model error. A paradox arises when the errors are not adequately known. For example, when model errors are large or a perfect model is assumed, the ensemble spread must be unrealistically large to include the truth as a realistic member. This is an undesirable characteristic of an ensemble, and can only be overcome by further study of the relative contributions of model and initial condition uncertainty, as is discussed in the next chapter.
Chapter 6

Comparison of ensemble methods

A more thorough evaluation of the different perturbation methods used in Chapters 4 and 5 (ranked perturbations and perturbations correlated with analysis differences) will elucidate the advantages and disadvantages of each. If one method shows an overall advantage, it will be verified in another case study with independent data.

Chapter 4 compared normalized error variances of experiments PERT and PHYS+ MRF to determine the contributions of model and IC error for one particular forecast, initialized 10 Feb. 1999, relative to the surrounding days. The relative importance of model bias was also addressed by comparing perfect ensembles with ensembles verified against analyses. It was found that on the forecast initialized 10 Feb., the model error was unusually large.

Chapter 5 compared perfect ensembles in experiments DIFF-MRF and DIFF-GEM with random perturbations RND-MRF and RND-GEM to determine the relationship between operational analysis differences and analysis uncertainty. Because the DIFF experiments showed more desirable results, the perfect ensemble configuration of the DIFF experiments was compared with the DIFF ensembles verified against analyses. It was found that model uncertainty must be accounted for when constructing perturbations based on analysis differences.

For generalization of either method, a longer period of analysis error would have to be examined. The daily analysis errors are not considered in PERT, and long-term averages should be used to scale the perturbations. The relationship between the actual analysis errors and $\Psi_d$ over a longer period is unknown, and using it to control the magnitude of the perturbations may be ill-advised if $\Psi_d$ ever exceeds analysis error. Chapter 5 established that the spatial distribution of the analysis differences can provide useful information for an ensemble forecast system, but that the magnitude of the perturbations is too small compared to analysis errors during
the experimental period. How it relates over a long period is a subject for further study.

Experiments PERT and PHYS were initially completed with an older version of the MC2 model (version 4.6). Experiments DIFF-MRF and DIFF-GEM were completed with a later version (version 4.8). Also, as noted in Section 4.4, the verification domain changed slightly. Thus for direct comparison, PERT and PHYS were re-run with version 4.8 of the MC2, and the verification was done on a common sub-domain of the computational domain. Finally, because of the further adjustment noted on the ensembles in Section 4.2, the period of dynamic initialization was increased.

6.1 Perfect ensemble comparison

6.1.1 Error and spread

Comparing rmse and $S$ of experiments PERT, DIFF-MRF, DIFF-GEM, and FULL in the perfect configuration reveals some different characteristics of each ensemble (Fig. 6.1). One obvious difference is at initialization time, when $S$ and $\text{rmse}$ of experiment PERT are less than the other experiments. The initial perturbation magnitude is arbitrarily chosen in PERT, while the MRF—GEM analysis differences control the perturbations in the DIFF experiments. Figure 5.5 showed that the analysis differences (and thus $S$ for the DIFF experiments) are smaller than actual analysis errors when verified against rawinsonde observations. This is a simplistic approach and only looks at local analysis error for a few days, but it suggests that the perturbations in experiment PERT could be more than twice as large and still be less than analysis error.

How $S$ and $\text{rmse}$ are related to each other is another obvious difference between experiment PERT and the rest (Fig. 6.1). For PERT, $S > \text{rmse}$, but $\text{rmse} > S$ for all the other experiments. Ensembles DIFF-MRF and DIFF-GEM appear under-dispersive in the perfect configuration, and ensemble PERT appears over-dispersive. Ensemble FULL comes the closest to $S = \text{rmse}$, a desirable characteristic in an ensemble that is to be used for predicting error.

Finally, the updated experiment PERT does not appear to undergo any reduction in $S$ in the first 12 h, as the rest of the experiments do. This behavior is likely related to how well the perturbations are balanced relative to the equations in the MC2. Generally, PERT strengthens and weakens physical structures in the analysis. The perturbations are only balanced in the sense that a positive temperature perturbation will coexist with a positive geopotential height perturbation. Any wind maxima...
Figure 6.1: Comparison of experiments (a) PERT, (b) DIFF-MRF, (c) DIFF-GEM, and (d) FULL 50.0 kPa geopotential height $S$ and $rmse$ in the perfect ensemble configuration.
resulting from the associated pressure gradient will also be increased. No effort was made to explicitly balance the perturbation, and the magnitude of temperature increase only relates linearly to the magnitude of geopotential height increase because the same constraints are imposed on the field variances. The same is true of any relationship between pressure gradient and wind speed.

It was suggested in Chapter 5 that perturbations based on $\Psi_d$ might be balanced as long as $\Psi_M$ and $\Psi_G$ are balanced. Figures 6.1b, c, and d show the spread of the DIFF ensembles reducing over the first 12 h, suggesting some adjustment of imbalances in the perturbations. $S$ in experiment CNTL (Fig. 5.8) does not decrease over the first 12 h. Thus $\Psi_d$ by itself may be balanced. The remaining members of the DIFF ensembles are generated by randomly perturbing $\Psi_d$. This is an unexpected result, but it implies that the comparison between DIFF and RND ensembles in Chapter 5 is valid. After the dynamic initialization procedure, both groups of forecasts may have imbalances, so the DIFF ensembles do not have an unfair advantage.

Time derivatives of the $S$ curves show how the growth rates evolve over the 72 h forecast period (Fig. 6.2). Values at each time represent the change in $S$ over the previous 12 hours. As noted on Figure 6.1, ensemble PERT is the only one that maintains a positive growth rate throughout the forecast period. But from 24 – 60 h, the growth rate is smaller than for the DIFF ensembles. Of the three DIFF ensembles, FULL shows the highest growth rate throughout the forecast. Figure 6.2 also highlights the fact that the spread in DIFF-GEM consistently grows faster than in DIFF-MRF, despite the smaller initial $S$.

The perturbations in both DIFF-GEM and DIFF-MRF are based on $\Psi_d$, and the dynamic initialization is responsible for the initial DIFF-GEM spread being smaller. If the perturbations are partially representative of analysis uncertainty, as they appear to be (Section 5.2), then small errors in the GEM analysis will grow more rapidly than in the MRF analysis. This suggests that the MRF analysis is better during this period. Though a small sample, this is supported by the slightly higher GEM $r m s e$ against rawinsonde observations (Fig. 5.5).

The results of this section favor experiments PERT and FULL in the perfect ensemble configuration. Though over-dispersive, PERT appears better-balanced and shows positive spread growth through the entire 72 h forecast. FULL shows the best agreement between $r m s e$ and $S$.

### 6.1.2 Spread-error correlation

The correlations between error and spread ($E$ and $S$ here) are smallest for experiment PERT (Fig. 6.3), but it is still growing at 72 h. Given that larger spread should
Figure 6.2: Spread growth rates \([\text{m} \ (12 \ \text{h})^{-1}]\) for experiments PERT, DIFF-MRF, DIFF-GEM, and FULL 50.0 kPa geopotential height in the perfect ensemble configuration.

result in better correlations (Table 4.1 and Whitaker and Loughe 1998), this is not a surprise. With its smaller initial spread, DIFF-GEM also begins with a smaller \(S-E\) correlation than DIFF-MRF. By 72 h, the correlation is greater than DIFF-MRF, as is the spread. The pattern does not hold true for FULL, which finishes with the smallest correlations of any DIFF experiment despite its large spread.

6.2 Verification against analyses

6.2.1 Ensemble mean errors

Up to now, the performance of the ensemble mean has been ignored for one reason: it is arguably not the best measure of ensemble performance. It is true that an ensemble mean forecast should be better than a categorical forecast, but it is often better only because an ensemble mean effectively filters smaller, less predictable scales from the verification.

In these experiments, the ensemble-mean 72 h forecast systematically shows less variance than the categorical forecast between domain wavenumbers 3 and 9. Differences are on the order of 1%, which may be negligible, but they are persistent and similar in all of the ensembles. Averaging has no affect at \(k = 2\), and \(k > 9\) may
contain mostly noise. When comparing $rmse$ of an ensemble mean and a categorical forecast, the categorical forecast will have more variance at active scales that include baroclinic storm systems.

Nevertheless, a common goal of ensemble forecasting is to produce an ensemble mean that is superior to any single forecast. Theoretically, this is only guaranteed with an infinite ensemble size ($I = \infty$), unless an additional regression step is applied to the ensemble mean (Leith 1974). With smaller ensembles, verification is required.

Figure 6.4 shows $rmse$ of the categorical forecast and ensemble-mean forecast, averaged over the experiment period. The verification is against GEM analyses, which causes the initialization error of DIFF-MRF to be greater than that of DIFF-GEM (panels b and c). In both cases, the difference between the forecast categorical and ensemble mean error cannot be seen until the end of the forecast period, when both ensemble mean forecasts are slightly better than the categorical forecasts. Ensemble FULL shows the greatest advantage for the ensemble mean, while ensemble PERT actually shows a reduction in skill through the entire forecast.

Figure 6.5 shows results of verification against rawinsonde observations, rather than analyses. Model forecasts are bilinearly interpolated to rawinsonde locations for the $rmse$ calculations. For most verification times (every 12 h), about 20 observations were available within the verification sub-domain. The skill is approximately 25% worse than shown in Figure 6.4 at initialization, but the difference is smaller at 72 h. The higher skill shown by the ensemble-mean FULL forecast (Fig. 6.4d) nearly
Figure 6.4: Comparison of experiments (a) PERT, (b) DIFF-MRF, (c) DIFF-GEM, and (d) FULL 50.0 kPa geopotential height rmse verified against GEM analyses. Thin lines show categorical forecasts and thick lines show ensemble-mean forecasts.
disappears. The disadvantage of the ensemble mean of PERT no longer stands out. Truncation, the DA cycle, and the small number of verifying observations combine to cause the skill shown in Figure 6.4 to decay more smoothly than it does in Figure 6.5. First, GEM analyses are interpolated to the verification sub-domain before the skill is calculated, removing small-scale features that will affect a single observation. This error of representativeness can contribute to the wiggles in the curves in Figure 6.4. Second, DA systems have an inherent consistency because they are built around a forecast model. Third, verification against rawinsonde observations typically included approximately 20 verification points spread irregularly, but verification against the GEM analyses included $50 \times 50 = 2500$ verification points, spread evenly across the grid. Figure 6.4 is statistically more robust, but each verification pair (analysis and forecast at one grid point) is perhaps more spatially correlated than the observations used in Figure 6.5. The importance of the difference decreases later in the forecast as the error grows large enough to dominate.

This simple look at ensemble mean error and categorical forecast error is inconclusive. Only ensemble FULL shows an improvement in skill, and verification against observations throws that into question. Ensemble-mean PERT forecasts show lower skill than categorical forecasts when compared with analyses, but the reduction nearly disappears when compared with observations. All of the differences between the ensemble mean and the categorical forecast skill are small compared to the actual error. Thus, the best ensemble method must be judged another way.

### 6.2.2 Spread-error correlation

When $S-E$ correlations are computed against GEM and MRF analyses, the values are smaller than the perfect ensemble configuration, as expected (Fig. 6.6 can be compared to Fig. 6.3). Large differences between the 0 h coefficients can be explained by the perturbation strategy. Only experiment PERT shows an increasing correlation coefficient at 72 h.

Because of the perturbation strategy of DIFF, DIFF-MRF has the largest positive correlation at initialization in panel (a), but DIFF-GEM and FULL both show negative correlations when compared to GEM analyses. An MRF analysis, with perturbations correlated to MRF-GEM differences, will show spread in the same regions where the MRF and GEM analyses do not agree. Perturbations to a GEM analysis are unrelated to the actual analysis error at any given location. The opposite scenario occurs when the verification is against MRF analyses in panel (b). Thus, the initial correlations are dependent upon the verification chosen, illustrating a problem with all verification against analyses. Ensemble FULL shows correlations that are in
Figure 6.5: Comparison of experiments (a) PERT, (b) DIFF-MRF, (c) DIFF-GEM, and (d) FULL 50.0 kPa geopotential height rmse verified against rawinsonde observations. Thin lines show categorical forecasts and thick lines show ensemble-mean forecasts.
between DIFF-GEM and DIFF-MRF at 0 h.

Correlations for the DIFF ensembles peak at or before 48 h, but the PERT correlation continues to grow, just as in the perfect configuration. Because both DIFF-MRF and PERT use the MRF analysis as a distribution mean for perturbation, comparing these suggests that the ranked perturbations result in a better evolution of spread-error correlation. DIFF-MRF shows the lowest correlation at the end of the forecast period in both cases. PERT is better even though it consistently shows smaller spread.

Ensemble FULL shows correlations that lie between DIFF-MRF and DIFF-GEM at all times. Except in the first 24 h, the FULL curves look similar in both Figure 6.6a and b. Similar to PERT, its performance appears more resistant to the particular verifying analysis chosen.

### 6.2.3 Spectral dependency

Another verification consideration is the scale dependency of spread and error. Chapter 5 concluded that the DIFF experiments do not fully account for forecast error, and section 6.1.1 showed that the PERT spread was even smaller. Those sections looked at total error and spread. In this section, the totals are broken down spectrally. Again, errors are determined by comparison to GEM analyses.

Figure 6.7a shows spectra for experiment PERT. It is calculated by performing an
FFT on the spatially-variable spread determined with equation 2.4. At initialization, the S curve is nearly parallel to the rmse curve, indicating that ranking the perturbations generates spread that captures the overall shape of the error spectrum. By targeting strong and weak features equally, the initial spread does not show the kink near the baroclinic wavelength (3-4000 km) at $k = 4$. Figure 6.7a also shows that the ranked perturbations could be nearly twice the magnitude of PERT and still compare to analysis error, which is greater than the initial rmse shown here (Fig. 5.5). At 72 h (panel b), the spread does not account for the forecast error, but the shape is similar at $k < 10$. The ensemble-mean forecast actually shows slightly less error variance than the categorical forecast, contradicting the slightly higher ensemble-mean error shown in Figures 6.4a and 6.5a.

As the forecast progresses, the rmse and S spectra grow, and the S spectra changes shape (Fig. 6.7c). By 24 h, peaks in the variance of S have developed at $k = 4$ and at $k = 10$. Domain wavenumber $k = 10$ approximately corresponds to a wavelength of 1000 km — smaller than the variance peak attributed to baroclinic, midlatitude storms. Past 24 h, variance in S at wavenumbers greater than 10 grows slowly, suggesting that the spread is saturated at a value well below the error. Though undesirable, these wavenumbers do not contribute much to the total S and rmse. The variance at $k = 10$ is $O(1)$ $m^2$, while the variance at the baroclinic wavenumbers is an order of magnitude greater.

Variances in S at smaller $k$ continue to grow throughout the 72 h forecast period, but the values are always below the rmse variances. Through 72 h, the spread under-predicts the error. Because S and rmse at $k < 10$ are still growing, the error and spread have not yet saturated at twice the climatological error, as theory predicts might occur on the order of 10 days (Leith 1974).

Figure 6.8 shows that the DIFF-MRF perturbations have a spectral shape that is different from PERT. At initialization, the baroclinic kink is clearly visible, even though $\Psi_d$ does not show a clear kink there (Fig. 5.4). The dynamic initialization routine organizes the perturbations to project onto baroclinic modes. The fact that PERT, which has a smoother initial spectrum, does not show the same behavior suggests that $\Psi_d$ does have at least a weak projection onto those modes — one that can grow very quickly.

Another difference is that S better approximates rmse at $k > 3$ in ensemble DIFF-MRF at initialization. Thus, the initial deficit of S results almost entirely from the smallest wavenumbers, where the rmse is more closely related to the bias than the error variance $E$. The dynamic initialization may introduce bias quickly. At 72 h, the shape of the S spectrum does not resemble the error spectra as much as the results from PERT.
Figure 6.7: Experiment PERT period-averaged 2D FFT spectra of 50.0 kPa geopotential height spread $S$, ensemble mean $rmse$, and categorical forecast $rmse$ at (a) initialization and (b) 72 h. Also shown are the time evolution of the (c) spread, and (d) ensemble mean error. The legends in the left column are valid in the adjacent frames. Wavenumbers $k = 2, 10,$ and 60 correspond roughly with wavelengths 6000, 1200, and 200 km, respectively.
Figure 6.8: Experiment DIFF-MRF period-averaged 2D FFT spectra of 50.0 kPa geopotential height $S$, ensemble mean rmse, and categorical forecast rmse at (a) initialization and (b) 72 h. Also shown are the time evolution of the (c) spread, and (d) ensemble mean error. The legends in the left column are valid in the adjacent frames.
Later in the forecast, DIFF-MRF actually has a weaker baroclinic peak than PERT, but the behavior at \( k > 10 \) is similar, with variances much smaller than the \( \text{rmse} \) curves show. Also agreeing with Figure 5.7, the total \( S \) is too small to account for \( \text{rmse} \).

Ensemble DIFF-GEM is similar to DIFF-MRF, as expected (Fig. 6.9). Unlike section 5.3, where DIFF-GEM was verified against MRF analyses, this error was determined by comparison with GEM analyses. At initialization, \( S \) almost perfectly matches \( \text{rmse} \) where \( k > 3 \). Before dynamic initialization, the \( \text{rmse} \) is identically zero, and the GEM analysis is one of the ensemble members (perfect ensemble). Afterward, \( S \) accounts for the error at \( k > 3 \), but not the bias that is introduced at \( k < 3 \). The evolution of the spectra over 72 h is similar to DIFF-MRF.

Because ensemble FULL includes both DIFF-MRF and DIFF-GEM, its initial spectra are similar to those (Fig. 6.10). But at the largest domain wavenumbers, \( S \) shows more variance than \( \text{rmse} \). This is noise, and because the values are so small it hardly contributes to the total \( S \). Throughout the forecast period, the shape of the \( S \) curves agree better with the \( \text{rmse} \) curves than PERT, DIFF-MRF, or DIFF-GEM. The variances are still too small, but the baroclinic kink is sharper and the spread at smaller scales, though not growing, does not disappear.

The results of this section suggest that ensembles PERT and FULL have the best spectral properties. They both show more persistent growth across the dominant scales than DIFF-MRF and DIFF-GEM. The shape of the forecast \( S \) spectra also better agrees with the forecast error.

### 6.2.4 Vertical error and spread structure

Except in section 5.1.2, vertical variation of spread and error has been ignored. Following much of the literature, the focus has been restricted to 50.0 kPa geopotential height verification. In this section, the verification is expanded to include the 100.0, 85.0, and 25.0 kPa levels, putting the 50.0 kPa results into perspective.

The ensemble-mean \( \text{rmse} \) at initialization shows little vertical variation (Fig. 6.11), as does the initial spread (Fig. 6.12). As in Figure 6.4, ensembles DIFF-GEM and FULL have the lowest initial \( \text{rmse} \). As in Figure 6.7a, the initial spread in PERT is too small compared to the error, but the shape of the \( S \) profiles are similar to the shape of the \( \text{rmse} \) profiles, with higher values at 100.0 and 25.0 kPa.

The vertical variation of \( \text{rmse} \) evolves similarly in all of the ensembles (Fig. 6.11), but corresponding growth in \( S \) is absent (Fig. 6.12). Spread grows faster at 100.0 kPa than at 85.0 kPa, just as error, but at 25.0 kPa the rapid error growth is not evident in the spread. Because the initial \( S \) profiles do not differ greatly from the
Figure 6.9: Experiment DIFF-GEM period-averaged 2D FFT spectra of 50.0 kPa geopotential height $S$, ensemble mean $\text{rmse}$, and categorical forecast $\text{rmse}$ at (a) initialization and (b) 72 h. Also shown are the time evolution of the (c) spread, and (d) ensemble mean error. The legends in the left column are valid in the adjacent frames.
Figure 6.10: Experiment FULL period-averaged 2D FFT spectra of 50.0 kPa geopotential height $S$, ensemble mean $rmse$, and categorical forecast $rmse$ at (a) initialization and (b) 72 h. Also shown are the time evolution of the (c) spread, and (d) ensemble mean error. The legends in the left column are valid in the adjacent frames.
rmse profiles, three possibilities exist for the difference: (1) the MC2 contains a large bias at 25.0 kPa, (2) the common boundary conditions impose a greater limitation aloft on $S$, and (3) sensitivity to small perturbations at jet level is much greater in the real atmosphere than in the MC2. Given that the fastest tropospheric winds are near 25.0 kPa, (2) surely contributes. Information that is common to all members of the ensemble will contaminate the ensemble most rapidly where the winds (and thus energy transport) is greatest.

Breaking down the rmse approximately into its components, $E^\frac{1}{2}$ and bias, shows that the bias does contribute to the large 25.0 kPa rmse, but most of the error is due to $E$. For example, figure 6.13 shows 72 h forecast results from ensemble FULL. The sum of bias and $E^\frac{1}{2}$ in 6.13 do not equal the rmse shown in the other figures in this section because of the order of operations, but it is a good approximation. The bias accounts for less than 30% of the total error at all levels, and the overall structure of the profiles in Figure 6.11 follow the structure of the $E^\frac{1}{2}$ profile in Figure 6.13.

Rearranging the profiles in Figures 6.11 and 6.12 clarifies the error and spread growth for each ensemble. While the growth in $S$ does not vary much above 85.0 kPa (Fig. 6.14), the rmse grows most rapidly at 25.0 kPa in all of the ensembles (Fig. 6.15), as explained above. The error has not reached saturation at any level by 72 h, as its growth continues up to that point. Similarly, the spreads grow through the entire forecast at all levels, but the rates cannot account for the error growth.

Though all of the ensembles show comparable error growth rates over each 24-h period (Fig. 6.15), the spreads of the ensembles show different evolution (Fig. 6.14). Experiment PERT shows the largest change from 0-24 h at all levels, and DIFF-MRF shows early spread increases only at 85.0 and 100.0 kPa. From 24-72 h for each ensemble, the spread grows at approximately the same rate at all levels. Though the 25.0 kPa spread of PERT does not grow much from 48-72 h, the fact that all of the ensembles continue to grow through the forecast suggests that no hard limit has been reached. All levels show slower growth from 48-72 h than from 24-48 h, suggesting that the growth curvature has changed from positive to negative (as in Figure 6.2).

Three reasons why the spread aloft might not behave similarly to the error were proposed above. Figure 6.13 showed that (1) a large bias is not responsible, but separating (2) a boundary-condition limit on spread from (3) inadequate sensitivity in the MC2 is more difficult. It is true that the spread is still growing, but the spread doubling time is between 24 and 48 h for the errors and around 48 h for the spread.

Consider that the spread at one grid point at 25.0 kPa, for example, is a combination of spread that advects or propagates to that point, and spread generated locally by development processes (i.e. $\frac{d^2 S}{dt^2}$ is the total derivative). If boundary conditions limit
Figure 6.11: Vertical profiles of geopotential height ensemble-mean rmse (against GEM analyses) for each of the ensembles, valid at forecast hour (a) 0, (b) 24, (c) 48, and (d) 72.
Figure 6.12: Vertical profiles of geopotential height ensemble spread for each of the ensembles, valid at forecast hour (a) 0, (b) 24, (c) 48, and (d) 72.
the spread that is advected or propagated, spread can still be generated locally, but the total growth rate will be smaller. It is equally plausible that too much diffusion aloft is limiting the growth in spread.

On time scales of midlatitude storm development, energy can be transported vertically through the troposphere. If the jet-level spread is limited by either false insensitivity or common boundary conditions, spread at lower levels could also be limited. Because determining the cause of this is beyond the scope of this research, it will be assumed that both of these factors are combining to control the growth in spread. While a strict comparison to true error growth may be questionable, a comparison between ensemble strategies is still valid.

The large error growth at 100.0 and 25.0 kPa, compared to 50.0 kPa, has further consequences for the results of this research. Using 50.0 kPa geopotential height as an indicator of forecast error could systematically underestimate forecast error. This cannot be stated with certainty because it intrinsically depends on the definition of forecast error. One could select a metric that includes only the 50.0 kPa geopotential height as done here, an integrated quantity such as kinetic energy, or a sensible weather parameter such as precipitation, all of which are present in the literature. Each metric has different error growth characteristics, and will display different growth in spread. The spread and error can only be assured of matching, regardless of the metric chosen, if a model and its ensemble are perfect. The dif-

Figure 6.13: Vertical profiles of ensemble-mean geopotential height rmse (against GEM analyses) components $E^{1/2}$ and bias.
Figure 6.14: Vertical profiles of geopotential height ensemble spread every 24 h, for each ensemble (a) PERT, (b) DIFF-MRF, (c) DIFF-GEM, and (d) FULL.
Figure 6.15: Vertical profiles of geopotential height ensemble-mean rmse (against GEM analyses) every 24 h, for each ensemble (a) PERT, (b) DIFF-MRF, (c) DIFF-GEM, and (d) FULL.
ferences between the perfect configuration and verification against analyses indicate weaknesses in both. In this case, the slowing of growth at 50.0 kPa is consistent with other levels. More implications are discussed in Chapter 8.

6.3 Further variations on ensemble experiments

The last two sections discussed results that suggest ensembles PERT and FULL have the most desirable characteristics. PERT shows positive spread growth through the entire forecast period, though it is too fast in the perfect-ensemble configuration. Its spread-error correlation is lower than the DIFF experiments, but it continues to grow through the forecast. Against analyses, experiment FULL shows the greatest improvement in ensemble mean skill compared to categorical skill when verified against analyses, but none of the experiments showed real improvement when verified against observations. Both PERT and FULL show consistent spread-error correlations when verified against either GEM or MRF analyses. PERT and FULL also show the best spectral evolution of error variance and spread.

Experiment PERT, with arbitrarily-chosen perturbation size, shows initial spread that is about half of analysis error. Doubling the initial spread may result in a better representation of the ranked perturbation strategy. This is easily accomplished by loosening the variance restriction (section 2.4.1), and is called PERTx2.

Ensemble FULL has an unfair advantage over all of the other ensembles simply because it has twice the number of members \( I \). Leith (1974) explained that skill improvement with an ensemble mean forecast is only guaranteed with an infinite number of members. With the exception of dynamically-conditioned ensemble methods, more recent studies have shown that more members can result in better ensemble forecasts (e.g. Houtekamer and Derome 1995). Ziehmann (2000) noted that larger ensembles are capable of higher spread-error correlations, and suggested that they have a better chance of capturing the complex evolution of the error in phase space. Thus, reducing FULL from 16 to eight perturbed members may be a fairer comparison against the other ensembles. This is easily accomplished by randomly selecting half of the perturbed members of FULL, called FULL8. The DIFF-MRF categorical forecast is used as the FULL8 categorical forecast.

Ensemble FULL8 is also motivated by efficiency concerns. Much of the literature uses simplified NWP models because running state-of-the-art models is computationally expensive. But the sacrifice in physics and/or resolution in these simplified models might introduce errors that partially offset any gains by having a greater number of ensemble members. If FULL8 performs comparably to FULL, it may be

96
used instead without sacrifices in physics or resolution.

The next two sections report an abridged evaluation of PERTx2 and FULL8. Ensemble PERT is compared to PERTx2, and ensemble FULL is compared to FULL8.

6.3.1 Doubling initial spread

A summary of performance measures for ensemble PERTx2 is shown in Figure 6.16. Panel (a) shows spread and skill for the perfect ensemble configuration, and (b)-(d) show direct a direct comparison between PERT and PERTx2.

In the perfect ensemble configuration, ensemble PERT showed too much spread relative to \textit{rmse} (Fig. 6.1a), but ensemble PERTx2 shows better agreement between spread and error (Fig. 6.16). In the perfect ensemble configuration, the agreement between \( S \) and \textit{rmse} is important because it shows that the ensemble spread is properly accounting for the error imposed on the initial conditions. Despite the higher \textit{rmse}, PERTx2 is an improvement over PERT in the perfect configuration.

Comparison against analyses also shows differences between experiments PERT and PERTx2. Figure 6.16b shows lower \textit{rmse} for ensemble PERTx2 after 24 h. Doubling the initial spread forms an ensemble that better accounts for growing errors. The initial spread is almost doubled (Fig. 6.16c), and after 24 h the PERTx2 spread grows parallel to the PERT spread.

Figure 6.16d shows that experiment PERTx2 has a higher spread-error correlation at all verifying times except 48 h, and it is growing faster at 72 h. The larger spread will contribute to higher correlations, but the lower error will reduce them. The higher spread dominates in this case.

The results presented in Figure 6.16 suggest that PERTx2 is an improvement over PERT. In particular, the better agreement between \( S \) and \textit{rmse} in the perfect configuration points to more reliable uncertainty estimates, and the lower ensemble-mean \textit{rmse} values points to higher ensemble-mean skill. The initial spread is closer to real analysis error, which is a satisfying property, though it could still be safely higher. The lower spread-error correlations appear to be a consequence of the lower error. The improvements make PERTx2 an attractive option for further study and evaluation against an independent data set.

6.3.2 Reducing the number of ensemble members

A summary of performance measures for ensemble FULL8 is shown in Figure 6.17. Panel (a) shows spread and skill for the perfect ensemble configuration, and (b)-(d) show direct a direct comparison between FULL and FULL8.
Figure 6.16: Summary of ensemble PERTx2 50.0 kPa geopotential height performance. The perfect ensemble configuration spread and skill is shown in (a). The other panels show (b) rmse, (c) S, and (d) S-E correlation of ensembles PERT and PERTx2.
In the perfect configuration, ensemble FULL showed a good agreement between spread and skill (Fig. 6.1d), but FULL8 shows greater error than spread (Fig. 6.17a). The \( \text{rms}e \) at all forecast times is greater in FULL8 than in FULL, but \( S \) does not noticeably change when the ensemble size is halved. The lack of spread relative to ensemble error shows that the smaller ensemble does not account for error growth in all the most-rapidly growing directions in phase space. Because the spread in FULL8 is greater than in ensembles DIFF-MRF and DIFF-GEM (Fig. 6.1), it captures more error growth that either of those, but FULL8 also shows greater error.

Comparison against analyses shows little difference between experiments FULL and FULL8. Against the GEM analyses, the \( \text{rms}e \) changes negligibly, and the curves in Figure 6.17b lie nearly on top of each other. The spread increases slightly from FULL to FULL8, but not as much as in the perfect configuration. Because the number of ensemble members is orders of magnitude below the dimension of the system (even just the 50.0 kPa geopotential height), such a small difference can easily be caused by the specific members that are included in FULL8. Finally, the spread-error correlation drops slightly when the number of members is reduced, agreeing with the assertion by Ziehmann (2000) that small ensembles should have a small spread-error correlation. Both ensembles FULL and FULL8 show decreasing coefficients at the end of the forecast.

The results presented in Figure 6.17 suggest that halving the ensemble negatively affects the ensemble performance. The greatest difference is apparent in the perfect configuration, and comparison with analyses shows small differences. Considering that FULL8 accounts for the same portion of analysis error as FULL, one must conclude that FULL better accounts for forecast error that may result from operational analysis errors, but when model error is included, the differences are not as important to the overall forecast skill.

Ensemble PERTx2 performs better than FULL8, and it will be used with another independent data set to verify its performance. Ensemble PERTx2 is clearly an improvement over ensemble PERT, while ensemble FULL8 does not perform as well as ensemble FULL. Ensemble FULL is computationally impractical with a model as complex as the MC2. For these experiments, about 45 minutes of wall-clock time is required for each 72 h forecast, using eight 700 MHz Pentium III processors. This is not much by today’s standards, but it would be ideal to increase the domain size by a factor of eight to maintain resolution and include the entire hemisphere to eliminate boundary problems.

The spectral characteristics and vertical structure of spread and error in ensemble PERT are similar to ensemble FULL, but the magnitudes are too small. PERTx2 successfully increases the spread magnitudes, making it competitive. In the perfect
Figure 6.17: Summary of ensemble FULL8 50.0 kPa geopotential height performance. The perfect ensemble configuration spread and skill is shown in (a). The other panels show (b) rmse, (c) $S$, and (d) $S-E$ correlation of ensembles FULL and FULL8.
configuration, the agreement between $S$ and $\text{rmse}$ is better for PERT than for FULL8. Against analyses, the spread-error correlation coefficients of ensemble PERT continue to improve through the forecast, but for ensemble FULL8 the coefficients are dropping at 72 h. Thus, PERTx2 will be used in further experiments to determine if its behavior is robust.

6.4 Summary of findings and recommendations

It is found in this chapter that a scaled ranked-perturbation approach results in the most desirable forecast characteristics of any IC ensemble studied here. This includes a spread-error correlation that grew through the 72 h forecast period, an improved ensemble-mean forecast, and $\text{rmse}$-spread agreement in the perfect-ensemble verification.

The results also show that the spread in short-range, IC-based, ensembles is not enough to account for forecast error. As in Chapter 5, an estimate of model uncertainty should also be included. This is further addressed in the next chapter.

Independent verification on specific pressure levels, as done here, is useful to understand the vertical error structure, but 50.0 kPa verification likely does not provide an accurate estimate of true forecast error or uncertainty. An integrated quantity such as total-domain kinetic energy may be better. Conversely, a sensible-weather verification (such as precipitation) is useful for defining the bottom-line of forecasting: its impact on people. But because error propagates vertically and affects other scales, much research is still necessary. Studying the vertical variation of the error and uncertainty may lead to problems that can be addressed specifically.
Chapter 7

Independent Performance Tests

7.1 Description of data set

It was demonstrated in the previous chapters that the ranked perturbations produce an ensemble with the best performance over the case-study period in Feb. 1999. This chapter further tests the ranked perturbations for robustness by applying it to more forecast days over a longer period in a different year, and using different analyses and categorical forecasts. If the behavior is similar to what was shown in Chapter 6, then the ranked-perturbation approach can be considered useful.

So that no three-day forecast periods overlap, and each forecast is nearly independent, forecast days are chosen for this verification based on a separation of analyses by at least four days. Analyses from the Aviation model (NCEP) are used as the IC categorical analyses for these experiments. Aviation (AVN) analyses were collected daily from the National Atmosphere and Oceanic Administration (NOAA) server between 20 March 2001 and 5 July 2001. Files from some days were not transferred properly, so the case days are not exactly four days apart.

A total of 12 cases were available for verification during this period (Table 7.1). Tribbia and Baumhefner (1988) estimated that more than 30 cases are required to demonstrate statistical significance, but smaller samples are still valid for giving a good approximation of results. Thus, while the ranked-perturbation approach will be shown to be robust, this finding likely does not qualify as statistically significant.

The resulting forecasts are verified against the Eta model (NCEP) analyses, valid at 00 UTC on each day. Eta analyses are also from the NOAA server. The forecast length, forecast domain, and verification sub-domain are the same as in Chapter 6, as described in Chapter 2. The Eta domain is too small for a spectral comparison,
Table 7.1: List of dates included in ensembles AVNPERT and AVNPHYS.

<table>
<thead>
<tr>
<th>Day</th>
<th>Month</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>March</td>
<td>2001</td>
</tr>
<tr>
<td>25</td>
<td>March</td>
<td>2001</td>
</tr>
<tr>
<td>8</td>
<td>April</td>
<td>2001</td>
</tr>
<tr>
<td>17</td>
<td>April</td>
<td>2001</td>
</tr>
<tr>
<td>26</td>
<td>April</td>
<td>2001</td>
</tr>
<tr>
<td>4</td>
<td>May</td>
<td>2001</td>
</tr>
<tr>
<td>22</td>
<td>May</td>
<td>2001</td>
</tr>
<tr>
<td>27</td>
<td>May</td>
<td>2001</td>
</tr>
<tr>
<td>7</td>
<td>June</td>
<td>2001</td>
</tr>
<tr>
<td>15</td>
<td>June</td>
<td>2001</td>
</tr>
<tr>
<td>24</td>
<td>June</td>
<td>2001</td>
</tr>
<tr>
<td>1</td>
<td>July</td>
<td>2001</td>
</tr>
</tbody>
</table>

so the AVN analyses are used here.

Ensemble AVNPERT is constructed using the ranked-perturbation approach, with perturbations scaled the same as ensemble PERTx2. Ensemble AVNPHYS is generated in the same way as PHYS+MRF. Eight members are run with the MC2 model, using different microphysical and convective parameterization schemes, and the ninth member is the AVN forecast itself.

The AVN 50.0 kPa spectrum (Fig. 7.1) is smoother than the MRF spectrum (Fig.5.3). The spectrum is averaged over a total of 37 days between 20 March 2001 and 13 July 2001, which includes the spring-summer transition and many different regimes. This is in contrast to the case study used in Chapters 3 – 6, which is averaged over eight days and dominated by one regime. But the spectrum shows the expected $k^{-3}$ relationship, and the ranked-perturbation approach will generate perturbations that have a similar spectrum.

The year 2001 variances of the AVN analyses shown in Figure 7.1 are lower than the variances for the previous case-study period in 1999. The MRF and GEM analysis total 50.0 kPa variances were approximately $9 \times 10^4$ m$^2$ for the Feb. 1999 case study (Table 5.1), but the corresponding AVN analysis variance averaged over the period between 20 March and 13 July 2001 was $4.09 \times 10^4$ m$^2$. Of all the AVN analyses successfully retrieved from the NOAA server, the largest 50.0 kPa variance was $7.6 \times 10^4$ m$^2$. The coefficients of matrix $W$ will be smaller when the variance is smaller,
and the ranked perturbations will consequently be smaller. Thus, initial spread will be smaller than in the case study, and decreasing from spring to summer.

One reason for the variance difference may be related to the decreased predictability during winter 1999, compared to winter 1998, noted by Shapiro et al. (2001). Though they did not explicitly calculate the 50.0 kPa geopotential height variance, they showed much greater meridional momentum and temperature eddy fluxes over the North Pacific in winter 1999. It is reasonable to expect that large 50.0 kPa geopotential height variances coexisted with these fluxes. Another reason is that the 50.0 kPa geopotential height variance is also expected to decrease with the normal seasonal transition from spring to summer in this set of analyses.

Chapter 6 established that initial spreads in ensemble PERT could be doubled and still be below the apparent 50.0 kPa geopotential height analysis error during the Feb. 1999 case study period. Ensemble PERTx2 showed better spread characteristics. For this new portion of the study, the characteristics of the initial spread relative to an estimate of analysis error is not considered. But if a decrease in analysis error accompanies a decrease in variance, then a reduction in initial (and forecast) spread in an IC ensemble is appropriate. The next two sections evaluate ensembles AVNPERT and AVNPHTHS, and compare them to results from the Feb. 1999 case study.
7.2 Ensemble verification and comparison with the 1999 case study

Both the perfect and imperfect configurations show smaller initial and forecast spread for ensemble AVNPERT (Fig. 7.2), compared with ensemble PERTx2 during the 1999 case study (Fig. 6.16a). In the perfect ensemble configuration (Fig. 7.2a), the spread and error are nearly identical. Though the spread in ensemble AVNPERT slightly decreases during the first 12 h in the perfect configuration, the growth in spread from 12 to 72 h is about the same as that displayed by ensembles PERT and PERTx2 (Fig. 6.16c).

Compared with Eta analyses, the spread grows much slower than the error (Fig. 7.2b). The reason for the large difference is bias. Figure 7.3 shows a spread that has been corrected to include bias and the spread of ensemble AVNPHYS. The corrected spread is larger than the error throughout the forecast, but the growth rates are similar. Table 7.2 confirms that the bias dominates at 72 h.

These results can be compared to Figure 5.10 and Table 5.2. The bias at initialization is a reflection of the difference between the AVN and Eta analyses. The difference is initially smaller here than for ensemble FULL, but it grows more rapidly. By 72 h, the case study period in 1999 shows greater spread from the IC-based ensemble,
Figure 7.3: Comparison of ensemble error and the corrected spread that includes the AVNPERT $S$, its bias, and the AVNPHYS $S$.

Table 7.2: Contributions to the total period-averaged 50.0 geopotential height spread (m) shown in Figure 7.3.

<table>
<thead>
<tr>
<th>Forecast h</th>
<th>AVNPERT $S$</th>
<th>AVNPHYS $S$</th>
<th>bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>11.21</td>
<td>2.45</td>
<td>3.75</td>
</tr>
<tr>
<td>24</td>
<td>11.93</td>
<td>6.64</td>
<td>11.22</td>
</tr>
<tr>
<td>48</td>
<td>13.64</td>
<td>10.71</td>
<td>14.69</td>
</tr>
<tr>
<td>72</td>
<td>15.60</td>
<td>15.60</td>
<td>18.33</td>
</tr>
</tbody>
</table>
with increasing contributions from ensemble PHYS+MRF, and relatively small bias. Conversely, bias is larger than spread in either ensemble AVNPERT or AVNPHYS at 72 h for the forecasts in Table 7.1.

The reduced uncertainty represented by ensemble AVNPERT compared to ensemble FULL suggests that the reduced predictability noted by Shapiro et al. (2001) during winter 1999 is largely IC error projecting onto sensitive modes. This is an intuitive result, because drastic improvements in the physics and numerics of operational models are infrequent. At 72 h, the uncertainty represented by AVNPERT and AVNPHYS are similar over these cases, where ensemble FULL shows more spread than ensemble PHYS+MRF, further supporting this assertion.

A reduction in sensitivity to initial conditions also implies that the model solutions are more closely tied to the boundary conditions, including upper and lower, and to internal model physics. Mean zonal flow during spring and summer 2001 will be weaker for these cases compared to the winter 1999 cases, so lateral boundaries should be less important in spring and summer. But the influence of surface fluxes, precipitation events, and interaction with topography may become greater, particularly in the spring. It is well known that convective activity over land in the northern hemisphere increases during the spring and summer months.

The spread in ensemble AVNPHYS is slightly greater than the spread in ensemble PHYS+MRF at 72 h, but the difference is not large. Rather, increased latent heating or another physical process that is closely tied to the dynamics may introduce common bias in all of the members of AVNPHYS and AVNPERT.

Although the perfect-ensemble verification shows positive spread-error correlation (Fig. 7.4a), one negative result of the reduced spread in ensemble AVNPERT is a reduced spread-error correlation when forecasts are verified against Eta analyses (Fig. 7.4b). In particular, the correlation is negative at 24 h, at the end of a period of possible spread reduction. Figure 7.4b also shows the spread-error correlation for ensemble AVNPHYS. Values remain positive through the entire forecast, and are similar for both the perfect and imperfect configurations.

Whitaker and Loughe (1998) suggest that average spread can never provide high spread-error correlation. Given that during less-active regimes, spread in IC-based ensembles is less apt to show daily variability, this imposes an undesirable limitation on short-range ensembles. That is, spread-error correlation can only be trusted during particularly active periods. Forecasts of extreme events can benefit, but to fully realize the potential of ensemble forecasts, it would be useful to find a way to boost the spread-error correlation for all regimes.
7.3 Spectral characteristics

The variance spectra of spread and error for ensemble AVNPERT also show the reduced spread (Fig. 7.5). Because the perturbation amplitude was increased as discussed in section 6.3.1, the initial spread is greater than the error at all wavenumbers (panel a). By 72 h (panel b), the spread underestimates the error as seen in Figure 7.2, but the ensemble-mean error shows lower variance than the categorical error, suggesting better forecast skill for the ensemble.

Larger wavenumbers are responsible for much of the reduction in spread over the first 24 h, but the entire spectrum does reduce (Fig. 7.5c). Spread growth is positive, but slow, from 24 to 72 h, as shown in Figure 7.2. The behavior suggests that the reduced uncertainty is not confined to a particular scale. The error (panel d) grows through the entire forecast, as expected.

One notable difference between the spectra shown in Figure 7.1 and those shown in Chapter 6 is the weaker signal at baroclinic wavelengths \(k = 4\). As discussed above, this results from the loss of baroclinic activity toward the summer season.

Figure 7.6 shows similar results for ensemble AVNPHYS, except that the spread growth better reproduces the error spectra. The initial categorical error variance is zero (panel a) because this comparison is against AVN analyses. The ensemble-mean
Figure 7.5: Experiment AVNPERT period-averaged 2D spectra of 50.0 kPa geopotential height $S$, ensemble mean rmse, and categorical forecast rmse at (a) initialization and (b) 72 h. Also shown are the time evolution of the (c) spread, and (d) ensemble mean error. The legends in the left column are valid in the adjacent frames.
forecast beats the categorical forecast for most scales at 72 h, but the improvement is not as noticeable as for ensemble AVNPERT.

7.4 Summary and conclusions

This chapter presents the ensemble verification of a new data set, over a different season. Categorical and verifying analyses are from the AVN and Eta models, respectively, instead of the MRF and GEM models. In general, ensemble AVNPERT behaves similarly to ensemble PERTx2, but with an expected spread reduction because of lower geopotential height variances and more predictable regimes.

The results here also confirm that including model uncertainty in a short-range ensemble is essential. Compared to the IC-based ensemble, the spread growth shown by the physics-based ensemble better represents error growth. This is even more important when IC error is less likely to produce large forecast errors. Removing the bias, and combining model and IC uncertainty, produces a corrected spread that can account for model error. The corrected spread is actually too high.

The generation of a physics-based ensemble is arbitrary, and the magnitude of uncertainty that it represents is dependent upon the specific parameters or parameterizations that are chosen to vary. Similarly, the perturbations in these IC-based ensembles are approximately scaled by apparent analysis error from the period 5–12 Feb. 1999, and those values may not be valid for the 2001 forecasts or for the AVN analyses. Fine-tuning of both the perturbation magnitude and the components of the physics-based ensemble, over many more forecasts, could produce a corrected spread that almost exactly matches the forecast error.

Possible reduced spread variability in AVNPERT compared to the Feb. 1999 case study, and ensemble PERTx2, produces poor spread-error correlations for ensemble AVNPERT. The results of Whitaker and Loughe (1998) and Shapiro et al. (2001) suggest this may be expected. This is one area where improvement is needed to gain more value from short-range ensemble forecasts.

In conclusion, it is confirmed that the scaled ranked-perturbation approach results in short-range (72 h) ensemble forecasts that are better than categorical forecasts, even for ensembles of only eight members. Based on error-variance spectra, the ensemble-mean forecasts perform better than categorical forecasts before 24 h, and their advantage grows through 72 h.
Figure 7.6: Experiment AVNPHYS period-averaged 2D spectra of 50.0 kPa geopotential height $S$, ensemble mean $rmse$, and categorical forecast $rmse$ at (a) initialization and (b) 72 h. Also shown are the time evolution of the (c) spread, and (d) ensemble mean error. The legends in the left column are valid in the adjacent frames.
Chapter 8

Summary, Conclusions, and Implications

8.1 Summary of the approach

A major North-Pacific weather-forecast bust, where strong winds and heavy precipitation were forecast but did not occur, provided the motivation for an investigation into the causes of the poor NWP guidance. The MC2 model was selected as representative of current-generation NWP, and it was shown to forecast a realistic 50.0 kPa geopotential height spectrum on the limited-area experiment domain. Several ensembles were run with the MC2 model, initialized at 00 UTC on each day of an eight-day period from 5–12 Feb. 1999. Forecast length was 72 h.

A scaled ranked-perturbation approach was developed that explicitly targets strong and weak features as determined by a wavelet-like transform. Eight-member ensembles, generated with ranked perturbations applied to MRF analyses, were called PERT, and were used to identify analysis uncertainty. Another eight-member ensemble (PHYS) was created by initializing each member with an unperturbed MRF analysis, but using different convective and explicit precipitation parameterization schemes to approximate model error. The MRF forecast was included in this ensemble to create ensemble PHYS+MRF. These ensembles were evaluated both against GEM analyses and under the perfect-ensemble assumption to determine the relative importance of IC and model error during that period.

Another group of ensembles were created for the same period, with the goal of isolating one possible source of daily analysis uncertainty: the differences between operational analyses. These experiments were motivated by the reported success of
ensembles consisting of different models run with different ICs, such as a group of operational models. Ensemble PHYS+MRF was also used here to account for model uncertainty.

Chapter 6 compared all of the IC ensemble strategies, and developed two modified ensembles (FULL8 and PERTx2) based on the results. When initial spread was commensurate with observed analysis error, the scaled ranked-perturbation approach showed the most-desirable characteristics, including initial spread and growth, spread-error correlation, and spectral properties. It was tested over a longer period of uncorrelated forecasts during a different season and year, and its performance was shown to be robust.

The next few sections outline the conclusions of this research. First, the case-study forecast bust is addressed, followed by general conclusions. Some implications of this research are discussed, particularly related to the North Pacific "data void", ensemble prediction, and suggestions for future predictability research.

8.2 A North Pacific bust

8.2.1 Summary of findings related to the case-study storm

For the Feb. 1999 "Storm that Never Came" to Vancouver, Canada, it is found that:

- Throughout the troposphere, the amplitude of the synoptic-scale wave that was associated with the approaching storm was persistently under-analyzed and under-forecast. The storm was forecast to track further to the southeast, while it actually tracked northeast. Also, the front actually stalled offshore in the amplifying pattern, instead of progressing inland as forecast.

- During the eight-day experiment period, normalized error-variances of ensemble PERT were greater than for ensemble PHYS+MRF, showing that IC error was generally more important than model error.

- However, for one critical short-range forecast initialized 00 UTC 10 Feb., model uncertainty was much greater than for the other days in the period, suggesting that model error was unusually large for that one forecast. For the same forecast, IC uncertainty was unusually low compared to the other forecasts in the period.

- A perfect-ensemble verification showed that bias in the MC2 was unusually large for the forecast initialized 00 UTC 10 Feb.
8.2.2 Implications

The most important implication of this portion of the study is that major errors still exist in NWP guidance to forecasters, even at short lead times. As more of the forecast burden is placed on NWP models, it becomes more critical to understand the nature of NWP errors, from both models and initial conditions.

The forecast for 10 Feb. 1999 showed that model error can lead to a poor forecast out to 72 h. Orrell et al. (2001) found that model error dominates in the short-range, but those results are mostly confined to the first few hours. At longer lead times, the interaction between model and IC error is murky because model errors can feed dynamic error growth. The results here are only valid at short lead times, while error growth is nearly linear.

But IC error can also dominate a short-range forecast. During the eight-day period, IC uncertainty was persistently greater than model error. It may be that model error or IC error is the largest contributor to forecast error under different synoptic regimes. For example, the forecast initialized 10 Feb. may have produced more precipitation in the verification domain, with associated latent heating that was misrepresented in the models. Different convective and microphysics parameterizations will have different errors in this case, leading to large spread in the PHYS ensemble. The other forecasts may have been dominated by dry development processes, leading to larger spread in the IC ensembles.

As models and DA systems become more intertwined, separating model and IC error will become more difficult. Because of the reliance on a short-range forecast to provide a first-guess field for subsequent DA cycles, a forecast with unusually-large model error can affect several cycles, potentially causing streaks of poor forecasts. To compound the difficulty, a poor short-range forecast may not be corrected in data-sparse regions such as the North Pacific, thereby increasing the likelihood of streaks of poor forecasts.

Current satellites cannot measure below thick clouds, so stormy areas are left to drift far from truth. Algorithms that retrieve temperature profiles from satellite-observed radiances require a first-guess profile that is often taken from a model forecast. Hence, satellite-derived soundings are partially dependent on an NWP forecast which may be erroneous. The problem is exacerbated where satellite soundings are expected to have large errors, such as stormy regions. All of these factors would increase forecast errors, and cause the errors to propagate from one forecast to the next.
8.3 Broader implications about short-range numerical predictability

8.3.1 Analysis error and forecast uncertainty

Findings that are related to short-range ensemble verification, and its relationship to analysis error and uncertainty are:

- Actual analysis error is greater than the difference between analyses from different operational NWP centers.
- The difference between operational analyses is not random. It appears to be related to IC uncertainty.
- At short ranges, model uncertainty must be included with IC uncertainty in an ensemble that better accounts for forecast error.
- The 50.0 kPa geopotential height error is not necessarily representative of overall forecast error.

During the study period of 5–12 Feb. 1999, the difference between either the GEM or MRF analysis and rawinsonde observations at 50.0 kPa was nearly twice the MRF–GEM difference. If this situation is generally true, ensembles of different models run from different analyses do not begin with a spread that is representative of analysis error. Studies have documented that such ensembles of forecasts from several centers perform as well or better than other IC ensembles (e.g. Ziehmann 2000 and Hou et al. 2001). This research confirms that those ensembles perform well because model uncertainty is also included in the ensemble.

Large short-range model error and linear error growth from IC errors combine to make it necessary to include model uncertainty. Whitaker and Loughe (1998) asserted that in perfect ensembles, the statistical nature of the IC error dominates in the first 2–3 days, while nonlinear error growth dominates at medium ranges. At medium ranges, nonlinear error growth results from a combination of IC error and model error near the beginning of a forecast, causing spread in an ensemble that contains both IC and model error. An ensemble that contains perturbations to the most sensitive modes will spread too quickly. This has led to ensembles initialized with spreads that are much smaller than realistic analysis error, because they are tuned to contain spread that is realistic at an optimization time that ranges from 24–72 h (Molteni and Palmer 1993, Mureau et al. 1993, Toth and Kalnay 1993, Buizza 1994a, Buizza 1994b,
Buizza 1995). In an ideal ensemble — one that fairly assigns uncertainty to either the ICs or the model on any given day — the initial spread of perturbed members must simulate realistic analysis error, and appropriate “model perturbations” must simulate realistic model error. Only then can one have confidence that the ensembles are providing true uncertainty information.

The experiments in Chapter 5 also show that analysis differences are not random. The differences are greatest over regions that are devoid of in-situ observations, such as the North Pacific. Ensembles from analysis differences (DIFF-MRF, DIFF-GEM, and FULL) show more uncertainty growth than the random (RND) ensembles. Furthermore, a two-member ensemble of control forecasts from MRF and GEM analyses (CNTL) has $rmse$ and spread similar to the FULL forecast. The FULL forecast includes the same information as CNTL, plus a random component. The similarity between FULL and CNTL shows that the analysis differences are more important than random components in IC uncertainty.

The analysis differences suggest that current DA systems make better use of in-situ than remote observations, and/or that the utility of satellite observations is limited. Likely, both are true, but the relative importance of each is open for debate. Until those issues are resolved, in-situ observing strategies appear to be the most promising for instantly reducing IC uncertainty in data-sparse regions, thereby allowing improved forecasts.

One other conclusion is that the 50.0 kPa geopotential height may not be representative of overall forecast error. Errors grow more near the surface (100.0 kPa) and near the tropopause (25.0 kPa). This is likely a combination of the smaller scales that can exist at these levels compared to 50.0 kPa, and atmospheric dynamics, thermodynamics, and turbulence that are more sensitive at those heights.

### 8.3.2 Ensemble performance

Research on perturbation methods concluded that:

- A scaled ranked-perturbation approach generates an ensemble with desirable spread-error correlations and an ensemble-mean short-range forecast that is slightly better than a categorical forecast for forecasting 50.0 kPa geopotential height, even for a small number of ensemble members.

Among the eight-member ensembles in this research, the ranked-perturbation approach, with initial spread commensurate with analysis error (PERTx2), showed the best performance of all the perturbation strategies. This included a spread-error
correlation that grew through the 72 h forecast period, agreement between spread and error in the perfect-ensemble verification, and an ensemble-mean rmse that was better than the categorical forecast. It was also the most resistant to the choice of verifying analysis.

Houtekamer and Derome (1995) compared several ensemble methods, including OSSE-Monte-Carlo forecasts and optimal perturbations, showing little or no error reduction for the first 72 h of eight-member ensembles. Stensrud et al. (1999) actually showed a reduction in forecast skill of the 50.0 kPa geopotential height over the first 72 h in ensembles of bred modes. Those results compared the higher-resolution meso-Eta ($\Delta X = 29$ km) and an ensemble of coarser ($\Delta X = 80$ km) runs, so it is not quite the same. Hou et al. (2001) compared a categorical and ensemble forecast for an ensemble of 25 members that included both IC and model uncertainty, finding that the ensemble showed better hit rates. But the smaller ensembles that accounted for only IC or model uncertainty did not perform as well.

Given its performance, the scaled ranked-perturbation of this study appears to be a viable alternative to the short-range IC perturbation methods prevalent in the literature. More sophisticated verification could confirm or refute this assertion, but the evidence here is promising. The spectral shape and vertical variation of the initial spread looks realistic. The vertical and spectral variation of growth also mimics the error growth.

The flexibility of the W transform is another advantage. Perturbations could easily be targeted to specific regions. They could also be scaled to any magnitude. Ideally, this approach will be tested on a large sample that includes more than one season.

For those researchers wishing to utilize the scaled ranked-perturbation approach demonstrated here, the steps are as follows:

1. Transform a 2D field as shown in equation 2.5, and determine the pseudo-variance structure.

2. At each scale, bin the W coefficients. The number of bins is half the number of desired ensemble members.

3. Choose a variance change from the original field that is acceptable and relates to an expected analysis error.

4. For each pair of ensemble members, multiply the coefficients in one bin by a number smaller than one and a number larger than one. The number chosen will determine the magnitude of the first-guess perturbations.
5. Check the new variance and compare to the original variance.

6. For each horizontal scale, repeat the perturbation, increasing or decreasing the multiplication factor until the difference between variance of the original field and the perturbed field is close to the prescribed constraint.

8.4 Prediction in western North America — some recommendations

The predictability research community often addresses one bottom-line question: where should resources go to make the largest positive impact on NWP forecasts, into better models or better observing/assimilation systems? During the period 5–12 Feb. 1999, both model and IC errors were large on different days, leading to one answer: better models and better observing/assimilation systems are necessary, perhaps each just as badly as the other right now. When one improves, it exposes weaknesses in the other, and their symbiotic relationship will continue for the foreseeable future.

It was noted that commercial airlines often take observations near the tropopause, but only on well-travelled flight paths. Either they are observing the wrong parameters, DA systems do not handle those observations well, or the coverage is not adequate over the Pacific. But even with better observations, tropopause dynamics may limit the predictability at jet level. The 3D nature of the atmosphere means that errors will propagate downward in the troposphere, eventually affecting predictability at all levels. More research should be completed to better describe the evolution of the vertical error structure.

As a community, we have a better chance of gaining in-situ observations of the lower troposphere than at jet level. The results here suggest that an observing strategy that covers the lowest 1-2 km of the troposphere may reduce uncertainty at those levels. This is of course attractive because, with few exceptions, the wind we feel on our faces is this low. But such a strategy may also be confounded by vertical error propagation from above. In cases where the lower troposphere controls storm development, such an observing system could be valuable. In cases where storm development is initiated at the tropopause, a lower-troposphere observing strategy may only serve to constrain a few details of the development. Thus, an ideal observing system will span as much of the depth of the troposphere as possible.

Conversely, all the experiments here show that model error is at least as important, and can sometimes dominate forecast error. Research aimed at directly improving model performance is justified.
As models improve or ICs improve, the balance between IC and model error will change, and the performance of multi-center ensembles will change with it. The advantage of a combination of ensembles such as PERT and PHYS+MRF is that it can be run at one site, and the performance continually evaluated. Representation of model and IC uncertainty can be tuned to match current capabilities. Here, the MRF is part of the PHYS+MRF ensemble, and PHYS by itself appears to underestimate model uncertainty. More physical parameterization schemes could be modified or exchanged to increase the spread in PHYS, making the MRF forecast an unnecessary component.

One thing is certain: much more research on the nature of model and IC uncertainty is necessary to complete the picture. How models and the atmosphere represent vertical error propagation, and “cascades” of error across a spectrum, are not understood. The real consequences of errors in physical parameterizations is only now being researched. Finally, the true value of ensemble forecasts may not yet be realized, and a lack of understanding of their relationship to real forecast error prevents it from completely maturing.
Appendix A

Acronym definitions

Table A.1: Miscellaneous acronym definitions.

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DA</td>
<td>Data assimilation</td>
</tr>
<tr>
<td>EnKF</td>
<td>Ensemble Kalman Filter</td>
</tr>
<tr>
<td>HPC</td>
<td>High-pressure center</td>
</tr>
<tr>
<td>LAM</td>
<td>Limited-area model</td>
</tr>
<tr>
<td>LPC</td>
<td>Low-pressure center</td>
</tr>
</tbody>
</table>

Table A.2: Operational forecasting centers (country location).

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMC</td>
<td>Canadian Meteorological Center (Canada)</td>
</tr>
<tr>
<td>ECMWF</td>
<td>European Center for Medium-Range Weather Forecasting (Britain)</td>
</tr>
<tr>
<td>NCEP</td>
<td>National Centers for Environmental Prediction (USA)</td>
</tr>
</tbody>
</table>
### Table A.3: Model acronyms.

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVN</td>
<td>Aviation (NCEP)</td>
</tr>
<tr>
<td>GEM</td>
<td>Global Environmental Multiscale (CMC)</td>
</tr>
<tr>
<td>MC2</td>
<td>Mesoscale Compressible Community</td>
</tr>
<tr>
<td>MRF</td>
<td>Medium Range Forecast (NCEP)</td>
</tr>
</tbody>
</table>

### Table A.4: Ensemble experiments.

<table>
<thead>
<tr>
<th>Name</th>
<th>Description (preturbation source)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVNPERT</td>
<td>PERTx2 around AVN (IC)</td>
</tr>
<tr>
<td>AVNPHYS</td>
<td>PHYS with additional AVN forecast (Model)</td>
</tr>
<tr>
<td>DIFF-GEM</td>
<td>Perturbations around GEM from analysis differences (IC)</td>
</tr>
<tr>
<td>DIFF-MRF</td>
<td>Perturbations around MRF from analysis differences (IC)</td>
</tr>
<tr>
<td>FULL</td>
<td>DIFF-MRF and DIFF-GEM combined (IC)</td>
</tr>
<tr>
<td>FULL8</td>
<td>Randomly-chosen members from FULL (IC)</td>
</tr>
<tr>
<td>PERT</td>
<td>Ranked perturbations (IC)</td>
</tr>
<tr>
<td>PERTx2</td>
<td>PERT with twice the initial spread (IC)</td>
</tr>
<tr>
<td>PHYS</td>
<td>Varied Cumulus and microphysics (Model)</td>
</tr>
<tr>
<td>PHYS+MRF</td>
<td>Additional member is MRF forecast (Model)</td>
</tr>
<tr>
<td>RND-GEM</td>
<td>Random perturbations around GEM (IC)</td>
</tr>
<tr>
<td>RND-MRF</td>
<td>Random perturbations around MRF (IC)</td>
</tr>
</tbody>
</table>
Bibliography


Nychka, D., C. Wilke, and J. A. Royle: 1999, Large spatial prediction problems and nonstationary random fields, manuscript.


