Ensemble Experiments on Numerical Weather Prediction Error and Uncertainty for a North Pacific Forecast Failure

JOSHUA P. HACKER

National Center for Atmospheric Research, Boulder, Colorado

E. SCOTT KRAYENHOFF AND ROLAND B. STULL

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

(Manuscript received 27 June 2001, in final form 2 August 2002)

ABSTRACT

An intense maritime cyclone off the northwest coast of North America during 9–14 February 1999 was remarkable in the repeated inability of numerical weather prediction (NWP) models to forecast its evolution. Each day during that period, the operational and research models of the United States and Canada were initialized with new analysis fields. Each day, the new short-term NWP forecasts predicted the storm to strike the densely populated Lower Mainland (Vancouver) area of southwest British Columbia, Canada, during the next 24–48 h. NWP guidance prompted the local forecast office to issue storm warnings including one or more of the following for Vancouver: heavy snow, heavy rain, and strong winds. Satellite imagery clearly showed the storm off the coast, but the storm did not strike Vancouver until much later and in a decayed state. This synoptic case is studied with an aim to understand the source of the NWP error, and an ensemble of research model runs is made to address three possibilities for failure: 1) initial condition (IC) error, 2) model error for a particularly nonlinear or sensitive event, and 3) sympathetic data denial. To estimate the effect of IC uncertainty, a short-range ensemble system is developed and tested on a limited-area model for a sequence of successive 3-day reforecasts covering the 10-day period surrounding the storm. This IC ensemble shows some correlation between spread and skill and provides one estimate of IC uncertainty. To estimate the effect of model uncertainty, a physics-based ensemble is run for the same period. The effect of data denial is investigated by comparing forecasts made with the same model but from analyses created at different operational centers. Results suggest that if the runs initialized at 0000 UTC 10 February 1999 were used for guidance, model uncertainty was likely responsible for the forecast failure. It had larger-than-average model error but lower-than-average IC error. Subsequent forecast errors were likely dominated by IC uncertainty. An attempt at assessing sympathetic data denial is inconclusive.

1. Introduction

In early February 1999, residents of the Lower Mainland (Vancouver) area of southwestern British Columbia (BC), Canada (Fig. 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 February, local time. *Geostationary Operational Environmental Satellite-10* (*GOES-10*) imagery showed a mature marine cyclone centered at 45°N, 160°W at 0000 UTC 10 February (Fig. 2a).

On the morning of 10 February [1000 Pacific standard time (PST), 1800 UTC], the local forecast office issued both a wind warning (forecast to reach 60–80 km h^{-1}) for later that day, and a snowfall warning (accumulations of 4–8 cm) for the evening. The numerical weather prediction (NWP) guidance indicated that the storm would continue into 11 February, leading local forecasters to modify their afternoon forecast update issued on 10 February to include wind and snowfall warnings, with winds increasing to 70–90 km h^{-1} and liquid-equivalent accumulations of 50 mm by 11 February.

The low pressure center (LPC) was forecast to be a few hundred kilometers off of the central BC coast by the evening of 10 February (0000 UTC 11 February), with an associated occluded front forecast to cross the Lower Mainland during the overnight hours. Satellite imagery showed the storm approaching the coast (Fig. 2b), but the cyclone was too far offshore for the Lower Mainland to experience precipitation by that time.

The storm did not strike the Lower Mainland on 10 or 11 February. Winds of 20 km h^{-1} and negligible precipitation were recorded at Vancouver International

Corresponding author address: Dr. Joshua P. Hacker, NCAR/ASP, P.O. Box 3000, Boulder, CO 80307-3000.
E-mail: hacker@ucar.edu

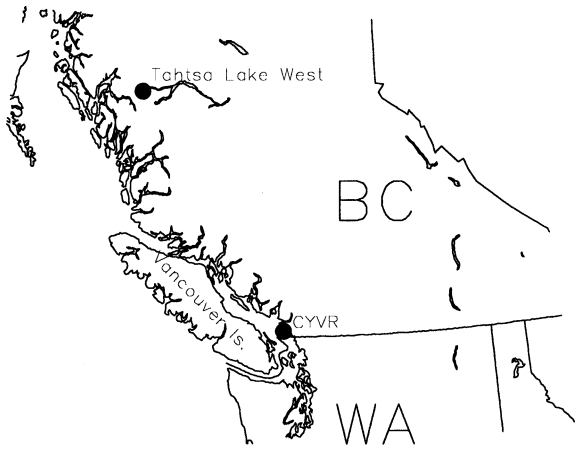


FIG. 1. Map of southern BC, Canada, and northern WA, showing Vancouver (CYVR) in the Lower Mainland, Tahtsa Lake West, and Vancouver Island.

Airport (CYVR) at 1600 PST (UTC = PST + 8 h) on 11 February (Fig. 3). What made this storm remarkable was not just the forecast ''bust'' on 10 and 11 February, but also that subsequent numerical model runs initialized on 11 and 12 February each produced revised guidance that incorrectly brought the front with heavy precipitation and strong winds into the Lower Mainland. The region did not receive significant precipitation until after 0400 PST (1200 UTC) on 13 February. Satellite imagery shows the system with an occluded front just to the west of Vancouver at 2300 UTC February 11, and the decayed system beginning to move over the forecast region at 0000 UTC 13 February (Figs. 2c and 2d). Figure 2c suggests two circulation centers near 45°N, 140°W and 55°N, 140°W.

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 NWP models from 0000 UTC 10 February 1999 demonstrate a wide range of solutions for a single, strong LPC (Fig. 4). Verifying hand analyses produced by the Pacific Weather Centre (PWC; Pacific and Yukon Region, Meteorological Service of Canada) show the cyclone splitting into two LPCs (Fig. 4).

The storm was indeed as vigorous as forecast but the forecast track, the 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 central BC coast, farther north, received the brunt of the storm. For example, 500 km north of Vancouver, the village of Tahtsa Lake West, in the coast mountains near the Alaska Panhandle, recorded a BC 1-day snowfall record of 145.0 cm on 11 February. For the main urban and economic center of southwestern BC, however, the forecast was a repeated bust.

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 midlatitude oceans can lead to numerical forecast degeneration downstream over continents. The lack of data exacerbates forecast difficulties associated with the underlying nonlinearity of the atmosphere, limits the range of predictability, and highlights sensitivity of the real atmosphere (and model atmosphere) to disturbances or error in initial conditions (Lorenz 1963). Shapiro et al. (2001) note unusually high NWP forecast error from 26 January through 10 February 1999 and suggest that the error may be related to the El Niño–Southern Oscillation (ENSO). For this study, the forecast bust provides motivation to examine forecast uncertainty and the possible causes of NWP error.

Three possible factors contributing to the failure of this forecast are 1) initial condition (IC) error, 2) model error for a particularly nonlinear or sensitive event, and 3) sympathetic data denial. Determining the possible contribution of each factor to the bust can help to evaluate the state of NWP in the North Pacific and western North America and possibly to indicate 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 ensemble perturbation generation 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). Note that only *uncertainty* can be measured via ensemble techniques. It must be interpreted as an estimate of error, which can never be precisely known. Factor 3 can be 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.

This study attempts to quantitatively determine the contribution of each factor to the NWP error of this storm event in the North Pacific. A new IC perturbation method is developed that is flexible, fast, and simple to implement, giving an idea of the forecast uncertainty that might result from IC uncertainty. A physics-based ensemble and forecasts from three entirely independent models provide an estimate of model uncertainty. Comparison of analysis fields from the Canadian Meteorological Centre (CMC) and the U.S. National Centers for Environmental Prediction (NCEP) will be used as a surrogate for data denial information. Comparing error variances and spread of these three types of ensembles indicates which is more important.

The next section describes the synoptic situation and provides a qualitative assessment of analysis and forecast error around one critical forecast, initialized 0000 UTC 10 February. Section 3 describes IC perturbation experiments, including the data and models, the perturbation method, and experiment design. It also ex-

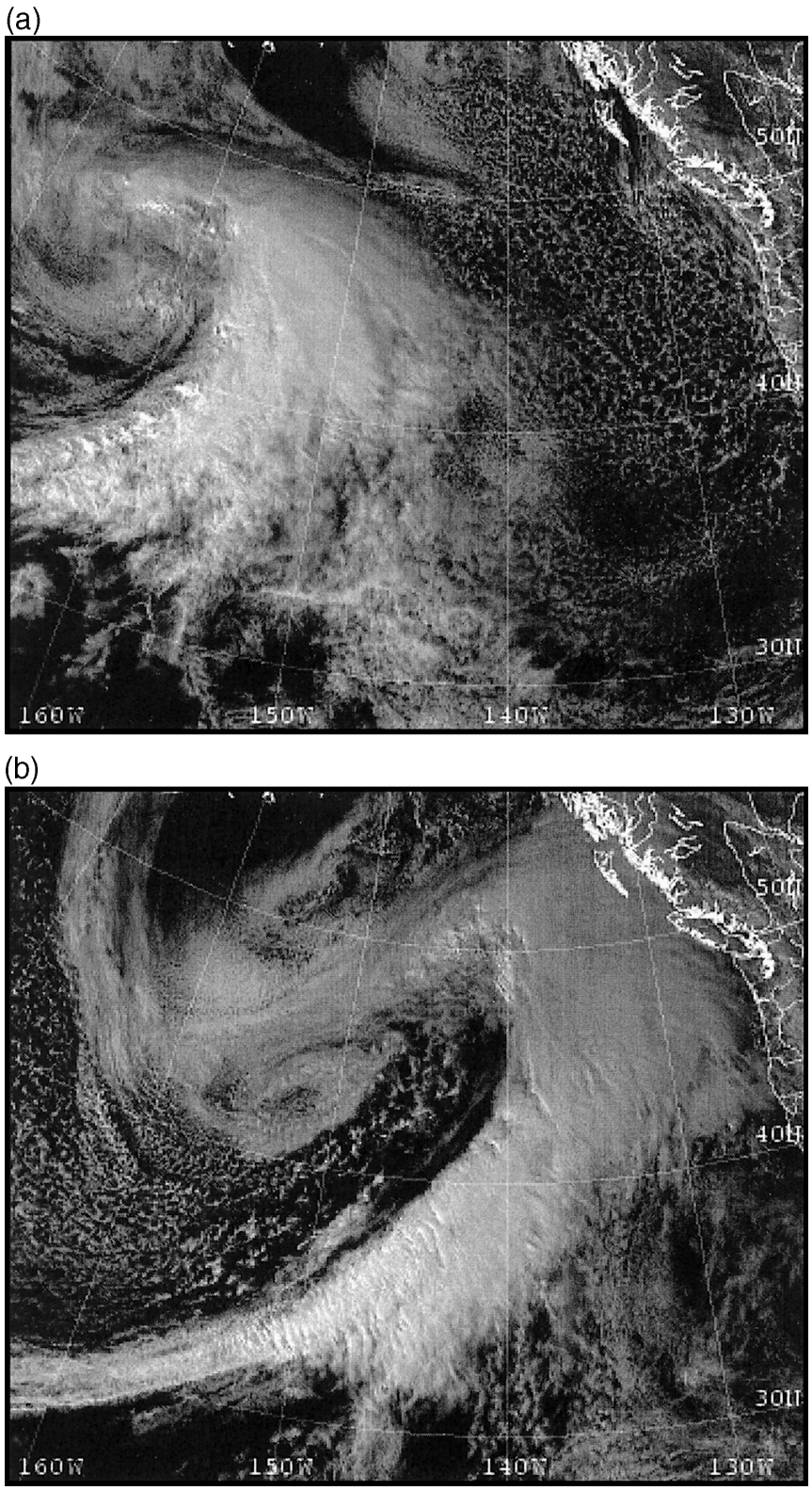
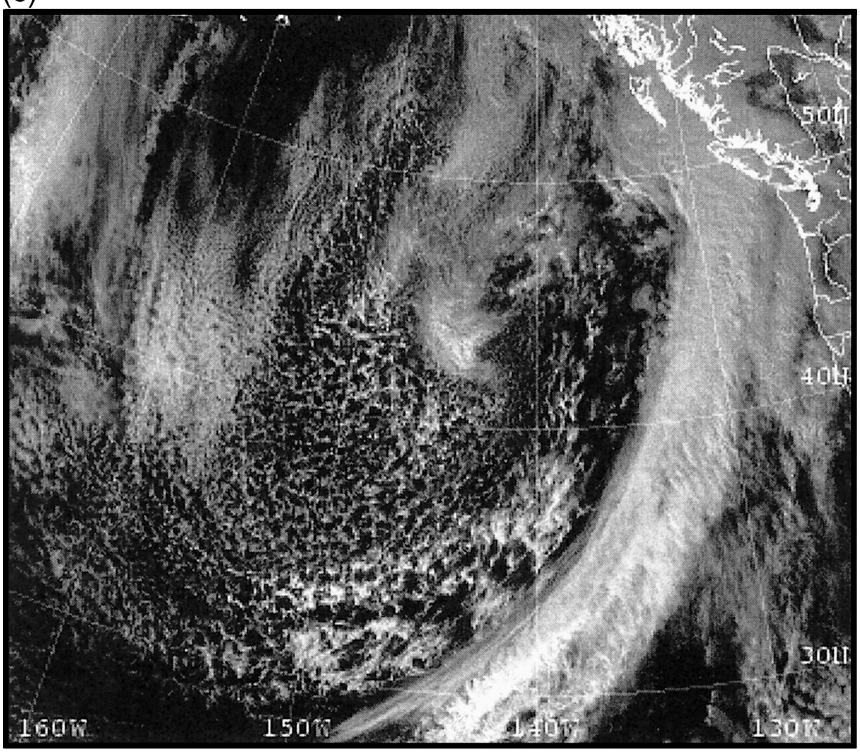


FIG. 2. *GOES-10* visible images of the North Pacific, valid (a) 0000 UTC 10 Feb, (b) 0000 UTC 11 Feb, (c) 2300 UTC 11 Feb, and (d) 0000 UTC 13 Feb 1999.

(c)



(d)

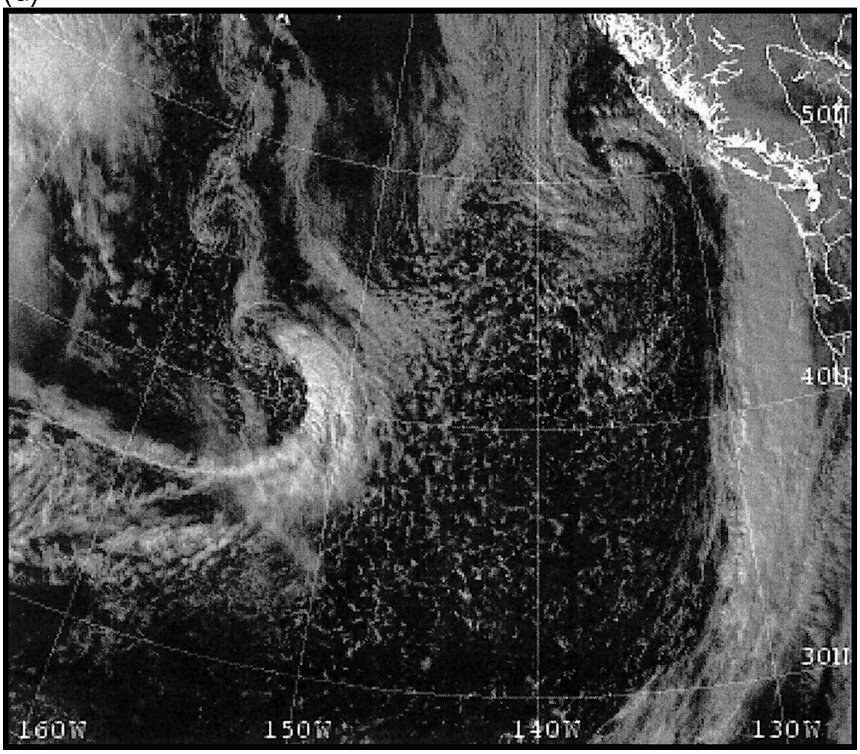


FIG. 2. (*Continued*)

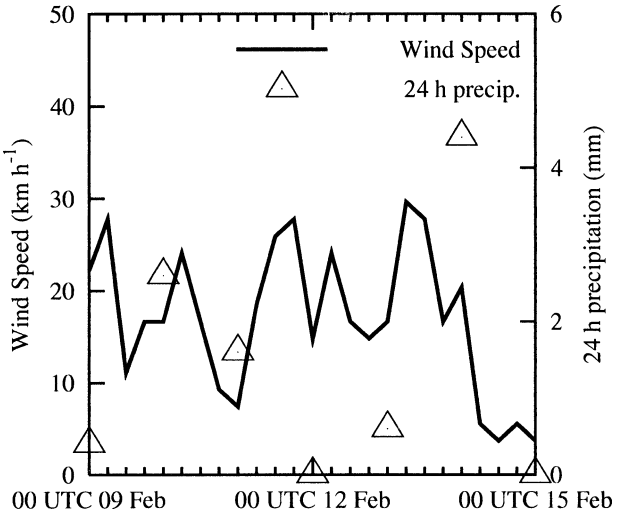


FIG. 3. CYVR precipitation and wind speed observations for the period 9–16 Feb 1999. The front passed at approximately 1200 UTC 13 Feb, and precipitation prior to that was a convective air mass. Precipitation values are 24-h accumulations from 1200 to 1200 UTC around the 0000 UTC date of each mark.

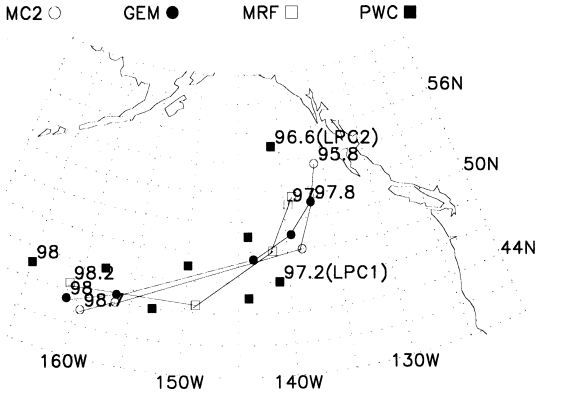


FIG. 4. Analyzed and 48-h forecast tracks from 0000 UTC 10 Feb 1999. Central pressures (kPa) are shown valid at 0 and 48 h. Forecasts from the MRF, GEM, and MC2 models are shown with the PWC hand-analyzed storm tracks. Two low pressure centers (LPC1 and LPC2) are shown in the PWC analysis at 48 h, whereas the model forecasts only evolved one low pressure center.

amines characteristics of the ensemble with the goal of determining how well it represents IC uncertainty. The three possible causes of forecast failure identified above are investigated in section 4, and a short discussion of the effect of boundary conditions is presented. The conclusions are summarized in section 5.

2. The storm that never came

This section describes the synoptic development of the storm and qualitatively compares Medium Range Forecast (MRF) model and Global Environmental Multiscale (GEM) analyses with each other and with the few available in situ observations. It will be shown that both analyses incorrectly located the storm center. The ensuing forecasts persistently brought the upper-level wave and associated surface LPC onshore too soon and with an incorrect trajectory. The cyclone split into an active LPC, developing near the frontal triple point downstream, and the occluded remnants of the original LPC that remained aligned with the 50-kPa trough. The occluded LPC was later responsible for increased warm-air advection ahead of the front, amplifying the pattern and slowing frontal progression. NWP forecasts filled the occluded LPC too rapidly and did not show its effects on the downstream ridge.

The synoptic development at the 100-, 50-, and 25-kPa isobaric levels, as analyzed by the MRF and GEM data assimilation (DA) systems, is shown in Figs. 5–7. PWC hand-analyzed fronts are also shown on the 100-kPa charts. At 0000 UTC 9 February (panels a and b in Figs. 5, 6) the 100- and 50-kPa geopotential height contours suggest geostrophic wind veering and concomitant warm-air advection near 40°N, 170°W, ahead of the incipient LPC evident just upstream. At 50 kPa an associated trough aloft was not yet visible within the domain (Figs. 6a and 6b), but the exit of an encroaching jet streak was just evident on the edge (Figs. 7a and 7b).

Between 0000 UTC 9 February and 10 February, the LPC deepened at a rate of 1.0 Bergerons [0.1 kPa h^{-1} for 24 h adjusted for latitude (Sanders and Gyakum 1980)] to become a mature cyclone centered near 45°N, 160°W (Figs. 5c,d, and 2a). The cyclone was supported by a 50-kPa short-wave trough (Figs. 6c,d) and a jet streak with winds in excess of 120 kt (Figs. 7c,d). The final two panels of Figs. 5–7, valid 0000 UTC 11 February, show an amplified pattern aloft supporting an elongated LPC near the surface that was oriented with the 50-kPa steering flow. The 0000 UTC 11 February analyses, corresponding to Fig. 2b, were valid at approximately the same time that the heavy precipitation and strong winds were initially forecast to begin in Vancouver. From Figs. 5e and 5f it is clear that the fronts were still well offshore.

Figures 5–7 also show observation − analysis (MRF) differences at the observation locations in the left column and the MRF − GEM difference in the right. Observation − analysis differences for GEM (not shown) look similar to those discussed here. Very few in situ observations exist over the Pacific, making it difficult to draw conclusions about the quality of the analysis there, but three dropsonde observations were taken over the central Pacific on both 9 and 10 February. At 100 kPa on 9 February (Fig. 5a), those three observation − analysis differences were positive, suggesting analyzed geopotential heights that were too low. The values were small but are larger toward the west, closer to the incipient LPC. This could have resulted from an underanalyzed high pressure center or an erroneous eastward shift in the near-surface pattern. The MRF − GEM differences (dashed lines in Fig. 5b) show only small values in the region of the developing LPC.

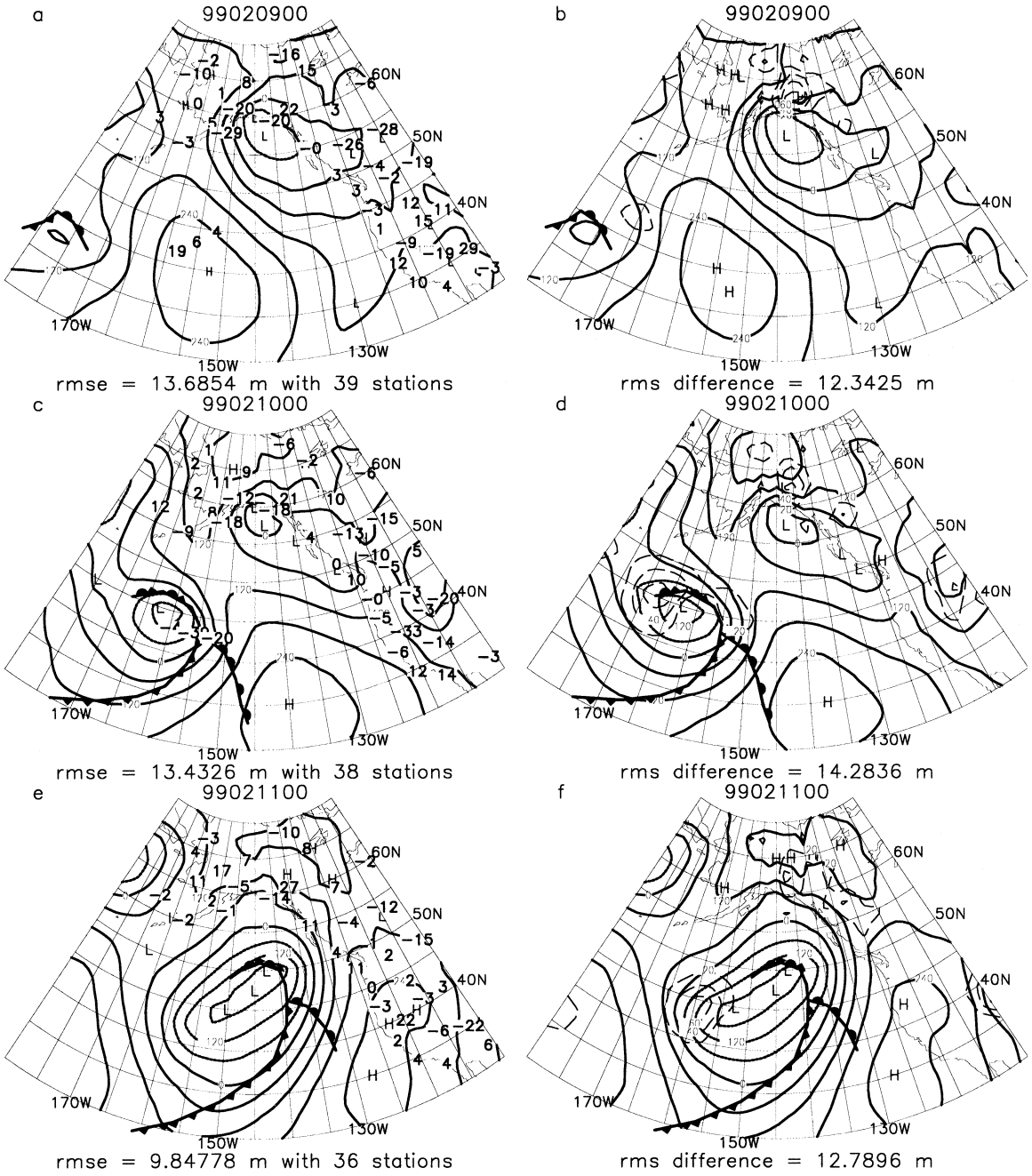


FIG. 5. (left) MRF and (right) GEM 100-kPa geopotential height analyses (solid lines), valid 0000 UTC on (a), (b) 9, (c), (d) 10, and (e), (f) 11 Feb 1999. Contour interval is 60 m. The left column also shows valid observation − analysis differences. The right column shows the difference between the MRF and GEM analyses (dashed lines) at contour intervals of 10 m. Below each plot on the left, the rmse is shown, calculated with the stations shown in the plot. On the right, the rms difference between the MRF and GEM analyses is shown, using the grid points within the domain shown. PWC hand-analyzed fronts are superimposed.

One day later (Fig. 5c) the differences in the same region were negative and larger in magnitude, suggesting that analyzed geopotential heights were too high. The more-mature LPC was analyzed too far west, or its central geopotential height was erroneously high. This is generally the opposite case from the previous day. If those observations are accurate, the MRF forecast/DA system did not track the early development of the system well. Greater disagreement between the MRF and GEM analyses is evident in Fig. 5d. With stronger analyzed geopotential height gradients, small central location discrepancies will show up as large differences, which in this case were greater than 30 m.

At 0000 UTC 11 February, no in situ observations

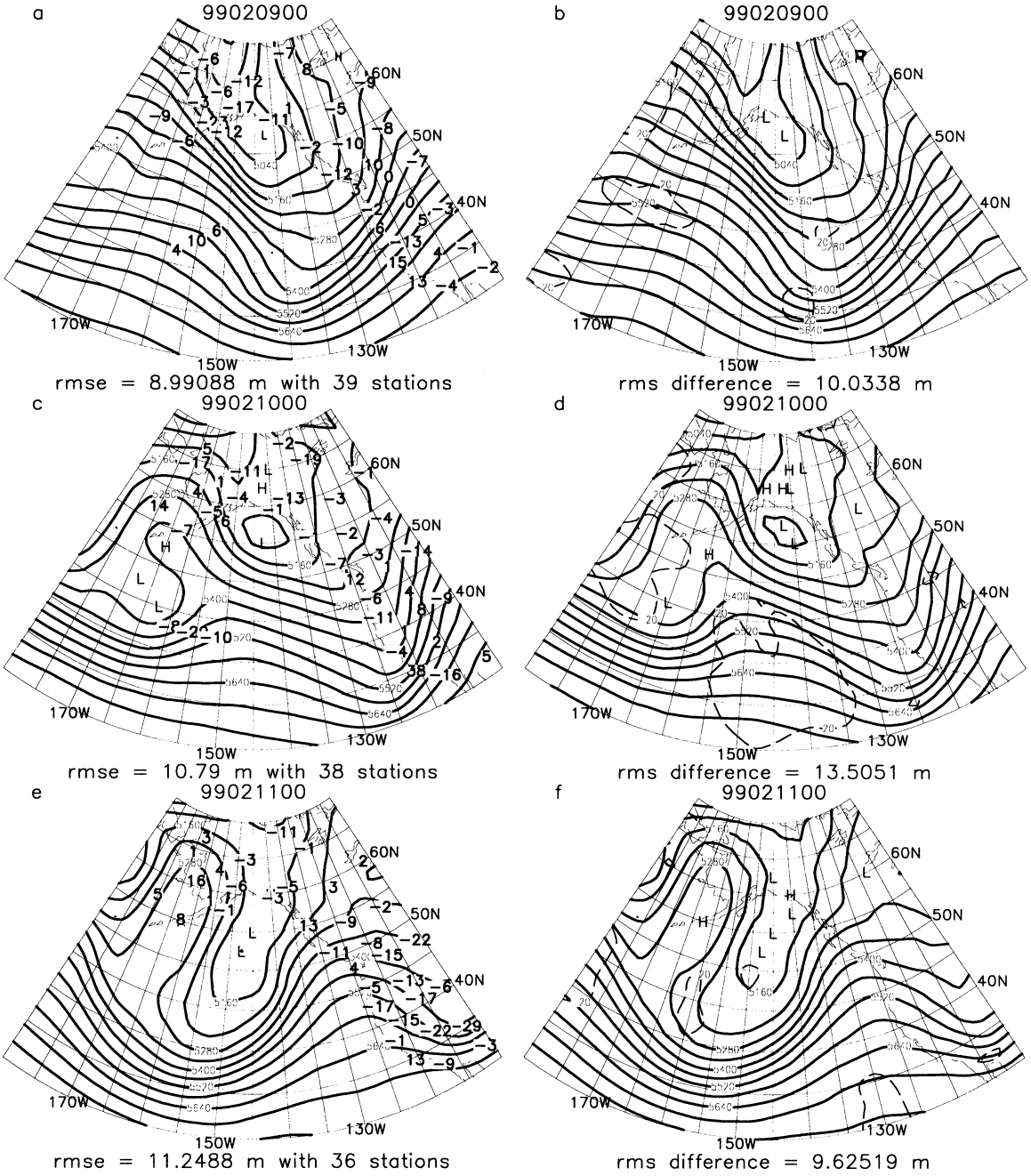


FIG. 6. Same as in Fig. 5 but valid at 50 kPa.

were available over the central North Pacific, but coastal observations provide some clues. Along the BC coast, positive coastal rawinsonde observation − analysis differences suggest that the analyzed LPC was too far east, while negative values along the southern Alaskan coast suggest that it was analyzed too far south. The downstream high pressure center may have also been underanalyzed, as suggested by the difference of 22 m near the center of its northern lobe. The MRF − GEM differences were about the same magnitude as the day before, and the largest differences were confined to the trailing end of the elongated LPC near where the original LPC had occluded. The shape of the LPC at this time suggests some tendency to split into two circulation centers, which hand analyses captured but which neither the MRF nor GEM analyses adequately represented.

Comparing the same central North Pacific dropsonde observations with analyses shows qualitatively similar error characteristics at 50 kPa. The differences on 9 February were small, but the differences on 10 February

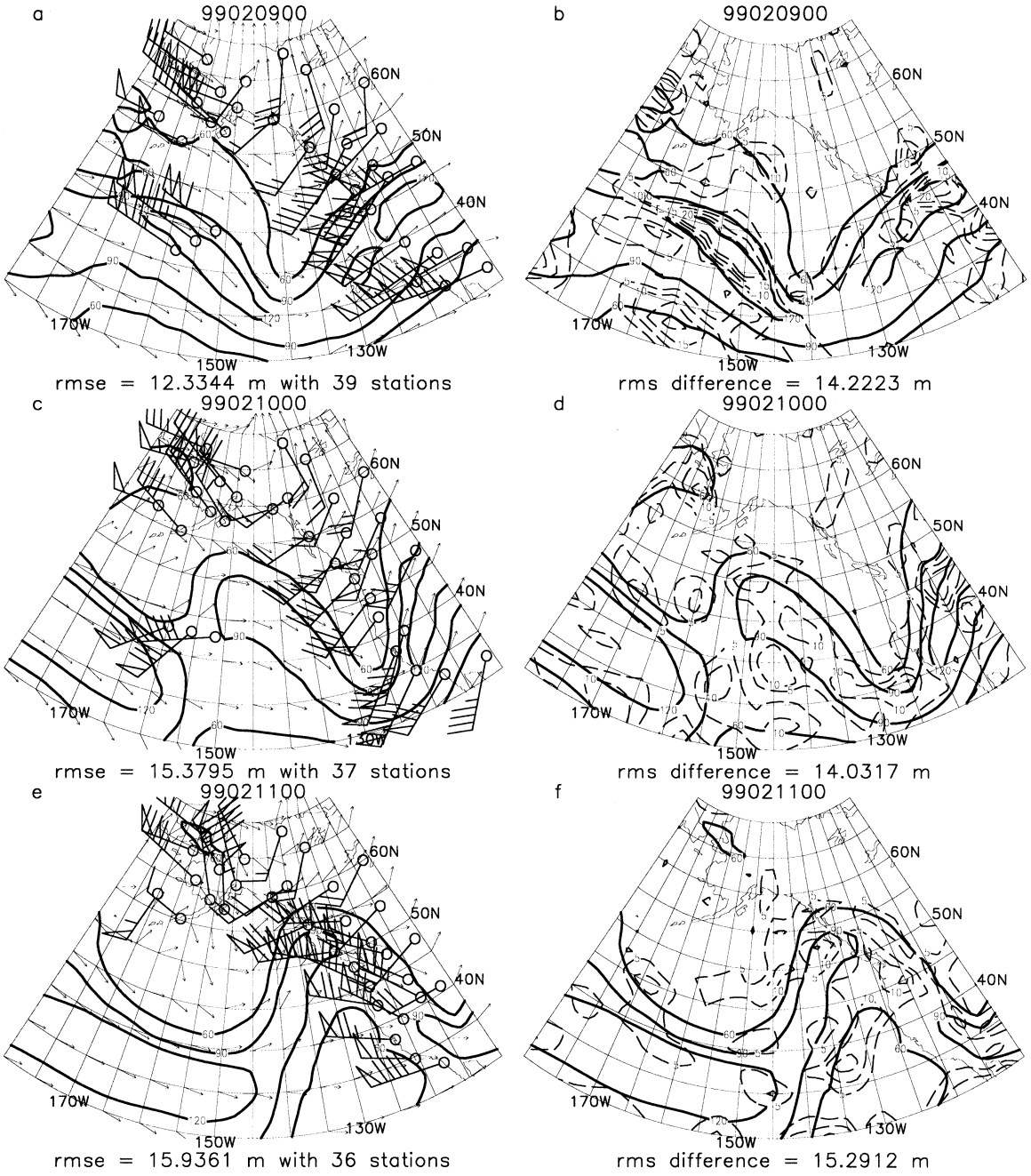


FIG. 7. Same as in Fig. 5 but for winds valid at 25 kPa. The vectors denote analyzed wind direction, and the isotachs denote analyzed wind speed (kt).

were larger and negative, showing that the analyzed geopotential heights were too high. Negative differences in the downstream trough over the coast of North America show that the trough was not deep enough (Fig. 6c). MRF − GEM differences were greater than 20 m over large areas in both the trough and the downstream ridge (Fig. 6d). At 0000 UTC 11 February the observed 50-kPa geopotential heights were lower than analyzed along the North American coast (Figs. 6e and 6f). This could result from erroneous ridge axis placement or an overdeveloped ridge over southwestern BC. Conversely, the observation − analysis value of 13 m over the central BC coast shows that the ridge was underdeveloped there. MRF − GEM differences were smaller than the day before, and qualitatively the MRF and GEM analyses more closely agreed.

The 25.0-kPa winds shown in Fig. 7a indicate that wind speeds were generally underestimated in the 9 Feb-

ruary analyses. The available in situ observations were all downstream of the developing jet core at 0000 UTC 9 February and they are of limited value in understanding the error. The observations valid at 0000 UTC 10 and 11 February are also not helpful because they do not coincide with an analyzed jet core, but the differences between the MRF and GEM analyses are widespread. The importance of jet-level dynamics to storm development (e.g., Uccellini et al. 1985) leads one to suspect that two model forecasts from these different analyses could diverge substantially.

Rms values are reported below each plot in Figs. 5–7 for two reasons: to see the daily variability in one simple estimate of analysis error, and to statistically compare the error with the analysis differences. In the left columns, which compare MRF analyses with in situ observations, the rmse of geopotential height at that level is reported. The number of stations included in the computation, which is simply the total number shown in the plot, is also shown. The rmse estimates are not robust because of the limited number of comparison points, but they are rough indicators of error. Below the plots in the right columns an rms difference in geopotential height between the MRF and GEM analyses is reported. That calculation includes all available grid points in the domain shown. In each of the Figs. 5–7 it is easy to see that the rms difference is similar in magnitude to the corresponding rmse. The daily variability in both rmse and rms differences is small, and they do not vary together. Last, we note that the rms values are generally greater near the surface and at jet level than they are at 50 kPa. From this glimpse of error, the analyses appear better in the midtroposphere.

We next turn from analyses to forecasts. No in situ observations were reported over the Pacific after 0000 UTC 10 February, but the observations available over the North American continent are enough to show that the 48-h forecasts valid at 0000 UTC on 11 (Fig. 8) and 12 (Fig. 9) February both brought the storm onshore too rapidly. Near the surface at 0000 UTC 11 February (Fig. 8a), positive observation − forecast values indicate forecast geopotential heights that were too low. This is the opposite of the analysis error shown in Fig. 5a. The largest error was near the long semiaxis of the LPC (note 25-m value to the northeast of the LPC). Farther southeast along the coast and inland, the errors decreased and some were negative, indicating smaller error where the gradients were weaker and regions where the forecast geopotential heights were too high. A north-westward shift of the forecast pattern would diminish much of the error.

At 50 kPa (Fig. 8b), the situation is similar with large positive error along the north coast and negative toward the southeast. The jet-level forecast does not elucidate the situation and is shown for completeness. The forecast rmse values shown are much higher than the analyzed values shown in Figs. 5–7, as expected, and the error growth rates are fastest near the tropopause in this case.

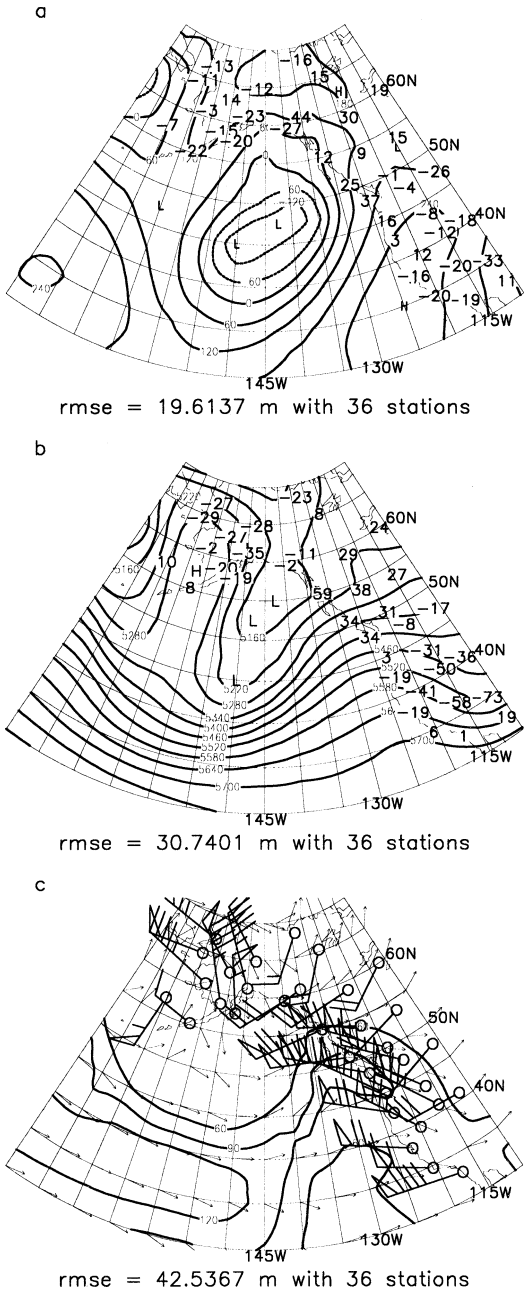


FIG. 8. MRF 48-h forecast initialized 0000 UTC 9 Feb, and valid 0000 UTC 11 Feb 1999. Forecasts of (a) 100-kPa geopotential height with observation − forecast values, (b) 50-kPa geopotential height with observation − forecast values, and (c) 25-kPa isotachs with observed wind barbs are shown. Also reported are geopotential height forecast rmse values.

Errors in the next 0000 UTC forecast, valid 0000 UTC 12 February for a 48-h forecast, continue to show the forecast pattern too far east. The stronger gradients that had invaded the coast by this time contributed to larger positive observation − forecast differences over the

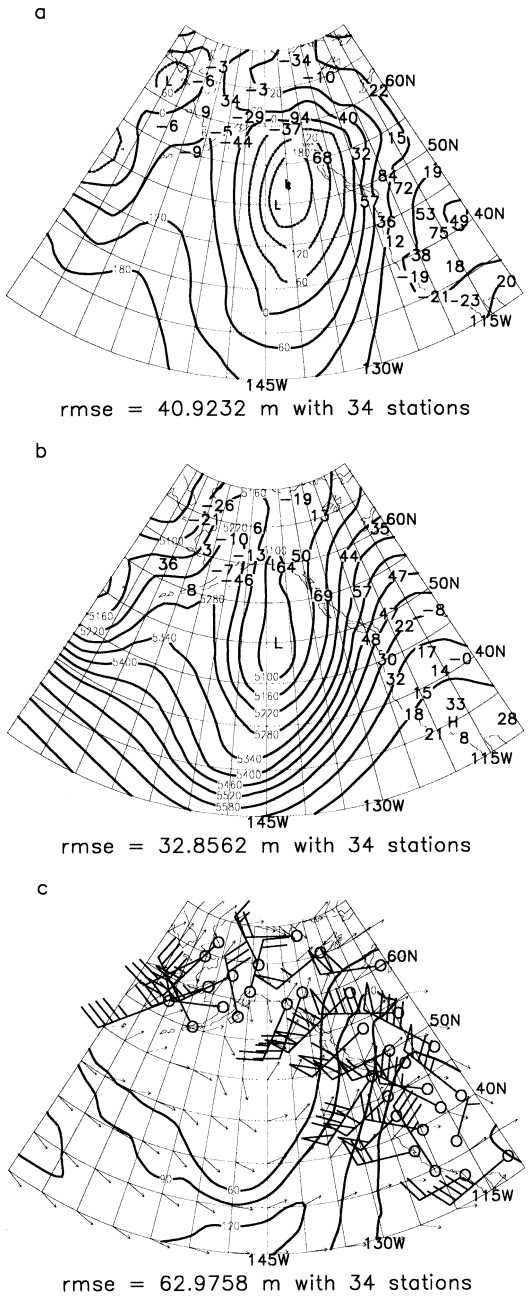


FIG. 9. Same as in Fig. 8 but for forecast initialized 0000 UTC 10 Feb and valid 12 Feb 1999.

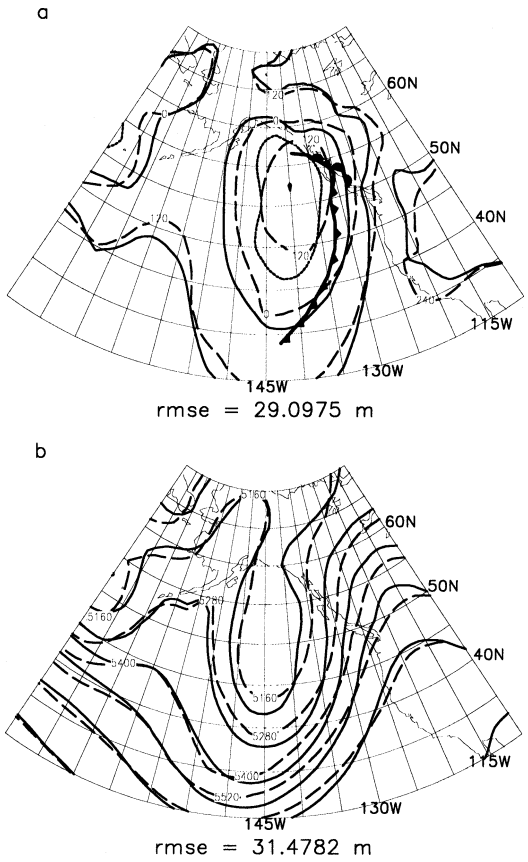


FIG. 10. MRF 48-h forecast (dashed) and verifying analysis (solid) valid 0000 UTC 12 Feb 1999. (a) The 100-kPa isobaric surface, and (b) the 50-kPa surface. Contour interval is 120 m. PWC hand-analyzed fronts are superimposed on the 100-kPa plot.

western United States and Canada, particularly at 100 kPa (Fig. 9a). As expected the rmse is much larger than for the previous 48-h forecast. At 50 kPa (Fig. 9b), the rmse value is slightly higher and the error over the continent is all positive, showing that the forecast geopotential heights were all too low there. Again, the error growth near the tropopause was the greatest.

Comparing 48-h MRF forecast geopotential heights with verifying analyses further clarifies the error. Figure 10a shows a northeasterly skewed forecast LPC, but no skewness in the verifying LPC. Though the contouring does not adequately show it, the analysis has two LPCs rather than one elongated LPC as forecast. Satellite imagery confirms two LPCs (Fig. 2c). The 50-kPa pattern shows that the forecast downstream ridge was underamplified. Consistent with the surface development, the forecast trough axis shows a greater eastward tilt (with latitude) than the verifying analysis.

Measuring error at grid points rather than at a few station locations causes the rmse values reported in Fig. 10 to be lower than in Fig. 9 at 100 kPa. This illustrates a problem with verifying against analyses rather than observations. The near-surface analysis in data-sparse regions is prone to large error arising from poor background forecasts that are never properly corrected in the DA cycle, and the forecast appears more favorable.

Another way to view the forecast storm development is through positive vorticity advection by the thermal wind, which indicates upward motion and possibly development (Trenberth 1978). Figure 11a shows the MRF 48-h forecast 100–50-kPa thickness and 60-kPa positive vorticity advection by the thermal wind (PVA_T) valid 0000 UTC 12 February, and Fig. 11b shows the verifying analysis. Near 60°N, 142°W, the forecast advection

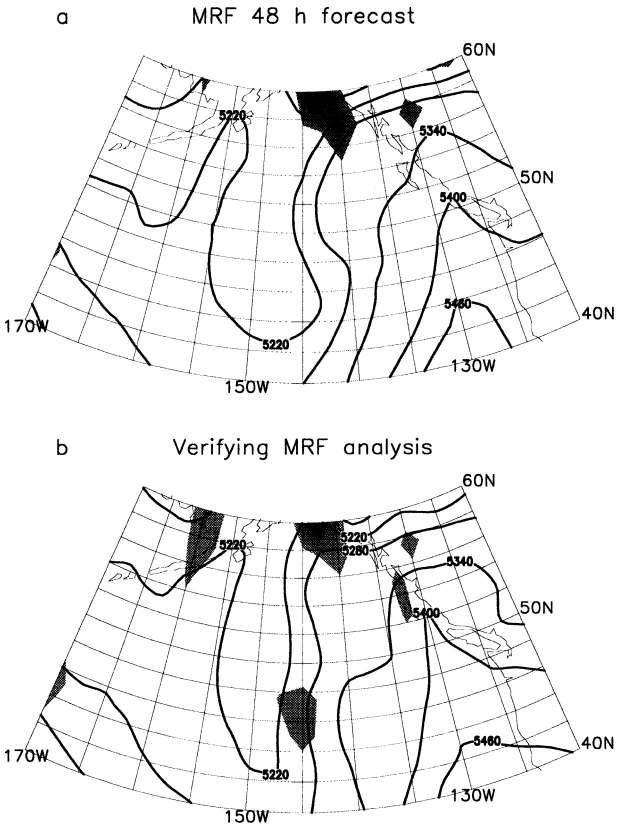


FIG. 11. Lower-tropospheric positive vorticity advection by the thermal wind (PVA_T; shaded values greater than 2×10^{-9} m s^{-2}), valid 0000 UTC 12 Feb 1999, and 100–50-kPa thickness (thick lines at 60-m intervals). (a) The 48-h MRF forecast, and (b) the verifying analysis.

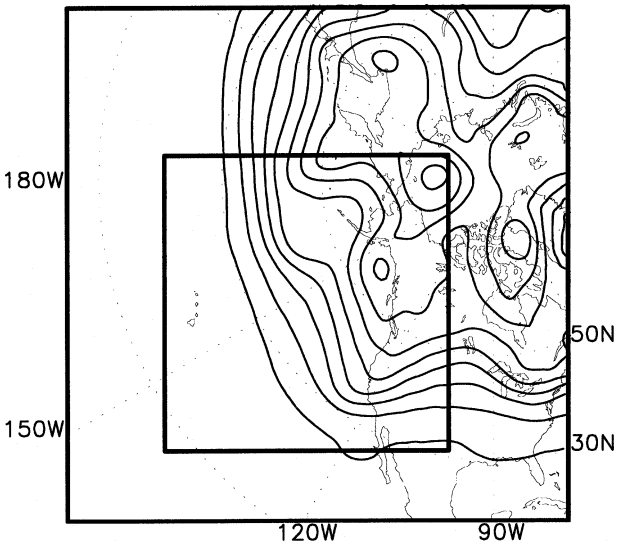


FIG. 12. Complete MC2 domain (outer box) with verification region (inner box) for the MC2, MRF, and GEM models.

by the thermal wind agrees well with the analysis, but analyzed PVA_T near 46°N, 142°W is absent in the forecast. Consequently the single, elongated, LPC that was forecast by the MRF is only associated with development along the thermal ridge axis to the north. The secondary, southern center of PVA_T is associated with upward motion and a secondary LPC at its downstream edge.

The southern vorticity center, which comprised the remnants of the original occluded LPC, generated warm-air advection ahead of it, allowing the thermal ridge to develop further (note 5340-m contour). In this case, the thermal ridge in the verifying analysis is farther west and oriented more northwesterly than the forecast thermal ridge. The lack of a strong forecast secondary circulation center causing warm-air advection downstream may have contributed to the underamplification of the downstream ridge shown in Fig. 10b.

The above qualitative description of analysis error shows that the storm location was not well analyzed. While verifying analyses showed the occluded (southern) LPC played a continuing role by increasing warm advection and possibly amplifying the ridge downstream, the numerical forecasts filled it and diminished its influence. The resulting forecast pattern was not sufficiently meridional and the predicted storm made landfall well before the actual storm. Significant weather did not strike until 13 February, and it was a fraction of what had been expected over the previous 3 days. Frontal passage on 13 February 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 1200 UTC 14 February. The surface temperature drop, associated with the cold air mass behind the front, occurred in the early hours of 14 February, and was likely delayed by interactions with orography and the formation of a trapped marine boundary layer. This assessment shows only that identifiable analysis and forecast differences and error existed. However, determining a dominant factor is nearly impossible without a quantitative approach, as described in the next section.

3. Design of the ensemble experiments

a. Data, models, and experiment overview

Forecasts from the GEM model (Côté et al. 1998), the Canadian Mesoscale Compressible Community (MC2) model (Benoit et al. 1997), and the MRF model make up a suite of experiments that address forecast uncertainty and provide estimates of error. The GEM and the MRF were operationally run on a global domain, and the respective operational centers provided the output grids for this study. The MC2 was run on a limited domain (outer box in Fig. 12) by the authors at the University of British Columbia (UBC), initialized at 0000 UTC on each day during 5–12 February 1999. For comparison, all of the 72-h forecasts were interpolated to a common 50×50 point grid (inner box in Fig. 12).

To study the contributions of analysis and model error,

different ensembles were run with the MC2 model for each forecast period. Experiment PERT used perturbations to the initial conditions, experiment PHYS used different physical parameterization schemes, and experiment DENY used identical runs with the MC2 started from the different NCEP and CMC analyses. Generation of these ensembles is discussed in sections 3b and 3c below. Categorical (unperturbed) ICs and common boundary conditions for the ensemble experiments were 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 were introduced to the ICs in this experiment. When perturbations are introduced (PERT), IC differences appear. The ''perfect model'' assumption (e.g., Mullen and Baumhefner 1989) is used to isolate the roles of IC and model uncertainty by comparing them against the MC2 categorical forecast. Statistical verification against CMC analyses valid every 6 h for the period 5–15 February helps to demonstrate the spread–error characteristics of the PERT experiment without invoking the perfect-model assumption.

To study analysis error such as could occur because of sympathetic data denial, two analyses produced independently by CMC and NCEP were used as initial conditions for the MC2 forecast runs. While these analyses had access to the same observation data, the different operational NWP centers use different data cutoff times, quality control systems, background fields, and background error covariances, all of which lead to different weights when assimilating the observations.

b. IC perturbation method

Several ensemble-generation methods have been reported in the literature, each with a different purpose. For medium-range forecasts, methods that perturb along a ''bred vector'' are intended to exploit nonlinear error growth (Toth and Kalnay 1993), and methods that perturb along a singular vector are usually optimized to provide the correct spread at 48 or 72 h (Palmer 1993). The structure of the perturbations is determined by error growth rather than an estimate of actual analysis error, and those methods 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 4–10, the ensemble-average skill can be worse for 1–3-day short-range forecasts. Conceivably, singular vectors could be constructed using an estimated analysis-error metric (i.e., arising from a DA system) rather than a forecast-error metric. The resulting perturbations may contain some information regarding analysis error, but to our knowledge this approach has not been implemented within a singular-vector ensemble prediction system.

The results of Whitaker and Loughe (1998) show that error in short-range forecasts is dominated by the statistical properties of IC error, while the error in medium-range forecasts is dominated by nonlinear error growth. Although both effects operate synergistically to reduce forecast skill, correctly sampling the IC error is more important for SREF systems than identifying regions prone to IC error growth. Hence we cannot use singular-vector (with a forecast-error metric) or bred-vector methods for our SREF experiments. The perturbation method here is akin to a Monte Carlo method, but ensemble members are chosen so that every feature in an analysis is represented equally with a limited number of ensemble members. It is based on a waveletlike transform of the analysis fields (see the appendix), and following Nychka et al. (2002) we call it the W transform.

Daley and Mayer (1986) provide a quantitative estimate of the analysis error spectrum using a then state-of-the-art model, explaining that the spectrum is largely flat. They found a maximum in analysis accuracy at wavenumbers coincident with baroclinic waves, but they suggest that no useful information exists at wavenumbers $k > 30$. 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 error may exist at different wavenumbers for each analysis. Mullen and Baumhefner (1989) exploited the results of Daley and Mayer (1986), employing a white noise perturbation strategy at $k < 30$, and saturated random phase error above.[1]

While the overall magnitude of the error may be smaller today, 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. Here, we assume that the control analysis (MRF/NCEP) is the mean of some IC distribution with an unknown shape. Perturbations were 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 (see the appendix for details).

These perturbation experiments used a model grid with $\Delta x = 100$ km, resolving waves that may be within the flat part of the error spectrum today. The perturbations targeted all resolvable scales similarly. An attempt was made to account for the possibility that daily variability may exist in the scale and location of analysis error. The goal here was to perturb different features at all scales in each ensemble member.

For one ensemble member, each analyzed 2D isobaric

[1] In this context, saturation refers to the state in which the amplitude of the perturbation is as large as the amplitude of the field itself for a particular scale. Thus, those analyzed scales are indistinguishable from a random selection from climatological data.

field of temperature, geopotential height, and U and V wind components were perturbed independently. Surface and sea surface temperature were also perturbed. This experiment made no attempt to perturb boundary conditions and, instead, relied on a very large domain to minimize boundary condition contamination of the results. This is further addressed in section 5b. The mass and momentum fields are unbalanced after perturbation, so a dynamic initialization algorithm that is built into the MC2 code ran for 8 time steps (forward and backward, for a total of 16), each of 200 s. This was an effort to remove spurious gravity waves before the experiment runs with full model physics.

c. Model physics perturbation method

Ensemble PHYS was created by using different microphysical and convective parameterization schemes for different ensemble members. The authors have observed sensitivity in a similar ensemble in real time, and sensitivity to these physics has been shown in other studies (Mullen and Baumhefner 1988; Hou et al. 2001). Additional PHYS members were generated by including the MRF forecasts in the PHYS ensemble (PHYS + MRF), and including a run where the subgrid parameterization schemes (i.e., microphysics, cumulus, turbulence, radiation, and gravity wave drag parameterizations) are turned off entirely (EXT). PHYS + MRF is perhaps the best measure of model uncertainty, because it includes many other factors besides convective and microphysical parameterization schemes. However, this ensemble is likely not representative of model *error* in any sense (indeed, little is known regarding model error). EXT is an example of extreme model uncertainty—beyond what is realistic today—but is provided here as a benchmark.

d. Spread–error characteristics

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 will be correlated. Furthermore, a forecast with spread similar to the climatological spread will yield little information on forecast confidence because it will not be related to improved or diminished accuracy. With the same assumptions, their arguments can be reversed. Namely, if an ensemble system displays proper error versus spread characteristics, it provides evidence that the initial probability density function (PDF) is partially sampled.

One way to measure the spread–error correlation is to construct a contingency table of ranked spread and error. Limiting the data used to find these statistics to the verification subdomain makes the results less dependent on boundary conditions, but retains enough points for robustness (2500). The spread is defined as in Whitaker and Loughe (1998):

$$S(x, y) = \left\{\frac{1}{N}\sum_{j=1}^{N}[\psi_j(x, y) - \overline{\psi}(x, y)]^2\right\}^{1/2}, \quad (1)$$

where S is the spread, N is the number of ensemble members (eight perturbed plus one categorical), j is an ensemble index, ψ is any variable such as geopotential height, and $\overline{\psi}$ is the ensemble mean. It is easy to see that this is simply the standard deviation of the ensemble forecast at any given grid point.

The ensemble mean error is defined as

$$E(x, y) = \{[\psi_a(x, y) - \overline{\psi}(x, y)]^2\}^{1/2}, \quad (2)$$

where E is the error and ψ_a is the value of the verifying analysis at any grid point. 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. Results are normalized by the total number of verification points to give a percentage of verification points where the spread and error fall into the same quintile. If there was no correlation between spread and error, each value in the table would be 0.2. A contingency table that indicates perfect spread–error correlation would have 1 in the diagonal and 0-everywhere else. The results of this analysis are presented in the next section.

4. Ensemble results

a. Spread–error correlation

A contingency table verifying 50-kPa geopotential height (Table 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 confident information may be gained later in the forecast. 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.

b. IC error versus model error

The ensembles show a correlation between ensemble spread and ensemble error, suggesting that this perturbation method is approximating 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. Similar to (2), the error of the jth forecast in the ensemble is

TABLE 1. Contingency tables of ranked forecast spread (rows) and ranked error (columns) for 0-, 24-, 48-, and 72-h 50.0-kPa geopotential height forecast over the entire experiment period.

0 h	0%–20%	20%–40%	40%–60%	60%–80%	80%–100%
0%–20%	0.26	0.20	0.18	0.22	0.14
20%–40%	0.22	0.21	0.22	0.19	0.15
40%–60%	0.19	0.22	0.22	0.19	0.19
60%–80%	0.18	0.19	0.19	0.20	0.24
80%–100%	0.15	0.19	0.19	0.20	0.28
24 h					
0%–20%	0.38	0.24	0.19	0.12	0.07
20%–40%	0.20	0.23	0.19	0.19	0.18
40%–60%	0.17	0.19	0.20	0.20	0.24
60%–80%	0.14	0.18	0.21	0.22	0.25
80%–100%	0.11	0.16	0.20	0.27	0.26
48 h					
0%–20%	0.32	0.32	0.22	0.09	0.05
20%–40%	0.25	0.22	0.22	0.18	0.13
40%–60%	0.16	0.16	0.19	0.24	0.26
60%–80%	0.13	0.14	0.18	0.25	0.29
80%–100%	0.14	0.16	0.18	0.24	0.27
72 h					
0%–20%	0.37	0.32	0.19	0.08	0.04
20%–40%	0.21	0.25	0.27	0.18	0.08
40%–60%	0.17	0.19	0.23	0.21	0.20
60%–80%	0.13	0.13	0.17	0.24	0.33
80%–100%	0.12	0.11	0.13	0.29	0.35

$$E_j(x, y) = \{[\psi_a(x, y) - \psi_j(x, y)]^2\}^{1/2} \quad (3)$$

and the error variance is then the variance of the distribution of all E_j across the ensemble members. It can be normalized by an estimate of the climatological variance of ψ and averaged over all verification locations. Mullen and Baumhefner (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 error for each ensemble.

Results averaged over the entire period are shown in Figs. 13a and 13b. DENY is not included because the variance of a two-member ensemble does not make sense. The PERT error variance decreases between forecast hours 0 and 12 during a period of further dynamic adjustment (Fig. 13c). Later, it shows an error growth rate that is greater than the PHYS error growth rate. The error values (Fig. 13a) are also much higher than those for PHYS. The PHYS + MRF error growth rate is larger than the PHYS throughout the forecast period, and 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). For clarity the standard deviations are not shown, but 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 h. If the PERT initialization is a realistic representation of analysis error, and the MC2 is a realistic representation of the atmosphere, one cannot claim that the differences are an artifact of the ensemble strategy.

The results suggest that through the first 66 h of model forecast during this case, initial condition uncertainty is more important than model uncertainty as measured by forecasts with both the MC2 and the MRF. In the last 6 h, 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 only valid for this particular synoptic regime and period, and to the extent that the MC2 realistically reproduces atmospheric phenomena.

The IC uncertainty and model uncertainty for the forecast initialized 0000 UTC 10 February are shown in Figs. 13c and 13d. 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. 13a). The growth rate is also smaller (Fig. 13d). The PHYS and PHYS + MRF curves for 10 February demonstrate greater values and growth rate, dominating by hour 54 with values much greater than

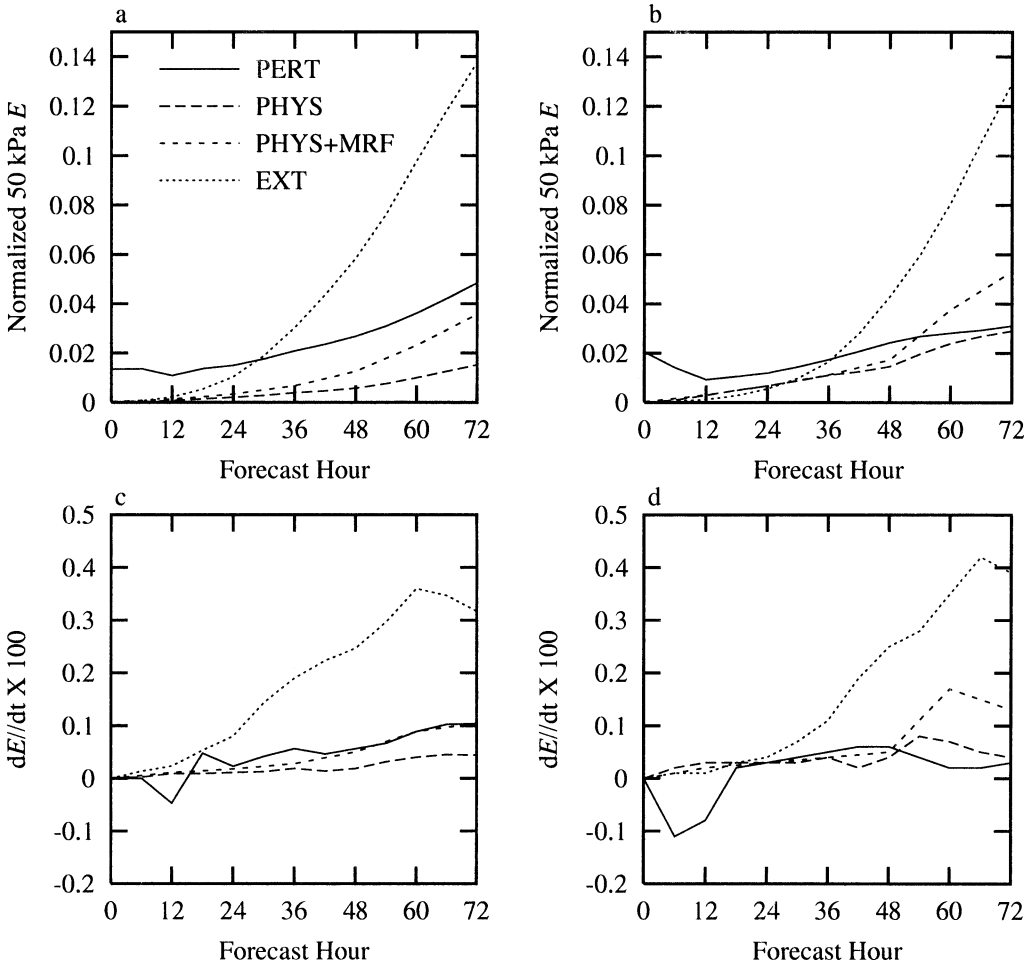


FIG. 13. Normalized 50-kPa geopotential height error variance (a) calculated for the entire period and (b) for the forecast initialized 0000 UTC 10 Feb 1999. Growth rates (c) for the entire period and (d) for the 10 Feb forecast. Results from all the experiments except DENY are shown.

the period averages (Fig. 13b). The large reduction in PERT error variance over the first 12 h 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 subdomain was not subject to dynamically sensitive modes despite the fact that the LPCs were inside the subdomain at initialization. This could occur if the storm had already reached maturity, and baroclinicity was already diminished.

Looking at the daily variability of error variances throughout the experiment period also supports this conclusion (Fig. 14). At 72 h, the IC uncertainty is by far the greatest for the forecast initialized 0000 UTC 8 February, while the model uncertainty is the greatest for the forecast initialized 0000 UTC 10 February. 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 February does not become obvious until 72 h. A particularly sensitive feature dominating the verification domain at this time could increase the spread, but the lack of a spike at 48 h for the forecast initialized 9 February suggests this is not the case. Experiment PHYS displays a tendency for relatively high error variances as early as 24 h for the forecast initialized 0000 UTC 10 February. For completeness, we also note that Figs. 13 and 14 suggest error doubling times of approximately 1.5–2 days.

Viewed this way, it appears that model uncertainty is more important than IC uncertainty for the forecast initialized 0000 UTC 10 February. That is, the uncertainty represented by the PERT experiment for this 1 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 high confidence in the forecast if the models were perfect. The large uncertainty in PHYS and PHYS + MRF suggests this is not the case, illustrating a dilemma for operational forecasters using ensemble forecasts, which

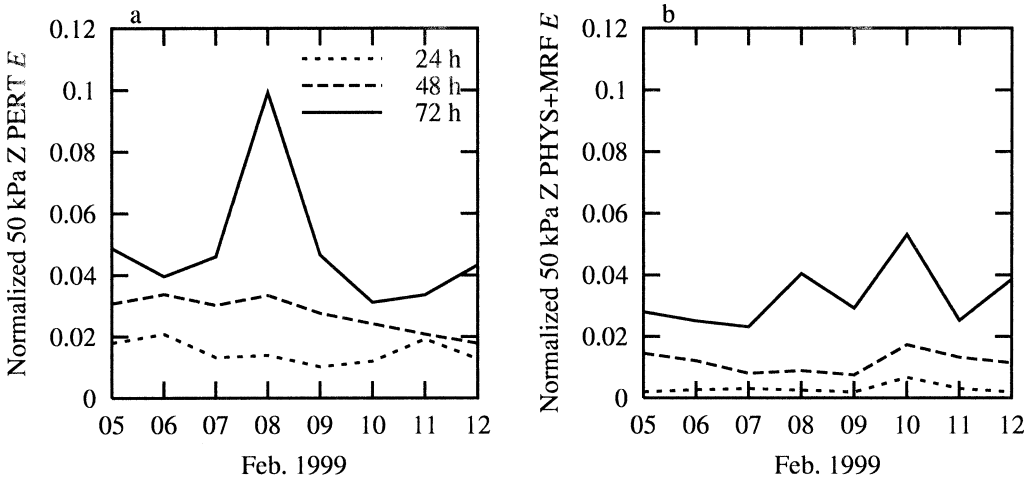


FIG. 14. Normalized 50-kPa geopotential height error variance of each ensemble forecast at 24, 48, and 72 h. (a) Results from experiment PERT, and (b) results from experiment PHYS + MRF. The date on the abcissa refers to the initialization date of each initialization, at 0000 UTC.

often include both runs with different ICs and different models.

Qualitatively, the IC and model uncertainty can be seen by tracking the low pressure center through the forecast from 0000 UTC 10 February. Figure 15a shows the PERT forecast and Fig. 15b 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 vary much more in the PERT ensemble than in the PHYS ensemble. When the MRF forecast is included, the spread at 48 h increases.

c. *Sympathetic data denial*

Although the behavior of PERT and PHYS + MRF are noticeably different between the forecast initialized 0000 UTC 10 February and for the whole experiment period, the DENY results from that day are similar to the whole experiment period (Fig. 16). If this comparison is an appropriate first-order surrogate for data denial, it suggests that any denial for the forecast period beginning 0000 UTC 10 February does not have a strong negative or positive impact on the forecast, relative to other forecasts in the experiment period. This assumption has several weaknesses because many other factors affect the analyses, including background fields and error statistics in the data assimilation systems. Substantial day-to-day variability in spread and spread growth did exist in DENY (Fig. 16b), suggesting that these factors combine to produce significant differences between analyses. Systematic error caused by poorly specified error covariances would tend to reduce daily variability in the spread and should mostly determine scale response and total error. Only part of the period is shown in Fig. 16b for comparison with Fig. 17 below.

A look at NCEP observations that were ingested by the data assimilation system does not provide any further insight. Rawinsonde and shipborne 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 more dense, and NCEP quality control flags give information about whether an observation

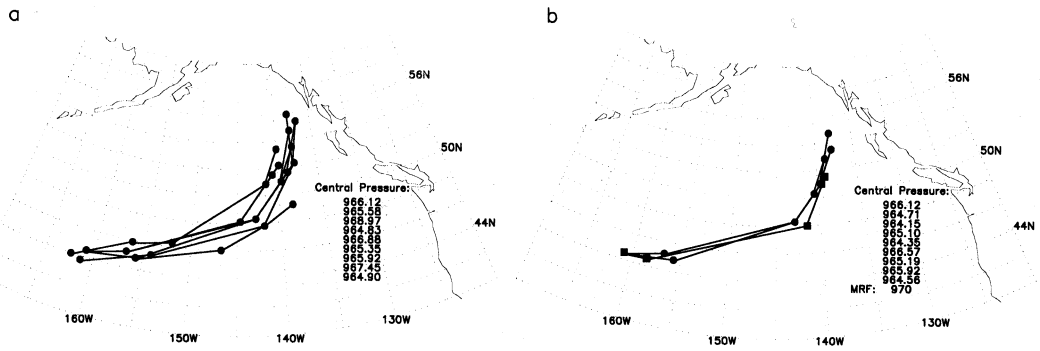


FIG. 15. (a) PERT ensemble and (b) PHYS ensemble storm tracks and 48-h central pressures from the forecast initialized 0000 UTC 10 Feb 1999. The MRF forecast is also shown in (b) as squares.

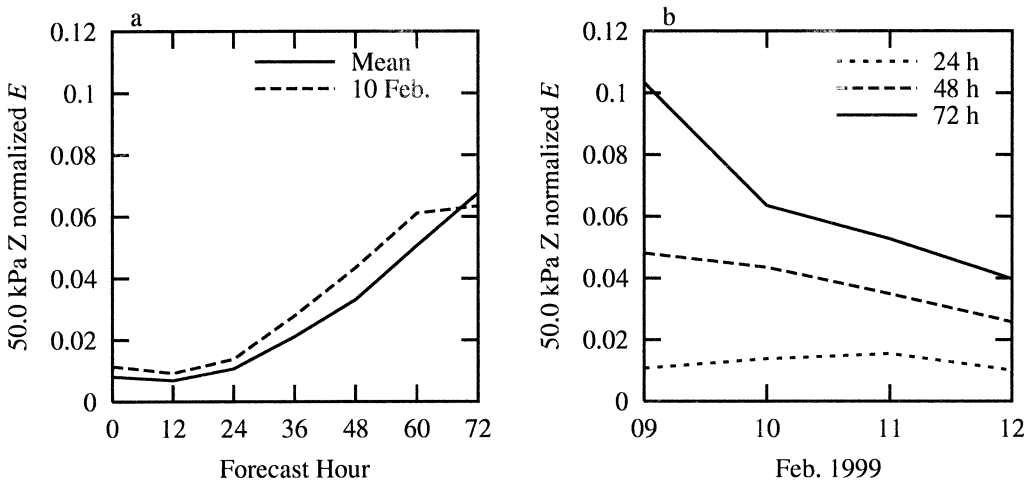


FIG. 16. (a) Ensemble spread of DENY 50-kPa geopotential height for the entire experiment period (solid), and for the forecast initialized 0000 UTC 10 Feb 1999 (dashed). (b) Same as in Fig. 13 but for expt DENY, and only including the time period corresponding to the time period in Fig. 17.

was appropriately used. A plot of the number of temperature observations from aircraft within the box bounded by 35°–60°N, and 120°–160°W shows substantial variation in observations (Fig. 17). Observations were deemed ''good'' when the data assimilation system did not explicitly reduce the weights or discard the data because of suspect quality. Within the period shown, three times with very few observations are evident—all between 0600 and 1200 UTC—in the middle of the night local time. Presumably few aircraft are flying during these hours on any day.

Comparing Fig. 17 with Fig. 16b does not suggest any relationship between aircraft observations and the differences that result in experiment DENY. Relatively large spread for the forecast initialized 0000 UTC 11 February does not correspond with a period of few aircraft observations. The same is true for the large 72-h spread for the forecast initialized 0000 UTC 9 February. Thus either experiment DENY is not an appropriate surrogate for data denial experiments, or variation in aircraft observations is due to air traffic patterns and does not have a major impact on data assimilation systems. The latter is likely for two reasons: 1) increased reliance on satellite observations over the ocean, and 2) the fact that most aircraft observations are at the cruising altitude of commercial jetliners on specific flight paths. 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.

d. The effect of boundaries

Boundary conditions that are common among all the ensemble members limit the possible spread. This ex-

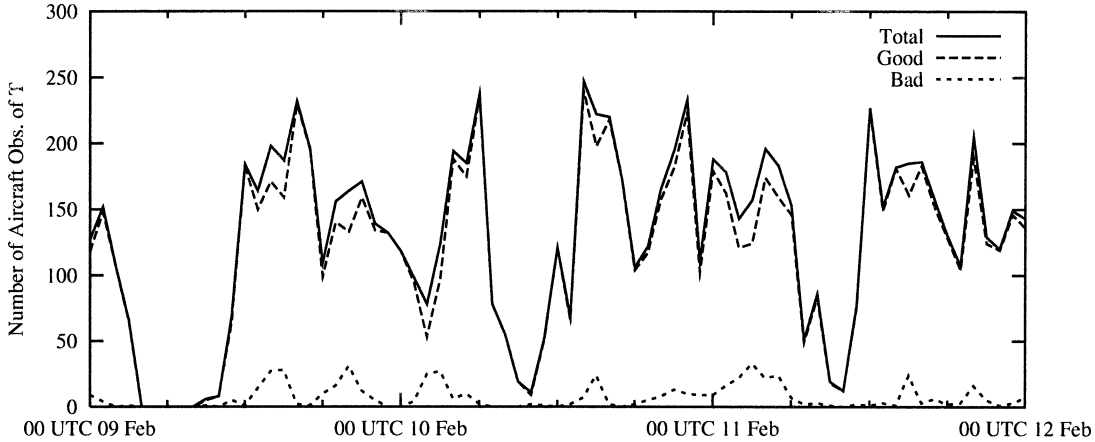


FIG. 17. Number of aircraft observations within the box bounded by 35°–60°N, and 120°–160°W. The data assimilation system at NCEP did not modify or discard the observations deemed ''good.''

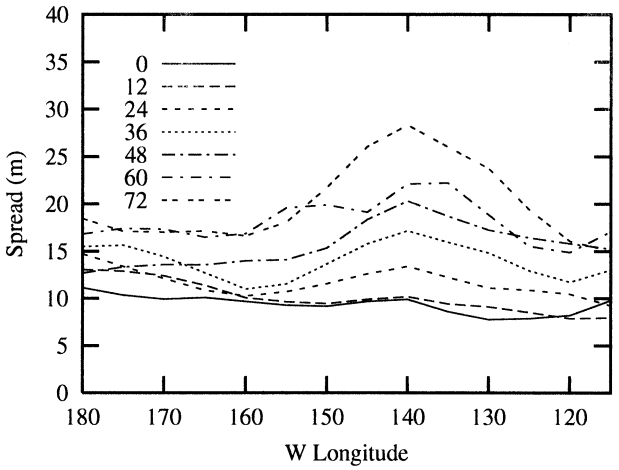


FIG. 18. Zonal-mean 50-kPa geopotential height spread vs longitude. Each curve represents a different forecast time, given in the legend. The mean is determined by averaging all values between 30° and 60°N.

periment makes no attempt to perturb boundary conditions, and the effects can be revealed by looking at the longitudinal variation of spread in the 50-kPa geopotential height field.

At each forecast time, zonal-mean spread values are calculated from 180° to 115°W in 5° bins. Note that the verification domain used in the previous sections is only a subset of the full zonal domain shown here (Fig. 12). To keep the number of values consistent, only 30°–60°N is considered. Results averaged over all the forecast cycles in the experiment are shown in Fig. 18, 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 progress; 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 the downstream (right) boundary. There is no persistent feature to which maximum spread is tied, such as a long-wave trough at 140°W that might maintain a spread maximum, and yet the characteristic shape of these curves is persistent through the experiment period.

Although 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. Though the verification domain used above is mostly around the peak in Fig. 18, the boundaries clearly limit the spread for experiment PERT, making it a lower-bound estimate of IC uncertainty.

For comparison, we note that Langland et al. (2002) estimated zonal error propagation at approximately 20 m s^{-1}, which was more than the jet-level phase velocity (5–10 m s^{-1}) of wind anomalies in their case study. This is suggestive of downstream error propagation associated with downstream development, and that lateral boundaries are more limiting than phase speeds would indicate.

5. Conclusions

A poor forecast for a landfalling front associated with a North Pacific midlatitude cyclone prompted an investigation into the cause of the poor numerical model guidance from models initialized at 0000 UTC 10 February 1999. A qualitative examination shows that differences between two operational analysis–forecast pairs, as well as analysis and forecast error in both, are easily identified. The analyses did not properly locate the developing storm. Short-range forecasts did not maintain a secondary, trailing, low pressure center and did not adequately amplify the downstream ridge. The resulting public forecast for significant weather was a false alarm for a major metropolitan area.

Perturbations to discrete, waveletlike transform coefficients were used to generate an ensemble that shows some correlation between ensemble spread and forecast error. Ensembles were produced from control initial conditions for 72-h forecasts initialized 0000 UTC 5–12 February 1999 and run with the MC2 model (experiment PERT). The spread–error characteristics indicate that the ensembles show promise for approximating the IC PDF. Here they are interpreted as giving an estimate of IC uncertainty.

For comparison, another ensemble (PHYS) was generated by changing microphysics and cumulus parameterization schemes in the MC2. It can be interpreted as one estimate of model uncertainty. A better estimate of model uncertainty was proposed by including the MRF forecast in the ensemble (PHYS + MRF).

Comparing the normalized error variances of each experiment suggests that, over the entire experiment period, the IC uncertainty is significantly greater than the model uncertainty through the first 66 h of forecast time, though the spread is limited by the boundaries of the domain. The one forecast initialized at 0000 UTC 10 February shows different characteristics, with model uncertainty dominating by hour 54. This was perhaps the critical forecast providing numerical guidance to forecasters of a frontal passage that never materialized. Results of experiments that attempted to determine any impact of sympathetic data denial were inconclusive.

This study does not rigorously take model *error* into account and it does not consider the likely scenario that the ensemble does not bracket the verifying analyses. A discussion of these issues is presented in Orrell et al. (2001).

Although generalization of these results is unwarranted, this case study demonstrates that large problems still exist with both the initialization of NWP models

over the Pacific and also with model treatment of developing storms over the Pacific. A similar experiment, over a much longer time period and covering many different synoptic regimes, would provide a better assessment of where resources could best be placed to improve NWP forecasts for western North America.

Acknowledgments. The authors gratefully acknowledge Robert Benoit, Michel Desgagné, and others at Rechereche en Prévision Numérique (RPN) for their contributions to the development and maintenance of the MC2; the Canadian Meteorological Centre (CMC) for providing model analyses and forecasts; and NCEP for observation and model data. Henryk Modzelewski's computing support is also recognized. The thoughtful comments of two anonymous reviewers were critical to improving the readability and focus of the manuscript.

APPENDIX

Details of Perturbation Algorithm

To perturb each 2D field $\boldsymbol{\Psi}$, a 2D discrete $\mathbf{W}$ transform (Nychka et al. 2002) first decomposes the field, providing spectral information from gridpoint scales $(N/2)\Delta x$ to $8\Delta x$, recursively stepping by a factor of 2:

$$\mathbf{W} = \mathbf{T}^{8\times 8} \cdots \mathbf{T}^{M/2\times N/2}(\boldsymbol{\Psi}^{M\times N})\mathbf{T}^{M/2\times N/2} \cdots \mathbf{T}^{8\times 8^t}. \quad \text{(A1)}$$

Matrix $\mathbf{W}$ stores the transform coefficients, $\mathbf{T}$ contains both the scale and basis of the wavelet, $\boldsymbol{\Psi}$ is an analysis field, $N = M = 128$ is the dimension of the grid here, and t denotes transpose.

The basis and scale functions were the same as that used in Nychka et al. (2002) to model climatological precipitation distributions. The transform is similar to a discrete wavelet transform, where the basis function of Nychka et al. (2002) 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.

For each scale, the $\mathbf{W}$ 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 $n = 8$ ensemble members for each forecast, so $n/2 = 4$ quantiles are created (quartiles), each of which is used to generate a positive and negative perturbation, as is described next.

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 1. Multiplying them by a number less than one decreases them, creating ensemble member 2. 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 1 that defines the magnitude of perturbation to the wavelet coefficients w. This process is continued for the three other quartiles, resulting in eight ensemble members.

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. If a forecast does not diverge from the control forecast, then it is not sensitive to errors in the perturbed features or at the perturbed scale. An inverse transform builds the perturbed physical fields using the perturbed coefficients. Figure A1a shows an example of perturbed and unperturbed MRF 50-kPa geopotential height field.

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

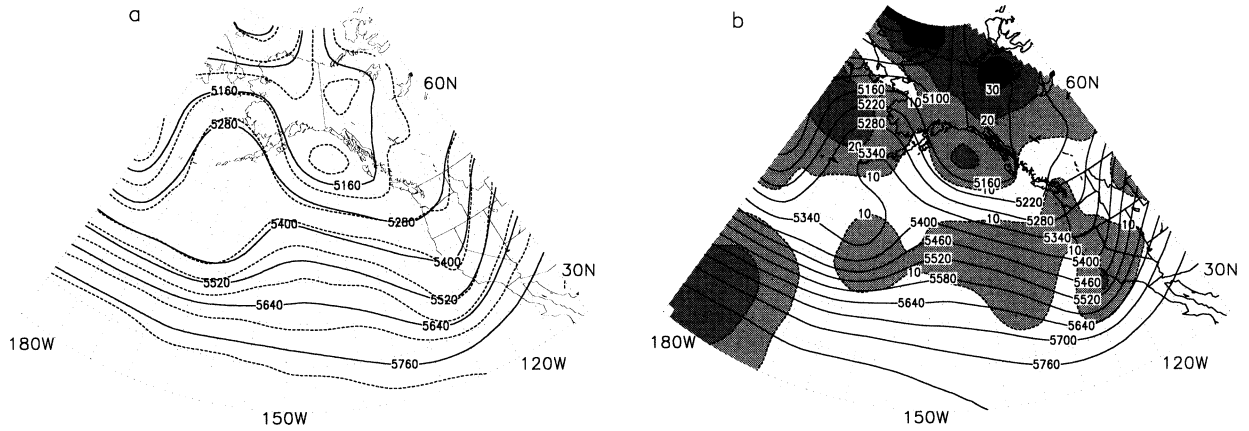


FIG. A1. Dynamically initialized perturbations of the 50-kPa geopotential height, valid 0000 UTC 10 Feb 1999. (a) The categorical initialization (solid) and ensemble member 1 (dashed), and (b) the ensemble mean (solid lines) and the rms spread (shaded, 10-m contour intervals).

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.

After generating the perturbations, the 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. At this point, the rms of the perturbations is smaller than before the initialization. For example, the 50-kPa-height rms ranges from zero to 30 m in Fig. A1b. Note 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, explicitly projecting the perturbations onto vertical modes. 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.

REFERENCES

Benoit, R., M. Desgagné, P. Pellerin, S. Pellerin, Y. Chartier, and S. Desjardins, 1997: The Canadian MC2: A semi-Lagrangian, semi-implicit wide-band atmospheric model suited for finescale process studies and simulation. *Mon. Wea. Rev.,* **125,** 2382–2415.

Côté, J., J.-G. Desmarais, S. Gravel, A. Méthot, A. Patoine, M. Roch, and A. Staniforth, 1998: The operational CMC–MRB Global Environmental Multiscale (GEM) model. Part I: Design considerations and formulation. *Mon. Wea. Rev.,* **126,** 1373–1395.

Daley, R., and T. Mayer, 1986: Estimates of global analysis error from the global weather experiment observational network. *Mon. Wea. Rev.,* **114,** 1642–1653.

Hou, D., E. Kalnay, and K. K. Drogemeier, 2001: Objective verification of the SAMEX '98 ensemble forecasts. *Mon. Wea. Rev.,* **129,** 73–91.

Langland, R. H., M. A. Shapiro, and R. Gelaro, 2002: Initial condition sensitivity and error growth in forecasts of the 25 January 2000 East Coast snowstorm. *Mon. Wea. Rev.,* **130,** 957–974.

Lorenz, E., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.,* **20,** 130–141.

Mullen, S. L., and D. P. Baumhefner, 1988: Sensitivity of numerical simulations of explosive oceanic cyclogenesis to changes in physics parameterizations. *Mon. Wea. Rev.,* **116,** 2289–2339.

——, and ——, 1989: The impact of initial condition uncertainty on numerical simulations of large-scale explosive cyclogenesis. *Mon. Wea. Rev.,* **117,** 2800–2821.

Nychka, D., C. Wilke, and J. A. Royle, 2002: Large spatial prediction problems and nonstationary random fields. *Stat. Model.,* in press.

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

Palmer, T. N., 1993: Extended-range atmospheric prediction and the Lorenz model. *Bull. Amer. Meteor. Soc.,* **74,** 49–64.

Sanders, F., and J. R. Gyakum, 1980: Synoptic-dynamic climatology of the ''bomb.'' *Mon. Wea. Rev.,* **108,** 1589–1606.

Shapiro, M. A., H. Wernli, N. A. Bond, and R. Langland, 2001: The influence of the 1997–99 El Niño Southern Oscillation on extratropical baroclinic life cycles over the eastern North Pacific. *Quart. J. Roy. Meteor. Soc.,* **127,** 331–342.

Smith, L. A., C. Ziehman, and K. Fraedrich, 1999: Uncertainty dynamics and predictabillity in chaotic systems. *Quart. J. Roy. Meteor. Soc.,* **125,** 2855–2886.

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

Trenberth, K. E., 1978: On the interpretation of the diagnostic quasi-geostrophic omega equation. *Mon. Wea. Rev.,* **106,** 131–136.

Tribbia, J. J., and D. P. Baumhefner, 1988: The reliability of improvements in deterministic short-range forecasts in the presence of initial state and modeling deficiencies. *Mon. Wea. Rev.,* **116,** 2276–2288.

Uccellini, L. W., D. Keyser, K. F. Brill, and C. H. Walsh, 1985: The Presidents' Day cyclone of 18–19 February 1979: Influence of upstream trough amplification and associated tropopause folding on rapid cyclogenesis. *Mon. Wea. Rev.,* **113,** 962–988.

Whitaker, J. S., and A. F. Loughe, 1998: The relationship between ensemble spread and ensemble mean skill. *Mon. Wea. Rev.,* **126,** 3292–3302.