The Relationship between Ensemble Spread and Ensemble Mean Skill

JEFFREY S. WHITAKER AND ANDREW F. LOUGHE

NOAA-CIRES Climate Diagnostics Center, Boulder, Colorado

(Manuscript received 29 September 1997, in final form 23 February 1998)

ABSTRACT

Statistical considerations suggest that 1) even for a perfect ensemble (one in which all sources of forecast error are sampled correctly) there need not be a high correlation between spread and skill, 2) the correlation between spread and skill should be larger where the day-to-day variability of spread is large, and 3) the spread is likely to be most useful as a predictor of skill when it is ''extreme,'' that is, when it is either very large or very small compared to its climatological mean value. The authors investigate the relationship between spread and skill in an operational setting by analyzing ensemble predictions produced by the National Centers for Environmental Prediction. The geographical dependence of the spread–skill relationship is found to be related to the geographical dependence of day-to-day variability of spread. Dynamical mechanisms for spread variability are investigated using a linear quasigeostrophic model. Problems associated with the sample size needed to define what constitutes an extreme value of spread at a given location are discussed.

1. Introduction

In general, the mean of an ensemble of forecasts will, on average, have a smaller error than the mean error of any of the individual forecasts comprising the ensemble (Leith 1974; Murphy 1988). Perhaps the simplest, and most widely used, method of utilizing an ensemble forecast is to treat the ensemble mean as a single forecast, representing the best available estimate of the future state of the atmosphere. As Leith (1974) pointed out, much, but not all, of the benefit realized by ensemble averaging can be achieved using a single forecast in conjunction with statistical correction techniques that utilize previous forecast verifications. However, an ensemble forecast provides an estimate of the forecast probability distribution of model variables, given an estimate of the probability distribution of analysis errors. Assuming that the forecast probability distribution is unimodal, variations in the width of the distribution from forecast to forecast may be related to the skill of the mean. The simplest measure of the width of the forecast probability distribution is the second moment of the ensemble, or the ensemble spread. Using this information, the utility of the ensemble mean as a forecast product can be significantly enhanced.

The utility of ensemble spread as a predictor of ensemble mean skill has traditionally been measured in terms of a linear correlation. In general, using operational forecast models, the correlation between spread and skill has been found to be positive for forecast lead times of less than a week or so (Kalnay and Dalcher 1987; Murphy 1988; Buizza 1997). Even in idealized ''perfect model'' experiments, in which the forecast model has no systematic biases, the correlation between spread and skill has been somewhat disappointing, generally less than about 0.5 (Barker 1991). This can be explained using a simple stochastic model of the spread–skill relationship used by Houtekamer (1993), based upon the statistical model of forecast error proposed by Kruizinga and Kok (1988). At a given grid point, it is assumed that the spread is a random variable with a lognormal distribution, that is,

$$\ln S = N(\ln S_{\mathrm{M}}, \beta), \qquad (1)$$

where S_M is the mean value of spread, β is the standard deviation of $\ln S$, and $N(\alpha, \gamma)$ denotes a random number drawn from a Gaussian distribution with a mean of α and standard deviation γ. The particular distribution chosen for S is not important; a lognormal distribution was chosen because it permits a simple analytical relationship for spread–skill correlation to be derived. If the ensemble forecast system is perfect, that is, the underlying distribution of analysis and model error is known and correctly sampled, and if the probability distribution of ensemble mean error E is Gaussian,[1] then E is completely determined by S. Under these assumptions $E = N(0, S)$, and the correlation between spread

Corresponding author address: Dr. Jeffrey S. Whitaker, NOAA-CIRES Climate Diagnostics Center, University of Colorado, Campus Box 449, Boulder, CO 80309.
E-mail: jsw@cdc.noaa.gov

[1] Strictly speaking, the ensemble mean error E need only be Gaussian after a suitable transformation.

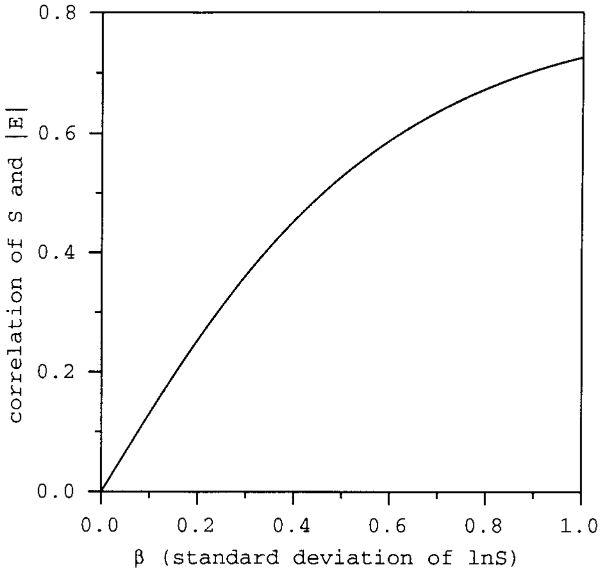

FIG. 1. Correlation of S and $|E|$ as a function of β, for the idealized statistical model given by (1) and $E = N(0, S)$.

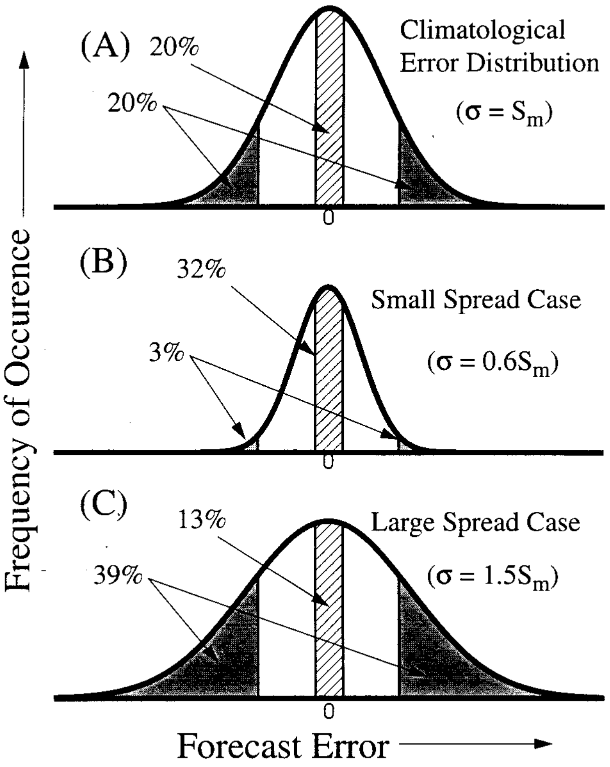

FIG. 2. Schematic Gaussian forecast error probability distributions. The area enclosed by the shaded (hatched) areas represents the cumulative probability that the forecast error will be among the largest (smallest) 20% of all cases ever observed. Here, (A) represents the climatological distribution of forecast error (for which the standard deviation is S_M). Also, (B) represents a situation in which the ensemble spread (and hence the width of the forecast error distribution in a perfect ensemble) is 60% of the climatological mean value (S_M). Finally, (C) represents a case in which the ensemble spread is 1.5 times the climatological mean value.

and skill (measured by $|E|$) may be expressed analytically [see Eq. (33) of Houtekamer 1993]. Figure 1 shows the correlation of S and $|E|$ as a function of β, which measures the temporal variability of spread at the grid point under consideration. When $\beta = 0$, the spread is constant, and the error is a random draw from a fixed distribution, so the correlation must be zero. As the temporal variability of spread increases, so does the correlation, asymptoting to a value of about 0.75 for large β. This simple example illustrates that even for a perfect ensemble, the correlation between spread and skill need not be large, and the magnitude of the correlation depends upon the day-to-day variability of spread.

Using contingency tables, Houtekamer (1993) also showed that the spread has the most predictive value when it is ‘‘extreme,’’ that is, when it is very large or small compared to its mean value. This can be understood with the aid of the schematic probability distributions shown in Fig. 2. The area enclosed by the hatched and shaded areas represent the cumulative probability that the forecast error will be among the largest (shaded) or smallest (hatched) 20% of all cases ever observed. The schematic Gaussian in Fig. 2a represents the climatological distribution of forecast error (for which the standard deviation is S_M), so the shaded and hatched areas each represent 20% of all observed cases. Figure 2b (2c) represents situations in which the ensemble spread is small (large) compared with its climatological mean value. For the small (large) spread case, the probability that the forecast error will be among the largest 20% observed is 3% (39%). Conversely, for the small (large) spread case, the probability that the forecast error will be among the smallest 20% observed is 32% (13%). Essentially, the predictability of skill depends upon how much the probability distribution of forecast error deviates from the climatological distribution. Thus, the more the spread departs from its climatological mean value, the more useful it is as a predictor of skill. When the spread happens to be close to the climatological mean value, then it has very little predictive value, since the forecast error is then essentially just a random draw from the climatological distribution. Therefore, in order to optimally exploit the relationship between spread and skill, it may be crucial to know the underlying climatological distribution of spread for a given ensemble configuration.

In this study, we examine output from operational ensemble predictions produced at the National Centers for Environmental Prediction (NCEP) to see if the relationship between spread and skill conforms with that expected from the simple statistical considerations outlined above. We are particularly interested in understanding the relationship between the day-to-day variability in ensemble spread and the geographical dependence of skill predictability in the operational system. Section 2 describes the dataset and analysis methods used. In section 3 we present results for two winters of operational predictions. A strong relationship is found between the day-to-day variability of spread and the geographical dependence of spread–skill correlation. Contingency tables of spread and error confirm that spread is only useful as a predictor of skill when it is extreme or very different from its mean value. Large differences are observed in the geographical dependence of the spread-skill relationship between the two winters studied, and these differences appear to be related to interannual variations in the geographical dependence of spread variability. In order to overcome problems in interpretation associated with the short data record of operational ensemble predictions, in section 4 we utilize an idealized model of spread variability based upon a linear quasigeostrophic model. The idealized model is used to estimate the climatological-mean geographical dependence of skill predictability and to understand the dynamical mechanisms responsible for this geographical dependence. Section 5 summarizes the results and their implications for operational skill prediction.

2. Datasets and analysis procedure

Gridded 250-hPa streamfunction data from the operational ensemble prediction system at NCEP are analyzed. Forecasts verifying during the periods 15 November 1995 to 15 March 1996 and 15 November 1996 to 15 March 1997 are used.[2] Details regarding the implementation of ensemble prediction at NCEP are available in Toth and Kalnay (1993). Briefly, the NCEP ensemble consists of 12 members initialized at 0000 UTC (one high-resolution run, one low-resolution control, and 10 perturbations of the low-resolution control) and five members initialized at 1200 UTC (a control run and four perturbations of the control). The five runs from 1200 UTC are included to yield an ensemble size of 17, although they are not weighted to account for the fact that they are 12 h older than the members initialized at 0000 UTC. The spatial resolution of the NCEP ensemble members is T62L18, except the 0000 UTC high-resolution run (which is T126L18 for the first 7 days of integration and T62L18 thereafter) and the 12 UTC control (which is T126L18 for the first 3 days of integration and T62L18 thereafter). The NCEP data are spectrally truncated to T35 resolution and output to a 2.5° lat–long grid. NCEP–NCAR (National Center for Atmospheric Research) reanalyses (Kalnay et al. 1996) are used for verification.

The ensemble mean error is defined as a root-mean-square (rms) distance between the analyzed and forecast fields, that is,

$$E(\lambda, \phi) = [(\psi_a(\lambda, \phi) - \overline{\psi}(\lambda, \phi))^2]^{1/2}, \qquad (2)$$

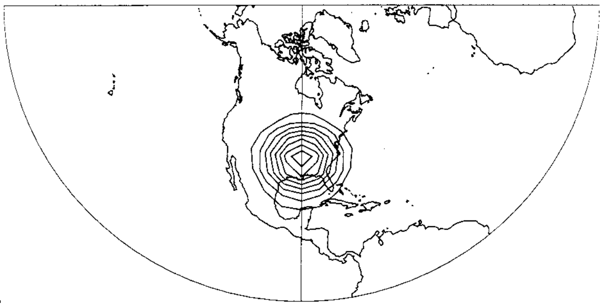

FIG. 3. Gridpoint filter weights for Gaussian spectral filter used to smooth S and E, for a selected point over North America. Contour interval is 1.8×10^{-3}.

where

$$\overline{\psi}(\lambda, \phi) = \frac{1}{N}\sum_{j=1}^{N} \psi_j \qquad (3)$$

is the ensemble average of N members and ψ_a is the verifying analysis. The spread is defined to be the average rms distance between the ensemble members and the ensemble mean, that is,

$$S(\lambda, \phi) = \left\{\frac{1}{N}\sum_{j=1}^{N} [\psi_j(\lambda, \phi) - \overline{\psi}(\lambda, \phi)]^2\right\}^{1/2}. \qquad (4)$$

Several investigators have used the anomaly correlation as a measure of spread and skill (Wobus and Kalnay 1995; Buizza 1997). We have chosen an rms definition of spread and skill for two reasons. First, since we will find it useful later on to interpret the day-to-day variability of spread in terms of the day-to-day variability of the "instability" of the atmosphere (that is, the day-to-day variability of the rate of growth of small perturbations), it is convenient to use a measure of spread that is quadratic and can be used as a measure of perturbation growth. Second, as noted by Arpe et al. (1985), Branstator (1986), Wobus and Kalnay (1995), and others, anomaly correlation is positively correlated with anomaly amplitude. Thus, as pointed out by Palmer and Tibaldi (1988), correlations between spread and skill using anomaly correlation may arise, in part, because spread and skill are mutually correlated with the magnitude of the forecast anomaly.

The relationship between error and spread is usually defined in terms of a spatial average, either for an entire hemisphere (Buizza 1997) or for specific geographical regions within a hemisphere (Wobus and Kalnay 1995). Rather than defining specific regions, we instead smooth both error and spread using a Gaussian spectral filter (Sardeshmukh and Hoskins 1984) of the form $\exp[-(n/12)^2]$, where n is the total wavenumber in spherical harmonic space. This is equivalent to averaging E and S over a circular region with a radius of about 1000 km. The geographical distribution of the filter weights in gridpoint space for a point centered over North America is shown in Fig. 3.

[2] Operational NCEP ensemble forecasts have been archived at the NOAA Climate Diagnostic Center since 1 November 1996 and are available from the authors upon request.

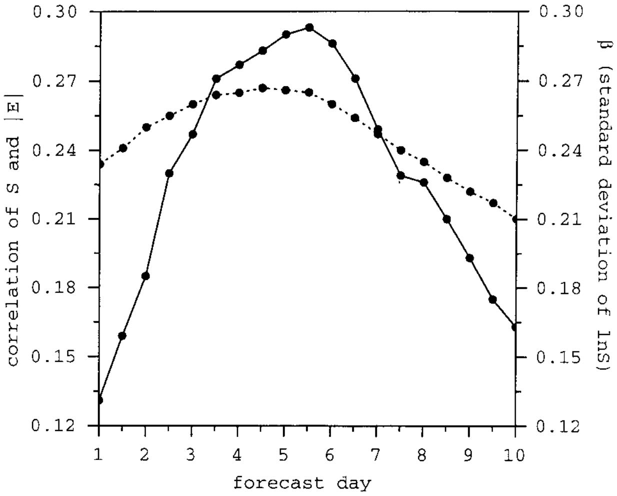

FIG. 4. Hemispheric mean values of spread/error correlation (solid) and β (dotted) as a function of forecast lead time. Average is computed for all grid points poleward of 20°N.

3. Spread/skill relationships in the NCEP operational ensemble

Figure 4 shows some hemispheric measures of the relationship between spread and error averaged over both seasons (comprising 203 forecasts). For brevity, we present results for the Northern Hemsiphere (poleward of 20°). Since the NCEP ensemble only attempts to sample analysis error, the relationship between spread and error is expected to be weaker in the summer hemisphere and in the Tropics, where model errors are expected to be relatively more important. The spread/error correlation is computed as a spatial average of the temporal correlations at each grid point. The parameter β is a measure of the day-to-day variability of spread and is simply the Northern Hemisphere average of the standard deviation of ln$S$ at each grid point.

The correlation between spread and skill in the NCEP ensemble peaks at a value of about 0.3 at day 5, and decreases to about 0.16 at day 10. Parameter β also peaks at about day 5 and decreases thereafter. Both β and spread/error correlation must decrease to zero at long forecast ranges, since the ensemble distribution approaches the climatological distribution of the model, which does not change from day to day. When the forecast distribution (and hence the spread) does not change from day to day (the $\beta = 0$ case shown in Fig. 1), the forecast error is simply a random draw from a fixed distribution.

For short forecast lead times, there is significant day-to-day variability of spread in the NCEP ensemble, but the correlation between spread and skill is low. In a perfect ensemble system, spread variability, and hence spread/error correlations, can arise from one of two sources: 1) day-to-day variations in analysis error amplitude, and 2) day-to-day variations in the growth rate of initial perturbations associated with variations in atmospheric instability. For very short forecast times, the former is likely to dominate, while the latter may dominate for longer forecast times. Figure 4 shows that although the ‘‘breeding method’’ (Toth and Kalnay 1993) used to generate initial perturbations for the NCEP ensemble does yield perturbations whose amplitude varies from day to day, these variations are not well correlated with day-to-day variations of short-range forecast error. Therefore, it is likely that the bred perturbations are not accurately sampling day-to-day variations in analysis error, and spread/error correlations in the NCEP ensemble are primarily associated with day-to-day variations in atmospheric instability. Since this mechanism cannot produce spread variability for short forecast times, and for long forecast times the spread-skill correlation must approach zero, the correlation between spread and skill must peak in the medium range.

TABLE 1. Contingency table of spread and error for 5-day forecasts. The entries in the table are the joint probability of obtaining the error and spread values in the indicated quintile. The columns are spread quintiles and the rows are error quintiles.

	0%–20%	20%–40%	40%–60%	60%–80%	80%–100%
0%–20%	0.35	0.24	0.19	0.14	0.09
20%–40%	0.23	0.23	0.22	0.19	0.14
40%–60%	0.18	0.21	0.22	0.21	0.18
60%–80%	0.15	0.18	0.20	0.24	0.23
80%–100%	0.09	0.14	0.17	0.23	0.36

Table 1 is a contingency table of E and S for 5-day forecasts, the forecast range for which the relationship between E and S is strongest. The entries in the table are the joint probability of obtaining the error and spread values in the indicated quintile. If there were no relationship between E and S, that is, if the correlation were zero, all entries in the table would be 0.2. If there were a perfect linear relationship, that is, if the correlation were unity, all the diagonal entries would be one and the off-diagonals would be zero. Most of the entries in the table are not very different from 0.2, except at the

TABLE 2. Joint probability that spread and error are in the top quintile (i.e., in the top 20% of all values ever realized) as a function of β for the stochastic model given by (1) and $E = N(0, S)$.

β	Probability
0.02	0.21
0.1	0.26
0.2	0.33
0.26	0.36
0.3	0.38
0.4	0.43
0.5	0.48
0.6	0.52
0.7	0.55
0.8	0.58
0.9	0.61
1.0	0.63
5.0	0.89

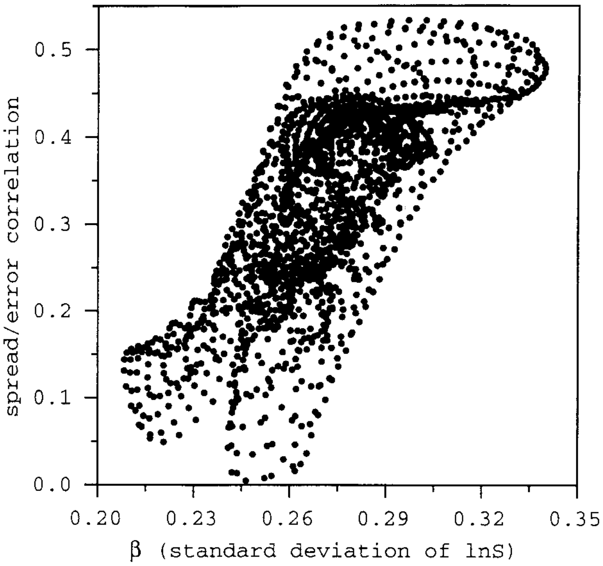

FIG. 5. Scatterplot of spread/error correlation versus β for 5-day forecasts, using all grid points poleward of 20°N.

corners. For example, if the spread is in the lowest quintile, there is a 3.5 times higher probability of the error is being in the lowest, rather than the highest, quintile. Therefore, as expected from simple statistical considerations, spread is much more useful as a predictor of skill when it is extreme. From this table, an optimist may conclude that the spread is a much better predictor of skill, when it is extreme, than the low linear correlations shown in Fig. 4 would suggest. However, a pessimist may also conclude that more than half the time, the spread is practically useless as a predictor of skill. Clearly, in order to maximize the practical utility of spread as a skill predictor, one must know the underlying climatological distribution of spread for a given ensemble forecasting system, so that what constitutes an extreme value of spread at a given location can be recognized. Using the simple stochastic model described in the introduction, contingency tables like that shown in Table 1 can be created for different values of the β parameter. Table 2 (which is similar to Table 4 of Houtekamer 1993) shows the probability that the forecast error is in the top quintile given that the spread is in the top quintile, as a function of β for the stochastic model. The deviation of the probability from 0.2 is a measure of the ability of the spread to identify bad forecasts. As the spread variability (measured by β) increases, so does the probability that forecast will be bad if the spread is large. This suggests that operational skill prediction using ensemble spread will be most useful in those geographical regions where the spread variability is large.

In accordance with the simple stochastic model, spread/error correlations in the NCEP ensemble are higher where the day-to-day variability of spread is larger. This is illustrated by Fig. 5, which is a scatterplot of spread/error correlation and β for 5-day forecasts, using all Northern Hemisphere grid points. The correlation is 0.63, so that about 36% of the spatial variation in spread/error correlation can be accounted for by spatial variations of β. Maps of spread/error correlation and β for 5-day forecasts are shown in Fig. 6. Both fields are maximum near Alaska, with a secondary maximum over Europe. There is a general tendency for spread/

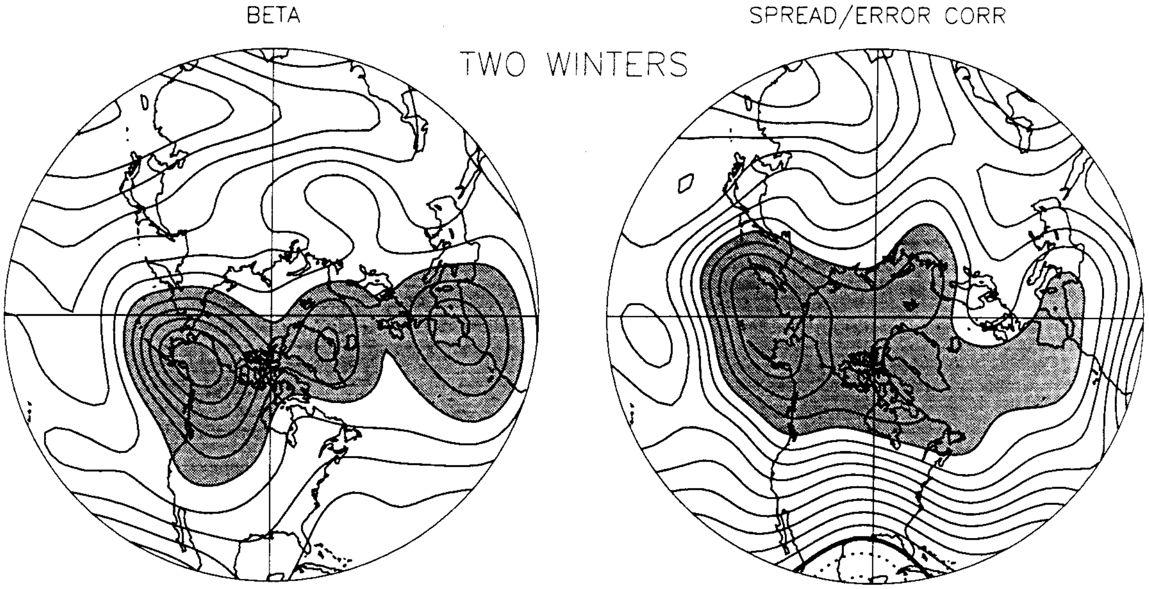

FIG. 6. Maps of β and spread/error correlation for 213 5-day forecasts made during two winter seasons. Contour interval for β is 0.01, with values greater than 0.28 shaded. Contour interval for correlation is 0.04 with values greater than 0.36 shaded. Negative contours are dashed, and the zero contour is thicker.

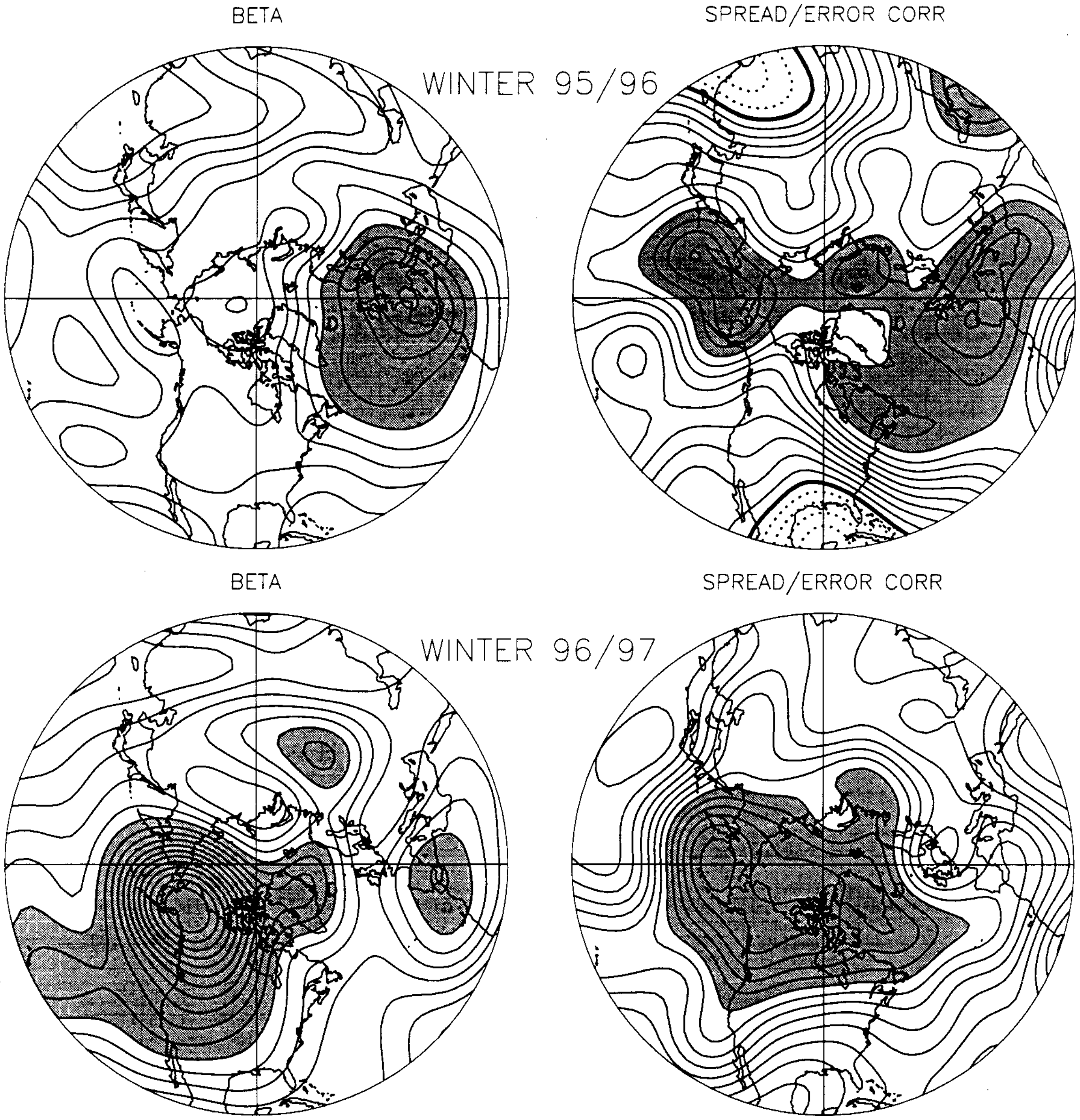

FIG. 7. As in Fig. 6, but for the two winter seasons separately.

error correlations and β to increase with latitude. However, as Fig. 7 shows, there is considerable interannual variability in both fields. Given the short data record, and the large degree of interannual variability in the geographical patterns of spread/error correlation and β, it is difficult to assess what regions possess the highest skill predictability. The short data record and limited vertical resolution of the available ensemble output also make if difficult to diagnose the dynamical mechanisms responsible for the geographical dependence of skill predictability. A long record of ensemble forecasts with a fixed ensemble configuration is needed to address these issues. Since such a dataset is not likely to be available in the near future, we have designed an idealized model of spread variability for this purpose. In the next section, we use this model to estimate the climatological mean patterns of spread variability and to investigate the dynamical processes responsible for that variability.

4. Understanding spread variability with a linear quasigeostrophic model

The dynamical model used in this study is the so-called Lorenz-P model (Lorenz 1960). The level of approximation is consistent with quasigeostrophy, but the retention of the spherical metric terms and the full variation of the Coriolis parameter distinguish it from the traditional β-plane quasigeostophic (QG) model and the

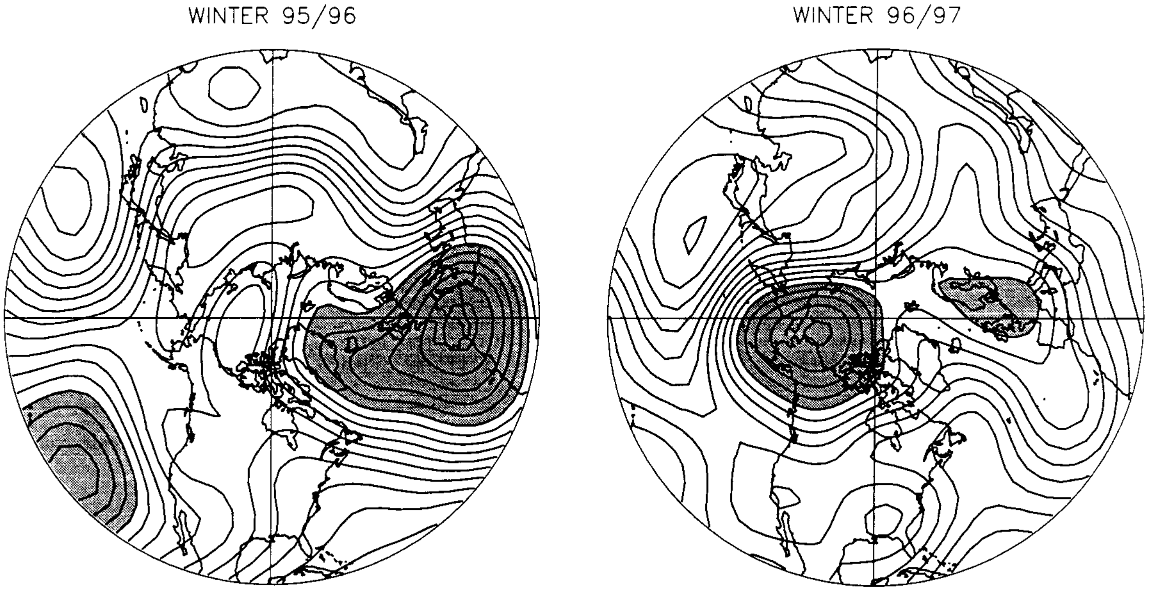

FIG. 8. Maps of β (computed using 300-hPa streamfunction) for two winter seasons, simulated at 3 days with the five-level linear QG model. Contour interval is 0.01, with values greater than 0.28 shaded.

spherical extension of the β-plane QG model used by Marshall and Molteni (1993). For reference, the vertically discretized equations and parameter settings are given in the appendix. For the results presented here, five vertical levels and a horizontal resolution of T31 are used.

a. Experimental design

We assume that day-to-day variability of spread in the NCEP ensemble is associated primarily with day-to-day variability in the growth of small perturbations and that day-to-day variations in the analysis error distribution are either unimportant or, more likely, not accurately sampled. In addition, we assume that the day-to-day variability of spread can be modeled using linear perturbation dynamics, at least for short forecast ranges. In other words, the evolution of spread in the NCEP ensemble can be simulated by a tangent linear model, linearized about the control forecast. This is supported by the fact that maps of β computed from the NCEP ensemble for 3-day forecasts (at which point the perturbation dynamics is just entering the nonlinear regime) look qualitatively similar to the 5-day forecast result shown in Fig. 6. Finally, we assume that linearizing about short segments of analyzed fields will yield results similar to linearizing about short global model forecast trajectories. This assumption would only be violated if the patterns of spread variability shown in Figs. 6 and 7 were somehow related to the evolution of *errors* within the control forecast and not to those aspects of the control forecast trajectory that were actually observed.

The five-level QG model is linearized about 3-day segments of 6-hourly NCEP reanalyses for 21 ‘‘winters’’ (the 121-day period starting on 15 November). Only the rotational wind field is needed for the basic state in the QG model. The 6-hourly analyses are interpolated linearly in time to the model time step (3 h). One hundred and twenty-one ensemble integrations, all starting at 0000 UTC, are performed for each winter, with an ensemble size of 10. The object of these integrations is to compute the 3-day forecast covariance matrix ($\mathbf{C}_3$), given the covariance matrix of analysis error ($\mathbf{C}_0$). The ensemble spread is given by the diagonal elements of $\mathbf{C}_3$. If $\mathbf{G}$ is the 3-day propagator of the tangent linear model, then $\mathbf{C}_3 = \mathbf{G}\mathbf{C}_0\mathbf{G}^T$. In principle, the matrix $\mathbf{G}$, and hence $\mathbf{C}_3$, could be computed directly, but for a model of this size a Monte Carlo approach to estimating $\mathbf{C}_3$ is much more efficient. Since we are computing spread variability associated with day-to-day variations in the growth of small perturbations, that is, day-to-day variations in $\mathbf{G}$, $\mathbf{C}_0$ is held fixed.

b. QG model results

Analysis error covariance is notoriously difficult to estimate accurately (Lonnberg and Hollingsworth 1986). Here we simply assume homogenous, isotropic streamfunction error statistics ($\mathbf{C}_0 = \mathbf{I}$). The resulting maps of β for the 1995–96 and 1996–97 winter seasons are shown in Fig. 8. Comparing Fig. 8 with Fig. 7, we see that the linear QG model with homogeneous, isotropic error statistics can simulate the geographical distribution of spread variability in the operational NCEP ensemble, at least qualitatively. We have used the differences between NCEP and European Centre for Medium-Range Weather Forecasts (ECMWF) reanalyses for the period 1979–93 to estimate $\mathbf{C}_0$, with qualitatively

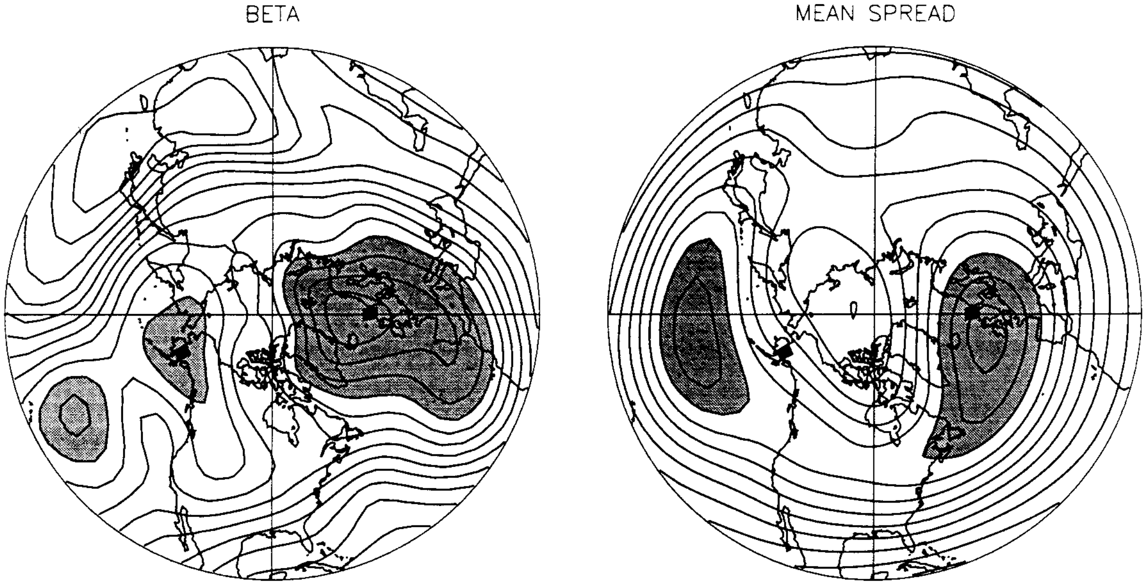

FIG. 9. Twenty-one winter mean 300-hPa streamfunction spread (S) and standard deviation of $\ln S$ (β), estimated from 3-day integrations of the five-level linear QG model. Here, S is normalized by the mean amplitude of the initial perturbations used in the ensemble integrations. Contour interval for β is 0.01, with values greater than 0.28 shaded. Contour interval for normalized S is 0.25, with values greater than 4 shaded. The filled rectangles indicate locations used to create correlation maps shown in Fig. 10.

similar results. The insensitivity of the results to the choice of $\mathbf{C}_0$ is due to the fact that all perturbations eventually evolve into the leading Lyapunov vector in the tangent linear model, regardless of their initial structure (Vannitsem and Nicolis 1997). Szunyogh et al. (1997) found that, by three or four days, arbitrary perturbations in a simplified version of the NCEP global forecast model evolved into structures resembling the leading Lyapunov vector. Although one can certainly propose analysis error structures that would impact the spread distribution at arbitrary lead times, it appears that beyond three days or so, the assumption that $\mathbf{C}_0 = \mathbf{I}$ produces a reasonable approximation to the spread distribution of the NCEP ensemble.

Figure 9 shows the 21 winter mean maps of 3-day spread and spread variability (β) computed with the linear QG model. The largest spread variability is located in the eastern North Atlantic, near or just downstream of the maximum in mean spread. In the eastern Pacific, the spread variability has two local maxima, one just to the north and east of the mean spread maximum, and one to the south and east.

Since we have shown that spread variability is generally associated with skill predictability in the NCEP ensemble, we would expect skill to be most predictable in regions where β is largest. In order to understand why these regions are favored, we have correlated the time series of $\ln S$ at the points indicated by the black rectangles in Fig. 9 with the corresponding time series of 3-day averaged 300-hPa streamfunction at all Northern Hemisphere grid points. The resulting correlation patterns (Fig. 10) closely resemble the leading modes of low-frequency variability in Northern Hemisphere winter (see, e.g., Wallace and Gutzler 1981). For example, spread variability in the eastern Atlantic is correlated with variability in the North Atlantic Oscillation (NAO), while spread variability in the eastern North Pacific is correlated with variability in the Pacific–North American Pattern (PNA). Thus, it appears that variations in the Pacific and Atlantic jets associated with the dominant modes of Northern Hemisphere low-frequency variability are primarily responsible for variations in the geographical patterns of spread from forecast to forecast. This explains why spread variability, and hence skill predictability, should be maximized in jet exit regions where low-frequency variability is strongest.

5. Summary and implications

Simple statistical considerations suggest that the more the ensemble spread departs from its climatological mean value, the more useful it is as a predictor of skill. Therefore, the correlation between spread and error should be related to the magnitude of spread variability. Examination of NCEP ensemble data for two winter seasons confirms that this is indeed true for an operational ensemble prediction system. In particular, spread/error correlations are higher where day-to-day variations of spread are largest, and spread is much more useful as a predictor of skill when it is extreme (very large or very small).

These results suggest that geographic variations of skill predictability are strongly related to geographic variations of spread variability. Spread variability can

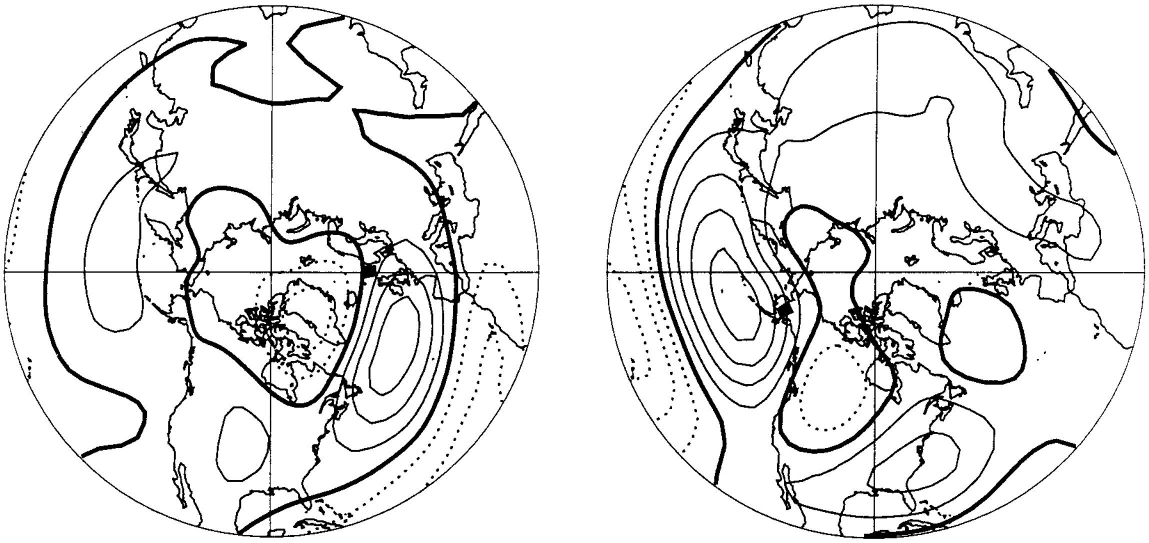

FIG. 10. Map of correlations between time series of ln$S$ at points indicated by the black rectangles and 3-day averaged 300-hPa streamfunction. Contour interval is 0.1, negative values are dashed, and the zero line is thick solid. Assuming there are 20 independent samples of ln$S$ for each winter season, correlations above 0.1 are locally significant at the 95% confidence level. Using the test described by Livezey and Chen (1983), the areas covered by correlations of 0.1 or greater are field significant at the 95% confidence level.

arise from two sources, 1) day-to-day variations in initial perturbation amplitude, and 2) day-to-day variations in the growth of these perturbations. For very short forecast times, the former is likely to dominate, while the latter will dominate for long forecast times. The NCEP breeding scheme for generating ensemble perturbations (Toth and Kalnay 1993) yields perturbations the magnitude of which can vary significantly from day to day. However, spread-error correlations are very small for short forecast times (1–2 days), indicating that these variations are not very representative of day-to-day variations of analysis error and that most of the useful skill predictability in the NCEP ensemble is associated with day-to-day variations in the perturbation growth.

With only two years of operational ensemble data it is difficult to determine what the climatological mean patterns of spread variability are, much less the dynamical mechanisms responsible for those patterns. Therefore, as a proxy for a long record of ensemble data, we have used data from a five-level QG model linearized about 3-day segments of NCEP reanalyses for 21 winter seasons. One hundred and twenty-one 3-day linear ensemble integrations are performed for each winter season, using random initial perturbations. The 21 winter mean maps of spread variability indicate that skill predictability should be highest just downstream of the maxima of climatological mean spread, over the eastern North Atlantic and Pacific oceans. Correlation analyses suggest that the dominant modes of northern winter low-frequency variability (i.e., the PNA and NAO modes) strongly modulate the growth of ensemble perturbations, leading to enhanced spread variability and skill predictability, just downstream of the jet exit regions.

As Table 1 demonstrates, spread is only useful as a skill predictor if it is extreme. Thus, it is crucial to know what the underlying distribution of spread is for a given ensemble configuration, so that extreme values of spread can be identified. From Fig. 8, it is clear that there are significant interannual variations of spread, so that more than two years of ensemble data are needed to accurately estimate the spread distribution. This suggests that in order to effectively utilize the information content present in the ensemble spread, a long record (10–15 yr) of ensemble integrations with the operational ensemble forecast system may be needed. This obviously would be an extremely expensive undertaking with a state of the art, high-resolution global forecast model, similar to those being used now for operational ensemble forecasting. However, for skill prediction it may be beneficial to run a simpler, lower-resolution ensemble for which a large dataset can be generated at a reasonable cost. The results presented here suggest that the potential benefit of having an accurate estimate of the climatological spread distribution may outweigh the loss of accuracy incurred by using a simplified forecast model.

Acknowledgments. Thanks to Zoltan Toth for help in obtaining the NCEP ensemble data. This work was sponsored by the National Science Foundation under Grant ATM-9503295.

APPENDIX

Simplified Dynamical Model of Spread Variability

After nondimensionalizing using the radius of the earth (a) as a length scale and the inverse of the earth's rotation rate (Ω) as a timescale, the governing equations for the Lorenz P model may be written

$$\frac{\partial\nabla^2\psi_j}{\partial t} + J(\psi_j, \nabla^2\psi_j + 2\mu) + \nabla\cdot 2\mu\nabla\chi_j = -(r_j^M + \upsilon\nabla^4)\nabla^2\psi_j \qquad (j = 1, 2, \ldots, N), \tag{A1}$$

$$\frac{\partial}{\partial t}(\phi_{j+1} - \phi_j) + \frac{1}{2}J(\psi_{j+1} + \psi_j, \phi_{j+1} - \phi_j) + \sigma_j\omega_j = -(r_j^T + \upsilon\nabla^4)(\phi_{j+1} - \phi_j) \qquad (j=1, 2, \ldots, N-1), \tag{A2}$$

$$\nabla^2(\phi_{j+1} - \phi_j) = \nabla\cdot 2\mu\nabla(\psi_{j+1} - \psi_j) \qquad (j = 1, 2, \ldots, N-1), \tag{A3}$$

where $J(A, B) = (\partial A/\partial\mu)(\partial B/\partial\lambda) - (\partial B/\partial\mu)(\partial A/\partial\lambda)$, ψ is streamfunction, ϕ is geopotential, χ is velocity potential, μ is the sine of latitude, $\omega = dp/dt$, and (r^M, r^T, υ) are damping parameters. The subscript j denotes the pressure level [$p_j = 1000 + (j - ½)\Delta p$ hPa] and Δp is the pressure difference between adjacent levels ($\Delta p = p_{j+1} - p_j = -1000/N$ hPa). The vertical velocity ω is staggered in the vertical with respect to ψ, ϕ, and χ. The variables ψ, χ, ϕ, and ω are nondimensionalized by Ωa^2, $(\Omega a)^2$, $(\Omega a)^2$, and $\Omega\Delta p$, respectively. The static stability σ_j is

$$\sigma_j = \frac{-\Delta\pi_j\Delta\Theta_j}{\Omega^2 a^2}, \tag{A4}$$

where $\Delta\pi_j = \pi_{j+1} - \pi_j$ is the difference between the Exner function [$\pi \equiv c_P(p/p_0)^{R/C_P}$] at adjacent levels, and $\Delta\Theta_j$ is the difference between the reference state potential temperature (Θ) at adjacent levels.

The horizontal boundary conditions are

$$\omega_N = 0 \quad \text{and} \quad \omega_0 = J(\psi_1, h), \tag{A5}$$

where h is topographic height (scaled by $\rho_0 g/\Delta p$), and ρ_0 is a reference value of density at 1000 hPa. The velocity potential χ is related to ω through the continuity equation

$$\nabla^2\chi_j = \omega_{j-1} - \omega_j. \tag{A6}$$

The vorticity equation (A1) is the prognostic equation for this model. Since geopotential and streamfunction are coupled through the balance equation (A3), elimination of the time derivatives in (A1) and (A2) using $\partial/\partial t$ of (A3) yields a diagnostic "ω equation" for the divergent flow. In order to derive the ω equation it is convenient to introduce a new variable, α, such that $\omega_j = -\nabla^2\alpha_j$ and $\alpha_{j+1} - \alpha_j = \chi_j$. In terms of α, the ω equation is

$$\sigma_j\nabla^4\alpha_j + \nabla\cdot 2\mu\nabla[\nabla^{-2}\{\nabla\cdot 2\mu\nabla(\alpha_{j+1} + \alpha_{j-1} - 2\alpha_j)\}] = \nabla\cdot 2\mu\nabla(F_{j+1}^{\psi} - F_j^{\psi}) - \nabla^2 F_j^{\phi} \qquad (j = 1, 2, \ldots, N-1), \tag{A7}$$

where

$$\nabla^2 F_j^{\psi} = -(r_j^M + \upsilon\nabla^4)\nabla^2\psi_j - J(\psi_j, \nabla^2\psi_j + 2\mu), \tag{A8}$$

and

$$F_j^{\phi} = -(r_j^T + \upsilon\nabla^4)(\phi_{j+1} - \phi_j) - \frac{1}{2}J(\psi_{j+1} + \psi_j, \phi_{j+1} - \phi_j). \tag{A9}$$

The boundary conditions for (A7) are $\alpha_N = 0$ and $\alpha_0 = -\nabla^{-2}J(\psi_1, h)$.

Here we set $N = 5$, $\upsilon = 2.338 \times 10^{16}$ m^4 s^{-1}, $r_j^M = 1$ days^{-1} for $j = 1$ (900 hPa) and zero otherwise, and $r_j^T = 0.1$ days^{-1} for all j. The reference state potential temperature profile Θ_j is computed from the climatological December–February mean in the latitude band 30°–60°N. The horizontal resolution of the model is T31, and a fourth-order Runge–Kutta time integration scheme is used.

REFERENCES

Arpe, K., A. Hollingsworth, M. S. Tracton, A. C. Lorenc, S. Uppala, and P. Kallberg, 1985: The response of numerical weather prediction systems to FGGE level IIb data. Part II: Forecast verifications and implications for predictability. *Quart. J. Roy. Meteor. Soc.,* **111,** 67–102.

Barker, T. W., 1991: The relationship between spread and forecast error in extended-range forecasts. *J. Climate,* **4,** 733–742.

Branstator, G., 1986: The variability in skill of 72-hour global-scale NMC forecasts. *Mon. Wea. Rev.,* **114,** 2628–2639.

Buizza, R., 1997: Potential forecast skill of ensemble prediction and spread and skill distributions of the ECMWF ensemble prediction system. *Mon. Wea. Rev.,* **125,** 99–119.

Houtekamer, P. L., 1993: Global and local skill forecasts. *Mon. Wea. Rev.,* **121,** 1834–1846.

Kalnay, E., and A. Dalcher, 1987: Forecasting forecast skill. *Mon. Wea. Rev.,* **115,** 349–356.

——, and Coauthors, 1996: The NCEP/NCAR 40–year reanalysis project. *Bull. Amer. Meteor. Soc.,* **77,** 437–471.

Kruizinga, S., and C. J. Kok, 1988: Evaluation of the ECMWF experimental skill prediction scheme and a statistical analysis of forecast errors. *Proc. ECMWF Workshop on Predictability in the Medium and Extended Range,* Reading, United Kingdom, ECMWF, 403–415.

Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.,* **102,** 409–418.

Livezey, R. E., and W. Y. Chen, 1983: Statistical field significance and its determination by Monte Carlo techniques. *Mon. Wea. Rev.,* **111,** 46–59.

Lonnberg, P., and A. Hollingsworth, 1986: The statistical structure of short-range forecast errors as determined from radiosonde data. Part II: The covariance of height and wind errors. *Tellus,* **38A,** 137–161.

Lorenz, E., 1960: Energy and numerical weather prediction. *Tellus,* **4,** 364–373.

Marshall, J., and F. Molteni, 1993: Toward a dynamical understanding of planetary-scale flow regimes. *J. Atmos. Sci.,* **50,** 1792–1818.

Murphy, J. M., 1988: The impact of ensemble forecasts on predictability. *Quart. J. Roy. Meteor. Soc.,* **114,** 463–493.

Palmer, T. N., and S. Tibaldi, 1988: On the prediction of forecast skill. *Proc. ECMWF Workshop on Predictability in the Medium and Extended Range,* Reading, United Kingdom, ECMWF, 253–310.

Sardeshmukh, P. D., and B. J. Hoskins, 1984: Spatial smoothing on the sphere. *Mon. Wea. Rev.,* **112,** 2524–2529.

Szunyogh, I., E. Kalnay, and Z. Toth, 1997: A comparison of Lyapunov and optimal vectors in a low-resolution GCM. *Tellus,* **49A,** 200–227.

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

Vannitsem, S., and C. Nicolis, 1997: Lyapunov vectors and error growth patterns in a T21L3 quasigeostrophic model. *J. Atmos. Sci.,* **54,** 347–361.

Wallace, J. M., and D. S. Gutzler, 1981: Teleconnections in the geopotential height field during the Northern Hemisphere winter. *Mon. Wea. Rev.,* **109,** 784–812.

Wobus, R. L., and E. Kalnay, 1995: Three years of operational prediction of forecast skill at NCEP. *Mon. Wea. Rev.,* **123,** 2132–2148.