23.5 FLOW-DEPENDENT CALIBRATION OF ENSEMBLE SPREAD USING FORECAST SPECTRA

Joshua P. Hacker* and David P. Baumhefner
The National Center for Atmospheric Research,† Boulder, CO

Abstract

This paper proposes a simple method for scale- and flow-dependent calibration of ensemble spread to account for excessive damping in a numerical weather prediction model. It is hypothesized that relationships between spatial variance and error growth can be applied to calibrate individual forecast periods. The relationships are transformed to spectral space and a simple example is used for concept demonstration. Applicability to individual forecasts is then tested on six independent cases by introducing numerical damping to the Weather Research and Forecasting (WRF) model. The results show that error growth estimated by the ensemble of imperfect damped forecasts can be calibrated to agree with the undamped model. The empirical correction factor is a function of the scale-dependent spatial variances in two model forecasts, and the flow of the day. Finally, the limitations of the calibration are demonstrated by comparing against a third model with very different error properties, and it is argued that the calibration provides a measure of the effect of those differences on ensemble spread. The calibration approach has potential application to ensemble forecasting systems, estimates of predictability limits, model-error diagnosis, and modern data assimilation systems.

1. Introduction

Numerical weather prediction model imperfections combine with initial-condition error to produce forecast error. Traditional ensemble forecasts attempt to predict the error by estimating the effect of initial-condition uncertainty on the forecast, but model imperfections still inhibit those estimates. One common source of model error results from numerical diffusion (damping), which produces stability while limiting the development of small-scale variance. An ensemble of forecasts with an overly-damped model will be under-dispersive and consistently under-predict error growth. The severity depends on the flow of the day and on the forecast lead time.

Error statistics forecast by an under-dispersive ensemble can be improved by posterior calibration through a regression against the climatological error (e.g. Hamill and Colucci 1998). Because the effects of damping varies according to the flow of the day, a calibration that relies on only the current forecast might provide further improvement and have applications to both operations and research. For example, ensemble-based data assimilation algorithms require accurate, flow-dependent, variance-covariance information from an ensemble forecast to be optimal, and can easily include model error covariances (e.g. Dee 1994). Research including observation-system simulation experiments for designing observing and data-assimilation systems, experiments to estimate the limits of predictability of phenomena or scales, and experiments to characterize initial condition and model error, could also benefit from improved error-growth estimates.

This study proposes and demonstrates a method for calibrating (correcting) flow-dependent ensemble spread by accounting for one type of model error: deficient spatial variance. Spatial variance and ensemble spread are related in the stationary climate limit (Leith 1974), explaining why an overly-damped model will produce climatologically under-dispersive ensembles. But daily variability in forecast spatial variance leads to daily variability in ensemble dispersion. Variances can be transformed to spectral space to obtain information about forecast amplitude, and a damped model will show smaller amplitudes at high wavenumbers. The overall effect of damping on ensemble spread then depends on the importance of high wavenumbers to the error growth in a forecast. We hypothesize that individual time-dependent forecast spectra can be used to correct the spread of an under-dispersive ensemble and achieve a spread appropriate for that particular forecast. Because damping effects will be unknown, and vary by case and forecast lead time, we must test the approach on individual forecast periods selected from different flow regimes. A refer-

Corresponding author address: Joshua Hacker, National Center for Atmospheric Research, P.O. Box 3000, Boulder, CO 80307. Email: hacker@ucar.edu

†The National Center for Atmospheric Research is sponsored by the National Science Foundation

ence model, which serves as truth, is used to make reference ensemble forecasts. Under-dispersive ensembles are generated by introducing model error in the form of a damping term. For any one forecast and lead time, a spectral calibration factor is computed from the reference and imperfect control forecast spectra. It contains the accumulated effects of damping on the spatial variance of the forecast, and can be used to correct the spread of the under-dispersive ensemble to agree with the reference ensemble.

The calibration is tested by applying a damping term to the model equations, but is aimed at any model error that results in a damped spectrum. It could be useful when a necessary component of model implementation, such as numerical diffusion or boundary conditions, is required because of expense or numerical stability. In the practical case that forecast spatial variance is limited because of computational constraints, a relatively inexpensive ensemble can be calibrated to obtain the error-growth statistics that would result from an expensive ensemble. For example, mesoscale models are often employed on limited domains to reduce costs, but recent studies reporting under-dispersive ensembles (Hamill and Colucci 1998; Hou et al. 2001; Grimit and Mass 2002; Hacker et al. 2002) suggest that at smaller scales and limited areas, truncation effects or other poorly-resolved scale interactions effectively damp the models and inhibit estimates of error growth. Multimodel ensemble techniques have shown limited success in accounting for this problem in specific applications (Hou et al. 2001; Stensrud et al. 2000; Ziehmann 2000; Grimit and Mass 2002), but their relationship to daily flow regimes are unclear. Flow-dependent calibration of single-model ensembles, generated with an amplitude-deficient model, is an alternative approach to improving error-growth estimates.

In the more profound case of estimating the error growth (e.g. doubling times) and the limit of predictability in the real atmospheric system, the method could be used with a large sample and observed atmospheric spectra[1] to obtain better estimates. Studies focusing on estimating the predictability of baroclinic scales in the atmosphere demonstrate that results are model-dependent. Lorenz (1965) estimated error doubling times of approximately 4 days, with an implied skillful prediction to 10 days. Charney et al. (1966) followed with an estimated 5-d error doubling time. With a coarse primitive-equation model, Smagorinsky (1969) shortened the estimate to 3 days. More recently, Lorenz (1982) used the ECMWF forecasting system to estimate 2.4 days. His comparison with analyses showed an upper-bound of 1.85 days. Even later, Savijärvi (1994) estimated a doubling time of 1.7 days for the MRF. Simmons et al. (1995) documented error doubling times in the ECMWF system from 1981 to 1994 (not monotonically decreasing), showing that they varied between 2.0 and 1.5 days. Simmons and Hollingsworth (2002) continued the analysis through 2002, confirming that estimates of error doubling times at baroclinic scales have decreased only slightly in recent years. At least part of the tendency toward shorter error-doubling times is attributable to better representation of small scales, leading to more realistic error growth. A fully-calibrated estimate of error growth would place this estimate at the limit of our ability to observe atmospheric spectra, and here we provide a partial calibration.

To explore flow-dependent calibration of ensemble spread we first derive a simple relationship to determine the under-estimation of error expected with an ensemble of damped forecasts. It relies on knowledge of a response function of one model relative to another, which can have complex structure but is easy to compute for discrete forecast times. We illustrate the effect with a simple smoother and a known response function that simulates a reduction in forecast amplitude. Later, two implementations of the Weather Research and Forecasting (WRF) model are used to make several ensemble forecasts. The WRF was chosen because it is a new model for which knowledge of ensemble behavior should benefit a large community in the future (e.g. Toth 2001), but the results are more general. Empirical, time-dependent, response functions are calculated for each forecast and used to explain differences in ensemble dispersion between ensembles of undamped and damped WRF runs. It is shown that the dispersion of the ensemble of damped WRF runs can be calibrated to agree with the ensemble of undamped WRF runs. Finally, ensembles with the Community Climate Model, version 3 (CCM3; Kiehl et al. 1998 and references therein) are compared to provide a counter example in which more comprehensive model error prevents a complete calibration. It is argued that the uncalibrated portion of ensemble spread is one measure the effects of the additional model error.

2. Ensemble response to damping

The goal of this section is to understand the expected response of an ensemble to damped forecast spectra. The smoother has a known spectral response function $R(k)$, where k is a wavenumber. Here, the response R is more generally the ratio between two spectra, from different models or a model and an analysis, and does not have to be a spectral filter response in a strict sense. It can be interpreted as a source of model error introduced to an otherwise perfect model, where the error affects the amplitude permitted in a forecast. The concepts are illus-

[1] By assuming that a robust atmospheric spectrum is possible to observe.

trated with an example designed to isolate the stationary impact of damping on ensemble dispersion.

For consistency with Leith (1974) we start with a perfect model, many forecast cases, and large ensembles. The operators averaging over cases and ensembles will be omitted for notational simplicity. One way to calculate ensemble spread is to average the spatial variance of the difference between all possible pairs of model forecasts for the same forecast period. One forecast pair is represented by $P(x,t)$ and $Q(x,t)$, with difference $S(x,t) = P - Q$. The ensemble spread is then the spatial variance of S, $\sigma_S^2(t) = \left\langle (S - \langle S \rangle)^2 \right\rangle$, where the operator $\langle * \rangle$ denotes a spatial average, and we understand that $\sigma_S^2(t)$ is really the average over *all* possible pairs. Initially $S(x,0)$ is very small, but it grows until the difference between any P and Q can no longer be distinguished from the difference between two random selections from climatology. Similarly, σ_S^2 is small and grows to saturation at $\sigma_S^2 = 2\sigma_P^2$, which is twice the climatological spatial variance of the model forecast. Because our model is perfect, $\sigma_Q^2 = \sigma_P^2$ in the climate average.

To further simplify the discussion and clarify the concepts, we temporarily drop the dependence on time. Invoking all of the simplifications, the ensemble spread can be written as

$$\sigma_S^2 = \left\langle (P - Q - \langle P - Q \rangle)^2 \right\rangle \ . \tag{1}$$

Expanding and using averaging rules,

$$\sigma_S^2 = \left\langle P^2 \right\rangle - \langle P \rangle^2 + \left\langle Q^2 \right\rangle - \langle Q \rangle^2 - 2\left(\langle PQ \rangle - \langle P \rangle \langle Q \rangle \right) \ , \tag{2}$$

which is just the sum of the spatial variances σ_P^2 and σ_Q^2 with twice the covariance $\mathrm{cov}(P,Q)$ subtracted (e.g. Strait 1989). Murphy (1988) outlines a similar procedure for mean-square error.

The model states can be viewed in discrete spectral space by defining the Fourier pair for $P(x)$, with $x = n\Delta x$:

$$P(n) = \sum_{k=0}^{N-1} \hat{P}(k) e^{\frac{i2\pi kn\Delta x}{N}} \ , \ \hat{P}(k) = \frac{1}{N} \sum_{n=0}^{N-1} P(n) e^{\frac{-i2\pi kn\Delta x}{N}} \ , \tag{3}$$

and similarly for Q, where $i = \sqrt{-1}$, and the location x varies discretely over $[0 \ldots N\Delta x]$. Using equation (3) to rewrite (2) in spectral space, and subtracting the mean, gives

$$\sigma_S^2 = \sum_{k=1}^{N-1} \hat{S}^2 = \sum_{k=1}^{N-1} \left(\hat{P}^2 + \hat{Q}^2 - 2\hat{P}\hat{Q} \right) \ . \tag{4}$$

The discrete spectral power of the difference field $P - Q$ is the ensemble spread as a function of wavenumber k and can be written

$$\hat{S}^2 = \hat{P}^2 + \hat{Q}^2 - 2\hat{P}\hat{Q} \ , \tag{5}$$

for $k = [0 \ldots N/2]$. This has been used commonly to compare scale-dependent ensemble dispersion with spatial variance $\hat{P}^2$ at discrete wavenumbers (e.g. Lorenz 1969; Boer 1984; Dalcher and Kalnay 1987; Errico and Baumhefner 1987; Savijärvi 1994). One measure of a useful forecast at a particular scale k is the condition $\hat{S}^2 < \hat{P}^2$, and we use $\hat{S}^2 = 2\hat{P}^2$ as the saturation point. For our purpose, equation (5) enables easy analysis of the spectral response of σ_S^2 to any change in the spectral characteristics of σ_P^2 or σ_Q^2, such as could result from damping or truncation.

Considering the spectra $\hat{P}, \hat{Q}$ as a reference, we can introduce one simulated source of model error in the form of a smoother that reduces the forecast amplitude. In spectral space, applying the smoother to the model states amounts to the operations:

$$\hat{p} = R\hat{P} \ , \ \ \hat{q} = R\hat{Q} \ . \tag{6}$$

With (6), the spread at wavenumber k that results from using damped model states is

$$\begin{aligned} \hat{S}_f^2 \quad &= \hat{p}^2 + \hat{q}^2 - 2\hat{p}\hat{q} \\ &= R^2 \left(\hat{P}^2 + \hat{Q}^2 - 2\hat{P}\hat{Q} \right) \\ &= R^2 \hat{S}^2 \ , \end{aligned} \tag{7}$$

Note that $\hat{S}$ describes the total ensemble dispersion, while $\hat{P}$ and $\hat{Q}$ provide information about amplitude only.

For the discrete model at all scales, the spread in an ensemble of damped model forecasts is then given by

$$\sigma_{S_f}^2 = \sum_{k=1}^{N-1} \hat{S}_f^2 = \sum_{k=1}^{N-1} R^2 \hat{S}^2 \ . \tag{8}$$

Thus if spread is determined by summing the spectral coefficients $\hat{S}^2$ or $\hat{S}_f^2$ for all pairs of model states, and the spectral response function R is known, then the instantaneous effect of damping on ensemble spread can be determined. The response R is a function of the damping error in spectral space, and can be computed via equation (6) when $\hat{P}$ and $\hat{p}$ are known.

As an example, consider a simple smoother applied to all 1-D forecast pairs $P(x,t)$, $Q(x,t)$ in a large ensemble on the domain from $x = 0$ to $N\Delta x$. A 1-2-1 smoother in physical space with a coefficient 0.5 gives the response in spectral space $R = \cos^2(\frac{\pi kn\Delta x}{N})$ (e.g. Haltiner and Williams 1980). Figure 1 shows both R and R^2 as a function of wavenumber, demonstrating the R^2 effect on the ensemble spread in spectral space. In this example R is constant but it need not be, as shown later.

A large sample of random statistical realizations are created to illustrate the consequences of damping a model. The parameters of the experiments are chosen to qualitatively represent observed geopotential height

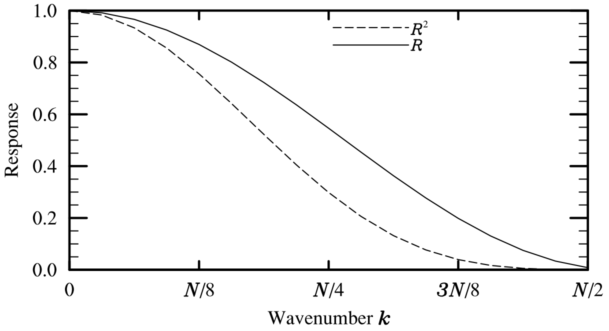

Figure 1: Filter response function (R) of a simple $2\Delta x$ smoother, as a function of wavenumber where N denotes the number of grid points in 1-D. The curve R^2 is also shown because it operates on the ensemble spread.

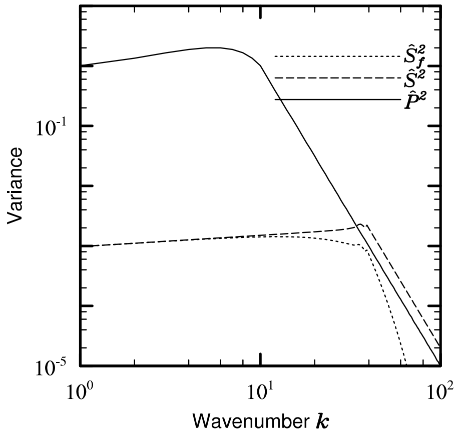

Figure 2: Variance as a function of wave number. Shown are the field itself ($\hat{P}^2$), the spread ($\hat{S}^2$), and the spread after a $2\Delta x$ smoother has been applied to the fields ($\hat{S}_f^2$).

spectra and error growth. In the midlatitudes, small scales become unpredictable sooner than large scales (Lorenz 1969), and $\hat{S}^2(k,t) \geq \hat{P}^2(k,t)$ at high wavenumbers. Thus the saturation wavenumber k_{sat} varies inversely with forecast lead time. For each experiment, 10^6 pairs of spectra are constructed by randomly rotating the complex Fourier coefficients of a smooth spectrum that has a slope of -5 at $k > 10$ (approximately baroclinic scale). These rotations are a simple way to represent the spectral signal of the difference between two model forecasts with the same model. In the limit of infinite ensemble members, they average to zero and the mean spectrum of all P and Q exactly matches the smooth spectrum. But for an infinite number of pairs of forecasts the difference is finite and the spectrum of differences shows the ensemble spread, which can be scaled to represent any shape of spread spectrum. An example of $\hat{P}^2$, $\hat{S}^2$, and $\hat{S}_f^2$ when $k_{sat} = 32$ for $N = 256$ is shown in Fig. 2. This type of error is qualitatively consistent with Leith (1974), Boer (1984), and Daley and Mayer (1986), and should result in realistic error growth curves. Because k_{sat} is a proxy for forecast lead time, spreads of damped and undamped forecast pairs are calculated for every discrete $k_{sat} = [0 \ldots N/2]$, thereby estimating the effect of the smoother R on the ensemble spread from initialization to its asymptotic limit.

To illustrate the effect of damping at a range of scales, $\sigma_{S_f}^2/\sigma_S^2$ is shown in Fig. 3a for a varying number of grid points in the physical domain: $N = (32, 64, 128, 256, 512, 1024)$. With a constant domain size, increasing N increases the model resolution in physical space. In real NWP models, smoothing coefficients are normalized by the grid spacing and are therefore explicitly tied to a physical scale. But in our case R is tied to the grid, and by varying N we are actually varying the scale selection of the smoother as if we were changing the smoothing algorithm or coefficients. The curves for the sharper R are toward the left (note the abscissa is reversed from the spectra plots). All of the curves asymptote as $k_{sat} \rightarrow 1$, but the spreads for a broader R (smaller N) asymptote at a lower level. Figure 3b shows that the growth rates in Fig. 3a collapse when k_{sat} is normalized by the resolution determined by N, but the asymptotic spread does not recover. The same can be expected from choosing a sharper R, which will weaken the impact on ensemble spread but not remove it entirely.

To further investigate the time-integrated effect of damping on ensemble spread, σ_S^2 and $\sigma_{S_f}^2$ can be separated from the individual curves in Fig. 3a and transformed to show time evolution. We assume exponential error growth according to a specified error-doubling time of two days to approximate recent research results (Savijärvi 1994; Simmons et al. 1995; Simmons and Hollingsworth 2002), resulting in the ensemble dispersion for $N = 128$ and 256 shown in Fig. 4. It is clear that damping inhibits both the growth rate and the asymptotic limit of ensemble spread, and that the effect is greater for the R with broader scale selection ($N = 128$). The dependency on scale selection can be explained by recognizing that, when R exhibits sharper scale selection, it acts on smaller scales containing less energy and saturating earlier in the forecast. The asymptotic limit is lower because the damped spread will saturate relative to the damped model, which has spectral amplitude below that of the perfect model (Fig. 2), and the same dependency on scale selection applies. As N increases and R is more selective, the damped dispersion curves converge to the undamped curve.

This example shows that an ensemble will be underdispersive if it is comprised of model forecasts that are

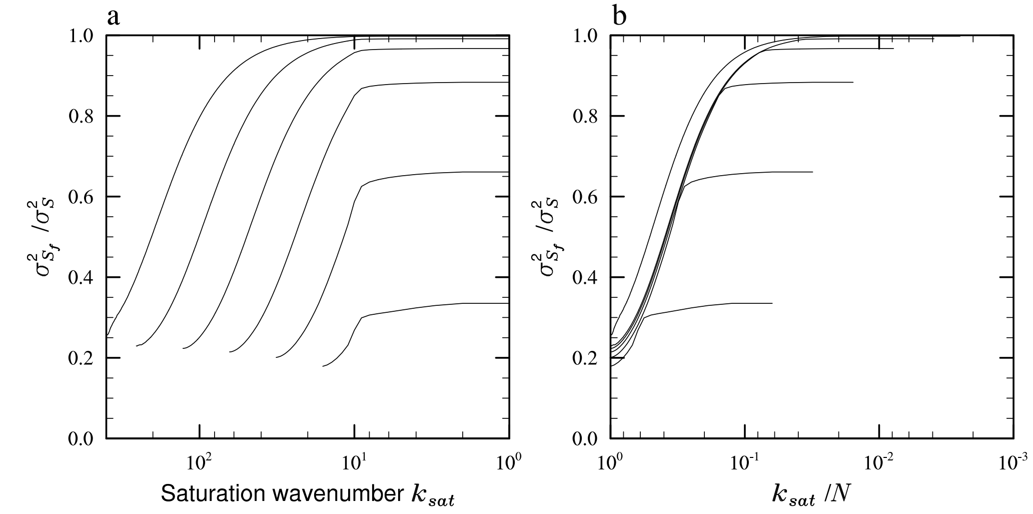

Figure 3: The response of the ensemble spread to damping in a model, as a function of saturation wavenumber. The abscissa denotes the scale of error saturation, increasing (going up scale) to the right. Each curve in panel (a) is for a different resolution, with higher resolution to the left, and the curves in panel (b) are normalized by the number of grid points.

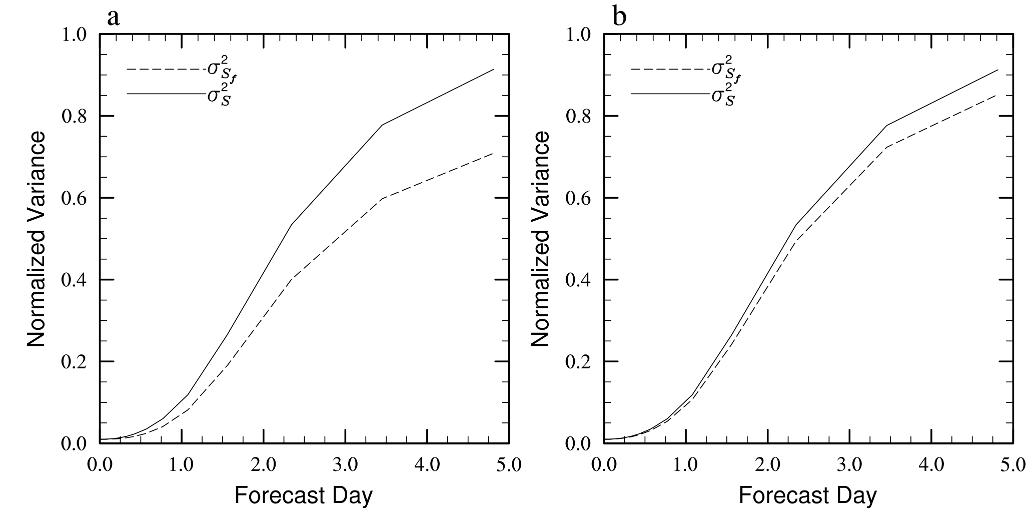

Figure 4: The effect of damping on error growth as estimated by ensemble spread for (a) a relatively low-resolution ($N = 128$) and (b) double resolution ($N = 256$).

unrealistically damped compared to the real atmosphere or a superior model, and that the effects are worse with an unwise filter selection. The instantaneous effects are demonstrated with a constant R that has no interaction with the flow and has limited utility for a real forecasting system, but conceptually and mathematically the relationships are the same.

When a reference model or analysis is available for computing $\hat{P}$, then $R(k,t)$ can be empirically estimated by computing $\hat{p}$ and using equation (6). In an ensemble experiment, $\hat{S}^2_f(k,t)$ is also easily measured. Then $\hat{S}^2(k,t)$ can be computed with equation (7) and the dispersion with equation (4), thereby calibrating an amplitude-deficient model.

The time- and flow-dependent response function R computed with real NWP forecasts measures the time-integrated effect of damping the spectra throughout a forecast. This may result from a combination of implicit numerical damping, explicit numerical filtering, subgrid-scale parameterization schemes that explicitly mix (including turbulence and convection schemes), and boundary conditions. It should have complex structure that does not at all resemble Fig. 1. The next section applies the variable R to calibrate an ensemble of damped model forecasts, and compares resulting dispersion characteristics and calibration efficacy in the face of flow- and time-dependency. Limitations of the method resulting from additional deficiencies are addressed in section 4.

3. Flow-dependent calibration for amplitude deficiencies

The remainder of this paper is concerned with extending the concepts developed in section 2 to include time- and flow-dependent dynamics by testing them with real NWP models. The experiments here are controlled to isolate the effects of damping. To begin, spectra of the forecasts and dispersion characteristics of ensembles with both undamped and damped versions of the WRF are compared. All results shown are for 50.0 kPa geopotential height, interpolated to a Gaussian grid, and spectra are computed with spherical harmonics.

The focus of the WRF design is as both a research and operational model to be implemented with grid-spacing of 1-10 km, though it should also be accurate and efficient at larger scales. It contains a suite of modern physical parameterization schemes built around three options for dynamical core. Here we configure the WRF with full physics, and employ a mass-based vertical coordinate with split-explicit high-order numerical schemes as described in Skamarock et al. (2001). One potential advantage of the higher-order numerical schemes available in the WRF is that truncation and implicit diffusion from the numerics should be minimal and the effective resolution should approach the true resolution as determined by the grid spacing.

A model domain is chosen that covers the entire northern hemisphere, with horizontal grid spacing $\Delta X = \Delta Y =$ 90 km, true at 45°N on a polar-stereographic projection. The projection on a hemispheric domain introduces substantial grid distortion in the tropics, but it minimizes boundary effects. Relatively large grid spacing provides a fair agreement with gridded analyses in the midlatitudes, minimizing interpolation errors for initial and boundary conditions and facilitating comparison with analyses in future work. Here we use the NCEP final analyses, available twice daily at 1° grid spacing on a cylindrical equidistant grid.

We start by designating the undamped WRF, and ensembles run with it (denoted *WRF*), to be the perfect reference against which model changes are evaluated. As in the illustrative case of the last section, one source of simulated model error is introduced by applying a 2-D, second-order diffusion term that explicitly (and locally in time) reduces forecast spatial variance at small scales. We denote the resulting damped WRF with DMP. The horizontal diffusion coefficient is chosen as $K_h =$ 90000 m^2 s^{-1}, corresponding to ΔX. This type of diffusion is purely numerical and depends only on the grid and K_h. Ensembles run with DMP (denoted *DMP*) are compared to ensembles *WRF*, and R is used to explain the differences in ensemble spread. Here R is the ratio, in spectral space, of the DMP to the WRF control runs, and it is computed separately for each forecast period and lead time.

Random perturbations are applied with the Errico-Baumhefner technique (e.g.
Errico and Baumhefner 1987; Tribbia and Baumhefner 1988; Mullen and Baumhefner 1989, 1994; Stensrud et al. 2000), which approximates analysis errors as estimated by Daley and Mayer (1986), to generate 10-member ensembles of initial conditions. Control (unperturbed) initial conditions are given by the NCEP final analyses. The WRF model currently does not have an option for explicit balancing (such as a normal mode initialization). The adjustment, and the fact that the random perturbations are not constrained to lay on the model attractor, lead to spurious inertio-gravity waves that disperse over the first few hours of the forecast (Anderson and Hubeny 1997). As the forecasts regain balance with respect to the model equations, the ensemble spread decreases. Here we ignore that adjustment period and examine the dispersion characteristics after the minimum in spread.

Cases were selected by finding six different flow regimes during the 2001-2002 northern hemisphere cool season that are considered independent of each other (Table 1). We present averages over the six cases for sum-

Table 1: Forecast cases that comprise the averages presented in this paper.

Initialization Time
00 UTC 24 November 2001
00 UTC 20 December 2001
12 UTC 08 January 2002
00 UTC 30 January 2002
12 UTC 06 February 2002
00 UTC 16 February 2002

mary purposes only, but each case is treated independently. Figure 5 shows that error growth, as estimated by ensemble dispersion, varies widely between the cases.

A comparison of *DMP* and *WRF* ensemble dispersion is shown in Fig. 6. Spread in the *WRF* ensembles grows faster than spread in the *DMP* ensembles over the first 4.5 forecast days. For any single case, error growth estimated from the *WRF* dispersion curve is faster than that estimated from the *DMP* curve. Assuming this behavior continues to the asymptotic limit, a shorter estimate of the limit of predictability would result from using the *WRF* curve.

Comparing spectra of the DMP and WRF 4.5-day forecasts (Fig. 7) further elucidates the effect of damping. When undamped, the dispersion is saturated through the high-wavenumber part of the spectrum. But relative to the undamped WRF forecast the dispersion of DMP forecasts does not saturate at any scale and the total dispersion (area under the curve) of ensemble *DMP* lags below the dispersion of *WRF*. The ratio of control spectra results in $R < 1$ almost everywhere in Fig. 7b.

These results confirm that applying damping to component forecasts of an ensemble strongly modulates the ensemble spread in a manner similar to the effect demonstrated with a simple statistical model in section 2. The instantaneous effect is similar to Fig. 2, but the effect of the filter spreads up scale as the forecast progresses. Here the damping is $2\Delta X$ diffusion that produces the fully-developed, time-integrated response function shown in Fig. 7b, which accounts for all of the scale interactions leading to amplitude differences between the forecasts

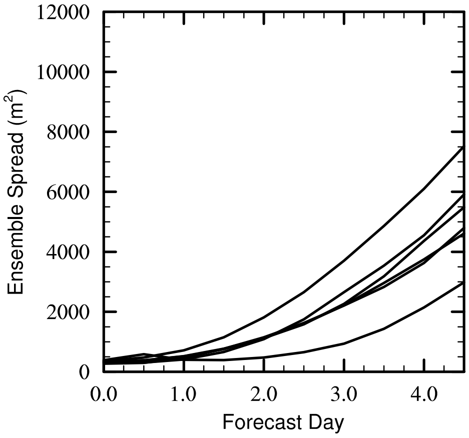

Figure 5: Dispersion of 50.0 kPa geopotential heights in the ensembles of undamped (*WRF*) forecasts for each of the six cases in Table 1.

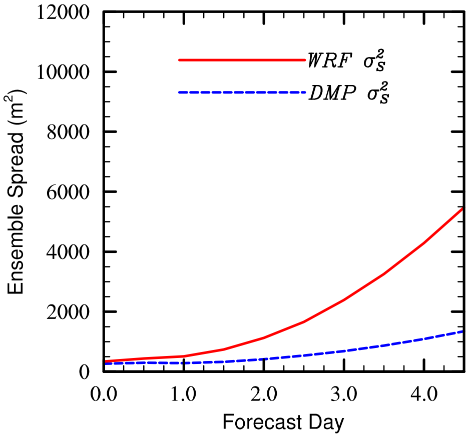

Figure 6: Dispersion of 50.0 kPa geopotential heights in the ensembles of undamped (*WRF*) and damped (*DMP*) forecasts.

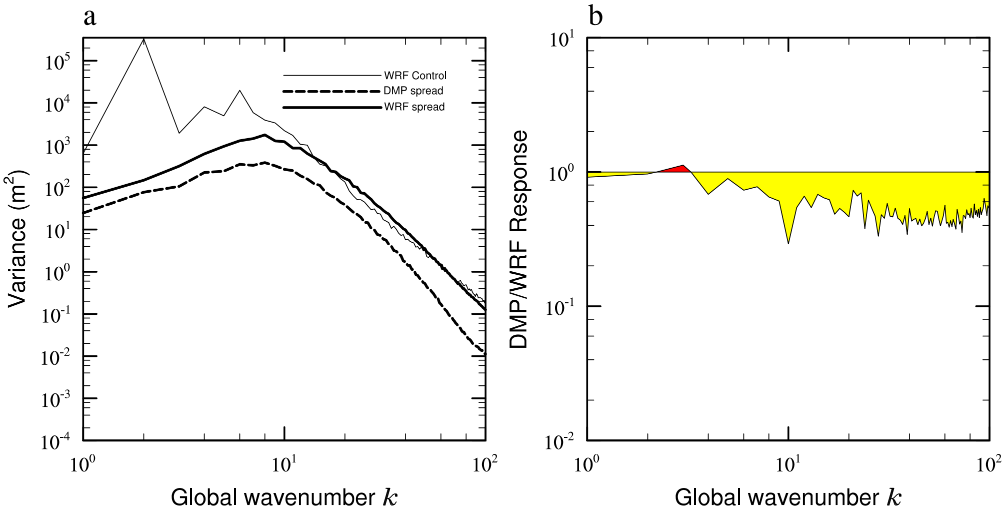

Figure 7: In (a), the WRF (thick dotted line) and DMP (thick solid line) 50.0 kPa geopotential height spread in spectral space, compared to the WRF control forecast spectrum (thin solid line). In (b), a response function is shown, calculated as the ratio of DMP to WRF forecast spectra. Results are for 4.5-day forecasts.

from two different models.

The flow- and time-dependent calibration potential of R can be easily checked by computing $R(k,t) = \hat{p}(k,t)/\hat{P}(k,t)$, and using equation (7) to correct the damped dispersion $\hat{S}_f^2(k,t)$ to $\hat{S}^2(k,t)$. The result, averaged over all the cases, is shown in Fig. 8. The corrected (*COR*) dispersion, closely follows the dispersion of the ensemble of undamped WRF runs. This implies that the time-dependent properties of R contain most of the time-integrated information required to relate ensemble dispersion with these two different models, and that an estimate of R can be used to successfully calibrate ensemble spread for a certain class of model error.

Distinguishing the effects of deficient spatial variance from the total ensemble spread helps guide discussion on the generality of these results. The fact that the difference between the *WRF* and *DMP* ensemble spreads can be almost entirely explained by the measured spectral response (Fig. 8) suggests that the accumulated effect of damping is focused on forecast amplitude at lead times up to 4.5 days. This calibration method will be most useful when two models differ primarily in the forecast spatial variance ($\hat{P}^2$). It should be applicable for model differences such as other numerical smoothers and perhaps some differences in physical parameterization schemes and boundary conditions. But the calibration is less likely to be successful for wholesale differences in numerical discretization schemes because they could have a large effect on other model error modulating the ensemble dispersion. The next section explores the limitations of the calibration by introducing more comprehensive model error.

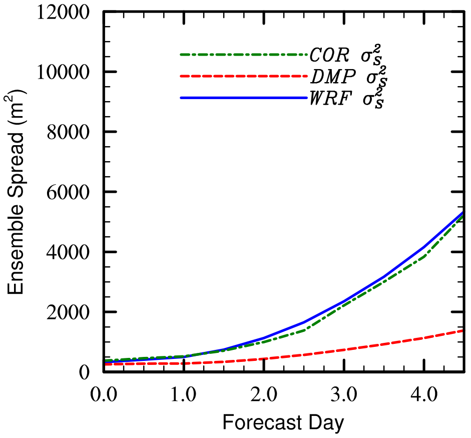

Figure 8: Dispersion of the ensemble with undamped (*WRF*) and damped (*DMP*) WRF forecasts. The corrected (*COR*) dispersion results from using the R^2 to predict the dispersion without damping.

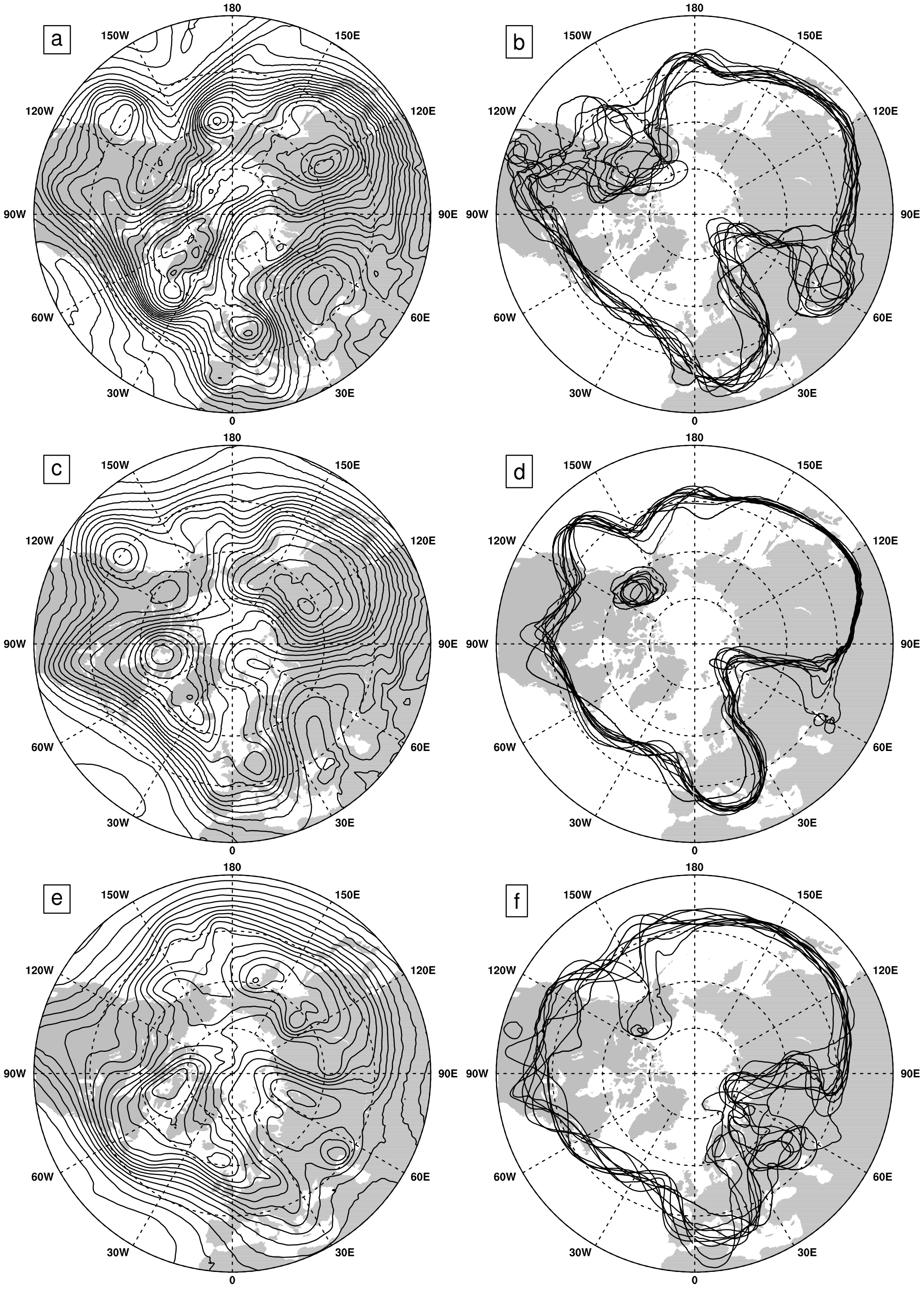

Figure 9: Day-six 50.0 kPa geopotential (m) height control forecasts (left column) and corresponding ensemble spaghetti diagrams (right column) for (a, b) the undamped WRF, (c, d) the damped WRF, and (e,f) the CCM, all valid 00 UTC 30 November 2001. Contour intervals are 60 m, and the 5880 m contour is plotted in the spaghetti diagrams.

4. Limitations of the calibration: other sources of model error

This section compares the WRF ensemble forecasts from the last section with CCM ensembles (denoted *CCM*) to expose limitations of the calibration method. The CCM (Kiehl et al. 1998) is discretized spectrally on the globe with energy-conserving schemes and the forecasts are not affected by lateral boundaries. Because of the discretization, the lateral boundaries in the WRF, and different physical parameterization schemes in each model, the difference between the WRF and the CCM is far greater than the difference between the damped and undamped WRF. Thus error-growth characteristics will also be different. Ensembles are run with the CCM for the same cases as the WRF, where ensembles *CCM* are generated with the same technique as ensembles *WRF*. The CCM is considered the perfect reference for erroneous model DMP, and the WRF the perfect reference for erroneous model CCM. The calibration will account for model error due to damping, and the uncalibrated spread measures the effect of the remaining model error on the total spread.

An example is useful to gain an appreciation for some of the differences between forecasts and ensembles with the different models. Figure 9 shows six-day forecasts from the WRF, DMP and CCM, valid at 00 UTC 30 November 2001. Comparison of the WRF to DMP control forecasts (panels a and c) shows a clear difference in forecast amplitude at all scales. Although some synoptic features are in slightly different locations, many of them are zonally collocated. For example, note the cutoff lows just off the west coast of North America and over Europe, and the ridging over western Asia. In both ensembles *WRF* and *DMP* (panels b and d) the locations of highs and lows also closely overlap, but the ensemble spread is greater in *WRF* with respect to the amplitude of features at all scales. The small ensemble spread in *DMP* demonstrates the physical effect of the applied damping. While the CCM forecast (panel e) shows synoptic amplitude that is similar to DMP, synoptic features are located differently. A cutoff low is forecast over western Asia and ridging is forecast off the west coast of North America. Comparing the spaghetti diagrams also shows that within the *CCM* ensemble, the locations of both small- and large-scale features have spread more than in either *WRF* or *DMP*. This example was arbitrarily chosen and similar behavior can be observed for all of the cases in Table 1.

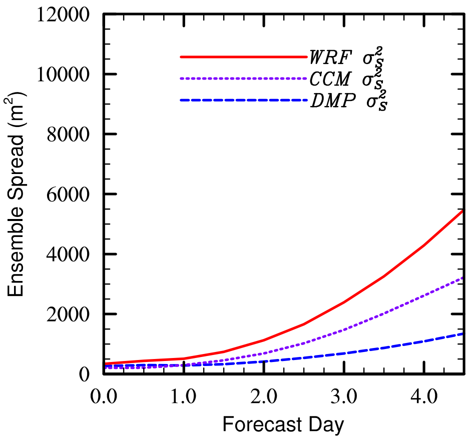

Figure 10: Ensemble dispersion versus forecast hour for the WRF (solid line), the damped WRF (dashed line), and the CCM (dotted line). The curves are the average of 50.0 kPa geopotential height ensemble dispersion for all the cases in Table 1.

The *CCM* ensemble dispersion curve, compared to the *WRF* and *DMP* dispersion, is shown in Fig. 10. Spread in the *WRF* ensembles grows faster than spread in the *CCM* ensembles over the first 4.5 forecast days, but spread in the *DMP* ensembles grows slower. Error growth estimated from the *WRF* (*DMP*) dispersion curve is faster (slower) than that estimated from the *CCM*.

Calibrating the *CCM* to the *WRF* curve or the *DMP* to the *CCM* curve[2] will not be effective given the CCM/WRF and DMP/CCM responses (Fig. 11). The CCM/WRF response shows magnitudes that are comparable to the DMP/WRF values of *R* in Fig. 7b through wide spectral bands despite the far smaller difference in spread between the *CCM* and *WRF* dispersion. Thus the calibration will overcorrect. Conversely, the DMP/CCM response remains near one throughout the spectral range, which will result in very little correction. This behavior is confirmed when the calibration is applied in Fig. 12.

When one model is a reference or truth, then one measure of the relative effect of model error on dispersion is the differences between the dispersion curves. After calibrating for amplitude differences, the results shown in Fig. 12 are useful to estimate the remaining relative error growth in the three models. We know from section 3 that this calibration corrects for deficient spatial variances represented by amplitude deficiencies. The forecast fields (Fig. 9) indicate similar amplitude in the DMP and CCM, and different amplitude in the WRF. The DMP/CCM similarity in forecast amplitude is also documented in Fig. 11b. Then from Fig. 12 we can deduce that the contribution to ensemble spread in *CCM* attributable to other sources, besides spatial variance, is greater than its contribution to the spread of *WRF*. The spaghetti plots in Fig. 9 are consistent with this. Figure 13 illustrates this concept, and it can be scaled for any

[2] Calibration can be performed in either direction and we choose to increase the ensemble spread for consistency with section 3.

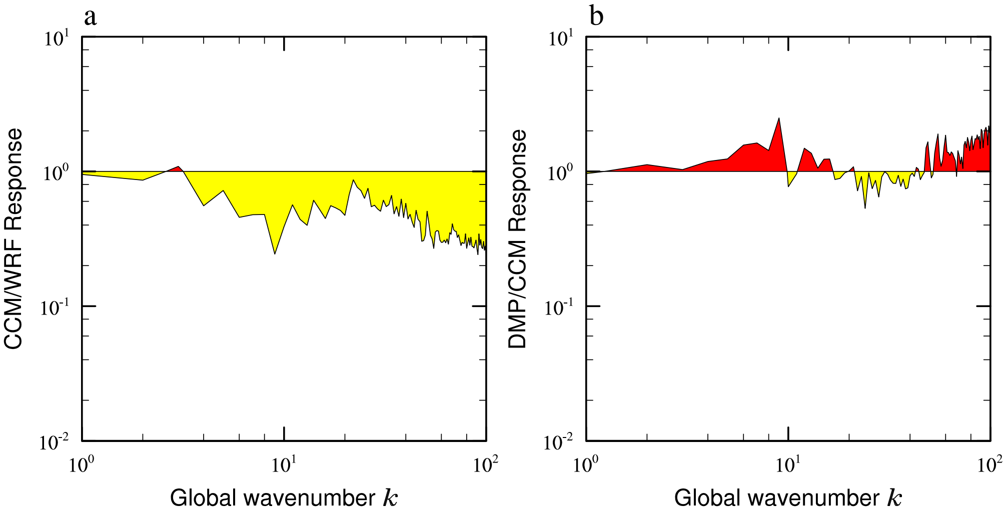

Figure 11: The 4.5-day response function R computed as the ratio of control forecast 50.0 kPa geopotential height spectra as (a) CCM/WRF and (b) DMP/CCM.

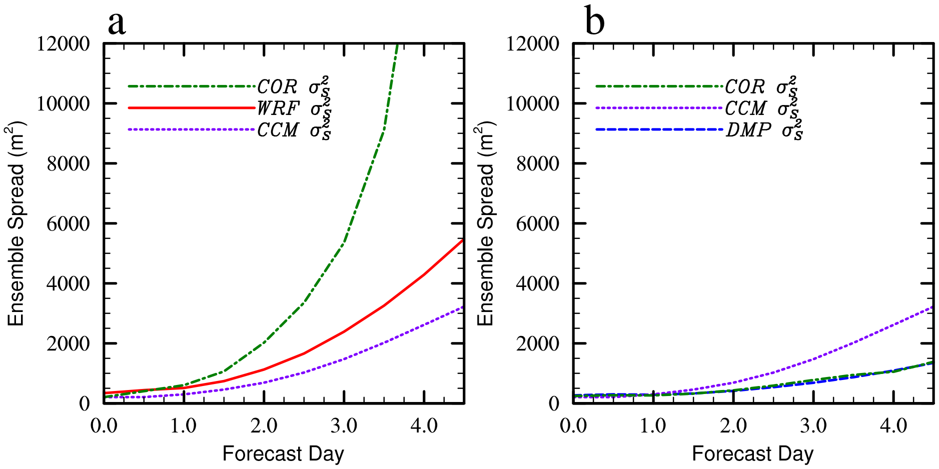

Figure 12: In (a), the 50.0 kPa geopotential height dispersion of ensembles *CCM* and *WRF*, with *COR* showing the *CCM* ensemble calibrated to the *WRF* ensemble using R in Fig. 11a. In (b), the dispersion of ensembles *DMP* and *CCM*, with *COR* showing the *DMP* ensemble calibrated to the *CCM* ensemble using R in Fig. 11b.

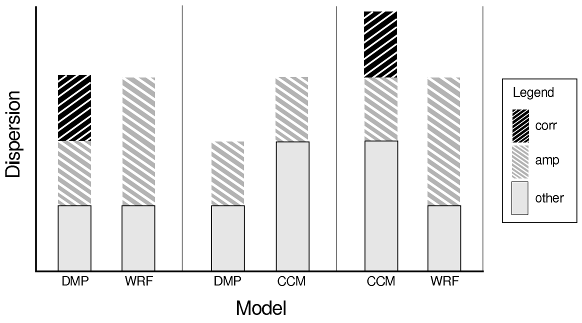

Figure 13: Schematic qualitatively showing the effects of model error on ensemble spread. Contributions from amplitude (amp) and other deficiencies to the total ensemble spreads are shown separately. The calibrated portion (corr) is related to the difference in amplitude dispersion for each pair.

forecast lead time with all of the dispersion bars becoming very small at initialization. Thus additional components to calibration will be required to account for all sources of model error.

A historical analysis could also be used to calibrate an overly-damped model and assign an absolute magnitude to the effects of other model error sources. For example, replace the WRF with a set of analyses and calibrate ensemble *CCM* with an *R* computed from CCM forecasts and verifying analysis spectra. If the calibrated ensemble spread grows faster or slower than an estimate of actual error growth, then the CCM has additional model error that affects its error prediction, and the effect is quantified by the difference between the *COR* curve and the actual error growth curve.

5. Summary and conclusions

The sensitivity of ensemble spread to flow-dependent forecast spatial variance was explored, and a calibration method was proposed to account for it. First, we derived the effect that damping an ensemble-generating model has on ensemble spread. This was achieved by extending the relationships explained in Leith (1974) to compute scale-dependent spatial variances (forecast amplitude) in spectral space. In the case of stationary climate statistics the relationship between forecast amplitude and ensemble spread is clear, and a statistical model served as an example to illustrate the consequences of deficient amplitude. We hypothesized that for an individual forecast the same relationships provide a flow-dependent correction for the spread of an under-dispersive ensemble. The hypothesis was tested by comparing spectral characteristics of two versions of the WRF model — one that was completely undamped and one to which a grid-dependent diffusion term was applied. Dispersion of an ensemble of undamped WRF forecasts was faster than an ensemble of damped WRF forecasts, and could be explained by more spatial variance in the undamped WRF model. An empirical time-dependent spectral response *R*, computed as the ratio of the damped to undamped WRF spectra for each of six independent cases, was used to correct the spread of the ensemble of damped model forecasts to agree with the spread of the ensemble of undamped model forecasts at all lead times. Finally, it was demonstrated that additional sources of model error contributing to ensemble spread is a limitation of the calibration method, but that the uncorrected portion of spread is a measure of the effects of that error.

The implications for predictability research and ensemble forecasting are greater when the goal is model trajectories that diverge at the same rate as trajectories in the real atmosphere, thereby allowing error prediction with model forecasts. Although this calibration does not change any trajectories (though in principle it could), it will facilitate error prediction that is consistent with the flow of the day when amplitude deficiencies are foremost. The reference for computing *R* can be either observations or a superior model. In the first instance, some stationarity of variance must be assumed and it only makes sense to calibrate hemispheric or global ensembles to hemispheric or global observations. A historical analysis could lead to a better estimate of scale-dependent error growth and limits of predictability in the atmosphere. In the second, a less expensive model (such as a LAM) can be used to generate ensembles that can be calibrated to a global model. Only one expensive global run is needed to compute an *R* that could account for boundary conditions or a lack of scale interactions.

Accurate error prediction in all situations requires that a calibrated ensemble accounts for the effects of all model error on ensemble spread. Section 4 documents the principal limitation of the proposed calibration: the case that an ensemble to be calibrated is generated from a model containing comprehensive error beyond what appears as damping. This arises from the fact that *R* contains information about amplitude only, while ensemble spread is modulated by all types of model error. Then partial calibration provides a measurement of the effect of the additional model error.

This calibration approach is a first step and additional research may extend its utility. A scale- and flow-dependent parameterization of model error would be useful for four-dimensional data assimilation methods (c.f. Dee 1994). More generally, a better understanding may lead to better-calibrated ensemble forecasts, a more robust method for predicting how model changes affect ensemble performance, and a useful diagnostic for model error.

REFERENCES

Anderson, J. and V. Hubeny, 1997: A reexamination of methods for evaluating the predictability of the atmosphere. *Nonlinear Processes Geophys.*, **4**, 157–165.

Boer, G., 1984: A spectral analysis of predictability and error in an operational forecast system. *Mon. Wea. Rev.*, **112**, 1183–1197.

Charney, J. G., R. G. Fleagle, H. Riehl, V. E. Lally, and D. Q. Wark, 1966: The feasibility of a global observation experiment. *Bull. Amer. Meteor. Soc.*, **47**, 200–220.

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

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

Dee, D., 1994: On-line estimation of error covariance parameters for atmospheric data assimilation. *Mon. Wea. Rev.*, **123**, 1128–1145.

Errico, R. and D. Baumhefner, 1987: Predictability experiments using a high-resolution limited-area model. *Mon. Wea. Rev.*, **113**, 488–504.

Grimit, E. and C. Mass, 2002: Initial results of a mesoscale short-range ensemble forecasting system over the Pacific Northwest. *Wea. and Forecast.*, **17**, 192–205.

Hacker, J., S. Krayenhoff, and R. Stull, 2002: Ensemble experiments on numerical weather prediction error and uncertainty for a North Pacific forecast failure. *Wea. Forecasting*, in press.

Haltiner, G. and R. Williams, 1980: *Numerical Prediction and Dynamic Meteorology*. John Wiley and Sons.

Hamill, T. and S. Colucci, 1998: Evaluation of Eta-RSM ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev.*, **126**, 711–724.

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

Kiehl, J., J. Hack, G. Bonan, B. Boville, D. Williamson, and P. Rasch, 1998: The National Center for Atmospheric Research Community Climate Model: CCM3. *J. Climate*, **11**, 1131–1150.

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

Lorenz, E., 1965: A study of the predictability of a 28-variable atmospheric model. *Tellus*, **17**, 321–333.

— 1969: The predictability of a flow which possesses many scales of motion. *Tellus*, **21**, 289–307.

— 1982: Atmospheric predictability experiments with a large numerical model. *Tellus*, **34**, 505–513.

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

— 1994: Monte carlo simulations of explosive cyclogenesis. *Mon. Wea. Rev.*, **122**, 1548–1567.

Murphy, A., 1988: Skill scores based on the mean square error and their relationships to the correlation coefficient. *Mon. Wea. Rev.*, **116**, 2417–2424.

Savijärvi, H., 1994: Error growth in a large numerical forecast system. *Mon. Wea. Rev.*, **123**, 212–221.

Simmons, A. and A. Hollingsworth, 2002: Some aspects of the improvement in skill of numerical weather prediction. *Quart. J. Roy. Meteo. Soc.*, **128**, 647–677.

Simmons, A., R. Mureau, and T. Petroliagis, 1995: Error growth and estimates of predictability from the ECMWF forecasting system. *Quart. J. Roy. Meteo. Soc.*, **121**, 1739–1771.

Skamarock, W., J. Klemp, and J. Dudhia, 2001: Prototypes for the WRF (Weather Research and Forecasting) model. *Ninth Conference on Mesoscale Processes*, Amer. Meteor. Soc., J11–J15.

Smagorinsky, J., 1969: Problems and promises of deterministic extended range forecasting. *Bull. Amer. Meteor. Soc.*, **50**, 99–164.

Stensrud, D., J.-W. Bao, and T. Warner, 2000: Using initial condition and model physics perturbations in short-range ensemble simulations of mesoscale convective systems. *Mon. Wea. Rev.*, **128**, 2077–2107.

Strait, P., 1989: *A First Course in Probability and Statistics with Applications*. Harcourt Brace Jovanovich, 599 pp.

Toth, Z., 2001: Ensemble forecasting in WRF. *Bull. Amer. Meteor. Soc.*, **82**, 695–697.

Tribbia, J. and D. 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.

Ziehmann, C., 2000: Comparison of a single-model EPS with a multi-model ensemble consisting of a few operational models. *Tellus*, **52**, 280–299.