2.2

CALIBRATION OF ENSEMBLE SPREAD USING FORECAST SPECTRA

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

1. Introduction

Estimating internal (atmospheric) error growth by observing the real atmosphere would be ideal, but the absence of good analogues prevents a robust estimate (Van Den Dool 1994) and we are forced to use models. An understanding of the second-moment statistics (spatial variance and its spectral decomposition) of a NWP model is one step toward full utilization of the model as a substitute for the real atmosphere, facilitating prediction and predictability experiments. Examples include 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 internal and external (model) error growth. Ensemble forecasting in some capacity lies at the heart of much of this research, but quantitative results from ensemble forecasts often rely on the *assumption* that the forecast model reproduces the first and second moments of the climate of the real atmosphere.

In this paper we explore the quantitative relationship between second-moment NWP model forecast statistics and ensemble spread, and the effect on resulting error-growth estimates. The inability of a model to produce realistic error growth in either phase or amplitude is referred to as a deficiency, and it is really one manifestation of model error. It is also different from a phase or amplitude error in the traditional sense, though it may be related. Following Leith (1974), we start by asserting that NWP forecasts with scale-dependent spatial variance (thus spectra) similar to the real atmosphere are necessary to comprise an ensemble intended for predictability studies. Forecast energy spectra describe forecast amplitude, and we will show that in the absence of other factors, an amplitude-deficient model will produce spread-deficient ensembles. Further, knowledge of the spectral properties of a forecast model can explain the ensemble deficiencies when phase deficiencies are small or irrelevant. The concepts are demonstrated with a simple, non-dynamical, statistical model, and tested with three modern NWP models. Empirically measuring the amplitude deficiencies in spectral space allows a correction, or calibration, of ensemble spread to account for them. It also provides a method for predicting the effect of model changes on ensemble spread when those model changes primarily affect its ability to produce dispersion in forecast amplitude.

2. Ensemble response to damping

Assume we have a perfect model, and many forecast cases over which it can be evaluated. Further assume large ensembles for each of those cases, where each ensemble member is a forecast with the same model but started from slightly different initial conditions. 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 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 (e.g. Leith 1974; subscript P denotes any model forecast for a particular time period), and is fixed for a given calendar date. Because our model is perfect, $\sigma_Q^2 = \sigma_P^2$ in the climatological average, and we assert that they are both equal to the observed spatial variance of the atmosphere.

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

**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

be written as

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

In Fourier space (with coefficients denoted $\hat{*}$), the discrete spectral power of the difference field $P - Q$ is the ensemble spread as a function of wavenumber k and can be written

$$\hat{S}^2 = \hat{P}^2 + \hat{Q}^2 - 2\hat{P}\hat{Q} . \qquad (2)$$

Within our perfect model framework, we introduce one simulated source of model error in the form of a smoother R that reduces the forecast amplitude. In spectral space, applying the smoother to the model states amounts to the operations:

$$\hat{p} = R\hat{P} , \;\; \hat{q} = R\hat{Q} . \qquad (3)$$

With (3), the spread at wavenumber k that results from using damped model states is

$$\begin{aligned} \hat{S}_f^2 &= \hat{p}^2 + \hat{q}^2 - 2\hat{p}\hat{q} \\ &= R^2\left(\hat{P}^2 + \hat{Q}^2 - 2\hat{P}\hat{Q}\right) \qquad (4) \\ &= R^2\hat{S}^2 , \end{aligned}$$

Damping in a model reduces the ensemble dispersion by a greater fraction than it reduces the amplitude of the model states. For the discrete model at all scales, the spread in an ensemble of damped model forecasts is then given by

$$\sigma_{S_f}^2 = \sum_{k=1}^{N-1} \hat{S}_f^2 = \sum_{k=1}^{N-1} R^2\hat{S}^2 , \qquad (5)$$

where N is the total (unfolded) number of Fourier coefficients.

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 for a particular forecast case, then the instantaneous effect of damping on ensemble spread can be determined. If R is interpreted as a manifestation of model error, the instantaneous effect of that model error is known, and the spread of an ensemble of damped (erroneous) states can be calibrated to agree with the ensemble of undamped (perfect) states.

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 $R = \cos^2(\frac{\pi k n \Delta x}{N})$ (e.g. Haltiner and Williams 1980). To measure the effect of damping P and Q, and saturating the error spectra at $k \geq k_{sat}$, where k_{sat} is the largest scale for which the error is saturated, a large sample of random statistical realizations are created. The parameters of the experiments are chosen to qualitatively represent observed geopotential height spectra and error growth.

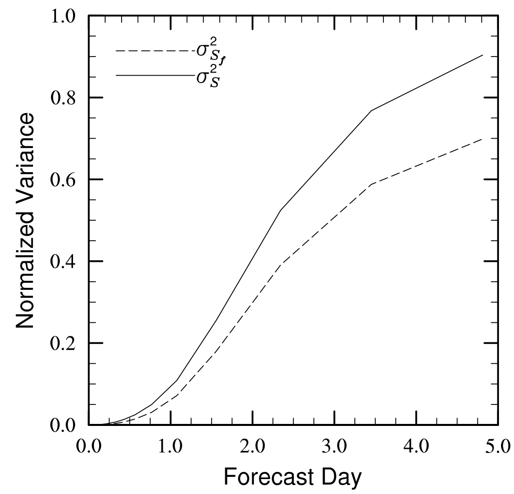

Figure 1: The effect of damping on error growth as estimated by ensemble spread for $N = 128$.

The error growth in time can be estimated by assuming an error doubling time of 2 d. Thus $\sigma_{S_f}^2$ and σ_S^2 are evaluated for $N = 128$ (Fig. 1). It is clear that damping inhibits both the growth rate and the asymptotic limit of ensemble spread. Repeating this computation for a range of N (not shown) reveals that the effect diminishes with increasing N (finer resolution for this fixed-size domain). 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, 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. Recall that the model here is assumed perfect and the climatological ensemble spread shown by the solid curve in Fig. 1 is equivalent to climatological error variance. Therefore, the real consequence of an under-dispersive ensemble resulting from using an overly-damped model in experiments is overly-optimistic estimates of predictability, and convergence toward the undamped curve mimics the convergence of error-growth estimates evident in the literature.

This example shows that an ensemble will be under-dispersive if it is comprised of model forecasts that are unrealistically damped or truncated compared to the real atmosphere. It also shows that instantaneous ensemble dispersion calculated with two models that have climatologically different spectra are linked by their spectral properties. This approach cannot address model differences in phase dispersion properties of a global model.

The spread in phase of an ensemble partially determines $\hat{S}^2$ without necessarily affecting $\hat{P}^2$ or $\hat{Q}^2$. Furthermore, this example uses a time-invariant response R that explains the instantaneous behavior of damped ensemble spread, and does not include dynamic scale interactions that contribute to dispersion and error growth.

With real NWP model runs, $\hat{p}$ and $\hat{q}$ are known, and R can be empirically estimated via (3) with two different model forecast or one model forecast and an analysis. In an ensemble experiment, $S_f^2(k)$ is also easily measured, and therefore $S^2(k)$ can be estimated with (4). The next section applies this procedure to the WRF model, and compares resulting dispersion characteristics.

3. Ensemble spread in two similar models

The rest of this paper is concerned with extending the concepts developed in section 2 to include time-dependent dynamics, and testing them with real NWP models run on a hemispheric domain with $\Delta X = 90$ km. We are interested in the spatial variance of the models, and the experiments here are controlled to isolate those. 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.

The climate of the WRF relative to the real atmosphere is not important for these experiments and we consider the undamped WRF, and ensembles run with it (denoted *WRF*), to be the references against which model changes are evaluated. One source of simulated model error is introduced by applying a 2-D, second-order diffusion term that explicitly (and locally in time) reduces spatial forecast variance at small scales. A response function R calculated from damped and undamped WRF runs reveals the effects of the model error in spectral space. Ensembles run with the damped version of the WRF (denoted *DMP*) are compared to ensembles *WRF*, and R is used to explain the time-varying differences in ensemble spread.

Ten-member ensembles are generated with the Errico-Baumhefner technique, which approximates random analysis errors. Control (unperturbed) initial conditions are given by the NCEP final analyses. To avoid the danger of obtaining regime-dependent results, all results are averaged over six forecast periods from the 2001-2002 cool season in the northern hemisphere.

If R carries time-integrated dynamics information on the effect of damping, (4) may be useful for predicting and explaining the dispersion of an ensemble. This can be easily checked by computing $R(k,t) = \hat{p}(k,t)/\hat{P}(k,t)$, and using (4) to correct the damped dispersion $\hat{S}_f^2(k,t)$ to $\hat{S}^2(k,t)$. Rather than computing dispersion in physical space, the dispersion spectrum is integrated in k to get the total dispersion at each time for each forecast case. The result, averaged over all the cases, is shown in Fig. 2. 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 on scale interaction that is required to relate ensemble dispersion with these two different models.

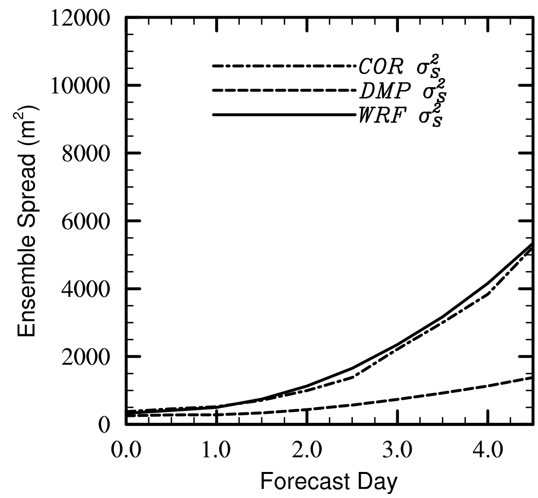

Figure 2: Dispersion of the ensemble with undamped (*WRF*) and damped (*DMP*) WRF forecasts. The corrected (*COR*) dispersion results from using the R^2 to predict the dispersion without damping.

Spread in the *WRF* ensembles grows faster than spread in the *DMP* ensembles over the first 4.5 forecast days. 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. In this case the correction applied to *DMP* could prevent the underestimation.

To better understand the effect of damping the WRF, we can compare it to the undamped WRF in spectral space for a 6-d forecast lead time (Fig. 3). In panel (a), the WRF control forecast spectrum is shown for comparison (thin solid line). When undamped, the dispersion is saturated through the high-wavenumber part of the spectrum. But the dispersion of damped forecasts does not saturate at any scale, relative to the undamped WRF forecast, and the total dispersion (area under the curve) of *DMP* lags below the dispersion of *WRF*. The spectral response function $R < 1$ everywhere in Fig. 3b. This is the ratio of damped to undamped spectra, and consistent with section 2 it appears to be responsible for the under-dispersion of the *DMP* ensemble that results from amplitude deficiencies in the damped WRF. If the

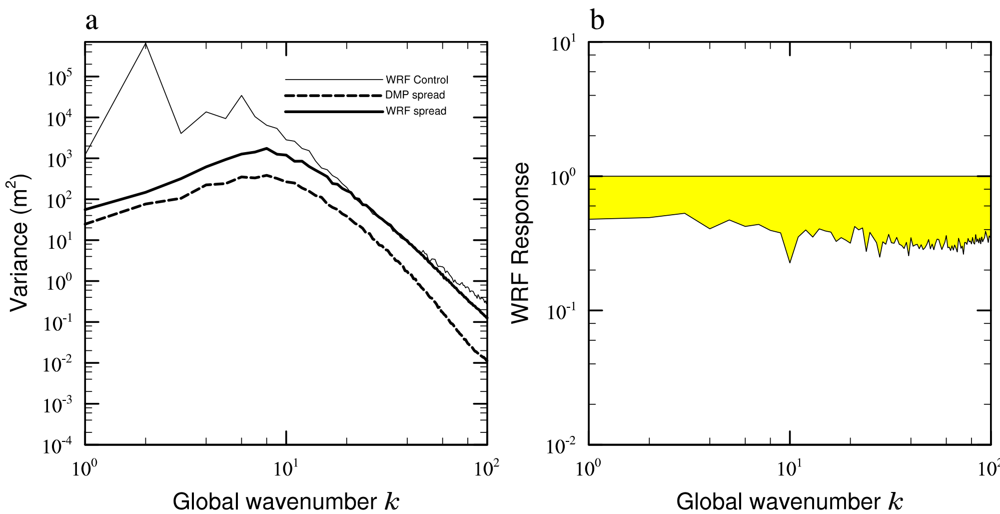

Figure 3: In (a), the undamped (thick dotted line) and damped (thick solid line) WRF 50.0 kPa geopotential height spread in spectral space, compared to the undamped WRF control forecast spectrum (thin solid line). In (b), a response function is shown, calculated from the ratio of damped to undamped WRF forecast spectra. Results are for 6-d forecasts.

forecast amplitude of the damped and undamped WRF were the same, averaged over these six cases, R would be identically one. But here a time-varying R is used to construct curve *COR* from curve *DMP* in Fig. 2.

This calibration was also tested against ensembles generated with the CCM3 model (Kiehl et al. 1998), which is fundamentally different from the WRF. Because phase differences are important, the calibration was not successful.

4. Conclusions

The most important product of this research is a method to calibrate ensembles with models that are deficient in spatial variance. In effect, a model can be calibrated for predictability research. The time-dependent spectral response is trivial to estimate using two categorical forecasts with different models. The fact that it can be used to almost entirely explain the differences between two ensemble dispersion curves demonstrates that it carries time-integrated spectral characteristics. Knowledge of it also allows prediction of ensemble dispersion changes that are expected from model changes affecting the spectral characteristics. This could be used, for example, to adjust the ensemble spread of a coarse-resolution model to account for the forecast by a high-resolution model.

The comparison of ensembles with damped and undamped versions of the WRF showed that knowledge of forecast spatial variance, and its spectral decomposition, is necessary to interpret error-growth estimates with ensembles. But the quantitative relationship between spectral response and dispersion did not hold for a model with different phase error properties, showing that correct spectral statistics in a forecast model are not sufficient for determining ensemble spread. Models with different phase-error growth characteristics result in ensembles with different phase-dispersion characteristics.

REFERENCES

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

Kiehl, J. T., J. J. Hack, G. B. Bonan, B. A. Boville, D. L. Williamson, and P. J. 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.

Van Den Dool, H., 1994: Searching for analogues, how long must we wait? *Tellus*, **46**, 314–324.