The impact of ensemble forecasts on predictability

By J. M. MURPHY
Meteorological Office, Bracknell

(Received 8 January 1987; revised 10 September 1987)

SUMMARY

An estimate of the mean effect of ensemble-averaging on forecast skill, under idealized 'perfect model' conditions, is obtained from a set of eight independent 50-day winter ensemble forecast experiments made with a hemispheric version of the Meteorological Office (UKMO) 5-level general circulation model. Each ensemble forecast consisted of seven individual integrations. Initial conditions for these were obtained by adding spatially correlated perturbations to a given wintertime analysis, and a further integration created in the same manner was used to represent nature, giving the perfect model approach.

The ensemble-mean forecast shows a clear improvement in amplitude and phase skill compared with individual forecasts, the period of significant predictability for daily fields being increased by 50%. The improvement in skill is consistent with simple theoretical estimates based on the perfect model assumption. These calculations are used to deduce how ensemble-mean forecast skill should vary with the size of ensemble. The superiority of the ensemble-mean is maintained when forecasts are spatially smoothed or time-averaged.

The spread of an ensemble distribution can in principle give an a priori indication of forecast skill. A moderate level of correlation between ensemble spread and the forecast skill of the ensemble-mean is found on the hemispheric scale.

The extent to which the potential benefits of ensemble forecasting may be achieved in reality depends on the model's practical forecast skill. Since the practical skill of the 5-level model is rather low, an ensemble-mean forecast is on average no better than an individual forecast up to the normal limit of deterministic predictability. However, in four experiments where the individual forecasts show skill beyond this point, the ensemble-mean forecast does give increased skill.

Spatial variations in both the practical and perfect model skills of an ensemble-mean anomaly field are found to be related to corresponding variations in the statistical significance of the anomaly field. For example, the average perfect model skill, in regions where the ensemble-mean anomaly is significantly different from zero, exceeds the full field skill in all experiments for forecast days 1–15, and in all but two cases for days 16–30.

1. INTRODUCTION

Two stumbling blocks hinder the production of useful weather forecasts beyond a few days ahead. Firstly, no numerical weather prediction (NWP) model with a finite grid size can ever represent the full range of atmospheric motions with perfect accuracy; and secondly, any observed atmospheric state used as the initial conditions for a forecast cannot be measured precisely. The instability of the atmospheric equations of motion to small perturbations ensures that small errors arising from either source will grow with time until they reach saturation level. Lorenz (1982) puts the limit of useful skill for individual forecasts of instantaneous weather patterns at ~10 days. To produce useful extended-range predictions we must find a means of forecasting beyond this point.

The methods to consider depend on the time scale of interest. For a seasonal forecast we expect the initial conditions to be unimportant and would seek the atmospheric response to some best estimate of the anomalous boundary conditions (sea surface temperature, soil moisture, snow and sea-ice cover) likely to prevail over the forecast period. On the monthly time scale the boundary conditions would undoubtedly play a role, but the initial conditions might also be important. In this paper we shall consider a means of increasing our ability to extract any predictability arising from the initial conditions.

Atmospheric predictability is a function of wavenumber, since errors grow more quickly in synoptic scales than in planetary scales (Shukla 1981; Bengtsson and Simmons...
1983; Boer 1984). Thus, removing the smaller scales from individual forecasts may provide a modest extension of the predictability limit. Potentially a more powerful approach is to run an ensemble of individual forecasts, as explained below. For ease of discussion it is assumed that the forecast model is perfect, so that forecast errors arise solely from incorrect initial conditions, and no climate drift occurs in the model.

The state of a system may be represented by a point \( u \) in a suitable phase space, each dimension of which corresponds to a system variable (e.g. Gleeson 1970). Thus the phase space for the atmosphere has an infinite number of dimensions, whereas for an NWP model the dimensionality is large but finite, typically \( \sim 10^5 \). An ensemble of independent analyses of a given atmospheric state, each based on a separate set of observations, corresponds to a 'cloud' of points in phase space centred (in the absence of bias in the observing system and analysis method) about the true state with a spread reflecting observational and analysis errors. This cloud approximates the continuous probability density function (p.d.f.) \( \psi(u; t=0) \) for the true initial state, which we approach as the number of analyses tends to infinity. Running a forecast from each analysis yields a set of trajectories in phase space giving the time evolution of the cloud of phase points, representing a sampling approximation to the evolution \( \psi(t) \) (dropping the explicit dependence on \( u \)), of the p.d.f. for the true state. Due to atmospheric instability the cloud expands with time until eventually its centroid becomes indistinguishable from the climate mean and the average separation of phase points becomes equal to the separation of randomly chosen atmospheric states, at which stage it is no longer possible to forecast above the level of chance.

Thus, uncertainty in the initial state measurement gives the forecasting problem a stochastic nature, such that our knowledge of the initial atmospheric state merely determines a probability distribution for its state at future times. In the most general sense our task in making a forecast is to specify that distribution as completely as possible. If we wish to present the forecast as a single best-estimate prediction, an estimate of its centroid is required, which will be more accurate if it is based on more than one integration. Also, the spread of an ensemble of forecasts gives a measure of the uncertainty associated with the ensemble-mean, which represents an advance estimate of the likely skill of the forecast. The ensemble forecast is a practical approximation to a general method of stochastic dynamic prediction (Epstein 1969) in which the evolution of \( \psi(t) \) is found by solving the continuity equation for probability in phase space.

The potential benefits of an ensemble forecast are readily apparent if we assume a perfect model. In practice these advantages will be eroded to a degree dependent on the extent of the model's deficiencies. Previous results from ensemble forecast experiments have mainly been obtained under perfect model conditions. Leith (1974) demonstrated the superiority, in an r.m.s. sense, of an ensemble-mean forecast compared with an individual forecast. Seidman (1981) compared the effects of space, time and ensemble averaging on predictability using a single ensemble of 34 integrations produced with a 3-level GCM. The divergence of small groups of individual GCM integrations created by adding small perturbations to a given observed state has been studied by Spar et al. (1978) and Shukla (1981). One aim of the present paper is to determine the mean impact of ensemble forecasts on predictability from a sizeable number of independent experiments, conducted under perfect model conditions using a realistic GCM. The practical skill of the forecasts is also considered. Hoffman and Kalnay (1983) avoided the perfect model assumption by using a low-order primitive equation model to represent nature while creating forecasts with a quasi-geostrophic version of the model. Ensembles based on successive analyses (see next section) gave better results than those based on perturbations, both in terms of forecast skill, and of skill prediction using ensemble
spread. Recently Miyakoda et al. (1986) have reported a set of practical monthly forecasts for eight January cases using a global 9-level GCM, where each forecast was a three-case ensemble initialized using analyses from three different meteorological centres. They find that ensemble-averaging reduces the forecast r.m.s. errors, and also note a slight increase in the anomaly correlation score of the ensemble-mean compared with its constituent individual forecasts.

2. ENSEMBLE FORECAST METHOD

The set of initial conditions for an ensemble forecast must reproduce \( \psi(o) \) as closely as possible within sampling limits. In practice we have only a single imperfect initial analysis plus an idea of its associated errors from our knowledge of the observing system and analysis method. Two methods of obtaining estimates of \( \psi(o) \) are the random perturbation (RP) approach (used in this paper), and the lagged-average forecast (LAF) approach.

(a) Random perturbation approach

We may generate a cluster of possible initial states by adding a series of independent perturbations to the analysed initial state consistent with observation and analysis errors. Provided that the analysed state is free of systematic bias we end up with a distribution \( \psi'(o) \) of possible true states centred about the analysed state. But \( \psi(o) \) actually represents the distribution of possible analysed states centred about the true state. Thus in general the centre of gravity of \( \psi'(o) \) is not coincident with the true state.

The simplest approach is to perturb the analysed state independently at each grid point according to the size of typical observation and analysis errors (e.g. Seidman 1981; Shukla 1981; Spar et al. 1978). However, this takes no account of covariances between errors at neighbouring grid points. Spurious high frequency oscillations are set off by consequent imbalances in each perturbed initial state, and their dissipative effect on the perturbation field tends to result in an unrealistic decrease in the spread of the ensemble during the early stages of the forecast. Thus the 'effective' spread of the initial state distribution is actually smaller than intended.

In the present work, to ameliorate this effect, each perturbed state is created by taking a linear combination of the analysed state and an independent randomly chosen winter analysis. The weights for the linear combination are calculated by demanding that the r.m.s. 500 mb height difference between the perturbed state and the original analysed state should be 30 m, taken to represent a typical analysis error. The linear combination so deduced is then applied to all model variables to generate the perturbed state. The resulting perturbation field is then such that the perturbations at neighbouring points are correlated. A small penalty is paid for this advantage in that the centre of gravity of the distribution \( \psi'(o) \) obtained by this approach is not coincident with the analysed state, being slightly biased towards the climate mean by an amount typically corresponding to about one half of the assumed analysis error. This, however, is felt to be more than offset by the benefit of obtaining a more realistic effective spread of the initial state distribution.

In order to set up an RP ensemble of seven individual forecasts six perturbed states are created in the manner described, using a different winter analysis to provide the perturbation field in each case. These six states plus the original analysed state constitute a set of seven initial conditions for the ensemble. This crude random perturbation method takes no account of systematic errors, the variation of typical error size with position or of the off-diagonal elements of the error covariance matrix specifying correlations
between errors. To do so would be a complex exercise which we have chosen to avoid. It is to be hoped that the prospects for extended-range prediction do not depend crucially on such details of the perturbation procedure.

(b) Lagged-average forecast approach

An alternative approach is to use forecasts made from a number of consecutive analyses, each separated by a constant time interval, with the latest corresponding to the start of the forecast period (Hoffman and Kalnay 1983). We may interpret this procedure as the use of the dynamical model to extrapolate previous analyses up to the base analysis time to obtain an estimate of $\psi(0)$, with the subsequent evolutions of the individual integrations forming the ensemble forecast.

In general it is not clear which method is better (Hoffman and Kalnay 1983; Murphy 1986a). In our case, however, the issue is clear cut. Only midnight analyses were available, so the interval, $\tau$, between successive members of an LAF ensemble would be 24 hours. As the model used (see section 4) has relatively poor short-range skill (Mansfield 1986), the prediction errors incurred in extrapolating the lagged observations for several days would be so large that the spread of $\psi'(0)$ would seriously overestimate the true uncertainty in the initial state (Murphy 1986a). In this study we therefore use the RP method.

(c) Presentation of ensemble forecast

Running an integration from each initial state contained within $\psi'(0)$ gives an estimate $\psi'(t)$ of $\psi(t)$. From a statistical viewpoint, the most natural approach might be to present the forecast as a series of alternative evolutions, each with an associated probability. For example, following the work of Charney and DeVore (1979), interest has grown in the idea that the atmosphere may possess a number of distinct circulation regimes, as discussed by Leith (1983). If the atmosphere spent most of its time locked in one or other of these regimes, with intermittent rapid transitions between them, one could imagine that $\psi(t)$ would typically exhibit several distinct modes, each corresponding to a different regime. In such a situation the assignment of a probability to each mode of $\psi'(t)$ would be the best way to present an ensemble forecast.

On the other hand if $\psi(t)$ typically resembles a normal distribution (Seidman 1981), all the important information is contained in the first and second moments, which then furnish respectively a single best-estimate forecast and an indication of its associated uncertainty. Expression of the ensemble forecast in this manner, which corresponds to the traditional analysis of predictability in terms of the progressive submergence of deterministic information by the steady growth of small errors, is the approach we choose to adopt. A search for clustering in the forecast distributions is beyond the scope of the present work, but is being pursued (Murphy and Palmer 1986).

In general the individual integrations within an ensemble may have different associated degrees of uncertainty, in which case different weights must be assigned to them when forming the ensemble-mean. For an RP ensemble all the individual forecasts made from perturbed initial conditions possess the same error statistics and are therefore given equal weight. The original analysed state is a better estimate of the true state, so strictly the forecast from that state should be weighted more highly. However, beyond the short-range period this factor is unlikely to be important, so this forecast is given equal weight.
3. THEORETICAL AND PRACTICAL PREDICTABILITY

(a) Individual forecasts

Theoretical predictability studies consider the divergence of pairs of model realizations caused by the amplification of small initial differences. This error growth arises solely from the internal dynamics of the model, hence the results represent an upper limit to the degree of skill obtainable in practice, where the effects of model imperfection accelerate the decay of predictability. The gap between the practical and theoretical limits shows the scope for further improvement through the development of more accurate forecast models.

There is no guarantee that error growth statistics observed for pairs of closely related model states will match the required, but unknown, error growth statistics for pairs of closely related atmospheric states. In fact, whilst estimates of the practical limit have increased due to the development of superior models, corresponding estimates of the theoretical limit have decreased. Lorenz's (1982) results from the ECMWF model place the practical limit at \( \sim 10 \) days and the theoretical limit at \( \sim 14 \) days for predictions of instantaneous weather patterns.

(b) Ensemble forecasts

In the present paper we wish to measure both the practical and theoretical predictability of ensemble-mean forecasts relative to that of individual forecasts. The theoretical argument for the superiority of the ensemble-mean forecast has already been discussed; however, to achieve increased practical skill, the model must be sufficiently skilful that the distance in the model phase space between the ensemble-mean forecast and the state of the real atmosphere is of the same order as the ensemble spread. The ensemble-mean is then closer to the real state than a typical individual forecast and therefore has more skill. If, however, the forecast model is poor, the real atmospheric state typically diverges from the ensemble-mean forecast very quickly compared with the rate at which the ensemble spreads out (see Fig. 1), so that the ensemble-mean forecast is little or no better than an individual forecast. Since our model has only moderate forecast skill our results are likely to be pessimistic in terms of what might be achievable.

![Figure 1](image)

Figure 1. Schematic typical evolution of ensemble forecast and real atmosphere for: (a) a poor forecast model, and (b) a good forecast model. The axes represent the model phase space. \( \times \) stands for ensemble-mean forecast and \( \circ \) for the real atmosphere. Each dot represents an individual forecast. \( t = 0 \) corresponds to the start of the forecast period and \( t = t_f \) represents a later point at which the dispersion of the ensemble has increased but has not yet reached saturation.
with an up-to-date NWP model. Thus we may regard our estimates of the theoretical and practical predictability of ensemble-mean forecasts, relative to individual forecasts, as upper and lower extremes (respectively) of the impact on practical forecast skill attainable in future with superior forecast models.

(c) Variation of predictability

The predictability of the atmosphere varies considerably between different synoptic situations. Evidence has been presented (e.g. Miyakoda et al. 1972; Mansfield 1986) of cases where individual model forecasts showed considerable skill at ranges well beyond the usual deterministic limit, and Bengtsson and Simmons (1983) have demonstrated the variability in skill of the ECMWF medium-range forecasts.

This variability has important implications for the role of ensemble forecasts. In forecasting for a month ahead, we must recognize that while ensemble-mean forecasts may provide some skill beyond the deterministic predictability limit for single forecasts, the intrinsic instability of the atmosphere will often still prohibit any possibility of achieving skill beyond about two weeks. Thus the performance of ensemble forecasts in unusually predictable situations is of particular interest, especially if such cases can be identified in advance. In principle the skill of an ensemble-mean forecast can to some extent be predicted from the spread of its distribution measured, for example, by its variance (Hoffman and Kalnay 1983). However, the correlation between ensemble spread and practical skill will be reduced by the effect of model imperfection, and may be negligible for relatively poor forecast models. The correlation between theoretical skill and ensemble spread (see sections 5 and 6) will indicate the maximum level of correlation attainable in practical forecasts with more accurate models.

Returning to the phase space description of atmospheric evolution, we may regard instances of abnormally high predictability as occasions where for some reason the rate of divergence of individual model forecasts from the real state is smaller than normal, so that the predicted phase paths remain close to the actual phase path for a longer period than usual. If the typical performance of the model were represented by a diagram like Fig. 1(a), then its performance in an unusually predictable situation might be represented by Fig. 1(b). Thus the practical skill of an ensemble-mean forecast might significantly outstrip that of an individual forecast in such cases, even if the normal performance of the model was so poor that an ensemble-mean forecast usually offered little or no advantage.

Mansfield (1986) demonstrates that, despite its generally moderate level of practical predictability, our model has on several occasions produced forecasts which showed skill out to 50 days, even without including anomalous boundary forcing conditions. We have therefore been able to set up experiments to consider both the average skill of an ensemble-mean forecast and also its skill in unusually predictable situations.

(d) Specification of ‘nature’ in theoretical predictability experiments

To measure theoretical forecast skill, we must use the model to generate a simulated evolution of the ‘real atmosphere’ against which to verify the forecast. The initial state for this integration, termed a ‘nature run’, is obtained by perturbing the analysed state in the same manner employed in the creation of initial conditions for an RP ensemble.

By generating such a nature run for each RP ensemble forecast we can use the same set of experiments to study both practical and theoretical predictability (Fig. 2). The only difference between a nature run and the integrations from perturbed initial conditions in the corresponding RP ensemble forecast is that we arbitrarily assume the initial state for the nature run to be the true state. (The unperturbed analysis cannot be used as the
initial conditions for the nature run, unless it is intended to regard the ensemble members as having arisen from independently obtained analyses, rather than independent perturbations of a single analysis.)

4. **Numerical Experiments**

   (a) **Model**

   The model used was a hemispheric version of the UKMO 5-level GCM (Corby et al. 1977) which has a quasi-uniform horizontal grid with a resolution of approximately 330 km. The initial data for the experiments were obtained from Meteorological Office operational analyses. Further details of the model and initialization procedure are given by Mansfield (1986). Climatological boundary forcing conditions were employed in all integrations.

   (b) **Experiments**

   A series of 50-day ensemble forecasts, each containing seven individual integrations, was produced using winter initial conditions taken from several different years. Eleven RP ensemble forecasts were created in all (see Table 1). For each of these a further individual integration was produced to act as the nature run (see above). Of the eleven RP cases, eight (hereafter referred to as RP\textsubscript{A} ensemble forecasts) were created from independent winter initial conditions chosen at random, to obtain a representative

|---------------------|---------|---------|---------|---------|---------|---------|---------|---------|

<table>
<thead>
<tr>
<th>Forecast category</th>
<th>RP\textsubscript{A}</th>
<th>RP\textsubscript{A}</th>
<th>RP\textsubscript{A}</th>
<th>RP\textsubscript{A}</th>
<th>RP\textsubscript{A}</th>
<th>RP\textsubscript{A}</th>
<th>RP\textsubscript{A}</th>
<th>RP\textsubscript{A}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initialization date</td>
<td>23 Dec.</td>
<td>23 Dec.</td>
<td>31 Dec.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1974</td>
<td>1975</td>
<td>1976</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Forecast category</td>
<td>RP\textsubscript{B}</td>
<td>RP\textsubscript{B}</td>
<td>RP\textsubscript{B}</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
estimate of the average impact of the ensemble-mean forecast on predictability. A further three RP ensemble forecasts (hereafter denoted by \(RP_B\)) were created from initial conditions specially picked out as cases where an individual forecast run previously had shown an unusually high degree of skill, and were thus considered separately.

(c) **Model climatology**

In general the climate of a GCM drifts with time away from the true atmospheric climate as systematic errors in the model circulation develop. Mansfield (1986) has shown that this is certainly true of the model used in this study. As in Mansfield's paper, we account empirically for the explicit effect of climate drift by calculating the forecast anomalies relative to a suitable estimate of the model's own climatology, created by averaging the eight nature runs corresponding to the eight \(RP_A\) experiments.

5. **Statistics of Ensemble Forecasts**

In this section we examine the statistical properties of ensemble forecasts in order to set up a theoretical background to the experimental results presented subsequently. Estimates of forecast skill and its relationship to ensemble spread under perfect model conditions may be obtained using certain basic assumptions:

1. the model is perfect;
2. each individual integration in an ensemble is equally likely to represent the true state evolution;
3. \(\psi'(o)\) is identical to \(\psi(o)\) within sampling limits.

Limitations to these assumptions have been discussed in sections 2 and 3.

An individual model forecast \(j\), selected from the infinite number making up a particular forecast p.d.f. \(\psi(t)\), may be represented by a point \(u_j(t)\) in phase space. At any forecast time \(t\) we may take the origin of phase space to be coincident with the model climate mean, in which case the elements of the vector \(u_j\) represent anomalies. If \(\langle \cdot \rangle\) denotes an average over an infinite number of integrations, chosen at random from an infinite selection of independent forecast p.d.f.s, then \(\langle u_j(t)\rangle = 0\). We may term \(\langle \cdot \rangle\) a 'climatic' average, as it represents an average over states selected randomly from the model's climate distribution. We shall use a hat to denote an average over the \(M\) integrations of an ensemble forecast, so that \(\hat{u}(t)\) is the ensemble-mean, or centroid, of \(\psi'(t)\). An average over a forecast p.d.f. is represented by an overbar, so that \(\overline{u}(t)\) is the centroid of \(\psi(t)\), towards which \(\hat{u}(t)\) tends as \(M\) approaches infinity, under assumption (3).

Under our assumptions, at \(t = 0\), the squared distance \(d_0 = (u_0 - \overline{u})^2\), from the true state \(u_0\) to the centroid of \(\psi(o)\), is zero. However, if \(d_j = (u_j - \overline{u})^2\) is the corresponding quantity for forecast \(j\), then \(D = \overline{d_j}\) has some finite non-zero value at \(t = 0\), reflecting the mean size of observation and analysis errors. Despite this, we shall assume for simplicity that \(\langle d_0 \rangle = \langle D \rangle\) for \(t > 0\), which should be reasonable beyond the very early stages of the forecast period. Our idealized framework is then effectively identical to that used in the discussions of ensemble forecast statistics given by Leith (1974) and Hayashi (1986). However, we shall consider skill scores different from that used by the above authors, and shall also consider the correlation between forecast skill and ensemble spread. (The reader should note that our notation is not identical to that used in either of the above papers.) Although forecast p.d.f.s may in general take complex forms (section 2), we restrict ourselves to consideration of the first two moments (following Leith 1974), thus making no attempt to distinguish \(\psi(t)\) from a normal distribution.
We shall omit detailed derivations of the results presented. These may be found in Murphy (1986b), see also the appendix. We shall also omit explicit time dependences, although all the statistics discussed are, in general, functions of \( t \).

(a) \textit{Amplitude measure of spread and skill}

The quantity \( D \) defined above represents the variance of \( \psi \), which is a suitable measure of its spread in 'amplitude' terms. An estimate of \( D \) is available from an ensemble of size \( M \) through the ensemble variance \( s_M \), given by

\[
s_M = \left\{ \sum_{j=1}^{M} (u_j - \bar{u})^2 \right\} / M = \{(M - 1)/M\} \hat{d} + (2/M^2) \sum_{j=1}^{M} \sum_{l=j+1}^{M} (u_j - \bar{u}) \cdot (u_l - \bar{u}). \quad (1)
\]

The second term in Eq. (1) represents a sum of random covariances, whose average value is zero, so that if we draw a large number of independent ensembles from \( \psi \) we find

\[
S_M = \bar{s}_M = \{(M - 1)/M\} D.
\]

At initialization time the size of \( D \) reflects the observation and analysis errors. As the forecast proceeds \( D \) increases and eventually approaches the variance of the model's climate distribution, at which point the limit of amplitude predictability is reached. If \( w_j = u_j^2 \) is the 'anomaly intensity' of a forecast selected from this climate distribution, the climate variance may be written as \( \langle w_j \rangle \). The variance ratio \( F = D / \langle w_j \rangle \) represents the predictability information carried by \( D \). In practice we must estimate \( \langle w_j \rangle \) from a set of \( N \) independent integrations \( u_k(t), k = 1 \rightarrow N \), representing the model climatology, which possesses a variance \( \nu \) given by

\[
\nu = \sum_{k=1}^{N} (u_k - \bar{u}_k)^2 / N
\]

where the tilde denotes an average over the \( N \) integrations. Thus

\[
\langle \nu \rangle = \langle w_k \rangle - \langle (\bar{u}_k)^2 \rangle = \{(N - 1)/N\} \langle w_k \rangle.
\]

In each of our RP ensemble experiments we may measure \( s_M \) and use

\[
f = M(N - 1)s_M / N(M - 1)\nu
\]

as an unbiased estimate of \( F \).

An obvious skill score which corresponds to the ensemble variance is the error variance, the square of the distance in phase space between the forecast and the actual state. The error variance \( e_M \) of an ensemble-mean forecast is

\[
e_M = (\bar{u} - u_0)^2 = (\hat{d}/M) + d_o - 2(\bar{u} - \bar{u}) \cdot (u_0 - \bar{u}). \quad (3)
\]

Thus

\[
E_M = \bar{e}_M = D / M + d_o, \quad (4)
\]

and

\[
\langle E_M \rangle = \{(M + 1)/M\} \langle D \rangle = \{(M + 1)/2M\} \langle E_1 \rangle. \quad (5)
\]

We may argue from Eq. (5) that since \( \langle E_M \rangle \) is smaller than \( \langle E_1 \rangle \) an ensemble-mean forecast is superior to an individual forecast (Leith 1974). Although true in a least-squares sense Eq. (5) is somewhat misleading since part of the reduction in error variance is caused simply by the reduction in anomaly intensity (smoothing) resulting from the
formation of an ensemble-mean. Smoothing a forecast towards the climate mean does not by itself improve the signal-to-noise ratio of the forecast. The real benefit lies in the fact that compared with an individual forecast, \( \tilde{u} \) is a better estimate of \( \bar{u} \). We may suppress the effect of such smoothing in our measurement of skill by expressing \( e_M \) as a fraction of the mean error variance between ensemble-mean forecasts and unrelated true states. This normalized error variance \( e'_M \) is thus \( e'_M = e_M / (w_M + w_o) \), where \( w_M \) and \( w_o \) are the anomaly intensities \( \tilde{u}^2 \) and \( \bar{u}_0^2 \) of an ensemble-mean forecast and a true state respectively. (Note that \( \langle w_o \rangle = \langle w_j \rangle \), the climate variance.) Noting that the climatic average normalized error variance for individual forecasts is \( \langle e'_1 \rangle = \langle D \rangle / \langle w_o \rangle \), it may be shown that

\[
\langle e'_M \rangle = \frac{\langle e'_1 \rangle}{\langle e'_1 \rangle + (2M/(M + 1))(1 - \langle e'_1 \rangle)}.
\]

Thus \( \langle e'_M \rangle \) decreases as \( M \) increases, and \( \langle e'_M \rangle \to \langle e'_1 \rangle / (2 - \langle e'_1 \rangle) \) as \( M \to \infty \). This increase in skill is attributable solely to the decrease in sampling error as the ensemble size increases, in contrast to Eq. (5). Note that as \( t \to \infty \), \( \langle e'_1 \rangle \) and \( \langle e'_M \rangle \to 1 \), representing complete loss of predictability.

(b) Phase measure of spread and skill

We now consider the complementary question of the correspondence in phase between the forecast and actual anomaly patterns. In any particular forecast p.d.f., it is possible that phase predictability may remain once amplitude predictability has disappeared, or vice versa, so we must measure both to determine fully the skill of a given forecast.

The phase dispersion, \( R \), of \( \psi \) may be defined as the square of the anomaly correlation between \( \bar{u} \) and all the individual forecasts \( u_j \), within \( \psi \), taken together. Thus, if \( w_\infty = \bar{u}^2 \),

\[
R = \frac{\langle u \cdot u_j \rangle^2}{w_\infty w_j} = w_\infty / w_j.
\]

From an ensemble of \( M \) forecasts we may estimate \( R \) using \( a_M \), the square of the anomaly correlation between the ensemble-mean and all the individual forecasts in the ensemble, taken together, which is given by

\[
a_M = \left( M^{-1} \sum_{j=1}^{M} \tilde{u} \cdot u_j \right)^2 / w_M w_j = w_M / w_j.
\]

The relationship between \( R \) and \( A_M = \tilde{a}_M \), the mean phase dispersion of size \( M \) ensembles drawn independently from \( \psi \), follows from the above definitions:

\[
A_M = \frac{1 + (M - 1)R}{M}
\]

assuming that sampling variations \( \delta w_j = \hat{w}_j - \bar{w}_j \) of \( w_j \) are small. As for the amplitude case (Eq. (2)), the ensemble dispersion is not an unbiased estimate of the dispersion of the forecast p.d.f. The amplitude and phase measures of spread are related through \( a_M = 1 - s_M / \hat{w}_j \) and

\[
R = 1 - D / \bar{w}_j = 1 - F(w_j) / w_j.
\]

Typically \( R \) has a value just below unity at initialization time. During the forecast \( R \) decreases due to the movement of \( \bar{u} \) towards the climate mean concomitant with the spreading of the ensemble. Eventually \( R \to 0 \) as \( t \to \infty \). Equation (5) demonstrates mathematically the complementary nature of amplitude and phase predictability discussed above. Even if \( F = 1 \) we may still have \( R > 0 \) if \( \langle w_j \rangle / w_j < 1 \). Similarly if \( R = 0 \) we may
still have \( F < 1 \) if \( \langle w_j \rangle / \bar{w}_j > 1 \). However, in the climatic average sense we may deduce the amplitude predictability from the phase predictability and \textit{vice versa}, since \( \langle R \rangle = 1 - \langle F \rangle \).

Ensemble-mean forecast skill may be measured by the anomaly correlation \( c_M \) between the ensemble-mean and the true state:

\[
c_M = \hat{u} \cdot \mathbf{u}_0 / (w_M w_0)^{1/2}.
\]

From this definition we may determine the variation of \( \langle c_M \rangle \) with ensemble size. In fact, assuming again that variations in anomaly intensity are small,

\[
\langle c_M \rangle = M^{1/2} \langle c_1 \rangle / (1 + (M - 1) \langle c_1 \rangle)^{1/2} \rightarrow \langle c_1 \rangle^{1/2} \quad \text{as} \quad M \rightarrow \infty
\]

where \( c_1 = \mathbf{u}_j \cdot \mathbf{u}_0 / (w_j w_0)^{1/2} \) is the corresponding anomaly correlation for an individual forecast.

(c) \textit{Amplitude correlation between spread and skill}

As discussed previously, an important issue concerns the extent to which case-by-case variations in skill may be predicted by corresponding variations in ensemble spread. We seek theoretical estimates of this which may subsequently be compared with experimental results.

We may use a correlation coefficient (defined below) to measure the relationship between the ensemble spread \( s_M \) and the skill \( e_M \) of the ensemble-mean forecast. For simplicity we use error variance \( e_M \) rather than normalized error variance \( e'_M \) as the skill score, since the correlation remains unaltered if we normalize the error variance. The correlation \( \rho_A \) between \( s_M \) and \( e_M \) is defined as

\[
\rho_A = \frac{\langle (e_M - \langle e_M \rangle)(s_M - \langle s_M \rangle) \rangle}{\left[ \langle (e_M - \langle e_M \rangle)^2 \rangle \langle (s_M - \langle s_M \rangle)^2 \rangle \right]^{1/2}}.
\]

The skill of different ensemble-mean forecasts may vary for two reasons. One is the sampling error incurred through approximating a forecast p.d.f. with an ensemble of finite size, and the other is the genuine variation in predictability between different forecast p.d.f.s, which we may term climatic variation. If \( \delta c_D = D - \langle D \rangle \) represents the deviation in the variance of a forecast p.d.f. from the climatic average value, then the quantity \( \langle (\delta c_D)^2 \rangle \) may be used to characterize the mean climatic variation in predictability. Sampling variations in estimating predictability may be represented by \( \langle (\delta d_j)^2 \rangle \), where \( \delta d_j = d_j - D \) shows the amount by which the square of the distance in phase space from forecast \( j \) to the centroid of \( \psi \) differs from the mean value for all forecasts within the p.d.f. Thus we may consider \( \langle (\delta c_D)^2 \rangle \) to represent the ‘predictable’ proportion of case-by-case variation in skill, whereas \( \langle (\delta d_j)^2 \rangle \) represents a random, ‘unpredictable’ proportion.

It turns out that \( \rho_A \) may be expressed in terms of the ratio \( \alpha \) of these quantities, where \( \alpha = \langle (\delta c_D)^2 \rangle / \langle (\delta d_j)^2 \rangle \). The result is

\[
\rho_A = \frac{M^{-2} + \langle (M + 1)/M \rangle \alpha}{\left[ (M^{-1} + \alpha)(1 + M^{-3} + \langle (M + 1)/M \rangle^2 \alpha) \right]^{1/2}}.
\]

In deriving Eq. (11) (see appendix), we assume for simplicity that the random covariance terms in Eqs. (1) and (3) are negligible in size. (Clearly their average values are zero.) If these terms are significant, the value of \( \rho_A \) would be smaller than indicated by the calculation, which should be taken as representing an upper limit. Using values of \( \alpha \) determined from experiment, theoretical values of \( \rho_A \) may be calculated from Eq. (11)
for comparison with experimental correlations. If the two agree well, verifying the assumptions implicit in Eq. (11), the latter may be used to deduce how \( \rho_A \) should vary with \( \alpha \). This is of interest for several reasons. The extent of case-by-case variations in predictability, measured by \( \alpha \), might vary seasonally, and also according to the amount of spatial or temporal filtering applied to the forecast fields. Furthermore experimental estimates of \( \alpha \) are subject to sampling error due to the finite number of experiments (see section 6).

Another consideration is the effect of varying the ensemble size. Clearly, as \( M \) increases our estimate \( (M/(M-1))s_M \) of the true spread \( D \) of the forecast p.d.f., becomes more precise through the decrease in sampling error. Consequently our prediction of forecast skill also becomes more precise. The correlation \( \rho_A \) is unfortunately not suitable to quantify this increased precision, since the proportion of the case-by-case variation in ensemble-mean forecast skill predictable from corresponding variations in spread is itself a function of ensemble size (see appendix). However, if we consider the mean error variance \( E_1 = \bar{e}_1 \) of individual forecasts within a forecast p.d.f., the 'predictable' proportion of the case-by-case variation in this measure of skill is independent of \( M \) (see appendix). Thus from the correlation \( \rho_A' \) between \( E_1 \) and \( s_M \) we can obtain an unbiased indication of how our ability to predict forecast skill varies with ensemble size, which is not the case with \( \rho_A \). In the appendix it is shown that

\[
\rho_A' = \frac{2\alpha}{[(\alpha + 1/M)(4\alpha + 1)]^{1/2}}. \tag{12}
\]

The implications of Eqs. (11) and (12) are discussed in section 6.

(d) Phase correlation between spread and skill

We may define a correlation coefficient \( \rho_p \), analogous to \( \rho_A \), to measure the relationship between phase spread and ensemble-mean skill. For this purpose we measure phase skill by \( c_M' = \pm |c_M| \), choosing the negative sign if \( c_M \) is negative, to distinguish cases of positive and negative anomaly correlation.

\[
\rho_p = \frac{\langle (c_M' - \langle c_M' \rangle)(a_M - \langle a_M \rangle) \rangle}{\langle (c_M' - \langle c_M' \rangle)^2 \rangle^{1/2} \langle (a_M - \langle a_M \rangle)^2 \rangle^{1/2}}. \tag{13}
\]

An expression analogous to Eq. (11) for \( \rho_p \) in terms of sampling and climatic variations in correlation could be derived. However, in order that such a theoretical estimate remains accurate in the extended-range as \( \bar{u} \) tends to zero, and correlations consequently become small, random covariance terms would have to be retained in the calculation (in contrast to the amplitude case), which would then become complicated and difficult to verify experimentally. Such a calculation is therefore not included.

6. Perfect model results

Turning now to experimental results we first consider the theoretical or 'perfect model' predictability of the RP\(_A\) ensemble forecasts. All the results are grid-point averages from the model 500 mb height field in the region 30–85\(^\circ\)N. We assume that forecast statistics calculated using this subset of the model phase space are representative of those we would obtain if all model variables were taken into account.
(a) Evolution of ensemble distribution

Initially we examine the time variation of the ensemble spread to establish the period for which an ensemble forecast ‘remembers’ its initial conditions. By calculating $f$ for each RP$_A$ ensemble forecast and averaging over the eight experiments we obtain an estimate of the climatic average ratio $\langle F \rangle$ between the ensemble variance and the model climate variance. (In this section we shall use $\langle \rangle$ to denote an average over the RP$_A$ experiments, although values obtained by such an operation are only estimates of true climatic average values, for which we would require an infinite number of experiments.) Figure 3(a) shows the time variation of $\langle F \rangle$ for daily forecast fields. The perturbation growth rate of $d\langle F \rangle/dt$ shows a maximum near $\langle F \rangle = 0.5$, in agreement with the results of Lorenz (1982). The predictability limit may be estimated by using a $t$-test to find the point on the curve where the difference between $\langle F \rangle$ and unity is no longer statistically significant. In doing so we assume our estimate $N\bar{v}/(N-1)$ of the true model climate variance to be perfectly accurate, even though it is actually subject to sampling error. The $t$ value is

$$t_A = (N-1)^{1/2}(1 - \langle F \rangle)/s_f,$$

where

$$s_f = \left\{ \sum_{k=1}^{N} (f_k - \langle F \rangle)^2/N \right\}^{1/2}$$

is the standard deviation of the values $f_k$ obtained from the $N(=8)$ RP$_A$ experiments, at the appropriate point on the curve. Setting the significance level at 5% we find a predictability limit of 26 days.

We may obtain a second predictability estimate using the phase ensemble spread. In each RP$_A$ experiment $r = (Ma_M - 1)/(M - 1)$ is an unbiased estimate of $R$ (see Eq. (7)). Thus the mean value of $r$ over the RP$_A$ experiments forms an estimate of $\langle R \rangle$, the climatic average phase ensemble spread. Since the climatic average phase and amplitude spread may be deduced from each other (Eq. (8)), the phase and amplitude predictability limits should be the same. When calculating $a_M$ in each experiment, the required anomaly fields are formed relative to the available estimate of the model climate mean (see section 4). The sampling error associated with this estimate biases the measured values $(a_M)_\text{meas}$ of $a_M$. In fact it may be easily shown that an unbiased estimate $(a_M)_\text{true}$ of the correct value of $a_M$ is given by

$$(a_M)_\text{true} - (a_M)_\text{meas} = [(1 - 2\hat{e}_1)\{(a_M)_\text{meas} - 1\}]/N,$$

where $\hat{e}_1$ is the mean skill of individual forecasts in the ensemble. ($\hat{e}_1$ appears because the model climatology comprises the set of nature runs, rather than a completely independent set of integrations.) This statistical correction is applied to the measured values of $a_M$ prior to the calculation of values of $r$.

From the experimental curve $\langle R \rangle(t)$ (see Fig. 3(a)), we may determine the predictability limit by testing the significance of the difference between $\langle R \rangle$ and 0. The quantity $r$, obtained from the squared correlation between the ensemble-mean and the individual forecasts within the ensemble, is clearly not normally distributed in general. However, provided the standard deviation $s = \{(\sum_{k=1}^{N} (r_k - \langle R \rangle)^2/N)^{1/2}$ plotted in Fig. 3(a) is small, deviations from normal should not be significant. Thus we may use a $t$-test based on $r_p$, defined by $t_p = (N - 1)^{1/2}(R)/s_r$, to determine the predictability limit. Using this test we find a 5% level predictability limit of 27 days, agreeing closely with the amplitude estimate.

These figures relate to the predictability of daily forecast fields, but from a long-range forecasting viewpoint we are more interested in the predictability of time-means. Accordingly, Fig. 3(b) shows values of $\langle R \rangle$ and $\langle F \rangle$, where all relevant forecast and climate
Figure 3 (a). Time variation for daily fields of: (a) ratio \( F \) of variance of forecast p.d.f. to model climate variance; (b) squared anomaly correlation \( R \) between centroid of forecast p.d.f. and all the individual forecasts within the p.d.f. Values plotted are estimates obtained from corresponding ensemble forecast statistics, averaged over the RF_{A} experiments, with associated standard deviations (dashed lines).

Figure 3 (b). As Fig. 3(a) but for 15-day-mean fields centred about points marked \( \times \).

variances have been calculated from 15-day-mean fields. Values of \( R \) are higher and values of \( F \) lower than their corresponding values for daily fields at the same time levels, suggesting that the 15-day means are potentially more predictable, and a t-test on the \( \langle R \rangle(t) \) curve suggests that the predictability limit lies between the points centred on days 38 and 43, although a t-test on the \( \langle F \rangle(t) \) curve does not reveal a corresponding increase relative to the limit for daily fields. It should be emphasized that since all \( F \) values remain below unity, and all \( R \) values above zero, some potential predictability may exist for 15-day-mean fields throughout the forecast period, even though it cannot be demonstrated unequivocally by a statistical test.
The other point of interest from Fig. 3(b) concerns the variation of spread from case to case. This variation, which is larger for 15-day-mean fields than for daily fields, is in general due partly to sampling effects and partly to genuine variations in predictability between the different experiments. This point is developed further in subsequent discussion of the correlation between spread and skill.

(b) Forecast skill

Forecast skill is determined by verification against the corresponding nature runs. In each experiment the ensemble-mean forecast may be compared with the individual forecast made from the unperturbed initial state. Figure 4(a) shows the daily normalized error variance and anomaly correlation scores averaged over all the RP_A experiments. Initially, while the ensemble spread remains small, the difference in skill is also small. Beyond the short-range period the superiority of the ensemble-mean forecast becomes more marked as ensemble spread increases.

![Graph showing forecast skill](image-url)

Figure 4 (a). Daily perfect model forecast anomaly correlation, (a), and normalized error variance, (b), averaged over RP_A experiments. ———: Individual forecast from unperturbed initial state. -----: Ensemble-mean forecast.

The upper limit of significant non-zero forecast skill may be deduced by applying a t-test to the curves of Fig. 4(a). From the anomaly correlation curves$^1$ we find that the skill limit for individual forecasts is 18 days and for $M = 7$ ensemble-mean forecasts is 27 days, representing a 50% increase. Using this criterion may yield a limit which is somewhat optimistic compared with the limit of useful skill, however, it is not clear which level of correlation or error variance corresponds to such a limit. In terms of, say, the time taken to reach a correlation level of 0.5, the percentage increase is somewhat smaller.

---

$^1$ For t-test purposes, anomaly correlations $c$ are converted to the Fisher (1958) z-statistic, defined by $z = \frac{1}{2} \ln((1 + c)/(1 - c))$, as $z$ is more nearly normally distributed than $c$. 
In Fig. 4(b) we show the theoretical time variation of $\langle e'_M \rangle$ and $\langle c_M \rangle$, deduced using Eqs. (6) and (9) from experimental curves showing the mean skill of all the individual forecasts within each ensemble, averaged over the RP$_A$ experiments. For both the normalized error variance and anomaly correlation skill scores the theoretical and experimental curves (copied from Fig. 4(a)) match up quite well. In fact the correspondence is probably even better than shown in Fig. 4(b), as the tendency for theoretical $\langle e'_M \rangle$ values to be slightly lower, and theoretical $\langle c_M \rangle$ values slightly higher, than their experimental counterparts, can be explained by small biases in the measured values caused, as for the $\langle a_M \rangle$ values discussed earlier, by sampling error in the model climatology. Since the required corrections tend to zero at large forecast times, thus having no effect on the asymptotic level of skill (cf. $a_M(i)$), they are omitted for simplicity.

We may therefore use Eqs. (6) and (9) to show how the skill of ensemble-mean forecasts varies with $M$. Figure 4(b) shows the skill we would obtain with $M = \infty$. For amplitude skill almost all the improvement we would obtain with $M = \infty$ is realized with $M = 7$. This agrees with Seidman's (1981) experiment in which it was demonstrated using an r.m.s. measure of skill that between four and eight individual forecasts gave a sufficiently accurate estimate of the centroid of one particular forecast p.d.f. For phase skill the situation is slightly different. Figure 4(b) shows that while an individual forecast still has a large degree of skill, say $\langle c_M \rangle \geq 0.5$, $M = 7$ gives almost as large an improvement as $M = \infty$. However, at longer range when individual forecasts have lower but still non-zero skill the difference between $M = 7$ and $M = \infty$ becomes more marked. Thus the phase skill is more sensitive to ensemble size than the amplitude skill. Our assessment of a reasonable value of $M$ to use therefore depends in general on our view of the relative importance of phase and amplitude skills.

![Figure 4 (b). Daily perfect model forecast anomaly correlation, (a), and normalized error variance, (b), averaged over RP$_A$ experiments. Solid curves 1 show average skill of all individual forecasts within ensemble. Solid curves 2 and 3 show theoretical skill of $M = 7$ ensemble-mean forecast, and $M = \infty$ ensemble-mean forecast, calculated from curves 1. --- shows experimental skill of $M = 7$ ensemble-mean forecast.](image-url)
(c) Effect of spatial and temporal averaging

We now consider the possibility of increasing forecast skill by suppressing the shorter, less predictable scales of motion by space or time smoothing. Shukla (1981) presented evidence based on control and perturbation integrations from three different initial conditions to suggest that the planetary waves may be more predictable than smaller scale components of the flow, and Seidman (1981) found that space–time averaging increased the predictability time in a single 34-case ensemble.

In Fig. 5 we compare the mean anomaly correlation of the planetary-scale component (zonal waves 0–3) of the individual forecasts to that of the corresponding unfiltered fields. An improvement is apparent throughout the period for which the unfiltered forecasts show skill, although no significant increase in the predictability limit results. Up to about day 16 the beneficial effect of this spatial smoothing is approximately equal to that of ensemble-averaging the unfiltered fields. To some extent these two smoothing operations are similar in that they both suppress unpredictable noise in the forecast patterns. However, the ensemble average contains more information as it also represents an average over a spread of possible long-wave configurations. Thus we obtain a further increase in skill if we consider only waves 0–3 of the ensemble-mean forecast.

![Figure 5](image-url)

Figure 5. Daily perfect model forecast anomaly correlation, averaged over RP4 experiments. ---: Unfiltered individual forecast from unperturbed initial state. ....: Individual forecast, zonal waves 0–3 only. ---: Unfiltered ensemble-mean forecast. ---: Ensemble-mean forecast, zonal waves 0–3 only.

From the ensemble spread statistics (Figs. 3(a) and (b)) we have seen that 15-day-mean forecast fields are potentially more predictable than daily fields. Figure 6 shows that this potential is fulfilled in terms of extra forecast skill. Time-averaging an individual forecast yields some improvement, but the level of skill of time-averaged ensemble-mean forecasts is substantially higher. The combined effect of time and ensemble averaging can be gauged by considering the mean skill scores of an unfiltered daily forecast and a time-averaged ensemble-mean forecast at day 18. This marks the mean predictability limit for the former, which shows an average anomaly correlation of 0.26 at this point, whereas the average score for the latter is 0.58, clearly indicative of a useful degree of skill.
Figure 6. Perfect model forecast anomaly correlation, averaged over RP4 experiments. ---: Daily individual forecast from unperturbed initial state. —×—×—: 15-day-mean individual forecast, centred about points marked ×. ----: Daily ensemble-mean forecast. —×—×—: 15-day-mean ensemble-mean forecast.

<table>
<thead>
<tr>
<th>Central point of averaging period</th>
<th>Length of averaging period (days)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Day 16</td>
<td>0.51</td>
</tr>
<tr>
<td>Day 21</td>
<td>0.30</td>
</tr>
<tr>
<td>Day 26</td>
<td>0.19</td>
</tr>
<tr>
<td>Day 31</td>
<td>0.10</td>
</tr>
</tbody>
</table>

In Table 2 we show how the anomaly correlation score varies with length of averaging period, for time-means centred about various points in the extended-range period of the forecast. The skill score invariably increases monotonically with length of period, but in principle this might be trivially explained by the inclusion of earlier, more skilful parts of the forecast in the time averages. However, the mean score for the daily fields comprising the time averages in Table 2 is always close to the score in the first column, confirming that almost all the improvement in skill is genuinely achieved through the filtering effect of the time average. Nevertheless, we must not infer from Table 2 that, for example, 31 days is a better time-averaging period than five days for use in the presentation of extended-range forecasts, even though 31-day-mean forecasts have more skill, since excessive smoothing may suppress useful predictability contained in higher frequencies and wavenumbers. As pointed out by Shukla (1981), determination of an optimum time-averaging period would require detailed knowledge of the amplitude and predictability statistics of all the different components of the space–time spectra of forecast flow patterns.

In the following the theoretical prediction of skill from ensemble spread is considered.
in the context of these perfect model results. Readers wishing to study the practical skill of the model integrations can skip to section 7 without loss of continuity.

(d) Prediction of skill

We have defined a correlation coefficient $\rho_A$ to measure the relationship between the amplitude ‘skill anomaly’ $(e_M - \langle e_M \rangle)$ of an ensemble-mean forecast and the ensemble ‘spread anomaly’ $(s_M - \langle s_M \rangle)$. Figure 7 shows the time variation of $\rho_A$ for daily fields, with averages $\langle \rangle$ over the model climate distribution (see Eq. (10)) again approximated by averages over the RP$_A$ experiments. The correlation $\rho_A$ remains positive for the first 25 days suggesting that while the forecast retains a significant degree of skill, the ensemble spread gives some indication of the likely deviation of the skill score from the climate average value. However, the level of correlation is not particularly high, the mean value over days 1–25 being 0.34.

To compare experimental and theoretical $\rho_A$ values, estimates of $\alpha$, representing the ratio between predictable and unpredictable variations in skill (see Eq. (11) and accompanying discussion) must be obtained from the experimental data. The mean sampling variation $\langle \langle (d_j)^2 \rangle \rangle$ of the spread of a forecast p.d.f. may be obtained from the quantity $\langle v_s \rangle$, given by

$$\langle v_s \rangle = \left\langle \sum_{j=1}^{M} ((u_j - \bar{u})^2 - s_M)^2 / M \right\rangle,$$

where $v_s$ represents the ensemble variance of $(u_j - \bar{u})^2$, the square of the distance in phase space from an ensemble forecast member to the ensemble-mean. $\langle v_s \rangle$ and $\langle \langle (d_j)^2 \rangle \rangle$ are related (Murphy 1986b) by

$$\langle v_s \rangle = \frac{(M - 1)(M - 2)^2}{M^3} \langle \langle (d_j)^2 \rangle \rangle.$$

In deriving this equation it is again necessary to assume that random covariances, of the type shown in Eq. (1), may be safely neglected. An estimate of $\langle v_s \rangle$, and hence

![Figure 7. Correlation $\rho_A$, calculated from RP$_A$ experiments, between perfect model error variance of ensemble-mean forecast, and ensemble variance. ——: Daily fields. ---•---: 15-day-mean fields centred about points marked •.](image-url)
\((\langle \delta d_i \rangle)^2\), for each time level was found by calculating \(v_i\) for each RP\(_A\) ensemble and then averaging over the eight experiments. By estimating the mean-square variation in forecast spread \(\langle \langle s_M - \langle s_M \rangle \rangle^2 \rangle\) from the experiments in similar fashion, an estimate of \(\langle \langle \delta c \rangle^2 \rangle\) and hence \(\alpha\) could be obtained using Eqs. (A7) (see appendix). The daily values of \(\alpha\) thus deduced show large fluctuations from day to day, reflecting the considerable uncertainty involved in calculating the necessary fourth-order statistics. A larger number of experiments would be required to obtain an accurate representation of the time variation of \(\alpha\). Nevertheless the mean value over days 1–25 of 0·16, corresponds to a theoretical \(\rho_A\) value of 0·33, which agrees well with the experimental value.

For long-range forecasting purposes the values of \(\rho_A\) for time-mean fields are of interest. Figure 7 also shows the experimental values for 15-day means, which turn out to be very small. The mean value of \(\alpha\) for the four points plotted is only 0·06, which would explain the low level of correlation. If this value is truly representative of the climatic distribution of predictability then there is very little variation in the amplitude predictability of time-averaged ensemble-mean forecasts made from different initial conditions. However, since we are approximating this distribution with only eight experiments, the possibility exists that the true value of \(\alpha\) may be larger. Furthermore, as discussed in section 5, different values of \(\alpha\) may apply for different seasons of the year, or for fields subject to different degrees of spatial or temporal filtering, and \(\alpha\) may also be model dependent. In Table 3 we therefore give the variation of \(\rho_A\) with \(\alpha\), based on Eq. (11). Not surprisingly, higher correlations are obtained with higher values of \(\alpha\). If, for example, the climatic variations are equal to the sampling variations (\(\alpha = 1\)), a correlation of greater than 0·7 applies.

### Table 3. Theoretical Variation of \(\rho_A\) with \(\alpha\) for an Ensemble Size of Seven

<table>
<thead>
<tr>
<th>(\alpha)</th>
<th>(\rho_A (M = 7))</th>
<th>(\alpha)</th>
<th>(\rho_A (M = 7))</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0·14</td>
<td>0·5</td>
<td>0·57</td>
</tr>
<tr>
<td>0·05</td>
<td>0·17</td>
<td>1·0</td>
<td>0·72</td>
</tr>
<tr>
<td>0·1</td>
<td>0·26</td>
<td>(\infty)</td>
<td>1·0</td>
</tr>
<tr>
<td>0·3</td>
<td>0·46</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(\rho_A\) is the correlation between perfect model ensemble-mean forecast error variance, \(e_M\), and ensemble variance, \(s_M\). \(\alpha\) is the ratio of climatic to sampling variations in the spread of a forecast p.d.f.

Another question of interest is that of how large an ensemble size we need to obtain an accurate measure of the spread of a forecast p.d.f. To answer this we use the correlation \(\rho'_A\) between ensemble spread and the mean skill \(E_1\) of individual forecasts within the forecast p.d.f. (see section 5).

Table 4 shows the variation of \(\rho'_A\) with \(M\) for different values of \(\alpha\) as deduced from Eq. (12). The extent to which an ensemble size of seven is sufficiently large to give an accurate indication of case-by-case variations in predictability depends on the value of \(\alpha\). For small values \(M = 7\) is clearly insufficient, whereas for \(\alpha = 1\) there is only a 5\% difference between the levels of correlation appropriate to \(M = 7\) and \(M = \infty\). The correlation \(\rho'_A\) cannot be measured experimentally, as we would require ensembles of a very large size to obtain sufficiently accurate estimates of \(E_1\). Its value is didactic, illustrating how our knowledge of the predictability of a forecast p.d.f. increases with ensemble size. The correlation coefficient \(\rho_A\), while unsuitable for this purpose, has a more obvious practical value however, as it shows the (maximum) extent to which we
THE IMPACT OF ENSEMBLE FORECASTS

TABLE 4. THEORETICAL VARIATION OF $\rho_\lambda$ WITH $\alpha$ AND $M$

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>$\rho_\lambda(M = 3)$</th>
<th>$\rho_\lambda(M = 7)$</th>
<th>$\rho_\lambda(M = 15)$</th>
<th>$\rho_\lambda(M = \infty)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>0.05</td>
<td>0.15</td>
<td>0.21</td>
<td>0.27</td>
<td>0.41</td>
</tr>
<tr>
<td>0.1</td>
<td>0.26</td>
<td>0.34</td>
<td>0.41</td>
<td>0.53</td>
</tr>
<tr>
<td>0.3</td>
<td>0.51</td>
<td>0.61</td>
<td>0.67</td>
<td>0.74</td>
</tr>
<tr>
<td>0.5</td>
<td>0.63</td>
<td>0.72</td>
<td>0.77</td>
<td>0.82</td>
</tr>
<tr>
<td>1.0</td>
<td>0.77</td>
<td>0.84</td>
<td>0.87</td>
<td>0.89</td>
</tr>
<tr>
<td>2.0</td>
<td>0.87</td>
<td>0.91</td>
<td>0.93</td>
<td>0.94</td>
</tr>
<tr>
<td>$\infty$</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

$\rho_\lambda$ is the correlation between the mean perfect model error variance $E_i$ of individual forecasts within a forecast p.d.f., and the ensemble variance $s_M$. $\alpha$ is as indicated in Table 3, and $M$ is ensemble size.

Figure 8. Correlation $\rho_p$, calculated from RP$_\lambda$ experiments, between the square of the perfect model anomaly correlation score of the ensemble-mean forecast (negative values assigned where the linear correlation is negative), and the square of the anomaly correlation between the ensemble-mean and all individual forecasts within the ensemble. ---: Daily fields. -x-x-x: 15-day-mean fields centred about points marked x.

could diagnose in advance the skill of our best-estimate prediction (the ensemble-mean) if we were to produce regular forecasts based on ensembles of $M$ integrations.

Figure 8 shows the time variation for daily fields of $\rho_p$ (see Eq. (13)), the correlation between the phase spread and skill anomalies observed from the RP$_\lambda$ experiments. The level of correlation is a little higher than for amplitude skill, being 0.40 for days 1–25 (cf. 0.34 for $\rho_\lambda$). In particular $\rho_p$ values for days 1–10 are considerably higher than corresponding $\rho_\lambda$ values, suggesting that phase skill may be a more sensitive indicator of case-by-case variations in predictability than amplitude skill. The $\rho_p$ values for 15-day-mean fields are no higher than contemporaneous daily values however, and the correlation drops to zero beyond day 18 (see Fig. 8), showing that we may only expect to obtain a non-zero correlation between phase spread and skill for about one half of the period over which the ensemble-mean forecast shows non-zero skill.
(e) Prediction of spatial variation of skill

In general forecast error growth is not uniformly distributed in space (Miyakoda et al. 1972). In any particular case errors may originate in certain localized regions and only later spread to affect the remaining areas. Spatial variations in an ensemble forecast distribution can in principle be used to predict corresponding spatial variations in ensemble-mean forecast skill, just as the space-integrated spread provides an indication of the space-integrated skill.

To investigate this we require a method of determining point-by-point the statistical significance of relevant differences between an ensemble distribution and the model climate distribution. If we measure skill by anomaly correlation, we must establish the significance of the ensemble-mean forecast anomalies. This is done using a $t$-test on the anomaly at each grid point. The $t$ statistic is defined as

$$(t_{\text{anom}})_i = \frac{(\bar{u})_i/[(M^{-1} + N^{-1})\{M(s_M)_i + N(v)_i\}]}{(M + N - 2)}^{1/2}$$

with $(M + N - 2)$ degrees of freedom, where $(\bar{u})_i$, $(s_M)_i$, and $(v)_i$ denote the values of the ensemble-mean anomaly, the ensemble variance and the climate variance of some model variable at grid point $i$.

Setting a null hypothesis of no difference between the ensemble-mean and the model climate mean, a field of $(t_{\text{anom}})_i$ values may be converted into a map showing the spatial variation of the significance levels associated with the relevant ensemble-mean anomaly field (see Fig. 11). Such maps were plotted for the 15-day-average ensemble-mean 500 mb height anomaly fields for forecast days 1–15, 16–30 and 31–45 for each of the eleven RP ensemble forecasts (i.e. both RP$_A$ and RP$_B$ cases), excluding forecasts where the full field skill score had dropped below zero. From each map, regions within which all grid points were significant at the 5% level or greater were picked out, and anomaly correlation scores found for each region. The area-weighted average of these limited-region scores was then compared with the full field score for the map.

On average the skill within the significant regions exceeds the full field skill (Table 5). Of the eleven ensemble-mean forecasts, ten show an improvement for days 1–15, nine for days 16–30, and five out of the eight cases still showing positive full field skill at days 31–45 also give an improvement. The total area covered by significant anomalies decreases with time, as the ensemble distributions approach climatology.

<table>
<thead>
<tr>
<th>Period</th>
<th>No. of forecasts with positive full field skill (max. 11)</th>
<th>Full field anomaly correlation</th>
<th>Significant region anomaly correlation</th>
<th>Fractional area covered by significant anomalies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Days</td>
<td>11</td>
<td>0.87</td>
<td>0.93</td>
<td>0.39</td>
</tr>
<tr>
<td>Days</td>
<td>11</td>
<td>0.57</td>
<td>0.64</td>
<td>0.18</td>
</tr>
<tr>
<td>Days</td>
<td>8</td>
<td>0.40</td>
<td>0.50</td>
<td>0.08</td>
</tr>
</tbody>
</table>

Columns 2–4 show figures averaged over those cases where the ensemble-mean forecast still shows a positive full field anomaly correlation.
A complementary approach, not attempted in the present paper, would be to repeat the analysis for amplitude skill by using a variance ratio test to determine significant regions and then looking for a reduction in normalized error variance compared to the full field value. In general the significant areas would differ from those found using the $t$-test. In particular, areas of small anomaly not significant in the phase sense might be found to be significant in the amplitude sense.

7. **Practical forecast skill**

In Fig. 9 we compare the practical forecast skill of the unperturbed individual forecast with that of the ensemble-mean forecast, averaged over the RP$_A$ cases. The latter affords no improvement in either amplitude or phase skill, for either daily or time-mean fields (not shown). This ties in with the results from four practical LAF ensemble experiments carried out with the same model, described elsewhere (Murphy 1986a). It is clear from Fig. 9 that the model does not possess good short- to medium-range skill. The anomaly correlation score drops to 0.2 by day 7. As explained in section 3, we would not normally expect an ensemble-mean forecast to offer any significant benefit under such circumstances. Neither would we expect to observe a significant correlation between ensemble spread and forecast skill on a hemispheric scale, this being corroborated by our experimental results. Dalcher et al. (1985) have recently reported a low level of spread/skill correlation using global verification in medium-range LAF experiments, and Kalnay and Dalcher (1987) have obtained similar results for hemispheric verification, with ensembles created using a perturbation approach.

In particular cases where individual forecasts show skill beyond the normal deterministic limit, it is possible that an ensemble-mean forecast may give an improvement in skill (see section 3). In two of the eight RP$_A$ cases the average 15-day-mean anomaly correlation score for individual forecasts within the ensemble remains positive throughout.

![Figure 9](image-url)

*Figure 9. Daily practical forecast anomaly correlation, (a), and normalized error variance, (b), averaged over RP$_A$ experiments. ---: Individual forecast from unperturbed initial state. ---: Ensemble-mean forecast.*
the 50-day forecast period. To these we might add the RP₈ cases, specially chosen as occasions where an individual forecast run previously is known to have shown an unusually high degree of skill.

In one of the three RP₈ cases (initialization date 23 December 1975), although the original unperturbed forecast shows positive skill out to day 50, the forecasts from the perturbed initial states generally lose skill well before this stage, and the average individual forecast score drops to zero at day 18. The skill of the unperturbed forecast must therefore be regarded as fortuitous in this particular case, since other forecasts made from initial states nominally compatible with observation errors do not show the same skill. Excluding this experiment from our set of 'predictable' cases leaves us with four experiments in which the average individual forecast anomaly correlation remains positive through to day 50. Figure 10 shows the anomaly correlation scores for 15-day-mean fields, averaged over these four cases, of the ensemble-mean forecast compared with the mean score of individual forecasts within the ensemble. There is a modest but nonetheless consistent improvement in skill at all time levels. For comparison corresponding scores averaged over the remaining seven experiments (six RP₈ cases plus 23 December 1975) are also given. (Note that Fig. 10 shows that in the medium-range forecast period, even in the 'unpredictable' cases, the ensemble-mean gives a slight improvement in skill compared with the average score of individual forecasts within the ensemble. However, it does not improve on the skill of the unperturbed individual forecast, which is slightly superior to the perturbed individual forecasts at this range.)

![Diagram](image)

**Figure 10.** Practical forecast anomaly correlation, for 15-day-mean fields centred about points marked ×. (a) Average over four 'predictable' cases. (b) Average over seven 'unpredictable' cases. ×−×−×−×−: Average skill of individual forecasts within ensemble. ---×---×---: Ensemble-mean forecast skill.

Having established the lack, with this particular model, of a significant correlation between ensemble spread and practical skill on a hemispheric scale, other a priori methods of skill prediction must be considered. One possibility is that the predictable occasions correspond to episodes of unusually high persistence in the atmosphere. If this is so then presumably the model forecast should also show above average persistence on these occasions. (This is not of course to hypothesize that every time the model shows
above average persistence, the atmosphere does also.) Table 6 shows the extent to which the anomalies developed in the first 15 days of the ensemble-mean forecast persist throughout the remainder of the forecast period, confirming that on average the degree of persistence is somewhat greater in the predictable cases, although the difference is not large enough to establish such a measure as a reliable predictor of skill.

<table>
<thead>
<tr>
<th>Period (days)</th>
<th>6–20</th>
<th>11–25</th>
<th>16–30</th>
<th>21–35</th>
<th>26–40</th>
<th>31–45</th>
<th>36–50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predictable cases</td>
<td>0.83</td>
<td>0.62</td>
<td>0.45</td>
<td>0.36</td>
<td>0.30</td>
<td>0.26</td>
<td>0.29</td>
</tr>
<tr>
<td>Unpredictable cases</td>
<td>0.83</td>
<td>0.58</td>
<td>0.40</td>
<td>0.24</td>
<td>0.06</td>
<td>-0.05</td>
<td>-0.05</td>
</tr>
</tbody>
</table>

The second row shows scores averaged over the four ‘predictable’ cases (see text), and the third row shows averages over the remaining seven RP experiments.

(a) Spatial variation of skill

In section 6 we used a t statistic (see Eq. (14)), to determine regions within which point-by-point 15-day-average ensemble-mean forecast anomalies were each significantly different from zero at the 5% level, comparing the average perfect model forecast anomaly correlation score in such regions with the full field score. Table 7 shows a repeat of this analysis for practical forecast skill.

For days 1–15 nine of the eleven ensemble-mean forecasts show an improved score in their significant regions, and those forecasts which retain positive skill at longer range also, on average, show more skill in these areas. The improved skill at days 1–15 is probably largely due to persistence, since regions of large anomaly in the initial conditions will show a degree of persistence both in the real atmosphere and in the model, guaranteeing some positive level of phase skill in such areas, even if the rate of divergence between forecast and reality is relatively high, whereas in regions of initially small anomalies no such guarantee applies. Kalnay and Dalcher (1987) have recently found that variations in the forecast anomaly correlation for limited areas can be well predicted in 5-day forecasts using the mean correlation between pairs of integrations in a five-member ensemble as a measure of spread. Those forecasts which show skill at days 16–30 and beyond possess above average predictability, so that the distribution of model states is likely to bear a greater similarity to the distribution of possible atmospheric states than normal, so it is not unreasonable that spatial variations in the ensemble distribution should be related to spatial variations in forecast skill in such cases, as shown in Table 7.

<table>
<thead>
<tr>
<th>Period</th>
<th>No. of forecasts with positive full field skill</th>
<th>Full field anomaly correlation</th>
<th>Significant region anomaly correlation</th>
<th>Fractional area covered by significant anomalies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Days</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1–15</td>
<td>11</td>
<td>0.48</td>
<td>0.56</td>
<td>0.39</td>
</tr>
<tr>
<td>Days</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16–30</td>
<td>8</td>
<td>0.29</td>
<td>0.37</td>
<td>0.17</td>
</tr>
<tr>
<td>Days</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31–45</td>
<td>4</td>
<td>0.34</td>
<td>0.52</td>
<td>0.07</td>
</tr>
</tbody>
</table>
An example of an unusually predictable forecast, the 14 December 1976 case; is given in Figs. 11(a)–(c). (See also Mansfield (1986), who describes the effect of including observed sea surface temperatures on forecast skill for this case.) During the forecast period a strong pattern of the PNA type (Wallace and Gutzler 1981) developed, showing a low anomaly in the Pacific with a downstream high over the west coast of North America and a further downstream low over the Great Lakes. Also present throughout the period was an extensive area of high anomaly at high latitudes. Although showing a fair degree of skill during the first 15 days (Table 8), the ensemble-mean forecast misses the initial development of the area of high anomaly centred over the west coast of North America. At days 16–30 the model has started to develop a high anomaly over Alaska but its downstream low is centred much too far to the east. The main areas of skill at this stage are over the Pacific and at high latitudes, where areas of low and high anomaly respectively are correctly forecast. The area covered by significant anomalies has decreased from days 1 to 15, as the individual forecasts within the ensemble diverge, but the skill in the remaining significant areas is, on average, higher than the full field skill (Table 8). In the next 15-day period the forecast flow finally develops a structure resembling the full PNA pattern, although the high and downstream low are centred somewhat too far east. The model also correctly maintains high anomalies in polar
Figure 11 (b). As Fig. 11(a) but for days 16–30.

Figure 11 (c). As Fig. 11(a) but for days 31–45.
regions. Although the total area of significant anomalies is now small, being only slightly above the level of chance, it is encouraging that the centres in the forecast PNA pattern, an area where the forecast shows skill, are all significant, whereas the centres of two downstream areas of high anomaly, which have no counterparts in the actual flow, are not significant.

8. Conclusions

A set of eight ensemble forecasts, each containing seven 50-day integrations of a hemispheric version of the Meteorological Office 5-level GCM, and employing climatological boundary conditions, has been used to study the effect of ensemble-averaging on theoretical and practical forecast skill. Initial conditions for the integrations in each ensemble were created by adding spatially-correlated random perturbations of a size consistent with typical observation and analysis errors to a given winter analysed state. Since our model possesses only moderate short- and medium-range practical forecast skill, an ensemble-mean forecast does not, on average, offer any improvement in skill compared with an individual forecast. Therefore, to demonstrate the potential of the method and establish an upper bound to the improvement in skill attainable in future practical forecasts using more skillful models, the ensemble forecasts were also verified under ‘perfect model’ conditions by using a model integration, made from initial conditions created by perturbing the analysed state as described above, to represent the evolution of ‘nature’ in each experiment.

Under these conditions the ensemble-mean forecast is clearly more skillful than an individual forecast, yielding a 50% increase in the predictability limit. Spatial smoothing or time-averaging improves the skill of both an individual and an ensemble-mean forecast, but the ensemble-mean forecast remains superior. The improvement in skill is consistent with theoretical calculations, relevant to perfect model conditions, in which an ensemble forecast is viewed as a sampling approximation to the infinite set of possible forecasts consistent with the errors associated with the single available analysis of the initial atmospheric state. These calculations are used to assess how ensemble-mean forecast skill is expected to vary with ensemble size.

A positive correlation between ensemble spread and perfect model ensemble-mean forecast skill is found, although this correlation essentially disappears after about one half of the period over which the ensemble-mean retains significant skill. However, with only eight experiments available to estimate the model’s climatic distribution of intrinsic predictability, these results are subject to considerable uncertainty. Further theoretical calculations are used to estimate the levels of correlation that would be obtained with different climatic distributions of predictability.

Although the practical skill of an ensemble-mean forecast is normally no higher than that of an individual forecast with this model, it is found that in four experiments (consisting of two from the original eight plus two from a further three specially selected
for the purpose), in which the average skill of the individual forecasts in the ensemble remains positive beyond the normal limit of deterministic predictability, the ensemble-mean forecast does show improved skill. In each of these four cases the anomaly pattern developed in the ensemble-mean forecast during the first 15 days was found to be unusually persistent during the remainder of the forecast period.

It is shown that spatial variations in the statistical significance of an ensemble-mean anomaly field are related to corresponding spatial variations in ensemble-mean forecast anomaly correlation, for both perfect model and practical forecasts, as long as the anomaly correlation of the full field remains positive.

This study shows that, given a model with superior skill up to the normal limit of deterministic predictability, the ensemble forecast technique can improve significantly on the skill of individual forecasts and therefore provides some extension to this limit. Furthermore, in those situations where skill at longer range is possible, the impact of the ensemble forecast should be correspondingly larger. This factor may be particularly important if boundary forcing anomalies, not considered in this paper, are included in the extended-range model forecasts.

9. ACKNOWLEDGMENTS

I am grateful to Drs D. A. Mansfield and T. N. Palmer for their encouragement and many helpful suggestions.

APPENDIX

Correlation between spread and skill

We define the correlation coefficient $\rho_A$, given by

$$\rho_A = \frac{(\langle e_M - \langle e_M \rangle \rangle (s_M - \langle s_M \rangle))}{\langle (e_M - \langle e_M \rangle)^2 \rangle^{1/2} \langle (s_M - \langle s_M \rangle)^2 \rangle^{1/2}} \quad (A1)$$

to represent the relationship between the error variance $e_M$ of an ensemble-mean forecast and the ensemble variance $s_M$ (see Eqs. (1), (3) and (11), main text).

$s_M$ and $e_M$ are subject to sampling variation represented by $\delta s_M = s_M - \langle s_M \rangle$ and $\delta e_M = e_M - \langle e_M \rangle$. The climatic deviation of the mean spread $S_M$ and skill $E_M$, of size $M$ ensembles drawn randomly from a particular forecast p.d.f., is written as $\delta c S_M = S_M - \langle S_M \rangle$ and $\delta c E_M = E_M - \langle E_M \rangle$.

We may split the terms in Eq. (A1) into uncorrelated sampling and climatic contributions, giving

$$\rho_A = \frac{(\langle \delta e_M \delta s_M \rangle + \langle \delta c E_M \delta c S_M \rangle)}{\langle (\langle \delta e_M \rangle^2 + \langle \delta c E_M \rangle^2 \rangle (\langle \delta s_M \rangle^2 + \langle \delta c S_M \rangle^2 \rangle)^{1/2}} \quad (A2)$$

From Eqs. (1)–(4), we have

$$\delta s_M = \{(M - 1)/M\} (\hat{d}_j - D) = \{(M - 1)/M\} \delta \hat{d}_j \quad (A3)$$

$$\delta e_M = \delta \hat{d}_j / M \quad (A4)$$

where the random covariance terms in Eqs. (1) and (3) are assumed negligible for simplicity. The validity of this assumption would depend on the typical covariance
structures apparent within phase space vectors such as \((\mathbf{u}_i - \mathbf{u})\), representing the deviation of a forecast from the centroid of the corresponding forecast p.d.f. For a model with a large number of degrees of freedom, such an assumption should be reasonable. If such covariance terms cannot in reality be neglected, the terms \(\langle (\delta e_M)^2 \rangle\) and \(\langle (\delta s_M)^2 \rangle\) would be larger, and \(\rho_A\) therefore smaller, than indicated by our calculation. For climatic deviations,

\[
\begin{align*}
\delta_c S_M &= \frac{(M - 1)}{M} \delta_c D \\
\delta_c E_M &= \frac{(\delta D/M)}{M} + (d_o - \langle D \rangle).
\end{align*}
\tag{A5}
\tag{A6}
\]

From Eqs. (A3)–(A6) the terms in Eq. (A2) may be calculated (Murphy 1986b), giving

\[
\begin{align*}
\langle \delta e_M \delta s_M \rangle &= \frac{(M - 1)}{M^3} \langle (\delta d_j)^2 \rangle \\
\langle \delta_c E_M \delta_c S_M \rangle &= \frac{(M + 1) (M - 1)}{M^2} \langle (\delta_c D)^2 \rangle \\
\langle (\delta e_M)^2 \rangle &= \frac{\langle (\delta d_j)^2 \rangle}{M^3} \\
\langle (\delta_c E_M)^2 \rangle &= \frac{(M + 1)}{M} \langle (\delta_c D)^2 \rangle + \langle (\delta d_j)^2 \rangle \\
\langle (\delta s_M)^2 \rangle &= \frac{(M - 1)}{M^3} \langle (\delta d_j)^2 \rangle \\
\langle (\delta_c S_M)^2 \rangle &= \frac{(M - 1)}{M^2} \langle (\delta_c D)^2 \rangle.
\end{align*}
\tag{A7}
\]

Thus the correlation \(\rho_A\) is determined by the climatic, \(\langle (\delta_c D)^2 \rangle\), and sampling, \(\langle (\delta d_j)^2 \rangle\), variances of the variance \(D\) of the forecast p.d.f., for a given ensemble size \(M\). Defining \(\alpha = \langle (\delta_c D)^2 \rangle / \langle (\delta d_j)^2 \rangle\), we may substitute Eqs. (A7) into Eq. (A2) to give

\[
\rho_A = \frac{M^{-2} + \frac{(M + 1)}{M} \alpha}{\left[ (M^{-1} + \alpha) \{1 + M^{-3} + \langle (M + 1)/M^2 \alpha \} \right]^{1/2}}.
\]

Two terms contribute to the case-by-case variation in \(E_M\) (see the fourth of Eqs. (A7)). The first, \(\langle (M + 1)/M \rangle \langle (\delta_c D)^2 \rangle\), represents the ‘predictable’ element due to climatic variation in spread, whereas the second, \(\langle (\delta d_j)^2 \rangle\), is an ‘unpredictable’ element arising from the prior uncertainty of the identity of the true state \(u_o\). The ratio of the two terms varies with ensemble size \(M\), such that as \(M\) increases, the ‘predictable’ proportion of the case-by-case variation in ensemble-mean forecast skill decreases.

If we consider instead the measure of skill \(E_1\), the mean error variance of individual integrations in the forecast p.d.f., we may define a second correlation coefficient \(\rho_A\), by

\[
\rho_A' = \frac{\langle (E_1 - \langle E_1 \rangle) (s_M - \langle s_M \rangle) \rangle}{\langle (E_1 - \langle E_1 \rangle)^2 \rangle \langle (s_M - \langle s_M \rangle)^2 \rangle}^{1/2},
\]

whence

\[
\rho_A' = \frac{\langle \delta_c E_1 \delta_c S_M \rangle}{\langle (\delta_c E_1)^2 \rangle \langle (\delta s_M)^2 \rangle}^{1/2}.
\]

Since it may be shown (Murphy 1986b) that \(\langle \delta_c E_1 \delta_c S_M \rangle = 2 \langle (M - 1)/M \rangle \langle (\delta_c D)^2 \rangle\) and also \(\langle (\delta_c E_1)^2 \rangle = 4 \langle (\delta_c D)^2 \rangle + \langle (\delta d_j)^2 \rangle\), (i.e. independent of \(M\)), it follows that

\[
\rho_A' = \frac{2 \alpha}{\langle (\alpha + 1/M) (4 \alpha + 1) \rangle^{1/2}}.
\]
REFERENCES


Epstein, E. S. 1969 Stochastic dynamic prediction. Tellus, 21, 739–759


Mansfield, D. A. 1982 Atmospheric predictability experiments with a large numerical model. Tellus, 34, 505–513


Murphy, J. M. 1986a 'Lagged-average forecast experiments with a 5-level general circulation model'. Long-Range Forecasting and Climate Research Memorandum No. 8. Meteorological Office, Bracknell

Murphy, J. M. and Palmer, T. N. 1986b 'Statistical aspects of ensemble forecasts'. Long-Range Forecasting and Climate Research Memorandum No. 9. Meteorological Office, Bracknell


Wallace, J. M. and Gutzler, D. S. 1981 Teleconnections in the geopotential height field during the northern hemisphere winter. Ibid., 109, 784–812