Long-Range Forecasting and Climate Research

Statistical aspects of ensemble forecasts

by

J.M. Murphy

LRFC 9

July 1986

Meteorological Office (Met. O. 13)
London Road
Bracknell
Berkshire RG12 2SZ

LONG-RANGE FORECASTING AND CLIMATE RESEARCH MEMORANDUM NO. 9

LRFC9

STATISTICAL ASPECTS OF ENSEMBLE FORECASTS

By J M Murphy

July 1986

Met O 13 (Synoptic Climatology)
Meteorological Office
London Road
BRACKNELL
Berkshire RG12 2SZ

(This paper has not been published. Permission to quote from it must be obtained from the Assistant Director of the above Meteorological Office Branch).

Abstract

Methods of measuring the spread and skill of ensembles of numerical forecasts, which emphasise either amplitude or phase predictability, are described. An ensemble forecast is regarded as a sampling approximation to a continuous distribution representing the complete set of possible predictions consistent with the errors associated with some particular initial analysis. Assuming that the forecast model is perfect, theoretical estimates are obtained of the improvement in forecast skill obtainable through ensemble-averaging a number of individual integrations.

In theory ensemble spread and forecast skill should be related, and correlation coefficients are introduced to quantify the extent to which the skill may be predicted from the spread. In the case of amplitude predictability expressions for two such coefficients are derived in terms of the ratio of genuine variations in predictability between independent forecast distributions (so-called 'climatic' variations), to random sampling variations within a particular forecast distribution.

1. INTRODUCTION

The idea of the ensemble forecast arises from the uncertainty associated with an analysis of an atmospheric state used as the initial conditions for a numerical forecast. This uncertainty implies that we can conceive of an infinite number of states, each consistent with observation and analysis errors, which are equally likely to represent the true state. Running an integration from each of these states would yield an infinite set of alternative predictions, each equally likely to correspond to the true evolution of the atmosphere. Such a distribution of forecasts may be represented by a continuous probability density function (p.d.f.) in the model phase space (see next section).

Due to the inherent instability of the equations of motion, the spread of the forecasts would increase with time, until eventually the mean separation between pairs of forecasts became equal to that between randomly-chosen states from a suitable model climate distribution, and the mean of the forecast p.d.f. became coincident with the model climate mean. At an intermediate stage, with the spread significantly greater than its initial value, but still short of the aforementioned saturation level indicative of complete loss of predictability, the mean of the forecast p.d.f. would, on average, be closer to the true state than would its constituent individual forecasts (Leith, 1974).

We cannot, in practice, run an infinite number of forecasts. However, if we produce a finite ensemble of forecasts from a number of appropriate initial conditions, the ensemble-mean will give a more accurate estimate of the mean of the forecast p.d.f. than will an individual forecast. In addition, the predictability information carried by the ensemble spread may in principle furnish a useful a priori indication of forecast skill (Hoffman and Kalnay, 1983; Dalcher et al, 1985; Murphy, 1986).

The extent to which these benefits are realised in practice depends on the deficiencies of the forecast model (Murphy, 1986). The maximum benefit, both in terms of improving skill through ensemble-averaging, and also of the a priori prediction of skill, would be achieved with a perfect model, i.e. one which, given a perfect analysis, would produce a perfect forecast of the subsequent atmospheric evolution.

In this paper we shall describe simple measures of forecast skill and ensemble spread, and derive estimates of the improvement in skill through ensemble-averaging and also of the correlation between ensemble spread and forecast skill, under perfect model conditions.

2. PHASE SPACE FORMALISM, ASSUMPTIONS AND TERMINOLOGY

The statistical description of ensembles is facilitated by visualising the evolution of an individual forecast as the path of a point in a multi-dimensional phase space, defined so that each dimension corresponds to a prognostic variable of the model (see Epstein, 1969; Gleeson, 1970; Leith, 1974). At any instant such a point may be represented by a general state vector **u**, each element of which corresponds to a phase space dimension.

If we were to make a large number of analyses, based on independent observations of the atmospheric state at some instant t=0, we would obtain a cloud of such phase points centred, in the absence of bias in the observing system and analysis method, around the true state, with a dispersion reflecting measurement errors. In the limit, as the number of analyses tends to infinity, the cloud becomes a continuum which may be described by a p.d.f. $\psi(\mathbf{u},t=0)$.

In practice we have only a single analysis of the atmospheric state, and must therefore generate alternative initial states artificially if we wish to run an ensemble forecast. This may be achieved either by perturbing the single analysis in some way consistent with our knowledge of observation and analysis errors (Seidman, 1981; Shukla, 1981; Murphy, 1986), or by extrapolating previous analyses up to the start of the forecast period using the forecast model (Hoffman and Kalnay, 1983). With either method we obtain a finite ensemble of phase points, $\psi'(t=0)$ say (dropping the explicit dependence on $\mathbf{u}$), which constitutes the set of initial conditions for the ensemble forecast.

Ideally $\psi'(0)$ should be identical to $\psi(0)$ within the sampling limits imposed by the finite size of the ensemble. However, if we use the perturbation approach $\psi'(0)$ will not in general be centred about the true state (Leith, 1974), and if we use the lagged-average forecast approach the spread of $\psi'(0)$ will be too large due to the growth of prediction errors during the extrapolation of lagged analyses up to the start of the forecast period. Nevertheless we shall assume in subsequent discussion that $\psi'(0)$ is identical to $\psi(0)$ within sampling limits. Each individual integration comprising the ensemble forecast $\psi'(t)$ is assumed equally likely to represent the true atmospheric evolution. This is obviously not strictly true for lagged-average forecasts although, if the interval between successive analyses is sufficiently short and, say, only the long-wave component of the forecasts is considered, the differences in skill due to time-lag should be small. Even for ensembles created by perturbation methods the assumption is not perfectly valid, if the unperturbed initial state provides one member of the ensemble, although the difference in skill between this integration and those from the perturbed initial conditions should be unimportant beyond the initial stages of the forecast (Murphy, 1986). Following Leith (1974), we shall consider only the first two moments of $\psi(t)$, thus treating it as effectively a normal distribution, although in practice its structure may in some cases be more complex (Murphy and Palmer, 1986).

In what follows, we shall take the origin of phase space to be coincident with the model climate mean at all forecast times. (The climate distribution appropriate to a given forecast may be obtained by running a large number of independent integrations using initial data from the same point in the seasonal cycle in different years. In general the resulting model climatology will vary with forecast time if, for example, the model itself contains the seasonal cycle). The elements of the state vector $\mathbf{u}$ are then anomalies from the climate mean, and if <> denotes an average over an infinite number of states chosen at random from the model climate distribution, then $\langle \mathbf{u} \rangle=0$ for all t. An average over an ensemble of M individual forecasts shall be signified by ^, so that $\hat{\mathbf{u}}(t)$ is the ensemble-mean forecast, or centroid of $\psi'(t)$. Averages over the full forecast p.d.f. $\psi(t)$ shall be denoted by $\bar{}$, so that, for example, its centroid is represented by $\bar{\mathbf{u}}(t)$. Thus as $M\rightarrow\infty$, $\hat{\mathbf{u}}(t)\rightarrow\bar{\mathbf{u}}(t)$, under our assumptions.

In this respect our interpretation of ψ' is different from that of Leith (1974) who, in a similarly-motivated study of forecast error statistics, dealing specifically with perturbation-style (or Monte Carlo) ensembles, considered ψ' as a distribution of possible true states centred around the single analysed state $\mathbf{u}_{obs}$, rather than vice versa. Thus whereas under our assumptions, at $t=0$, $(\mathbf{u}_o-\mathbf{u})^2 = d_o = 0$ in the limit as $M\rightarrow\infty$, where $d_o = (\mathbf{u}_o-\bar{\mathbf{u}})^2$ is the squared distance in phase space from the true state $\mathbf{u}_o$ to the centroid of ψ, under Leith's interpretation $(\mathbf{u}_o-\hat{\mathbf{u}})^2\rightarrow(\bar{\mathbf{u}}-\mathbf{u}_{obs})^2$ as $M\rightarrow\infty$. If $\mathbf{u}_j$ is an arbitrary state within ψ, then $\langle d_j\rangle=\langle(\mathbf{u}_j-\bar{\mathbf{u}})^2\rangle=\langle(\mathbf{u}_{obs}-\bar{\mathbf{u}})^2\rangle$. Therefore, if we assume that for $t>0$, $\langle d_o(t)\rangle=\langle d_j(t)\rangle$, our framework becomes effectively identical to that of Leith. Since $\mathbf{u}_o$ evolves away from $\bar{\mathbf{u}}$ during the forecast due to the non-linear nature of the dynamics (Lorenz, 1965), this assumption should be reasonable beyond the short range.

As well as by Leith, the statistical interpretation of ensemble forecasts is discussed by Hayashi (1986). However, whereas these authors use error variance to quantify forecast skill, we use a normalised version of this amplitude-based score, and also consider the phase-emphasising anomaly correlation score, in addition to investigating the a priori prediction of skill. The reader should note that our notation is not in general the same as that used in either of the above papers.

3. MEASURES OF FORECAST SPREAD AND SKILL

The simplest measure of the spread of the forecast p.d.f. ψ is its variance D, given by $D = \bar{d}_j=\overline{(\mathbf{u}_j-\bar{\mathbf{u}})^2}$. Initially D is small, reflecting observation and analysis errors. As the forecast proceeds D becomes larger, eventually approaching the climate variance as predictability disappears. From an ensemble of size M we may estimate D through the ensemble variance s_M, given by

$$s_M = \left(\sum_{j=1}^{M}(\mathbf{u}_j-\hat{\mathbf{u}})^2\right)/M=\hat{d}_j-(\hat{\mathbf{u}}-\bar{\mathbf{u}})^2 \qquad (1)$$

$$=((M-1)/M)\hat{d}_j+(2/M^2)\sum_{j=1}^{M}\sum_{k=j+1}^{M}(\mathbf{u}_j-\bar{\mathbf{u}}).(\mathbf{u}_k-\bar{\mathbf{u}}).$$

The second term in equation (1) is a sum of random covariances, whose average value is zero.

Thus if we consider an infinite number of size M ensembles drawn independently from ψ, we have

$$S_M = \bar{s}_M = ((M-1)/M)D, \qquad (2)$$

so that from a single ensemble of M forecasts, $(M/(M-1))s_M$ is an unbiased estimate of the 'amplitude' dispersion D of ψ.

A measure of skill corresponding to D is the error variance, the squared distance in phase space from the forecast to the true state. The error variance of an ensemble-mean forecast is

$$e_M=(\hat{\mathbf{u}}-\mathbf{u}_o)^2=(\hat{\mathbf{u}}-\bar{\mathbf{u}})^2+(\mathbf{u}_o-\bar{\mathbf{u}})^2-2(\hat{\mathbf{u}}-\bar{\mathbf{u}}).(\mathbf{u}_o-\bar{\mathbf{u}})$$

$$=\hat{d}_j/M+d_o-2(\hat{\mathbf{u}}-\bar{\mathbf{u}}).(\mathbf{u}_o-\bar{\mathbf{u}}). \qquad (3)$$

Thus E_M, the mean error variance of size M ensembles drawn independently from ψ, is given by

$$E_M=\bar{e}_M=\bar{d}_j/M+d_o=D/M+d_o. \qquad (4)$$

If we average over an infinite number of independent forecast p.d.f.'s, created from initial state p.d.f.'s sampled from an appropriate climatological distribution (see section 2), we obtain

$$\langle E_M\rangle=((M+1)/M)\langle D\rangle. \qquad (5)$$

Alternatively, we may express the right hand side in terms of the skill of individual forecasts, in which case

$$\langle E_M\rangle=((M+1)/2M)\langle E_1\rangle. \qquad (6)$$

On average, ensemble-mean forecasts are therefore closer in phase space to the true state than are individual forecasts. From this result, first obtained by Leith (1974), we could argue that ensemble-mean forecasts are consequently more skilful than individual forecasts. However, a reduction in the error variance of an individual forecast may also be achieved simply by smoothing it towards the climate mean. For example at some large forecast time, when all predictability has disappeared, and $\bar{\mathbf{u}}$ is thus coincident with the climate mean, the mean error variance of a climatology forecast, $\langle u_o^2\rangle$, is one half of the mean error variance of an individual forecast chosen at random from ψ. Clearly, however, the act of smoothing such a forecast towards climatology would not render the forecast useful in any practical sense, despite the reduction in error variance.

A more appropriate way of measuring the usefulness of a prediction is to use a 'normalised' error variance e'_M for the verification of ensemble-mean forecasts, where

$$e'_M=e_M/(\langle w_M\rangle+\langle w_o\rangle),$$

and hence

$$\langle E'_M\rangle=\langle E_M\rangle/(\langle w_M\rangle+\langle w_o\rangle), \qquad (7)$$

where $E'_M=\bar{e}'_M$, and w_o and w_M are the 'anomaly intensity' $\mathbf{u}_o^2$ and $(\hat{\mathbf{u}})^2$ of a true state and an ensemble mean forecast respectively. The denominator is equal to the 'no-skill' level of error variance calculated by verifying M-member ensemble-mean forecasts against randomly chosen, unrelated true states.

By normalising the error variance in this way, we express forecast skill essentially in signal-to-noise terms. When the signal (i.e. predictability) is large, e'_M is small, and as predictability disappears and ψ approaches the model climate distribution, $e'_M\to 1$. This skill score should therefore reflect aptly the benefit of ensemble-averaging, namely the separation of the signal, $\bar{\mathbf{u}}$, from the noise, D, in the forecast p.d.f. On this basis we may expect the normalised error variance to decrease with increasing M, as $\hat{\mathbf{u}}$ approximates $\bar{\mathbf{u}}$ more precisely.

To derive the variation of $\langle E'_M\rangle$ with M, we first combine equations (6) and (7) to give

$$\langle E'_M\rangle=((M+1)/M)\langle E'_1\rangle/(1+\langle w_M\rangle/\langle w_o\rangle).$$

Since

$$\langle w_o\rangle=\langle(\mathbf{u}_o-\bar{\mathbf{u}})^2\rangle+\langle\bar{\mathbf{u}}^2\rangle=\langle D\rangle+\langle w_\infty\rangle, \qquad (8)$$

where $w_\infty=\bar{\mathbf{u}}^2$, and similarly

$$\langle w_M\rangle=\langle D\rangle/M+\langle w_\infty\rangle, \qquad (9)$$

we have

$$\langle w_M\rangle/\langle w_o\rangle=1-((M-1)/M)\langle D\rangle/\langle w_o\rangle$$
$$=1-((M-1)/M)\langle E'_1\rangle,$$

giving

$$\langle E'_M\rangle=\langle E'_1\rangle/(\langle E'_1\rangle+(2M/(M+1))(1-\langle E'_1\rangle)). \qquad (10)$$

Equation (10) confirms the improvement in skill (i.e. the decrease in $\langle E'_M\rangle$), as the size of ensemble increases. As $M\to\infty$, $\langle E'_M\rangle\to\langle E'_1\rangle/(2-\langle E'_1\rangle)$.

In addition to these amplitude-based statistics, we may also measure predictability in terms of the phase correspondence between members of a forecast p.d.f. An appropriate measure of the dispersion of ψ in phase terms is R, the squared anomaly correlation between $\bar{\mathbf{u}}$, and all the individual forecasts in ψ taken together, which is given by

$$R=(\overline{\bar{\mathbf{u}}.\mathbf{u}_j})^2/w_\infty\bar{w}_j=w_\infty/\bar{w}_j. \qquad (11)$$

At initialisation time R is close to unity. As errors grow during the forecast R decreases, eventually tending to zero as $\bar{\mathbf{u}}$ approaches the climate mean, at which time complete loss of phase predictability has occurred. We may obtain an estimate of R from an ensemble of M forecasts, using the squared anomaly correlation between the ensemble-mean, and all the individual forecasts within the ensemble taken together, given by

$$a_M=(M^{-1}\sum_{j=1}^{M}\hat{\mathbf{u}}.\mathbf{u}_j)^2/w_M\hat{w}_j=w_M/\hat{w}_j. \qquad (12)$$

If we write $\delta a_M=a_M-\bar{a}_M$ and $\delta\hat{w}_j=\hat{w}_j-\bar{w}_j$ to represent sampling variations in a_M and $\hat{w}_j$, then from equation (12) we have

$$\bar{a}_M \, \bar{w}_j + \overline{\delta a_M \delta \hat{w}_j} = \bar{w}_M.$$

We shall assume that the second term, representing a correlation between sampling errors, may be neglected, so that

$$\bar{a}_M \, \bar{w}_j = \bar{w}_M.$$

Since $\bar{w}_M = D/M+w_\infty$, and $\bar{w}_j=D+w_\infty$, it follows that

$$A_M=\bar{a}_M=(1/M)(1+(M-1)R). \qquad (13)$$

As for the amplitude case, the ensemble phase dispersion a_M is not an unbiased estimate of the dispersion R of the forecast p.d.f.

From the definitions of the phase (a_M) and amplitude (s_M) ensemble dispersion, it follows that the two are related by

$$a_M=1-(s_M/\hat{w}_j).$$

Similarly, from the definitions of the corresponding quantities R and D relating to the forecast p.d.f., we have

$$R=1-(D/\bar{w}_j)=1-F\langle w_j\rangle/\bar{w}_j, \qquad (14)$$

where the variance ratio $F=D/\langle w_j\rangle$ expresses the amplitude predictability of ψ as a fraction of the model's climate variance. Equation (14) shows that, for any particular forecast p.d.f., phase and amplitude predictability may be regarded as complementary. For example if F=1, representing complete loss of amplitude predictability, we may still have phase predictability (i.e. R>0), if $\bar{w}_j>\langle w_j\rangle$. Similarly if R=0, we may still have F<1 if $\bar{w}_j<\langle w_j\rangle$. If we average over many different forecast p.d.f.'s, the mean phase and amplitude predictability may be deduced from each other, since $\langle R\rangle=1-\langle F\rangle$.

The phase skill of a forecast is measured by the anomaly correlation score, which, for an ensemble-mean forecast, is given by

$$c_M=\hat{\mathbf{u}}.\mathbf{u}_o/(w_M w_o)^{1/2}. \qquad (15)$$

Using δ to represent sampling deviations, so that $\delta c_M=c_M-\bar{c}_M$, $\delta w_M=w_M-\bar{w}_M$, and $\delta w_o=w_o-\bar{w}_j$ (note that $\langle w_o\rangle=\langle\bar{w}_j\rangle$), we may rewrite (15) as

$$(\bar{c}_M+\delta c_M)(\bar{w}_M+\delta w_M)^{1/2}(\bar{w}_j+\delta w_o)^{1/2}=\hat{\mathbf{u}}.\mathbf{u}_o.$$

We shall assume that such sampling deviations in anomaly intensity are negligibly small, so that $\overline{(\bar{w}_M+\delta w_M)^{1/2}}=\bar{w}_M{}^{1/2}$, and $\langle(\bar{w}_j+\delta w_o)^{1/2}\rangle=\langle\bar{w}_j{}^{1/2}\rangle$, giving

$$\langle\bar{c}_M \, \bar{w}_M{}^{1/2}\bar{w}_j{}^{1/2}\rangle=\langle\bar{\mathbf{u}}.\mathbf{u}_o\rangle=\langle w_\infty\rangle.$$

If we assume also that 'climatic' variations, in anomaly intensity, such as $\delta_c\bar{w}_M=\bar{w}_M-\langle\bar{w}_M\rangle$, are also small, we may write, noting that $\langle c_M\rangle=\langle\bar{c}_M\rangle$, $\langle w_M\rangle=\langle w_M\rangle$ and $\langle w_j\rangle=\langle\bar{w}_j\rangle$,

$$\langle c_M\rangle=\langle w_\infty\rangle/(\langle w_M\rangle\langle w_j\rangle)^{1/2}. \qquad (16)$$

From equations (8) and (9),

$$\langle w_M\rangle=(\langle w_j\rangle+(M-1)\langle w_\infty\rangle)/M.$$

Substitution into equation (16) gives

$$\langle c_M\rangle=M^{1/2}(\langle w_\infty\rangle/\langle w_j\rangle)/(1+(M-1)(\langle w_\infty\rangle/\langle w_j\rangle))^{1/2}. \qquad (17)$$

With M=1 (i.e. for individual forecasts), equation (17) gives

$$\langle c_1\rangle=\langle w_\infty\rangle/\langle w_j\rangle,$$

so that equation (17) may be written

$$\langle c_M\rangle=M^{1/2}\langle c_1\rangle/(1+(M-1)\langle c_1\rangle)^{1/2}, \qquad (18)$$

showing the variation with M of the ensemble-mean forecast anomaly correlation. Clearly $\langle c_M\rangle$ increases with M, and as $M\to\infty$, $\langle c_M\rangle\to\langle c_1\rangle^{1/2}$. This reflects the elimination, through ensemble-averaging, of random phase errors attributable to the 'noise' R in the forecast p.d.f.

Equation (18) is found to predict well the mean improvement in phase skill observed in a series of seven-member perfect model ensemble forecast experiments carried out using a 5-level GCM (Murphy, 1986), and the improvement in amplitude skill also corresponds closely to equation (10). Further discussion of the implications of these results is given in the above paper.

4. CORRELATION BETWEEN FORECAST SPREAD AND SKILL

In the previous section we indicated the two possible causes of variation in skill between ensemble forecasts created from independent initial state p.d.f.'s. One is the sampling error incurred, in any particular instance, by approximating the forecast p.d.f. ψ with a finite ensemble of size M (≥ 1), and the other is the climatic variation caused by genuine case-by-case differences in ψ, in terms of mean and/or spread.

The extent to which the variation in skill may be predicted in advance from the corresponding variation in ensemble spread depends on the relative magnitude of the sampling and climatic contributions, with the former representing the 'unpredictable' proportion of the variation and the latter the 'predictable' proportion.

We may use a correlation coefficient to measure the relationship between ensemble spread and the skill of an ensemble-mean forecast. In terms of amplitude predictability, if we measure the spread of an ensemble by its variance s_M, and its skill by the error variance e_M of the ensemble-mean forecast, then the correlation ρ_A between these two quantities is

We may use a correlation coefficient to measure the relationship between ensemble spread and the skill of an ensemble-mean forecast. In terms of amplitude predictability, if we measure the spread of an ensemble by its variance s_M, and its skill by the error variance e_M of the ensemble-mean forecast, then the correlation ρ_A between these two quantities is

$$\rho_A = \frac{\langle (e_M-\langle e_M\rangle)(s_M-\langle s_M\rangle)\rangle}{[\langle (e_M-\langle e_M\rangle)^2\rangle\langle (s_M-\langle s_M\rangle)^2\rangle]^{1/2}} . \qquad (19)$$

The choice of e_M rather than the normalised error variance e'_M is made purely for the sake of simplicity, since ρ_A remains unaltered if e'_M is used.

We wish to express ρ_A in terms of the sampling and climatic variations alluded to above. The sampling deviations of s_M and e_M are

$$\delta s_M = s_M - S_M$$

$$\delta e_M = e_M - E_M .$$

The climatic deviations of the mean spread S_M, and skill E_M, of size M ensembles drawn independently from ψ, is

$$\delta_c S_M = S_M - \langle S_M\rangle$$

$$\delta_c E_M = E_M - \langle E_M\rangle .$$

Since sampling and climatic deviations are statistically independent, and noting that $\langle s_M\rangle = \langle S_M\rangle$ and $\langle e_M\rangle = \langle E_M\rangle$, we may rewrite equation (19) as

$$\rho_A = \frac{\langle \delta e_M \delta s_M\rangle + \langle \delta_c E_M \delta_c S_M\rangle}{[(\langle \delta e_M{}^2\rangle + \langle \delta_c E_M{}^2\rangle)(\langle \delta s_M{}^2\rangle + \langle \delta_c S_M{}^2\rangle)]^{1/2}} . \qquad (20)$$

The mean extent of sampling variation is represented by $\langle \delta d_j{}^2\rangle$, where $\delta d_j = d_j - D$ is the amount by which the squared distance in phase space from forecast j to the centroid of ψ differs from the average value D for all forecasts within ψ. We may express the sampling variations in s_M and e_M in terms of $\langle \delta d_j{}^2\rangle$ as follows.

From equations (1) and (2) we have

$$\delta s_M = ((M-1)/M)(\hat{d}_j - D) = ((M-1)/M)\delta\hat{d}_j , \qquad (21)$$

where we have assumed for the purposes of this calculation that the random covariance term in equation (1) may be neglected. The typical size of this term depends on the number of independent variables in the model, and the typical space scale on which, in a given forecast selected from ψ, deviations from the distribution mean at different grid points are correlated. This in turn would depend on the structure of initial state errors and the manner of their growth during the forecast. The degree of spatial or temporal filtering applied to the forecast fields would also be a factor. However, it is reasonable to suppose that, for a model with a large number of degrees of freedom, such a term will always be small (Clearly if this is not so, $\langle \delta s_M^2 \rangle$ is liable to be larger, and ρ_A smaller, than our theoretical estimates indicate.)

Thus, with $\delta\hat{d}_j = (\sum_{j=1}^{M} \delta d_j)/M$, we then have

$$\delta s_M^2 = ((M-1)^2/M^4)(\sum_{j=1}^{M} \delta d_j)^2.$$

Since the cross-terms in the expansion of the right hand side are uncorrelated, we obtain

$$\langle \delta s_M^2 \rangle = ((M-1)^2/M^3)\langle \delta d_j^2 \rangle. \qquad (22)$$

Similarly, from equation (3),

$$\delta e_M = (\delta\hat{d}_j/M) - 2(\hat{\mathbf{u}} - \bar{\mathbf{u}}).(\mathbf{u}_o - \bar{\mathbf{u}}). \qquad (23)$$

Again we shall assume that second term on the right hand side in equation (23), which represents a random covariance, may be neglected, in which case

$$\langle \delta e_M^2 \rangle = \langle \delta\hat{d}_j^2 \rangle / M^2 = \langle \delta d_j^2 \rangle / M^3. \qquad (24)$$

The covariance between δs_M and δe_M is found by multiplying equations (21) and (24) together, from which we find

$$\langle \delta e_M \delta s_M \rangle = ((M-1)/M^2)\langle \delta d_j^2 \rangle = ((M-1)/M^3)\langle \delta d_j^2 \rangle. \qquad (25)$$

(Note that it is not necessary to assume random covariance terms to be zero for the purpose of deriving equation (25)).

Thus sampling variations in spread are $\sim M^2$ larger than sampling variations in ensemble-mean forecast skill (equations (23) and (24)), since deviations of individual forecasts about an ensemble-mean are essentially M times larger than that of an ensemble-mean about the mean $\bar{\mathbf{u}}$ of the forecast p.d.f.

We may derive the climatic variation in spread from equation (2). Defining the climatic deviation in the variance D, of ψ, as $\delta_c D = D - \langle D \rangle$, we have

$$\delta_c s_M = ((M-1)/M)\delta_c D, \qquad (26)$$

from which we obtain

$$\langle \delta_c s_M^2 \rangle = ((M-1)/M)^2 \langle \delta_c D^2 \rangle. \qquad (27)$$

The climatic variation in skill may be deduced from equation (4), by writing

$$\delta_c E_M=(\delta_c D/M)+(d_o-\langle D\rangle)$$

$$=(\delta_c D/M)+(d_o-D)+(D-\langle D\rangle)$$

$$=((M+1)/M)\delta_c D+(d_o-D). \qquad (28)$$

Squaring and averaging equation (28) gives

$$\langle\delta_c E_M{}^2\rangle=((M+1)/M)^2\langle\delta_c D^2\rangle+\langle\delta d_j{}^2\rangle, \qquad (29)$$

assuming that, just as $\langle d_o\rangle=\langle d_j\rangle$, $\langle\delta d_o{}^2\rangle=\langle\delta d_j{}^2\rangle$ for $t>0$, with j representing a forecast selected at random from ψ. The covariance between spread and ensemble-mean forecast skill follows easily by multiplying equations (26) and (28), giving

$$\langle\delta_c E_M\delta_c S_M\rangle=((M+1)(M-1)/M^2)\langle\delta_c D^2\rangle. \qquad (30)$$

We may define $\alpha=\langle\delta_c D^2\rangle/\langle\delta d_j{}^2\rangle$ as an appropriate measure of the ratio of climatic to sampling variations, in which case, substituting the relevant results into equation (20), we find

$$\rho_A=[M^{-2}+((M+1)/M)\alpha]/[(M^{-1}+\alpha)(M^{-3}+1+((M+1)/M)^2\alpha)]^{1/2}. \qquad (31)$$

Equation (31) shows the correlation between ensemble spread, and the skill of the ensemble-mean forecast, that we may expect under perfect model conditions using ensembles of size M, in a scenario where a ratio α exists between the real case-by-case variation $\langle\delta_c D^2\rangle$ in the spread of the forecast p.d.f., and the mean extent $\langle\delta d_j{}^2\rangle$ of sampling error involved in the measurement of the spread. In practice the effect of model imperfection would render ρ_A smaller than predicted by equation (31), which represents an upper limit.

Clearly, for given M, we should expect ρ_A to increase with α, which is confirmed by equation (31) (see Table 3 and associated discussion in Murphy, 1986). It is important to appreciate the variation of ρ_A with α for several reasons. The value of α is likely to depend on season, and also on whether spatially or temporally filtered fields are being used. In addition, although we may estimate α from a finite number of model ensemble forecast experiments, the value will be subject to sampling error, and may also of course be model-dependent. The value of α determines how accurately we need to know the true spread of ψ in order to obtain a satisfactory prediction of skill. As ensemble size increases, the estimate of the true spread of ψ becomes more precise. The coefficient ρ_A is not suitable to quantify the effect of this increased precision however. This is because two terms contribute to the case-by-case variation $\langle\delta_c E_M{}^2\rangle$ in the average skill E_M of M-member ensemble-mean forecasts drawn independently from ψ (equation (29)), one being a 'predictable' element due to variations in the spread of ψ, and the other an 'unpredictable' element arising from the prior uncertainty of the distance in phase space from the true state to the centroid of ψ. As M increases, the size of the 'predictable' term $((M+1)/M)^2\langle\delta_c D^2\rangle$ decreases, which inhibits the increase in ρ_A that we would anticipate through having available a more precise estimate of spread.

We may better illustrate the benefit of this extra precision by considering E_1, the mean error variance of all individual forecasts within ψ. The correlation ρ'_A between ensemble spread and E_1 is given by

$$\rho'_A = \frac{<(E_1-<E_1>)(s_M-<s_M>)>}{[<(E_1-<E_1>)^2><(s_M-<s_M>)^2>]^{1/2}}$$

$$= \frac{<\delta_c E_1 \delta_c S_M>}{[<\delta_c E_1{}^2>(<\delta s_M{}^2>+<\delta_c S_M{}^2>)]^{1/2}} \; . \qquad (32)$$

From equation (4) we deduce

$$\delta_c E_1=\delta_c D+d_o-<D>=2\delta_c D+d_o-D \; . \qquad (33)$$

Thus

$$<\delta_c E_1{}^2>=4<\delta_c D^2>+<\delta d_o{}^2>=4<\delta_c D^2>+<\delta d_j{}^2>,$$

independent of ensemble size, as required. The covariance term $<\delta_c E_1 \delta_c S_M>$ follows from equations (26) and (33), giving

$$<\delta_c E_1 \delta_c S_M>=2((M-1)/M)<\delta_c D^2> \; .$$

Substitution into (32) gives the result

$$\rho'_A=2\alpha/[(\alpha+M^{-1})(4\alpha+1)]^{1/2} \; . \qquad (34)$$

The proportion of $<\delta_c E_1{}^2>$ predictable from variations in spread is constant, thus rendering ρ'_A a suitable statistic with which to demonstrate the increased precision in our prediction of skill achieved by increasing M. Table 4 of Murphy (1986) shows the increase of ρ'_A with M for various values of α, and the implications of the results are discussed further in that paper.

It is worth emphasising the complementary nature of the correlation coefficients ρ_A and ρ'_A. We should use ρ'_A to deduce an appropriate value of M which achieves a satisfactory degree of precision in our prediction of skill, and having done this, ρ_A then reveals the (maximum) extent to which we could predict variations in the skill of the best available forecast (the ensemble-mean), if we were to produce regular ensemble forecasts of membership M in practice.

The expression for ρ_A is easier to verify by experiment (see Murphy, 1986), than that for ρ'_A, since in the latter case we would require ensembles of a very large size to obtain sufficiently accurate estimates of E_1. A problem we face in verifying experimentally obtained values of either coefficient is to determine the appropriate value of α from the data. We cannot estimate $<\delta d_j{}^2>$ directly, since in any ensemble forecast experiment we do not know $\bar{\mathbf{u}}$, but an approximate value may be obtained by calculating, from each experiment, the quantity v_s given by

$$v_s = (\sum_{j=1}^{M} ((u_j-\hat{u})^2-s_M)^2)/M, \qquad (35)$$

which represents the ensemble variance of $(\mathbf{u}_j-\hat{\mathbf{u}})^2$. From the above definition, and equation (1), with the random covariance term again neglected, we obtain

$$v_s = \left(\sum_{j=1}^{M} (d_j-(1-2/M)\hat{d}_j-2\delta\mathbf{u}_j.\delta\hat{\mathbf{u}})^2\right)/M,$$

where $\delta\mathbf{u}_j=\mathbf{u}_j-\bar{\mathbf{u}}$, and $\delta\hat{\mathbf{u}}=\hat{\mathbf{u}}-\bar{\mathbf{u}}$. But we may write

$$\delta\mathbf{u}_j.\delta\hat{\mathbf{u}} = \left(\sum_{k=1}^{M} \delta\mathbf{u}_j.\delta\mathbf{u}_k\right)/M = d_j/M, \qquad (36)$$

again disregarding covariance terms $\delta\mathbf{u}_j.\delta\mathbf{u}_k$. Thus we have

$$\langle v_s\rangle = ((1-2/M)^2/M)\langle \sum_{j=1}^{M} (d_j-\hat{d}_j)^2\rangle$$

$$= (1-2/M)^2(\langle d_j{}^2\rangle-\langle(\hat{d}_j)^2\rangle). \qquad (37)$$

Since $\langle\delta d_j{}^2\rangle=\langle d_j{}^2\rangle-\langle D^2\rangle$, and $\langle\delta(\hat{d}_j)^2\rangle=\langle(\hat{d}_j)^2\rangle-\langle D^2\rangle$, equation (36) becomes

$$\langle v_s\rangle=(1-2/M)^2(\langle\delta d_j{}^2\rangle-\langle\delta(\hat{d}_j)^2\rangle)$$

$$=((M-1)(M-2)/M^3)\langle\delta d_j{}^2\rangle. \qquad (38)$$

Thus an estimate of $\langle\delta d_j{}^2\rangle$ may be recovered by measuring v_s in a large number of independent ensemble forecast experiments, assuming that the approximation concerning covariance terms is reliable. By also measuring $\langle(s_M-\langle s_M\rangle)^2\rangle$ experimentally, we may then deduce $\langle\delta_c D^2\rangle$, and hence α.

Phase correlation between spread and skill

We may define a correlation coefficient ρ_P, analogous to ρ_A, which shows the relationship between the phase spread a_M of an ensémble, and the phase skill of the ensemble-mean forecast measured by c'_M, where $c'_M=\pm|c_M|^2$, choosing the negative sign if c_M is negative so that we may distinguish between instances of positive and negative anomaly correlation.

$$\rho_P = \frac{\langle(c'_M-\langle c'_M\rangle)(a_M-\langle a_M\rangle)\rangle}{[\langle(c'_M-\langle c'_M\rangle)^2\rangle\langle(a_M-\langle a_M\rangle)^2\rangle]^{1/2}} . \qquad (39)$$

The phase correspondence between an individual forecast j, within ψ, and $\bar{\mathbf{u}}$, may be measured by

$$r_j = (\bar{\mathbf{u}}.\mathbf{u}_j)^2/w_\infty w_j . \qquad (40)$$

If we assume that both the covariance term $\bar{\mathbf{u}}.(\bar{\mathbf{u}}_j-\bar{\mathbf{u}})$, and the term $(r_j-\bar{r}_j)(w_j-\bar{w}_j)$, are negligible, then it follows from equation (40) that $\bar{r}_j=R$. Sampling variations in phase spread may then be characterised by $\langle\delta r_j{}^2\rangle=\langle(r_j-R)^2\rangle$, with climatic variations represented by $\langle\delta_c R^2\rangle=\langle(R-\langle R\rangle)^2\rangle$. In theory we might derive an expression for ρ_P in terms of $\langle\delta r_j{}^2\rangle$ and $\langle\delta_c R^2\rangle$, in analogy to the calculation for ρ_A leading to equation (31). A problem exists, however, in that at large forecast times, as $\bar{\mathbf{u}}$→0, $\langle\delta r_j{}^2\rangle$→0. This contrasts with the amplitude case, in which the corresponding quantity $\langle \delta d_j^2 \rangle$ remains finite as $t \to \infty$. Thus the necessary assumption , that random covariance terms may be safely ignored relative to such sampling terms, becomes invalid in the phase case as predictability attenuates with time. For this reason we do not perform the calculation for ρ_P analogous to that for ρ_A.

5. SUMMARY

We have derived simple theoretical estimates of the effect on forecast skill of ensemble-averaging a number of individual model predictions, based on the assumption of a perfect forecast model, namely that given the correct initial state, the model produces an error-free forecast. An ensemble forecast is a sampling approximation to a continuous probability density function (p.d.f.) composed of an infinite number of equally-likely forecasts, each consistent with the errors which inevitably accompany a given initial analysis. Although such a p.d.f. may in principle have a complex form we consider only its first two moments (thus making no attempt to distinguish it from a normal distribution), which give the best-estimate forecast and its associated uncertainty.

Two measures of predictability are considered, emphasising respectively the correspondence in amplitude and phase between the forecast and actual anomaly patterns. The amplitude measure is the forecast error variance, normalised by the mean 'no-skill' error variance between a forecast and an unrelated, randomly-chosen actual state. This normalisation is important as the no-skill error variance of an ensemble-mean forecast varies with the size of the ensemble, so that the unnormalised error variance gives a misleading indication of the variation of useful skill with ensemble size. The anomaly correlation score is used to measure phase predictability.

Equations (10) and (18), which give respectively the mean amplitude and phase skill of ensemble-mean forecasts in terms of individual forecast skill, show that skill increases with size of ensemble, reflecting the increased precision in the estimate of the mean of the forecast p.d.f. available from a larger ensemble. The implications of these results for extended range forecasting are discussed further in Murphy (1986), where the theoretical predictions are shown to compare well with experimentally-obtained perfect model results from seven-member ensemble forecasts. The use of a perfect model assumption means that our calculations will represent an upper limit to the improvement in skill likely to be obtainable in practical ensemble forecasting.

Another important question is that of the a priori prediction of forecast skill. In theory the spread of an ensemble should provide an indication of the likely skill of the forecast. Measures of amplitude and phase ensemble spread are discussed and correlation coefficients defined to measure the relationship between spread and skill. In the amplitude case expressions are derived for two correlation coefficients (distinguished by slightly different methods of measuring skill), in terms of 'unpredictable' sampling variation within a given forecast p.d.f., and 'predictable', so-called climatic, variation between independent forecast p.d.f.'s. Results show that larger correlations apply for larger ratios of climatic to sampling variation, and that the precision of the prediction of skill increases with ensemble size. In the phase case the corresponding calculation is omitted, since a necessary simplifying assumption, that random covariances between model state vectors may be neglected in comparison with other terms, whilst reasonable in the amplitude case, is not valid in the phase case at forecast times corresponding to the extended range period. Further discussion of the correlation between forecast spread and skill is given in Murphy (1986).

References

Dalcher, A., Kalnay, E. and Hoffman, R.N. 1985 Application of lagged average forecasting to medium-range forecasting. Seventh Conference on Numerical Weather Prediction, Boston, Mass., Amer. Met. Soc., 202-207.

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

Gleeson, T.A. 1970 Statistical-dynamical predictions. J. Appl. Meteor., 9, 333-344.

Hayashi, Y. 1986 Statistical interpretations of ensemble-time mean predictability. J. Met. Soc. Jpn., 64, 167-181.

Hoffman, R.N. and Kalnay, E. 1983 Lagged-average forecasting, an alternative to Monte Carlo forecasting. Tellus, 35, 100-118.

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

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

Murphy, J.M. 1986 The impact of ensemble forecasts on predictability. Submitted to Quart. J. Roy. Met. Soc.

Murphy, J.M. and Palmer, T.N. 1986 Experimental monthly long-range forecasts for the United Kingdom. Part II. A real-time long-range forecast by an ensemble of numerical integrations. Met. Mag., 115, 337-349.

Seidman, A.N. 1981 Averaging techniques in long-range weather forecasting. Mon. Wea. Rev., 109, 1367-1379.

Shukla, J. 1981 Dynamical predictability of monthly means. J. Atmos. Sci., 38, 2547-2572.

INDEX TO LONG-RANGE FORECASTING AND CLIMATE RESEARCH SERIES

1) THE CLIMATE OF THE WORLD - Introduction and description of world climate.

by C K Folland (March 1986)

2) THE CLIMATE OF THE WORLD - Forcing and feedback processess.

by C K Folland (March 1986)

3) THE CLIMATE OF THE WORLD - El Nino/Southern Oscillation and the Quasi-biennial Oscillation.

by C K Folland (March 1986)

4) THE CLIMATE OF THE WORLD - Climate change: the ancient earth to the 'Little Ice Age'.

by C K Folland (March 1986)

5) THE CLIMATE OF THE WORLD - Climate change: the instrumental period.

by C K Folland (March 1986)

6) THE CLIMATE OF THE WORLD - Carbon dioxide and climate (with appendix on simple climate models).

by C K Folland (March 1986)

7) Sahel rainfall,Northern Hemisphere circulation anomalies and worldwide sea temperature changes. (To be published in the Proceedings of the "Pontifical Academy of Sciences Study Week", Vatican, 23-27 September 1986).

by C K Folland, D E Parker, M N Ward and A W Colman (September 1986)

8) Lagged-average forecast experiments with a 5-level general circulation model.

by J M Murphy (March 1986)

9) Statistical Aspects of Ensemble Forecasts.

by J M Murphy (July 1986)

10) The Impact of El Nino on an Ensemble of Extended-Range Forecasts. (Submitted to Monthly Weather Review)

by J A Owen and T N Palmer (December 1986)