Tellus (1983), *35A*, 100–118

Lagged average forecasting, an alternative to Monte Carlo forecasting

By ROSS N. HOFFMAN and EUGENIA KALNAY, *Goddard Laboratory for Atmospheric Sciences, NASA/Goddard Space Flight Center, Greenbelt, MD 20771, U.S.A.*

(Manuscript received March 26; in final form August 2, 1982)

ABSTRACT

In order to use the information present in past observations and simultaneously to take advantage of the benefits of stochastic dynamic prediction we formulate the lagged average forecast (LAF) method. In a LAF, just as in a Monte Carlo forecast (MCF), sample statistics are calculated from an ensemble of forecasts. Each LAF ensemble member is an ordinary dynamical forecast (ODF) started from the initial conditions observed at a time lagging the start of the forecast period by a different amount. These forecasts are averaged at their proper verification times to obtain an LAF. The LAF method is operationally feasible since the LAF ensemble members are produced during the normal operational cycle.

To test the LAF method, we use a two-layer, f-plane, highly truncated spectral model, forced by asymmetric Newtonian heating of the lower layer. In the experiments, a long run is generated by the primitive equation version of the model which is taken to represent nature, while forecasts are made by the quasigeostrophic version of the model. On the basis of forecast skill, the LAF and MCF are superior to the ODF; this occurs principally because ensemble averaging hedges the LAF and MCF toward the climate mean. The LAF, MCF and ODF are all improved when tempered by a simple regression filter; this procedure yields different weights for the different members of the LAF ensemble. The tempered LAF is the most skillful of the forecast methods tested. The LAF and MCF can provide *a priori* estimates of forecast skill because there is a strong correlation between the dispersion of the ensemble and the loss of predictability. In this way the time at which individual forecasts lose their skill can be predicted.

The application of the LAF method to more realistic models and to monthly or seasonally averaged forecasts is briefly discussed.

1. Introduction

The probabilistic nature of the problem of numerical weather prediction (NWP) is well recognized; statistical considerations apply to the preparation of the data, to the model integration and to the regression filtering of the forecast. However, on an operational basis, the combination of statistics and dynamics during the model integration has not yet been accomplished, even though theoretical bases were discussed by Epstein (1969) and Gleeson (1970). A probabilistic approach to prediction has two immediate advantages over a deterministic approach. First, by explicitly considering the nonlinear evolution of error, forecast skill is improved in an r.m.s. sense. Second, by forecasting the expected dispersion about the forecast mean, an *a priori* estimate of the skill of the forecast mean is itself forecast.[1] The improvement in forecast skill may not be large when the basis of comparison is an optimally filtered deterministic forecast. The operational prediction of forecast skill, on the other hand, could greatly change our uses of and attitudes towards NWP products. A third advantage is that the

[1] These points are made in the stochastic dynamic context (Epstein, 1969). From the statistical-dynamical point of view (Gleeson, 1970) the forecast of the expected dispersion about the deterministic forecast, provides an *a priori* estimate of the skill of this forecast. This information could then be used to improve the forecast in terms of its skill, e.g., by forming a weighted average of the climate mean and the deterministic forecast.

forecast of the forecast error covariances could be used to improve the assimilation of data (Pitcher, 1977).

The advantages of a combined statistical and dynamical approach are most easily illustrated within the phase space description of an atmospheric model. In such a phase space each coordinate is the amplitude of one model variable. A point in the phase space is a model state; as the model evolves the point representing the model state describes a trajectory in phase space. (Discussions of the phase space description of meteorological dynamical systems can be found in papers by Fleming (1971), Epstein (1969), Gleeson (1970) and others.) Consider a particular point, C, and an ensemble of points uniformly covering a small sphere centered at C (Fig. 1a). We will call the forecast started from C the central forecast and the average of the forecasts started from the points on the sphere, the ensemble forecast. For short forecast times, $t < \tau_{NL}$, where τ_{NL} is the time when nonlinear effects can no longer be neglected, the difference between each member of the ensemble and the central forecast is governed by linear dynamics. The sphere deforms into an ellipsoid (Lorenz, 1965) and the central forecast and ensemble forecast coincide (Fig. 1b). Typically, for forced dissipative systems as the ellipsoid changes shape—flattening in some directions and expanding in others—it shrinks in volume while simultaneously the r.m.s. distance between its center and its surface increases. At longer forecast times, $t \gtrsim \tau_{NL}$, the central forecast and ensemble forecast differ because the ensemble loses its symmetry (Fig. 1c). The dispersion of the ensemble continues to increase in the nonlinear regime until some equilibrium range is attained. Now suppose that at the initial time, each point on the sphere is equally likely to be the true state. Of all possible *model* forecasts the central forecast is best for $t < \tau_{NL}$ because it minimizes the expected squared forecast error. However, the ensemble of forecasts contains more information: the ensemble forecast is a better forecast than the central forecast for $t > \tau_{NL}$ and the dispersion of the ensemble is indicative of the forecast skill. In general the initial true state will not be restricted to a spherical shell, but the above discussion applies to each surface of equal probability density in the phase space.

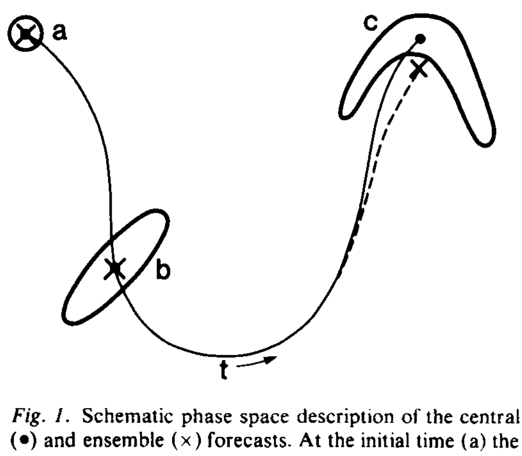

Fig. 1. Schematic phase space description of the central (•) and ensemble (×) forecasts. At the initial time (a) the central forecast is the center of the spherical ensemble. For $t < \tau_{NL}$ (b) the central and ensemble forecasts coincide; the ensemble is ellipsoidal. At later times, $t \gtrsim \tau_{NL}$ (c), the central forecast and the ensemble forecast diverge; the ensemble loses its symmetry.

Stochastic dynamic prediction (Epstein, 1969) offers a rational method of dealing with uncertainties in the initial data of an otherwise deterministic initial value problem by solving the continuity equation for probability which results from the generalization of the above discussion. It is not easy to include in this theory the effects of model imperfection; it is necessary to specify the statistical properties of the model errors (Fleming, 1972; Pitcher, 1977). This limits the reliability of the results of the method and it is likely that the higher order moments of the probability distribution are affected most. For this reason and for practical reasons prognostic equations are retained for only the lowest order moments of the probability distribution. This introduces a closure problem which may be treated by a variety of methods (Epstein, 1969; Fleming, 1971; Pitcher, 1977; Laurmann, 1978; Farago, 1978). Even so, implementation of the method is a formidable task for all but the simplest models: if the basic deterministic model has M prognostic variables then a stochastic dynamic model which ignores all but first and second moments has $O(M^2)$ prognostic variables. Furthermore one must derive and code the particular governing equation and parameterize the higher order moments.

One way of avoiding the difficulties of stochastic dynamic prediction is to use an ensemble forecast method, such as the Monte Carlo forecast (MCF) method (Epstein, 1969; Leith, 1974a; Seidman, 1981). In the simplest form of the MCF method an ensemble of N initial states are selected randomly from a collection of possible initial states which have a density in phase space proportional to the probability density function (p.d.f.) of the true state. Each ensemble member is then integrated in time and the ensemble of forecasts is used to calculate estimates of the desired statistics (Fig. 2a). Thus the MCF requires N times as much arithmetic as an ordinary dynamical forecast (ODF); ensemble forecast methods trade the closure problem of stochastic dynamic prediction for a sampling problem. Epstein used $N = O(1000)$, but Leith indicated that $N = 8$ may be sufficient.

There are two difficulties associated with the initial conditions needed for an MCF. The first difficulty is to specify the initial state p.d.f. The centroid of this p.d.f. is the best estimate of the true state; it may be obtained by an objective analysis procedure. It may be possible to estimate the higher order moments by using the statistics of the previous ensemble of forecasts, but, in order to obtain reliable second moments, the MCF sample size must be increased (Leith, 1974a, 1978a) and the external error growth due to the difference between model and nature must be parameterized by a stochastic forcing function (Fleming, 1972; Pitcher, 1977). The second difficulty is conceptual: no use is made of the ODF started from the best estimate of the true initial state.

To avoid the difficulties of the MCF method we propose a new ensemble forecast method which we call the lagged average forecast (LAF) method. In the LAF method, just as in the MCF method, sample statistics are calculated from an ensemble of forecasts. Each LAF ensemble member is an ODF started from the initial conditions observed at a time lagging the start of the forecast period by a different amount (Fig. 2b). Thus ODFs initialized with observations (or with an operational analysis) at $t = 0, -\tau, -2\tau, \ldots -(N-1)\tau$, are averaged, at their proper verification times, to obtain an LAF. The LAF method is operationally feasible since the LAF ensemble members are produced during the normal operational cycle at many meteorological centers. Like the recent approach of Gleeson (1981), the LAF uses only normally available operational forecast products. In the LAF method the most recent ODF is included in the ensemble and observations, extrapolated in time by the dynamical model, are used to define the initial uncertainty. If the model errors are small for a time of $O(N\tau)$, then the LAF ensemble will reflect the true initial uncertainty. The LAF may be considered an extension of an idea presented by Miyakoda and Talagrand (1971) in a way suggested by some results of Leith (1974b). Miyakoda and Talagrand, in a series of identical twin experiments, reduced the error of their standard objective analysis by averaging this analysis with model solutions started from prior

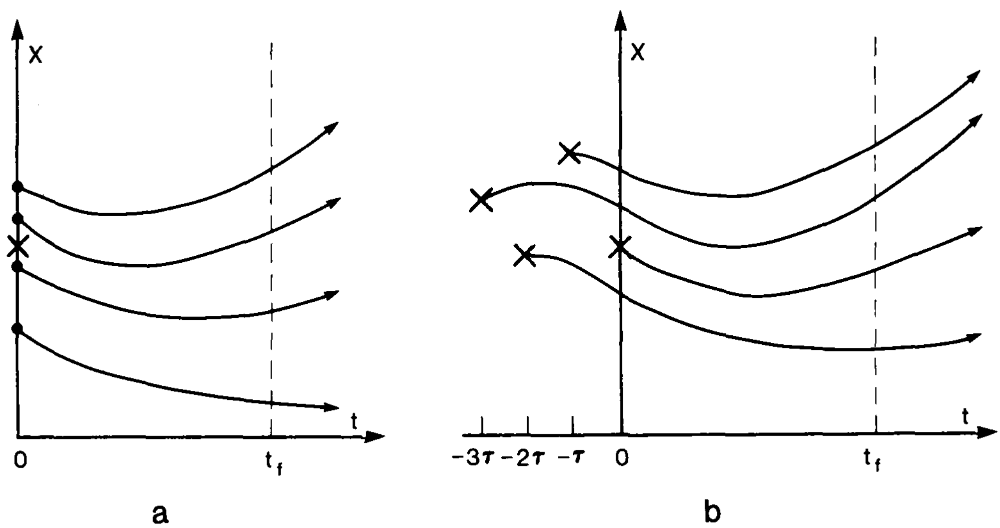

Fig. 2. Schematic time evolutions of the MCF (a) and LAF (b) ensembles. The abscissa is forecast time, t, and the ordinate is the amplitude of a model variable, X. X is observed (×) at intervals of time, τ; t_f is a particular forecast time.

initial conditions. Leith, in a statistical dynamical experiment, using a global spectral barotropic model to generate dynamical forecasts and real observations for verification, obtained better forecasts when his regression equations included as predictors information from dynamical forecasts initiated at two successive time levels. (These results are summarized in Fig. 2 of Leith (1978a).)

In this paper we test the LAF as an alternative to the MCF in the context of a highly simplified atmospheric model. As bases of comparison, we include the ODF and a persistence climatology forecast (PCF) in our tests of forecast skill. Following Leith (1974a), we have also included tempered or optimally filtered versions of the forecast methods in these tests. Tempering the forecast reduces the forecast deviations from the climate mean in order to minimize the expected mean squared forecast error. In other words, tempering hedges the forecast towards the climate mean.

The plan of this paper is as follows. We define the forecast methods in Section 2, summarize the dynamical models in Section 3, and describe the experiments and a simplified regression technique which we use to temper the forecasts in Section 4. The forecast skills of the different prediction methods are compared in Section 5. In Section 6 the spreading of the ensemble of forecasts is used to predict the average skill of the forecasts in the ensemble. Section 7 contains a summary of these results and a discussion of the application of the LAF method to more realistic models and to time averaged quantities.

2. Forecast methods

The forecast methods described here and listed in Table 1 range from purely dynamical to purely statistical. They may be applied to any finite dimensional dynamical system. The model state vector, $\mathbf{X}$, is presumed to be an adequate description of the true state of nature. It will be assumed that the required climate statistics are known with negligible error.

Table 1. *The forecast methods*

Acronym	Method
LAF	Lagged average forecast
MCF	Monte Carlo forecast
ODF	Ordinary dynamical forecast
PCF	Persistence climatology forecast
tLAF	Tempered lagged average forecast
tMCF	Tempered Monte Carlo forecast
tODF	Tempered ordinary dynamical forecast

2.1. *The ordinary dynamical forecast (ODF)*

The governing equation of the model is

$$\dot{\mathbf{X}} = \mathbf{F}(\mathbf{X}; t) \tag{1}$$

Eq. (1) states that the time rate of change of $\mathbf{X}$ is equal to a known function $\mathbf{F}$ of $\mathbf{X}$ and possibly of time t. The solution of (1) which starts from the initial conditions $\boldsymbol{\eta}$ at the time t_0 is denoted $\mathbf{X}(t; t_0, \boldsymbol{\eta})$. In order to conform to the notation used in what follows the ODF is defined as

$$\bar{\mathbf{X}}_1 = \mathbf{X}(t; 0, \mathbf{X}^o(0)) \tag{2}$$

where $\mathbf{X}^o(t)$ is the observation of $\mathbf{X}$ at time t. The time origin, $t = 0$, is the time of the most recent observation.

2.2. *The Monte Carlo forecast (MCF)*

Ensemble average forecasts of $\mathbf{X}$ are

$$\bar{\bar{\mathbf{X}}} = N^{-1} \sum_{i=1}^{N} \bar{\mathbf{X}}_i \tag{3}$$

where $\bar{\mathbf{X}}_i$ is the ith member of the ensemble of forecasts and N is the size of the ensemble. The ODF is an ensemble forecast where $N = 1$. For the MCF (Leith, 1974a)

$$\bar{\mathbf{X}}_i = \mathbf{X}(t; 0, \mathbf{X}^o(0) + \mathbf{Z}_i) \tag{4}$$

where the $\mathbf{Z}_i$ are random vectors with zero mean, uncorrelated with the observation $\mathbf{X}^o(0)$. The $\mathbf{Z}_i$ are chosen from a population with statistics which are the same as the statistics of the observing errors.

2.3. *The lagged average forecast (LAF)*

The LAF is given by (3) and

$$\bar{\mathbf{X}}_i = \mathbf{X}(t; t_i, \mathbf{X}^o(t_i)) \tag{5}$$

where the t_i are successive observing times. In the sequel $t_i = -(i-1)\tau$, for $i = 1, \dots N$, where τ is the interval between observing times. Thus the initial state for the ith lagged forecast, $\mathbf{X}^o(t_i)$, is observed at a time lag of $(i-1)\tau$.

2.4. *The tempered forecasts*

Leith (1974a) outlined an optimal filtering scheme in which any forecasting method may be made optimal in a least squares sense by performing a regression analysis of the true state in terms of the predicted state. This method was shown to be useful by Leith (1974b) and Lorenz (1977). Following Leith we will call such forecasts tempered. For an ensemble of forecasts, the tempered forecast, $\hat{\mathbf{X}}$, is given by

$$(\hat{\mathbf{X}} - \langle \mathbf{X} \rangle) = \sum_{i=1}^{N} \mathbf{A}_i(t)(\bar{\mathbf{X}}_i - \langle \mathbf{X} \rangle) \qquad (6)$$

where $\langle \mathbf{X} \rangle$ is the known climate mean, the $\mathbf{A}_i(t)$ are the regression matrices and the $\bar{\mathbf{X}}_i$ are obtained from (2), (4) or (5) for the tempered ODF (tODF), tempered MCF (tMCF) or tempered LAF (tLAF), respectively. For the tODF and tMCF, the $\mathbf{A}_i$ are independent of i; for the tLAF the regression naturally generates different weights for the individual forecasts. As formulated by Leith (1974a), the tempered forecast scheme is impractical because the size of the regression matrix is usually large compared to the size of the sample available to estimate the regression matrix. This problem is discussed by Lorenz (1977) who suggests that while screening regression may work well in some cases, a selection method based on physical considerations is preferable. In our experiments, this problem is avoided by introducing simplifying assumptions which limit the number of independent parameters to be estimated (Section 4).

2.5. *The persistence climatology forecast* (*PCF*)

The PCF models each variable X_j as a first-order autoregressive process; in the PCF the observed deviation from the climate mean for each model variable decays exponentially with time. The PCF, $\mathbf{X}^p$, is given by

$$\mathbf{X}^p(t) - \langle \mathbf{X} \rangle = \mathbf{q}^{t/\tau}(\mathbf{X}^o(0) - \langle \mathbf{X} \rangle) \qquad (7)$$

where the constant diagonal matrix $\mathbf{q}$, governing the decay of the observed deviations, is determined from climatology and τ is the interval between observations. The time constant for X_j is then $-\tau/\ln q_j$ where q_j is the jth entry on the diagonal of $\mathbf{q}$ and $0 \leqslant q_j \leqslant 1$.

A PCF provides a useful baseline against which to measure the other forecasts. For sufficiently long forecast times, the individual forecasts (i.e., the $\bar{\mathbf{X}}_i$) are not correlated with each other or with nature. When this holds, the expected mean squared forecast error of an ensemble forecast is greater than the expected mean squared forecast error of a PCF by a factor of $(1 + N^{-1})$.

3. Dynamical models used in the experiments

Lorenz (1960) formulated an adiabatic inviscid N-layer model of the atmosphere which has certain energy invariants. This model has three forms based on the primitive equations (PE), the nonlinear balance equations and the quasigeostrophic (QG) equations. We have used a 2-layer f-plane spectral version of this model (Hoffman, 1981; hereafter referred to as A). The following description of the model is complete but concise and the reader interested in the details of the model derivation is referred to Lorenz (1960) and to A. The models used here are identical to those used in the low forcing experiments reported in A except as noted below.

The variables of the 2-layer model are ψ and τ, half the sum and difference of the stream functions in the two layers, χ, the velocity potential in the lower layer, σ_0 the static stability and $\theta + \theta_0$, the potential temperature at the interface between the two layers. The scales of length, time and temperature are L, f^{-1} and $L^2 f^2 (C_p b)^{-1}$, where f is the Coriolis parameter, C_p is the specific heat of air at constant pressure and b is a constant ($\doteq 0.124$). For dimensional interpretations we assume f is (3 h)$^{-1}$ and L is the radius of the earth. At the top and bottom boundaries the material derivative of pressure is assumed to vanish; this implies that there are no external gravity waves and that the velocity potential in the upper layer is $-\chi$. Static stability, σ_0, does not vary horizontally. θ is the deviation from the horizontally averaged potential temperature at the interface. θ_0 is the difference between the horizontally averaged potential temperature at the interface and the horizontally averaged radiative equilibrium potential temperature forcing the lower layer. As in A, there is linear heating of the lower layer, linear friction at the lower surface and linear vertical exchanges of heat and momentum at the interface.[2] Each of θ, ψ, τ and χ is expanded in a complex Fourier series of the form

$$\zeta(x,y,t) = \sum_{\mathbf{I}} \zeta_{\mathbf{I}}(t) \exp(i\mathbf{I}\cdot\mathbf{R})$$

where $\mathbf{R} = (x,y)$ is the position vector and the sum is over all integer two-dimensional wave number vectors $\mathbf{I}$ excluding $\mathbf{0}$. The solutions are thus $2\pi L$ periodic in both x and y directions. In order that ζ be real, $\zeta_{-\mathbf{I}}$ must be equal to the complex conjugate of $\zeta_{\mathbf{I}}$ for all $\mathbf{I}$ and for all ζ. The equations of motion in spectral form for the PE and QG forms of the 2-layer model are then

$$\dot{\theta}_{\mathbf{I}} = -\sum \psi_{\mathbf{J}} c_{\mathbf{IJK}} \theta_{\mathbf{K}} + \sigma_0(-a_{\mathbf{I}}^2)\chi_{\mathbf{I}} - h_1(\theta_{\mathbf{I}} - \theta_{\mathbf{I}}^R)$$

$$\begin{aligned}(-a_{\mathbf{I}}^2)\dot{\psi}_{\mathbf{I}} = &-\sum \psi_{\mathbf{J}} c_{\mathbf{IJK}}(-a_{\mathbf{K}}^2)\psi_{\mathbf{K}} \\ &-\sum \tau_{\mathbf{J}} c_{\mathbf{IJK}}(-a_{\mathbf{K}}^2)\tau_{\mathbf{K}} + \delta\Big[\sum (-a_{\mathbf{J}}^2)\tau_{\mathbf{J}} d_{\mathbf{IJK}}\chi_{\mathbf{K}} \\ &+\sum (-a_{\mathbf{J}}^2)\chi_{\mathbf{J}} d_{\mathbf{IJK}}\tau_{\mathbf{K}} + \sum \chi_{\mathbf{J}} c_{\mathbf{IJK}}(-a_{\mathbf{K}}^2)\chi_{\mathbf{K}}\Big] \\ &-k_0(-a_{\mathbf{I}}^2)(\psi_{\mathbf{I}} - \tau_{\mathbf{I}})\end{aligned}$$

$$\begin{aligned}(-a_{\mathbf{I}}^2)\dot{\tau}_{\mathbf{I}} = &-\sum \psi_{\mathbf{J}} c_{\mathbf{IJK}}(-a_{\mathbf{K}}^2)\tau_{\mathbf{K}} \\ &-\sum \tau_{\mathbf{J}} c_{\mathbf{IJK}}(-a_{\mathbf{K}}^2)\psi_{\mathbf{K}} + (-a_{\mathbf{I}}^2)\chi_{\mathbf{I}} \\ &+\delta\sum (-a_{\mathbf{J}}^2)\psi_{\mathbf{J}} d_{\mathbf{IJK}}\chi_{\mathbf{K}} - k_1(-a_{\mathbf{J}}^2)\tau_{\mathbf{J}} d_{\mathbf{IJK}}(-a_{\mathbf{I}}^2)\psi_{\mathbf{I}}\end{aligned}$$

$$\begin{aligned}\delta(-a_{\mathbf{I}}^2)\psi_{\mathbf{I}} = &(-a_{\mathbf{I}}^2)\theta_{\mathbf{I}} - (-a_{\mathbf{I}}^2)\tau_{\mathbf{I}} \\ &-\delta\Big[\sum (-a_{\mathbf{J}}^2)\psi_{\mathbf{J}} d_{\mathbf{IJK}}\tau_{\mathbf{K}} + \sum (-a_{\mathbf{J}}^2)\tau_{\mathbf{J}} d_{\mathbf{IJK}}\psi_{\mathbf{K}} \\ &-(-a_{\mathbf{I}}^2)\sum \tau_{\mathbf{J}} e_{\mathbf{IJK}}\psi_{\mathbf{K}} - (-a_{\mathbf{I}}^2)\sum \psi_{\mathbf{J}} c_{\mathbf{IJK}}\chi_{\mathbf{K}} \\ &-\sum \chi_{\mathbf{J}} c_{\mathbf{IJK}}(-a_{\mathbf{K}}^2)\psi_{\mathbf{K}} - k_1(-a_{\mathbf{I}}^2)\chi_{\mathbf{I}}\Big]\end{aligned}$$

$$\dot{\sigma}_0 = -\sum_{\mathbf{K}} (-a_{\mathbf{K}}^2)\theta_{-\mathbf{K}}\chi_{\mathbf{K}} - h_0\sigma_0 - h_1\sigma_0 + h_1\theta_0$$

$$\dot{\theta}_0 = -h_1\theta_0 + h_1\sigma_0$$

where a dot over a variable denotes the time rate of change of that variable and $a_{\mathbf{I}}^2 = I^2$. We write $\sum$ for $\sum_{\mathbf{J}}\sum_{\mathbf{K}}$ and I^2 for $\mathbf{I}\cdot\mathbf{I}$. The interaction coefficients are

$$c_{\mathbf{IJK}} = (\mathbf{K}\times\mathbf{J})\, b_{\mathbf{IJK}}$$

$$d_{\mathbf{IJK}} = -2^{-1}(I^2 + K^2 - J^2)\, b_{\mathbf{IJK}}$$

$$e_{\mathbf{IJK}} = 2^{-1}(J^2 + K^2 - I^2)\, b_{\mathbf{IJK}}$$

where

$$b_{\mathbf{IJK}} = \begin{cases} 1, & \text{if } \mathbf{J} + \mathbf{K} = \mathbf{I} \\ 0, & \text{otherwise.}\end{cases}$$

The superscript R refers to the imposed radiative equilibrium potential temperature field;

$$\delta = \begin{cases} 0 & \text{for the QG equations} \\ 1 & \text{for the PE.}\end{cases}$$

As in A only 12 wave number vectors are retained, $\{\pm(0,2), \pm(0,4), \pm(-4,2), \pm(4,2), \pm(4,0), \pm(8,0)\}$, and the coefficients k_0, k_1, h_0, and h_1 which appear in the governing equations because of the parameterizations of the viscous and diabatic processes are given by

$$2k_0 = k_1 = 0.016$$

$$h_0 = h_1 = 0.018.$$

In contrast to A, the heating function imposed in the lower layer

$$\theta^R = \theta^*(2\cos 2y + 0.5\cos 4x)$$

is not zonally symmetric. We choose $\theta^* = 0.008$ (= 25 K dimensionally). Except for the departure from zonal symmetry in this heating function, the PE and QG models are dynamically identical to those used in the low forcing experiments reported in A. Since $\zeta_{-\mathbf{I}}$ is a diagnostic variable if $\zeta_{\mathbf{I}}$ is a prognostic variable, it is possible to define the model state vector $\mathbf{X}$ in terms of variables associated with only half of the retained wave number vectors. Also, in the QG model, τ and χ may be determined diagnostically. Therefore we are able to define $\mathbf{X}$ for the QG model as a vector with only 26 entries,

$$\mathbf{X} = (\sigma_0, \theta_0, \mathrm{Re}\,\psi_{(0,2)}, \mathrm{Im}\,\psi_{(0,2)}, \mathrm{Re}\,\psi_{(0,4)}, \mathrm{Im}\,\psi_{(0,4)}, \ldots \mathrm{Im}\,\psi_{(8,0)}, \mathrm{Re}\,\theta_{(0,2)}, \mathrm{Im}\,\theta_{(0,2)}, \ldots \mathrm{Im}\,\theta_{(8,0)})^T$$

4. Experiments

4.1. *Nature and observations*

By assuming that one model is nature and by using a different model to forecast nature we have avoided the unrealistic assumption of a perfect model which is made in so-called "identical twin" forecast experiments. The nature run generated by

[2] In A, θ_0 was defined to be the horizontally averaged potential temperature at the interface and the horizontally averaged radiative equilibrium potential temperature in the lower layer was denoted by $(\theta_0^R - \sigma_0^R)$. Because there is no heating of the upper layer, $(\theta_0^R - \sigma_0^R)$ has no dynamical significance and may be either arbitrarily chosen or eliminated from the governing equations as we have done here.

the PE model is 35 years long. The initial state for this run is a small perturbation superposed on the zonal steady-state solution of the PE forced by the zonally averaged heating function, $\theta^R = \theta^*$ 2 cos 2y. A time step, Δt, equal to f^{-1} was used in an alternating 4-cycle scheme (Lorenz, 1971).

Fifty cases were extracted from the nature run at intervals of 250 days; the forecast period for the first case starts at day 250, well after transients associated with the initial conditions have disappeared. The observations are separated by $\tau = 6$ h and the QG model state $\mathbf{X}$ is observed at each observing time. The observations are the true state of nature contaminated by a random error. The error $\boldsymbol{\varepsilon}$ at each observing time is independent of the error at all other observing times. The error ε_j for each variable X_j is independent of ε_k for all $k \neq j$; each ε_j has a Gaussian distribution with zero mean and a standard deviation approximately proportional to the natural variability, i.e., to the standard deviation of X_j. The constant of proportionality γ is 0.075 and the standard deviation of $\boldsymbol{\varepsilon}$ multiplied by 10^6 is (25, 10, 155, 195, 35, 35, 180, 180, 180, 180, 180, 180, 55, 55, 80, 80, 60, 60, 80, 80, 80, 80, 80, 80, 80, 80)^T.

The QG forecasts use the observation, $\mathbf{X}^o$, as initial conditions. This is sufficient for our purposes but there are other possible initial conditions. For example, we might have followed Leith's (1980) approach to the initialization of quasi-rotational dynamics in which the data is used to define a nonlinearly balanced PE state which is then projected onto the rotational manifold.

4.2. *The dynamical forecasts*

All dynamical forecasts were made using the QG model with a time step Δt of $2f^{-1} = 6$ *h*. For the MCF experiments, the $\mathbf{Z}_i$ have the same statistics as the $\boldsymbol{\varepsilon}$. For the LAF experiments, observations are separated in time by $\tau = 6$ h. In both MCF and LAF experiments $N = 8$; an ensemble of this size is suggested by some of Leith's (1974a) results. We conclude that $N = 8$ is a large enough ensemble since the results presented agree very well with results obtained for the same 50 cases for $N = 4$ and $\tau = 12$ h.

4.3. *The tempered forecasts*

The general optimal filtering technique involves estimating a large number of parameters and thus requires a very large homogeneous data sample (Lorenz, 1977). It is therefore an advantage to introduce simplifying assumptions which limit the number of independent parameters.

We will assume that only forecasts of a given variable may enter the regression equation of that variable. Then, with no loss of generality we may consider a single variable, x, measured as a deviation from its climate mean at a single time level. In place of (6), the tempered forecast is now given by

$$\hat{x} = \sum a_i \bar{x}_i \tag{8}$$

where $\bar{x}_i$ is the ith dynamical forecast of x and the a_i are the tempering weights to be determined. The a_i depend on the forecast time, the forecast method and the particular model variable.

If the model and natural climatology are the same, i.e., if $\langle \bar{x}_i \rangle = 0$ and $\langle \bar{x}_i^2 \rangle = \langle x^2 \rangle$, where x is the true state of nature and where the climate ensemble average indicated by angle brackets is for a fixed forecast time, then it may be shown that the a_i which minimize the climate ensemble average squared forecast error, $\langle (\hat{x} - x)^2 \rangle$, verify

$$\sum_{i=1}^{N} (1 - s_{ij}^2)\, a_i = 1 - r_j^2 \tag{9}$$

where

$$r_j^2 = \langle (x - \bar{x}_j)^2 \rangle (2 \langle x^2 \rangle)^{-1}$$

is the normalized forecast error variance of the jth forecast and

$$s_{ij}^2 = \langle (\bar{x}_i - \bar{x}_j)^2 \rangle (2 \langle x^2 \rangle)^{-1}$$

is the normalized variance of the difference between the ith and jth forecast. In theory, the climate ensemble contains all the normally occurring states of nature each weighted by its frequency of occurrence; in practice, the climate ensemble will be replaced by our sample of 50 cases. In our experiments the model and natural climatology are nearly identical. When the model and natural climatology are not identical (9) still holds if terms of order $(\langle \bar{x}_i^2 \rangle - \langle x^2 \rangle)(2 \langle x^2 \rangle)^{-1}$ may be neglected. As forecast time increases both s_{ij}^2 and r_j^2 are expected to increase monotonically and approach unity. (Here and below we only discuss s_{ij} for $i \neq j$ because s_{ii} is identically zero.) ds_{ij}^2/dt is proportional to the amplification rate of the squared forecast error due *only* to the internal dynamics of the model, while dr_j^2/dt is proportional to the amplification rate of the squared forecast error due to both the internal dynamics of the model *and* to the external error caused by the difference between the model dynamics and the natural dynamics. That is, s_{ij} grows because the model itself is not predictable while r_j grows for this reason and because the model is not perfect.

In order to determine the a_i we must estimate s_{ij} and r_j from our sample at each time level and for each model variable. In our experiments, we observed that s_{ij} and r_j grow nearly linearly in time until their maximum value of unity is reached. For example, the curve labeled ODF in Fig. 4a is the r.m.s. of twice the r_1^2 of all the model variables for the ODF. Also it is observed that s_{ij} and r_j for most of the variables have the same growth rate. We will assume that these observations hold exactly for each variable. Then all the a_i are determined once we estimate two parameters—$\dot{s}$ and $\dot{r}$, the growth rates of s_{ij} and r_j, respectively. After fitting the data from the dynamical experiments in several ways, we subjectively estimated these parameters to be 0.03 day^{-1} and 0.04 day^{-1}.

To complete the specification of s_{ij} and r_j their initial values must be determined. These initial values may be obtained by considering the nature of $\boldsymbol{\varepsilon}$ and of $\mathbf{Z}_j$. For the ODF and for the members of the LAF ensemble we have specified that $2r_j^2(t_j) = \gamma^2$ at the observing time t_j. For the MCF $r_j^2(0) = \gamma^2$ since $\boldsymbol{\varepsilon}$ and $\mathbf{Z}_j$ are uncorrelated. Similarly since $\mathbf{Z}_i$ and $\mathbf{Z}_j$ are uncorrelated, $s_{ij}^2(0) = \gamma^2$ for the MCF. To determine the initial s_{ij}^2 for the LAF, suppose $t_i < t_j \leq 0$ and consider $s_{ij}^2(t_j)$. Since $\boldsymbol{\varepsilon}$ at t_j is uncorrelated with the forecast $\hat{\mathbf{X}}_i$

$$s_{ij}^2(t_j) = r_i^2(t_j) + r_j^2(t_j)$$
$$= [2^{-1/2}\,\gamma + (i-j)\,\tau\dot{r}]^2 + 2^{-1}\,\gamma^2.$$

The a_i determined following the above procedure are the same for each model variable and are plotted as functions of forecast time in Fig. 3. Since the members of the Monte Carlo ensemble are not distinguishable all the a_i for the MCF are the same. Note that these weights are different from the weights which would be obtained by individually tempering each ensemble member using the tODF weights and then ensemble averaging. This is apparent in the figure; the tMCF weights are larger than N^{-1} times the tODF weights since the MCF is already hedged toward the climate mean by the ensemble averaging procedure. Because of correlations between the forecasts, some of the tLAF weights are negative. The average of the tLAF weights is not shown in Fig. 3; it would be indistinguishable from the tMCF weights. At long forecast times all the a_i are zero indicating that the best forecast is the climate mean.

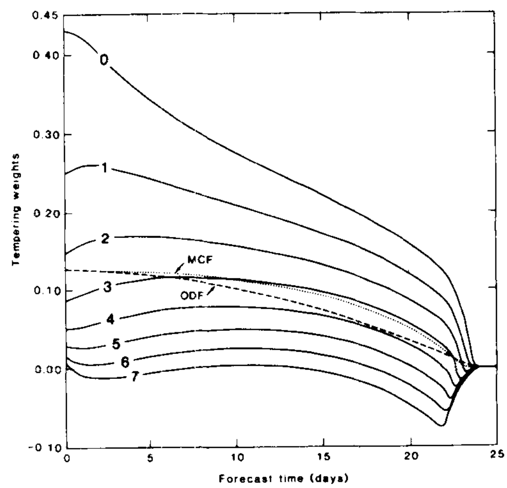

Fig. 3. The weights, a_i, for tempering the ODF (— — —), the MCF (·········) and the LAF (————). The a for the ODF are normalized by $N = 8$; the LAF a_i curves are labeled by $i - 1$, i.e., by the number of lagged intervals τ. The average of the LAF weights is very close to the MCF weights.

Since only two parameters have been estimated from a relatively large sample we expect that the tempering weights calculated here are not significantly affected by the size of the sample and that we may apply these a_i to this same sample to temper the forecasts. To support this last assumption, we repeated the ODF and LAF experiments for an independent sample using the tempering weights determined here. The same qualitative effect on the forecast skills is observed for both samples.

Actually during the first few days of the forecast our estimate of the growth rate of r_j could be improved. At the initial time the error is distributed among the exponentially growing and decaying modes of the linear equation governing the evolution of small errors. Thus the total error may actually decrease at first. Since we observe that the growth rate of r_j is approximately zero initially and then smoothly approaches $\dot{r}$, a good representation of $\mathrm{d}r_j/\mathrm{d}t$ would be of the form $\dot{r}(1 - \mathrm{e}^{-t/\beta})$ where β

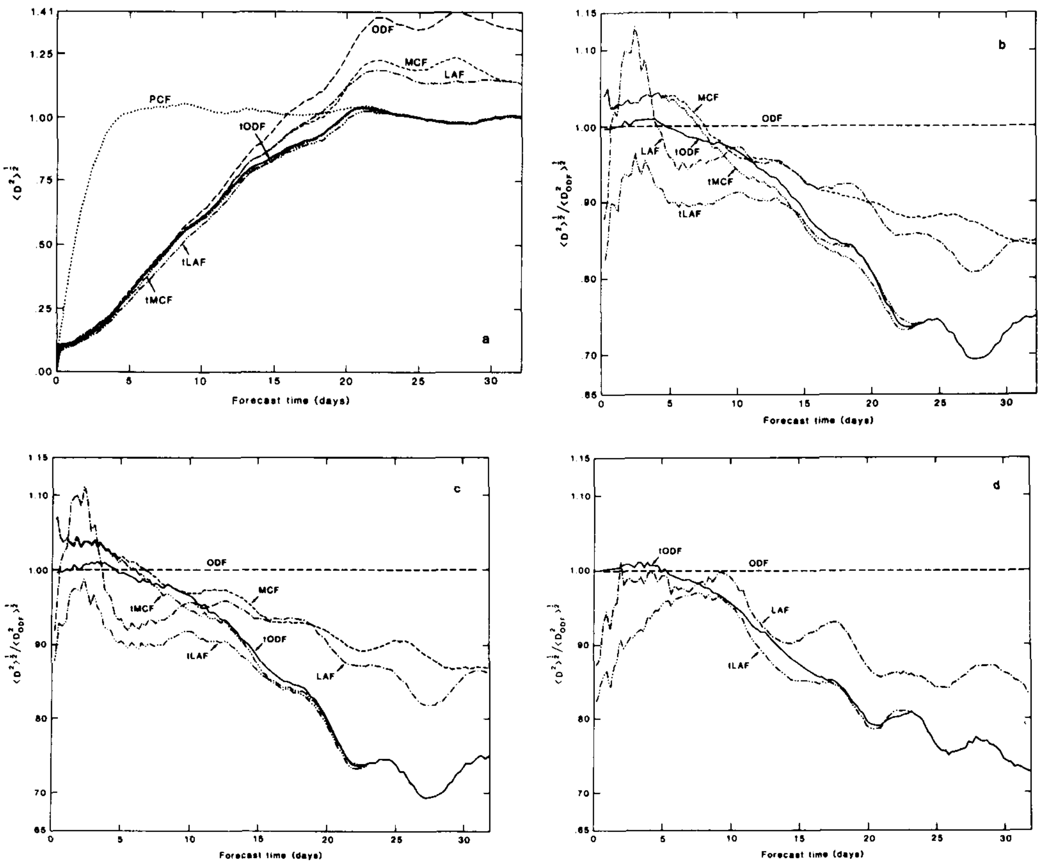

Fig. 4. Comparison of the skills of the different forecast methods. In (a) $\langle D^2 \rangle^{1/2}$ and in (b) $\langle D^2 \rangle^{1/2} / \langle D^2_{\text{ODF}} \rangle^{1/2}$ are plotted as functions of forecast time, t, for the ODF (———), MCF (------), LAF (—·—·—), PCF (·········), tODF (———), tMCF (—···—···), and tLAF (—··—··—). D is a skill score which is also the distance in the model phase space from forecast to observation. Results for a smaller ensemble ($N = 4$, $\tau = 12$ h) (c) and an independent sample (d) are shown in the same format as (b).

is O(2 days). This discussion also holds for the evolution of s_{ij}. If the predictions of an operational model are to be tempered, a more careful study of the covariances is warranted; especially one should look for the effect just described at short forecast time.

4.4. *The persistence climatology forecast (PCF)*

To calculate the PCF, the diagonal matrix, $\mathbf{q}$, must be specified in eq. (7). Ordinarily one might choose $q_j = \phi_{1j}$ where ϕ_{mj} is the lag m autocorrelation coefficient of X_j calculated directly from the nature run. However, the assumption of a first-order autoregressive (AR(1)) process is not really appropriate; the observed ϕ_{mj} decay more rapidly than the autocorrelations of a true AR(1) process for which $\phi_{mj} = (\phi_{1j})^m$. Thus the choice $q_j = \phi_{1j}$ would yield a PCF which decayed to the climate mean too slowly. There are two reasonable alternative ways of choosing q_j. The first alternative is to determine q_j by fitting the observed autocorrelations, e.g., by minimizing $\sum (\ln \phi_{mj} - m \ln q_j)^2$, where the sum is over m. The second alternative is to choose q_j to minimize the squared forecast error of the PCF, that is, to minimize the sum over all initial conditions of $\sum (X_j^o(m\tau) - X_j^p(m\tau))^2$. A necessary condition for a minimum in this case, if X^o is replaced by the true state of nature, is that q_j is a root of the polynomial $\sum mq^{2m-1} - \sum m\phi_{mj} q^{m-1}$. Since the correct root must be real and close to 1, it is easily found by Newton's method. We used this second alternative since we are more interested in the skill of the PCF than in fitting the autocorrelations. Both alternatives involved some arbitrariness since the sums over m must be assigned an upper limit and might reasonably be weighted by some function of m. We have not included a weighting function and chose the upper limit as either 25 or the minimum value of m for which $\phi_{m+1,j} < 0.25$, whichever is smaller. This choice limits the lags considered to lags for which PCF is expected to be considerably more skillful than a simple climate mean forecast. As expected the values of q_j obtained are substantially less than the corresponding ϕ_{1j}. The most extreme case is for σ_0, i.e. for $j = 1$: $q_1 = 0.9079$ while $\phi_{11} = 0.9837$.

5. Comparison of forecast skills

To compare the forecast skills of the different methods we have considered an integral measure of skill,

$$\Delta^2 = (\delta\sigma_0)^2 + (\delta\theta_0)^2 + 2^{-1}\overline{(\delta\theta)^2} + 2^{-1}\overline{(\delta\psi)^2}$$

In this expression $\delta\zeta$ is the perceived forecast error of ζ, and an overbar is an average over the horizontal domain. For skillful forecasts Δ will have relatively small values. As before, σ_0 is the horizontally averaged static stability, θ_0 is the horizontally averaged potential temperature measured as a deviation from a reference temperature, θ is the potential temperature measured as a deviation from the horizontally averaged potential temperature and ψ is the stream function—all defined at the interface between the two layers which may be considered to be at 500 mb. This definition of an error metric is a natural choice because in terms of the phase space of the QG model, Δ is the Euclidean distance from verification, $\mathbf{X}^v$, to forecast, $\hat{\mathbf{X}}$,

$$\Delta^2 = (\mathbf{X}^v - \hat{\mathbf{X}})^T(\mathbf{X}^v - \hat{\mathbf{X}})$$

The verification will normally be the observations, X^o. A normalized measure of skill is D, defined by

$$D^2 = \Delta^2\,\Delta_c^{-2}$$

where Δ_c^2 is the expected value of Δ^2 when the forecast is the climate mean and when $\mathbf{X}^v$ is the true state; Δ_c^2 is the sum of the climate variances of the model variables. The estimate, $\Delta_c = 7742 \times 10^{-6}$, obtained from the nature run, is used throughout this study.

The forecast skills of the different forecast methods are compared in Fig. 4. In Fig. 4a $\langle D^2\rangle^{1/2}$ is plotted as a function of forecast time, while in Fig. 4b $\langle D^2\rangle^{1/2}$ for each forecast method, normalized by $\langle D^2\rangle^{1/2}$ for the ODF, is plotted as a function of forecast time. Since the climate ensemble averages are calculated here as averages over the sample of 50 cases, each point plotted in Fig. 4b is the square root of the ratio of two sums of squares and is similar to an F statistic. This property could be used to test whether any one of the methods yields a significant improvement in forecast skill relative to that of the ODF. Unfortunately, it is not certain how many degrees of freedom to associate with each sum of squares. As a guide to the variability of $\langle D^2\rangle^{1/2}$, there is no reason to expect any difference between the skills of the ODF at different forecast times past the time ($t \sim 22$ days) when $\langle D^2\rangle^{1/2}$ or the ODF comes close to its asymptotic value of $2^{1/2}$. The oscillations after this time suggest that the sampling error of $\langle D^2\rangle^{1/2}$ is about 5% of the estimated value of $\langle D^2\rangle^{1/2}$ at any particular forecast time. Examination of $\langle D^2\rangle^{1/2}$ for an independent sample of ODFs supports this conclusion.

As can be seen in Fig. 4a–b the skill of the tLAF is superior to that of the other forecast methods at short range ($t \lesssim 1$ week) and medium range (1 week $\lesssim t \lesssim 2$ weeks). At short range all the other dynamical forecast methods are roughly equivalent in terms of forecast skill. At medium range and long range ($t \gtrsim 2$ weeks) all the ensemble and/or tempered forecast methods are more skillful than the ODF. At all forecast times the MCF and LAF have roughly the same skill. The skills of the tODF, tMCF, and tLAF are indistinguishable at long range. Similar results are obtained for the same 50 cases using a smaller sample size ($N = 4$) and a larger time increment ($\tau = 12$ h) (Fig. 4c). Qualitatively similar results are obtained for an independent sample of 50 cases for the ODF and LAF experiments with $N = 8$ and $\tau = 6$ h (Fig. 4d). In these experiments, no new parameters were estimated; the values of $\dot{r}$ and $\dot{s}$ estimated previously were used to calculate the tempering weights.

Since we expect a climate mean forecast to yield $\langle D^2\rangle^{1/2} = 1$ for any forecast time, the time when

$\langle D^2 \rangle^{1/2}$, for a particular forecast method, reaches 1 may be considered to be the predictability limit of that forecast method. According to this definition the predictability limit of the ODF is 15.5 days, ensemble averaging with $N = 8$ increases this limit to 18.5 days and tempering increases the predictability limit to 21.0 days. For the PCF $\langle D^2 \rangle^{1/2}$ nearly reaches its asymptotic value of unity by 5 days. It may appear that these predictability limits are too large. They *are* large compared to the predictability limit of the atmosphere but this is expected because of the extreme truncation used here. Only the largest scales are retained in our experiments while in the atmosphere it is probably the largest scales which have the longest predictability limits (Shukla, 1981). Furthermore, the absence of smaller scales neglects an important source of the growth of error in the larger scales (Lorenz, 1969). There is also some arbitrariness in our choice of time scale, $f = (3\ \mathrm{h})^{-1}$.

The qualitative nature and much of the quantitative nature of the above results can be anticipated from simple arguments similar to those used to obtain the tempering weights. Excluding the PCF, each of the forecasts, $\bar{\mathbf{X}}$, may be considered to be a weighted sum of the climate mean and K dynamical forecasts.

$$(\bar{\mathbf{X}} - \langle \mathbf{X} \rangle) = \sum_{i=1}^{\bar{K}} c_i (\bar{\mathbf{X}}_i - \langle \mathbf{X} \rangle)$$

where $K = 1$ or N and $c_i = 1$, $1/N$ or a_i as appropriate. Based on the assumptions of Section 4.3, and assuming that $\mathbf{X}^v$ is the true state, the expected value of D^2 is

$$\langle D^2 \rangle = 1 - 2 \sum_{i=1}^{K} c_i (1 - r_i^2) + \sum_{i=1}^{K} \sum_{j=1}^{K} c_i c_j (1 - s_{ij}^2)$$

Values of $\langle D^2 \rangle$ calculated in this manner for $N = 8$, $\tau = 6$ h were used to produce Fig. 5 which is to be compared with Fig. 4. The qualitative agreement between these two figures is good and is an *a posteriori* justification of the assumptions made in deducing the a_i. Quantitatively, the ensemble forecast methods are somewhat less skillful than expected at short and medium range.

Leith (1982) has pointed out that any sort of averaging has a tempering effect. Since the tODF and tMCF have roughly the same skill, the advantage of the MCF over the ODF at medium and long range is probably simply a reflection of the fact that ensemble averaging hedges the ensemble forecast towards the climate mean. This suggests that, in these experiments, ensemble averaging itself does not accomplish anything that simple statistical regression could not accomplish. This raises two questions: first, why does ensemble forecasting fail to produce the improvement in forecast skill anticipated in the Introduction?; second, what causes the superiority of the tLAF? To address the latter question first, the superiority of the tLAF is probably due to a better specification of the initial state. The tODF cal-

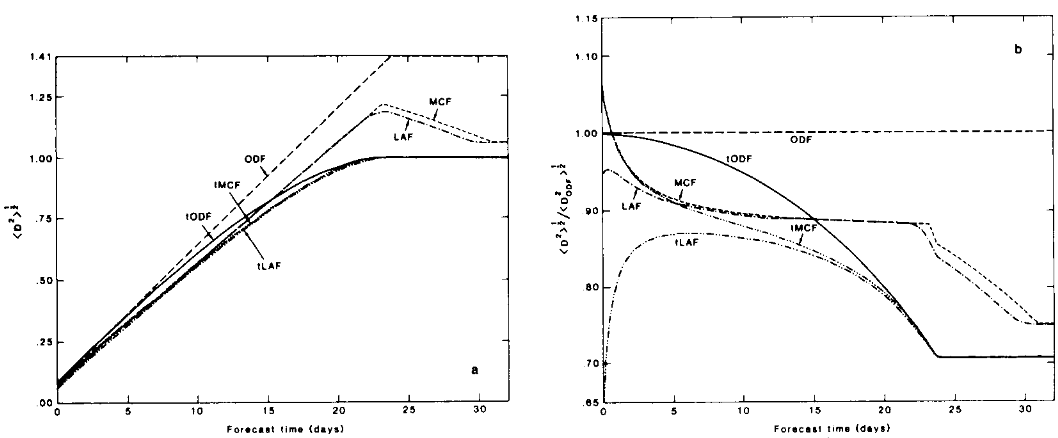

Fig. 5. Comparison of the expected skills of the different forecast methods. Same format as Fig. 4. Expected values of $\langle D^2 \rangle$ are calculated based on the assumptions discussed in the text.

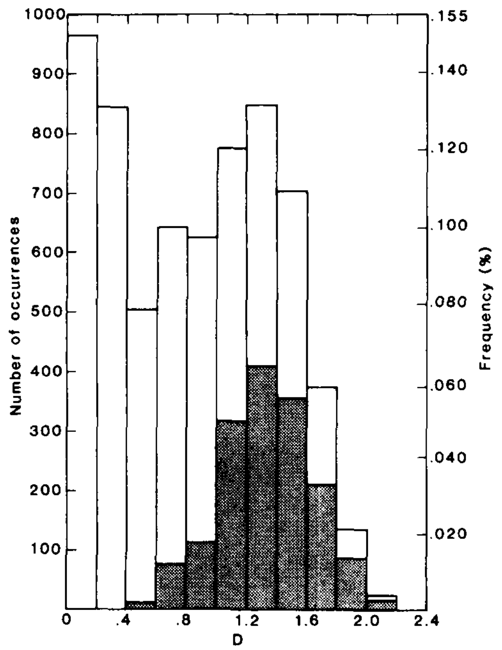

Fig. 6. Histogram plot of D for the ODF experiments. All data in the interval $0 \leqslant t \leqslant 32$ days was used. The contribution due to the time interval $24 < t \leqslant 32$ days is shaded. The total sample size is 6450.

culated from the tLAF initial conditions, i.e. using the ODF

$$\bar{\mathbf{X}}_1 = \mathbf{X}(t; 0, \hat{\mathbf{X}}_{\mathrm{LAF}}(0)),$$

has essentially the same skill as the tLAF. This forecast method is similar to that of Miyakoda and Talagrand (1971). This experiment also indicates that for short range the deviations from the ensemble mean are governed by linear dynamics. This observation answers, in part, the former question; at short range, in the linear regime ($t < \tau_{\mathrm{NL}}$), the ensemble forecast is identical with the central forecast and the ensemble forecast method has no advantage in terms of forecast skill. Furthermore, at long range all tempered forecasts are close to the climate mean and there cannot be much difference between them. Thus, when the forecast is good (roughly when $D \lesssim 1/3$) the deviations from the ensemble mean evolve linearly, and when the forecast is poor (roughly when $D \gtrsim 1$) tempering should eliminate the difference between the ODF and the ensemble forecasts. When the forecast has medium skill (roughly when $1/3 \lesssim D \lesssim 1$) it *is* possible that ensemble averaging will yield a significant improvement in forecast skill. However, the occurrence of medium skill is

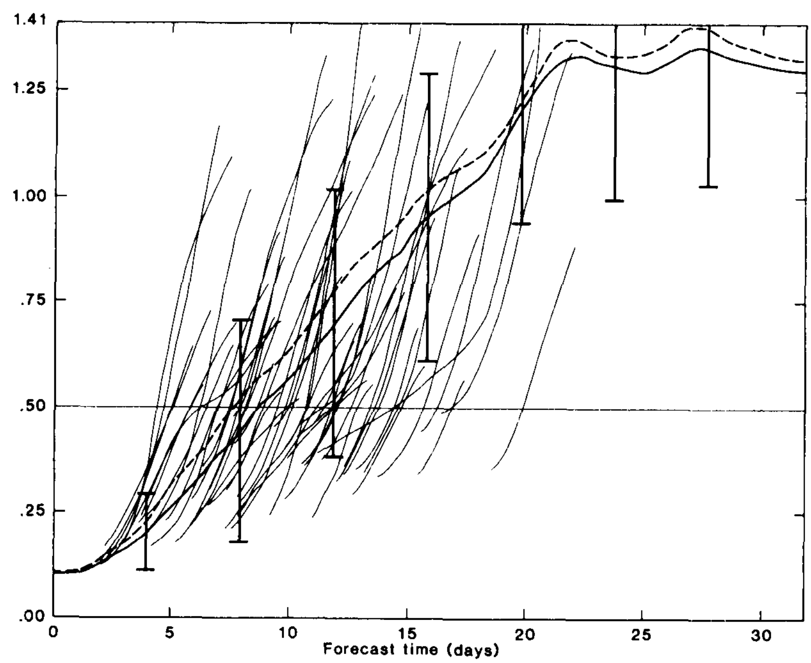

Fig. 7. Time evolution of D for the ODF experiments. Shown are selected portions of the evolution of D for each case (light ———). These curves have been smoothed to eliminate the roughness caused by the white noise observing error. Also shown are $\langle D^2 \rangle^{1/2}$ (— — —) and $\langle D \rangle$ (heavy ———). Error bars are calculated as $(\langle D^2 \rangle - \langle D \rangle^2)^{1/2}$.

relatively infrequent (Fig. 6). Actually, the time evolution of $\langle D^2 \rangle^{1/2}$ is not representative of the time evolution of D. In a typical time evolution of D, a period of very rapid growth follows an initial period of slow growth; that is, the skill of an individual forecast breaks down suddenly. In Fig. 7 we have drawn selected portions of the time evolution of D for each ODF. In each case the selected portion shows a period of explosive error growth through the skill level $D = 0.5$. Thus, in each case there is only a limited time interval during which the ensemble forecast might have a distinct advantage. Moreover, these time intervals occur at different forecast times—therefore the ensemble forecast can be advantageous, at any particular forecast time, in at most a small percentage of cases.

We remark, in closing this section, that Fig. 6 suggests an interesting approach to defining a standardized skill score. Note in Fig. 6 the apparently normal distribution of D for the forecast period 24 to 32 days, when the individual forecasts have little or no skill. This suggests that, given a particular forecast and its skill score, that is, some particular measure of its forecast error, we define a standarized skill as the probability that a larger forecast error would be obtained by a random forecast. By random forecast we mean an atmospheric state chosen at random from the climate ensemble. In the present case the probability that D, for a random forecast, will exceed a particular value could be estimated by fitting a normal p.d.f. to the shaded area of Fig. 6.

6. Prediction of forecast skill

The length of time for which a forecast is skillful is variable. Some situations are more difficult to forecast because the forecast is more sensitive than usual to the errors which are inevitably present in the specification of the initial state. This variability is clearly seen in Fig. 7; it was also pointed out by Lorenz (1965) in a study using a model similar to that used here (see his Fig. 2) and by Halem et al. (1982) in a study using a state of the art global NWP model (see their Fig. 11). As discussed in the Introduction, the spreading of the MCF or LAF ensemble should be related to the forecast error. Such a relationship would enable us to predict, *a priori*, forecast skill on a case by case basis since the spreading of the forecast ensemble may be calculated without knowledge of the future evolution of nature. In this section, given the spreading of the ensemble $\tilde{S}$, we will try to predict the forecast error $\tilde{D}$.

In practice there are several reasons why $\tilde{D}$ cannot be precisely predicted from $\tilde{S}$. Ideally the evolution of $\tilde{S}$ would be identical to the expected evolution of $\tilde{D}$. Unfortunately this ideal case is unattainable, entailing a perfect model, perfect initial conditions and an infinite ensemble. First consider the difference between forecast error and ensemble spread. Forecast error grows because of the natural intrinsic growth of error which is due to the uncertainty of the initial state and because of the effect of external error which is due to the uncertainty of the dynamics. The ensemble of forecasts spreads only because of the uncertainty of the initial state. Thus $\tilde{S}$ is generally smaller than $\tilde{D}$ because the growth of $\tilde{S}$ does not include the effect of external error. Moreover, since the dynamics governing nature are different from the model dynamics, the natural and model intrinsic growth of error are, in general, different. Second, even if the model were perfect, the intrinsic growth of error depends on the initial state, which is not known precisely. Finally, any evolution of $\tilde{S}$ calculated from a finite ensemble is only an approximation of the model's true intrinsic growth of error.

In this study $\tilde{D}$ is defined as the forecast ensemble r.m.s. of D,

$$N\tilde{D}^2 = \sum_{i=1}^{N} D_i^2$$

where D_i is the normalized distance in phase space, as defined in Section 5, between the ith forecast in the ensemble and the observed state of nature. The corresponding ensemble spread, $\tilde{S}$, is the r.m.s. normalized distance between all distinct pairs of forecasts in the ensemble and is defined by

$$N(N-1)\,\tilde{S}^2 = \sum_{i=1}^{N} \sum_{j=1}^{N} S_{ij}^2$$

where S_{ij} is the distance in phase space between the ith and jth forecast divided by Δ_c. Because a finite sample size affects estimates of second-order statistics more than estimates of means, we have considered only these global measures of forecast skill and ensemble spread. For the LAF we might have used weighted averages in defining $\tilde{D}$ and $\tilde{S}$. With the above definitions there is bound to be a close relationship between $\tilde{D}$ and $\tilde{S}$: since $\tilde{D}^2$ is a sum of squares it may be partitioned into a sum of squares due to the mean forecast, which is just D^2 for the LAF or MCF, and a sum of squares due to deviations about the mean forecast, which is exactly $2^{-1}(N-1)N^{-1}S^2$.

We found that $\tilde{D}$ and $\tilde{S}$ are indeed related but are far from identical and that to a certain extent the effect of model imperfections can be accounted for by a regression analysis. We tested a series of parameterizations of the form

$$\tilde{D} = \tilde{S} + bf_k t + et^{1/2} \tag{10}$$

where t is non-dimensional time, and f_k is the kth function specified in Table 2. The coefficient b was chosen to minimize the ensemble mean squared error $\langle e^2 \rangle$, i.e., b was determined by a linear regression analysis through the origin of $(\tilde{D} - \tilde{S})t^{-1/2}$ in terms of $f_k t^{1/2}$. Note in (10) that the regression residual is $et^{1/2}$; this scaling makes the size of the error, e, approximately independent of t. If we had not stabilized the variance of the error in this way, the estimates of b would be minimum variance estimates instead of unbiased minimum variance estimates (Neter and Wasserman, 1974, p. 131). We obtained less successful results for other predicted quantities (e.g., D_{ODF}) and for other forms relating $\tilde{D}$ to $\tilde{S}$. Eq. (10) includes a time dependence because the effect of external error increases with forecast time, at least during the initial part of the forecast. This dependence is taken to be linear because the climate mean and standard deviation of $(\tilde{D} - \tilde{S})$ as functions of forecast time are roughly proportional to forecast time.

As a basis for comparison we have included in Table 2 the null parameterization, $f_0 = 0$. The simplest parameterization, including only the linear time dependence, is $f_1 = 1$. Since the QG approximation is more or less appropriate depending upon the particular flow, we include in f_2 and f_3 the time averaged energy,

$$\bar{E}_x = t^{-1} \int_0^t E_x \, dt \tag{11}$$

where E_x is the horizontally averaged kinetic energy of the forecast ensemble averaged QG divergent wind. $\bar{E}_x$ is a proxy measure of the Rossby number averaged over the forecast period. Although both $\tilde{D}$ and $\tilde{S}$ have an expected value of $2^{1/2}$ for long forecast times, estimates of $\tilde{D}$ obtained from f_1 and f_2 increase with t indefinitely. In order to account for this asymptotic behavior, f_3 also includes the factor $(1 - 2^{-1/2}\tilde{S})^{1/2}$ which approaches zero as the ensemble spreads. The particular f_k we considered and (10) are somewhat arbitrary; however, the three phenomena discussed here are real and should be included in any parameterization of the effect of external error on the growth of forecast error: in general, the effect of external error increases with time initially, depends on the appropriateness of the model and eventually decays towards zero.

Useful predictions of forecast skill are obtained from (10), using f_3. For each case, we predicted the time for $\tilde{D}$ to reach a value of 0.5 and compared this to the time when $\tilde{D}$ actually reached this value. The results are presented in Fig. 8 for both the LAF and MCF. The prediction of the time for D to reach a value of 0.5 obtained using the LAF ensemble was more accurate than that obtained using the MCF ensemble, with correlations between the time predicted and the time observed of 0.79 and 0.68, respectively. This suggests that the p.d.f. of the initial state is better represented by the

Table 2. *Least squares fits of eq. (10) for the LAF and MCF ensembles. The data from all 50 cases, satisfying $\tilde{S} < 2$ in the interval $0 < t \leqslant 32$ days was used. The sample size is N_0. $\bar{E}_x$ is defined by eq. (11)*

		LAF ($N_0 = 6274$)		MCF ($N_0 = 6264$)	
k	f_k	b	$\langle e^2 \rangle \times 10^6$	b	$\langle e^2 \rangle \times 10^6$
0	0	—	1889	—	2947
1	1	716×10^{-6}	1073	982×10^{-6}	1413
2	$\bar{E}_x$	3358	808	4328	932
3	$\bar{E}_x(1 - 2^{-1/2}\tilde{S})^{1/2}$	5075	716	5822	800

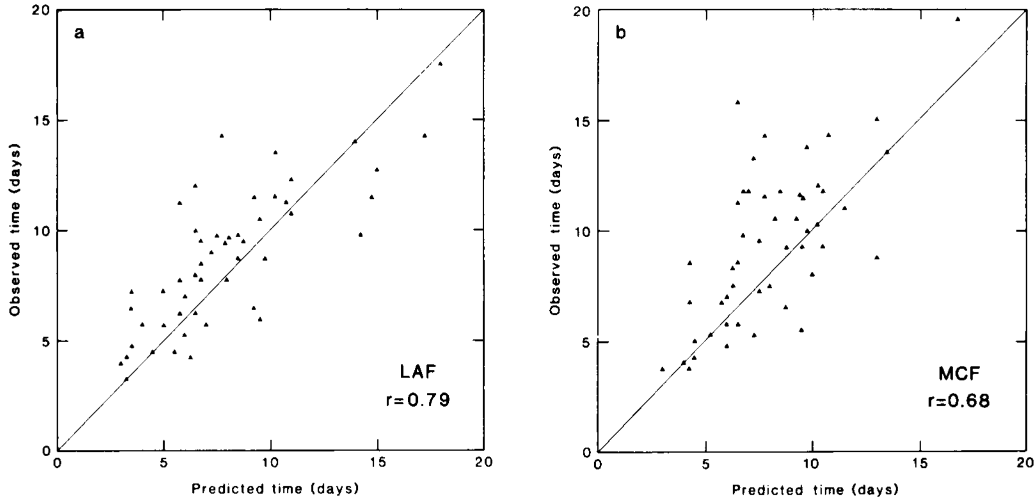

Fig. 8. Scatter plots of predicted versus observed time for $\tilde{D}$ to reach a value of 0.5 for the LAF (a) and the MCF (b). The predicted $\tilde{D}$ is obtained using eq. (10) with $k = 3$.

LAF ensemble than by the MCF ensemble. Using the same value of b for the same 50 cases, but taking $N = 4$ and $\tau = 12$ h, correlations of 0.67 and 0.66, were obtained for the LAF and MCF ensembles respectively. For an independent sample of LAF experiments, with $N = 8$ and $\tau = 6$ h, a correlation of 0.77 was obtained. The first of these observations is important for operational applications where forecasts are usually issued every 12 h.

Although we attempted only the simplest approaches to extract information from the spread of the ensemble for the purpose of predicting forecast error, we obtained the encouraging results just described. In this study our aim was only to show the forecast skill can usefully be predicted on a case by case basis using the spread of ensemble, not to obtain the best of such predictions. When this idea is applied to the forecasts of an operational model more care should be taken and several other approaches will probably suggest themselves. For example, each S_{ij} may be considered a separate piece of information about the intrinsic growth of error. Pooling all the S_{ij} into $\tilde{S}$ as we have done probably results in a loss of information.

The above results also suggest that the forecast of $\tilde{D}$ may be used to improve the tempered forecast; for example, when $\tilde{D}$ is predicted to be large the forecast is less reliable and more weight should be given to the climate mean. Forecasts tempered in this way might be called dynamically tempered forecasts because the tempering weights are chosen differently for each case, depending on the dispersion of the ensemble of forecasts. We obtained a dynamically tempered MCF by noting that the development of Section 4.3 holds if the angle brackets denote a climate ensemble associated with a particular value of $\tilde{S}$, a particular value of $\tilde{K}_x$ and a particular value of t instead of just a particular value of t. Since $a_i = a$, for the MCF we may average (9) over j and sum the result over all model variables to obtain

$$aN(1 - (1 - N^{-1})\, 2^{-1}\langle \tilde{S}^2 \rangle) = 1 - 2^{-1}\langle \tilde{D}^2 \rangle \qquad (12)$$

Now $\langle \tilde{S}^2 \rangle = \tilde{S}^2$ and we may estimate $\langle \tilde{D}^2 \rangle$ from (10). For a dynamically tempered LAF we used the a_i from Section 4.3 corresponding to the particular time level for which the average of the a_i is equal to the value of a obtained from (12). The dynamically tempered forecasts obtained in this way have approximately the same skills as the corresponding tempered forecasts. Apparently the estimates of $\langle \tilde{D}^2 \rangle$ obtained from (10) are not precise enough to be useful in this context.

7. Summary and concluding remarks

In order to use the information present in past observations and simultaneously to take advantage of the benefits of stochastic dynamic prediction we have formulated the lagged average forecast (LAF) method. The LAF is an average of an ensemble of forecasts, each at its proper verification time, started from the current and past observations, e.g., started at $t = 0, -12$ h, -24 h . . ..

To test the LAF method we have used a highly simplified atmospheric model. This model is a two-layer low order spectral model on a doubly periodic f-plane forced by asymmetric Newtonian heating of the lower layer. In our experiments there is a continuous external source of forecast error due to model imperfections since the forecast model is QG while "nature" is governed by the PE.

7.1. *Comparison of forecast skill*

On the basis of forecast skill, the LAF is slightly superior to the ODF, MCF and PCF (Fig. 4). The LAF and MCF are more skillful than the ODF at medium and long range ($t \gtrsim 1$ week) because as the ensemble spreads out the average of the ensemble approaches the climate mean. In this sense the ensemble forecasts are hedged. The ODF may also be hedged towards the climate mean by purely statistical techniques. The *tempered* ODF, LAF and MCF are the result of applying the simplest regression filter to the dynamical forecasts. The regression analysis naturally generates different weights for the members of the LAF ensemble (Fig. 3). At long range all three tempered forecasts have the same skill while at short and medium range ($t \lesssim$ 2 weeks) the tLAF is superior. In the present case, this superiority is probably due to a better specification of the initial state and not to the ensemble averaging procedure.

The results of the comparison of the forecast skills of the different forecast methods may be understood in terms of the phase space description introduced earlier. Recall that the ensemble forecast is better than the central forecast for $t > \tau_{NL}$, i.e., when nonlinear effects can no longer be neglected; the mean of the ensemble of forecasts is the estimate of the true state which is best in terms of minimum expected squared forecast error. To show the effects of tempering we must include in the phase space, descriptions of the climate manifold or attractor, represented schematically in Fig. 9 as the surface of a sphere, and the climate

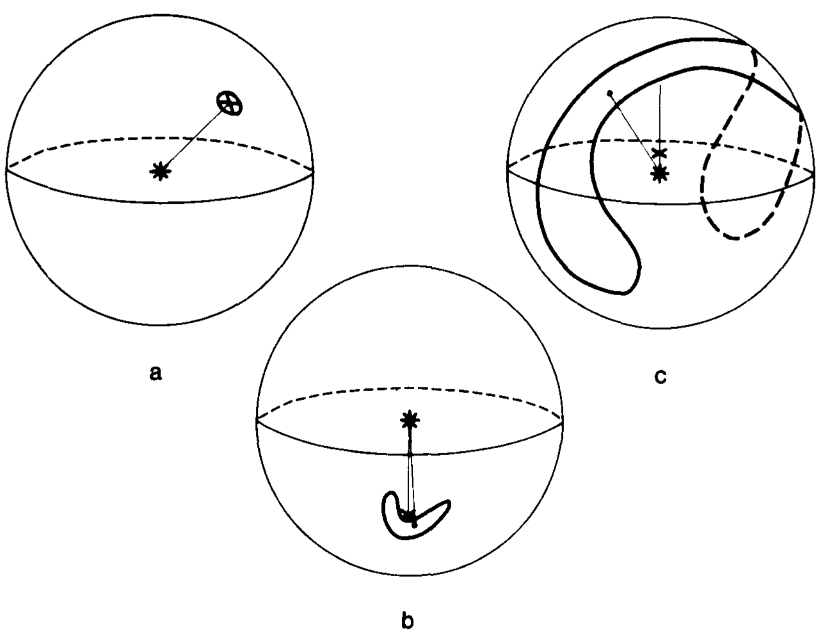

Fig. 9. Schematic phase space description of the central (•) and ensemble (×) forecasts, at the initial time $t = 0$ (a), at an intermediate time, $\tau_{NL} < t < \tau_{MIX}$ (b) and at a later time, $t \sim \tau_{MIX}$ (c). The dynamical forecasts always remain on the attractor (the surface of the sphere) while ensemble averaging and/or tempering hedge the forecast toward the climate mean (the * at the center of the sphere). For clarity, radii of the sphere are drawn through the forecasts.

mean, represented by the center of the sphere in Fig. 9. This schematic representation is a reminder that ensemble forecasts as well as statistically filtered forecasts are not necessarily typical or even attainable states of the system. For example, few, if any, synoptic charts are as zonal in character as are maps of climate normals. At the start of the forecast (Fig. 9a), the ensemble is represented by a small circle on the sphere and the mean of the ensemble and the initial conditions for the central forecast are coincident. In Fig. 9b, for $t \gtrsim \tau_{NL}$, the ensemble has spread and deformed but covers such a small part of the sphere that there is little advantage to hedging either the central forecast or the ensemble forecast towards the climate mean. For longer forecast times, as t approaches the mixing time of the system, τ_{MIX}, the ensemble spreads out over the entire attractor. In Fig. 9c for $t \sim \tau_{MIX}$ the ensemble forecast is close to the climate mean and its superiority over the central forecast may for the most part be eliminated by hedging the central forecast towards the climate mean. The possible advantage of ensemble forecasting is therefore confined to the interval between τ_{NL} and τ_{MIX}, that is, to situations similar to Fig. 9b. In the present experiments, ensemble forecasting does not have a clear superiority because τ_{NL} is apparently not much smaller than τ_{MIX}. Although the r.m.s. forecast error grows roughly linearly with time, we have seen that the typical evolution of forecast error for an individual case has a relatively short episode of explosive growth (Fig. 7). Furthermore, the timing of this forecast breakdown varies from case to case. Thus τ_{NL} and τ_{MIX} actually vary substantially from case to case but in all cases $\tau_{MIX} - \tau_{NL}$ is relatively small. It is quite possible that these results are model specific.

We caution the reader that we have only considered an r.m.s. measure of forecast skill in this study. We cannot rule out the possibility that different conclusions might have been reached if we had used some other measure of forecast skill such as a measure based on correlations. Certainly the tempered forecasts were designed with an r.m.s. measure of forecast skill in mind.

7.2. *Prediction of forecast skill*

Since the skill of a forecast model varies from case to case, an *a priori* prediction of forecast skill would be valuable. We have found that, on a case by case basis, $\tilde{S}$, the spreading of the ensemble, can be used to estimate $\tilde{D}$, the average forecast error of the forecasts making up the ensemble. Because the model is imperfect the difference between $\tilde{D}$ and $\tilde{S}$ increases with forecast time initially and is correlated with estimates of the inappropriateness of the modeling assumptions. When these factors are accounted for, good predictions of the time of forecast breakdown (i.e., the time when $\tilde{D}$ reaches 0.5) were obtained (Fig. 8).

The ability to forecast *a priori* the forecast skill of an individual forecast led us to attempt to use this information to temper the ensemble forecasts on a case by case basis. Essentially, the covariances needed to solve the regression equation were modeled as functions of the observed ensemble spread as well as of the forecast time. However, compared to the other tempered forecast methods, we were not able to improve the forecast skill by using this technique.

7.3. *Application to more realistic models*

Some of the implications of this study for realistic weather prediction models are immediate. We have shown that the advantages of ensemble forecasting can be obtained at virtually no cost by making simultaneous use of forecasts started from the latest as well as previous observed initial states. We found that most of the improvement of forecast skill due to ensemble forecasting can be obtained by hedging the latest forecast towards the climate mean. On the other hand, the spreading of the members of the ensemble can produce a useful *a priori* estimate of forecast skill on a case by case basis.

Since predictability varies with the scale of motion under consideration (Lorenz, 1969; Shukla, 1981) it might be useful to define the tempering weights, ensemble spread and forecast error as functions of wave number. This is not a consideration in the present study since all the wave number vectors retained have magnitudes in the range 2 to 8. On the other hand since predictability varies with place and season, tempering weights which depend on latitude, region and/or time of year should be considered. Similarly, measures of forecast skill which depend on latitude or region should be considered to define the ensemble spread and forecast error.

When an r.m.s. measure of forecast skill is a prime consideration, it is important that the forecast be hedged toward the *correct* climate mean. The most appropriate definition of earth's natural climate is given by Leith (1978b) in terms of a hypothetical ensemble of planets, each having geometry, composition and boundary conditions identical to those of the earth. Consider, for example, a formulation in which sea surface temperature, sea ice and snow cover are not predicted but are taken to be given boundary conditions. Then the natural climate which should be considered is the average over the hypothetical ensemble of earths having the same slowly varying boundary conditions. As an approximation, these boundary conditions may be taken to be fixed at the values observed at the start of the forecast. If the model correctly simulates the natural climate, then the ensemble forecast will be hedged towards the correct climate mean. If the boundary conditions are indeed almost constant during the forecast period this hedging will improve the forecast. In the present study this advantage of the ensemble methods had no effect because the boundary conditions were the same for each forecast. However if the model-simulated climate has systematic errors this advantage may become a liability. Even if the model cannot exactly simulate the earth's climate, predictions by the model of anomalies from the *model*'s climate mean may be good forecasts of the *actual* anomalies. Otherwise it may be better to use a statistical technique, such as the tempering procedure described here, to hedge the forecast towards the observed monthly or seasonal climate mean which is an average over many years and, inevitably, over different boundary conditions.

The LAF method may be useful in producing weekly, monthly or perhaps even seasonally averaged anomaly predictions. While there have been some research successes in the long-range forecasting of time-averaged quantities, operational forecasts have not been skillful (Nicholls, 1980). Two recent identical twin experiments (Seidman, 1981; Shukla, 1981) suggest that MCF of time-averaged quantities may be useful. In these experiments, the boundary conditions were fixed. In light of the discussion of the previous paragraph, ensemble averaging should have more impact when observed or ideally interactive boundary conditions forcing the atmosphere are used. If the LAF is applied to time-averaged quantities, tempering the forecast will still be useful; we expect that the tempering weights will decay much more slowly than those appropriate for single time quantities both as forecast time and the time lag of the initial conditions increase.

One aspect of the LAF method—the *a priori* prediction of forecast skill on a case by case basis—has been in qualitative use since the start of NWP. Traditionally, forecasters place more confidence in a numerical prognosis when it is similar to older forecasts for the same verification time. We hope that this study will provide the basis for reproducing this practice in an objective manner at operational centers.

8. Acknowledgments

This work was performed at the Modeling and Simulation Facility at the NASA Goddard Space Flight Center Laboratory for Atmospheric Sciences. Preliminary results of this work were presented at the Symposium on Current Problems of Weather Prediction, June, 1981, Vienna, Austria, and at the Seventh Conference on Probability and Statistics in Atmospheric Sciences, November 2–6, 1981, Monterey, CA., U.S.A. We are especially indebted to Milton Halem, Jim Edelmann and Michael Ghil for their help and encouragement. We thank Laura Rumburg and Jody Reckley for their excellent drafting and typing assistance. We also thank R. Livezey and two anonymous reviewers for their useful comments.

Ross Hoffman was a National Research Council/Resident Research Associate during the course of this study.

REFERENCES

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

Farago, T. 1978. Introduction of the characteristic function in stochastic dynamic modelling. *Tellus 30*, 335–340.

Fleming, R. J. 1971. On stochastic dynamic prediction. I. The energetics of uncertainty and the question of closure. *Mon. Weather Rev. 99*, 851–872.

Fleming, R. J. 1972. Predictability with and without the influence of random external forces. *J. Appl. Meteorol. 11*, 1155–1165.

Gleeson, T. A. 1970. Statistical-dynamical predictions. *J. Appl. Meteorol. 9*, 333–344.

Gleeson, T. A. 1981. A short procedure for statistical-dynamical predictions of weather probabilities. *Preprints Seventh Conference on Probability and*

Statistics in Atmospheric Sciences, November 2–6, 1981, Monterey, California. Boston, MA/U.S.A.: American Meteorological Society, 79–83.

Halem, M., Kalnay, E., Baker, W. E. and Atlas, R. 1982. An assessment of the FGGE satellite observing system during SOP-1. *Bull. Am. Meteorol. Soc. 63*, 407–426.

Hoffman, R. N. 1981. Alterations of the climate of a primitive equation model produced by filtering approximations and subsequent tuning and stochastic forcing. *J. Atmos. Sci. 38*, 514–530.

Laurmann, J. A. 1978. A small perturbation approximation for stochastic dynamic weather prediction. *Tellus 30*, 404–417.

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

Leith, C. E. 1974b. Spectral statistical-dynamical forecast experiments. *The GARP Programme on Numerical Experimentation, Report of the International Symposium on Spectral Methods in Numerical Weather Prediction, Copenhagen, 12–16 August, 1974.* (Report No. 7 of the GARP Working Group on Numerical Experimentation.) Geneva/ Switzerland: WMO, 445–467.

Leith, C. E. 1978a. Objeetive methods for weather prediction. *Annu. Rev. Fluid Mech. 10*, 107–128.

Leith, C. E. 1978b. Predictability of climate. *Nature 276*, 352–355.

Leith, C. E. 1980. Nonlinear normal mode initialization and quasigeostrophic theory. *J. Atmos. Sci. 37*, 958–968.

Leith, C. E. 1982. Statistical methods for the verification of long and short range forecasts. *European Centre for Medium Range Weather Forecasts, Seminar 1981, Problems and Prospects in Long and Medium Range Weather Forecasting, 14–18 September, 313–333.*

Lorenz, E. N. 1960. Energy and numerical weather prediction. *Tellus 12*, 364–373.

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

Lorenz, E. N. 1969. The predictability of a flow which possesses many scales of motion. *Tellus 21*, 289–307.

Lorenz, E. N. 1971. An N-cycle time-differencing scheme for stepwise numerical integration. *Mon. Weather Rev. 99*, 644–648.

Lorenz, E. N. 1977. An experiment in nonlinear statistical weather forecasting. *Mon. Weather Rev. 105*, 590–602.

Miyakoda, K. and Talagrand, O. 1971. The assimilation of past data in dynamical analysis. I. *Tellus 23*, 310–317.

Neter, J. and Wasserman, W. 1974. *Applied linear statistical models.* London/U.K.: Irwin-Dorsey, xvii + 842 pp.

Nicholls, N. 1980. Long-range weather forecasting: Value, status and prospects. *Rev. Geophys. Space Phys. 18*, 771–788.

Pitcher, E. J. 1977. Application of stochastic dynamic prediction to real data. *J. Atmos. Sci. 34*, 3–21.

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

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