Model error and ensemble forecasting: experiments with a simple model

D. Orrell

Centre for Nonlinear Dynamics, University College London, Gower Street, London, WC1E 6BT, UK

Camera-ready Copy for

Tellus

Manuscript-No. 0

Tellus (20**) *:101–115

Tellus
©European Geophysical Society 20**

Model error and ensemble forecasting: experiments with a simple model

D. Orrell

Centre for Nonlinear Dynamics, University College London, Gower Street, London, WC1E 6BT, UK

Received: ??? – Accepted: ???

Abstract. Error in weather forecasting is due to inaccuracy both in the models used, and in the estimate of the current atmospheric state at which the model is initiated. Because weather models are thought to be chaotic, and therefore sensitive to initial condition, techniques such as ensemble forecasting have been developed to address the latter effect. An ensemble of forecasts are made with perturbed initial conditions, the aim being to produce an estimate of the probability distribution function for the future state of the weather. Some ensemble schemes also include changes to the model, so as to account for the effects of model error. While the ensemble approach is quite widely adopted, however, its verification is complicated. Furthermore, recent results indicate that model error may be higher, and sensitivity to initial condition lower, than previously thought, so that model error is a dominant source of error over the first few days. It is therefore necessary to evaluate the effect of model error on ensemble forecasting. In this paper, we consider techniques to achieve this, based on the concept of shadow orbits. Two of the methods aim to establish the ability of the model to shadow (stay close to) the analysis, and therefore of the ensemble to represent the real weather. The third method tests whether the convex hull of a particular ensemble, as formed in the subspace spanned by the ensemble members, is near the analysis. The techniques are illustrated using a simple medium-dimensional system, which is tuned to simulate weather model errors. Comparisons with full weather models are also presented. It is shown that the presence of model error can severely limit the accuracy of ensemble schemes, while ensemble performance can also be used to deduce the presence of model error.

1 Introduction

Ensemble techniques have become established in recent years as a method for generating probabilistic weather forecasts. By running forward an array of slightly perturbed initial conditions, the ensemble forecast is intended to provide an approximation to the probability density function of the weather's future state (Ehrendorfer, 1997; Palmer, 2000). Techniques involving singular vectors (Molteni et al., 1996) or bred vectors (Toth and Kalnay, 1993) identify the fastest growing perturbations, so as to capture any rapidly growing modes. Ensemble shemes have proved to be essential tools in understanding the role of initial condition error.

The development of ensemble schemes was originally motivated by two ideas. The first was that the models were highly sensitive to initial condition: if the flapping of a butterfly's wings was enough to perturb the atmospheric flow and affect forecasts (Lorenz, 1963), then it followed that forecast accuracy was limited by the effects of chaos. This could be accounted for by running an ensemble of perturbed forecasts (Toth and Kalnay, 1993). The second idea, or working hypothesis, was that model error should be relatively small, at least for short forecast times: the 'perfect model' assumption (Buizza et al., 2000). Forecast error would therefore be dominated by the initial condition rather than the model (Toth et al., 1996).

Ensemble schemes have since evolved, and more recent schemes attempt to account, not only for perturbations in the initial condition, but changes in the model. One approach is the multi-model ensemble (Harrison et al., 1999), which includes forecasts using different models as well as initial conditions. Another method is to perturb the model itself, either by adding a stochastic error term to the model equations (Philips, 1986; Bennett and Budgell, 1987), or allowing changes to the model parameters or parameterisation schemes (Houtekamer et al., 1996; Buizza et al., 1997).

As ensemble useage has evolved, so have methods to verify their accuracy. However, standard techniques for ensemble verification, such as Brier scores, reliability diagrams, and Talagrand diagrams, are usually based on statistical approaches which, while providing useful diagnostics, do not specifically evaluate the role of model error. For example, it may be the case that an ensemble scheme does a good job of capturing the atmospheric variability, and yields promising statistical results, even though no ensemble member is close to the analysis after a couple of days (such an example will be considered in Section 6).

Our main concern here is that, while the ensemble approach was originally developed in a context of high sensitivity to initial condition and low model error, recent results indicate that model error is a major source of error over the first few days (Orrell et al., 2001). Furthermore, error doubling times due to sensitivity to initial condition appear relatively slow when measured in a global metric (Orrell, 2002). It is therefore necessary to evaluate the reliability of ensemble schemes when model error is significant, and determine whether models are sufficiently good that the ensemble approach is applicable. In this paper, our aim is to develop methods for gauging the effect of model error on ensemble forecasts, based on the concept of shadow orbits (Smith, 1996; Gilmour, 1998).

A model is said to shadow a specified target (in this case the analysis) for a shadow time τ if it remains within a specified radius r of the target over the shadow time. Shadow performance is primarily determined by model error. If a model does not shadow, then it is generally because model error is large; it may not therefore be appropriate to account for the error by perturbing the initial condition. Also, if no shadow exists, then no small change to the initial condition can result in a forecast near the analysis, or by implication the actual weather, so the ensemble can not strictly speaking be used to generate a probability distribution function for the atmosphere's future state. The existence of shadow orbits is therefore a test to determine whether the initial condition can be perturbed so that the predicted state is near the analysis, or whether the required correction is not in the initial condition, but in some aspect of the model itself (model perturbations will also be discussed below).

Three points: firstly, by probability distribution function, we mean for a particular forecast over different realisations of observation error (Ehrendorfer, 1997), as opposed to a probability distribution function over a large number of forecasts. If a model consistently predicts too high a value of, say, temperature during the winter, and too low a value during the summer, then it may appear to give reasonable results when averaged over the course of a year, but not for forecasts on particular days. This will be discussed further when we look at statistical methods for ensemble verification.

Secondly, shadow performance is metric-dependent, so for example a model may be able to shadow a local variable such as temperature in a specific location, even though it fails to shadow in a more general metric. An ensemble will therefore generate a spread of temperatures, and the correct answer can be expected to lie within that range. One could say that the ensemble approach has succeeded. Our point, though, is that, if a model does not shadow in a global metric, then that is because of model error, not the initial condition. Therefore, while perturbing the initial condition will result in a certain spread (as will any kind of perturbation), and stand a good chance of including the correct temperature, it will not address the underlying problem. An alternative method to obtain a similar result might be to simply add error bars to the predicted temperature, where the size of the error bars is determined from error statistics.

Finally, the purpose of this paper is of course not to imply that ensemble forecasters are unaware of the importance of model error. The history of weather forecasting is one of incremental improvements to the models and the data. However, it is also a historical fact that ensembles were designed to counter initial condition error, and only later adapted to account for model error. Since the two types of error are different in essence, it needs to be studied how ensembles perform if model error is large. Also, while model error is constantly being reduced by the designers of weather models, there is also a debate about whether resources should be allocated to improving the initial condition, running more ensemble members, and so on, or using/developing a better model. An improved understanding of model error is essential to this debate.

In this paper, three different methods for investigating the effects of model error on ensemble forecasts are discussed. The first two address the question of whether shadow orbits can, in principle, exist. The third is a simple test to determine whether a particular ensemble contains a shadow point after a certain time. The techniques are illustrated using a simple medium-dimensional model/system pair, which is tuned to simulate typical weather model error growth, and the results compared with weather model results. In conclusion, we consider strategies to account for model error, including the use of stochastic model error terms, and ask how good a model has to be for the ensemble approach to work efficiently.

2 The two-level system

To illustrate the validation of ensemble techniques, we will use a version of the Lorenz '96 system (Lorenz, 1996), that is designed to simulate a number of properties of weather model behaviour. The *two-level* system, which was introduced in (Orrell, 2002) to study the causes of forecast error growth, consists of 8 large-scale variables x_i and 32 coupled small-scale variables $y_{i,j}$, which can be viewed as atmospheric variables around a circle. The equations, which are given in the Appendix, simulate properties such as advection, damping, and forcing. Model error is provided by stochastic forcing terms which are present in the system, but absent in the model. In addition, the random component of analysis errors is simulated by adding a random noise component to each observation of the system. The magnitude of the noise term is set to $1.0m$ in the x variables, and $0.5ms^{-1}$ in the y variables. (In reality, neither model errors or analysis errors will be purely stochastic. Also, estimates of analysis error magnitude will be affected by model error, due to use of the model in the analysis procedure. The aim here is not to account for all these effects, but only to produce a reasonably plausible model/system pair which will illustrate the effect of model error on ensembles.)

We first re-cap some of the properties of the system. Because the x variables are large-scale and slow-varying, while the y variables are small-scale and fast-varying, the former resemble variables such as 500 hPa height, while the latter resemble more energetic variables like wind and temperature. By suitable choice of the scaling parameters, the errors can be brought to match those of a GCM (global circulation model, in this case the ECMWF operational model). The root-mean-square error growth is shown in Figure 1: the forecast errors of the large-scale x (solid line, top panel) agree reasonably well with GCM 500 hPa errors (+ symbol), while the small-scale y errors (bottom panel) are similar to GCM total energy errors. The two-level system also matches the GCM in terms of model error, as measured by the drift, and sensitivity to initial condition, as measured by lagged forecasts in a global metric (Orrell, 2002).

As discussed in the above reference, the drift can be computed from a sum of short forecast errors. For example, suppose the target point at time $t_j = t_o + j\Delta$ is $\tilde{\mathbf{x}}(t_j)$, and let $\mathbf{x_j}(t)$ for $t \geq t_j$ be the model trajectory initiated at the target point $\tilde{\mathbf{x}}(t_j)$. The drift at time t_K is then given by:

$$d(t_k) = \| \sum_{j=0}^{K-1} (\mathbf{x_j}(t_{j+1}) - \tilde{\mathbf{x}}(t_{j+1}))\| \qquad (1)$$

(the timestep Δ should be chosen sufficiently small that the calculation converges). Because the stochastic model error terms in the two-level system are uncorrelated, the drift grows in a square-root fashion like a random walk.

The size of the model error terms was determined by fitting the drift curves to those of the GCM. Figure 1 also shows for comparison the effect when the stochastic model error terms are reduced by a factor 10. When model error is high, the observation error has little effect on the calculations, but for low model error the observation error has a more significant effect. Two cases are therefore shown: in case 1, the observation error is at the normal value, while in case 2 it is tripled. In either case the rate of growth is significantly below that of the GCM.

Since the two-level model manages to approximate the basic properties of GCM error growth, it is reasonable to suppose that it should capture the essence of weather model ensemble behaviour. The upper panels of Figure 2 show errors for a 500-member ensemble generated by random perturbations. The left panel can be compared with any 500 hPa ensemble. The errors in the y variables (right panel) are similar but have a smaller spread.

The lower panels show Talagrand diagrams for the x and y variables for a 32-member ensemble, formed from taking perturbations in the positive and negative directions of the 16 leading singular vectors in the y metric, with an optimisation time of 2 days. These diagrams, which are discussed for example in (Ehrendorfer, 1997), provide a statistical test of the ensemble by counting the distribution of the true system relative to the ensemble predictions. Suppose we wish to predict the i'th x-component of the true system. Since the ensemble contains 32 members, it will provide 32 values x_i^j where j denotes the ensemble member. These can be ordered

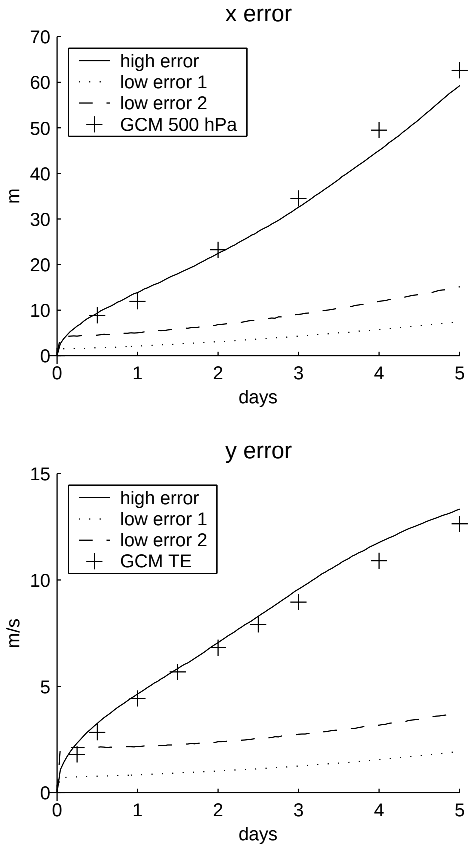

Fig. 1. Plot comparing root-mean-square errors for the two-level system with low and high model error. The observation error plays a larger role when model error is small than when model error is large, so two cases are shown. Upper panel shows error in x variables with high error (solid line), low model error and normal observation error (case 1, dotted line), and low model error with observation error increased by a factor three (case 2, dashed line). Results are compared with the GCM 500 hPa results (+ symbol). Lower panel compares the y errors for the different cases with GCM total energy.

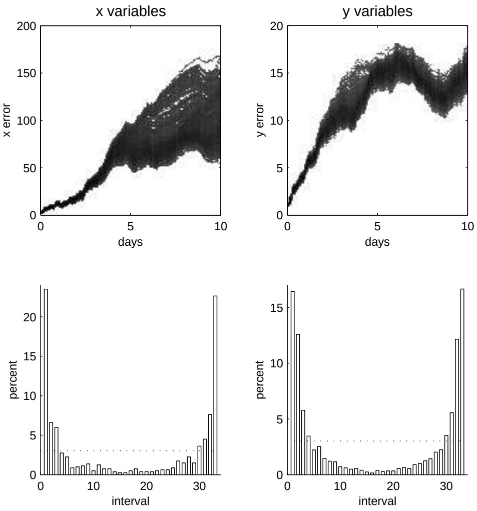

Fig. 2. Plot showing the ensemble performance of the two-level system. The top left panel shows errors in x variables, which can be compared to a 500 hPa ensemble for a weather model. The right panel shows the corresponding errors for the y variables. Lower panels show Talagrand diagrams of a 2-day forecast in x (left) and y (right) variables. Dotted line shows the ideal distribution. See text for discussion.

so that they form a partition into 33 bins, where the first bin corresponds to predictions smaller than x_i^1, the second bin corresponds to values between x_i^1 and x_i^2, and the last bin corresponds to values greater than x_i^{32}. We then note where the observed value of the true system lies, and repeat the experiment a number of times, for each index i. The result is a histogram of the position of each observation in the ensemble partition. If the ensemble gives a probability distribution function of the analysis, then the distribution should be flat, but here there is a distinct U-shape which indicates that the true values are often falling above or below the ensemble's range. The same effect, if to a slightly lower degree, is typically seen with GCM's (Strauss and Lanzinger, 1996).

Note that the Talagrand diagram only tests whether the ensemble gives a probability distribution function over many forecasts, which is a different question than whether it will do so over a single forecast, with different realisations of the observation error. In general, statistical verification schemes such as Talagrand diagrams provide a necessary condition for the ensemble to be an accurate representation of the atmosphere's future state; however, they are not a sufficient condition, since one could obtain a perfect Talagrand diagram by using an ensemble of randomly chosen climate states (Strauss and Lanzinger, 1996). For the same reason, while such diagrams are useful diagnostic tools in many respects, they are unsuited for assessing the effect of model error: as discussed in Section 6, a model can be quite bad, but yield a satisfactory Talagrand diagram, so long as it has been tuned to give a reasonable climatology.

A stronger necessary condition is that ensembles contain members where error at future times remain small, i.e. shadow orbits. As we will see below, the ability of a model to shadow is primarily a function of its model error. It is notable that none of the ensemble members in the upper panels shadow for long, since the errors of all members increase with time. For weather models, it is not possible to interpret ensemble diagrams so easily, because the ensemble is small relative to the dimension of the space, and the fact that ensemble errors increase with time does not necessarily imply that there do not exist other perturbations of equal or smaller magnitude that do shadow. In order to validate the ensemble approach, it is therefore necessary to find alternative methods for establishing the existence of shadow orbits. In the next sections, we consider some techniques for doing this.

3 Estimating shadow times

If ensembles are to be used to generate probability distribution functions of the weather at some future time τ, then at least one point, perturbed an amount smaller than or equal to the perturbation radius r of the ensemble, should manage to shadow the analysis within that radius r for time τ. In other words, if the target orbit (in this case the analysis) is $\tilde{\mathbf{s}}(t)$, the model trajectory is $\mathbf{s}(t)$, and the error vector $\mathbf{e}(t)$ is

$$\mathbf{e}(t) = \mathbf{s}(t) - \tilde{\mathbf{s}}(t), \qquad (2)$$

then we require an initial perturbation $\mathbf{e}(0)$ such that $||\mathbf{e}(t)|| \leq r$ for all $0 \leq t \leq \tau$. We will consider two independent methods to estimate a model's ability to shadow, which were also discussed in less detail in (Orrell et al., 2001). The first, which is direct but also expensive, will be to directly search for candidate shadow orbits. The second is based on the *shadow-drift law*, which relates shadow times to the drift. A third technique is a simple test to check whether a particular ensemble, created by either initial condition or model perturbations, can contain a shadow point after a period of time. Because model error is larger in the y variables, we will concentrate on shadowing in these variables: it is found that any orbit which shadows in y has only negligible errors in x (though the opposite is not true).

The most straightforward method of estimating shadow times is to search for orbits that shadow within a radius r. Ideally, with infinite computer resources, this would be done by testing all initial displacements $||\mathbf{e}(0)|| \leq r$ inside the shadow radius for one which remains within the shadow radius for the longest time. Failing that, an optimisation technique can be used to find the optimal initial condition; one method, dubbed the 'amoeba' method (Hansen, 1999), uses a simplex scheme (Press et al., 1993). The upper panel of Figure 3 shows a distribution of shadow times determined in this way for the two-level model with shadow radius $r = 2$.

While the amoeba technique is suitable for use with the two-level model, such a brute-force approach cannot be used with weather models, because of the high dimension of the space. Some more efficient scheme must therefore be applied. One approach might be to search in a limited subspace, for example the space of singular vectors, which would have the benefit of giving the maximum final displacement for the smallest initial displacement. A problem, though, is that model error is unlikely to be aligned with the singular vectors; indeed, because of the high dimension of the space, it is safe to assume that the two are orthogonal. Therefore the chances of finding an optimal shadow orbit in such a subspace is small.

However, if the adjoint is available, it can be used, not just to produce singular vectors, but to actually find the optimal initial displacement which will offset model error. Such a code has been used at ECMWF for so-called ‘sensitivity analysis’ (Rabier et al., 1996). Suppose that we set a shadow time τ, and wish to find an initial condition which stays within a minimal distance of the true system. (In effect, this is the inverse of the normal shadowing problem, since we set the time first and determine the radius, rather than vice-versa.) If we focus only on the final error, and ignore for the time being the intermediate points, then the problem can be phrased as

$$\text{minimise } C(\mathbf{e}(0)) = \frac{1}{2}\|\mathbf{e}(\tau)\|^2. \tag{3}$$

This optimisation problem has Hessian

$$\mathbf{M}^T(t)\mathbf{M}(t) \tag{4}$$

where $\mathbf{M}$ is the linear propagator of the forecast model, and the transpose symbol refers to the adjoint operator (Dimet and Talagrand, 1988). An optimal solution can be found in an iterative fashion by taking a sequence of steps in the Newton direction (Gill et al., 1981). Similar techniques are also employed in 4D-Var (Lewis and Derber, 1985; Courtier and Talagrand, 1994).

Figure 4 shows how this optimisation routine works in practice for the two-level model. At each iteration, the initial condition is perturbed in the Newton direction, which reduces the final error. The process is terminated when the initial and final errors are equal. Note that intermediate points do not exceed the shadow radius: this is typical of shadowing behaviour, but needs to be checked for. When performed for a variety of different shadow times τ, the result is a curve of shadow radius versus shadow times, as in Figure 3. Comparing the performance of the sensitivity method to the amoeba method, it seems that the method works well for $r < 4$ (diameter less than 8), but is less efficient at higher radii.

The sensitivity method can be improved by adapting it so that the optimisation is performed relative to the constraint on the size of the initial perturbation. In its usual form, the sensitivity method is solving an unconstrained problem, rather than a constrained one, so the solution found by terminating the iterative process when the initial error equals the final error does not yield an optimal result. This can be addressed by writing the problem instead as

$$\begin{aligned} \text{minimise } C(\mathbf{e}(0)) &= \frac{1}{2}\|\mathbf{e}(\tau)\|^2 \\ \text{subject to } \|\mathbf{e}(0)\| &\leq r \end{aligned} \tag{5}$$

where the constraint is on the magnitude of the initial condition. One approach to solving such problems is the penalty function method (Gill et al., 1981), which transforms the constrained problem into an unconstrained one by adding a penalty term:

$$\text{minimise } C(\mathbf{e}) = \frac{1}{2}\|\mathbf{e}(\tau)\|^2 + \lambda(\|\mathbf{e}(0)\| - r^2) \tag{6}$$

where λ is some suitably large constant. The above formulation will force the initial condition $\mathbf{e}(0)$ to have radius r; alternatively, the penalty function could switch on only if the radius exceeds the shadow radius.

There is a symmetry to the shadow problem, however, which doesn’t distinguish between the initial and final displacements; we could equally well minimise the initial displacement subject to the final displacement being within the shadow radius. A balanced approach, then, is to minimise the sum of the initial and final displacements

$$\text{minimise } C(\mathbf{e}) = \frac{1}{2}\|\mathbf{e}(\tau)\|^2 + \frac{1}{2}\|\mathbf{e}(0)\|^2. \tag{7}$$

The shadow radius r can then be taken as the maximum of these two values. We again assume that intermediate values will remain within bounds; this is easily checked for. The Hessian of the cost function is

$$\mathbf{M}^T(\tau)\mathbf{M}(\tau) + 2\mathbf{I} \tag{8}$$

which can be used to determine the Newton direction. We refer to this method as the ‘pinch’ method, since it involves minimising the initial and final displacements. Figure 3 compares its performance with the other methods; it is more efficient than the sensitivity technique for $r > 4$, but for smaller values there is little difference between the three methods.

Therefore while the pinch method is superior to the sensitivity method, and can also be implemented with fairly minor modifications, it appears that the existing sensitivity code should give reliable results so long as the shadow radius is sufficiently small.

4 The shadow-drift law

Another method to estimate shadow performance is through the shadow-drift law. This states that the expected radius within which the model can shadow the target system for a time τ is either approximately equal to, or greater than, the drift divided by 2. Equivalently, the shadow diameter is bounded below by the drift. Furthermore, if model error is large, then the bound is approached, so the shadow diameter is approximately equal to the drift. The lower panel of Figure 3 shows the drift in the y variables. It provides a fairly accurate estimate of the amoeba method shadow diameter for the first day or so, and gives an underestimate at higher times. Since the drift, which is just a sum of short forecasts, is easily computed, this is a cheap method to estimate shadow times, especially when model error is large.

The shadow-drift law was illustrated in (Orrell, 2001) for a variety of model/system pairs, including weather models. The general proof is given in the same reference, and will be the subject of a future paper. The argument rests on showing

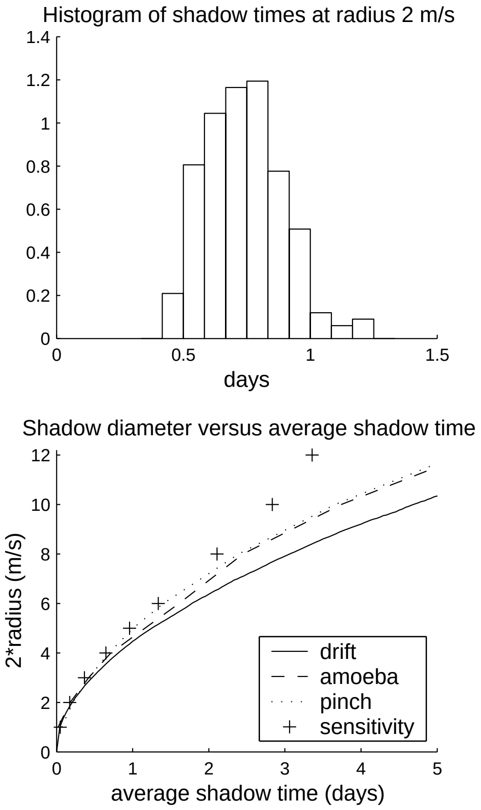

Fig. 3. Plot showing performance of different shadowing schemes. Upper panel shows a histogram of shadow times determined using the amoeba method at shadow radius $2ms^{-1}$. Lower panel compares average shadow times determined using the amoeba, pinch (minimise initial and final errors), and sensitivity (minimise final error only) methods. The pinch method gives results similar to the amoeba method, while the sensitivity method is accurate for shadow times up to 2 days. The solid line is the drift, which gives a lower bound for the expected shadow diameter.

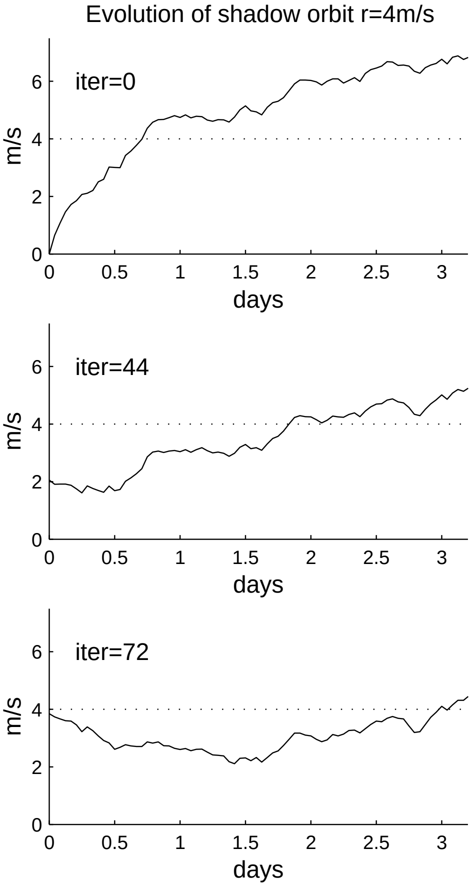

Fig. 4. Evolution of a two-level model shadow orbit using the sensitivity method. Upper panel shows the forecast error in y. The choppy nature of the curve is due to the fluctuations of the stochastic terms, and will be less evident in higher-dimensional systems. Middle panel shows the orbit at iteration 44, part-way through the optimisation procedure. The cost function is given by the square of the final displacement at time 3 days. Each iteration takes a step with direction determined using Newton's method. The procedure continues until (lower panel) the initial displacement equals the final displacement, here $4ms^{-1}$.

that, in a dissipative model, the net effect of a model's sensitivity to initial condition is small when averaged over a large number of experiments. Shadow performance is therefore dominated by the model error, as measured by the drift. We here present a more limited but rather simpler proof, based on the pinch method of searching for shadow orbits, which applies to a specific class of models.

We will first assume that the linearised dynamics in (Orrell et al., 2001) is exact, so error growth can be modelled as

$$\mathbf{e}(\tau) \approx \mathbf{M}(\tau, 0)\mathbf{e}(0) + \mathbf{d}(\tau). \tag{9}$$

(This approximation can be improved by using the *propagated drift* in place of the drift, as discussed in (Orrell, 2002); for shadow orbits the difference is small.)

Suppose that, using the pinch method, we find an initial condition $\hat{\mathbf{e}}_0$ which minimises the average of the initial and final displacements squared, so it solves the problem

$$\text{minimise } C(\mathbf{e}_0) = \frac{1}{2}||\mathbf{M}(\tau)\mathbf{e}_0 + \mathbf{d}(\tau)||^2 + \frac{1}{2}||\mathbf{e}_0||^2. \tag{10}$$

Let $\hat{r}^2 = C(\hat{\mathbf{e}})$. Then if r is the minimum attainable shadow radius, it follows that $r \geq \hat{r}$, i.e. no orbit can shadow (under the linearised dynamics) within a radius smaller than $\hat{r}$. For suppose that there exists an initial condition $\mathbf{e}_0$ for which the initial and final errors are within a radius $r \leq \hat{r}$. Then for this initial condition, we have

$$C(\mathbf{e}_0) = \frac{1}{2}r^2 + \frac{1}{2}r^2 \leq \hat{r}^2 \tag{11}$$

which violates the assumption that $\hat{\mathbf{e}}_0$ is optimal.

The solution of Eq. 10 therefore yields a radius $\hat{r}$ which is an underestimate of the true shadowing radius. We can solve directly for $\hat{r}$ by setting the gradient of the cost function C equal to zero:

$$\mathbf{M}^T(\tau)(\mathbf{M}(\tau)\hat{\mathbf{e}}_0 + \mathbf{d}(\tau)) + \hat{\mathbf{e}}_0 = 0. \tag{12}$$

Dropping the dependence on time for clarity, we have

$$\hat{\mathbf{e}}_0 = (\mathbf{M}^T\mathbf{M} + \mathbf{I})^{-1}\mathbf{M}^T\mathbf{d} \tag{13}$$

where $\mathbf{I}$ is the identity matrix.

We next write the linear propagator $\mathbf{M}$ in its singular value decomposition (SVD) form (Golub and Loan, 1989) as

$$\mathbf{M} = \mathbf{U}\mathbf{W}\mathbf{V}^T. \tag{14}$$

If $\mathbf{M}$ is an n by n matrix, then $\mathbf{U}$ and $\mathbf{V}$ are matrices of the same dimension with orthonormal columns, while $\mathbf{W}$ is a diagonal matrix with positive diagonal entries. Substituting into Eq. 13 then gives an initial displacement

$$\hat{\mathbf{e}}_0 = -\mathbf{V}(\mathbf{W}^2 + \mathbf{I})^{-1}\mathbf{W}\mathbf{U}^T\mathbf{d} \tag{15}$$

from which it follows that

$$\begin{aligned} \hat{r}^2 &= C(\hat{\mathbf{e}}_0) && (16) \\ &= \frac{1}{2}\mathbf{d}^T\mathbf{U}(\mathbf{W}^2 + \mathbf{I})^{-1}\mathbf{U}^T\mathbf{d} && (17) \\ &= \frac{1}{2}\sum_{i=1}^{n}\frac{(\mathbf{d}\cdot\mathbf{u}_i)^2}{1+\sigma_i^2} && (18) \end{aligned}$$

Fix the magnitude of the drift vector, and the multipliers σ_i, and assume that the components of the drift vector are uncorrelated with the direction of the singular vectors. Then if we take the expected value of the sum over all possible orientations of the singular vectors, the term $(\mathbf{d}\cdot\mathbf{u}_i)^2$ is a random variable of magnitude $\frac{||\mathbf{d}||^2}{n}$. Therefore

$$\begin{aligned} \langle\hat{r}^2\rangle &= \frac{1}{2}\langle\sum_{i=1}^{n}\frac{(\mathbf{d}\cdot\mathbf{u}_i)^2}{1+\sigma_i^2}\rangle && (19) \\ &= \frac{||\mathbf{d}||^2}{2n}\sum_{i=1}^{n}\frac{1}{1+\sigma_i^2}. && (20) \end{aligned}$$

(Note that this is slightly different from Eq. 9 of (Orrell et al., 2001), which was obtained by a geometric argument. The result here gives only a lower bound on the shadow radius.)

To demonstrate the shadow-drift law, we wish to show

$$\langle\hat{r}^2\rangle \geq \frac{||\mathbf{d}||^2}{4}, \tag{21}$$

which will occur if

$$\sum_{i=1}^{n}\frac{1}{(1+\sigma_i^2)} \geq \frac{n}{2}. \tag{22}$$

Now, an example of a volume preserving model is one where the magnitudes of the n singular vector multipliers σ_i, when arranged in descending order, follow a power law distribution, so that

$$\sigma_i = \sigma_1^{1-\frac{2i}{n}}. \tag{23}$$

The largest singular vector multiplier is therefore σ_1, and the smallest is $\sigma_n = \sigma_1^{-1}$. An equal number of directions contract as expand in phase space, and because the product of the multipliers is 1, such a model would preserve state space volume.

Given the ideal power law distribution, it is easily seen that

$$\sum_{i=1}^{n}\frac{1}{1+\sigma_1^{2(1-\frac{2i}{n})}} = \frac{n}{2}. \tag{24}$$

It therefore follows that, if a plot of the singular vectors lies beneath the power series distribution for some choice of σ_1, then the shadow-drift law will apply.

This argument therefore proves the shadow-drift law for only a specific class of models, namely those which are more dissipative than the power law case, in the sense described above. However, real models often have this characteristic. The upper panel of Figure 5 plots the average singular value multipliers for the two-level model, compared with a power law distribution (a straight line in the semilog scale). Results, which are averaged over 100 shadow experiments at shadow radius 4.0 ms^{-1}, always lie below the straight line, indicating that the model is more dissipative than the volume-preserving ideal power law case. The lower panel compares the root-mean-square value of each term in the sum Eq. 22 to the corresponding term for the power law; since the solid

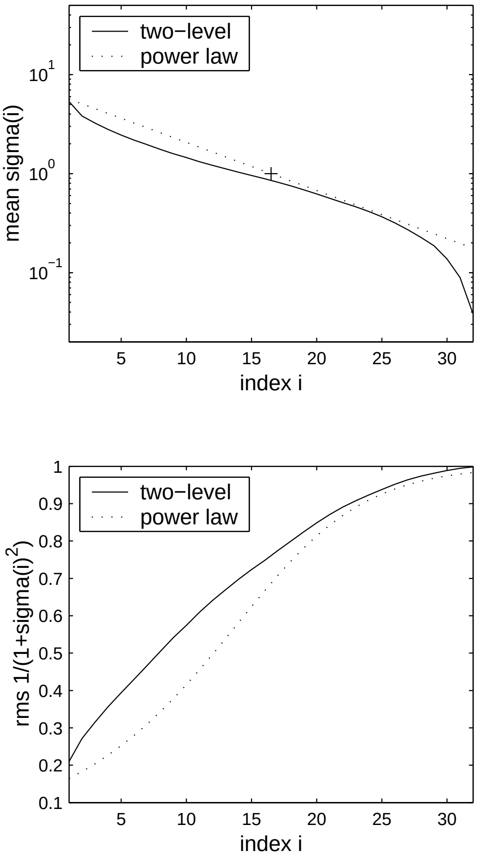

Fig. 5. Top panel shows the average singular value multipliers for the two-level model (solid line), compared with a power law distribution (dotted line, which is straight in the semilog scale). Results are averaged over 100 shadow experiments at shadow radius 4.0 ms^{-1}. The solid line is always below the straight line, indicating that the model is more dissipative than the volume-preserving ideal power law case. The lower panel compares the root-mean-square value of each term in the sum Eq. 22 to the corresponding term for the power law. Here the solid line is always above the dotted line.

line is always above the dotted line, it follows that the minimum shadow radius will be greater than that predicted by the shadow-drift law, so the law gives a lower bound as claimed. For weather models, where the full range of singular values is not calculated, we need to evoke the general proof, which applies to dissipative models.

The shadow-drift law can be used to estimate shadow times for a range of different shadow radii. The results can then be compared with actual shadow experiments, as in Figure 3. Because the shadow calculations and the drift calculations are performed in completely different ways, this gives two independent methods of estimating shadow times. Performing the calculations for a variety of different shadow times will also negate the possibility that the two tests agree by accident.

A drawback to these two methods is that they really apply only to ensembles formed by perturbating the initial condition, not the model itself. The next section presents a technique to measure the effect of model error on a given ensemble, however it is generated.

5 The mean projection test

The third method to gauge the effect of model error on ensemble forecasts is to check whether the convex hull of a particular ensemble contains a shadow point (i.e. a point within the shadow radius of the analysis) after a period of time. By convex hull, we mean as formed within the subspace spanned by the ensemble members, so if there are n ensemble members, the convex hull has dimension $n + 1$.

The reason for using the convex hull is that the ensemble is usually formed by taking each perturbation at a set magnitude. If all ensemble members fail to shadow, this still leaves the possibility that some linear combination of the initial perturbations, that lies within their convex hull, could shadow. If we assume that the model is roughly linear over short times (though see (Gilmour et al., 2001), then the image of that point at a set time would lie within or near the convex hull of the ensemble. Therefore, if the convex hull can be shown to be moving away from the analysis, it is a much clearer indication of model error than showing only that the ensemble members themselves fail to shadow.

This technique is less general than the shadow methods, since it applies only to a particular ensemble, and cannot be used to determine the overall shadow performance of a model. It is also a stronger condition, since it demands that the convex hull actually contains a shadow point, rather than asking whether a shadow point could in principle exist given the right perturbation. An advantage is that it is extremely easy to perform. It can also be applied to ensemble schemes which include perturbations to the model.

For a particular time τ, each ensemble member is expressed as an error field (say 500 hPa) over a grid. The ensemble mean error is then computed. Figure 6 is a schematic diagram showing the ensemble errors, mean error, and convex hull of the ensemble in a 2-D space. We claim that to check whether the convex hull contains members within a distance r of the origin, it suffices to take the projection of each ensemble error onto the mean error vector (solid line joining the ensemble mean to the origin in the figure). If the projections of the errors are all greater than the shadow radius, then the errors themselves must be greater than that radius. This holds for any point in the convex hull.

The method is called the *mean projection test*. Referring to the figure, if p is the magnitude of the projection of the nearest point $\mathbf{e}$ onto the mean error, then so long as p is greater than the shadow radius, no point in the convex hull can be a shadow point.

To see why this is the case, consider the plane which is orthogonal to the mean error and a distance p from the origin. This plane is indicated in the figure by the solid line which contains the point $\mathbf{e}$. If the mean projection test fails, then $p > r$, and all the ensemble errors, and therefore the entire convex hull, must be either on, or on the opposite side from the origin of, this plane. Since the minimum distance from the plane to the origin is p, it follows that no ensemble error is within the shadow radius r.

Figure 7 shows the mean projection test applied to a 20-

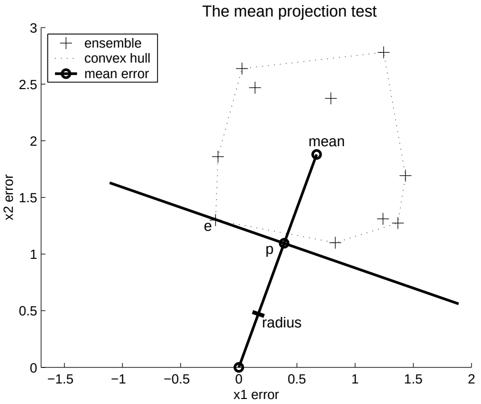

Fig. 6. Schematic diagram showing the mean projection test. Ensemble errors (in 2-D) are indicated by ’+’ symbols. Solid line starting at origin is the error of the ensemble mean. The projection of ensemble member **e** onto the mean error is the distance p. Since this projection is greater than the shadow radius (here 0.5) for all ensemble members, the mean projection test fails, and we deduce that the convex hull of the ensemble does not contain a shadow point.

member ensemble for the two-level model, in the x and y variables. The projections of the ensemble errors onto the mean error are shown each 3 hours. Since the projections are always greater than the shadow radius, indicated by the dotted line, it follows that this particular ensemble does not contain a shadow point for even 3 hours.

For the two-level model, all three techniques therefore give a similar answer: ensembles in the y variables (corresponding to total energy), with a perturbation radius of $0.5ms^{-1}$, will shadow at that radius for at most a few hours. The ensembles cannot therefore be considered a reliable probabilistic guide to the true state of the system. Since the two-level drift was chosen to match that of the GCM, the shadow-drift law implies that GCM shadow times will be similar; however, this must be confirmed by searching for actual shadow orbits, for example with the sensitivity code, and checking whether particular ensembles contain shadow points, with the mean projection test.

Because the mean projection test shows whether an ensemble is drifting away from truth, it can serve as a test for model error. Ensembles can therefore be used, not just as a forecasting tool, but as a method for determining or verifying the importance of model error relative to initial condition error.

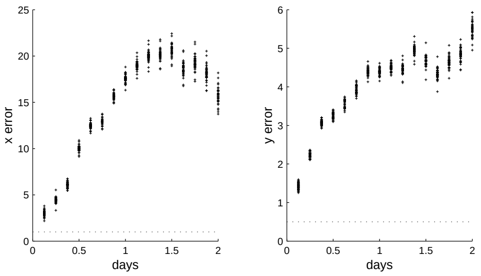

Fig. 7. Plot showing the mean projection test applied to the two-level system. The projection of the ensemble errors onto the mean error, sampled each 3 hours, is greater than the shadow radius (dotted line) at any time, implying that the convex hull of the ensemble does not contain a shadow point.

6 Comparisons with weather models

To show how the above techniques can be applied to real weather models, we briefly compare two weather model case studies with results for the two-level model. The intention is to illustrate that the methods are workable, and that the two-level model is capable of simulating actual ensemble schemes. The first case study is the inter-model comparison between the ECMWF T42 and TL159 models, which was presented in (Orrell et al., 2001). In this experiment, the T42 model was used to shadow a TL159 target orbit. The shadow radius at 2 days was estimated from the shadow-drift law, and the sensitivity code. We will compare the results with a modified version of the two-level system.

The solid line in the upper left panel of Figure 8 shows the total energy forecast error. The large initial error is due to the truncation operator between TL159 and T42. The drift at time 2 days was estimated from a sum of short forecast errors to be $1.8ms^{-1}$ (it is about equal to the total forecast error). From the shadow-drift law, the expected shadow radius is therefore half the drift, or $0.9ms^{-1}$.

The sensitivity code was then used to produce an actual shadow orbit. The dotted line shows the result after fifty iterations of the optimisation procedure. Because the radius at two days is slightly greater than the initial radius, it appears that the shadow orbit could have been improved by performing further iterations. The average of the initial and final errors is about $1.1ms^{-1}$, which is slightly greater than the drift over two, as expected by the shadow-drift law.

For comparison, the upper right panel shows a y variable error curve and typical shadow orbit for the two-level (medium error) model, where the model error terms were reduced by a factor 4, and observation errors by a factor 2, so as to match the inter-model error curve. The shadow orbit was calculated using the amoeba method with a shadow radius of $1.0ms^{-1}$.

The two methods for estimating the two-level shadow radius at time two days therefore give a radius in the region of $1.0ms^{-1}$. Since this is greater than the ensemble perturbation radius, it means that an ensemble is not expected to contain shadow points at 2 days. In fact, the drift becomes equal to twice the shadow radius after about half a day, so the ensemble should cease to contain shadow points after this time.

The lower panels show ensemble errors calculated for T42 (left) and the two-level model with medium error (right). The two-level ensemble contains a large sample of 400 perturbations, so it is safe to assume that no shadow point exists after at most half a day; for the T42 ensemble, which contains only 50 perturbations in the +/- directions of the leading 25 singular vectors, a mean projection test should ideally be performed, however this was not done at the time of the exper-

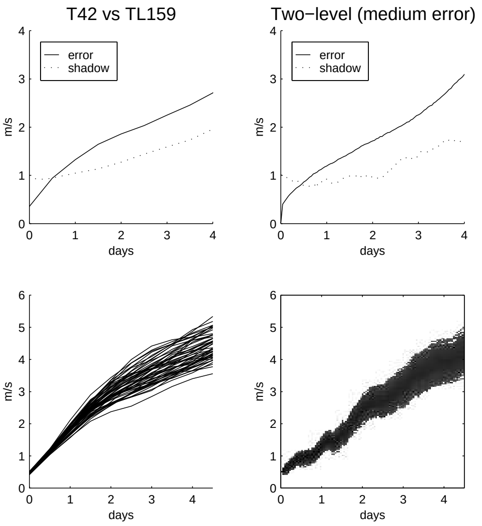

Fig. 8. Plot comparing shadow behavior for the T42/TL159 inter-model experiment, and the two-level system with medium error. Upper left panel shows T42 forecast error and shadow orbit optimised at 2 days. Lower left panel shows a T42 ensemble. Panels on right show same for the two-level model with medium error, where the model error terms have been reduced by a factor 4, and the observation errors by a factor 2, to match the inter-model data.

iment. The effect of model error is clear for either ensemble by the large initial slope.

It is interesting to note that the ensemble errors relative to T159 can be viewed as the orthogonal sum of the T42 forecast error, and errors of the T42 ensemble with respect to an unperturbed T42 control. This is shown in Figure 9. The upper panel shows the T42 ensemble errors e_{42} relative to T42. Because there is no model error in this situation, the errors are due to initial condition only. The lower panel shows the T42 forecast error f_{159} relative to T159 (dotted line). This error is due primarily to the model. The orthogonal sum of f_{159} with the T42 ensemble errors e_{42} is a good approximation to the total ensemble errors in Figure 8. This is a consequence of Eq. 9, which shows that the error can be approximated as the sum of two components, one due to the initial condition, and one to the model error. The reason the two-level model can approximate the weather model ensemble behaviour is because it has the right amount of model error and the right sensitivity to initial condition.

The second case study is an ensemble from the NOAA MRF model. The upper left panel of Figure 10 shows the ensemble errors at one-day increments in 500 hPa. The error of the ensemble mean is also shown. Note the large size of the initial perturbations. The average perturbation size is about $12m$, which we set as the shadow tolerance. Because of the large initial perturbations, and the increased shadow tolerance, the expected shadow times are longer than in the previous case study. The lower left panel shows a mean projection test for this ensemble. The radius of $12m$ is indicated by the dashed line, and the mean projection test appears to fail after about 2.5 days.

How does this compare with expected shadow behaviour?

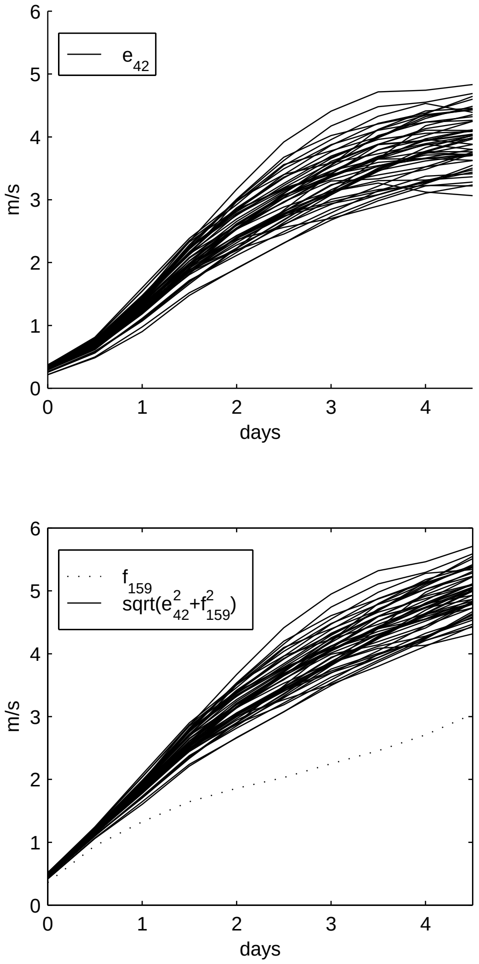

Fig. 9. Plot showing how inter-model ensemble errors can be viewed as the orthogonal sum of errors due to initial condition, and error due to the model. Top panel shows errors e_{42} of the T42 ensemble with respect to the T42 control forecast. Since there is no model error, the errors are due to initial condition only. Lower panel shows the forecast error f_{159} relative to the T159 model, which is primarily due to model error (dashed line). The orthogonal sum of this with the ensemble errors e_{42} gives a result (solid line) which closely approximates the ensemble errors relative to T159 in the lower left panel of Figure 8.

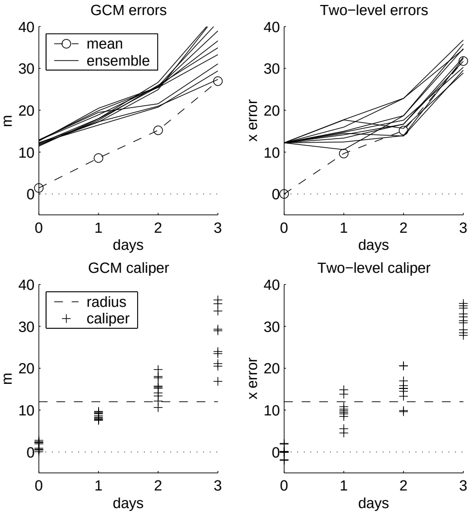

Fig. 10. Plot showing ensemble errors and mean projection test for the MRF model in 500 hPa, and the two-level model in x. The MRF ensemble was initiated at Oct 30, 2001. Top panels show root-mean-square errors, evaluated once per day. The perturbation size is $12m$ for the MRF ensemble. Lower panels show the corresponding mean projection test. In either case, the projected errors exceed the $12m$ line after about 2 days.

Since we are working in a non-global metric, statements about shadowing will depend to an extent on the other, unseen variables. Nevertheless, we can make some broad approximations. As a proxy for drift, we can use the ensemble mean error (if model error is large, then the drift accounts for most of the error over short times (Orrell, 2002)). An ensemble with perturbation size $12m$ should shadow at most until the mean error is about twice the radius, or $24m$. Extrapolating the error curve, this happens after about 3 days. Since this is the best expected shadow time for an optimal perturbation, it is reasonable that the ensemble should shadow a slightly shorter time.

For comparison, the right panels show a two-level model ensemble in x. The results of the mean projection test in the lower right panel again show that the ensemble contains no shadow point at radius $12m$ after about 2.5 days. Note that the reason shadow times are longer than in Figure 7 is because the shadow radius is $12m$ instead of $2m$. Loosely speaking, if drift varies approximately with the square-root of time, then expected shadow times will vary approximately with the square of shadow radius.

Together, these case studies show that the three methods based on drift, shadow experiments, and mean projection tests, are feasible techniques which can be applied to full weather models. In terms of computational expense, the drift is comparable to a single long forecast, while the mean projection test is similar to calculating root-mean-square errors. The studies also show that the two-level model can be adapted to simulate a number of different weather models.

7 Model perturbations

The above techniques will help to validate the performance of ensembles. A related question is how ensembles might be improved if model error is large. As discussed in the introduction, one approach that has been adopted is to use stochastic elements in the model equations, either by perturbing the model parameters, or by adding a stochastic forcing term.

Suppose for example that the errors are assumed to arise from sub-grid scale processes, and we decide to account for them by adding stochastic terms to the model. For the two-level model, we could do this by simply adding stochastic terms which have the same magnitude as those in the true system. The model and the system would then have identical equations, but different realisations of the stochastic forcing. Figure 11 shows the result. In the upper panels, the spread of the ensemble errors has increased relative to Figure 2. However, so has the mean error. The reason is that errors are dominated by the drift, which is the integral of the tendency error over time. The unperturbed model is in a sense an optimal choice, because the constant forcing minimises the expected value of the tendency error, and therefore the drift. If the model contains stochastic terms equal in magnitude to those of the system, then the expected tendency error is the expected value of the sum of the two stochastic terms, which represents an increase by a factor $\sqrt{2}$. Therefore, while the spread increases, so does the mean error, and the net effect is that overall accuracy is not improved: of the 400 points tested, none managed to shadow within a reasonable tolerance. We could not therefore say that adding the stochastic terms has helped ensemble performance, if the goal is to provide a probability distribution function.

The lower panels, however, tell a very different story. Because the model and the true system now have identical variability, the model yields a perfect Talagrand diagram, even though the ensemble is no closer to tracking the system. This shows the limitation of statistical techniques when discussing ensemble quality. The problem is that prediction of near-term errors, and the prediction of long-term error variability, are two completely different questions. Statistical verification methods are better suited to the latter than the former.

To further illustrate this point, suppose that we wish to predict the next number in a random string of 0's and 1's. To minimise the expected root-mean-square error, we should choose a value of 0.5 as our prediction, for which the RMS error is 0.5. This is essentially what we did with the model of the two-level system, where we used the average value of the forcing and ignored the stochastic terms. Another approach would be to use a coin, one face marked 0 and the other 1, as a 'model' for the random string. We toss the coin, and take that as the prediction. In the long-term, the model would perfectly replicate the climatology of the system, in that it would give 0's half the time and 1's half the time. However the expected RMS error for a single prediction, which is $\frac{1}{\sqrt{2}}$, has increased by a factor $\sqrt{2}$. This corresponds to the case

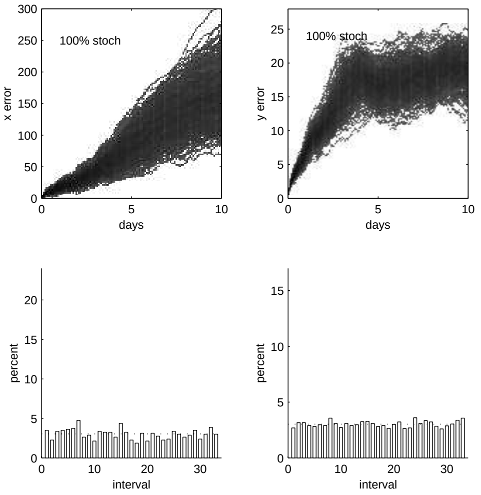

Fig. 11. Plot showing the effect on ensemble performance of adding stochastic terms to the model. Left panels show errors in x variables, right panels show errors in y variables. Upper panels are ensemble errors, lower panels show Talagrand diagrams of a 2-day forecast. While the addition of stochastic terms increases the spread of the ensemble, and yields a perfect Talagrand diagram, it actually detracts from overall accuracy.

where we added stochastic terms to the model.

For the model of a random string, it is at least the case that an ensemble with spread equal to the mean error, e.g. that consisting of 0 and 1, is guaranteed to contain the true value. In a high dimension space, the situation is worse. Any random perturbation is expected to be orthogonal to the model error, so will not correct it. The spread will increase, but so will the mean error. Therefore the ensemble's chance of containing shadow points will probably not improve.

In any case, the existence of a shadow orbit is a necessary, but not a sufficient, condition for ensembles to be effective. An ensemble consisting of all possible two-level states would contain a shadow point, but no useful information. The goal of ensemble forecasting is not to increase the spread by any means possible, since the usefulness of the resulting probabilistic forecast will vary inversely with the range of possibilities portrayed.

Perturbing, or changing, the model may well work better for weather models than for the two-level system, which is perhaps a worst-case example since the errors are entirely random. However, there remain some additional theoretical questions. Ensembles are suitable for simulating the effects of initial condition error because we know certain things about that error source. For example, the 'true' initial condition (i.e. the weather) is expected to exist within a certain distance of the analysis. Also, perturbations can be chosen in the direction of rapidly growing modes, to give the maximum spread. Model error is a completely different situation. There is no obvious 'model space' counterpart to singular vectors or bred vectors. There may even be no accessible set of equations that perfectly mimic the dynamics of the system (Smith, 2000; Judd and Smith, 2001). The ensemble methodology, which was designed to handle initial condition error, cannot necessarily be transplanted to deal with model error - at least until we know more about the nature of this error.

For ensembles to provide a reasonable probability distribution, model error should ideally be reduced below some threshold. It is interesting to ask, again in the somewhat idealised context of the two-level system, how good the model needs to be. It was seen with the inter-model case study that model error had a significant effect on ensemble performance when the model error in the two-level system was reduced by a factor 4. The two-level model was therefore again run with the stochastic model error terms this time reduced by a factor of 10. The upper panels of Figure 12 show errors of a typical ensemble; note the different vertical scale. The expected shadow times at a radius $0.5ms^{-1}$ are in the region of 4 days. The ensemble will therefore contain a shadow orbit for around this time: however, it still does not function very well as a probabilistic forecast. It appears that even a small amount of model error is capable of disrupting the performance of ensembles. The middle panels show the Talagrand diagrams at 2 days. The statistics are worse than for the stochastic model discussed above, even though ensemble accuracy is far improved! The lower panels show the results of the mean projection test applied every 12 hours over a 7-day period. The ensembles in either x or y variables cease to contain shadow points after about 4 days, as indicated by the fact that the projections onto the mean error are greater than the shadow tolerance (dotted line) past this time. Note that, if model error is reduced by a factor 10, then forecast error growth will be as shown in Figure 1, so the model has excellent predictive skill.

8 Conclusions and future work

The aim of ensemble schemes is to provide a probability distribution function of the weather's future state. If this is to occur, then a reasonable condition that must be satisfied is the existence of a shadow orbit. In this paper we have discussed independent tests for estimating shadow times. The first is based on the currently existing sensitivity code, which can be used to produce candidate shadow orbits. The second is based on the shadow-drift law, which relates expected shadow performance to model error as measured by the drift. Both these techniques have already been tested in inter-model experiments at ECMWF. By plotting a curve of shadow radius versus shadow times using both methods, the chances of the two agreeing by accident can be mitigated. The sensitivity code can also be improved by including the initial constraint, as in the pinch method. A third technique, the mean projection test, is a simple method to check whether a particular ensemble, however generated, contains a shadow point after a period of time. It can also serve as a method to determine, through ensemble behaviour, the effects of model error.

Ensemble schemes are essential tools for studying the effects of initial condition error. However, for the two-level model, it appears that model error must be small in order for

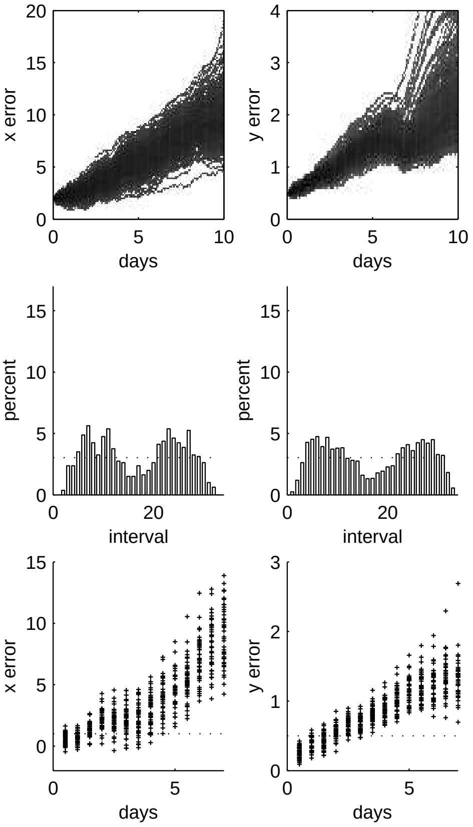

Fig. 12. Plot showing ensemble performance when stochastic error in the two-level model is reduced by a factor 10. Upper left panel is x errors, upper right panel is y errors. Middle panels show the 2-day Talagrand diagrams: note the difference in shape from the high-error case. Lower panels show the results of the mean projection test, which indicates that the convex hull of the ensemble ceases to contain shadow points after about 4 days. Shadow times for a y radius of $0.5ms^{-1}$, determined by the amoeba method, are also in the region of 4 days.

ensembles to yield accurate probabilistic forecasts. Similar effects may also occur with weather models. This is not to say that, when model error is high, a probabilistic approach to forecasting is no longer required, or that ensembles will not be applicable. Even if a model is incapable of shadowing the analysis, it may still be the case that ensembles are a useful tool for making forecasts, say of temperature in a particular region (any method of generating an ensemble will produce a certain spread, and the larger the spread, the greater the chance that the observed value will fall within the bounds of the ensemble). However, since in this case the forecast errors are not themselves primarily a result of changes in the initial condition, it follows that perturbing the initial condition may not be the most appropriate or efficient way to produce a probabilistic distribution. And, as discussed in Section 6, model perturbations are not without their difficulties. For these reasons, other techniques, such as the use of past error statistics, may provide an equally valid, and certainly cheaper, method to generate probabilistic forecasts.

The two-level system has been used to simulate a number of different weather models. While it manages to reproduce many aspects of GCM behaviour, however, it is only a stand-in for the real thing. As mentioned earlier, for example, the effect of analysis error, which is a complex convolution of observation error and model error, is not adequately represented by the two-level system. Nor does it fully capture the variability of error growth over different scales. However, the fact that a system can be produced which agrees reasonably well with GCM behaviour, but fails to yield accurate ensemble forecasts, demonstrates that weather models need to be carefully examined to validate the ensemble approach. The most direct method is to establish the existence of shadow orbits. Such experiments will reveal the effect of model error on current ensemble schemes.

Appendix A The two-level system

The two-level system is a scaled version of the Lorenz ’96 system, which was used in (Lorenz, 1996) to simulate error growth, with stochastic terms added. The equations are

$$\frac{dx_i}{dt} = x_{i-1}(x_{i+1} - x_{i-2}) - x_i + F - \sum_{j=1}^{4} y_{i,j} + N_x \tag{A1}$$

$$\frac{dy_{i,j}}{dt} = c^2 y_{i,j+1}(y_{i,j-1} - y_{i,j+2}) - c y_{i,j} + x_i + c N_y \tag{A2}$$

for $i = 1$ to 8, and $j = 1$ to 4. The indices are cyclic, so for example $x_{i+8} = x_i$ and $y_{i,j+4} = y_{i+1,j}$, and the variables can be viewed as atmospheric quantities around a circle. The parameter c is set to 10. The x variables are scaled by a factor 900 to put in units of m for comparison with GCM 500 hPa results, while the y variables are scaled by a factor 5.3 to put in units of ms^{-1} for comparison with total energy. Time is scaled by a factor 100 to put in days. $F = 14$ is a constant forcing term, while N_x and N_y are random variables with variance 2.5 and 7.5 respectively, updated every hour. In addition, the x and y variables are observed each hour with a stochastic error O_x and O_y, which have standard deviation $1.0m$ and $0.5ms^{-1}$ respectively. These terms are meant to simulate the random component of the analysis errors.

In the **medium error** system, N_x and N_y are reduced by a factor 4, and the observation error O_x and O_y by a factor 2. In the **low error** system, N_x and N_y are reduced by a factor 10, while observation error is unchanged unless otherwise specified.

The model has the same equations, but with no stochastic forcing, so $N_x = N_y = 0$, and no observation error so $O_x = O_y = 0$. The difference between the model and the system is therefore the stochastic forcing terms, and the observation error. Equations are solved using a Runge-Kutta scheme with time step of one hour. A long transient of 100,000 hours is run before making calculations.

Acknowledgements. Thanks to J. Barkmeijer and M. Leutbecher for help with the ECMWF weather model calculations, and to J. Hacker for the ensemble data. Thanks also to L. Smith for useful discussions.

References

Bennett, A.I. and Budgell, W.P., 1987, Ocean data assimilation and the Kalman filter: Spatial regularity, *J. Phys. Oceanogr., 17*,1583-1601.

Buizza, R., Miller, M., and Palmer, T.N., 1997, Stochastic simulation of model uncertainties in the ECMWF Ensemble Prediction Scheme, in *Predictability*, edited by T.N. Palmer, European Centre for Medium-Range Weather Forecasting, Shinfield Park, Reading UK.

Buizza, R., Barkmeijer, J., Palmer, T.N., and Richardson, D.S., 2000, Current status and future developments of the ECMWF Ensemble Prediction System, *Meteorol. Applic., 7*, 163-175.

Courtier, R. and Talagrand, 0., 1994, Variational assimilation of meteorological observations with the adjoint vorticity equation. Part 2: Numerical results., *Q.J.R. Meteorol. Soc., 113*, 1329-1347.

Dimet, J. L. and Talagrand, 0., 1988, Variational algorithms for analysis and assimilation of meteorological observations, *Tellus, 37A*, 97-110.

Ehrendorfer, M., 1997, Predicting the uncertainty of numerical weather forecasts: a review, *Meteorologische Zeitschrift, 6*, 147-183.

Gill, P., Murray, W., and Wright, M., 1981, *Practical Optimization*, Academic Press, New York.

Gilmour, I., 1998, Nonlinear model evaluation: ι-shadowing, probabilistic prediction and weather forecasting, *D. Phil. Thesis, Oxford University*.

Gilmour, I., Smith, L., and Buizza, R., 2001, On the duration of the linear regime: Is 24 hours a long time in weather forecasting?, *J. Atmos. Sci.*, under review.

Golub, G. and Loan, C. V., 1989, *Matrix Computations*, The John Hopkins University Press.

Hansen, J. A., 1999, private communication.

Harrison, M., Palmer, T., Richardson, D., and Buizza, R., 1999, Analysis and model dependencies in medium-range ensembles: two transplant case studies, *Q.J.R. Meteorol. Soc., 125*, 2487-2515.

Houtekamer, P., Lefaivre, L., Derome, J., Ritchie, H., and Mitchell, H., 1996, A system simulation approach to ensemble prediction, *Mon. Wea. Rev., 124*, 1225-1242.

Judd, K. and Smith, L., 2001, Indistinguishable states 1: The perfect model scenario, *Physica D, 151*, 125-141.

Lewis, J. and Derber, J., 1985, The use of adjoint equations to solve a variational adjustment problem with advective constraints, *Tellus, 37A*, 309-322.

Lorenz, E., 1996, Predictability - a problem partly solved, in *Predictability*, edited by T. Palmer, European Centre for Medium-Range Weather Forecasting, Shinfield Park, Reading UK.

Lorenz, E. N., 1963, Deterministic nonperiodic flow, *J. Atmos. Sci., 20*, 130-141.

Molteni, F., Buizza, R., Palmer, T., and Petroliagis, T., 1996, The ECMWF ensemble prediction system: Methodology and validation, *Q.J.R. Meteorol. Soc., 122*,73-119.

Orrell, D., 2001, Modelling nonlinear dynamical systems: chaos, error, and uncertainty, *D.Phil. Thesis, Oxford University*.

Orrell, D., Smith, L., Barkmeijer, J., and Palmer, T.N., 2001, Model error in weather forecasting, *Nonlin. Proc. Geo., 8*.

Orrell, D., 2002, Role of the metric in forecast error growth: how chaotic is the weather?, *Tellus*, in press.

Palmer, T.N., 2000, Predicting uncertainty in forecasts of weather and climate, *Reports on Progress in Physics, 63*, 71-116.

Philips, N., 1986, The spatial statistics of random geostrophic modes and first-guess error, *Tellus, 38A*, 314-322.

Press, W., Flannery, B., Teukolsky, S., and Vetterling, W., 1993, *Numerical Recipes in C: The Art of Scientific Computing*, Cambridge University Press, Cambridge.

Rabier, F., Klinker, E., Courtier, P., and Hollingsworth, A., 1996, Sensitivity of forecast errors to initial conditions, *Q.J.R. Meteorol. Soc., 122*, 121-150.

Smith, L., 1996, Accountability in ensemble prediction, in *Predictability*, edited by T. Palmer, European Centre for Medium-Range Weather Forecasting, Shinfield Park, Reading UK.

Smith, L., 2000, Disentangling uncertainty and error: On the predictability of nonlinear systems, in *Nonlinear Dynamics and Statistics*, edited by A.I. Mees, pp. 31-64, Birkhauser, Boston.

Strauss, B. and Lanzinger, A., 1996, Validation of the ECMWF prediction system, in *Proc. 1995 ECMWF Seminar on Predictability*, edited by T. Palmer, European Centre for Medium-Range Weather Forecasting, Shinfield Park, Reading UK.

Toth, Z. and Kalnay, E., 1993, Ensemble forecasting at NMC: the generation of perturbations, *Bull. Am. Meteorol. Soc., 74*, 2317-2330.

Toth, Z., Kalnay, E., Tracton, S., Wobus, S., and lrwin, J., 1996, A synoptic evaluation of the NCEP ensemble, in *Proc. 5th Workshop on Meteorological Operating Systems*, edited by T. Palmer, European Centre for Medium-Range Weather Forecasting, Shinfield Park, Reading UK.