QUARTERLY JOURNAL OF THE ROYAL METEOROLOGICAL SOCIETY
Q. J. R. Meteorol. Soc. **00**: 1–16 (0000)
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/qj.000

How good is an ensemble at capturing truth? : Using bounding boxes for forecast evaluation

Kevin Judd[a*] and Leonard A. Smith[bc] and Antje Weisheimer[cd]

[a] *School of Mathematics and Statistics, University of Western Australia, Perth*
[b] *Oxford Centre for Industrial and Applied Mathematics, Mathematics Institute, Oxford*
[c] *Centre for Analysis of Time Series, London School of Economics, London*
[d] *European Centre for Medium-Range Weather Forecasting, Reading*

Abstract: Ensemble prediction systems aim to account for uncertainties of initial conditions and model error. Ensemble forecasting is sometimes viewed as a method of obtaining (objective) probabilistic forecasts. How is one to judge the quality of an ensemble at forecasting a system? The probability that the bounding box of an ensemble *captures* some target (such as "truth" in a perfect model scenario) provides new statistics for quantifying the quality of an ensemble prediction system; information that can provide insight all the way from ensemble system design and to user decision support. These simple measures clarify basic questions, like, what the minimal size of an ensemble should be. To illustrate their utility, bounding boxes are used in the imperfect model context to quantify the differences between ensemble forecasting with a stochastic-model ensemble prediction system and a deterministic-model prediction system. Examining forecasts via their bounding boxes statistics provides illustration of how adding stochastic terms to an imperfect model may improve forecasts even when the underlying system is deterministitc. Copyright © 0000 Royal Meteorological Society

KEY WORDS Ensemble forecasting; Probability forecasting; Bounding box

Received ; Revised ; Accepted

1 Introduction

There are many uncertainties associated with operational forecasting. In numerical weather prediction (NWP), for example: observations of the atmosphere and oceans are incomplete and sometimes inaccurate; there are uncertainties introduced when data is assimilated into forecast models; the models themselves are only an approximate representation of the physical processes of weather; there are dynamical instabilities that amplify errors of the model and state, which further degrade forecasts. Some action must be taken to account for the uncertainties of NWP and ensemble forecasting appears to be the best available option, because an ensemble reflects known sources of uncertainty of a forecast. Presently operational NWP uses ensembles that are constructed by perturbing a best guess analysis in the direction of singular vectors (Buizza, 1995; Molteni et al., 1996; Mureau et al., 1993) or bred vectors (Toth and Kalnay, 1997) or model variations (Houtekamer et al., 1996); other proposed techniques for ensemble forecasting include random perturbations (Errico and Baumhefner, 1987; Tribbia and Baumhefner, 2003), ensemble Kalman filters (Evensen, 1994; Anderson, 1999; Houtekamer et al., 2005; Bishop et al., 2001; Hamil and Synder, 2000), and sets of indistinguishable states (Judd and Smith, 2004). All these methods are trying to account for uncertainties in a forecast. There is also growing interest in using ensemble forecasts to provide probabilistic forecasts (Petterssen, 1958; Anderson, 1996; Talagrand et al., 1997), which would be more valuable than best guess forecasts for many, if not all, forecast users.

Providing probabilistic forecasts presents extreme technical and practical difficulties. The most obvious difficulty for NWP is that of describing a joint probability distribution in millions of variables. In principle, any probability density can be represented with arbitrary accuracy by an ensemble that is a sufficiently large random sample *of that* density. The probability of any event can then be approximated by counting the frequency of the event in the ensemble, and the approximation will typically improve as the sample size is increased. To achieve an approximation of a joint distribution with a given accuracy, however, the size of the sample must increase exponentially with the number of variables. Although techniques like kernel density estimation (Silverman, 1988), ensemble dressing (Roulston and Smith, 2003; Gneiting et al., 2004) and weighted ensembles (Judd and Smith, 2001) can dramatically reduce the required size of an ensemble, the "curse of dimensionality" persists[†]. It has to be expected that ensemble/probabilistic forecasting can never provide

*Correspondence to: Email: Kevin.Judd@uwa.edu.au

[†]Silverman estimates that for kernel density estimation the size of the ensemble needs to increase as the fifth power of the number of variables (Silverman, 1988).

Copyright © 0000 Royal Meteorological Society

unbiased marginal distributions of more than a few uncorrelated variables, and perhaps a few more correlated variables. Even when restricted to marginal densities, the fact that our models are imperfect places serious and severe limitations on aspirations for probabilistic forecasting. An imperfect model does not provide a probabilistic forecast that is a description of the system (the weather), it is a description of the model's behaviour, and as such any probability forecast formulated for a real event can be quite different from the "probability of an event" itself.

The provision of probability forecasts is a complex issue, especially when imperfect models are considered (Smith, 2000). Ideally one would like to achieve *accountable* probability forecasts (Smith, 1995, 1997, 2000), but we are far from this goal. We therefore divide the problem into simpler questions. Some amount of quality control in ensemble forecasting can be ensured by requiring that ensembles have certain significant properties, and one minimal property we should require is that the ensemble *captures the target* with a high probability. We are yet to define exactly what capturing the target might mean, but one can imagine that if an ensemble somehow captures the target with high probability, then the marginal distribution computed from the ensemble will not be entirely misleading, in that it will not assign a very low, or zero, probability to what actually happens. On the other hand, one could presumably capture any target routinely by ensuring a widely dispersed ensemble, but such an ensemble would provide little useful probabilistic information.

In the perfect model scenario there is a true state, so capturing the target could mean that *truth* lies within the cloud of ensemble states. In the imperfect model scenario there is no true state, however, we can aim to capture some *target*. What target might mean is defined more carefully in section 3, for the present a target can be thought of as a collection of future observations, or a future analysis.

As we discuss below, insights and statistics derived via the bounding boxes provide information independent of and complementary to those from other commonly used tests, such as, 1-dimensional rank histograms (Anderson, 1996; Hamill, 2001; Talagrand et al., 1997), multi-dimensional rank histograms from minimum spanning trees (Smith, 2000; Smith and Hansen, 2004; Wilks, 2004), and other common skill scores (Wilks, 1995). We also demonstrate this test is useful. Independent of their value in evaluating forecast quality and ensemble design, bounding box statistics are of immediate use to decision makers in evaluating an ensemble prediction system's likely value

We will now assume there is some unique and well defined target state that we require the ensemble *captures* with high probability. One way an ensemble might capture the target is by having the target lie within the convex hull of the ensemble (figure 1a), that is, given an ensemble $\mathcal{E}$ (a finite set of states) and the target state $x^\star$, there exists $0 \le \lambda_x \le 1$ for each $x \in \mathcal{E}$ such that

$$\sum_{x\in\mathcal{E}} \lambda_x = 1 \quad \text{and} \quad x^\star = \sum_{x\in\mathcal{E}} \lambda_x x. \tag{1}$$

In a d-dimensional space a convex hull requires at least $d+1$ states for it to contain any volume. This puts a lower bound on the size of an ensemble that for weather forecasting is impossible to achieve. An alternative capture criterion requires that the target lie within the bounding box of the ensemble (figure 1b), that is, each component $x_i^\star$ of the target vector lie between the minimum and maximum values of the corresponding components of the ensemble vectors

$$\min_{x\in\mathcal{E}} x_i \le x_i^\star \le \max_{x\in\mathcal{E}} x_i \quad \text{for all } i. \tag{2}$$

A bounding box is trivial to compute and is defined for any size ensemble, because two or more states is sufficient to define a bounding box[‡].

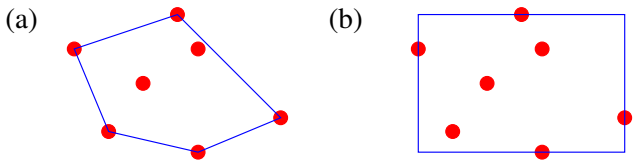

Figure 1. Schematic comparison of (a) the convex hull of an ensemble, and, (b) its bounding box. Capturing a target requires the target lie within the convex hull or bounding box.

The bounding box of a forecast ensemble can be thought of as providing simultaneous *prediction intervals* (Chatfield, 2001) for all variables, where a prediction interval is a forecasted range in which the target is *expected* to fall with a given probability. Bounding boxes are complementary to *rank histograms* (Anderson, 1996; Hamill, 2001; Talagrand et al., 1997). Traditional rank histograms ("Talagrand Diagrams") evaluate uni-variate forecasts (that is, one-dimension), while computations with the minimum spanning tree of an ensemble provides a rank order generalization to multivariate data (Smith, 2000; Smith and Hansen, 2004). Beyond dimension one, the bounding box statistics are quite different than those of rank histograms, which are primarily diagnostic, being more informative about the details of ensemble distributions whereas the bounding box considers volumes in state space directly and has design, diagnostic, and interpretive uses.

Despite the simplicity of the idea of bounding boxes capturing a target, it allows us to make a statement about the simultaneous properties of a large number of variables, and hence provide useful information

[‡]One should appreciate that figure 1 is a misleading representation of the actual situation for high dimensional state spaces with small ensembles; it would be more accurate to have only shown two points. Bounding boxes are not invariant under all change of coordinates, for example, rotation of axes. Bounding spheres are invariant under rotation, but not scaling. Bounding spheres are also significantly harder to compute. Non-invariance of bounding boxes is not necessarily a disadvantage, because many useful statistics are not invariant under change of coordinates.

Copyright © 0000 Royal Meteorological Society

that cannot be obtained from available marginal distributions, which can only make statements about for a few variables at a time, or rank histograms, which focus on the entire distribution. Other ideas and methods similar to bounding boxes have preceded us (Atger, 1999).

Our claim is not that examining bounding boxes provides a superior forecast evaluation to other techniques (rank histograms, probability skill scores, the relative operating characteristic and so forth). We demonstrate, rather, that the statistics extracted from bounding boxes provide different information. Statistics derived from bounding boxes take advantage of the finite-sample nature of ensemble prediction systems, and do not require the assumption of any explicit probability density. This information has application in three distinct areas: (1) the design of proposed ensemble prediction systems, (2) the evaluation of operational ensemble systems, and (3) decision-support for users of operational ensemble products. As an example of (1), we examine what the expected properties of bounding boxes can tell us about ensembles and probabilistic forecasts. In particular, we ask how large an ensemble needs to be to capture the target under various assumptions regarding the quality of the ensemble prediction system. We consider briefly the relationship between rank histograms and bounding boxes in section 5; a collection of useful numerical results are provided in section 6. As an instance of (2) we illustrate a number of ideas using data from the DEMETER experiment data (see Weisheimer et al. (2005)). We also include an application of bounding boxes to verify some recent theoretical assertions about how to make ensemble forecasts with imperfect models. We return to (3) briefly in the conclusions, and note the utility of bounding box statistics of operational ensemble systems.

2 Constructing ensembles

A number of different techniques are currently in use for constructing ensembles for forecast applications. For example, scalar multiples of the most significant singular and bred vectors are used in operational NWP (Molteni et al., 1996; Toth and Kalnay, 1997), however, this method of constructing an ensemble is not exactly what we have in mind here. Ensembles of singular and bred vectors are not necessarily trying to represent a probabilistic forecast; they may be viewed as trying to bound, or quantify, the maximum error growth. This is reasonable and compatible with the idea of trying to capture truth (or a target state) in the senses discussed. In fact, the technique used by ECMWF operational medium-range weather forecasting of taking positive and negative multiples of the singular vectors would capture the truth when the initial state is well placed and error growth was well behaved. These methods of constructing ensembles may not be consistent with the analysis that follows.

In the following analysis we will assume that the ensemble is constructed by selecting members randomly according to some probability density[§], which implies that the ensemble can be made arbitrarily large. Ideally, the probability density would represent exactly the probability distribution of the target state at a particular forecast time, given all the uncertainties of observations and model error. In practice, one would use an approximation to this, for example, one might make random perturbations (Errico and Baumhefner, 1987; Tribbia and Baumhefner, 2003) of an analysis, or other initial state, and then evolve this ensemble forward to the required forecast time, which is equivalent to selecting the ensemble members according to some probability density at the forecast time. A suitable process for constructing an ensemble from a collection of singular or bred vectors, would be to take random linear combinations of them, with the random coefficients generated according to some density, and then evolving these perturbations of the initial state forward to the required forecast time. It is not clear whether the following results on the size of an ensemble apply to ensembles as currently constructed from singular and bred vectors. The kind of ensembles the authors are most interested in are ensembles drawn from sets of indistinguishable states (Judd and Smith, 2001, 2004). These are compatible with ensembles obtained using Bayesian methods, such as a particle filter (Del Moral, 1995), only they are obtained more efficiently and can be obtained from imperfect models.

More formally, when we speak of constructing an ensemble, we are thinking of a process like the following. When dealing with a deterministic model, one might start from an initial ensemble, then evolve this forward to the forecast time, that is, if $\mathcal{E}_0$ is the initial ensemble, then at the forecast time t the ensemble is $\mathcal{E}_t = \{\phi_t(x_0) \colon x_0 \in \mathcal{E}_0\}$, where ϕ_t is the evolution operator over the time interval t. In a discrete time model one can define recursively $\mathcal{E}_t = \{f(x_{t-1}) \colon x_{t-1} \in \mathcal{E}_{t-1}\}$, where f is the model. It should be noted, however, that these methods are fully justified only when the model is perfect. If the model is imperfect, then forecasts can often be improved by taking model error into account explicitly (and are degraded by treating the model as perfect, see Judd and Smith (2004)). One way to account for model error is to assume the unknown model errors are random with some appropriate distribution η, then evolve the initial ensemble according to a stochastic evolution operator, even though the model is deterministic (Judd and Smith, 2004). For example, for the discrete time model $\mathcal{E}_t = \{f(x_{t-1}) + \epsilon \colon x_{t-1} \in \mathcal{E}_{t-1}, \epsilon \sim \eta\}$. In section 6 we investigate the performance of these ensembles for perfect and imperfect models.

[§]That is, we are assuming the probability measure is absolutely continuous.

Copyright © 0000 Royal Meteorological Society

3 The probability of capturing a target

There is considerable flexibility in what a target might be. The target might be a collection of future observations, for example, specific station temperatures, precipitation over a specified region and period, the 500mb height, an aerosol concentration, and so on. Note that observed quantities will have measurement errors, so the target is a random variable with some distribution, and this error distribution is typically well known; it being a property of the measurement instrument. The target may also be a future analysis, that is, a model state derived from observations using a data assimilation technique, which is essentially a projection of a history of observations into model space. Such a target is also a random variable, because the observations have measurement errors, but there is the additional complication that the state is under determined, so that the distribution of the target state variables is not simply related to measurement errors[¶]. Mathematically there is no distinction between a target being a set of observations or a verifying analysis.

In the following the word *target* almost always refers to the realization of a random variable (as determined by the observations) except in situations where we refer to the "distribution of the target" or the "expected value of the target", where the random variable associated with the target is meant.

In the following let $\mathcal{E} \subset \mathrm{R}^d$ be a finite ensemble of states, randomly drawn from some probability density ξ. Ideally ξ represents the probability distribution of the target $x^\star \in \mathrm{R}^d$, but usually it is only an approximation of this, and possibly a poor approximation[‖]. Typically, the ensemble is constructed from states selected at $t = 0$, then evolved to the forecast lead time $t = T$. The distribution ξ and the ensemble $\mathcal{E}$ should be understood to be defined at $t = T$ and our discussion refers to their properties at this lead time. At some subsequent time $t \geq T$ the actual target is determined. It should be noted that one cannot necessarily know all the detailed properties of ξ and $\mathcal{E}$, because the properties of ξ depend on how the ensemble is constructed and the evolution from $t = 0$ to $t = T$.

We introduce the notation $x \in bb(\mathcal{E})$ to indicate that x is contained in the bounding box of $\mathcal{E}$, and $|\mathcal{E}|$ to denote the number of members of the ensemble. It is too much to ask that the target state *always* lies within the bounding box of the ensemble, however, it is reasonable to ask for the probability of the target lying in the bounding box, because the ensemble is constructed randomly. One major convenience of considering the bounding boxes is that each coordinate can be considered separately. Hence, consider initially when the dimension of the state space $d = 1$. Let

$$p = \Pr(y < x^\star \colon y \sim \xi), \tag{3}$$

that is, p is the probability a y, randomly selected according to ξ, is to the left of (less than) $x^\star$. For the ensemble to capture the target, then some ensemble members must be to the left of (less than) the target, and some must be to the right of (greater than) the target. The only ways a randomly ξ-generated ensemble of size n can fail to capture the target are either all ensemble members are to the left of the target, which happens with probability p^n, or all ensemble members are to the right of the target, which happens with probability $(1-p)^n$. Hence, the probability that the bounding box of an ensemble of size n captures the target is,

$$\Pr(x^\star \in bb(\mathcal{E}) \colon d = 1, |\mathcal{E}| = n) = 1 - p^n - (1-p)^n, \tag{4}$$

that is, the ensemble must be neither entirely to the left of the target nor entirely to the right of the target.

When $d > 1$, let $p_i = \Pr(y_i < x_i^\star \colon y \sim \xi)$, where i refers to the coordinate. If the distributions of each coordinate are independent, then similar to (4) one obtains,

$$\Pr(x^\star \in bb(\mathcal{E}) \colon \text{independent coordinates}, |\mathcal{E}| = n) = \prod_{i=1}^{d} (1 - p_i^n - (1-p_i)^n). \tag{5}$$

We will say that the distribution of the ensemble is *unbiased* when $p_i = 1/2$ for all i (regardless of whether or not components are independent), that is, the target lies at the median of the ensemble's distribution ξ in each component[**]. In this case, one can derive from (5) that,

$$\Pr(x^\star \in bb(\mathcal{E}) \colon \text{unbiased}, |\mathcal{E}| = n) = \left(1 - \frac{1}{2^{n-1}}\right)^d. \tag{6}$$

If the ensemble is drawn from independent Gaussian distributions for each coordinate $N(x_i^\star + \beta_i, \sigma_i)$, $i = 1, \ldots, d$, for any d, then $p_i = \Phi(-z_i)$, where $z_i =$

[¶]The distribution of a target in this case is not necessarily what is often called analysis error, because the target analysis may use observations much further into the future than the verification time Judd and Smith (2001); Ridout and Judd (2001).

[‖]It is sufficient for our purposes to require that ξ is chosen so that a sufficiently large random sample $\mathcal{E}$ should capture the target state $x^\star$. That is, the probability density evaluated at the target is not zero.

[**]Amongst statisticians unbiased is usually taken to mean the expected mean of the ensemble is the target, where as, our definition requires the expected *median* to be the target. When the distribution ξ is symmetric the expected mean and median coincide, thus for the Gaussian distributions considered below there is no difference from the most common statistical meaning of unbiased. It should also be noted that one can not necessarily known at where the medians will lie or what states at $t = 0$ will evolve to have components corresponding to the median of ξ at $t = T$.

Copyright © 0000 Royal Meteorological Society

$|\beta_i|/\sigma_i$ and Φ is the cumulative probability of the standard normal density. It follows that,

$$\Pr(x^\star \in bb(\mathcal{E})\colon \text{Gaussian}, |\mathcal{E}| = n) = \prod_{i=1}^{d} (1 - \Phi(-z_i)^n - \Phi(z_i)^n)\,. \quad (7)$$

Throughout this paper, we use the term *bias* to indicate a statistical property of the distribution of ensemble members. Specifically, bias indicates that the median (or mean) of the ensemble is consistently different from that of the expected target location[††]. This bias may reflect *state-dependent systematic model error* (Smith, 2000; Orrell et al., 2001), even if such errors "average to zero" in some global sense. Bias has sources other than model error, for instance it can result from errors in ensemble formation, for example, by centering the ensemble distribution on a best guess analysis that happens to be sub-optimal or in error.

4 Minimum ensembles sizes

Knowing $\Pr(x^\star \in bb(\mathcal{E})\colon |\mathcal{E}| = n)$ allows one to estimate a suitable size for an ensemble, because if one wants to capture the target with probability α, then one should adjust[‡‡] n until $\Pr(x^\star \in bb(\mathcal{E})\colon |\mathcal{E}| = n) > \alpha$. The probabilities derived in the previous section may sometimes serve as lower bounds on $\Pr(x^\star \in bb(\mathcal{E})\colon |\mathcal{E}| = n)$, as indicated in the following sections. These bounds are useful; although using a lower bound to compute the ensemble size will over estimate the required ensemble size, it guarantees an ensemble of the computed size will capture the target with at least the probability α. Also note that assuming coordinates are independent when they are not, will also over estimate the size of the ensemble, because dependence means bounding one coordinate tends also to bound the dependent coordinates, and hence the effective dimension is less than d.

Unbiased ensembles: If the ensemble was drawn from a density that is unbiased, then one obtains from (6),

$$n > 1 - \frac{\log\left(1 - \alpha^{1/d}\right)}{\log(2)}, \quad (8)$$

that is, an ensemble of size n or larger can be expected to capture the target with a probability of at least α. Table I shows the lower bound on n for $\alpha = 0.95$ and various d, assuming coordinates are independent. When d is large one can use a series approximation in the inequality (8) giving,

$$n > 1 + \frac{1}{\log(2)}\left(\log(d) - \log(-\log(\alpha)) - \frac{\log(\alpha)}{2d} - \frac{1}{24}\left(\frac{\log(\alpha)}{d}\right)^2 - O(1/d)^4\right), \quad (9)$$

and hence n grows only as fast as $\log(d)$.

Table I. Minimum ensemble sizes in various dimensions for unbiased ensembles with a confidence level $\alpha = 0.95$.

d =	1	10	10^2	10^3	10^4	10^5	10^6	10^7	10^8
n ≥	6	9	12	16	19	22	26	29	32

It might at first be surprising that a modestly sized ensemble will capture the target in a state space with millions of independent dimensions , but on second thoughts it is not surprising. When the ensemble is drawn from an unbiased density, one expects nearly equal numbers of ensembles members either side of the target and it therefore should not be surprising that the ensemble need not be large, because it is highly unlikely that all the randomly selected ensemble members should be to one side of the target. For comparison ECMWF currently employs a 51 member ensemble for its operational medium-range weather forecasts where $d \approx 10^7$. An ensemble of this size would be more than ample to capture a target analysis if the forecasts were unbiased.

The modest size of the ensemble required to capture the target as D increases does not contradict the exponential growth in ensemble size required to estimate the forecast probability density. The capture of a target is a much weaker requirement, and consequently can be achieved with much smaller ensembles.

Biased ensembles: A forecast ensemble can be biased for many reasons, for example: as a result of the initial state that is perturbed to obtain the ensemble having inevitable errors due to limited and inaccurate observations; or as a result of the model dynamics being imperfect and moving all ensemble members away from the evolving target.

When the ensemble is biased, the bound on the size of the ensemble are less encouraging. Bias means that for some coordinates the ensemble members are more likely to be on one particular side of the target, and so there is an increased chance the bounding box will fail to capture the target. Consider the simplest situation where coordinates are independent and all have the same bias, that is, in (5) $p_i = p \neq 1/2$ for all i. Table II illustrates how the ensembles size varies with dimension d and p. Note that minimum ensemble sizes in situations where some coordinates are biased and some are not, or where different coordinates have different bias, can all be obtained by numerical solution of (5). On the other hand, a worst case estimate of

[††]The term bias is used variously in the meteorological literature, most frequently in the very specific sense of a fixed, global, state independent error. Statisticians apply the term to any distribution, not just global time invariant ones.

[‡‡]The choice $\alpha = 0.95$ is common; although one might not necessarily be able to make an ensemble with given capture probability α, but one can estimate α for a given ensemble prediction system. This is in itself a useful characterization of an ensemble for both users and in model development.

Copyright © 0000 Royal Meteorological Society

the ensemble size can be obtained from table II by taking $p = \min_i p_i$. Observe from table II how the size of the ensemble increases rapidly once bias p exceeds about $2^{-4} = 0.0625$. For a Gaussian distribution this corresponds to a shift of the mean away from the target by more than 1.53 standard deviations.

Table II. Minimum ensemble sizes in various dimensions given the probability p that truth lies to the left of every coordinate of every ensemble member. The confidence level is $\alpha = 0.95$.

	p =	2^{-1}	2^{-2}	2^{-4}	2^{-6}	2^{-8}
d=1	$n \geq$	6	11	47	191	766
d=10	$n \geq$	9	19	82	335	1348
d=10^2	$n \geq$	12	27	118	482	1936
d=10^3	$n \geq$	16	35	154	628	2524
d=10^4	$n \geq$	19	43	189	774	3113
d=10^5	$n \geq$	22	51	225	920	3701
d=10^6	$n \geq$	26	59	261	1066	4289
d=10^7	$n \geq$	29	67	296	1213	4878
d=10^8	$n \geq$	32	75	332	1359	5466

Gaussian ensembles: Expressing bias in terms of a probability is a little inconvenient, and one can obtain slightly more useful intuition by expressing bias as a shift of the median relative to the spread of the ensemble, which is easily done if the ensemble is drawn from a symmetric density, like a Gaussian. Suppose for the coordinate i that the ensemble coordinates have a distribution $N(x_i^\star + \beta_i, \sigma_i)$, that is, β_i is the bias and σ_i is the spread of the ensemble in this coordinate. It is convenient to work in terms of the *normalized* bias $z_i = |\beta_i|/\sigma_i$. Table III lists numerically computed minimum ensemble size n when $\alpha = 0.95$ for a fixed number of independent coordinates and various amounts of normalized bias where all coordinates have identical normalized bias $z_i = z$, or alternatively, the worst case estimate where $z = \max_i z_i$. It is clear from table III that once the bias exceeds more than one standard deviation the minimum ensemble size grows rapidly. For example, assuming the ECMWF operational medium-range weather forecasting 50 member ensembles are Gaussian, then they ought to be able to cope with bias on all 10^7 variables of up to 0.5 of a standard deviation, but not one standard deviation. If only one specific variable is required to be captured, then the ensemble can cope with a bias of more than 1.5 standard deviations. As another example, consider that it is common practice to optimize the initial ensemble spread so that at forecast time the ensemble spread matches the spread of the best-guess forecast error (Reynolds and Rosmond, 2003). If this is the case, then one should expect random ensemble bias, which 95% of the time will be less than about 1.5 standard deviations. This implies for an operational model of 10^6 variables, that ensembles of around 250 members ought to be able to ensure capture of all 10^6 variables of the target 95% of the time.

Figure 2 shows the probability of capturing the target for Gaussian ensembles with various numbers of independent coordinates and normalized bias z. It is clear from these figures that there is a very rapid transition from capturing the target to not capturing the target, as the bias or dimension is increased for a given size of ensemble. The sharpness of this transition is useful, because it could be used to measure a bias. For example, figure 2 shows that the bias threshold for consistent capture increases with ensemble size, so the ensemble size that just succeeds to consistently capture can provide an indication of the bias.

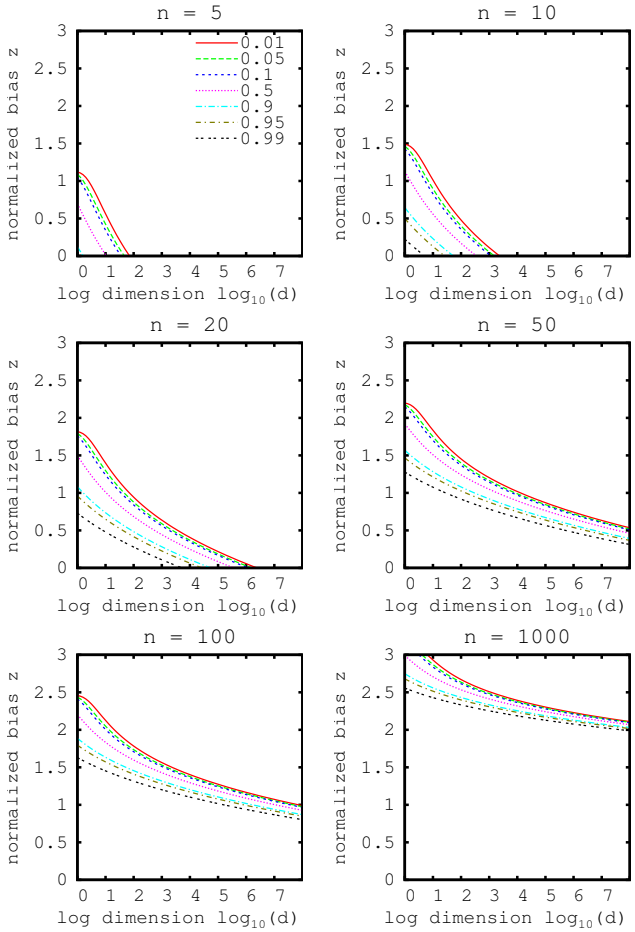

Figure 2. The probability contours for Gaussian ensembles of size n capturing the target in d independent coordinates (dimension) and various amounts of normalized bias $z = |\beta|/\sigma$. Small ensembles may not achieve capture at high probabilities. For large ensembles the transition from almost certain capture, to almost certain non-capture is very narrow.

5 The effect of analysis error

We now come to the results with potentially the most interest and significance to NWP. Up until now we have treated ensemble bias as being fixed, but suppose the ensemble is centred on some *primary forecast*, for example, the forecast from the current *best guess* state,

Copyright © 0000 Royal Meteorological Society

Table III. Minimum ensemble sizes in various dimensions for Gaussian ensembles with a confidence level $\alpha = 0.95$. The size of the ensemble is strongly dependent on the normalized bias of the ensemble $z = |\beta|/\sigma$. Here we assume the worst case where all coordinates have the same normalized bias.

	z =	0	0.2	0.5	1	1.5	2	2.5	3
d = 1	$n \geq$	6	6	9	18	44	131	481	2218
d = 10	$n \geq$	9	10	15	31	77	230	847	3906
d = 10^2	$n \geq$	12	14	21	44	110	330	1217	5609
d = 10^3	$n \geq$	16	19	27	58	143	430	1586	7313
d = 10^4	$n \geq$	19	23	34	71	177	530	1956	9018
d = 10^5	$n \geq$	22	27	40	84	210	630	2326	10722
d = 10^6	$n \geq$	26	31	46	98	243	730	2695	12427
d = 10^7	$n \geq$	29	35	52	111	277	830	3065	14131
d = 10^8	$n \geq$	32	40	58	124	310	930	3435	15836

or *analysis*. In reality this primary forecast will not exactly forecast the target, it will have some random error, even if the primary forecast is an unbiased estimate of the target. If the primary forecast has an error β relative to the target, then this means the ensemble centred on it has a bias β. Hence, the probability of capturing the target with an ensemble centred on the primary forecast is

$$\Pr(x^\star \in bb(\mathcal{E})| \text{ centred on primary forecast }) = \int \Pr(x^\star \in bb(\mathcal{E})|\beta)\rho(\beta)d\beta, \quad (10)$$

where $\rho(\beta)$ is the probability density of the primary forecast error β, which is a joint distribution of a d-dimensional bias vector. In general the properties of $\rho(\beta)$ are unknown, so we must again consider plausible models of this density to obtain bounds on possible outcomes.

One extreme model for $\rho(\beta)$ is to assume each component is independently distributed as $\rho_i(\beta_i)$. The probability of capturing the target by an ensemble of size n is the product of the probabilities

$$q_i = \int \left(1 - p_i(\zeta)^n - (1 - p_i(\zeta))^n\right) \rho_i(\zeta)d\zeta, \quad (11)$$

where $p_i(\zeta)$ is the probability the i-th component of an ensemble member is to the left of the target when the ensemble is centred on a primary forecast with bias ζ in its i-th component. Suppose a component of the primary forecast's error has a Gaussian distribution $N(\beta', \sigma'^2)$, where β' is the expected error in the primary forecast, and σ' the expected spread of the this error. Also suppose the ensemble is drawn from a distribution $N(\beta, \sigma^2)$. Table IV shows the computed minimum ensemble sizes for 95% capture probability for various d and various values for the key ratios β'/σ (the ratio of systematic primary forecast error to ensemble spread) and σ'/σ (the ratio of primary forecast error spread to the ensemble spread). If $\beta = \beta'$ and $\sigma'/\sigma = 1$, then the ensemble has the same distribution as the primary forecast.

The minimum ensemble sizes shown in table IV are too large to be feasible, and a moment's thought should reveal why. When each component is independent, then in a 10^6 dimensional space some components will have a bias far out in the tails of ρ_i. Consequently, a random sample from the ensemble distribution must be enormous to guarantee capture of this extreme. Either we must assume the bias of each component are not independent, or we must accept something less severe than capture of all components, for example, capture of 95% of components 95% of the time. We will consider both of these cases.

Dependent bias: Now consider the case where components of the bias β are dependent. In this case, capture of the component with the worst bias implies capture of all components, and, hence, we need only consider the situation where all components have the same bias. Suppose a Gaussian ensemble centred on the primary forecast is constructed so that the ensemble is drawn from the distribution $N(\beta, \sigma^2)$, where β is distributed as $N(\beta', \sigma'^2)$ for all variables. From equations (7) and (10) we have under these assumptions the following probability of capture of an ensemble of size n for d independent coordinates

$$\begin{aligned}\Pr(x^\star \in bb(\mathcal{E})|\beta', \sigma', \sigma, d, |\mathcal{E}| = n) &= \int \left(1 - \Phi(\beta/\sigma)^n - \Phi(-\beta/\sigma)^n\right)^d \\ &\quad \times \frac{1}{\sqrt{2\pi}\sigma'} e^{-\frac{1}{2}(\beta-\beta')^2/\sigma'^2} d\beta \\ &= \int \left(1 - \Phi(z)^n - \Phi(-z)^n\right)^d \\ &\quad \times \frac{1}{\sqrt{2\pi}(\sigma'/\sigma)} e^{-\frac{1}{2}(z-(\beta'/\sigma))^2/(\sigma'/\sigma)^2} dz\end{aligned} \quad (12)$$

Table V shows minimum ensemble sizes to achieve 95% probability of capture for various d and various values for the key ratios β'/σ (the ratio of systematic primary forecast error to ensemble spread) and σ'/σ (the ratio of primary forecast error spread to the ensemble spread). These ensemble sizes are feasible. The most important thing to notice is that random bias of the primary forecast significantly increases the minimum size of the ensemble, even when the primary forecast is unbiased ($\beta' = 0$).

Copyright © 0000 Royal Meteorological Society

Table IV. Independent bias: Minimum ensemble sizes for 95% capture assuming each variable has an independent random bias. Here each ensemble component is Gaussian $N(\beta, \sigma^2)$, where each bias β has a Gaussian distribution $N(\beta', \sigma'^2)$.

$\sigma'/\sigma = 0.8$	$\beta'/\sigma =$	0	0.5	1.0	1.5
$d = 1$	$n \geq$	20	31	93	368
$d = 10$	$n \geq$	95	208	903	5074
$d = 10^2$	$n \geq$	444	1274	7326	54229
$d = 10^4$	$n \geq$	9216	41029	367483	4222135
$d = 10^6$	$n \geq$	186033	1185158	15183623	249405438
$\sigma'/\sigma = 1.0$	$\beta'/\sigma =$	0	0.5	1.0	1.5
$d = 1$	$n \geq$	39	64	210	979
$d = 10$	$n \geq$	390	909	4943	35887
$d = 10^2$	$n \geq$	3900	12527	98427	1011764
$d = 10^4$	$n \geq$	389915	2138232	29595390	528514685
$d = 10^6$	$n \geq$	38991457	332930114	7242331687	202493201376
$\sigma'/\sigma = 1.2$	$\beta'/\sigma =$	0	0.5	1.0	1.5
$d = 1$	$n \geq$	95	154	548	2958
$d = 10$	$n \geq$	2252	5464	36230	338440
$d = 10^2$	$n \geq$	57222	199324	2093488	29808737
$d = 10^4$	$n \geq$	39116396	252891781	5342388040	148066679273
$d = 10^6$	$n \geq$	27617643176	301086929520	11045606005357	533114080047573

Table V. Dependent bias: Minimum ensemble sizes for 95% capture assuming the maximum bias of all variables has Gaussian distribution. Here the ensemble is Gaussian $N(\beta, \sigma^2)$, where the bias β has a Gaussian distribution $N(\beta', \sigma'^2)$.

$\sigma'/\sigma = 0.8$	$\beta'/\sigma =$	0	0.5	1.0	1.5
$d = 1$	$n \geq$	27	43	134	575
$d = 10$	$n \geq$	74	126	426	1916
$d = 10^2$	$n \geq$	129	223	764	3461
$d = 10^4$	$n \geq$	242	420	1448	6576
$d = 10^6$	$n \geq$	355	618	2132	9684
$d = 10^8$	$n \geq$	468	815	2815	12790
$\sigma'/\sigma = 1.0$	$\beta'/\sigma =$	0	0.5	1.0	1.5
$d = 1$	$n \geq$	39	64	210	979
$d = 10$	$n \geq$	116	197	698	3396
$d = 10^2$	$n \geq$	205	352	1260	6164
$d = 10^4$	$n \geq$	387	666	2393	11734
$d = 10^6$	$n \geq$	569	980	3524	17288
$d = 10^8$	$n \geq$	751	1294	4654	22836
$\sigma'/\sigma = 1.2$	$\beta'/\sigma =$	0	0.5	1.0	1.5
$d = 1$	$n \geq$	59	95	327	1634
$d = 10$	$n \geq$	181	306	1122	5845
$d = 10^2$	$n \geq$	323	550	2033	10648
$d = 10^4$	$n \geq$	611	1044	3868	20296
$d = 10^6$	$n \geq$	898	1536	5698	29911
$d = 10^8$	$n \geq$	1186	2028	7526	39514

Observe from table V the effect of the ratio of primary forecast error spread to ensemble spread (σ'/σ). In techniques like Ensemble Kalman filtering (Evensen, 1994; Anderson, 1999; Bishop et al., 2001; Hamil and Synder, 2000), the ensemble spread is tuned so that $\sigma'/\sigma = 1$. If current operational NWP were perfect and ensemble formation introduced no systematic procedural bias, so that $\beta' = 0$, then capture of the target with 95% probability would require 500–700 member ensembles. Furthermore, if the models, or procedures, were imperfect and had even small bias ($\beta'/\sigma = 0.5$), then the ensemble size almost doubles. Note also that smaller ensemble sizes can be used, if one is prepared to give the ensemble a larger spread than the actual primary forecast error ($\sigma'/\sigma < 1$), that is, the likelihood of capturing the target is increased by intentionally making the initial ensemble too wide. On the other hand, if the primary forecast error is under-estimated (under-dispersive), so that the ensemble spread is just 20% less than the primary forecast error spread ($\sigma'/\sigma = 1.2$), then minimum ensemble size almost doubles. Note on the other hand, that if one aims only to capture one specific variable, then a 50 member ensemble is good even when an unbiased ensemble is too narrow ($\sigma'/\sigma = 1.2$).

We conclude with a simple analysis to show that the increase in necessary ensemble size is primarily the result of the occasional large error of the primary forecast, that is, the larger ensemble is needed as a safe guard. First note that the effect of primary forecast error could have been roughly deduced from the tables of fixed bias, for example, table III, by the following argument. If the ensemble has the same spread as the primary forecast error, then 95% of primary forecast errors are less than 1.64 standard deviations, therefore, to capture this amount of random bias of the primary forecast requires an ensemble size somewhere between the $z = 1.5$ and $z = 2.0$ columns of table III, and

Copyright © 0000 Royal Meteorological Society

a linear interpolation gives approximately the same values as the $\beta'/\sigma = 0$ column of the $\sigma'/\sigma = 1.0$ subtable of table V. The implication is that the much larger ensemble size is required to capture the occasional large primary forecast error, where as from table III one can deduce that if primary forecast errors are unbiased and not under estimated, then ensembles of 300 members should capture 95% of the time. Hence, one can conclude most the cost of ensemble forecasting is in capturing *extreme errors*, that is, statistically rare situations where large initial perturbations are required to cope with large primary forecast errors.

Limited capture: As we have seen, complete capture of all components results in larger ensemble sizes principally to capture *extreme errors*, as defined at the end of the previous paragraph. An alternative to complete capture is to require capture a fraction α' of components a fraction α of the time. Such limited bounding boxes have interesting properties.

Again assuming the worst case where all bias components are independent, we can compute the minimum ensemble size as follows. Let $\mathcal{P}$ be the power set of the integers 1 to d. Then $K \in \mathcal{P}$ is a subset of these integers. Let $|K|$ denote the size of this set. Given q_i the probability in the component i, see equation (11), then to capture a fraction α' of components a fraction α of the time requires that

$$\sum_{\substack{K\in\mathcal{P}\\ |K|>\alpha' d}} \prod_{i\in K} q_i \times \prod_{j\notin K} (1-q_j) > \alpha. \tag{13}$$

If we deem all components as equally important, then all components should have the same capture probability, that is $q_i = q$ for all i. In which case condition (13) becomes

$$\sum_{k>\alpha' d} \binom{d}{k} q^k (1-q)^{d-k} > \alpha. \tag{14}$$

Table VI shows minimum ensemble sizes to capture 95% of components 95% of the time. A curious property here is the ensemble size increases as d increases, peaks, then decreases asymptotically to the $d = 1$ values. This phenomenon occurs because the bias components are assumed independent, and hence the distribution of a large number of components is equivalent to large sample from the one-dimensional distribution. The initial increase in ensemble size accounts for variance of the sampling, that is, there is a greater chance of failing capture a component when $d = 2$, than for $d = 1$, but for d very large one expects to fail to capture almost exactly 95% of components with little variance.

One should be clear about the assumptions used to construct table VI, in particular one cannot conclude from table VI that small ensembles are sufficient. The main problem is that when only 95% of components are captured, there is no guarantee that the most important components are captured. In practice one should expect that bias components are not independent, and so it could happen the 5% of components not captured may be the most important 5% of components. Consequently, in practice one should anticipate that the efficient ensemble size will be somewhere between the minimum sizes implied by tables V and VI

Table VI. Limited capture: Minimum ensemble sizes for capture of 95% of variables 95% of the time assuming all variables have independent random bias. Here the ensemble is Gaussian $N(\beta, \sigma^2)$, where the bias β has a Gaussian distribution $N(\beta', \sigma'^2)$.

$\sigma'/\sigma = 0.8$	$\beta'/\sigma =$	0	0.5	1.0	1.5
$d = 1$	$n \geq$	20	31	93	368
$d = 10$	$n \geq$	95	208	903	5074
$d = 10^2$	$n \geq$	37	68	241	1116
$d = 10^4$	$n \geq$	21	33	100	404
$d = 10^6$	$n \geq$	20	31	93	372
$\sigma'/\sigma = 1.0$	$\beta'/\sigma =$	0	0.5	1.0	1.5
$d = 1$	$n \geq$	39	64	210	979
$d = 10$	$n \geq$	390	909	4943	35887
$d = 10^2$	$n \geq$	100	188	779	4415
$d = 10^4$	$n \geq$	43	69	234	1107
$d = 10^6$	$n \geq$	40	64	213	991
$\sigma'/\sigma = 1.2$	$\beta'/\sigma =$	0	0.5	1.0	1.5
$d = 1$	$n \geq$	95	154	548	2958
$d = 10$	$n \geq$	2252	5464	36230	338440
$d = 10^2$	$n \geq$	338	650	3065	21146
$d = 10^4$	$n \geq$	105	172	630	3472
$d = 10^6$	$n \geq$	95	155	556	3005

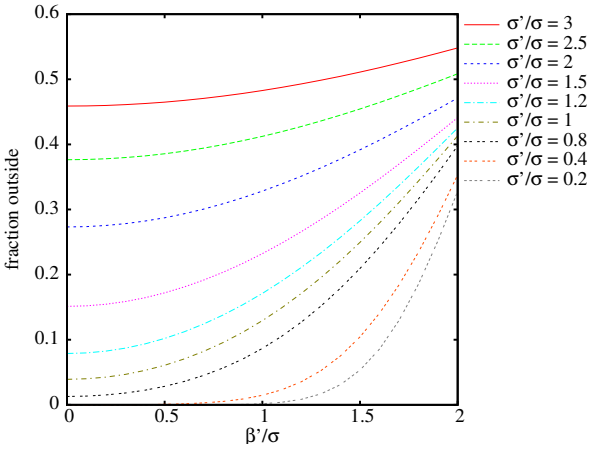

Figure 3. Expected number of components of a 50 member ensemble not captured assuming a independent random bias on each component. Various relative bias and ensemble spreads are shown.

Figure 3 shows an alternative view of limited capture, where we compute the expected maximum

Copyright © 0000 Royal Meteorological Society

fraction of components not captured of a 50 member ensemble. The fraction of components not captured does not provide definitive information about the properties of the ensemble. Figure 3 shows that if the spread is too narrow ($\sigma'/\sigma > 1$) then the fraction captured increasing rapidly for even small bias.

Rank histograms: It is useful to observe that $d = 1$ lines of tables IV, V and VI can be interpreted as statements about rank histograms, because when $d = 1$ capturing the target in the bounding box of an ensemble is equivalent to saying that the target does not lie in the extreme bins of a rank histogram of target and ensemble. For example, when $d = 1$, $\sigma'/\sigma = 1.0$, and $\beta'/\sigma = 0$, the target should be statistically indistinguishable from the ensemble members, and so the target is outside the bounding box, or equivalently, in the extreme bins of the rank histogram, with probability $2/(|\mathcal{E}| + 1)$, implying that capturing truth with at least probability α requires $|\mathcal{E}| \geq (2/\alpha) - 1$, which is well-known. Similarly, varying β'/σ shows the effect of forecast bias on the rank histogram, where as $\sigma'/\sigma < 1$ shows the effect of over-dispersion of the ensemble and $\sigma'/\sigma > 1$ shows the effect of under-dispersion. All these effects are consistent with known properties of 1-dimensional rank histograms (see Hamill (2001)).

6 Applications

Bounding boxes can be used as a simple means to verify ensemble forecasting schemes, by testing whether an ensemble captures the target or not. This section has three subsections. In the first subsection we illustrate some of the predictions of the calculations and examine experimentally some related issues using a simple quasi-geostrophic climate model in a perfect model scenario. In the second subsection we use bounding boxes to verify experimentally a claim made elsewhere (Judd and Smith, 2004) about appropriate ensemble forecasting schemes for imperfect models. In the third subsection we make some qualitative observations of bounding boxes for the DEMETER multi-model ensemble experiments (Palmer et al., 2004)

6.1 Perfect model scenario

In this subsection we examine bounding boxes in a geophysical modelling situation where the model is perfect. In particular we verify that initial Gaussian ensembles capture truth as expected for various amounts and forms of bias, and also investigate how the number of coordinates that fail to be captured changes with increasing bias.

The system used in the tests is a spectral three-layer quasi-geostropic (QG) atmospheric model (Weisheimer et al., 2003), which has been constructed primarily to study ultra-low-frequency climate variability in a minimum-complexity model of the extra-tropical circulation. The model solves the QG potential vorticity equation for three layers of equal mass and uses simplified parameterizations of orographic, frictional and diabatic processes. The effect of diabatic heating is simulated using a relaxation toward a radiative-equilibrium state with a corresponding equator-to-pole temperature difference of 60K in the middle troposphere and 30K in the upper troposphere/lower stratosphere. Two non-axisymmetric major mountain ridges with a maximum height of 1200m simulate the orographic forcing. Sensitivity studies with different parameter sets revealed that the quantitative model behaviour is sensitive to the amplitudes of the thermal and orographic forcing. The atmospheric flow becomes quasi-stationary for weak forcing; whereas a more realistic intensity of the orographic and thermal forcing leads to a more irregular chaotic flow (Dethloff et al., 1998). In contrast, the impact of dissipation variations is not that crucial. An Ekman dissipation e-folding time of 16 days and a Newtonian cooling time scale of 27 days following Weisheimer et al. (2003) have been used in this study. The QG equations are solved with a horizontal resolution of T20, corresponding to 32×64 grid points or a 5.2×5.2 degree resolution. The Runge-Kutta integration time step was 1 hour, with the output every 24 hours. On the 32×64 grid there are 6144 coordinates in total, but these are not independent, if for no other reason than the T20 resolution model has only 1386 spectral coordinates. All the calculations and results use non-dimensionalized stream function coordinates. There is, of course, a mapping ϕ from spectral coordinates to stream function cordinates, which will be employed subsequently. The spectral coordinates of the T20 model have a roughly power-law distribution with wave-number. The maximum standard deviation of any spectral component is around 4×10^{-4}, which provides a useful scale for comparison to the earth's atmosphere and a useful scale for the size of perturbations we apply in generating ensembles. In the following we describe results for a single 50 day truth trajectory as calculated from an initial condition $x_0^\star$, which is the final day of a 1000 day spin-up from a random state where all components were close to zero.

The T20 model has 1386 coordinates, so for this model table III implies that a 50 member Gaussian ensemble should be able to handle a normalized bias of more than 0.5 but not more than 1.0. This verified for the initial ensemble. What we are interested in is, given an initial Gaussian ensemble that captures truth, whether capture of truth continues as the ensemble is evolved forward to longer lead times, because the ensemble is unlikely to remain Gaussian (Gilmour et al., 2001; Reynolds and Rosmond, 2003), and may become biased, because, for example, the mean of the ensemble is not necessarily the mean of the distribution it samples.

There are a number of potential methods for generation of an initial ensemble from perturbations of

Copyright © 0000 Royal Meteorological Society

one initial state. The perturbations could be independent perturbations of stream function coordinates or the spectral coordinates. The perturbations could be uniform across all coordinates or scaled, for example, by the square root of the variance or covariances of the observed coordinates or an appropriate error field, like analysis errors. Since our objective here is to illustrate the use of bounding boxes we consider just two simple perturbation schemes: (iS-type) Gaussian perturbation of stream function coordinates that are independent and identically distributed for each coordinate; (iZ-type) independent Gaussian perturbation of spectral coordinates scaled by the square root of the coordinate's long-term variance.

An iS-type ensemble is generated about a state x_0 by perturbing each of the 6144 stream function coordinates by an independent random sample from the distribution $N(0, \sigma^2)$. In our experiments the perturbations are on the order of $\sigma = 2 \times 10^{-5}$, which is approximately 5% of the maximum variation. To make an ensemble with normalized bias z, relative to x_0, the perturbations were drawn from a distribution $N(\pm z\sigma, \sigma^2)$, where the sign is chosen with equal likelihood and independently for each component.

An iZ-type ensemble about a state x_0 is constructed using the mapping ϕ from spectral coordinates to stream function coordinates. To compute an ensemble with specified bias the members are of the form $x_0 + z\phi(v) + \phi(w)$, where v and w are random samples from $N(0, \Sigma^2)$, with v fixed for all ensembles members and w different for each. The covariance matrix Σ was chosen to be diagonal and proportional to the observed spectral coordinate variances. The matrix Σ can be scaled so that the stream function coordinates have a specified standard deviation σ. In the following comparison of the spread of iS-type and iZ-type ensembles is always in terms of this σ of the stream function coordinates.

Figure 4 shows the fraction of the 6144 true stream function coordinates that are outside the bounding box of a 50 member iS-type ensemble with $\sigma = 2 \times 10^{-5}$ and normalized bias z for various values between 0.0 and 2.0. For $z \leq 0.4$ the ensemble is capturing truth as table III predicts. At around $z = 0.6$, the forecast ensembles begin to fail to capture truth at most time steps, although the number of coordinates that are not captured is rather small, no more than 20 of 6144 out to day 47. Once $z > 0.6$, the fraction of coordinates not captured increases roughly linearly with z at any fixed time, although the actual fraction varies with the time.

Figure 5 shows the fraction of the 6144 true stream function coordinates that are outside the bounding box of a 50 member iZ-type ensemble. Once again we see that for $z > 0.6$ the fraction of coordinates not captured increase roughly linearly with z at any fixed time, although the actual fraction varies with the time. It is curious that the iS-type ensemble captures better than the iZ-type ensemble from day 10 to day 40. This behaviour does not appear to be significantly effected

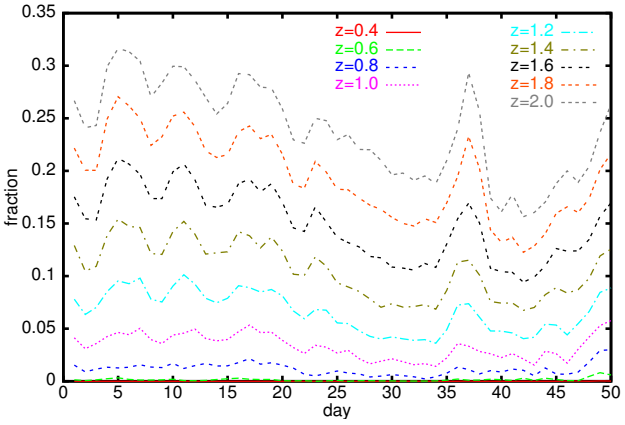

Figure 4. Effect of bias on a 50 member iS-type ensemble with an initial $\sigma = 2 \times 10^{-5}$. The graphs show the fraction of stream function coordinates not captured with forecast lead time given the specified bias. For a bias of $z = 0.4$ there is almost complete capture, specifically, at only 5 time steps did the ensemble fail to capture, and then only for one or two coordinates, and hence its graph is almost invisible on this plot.

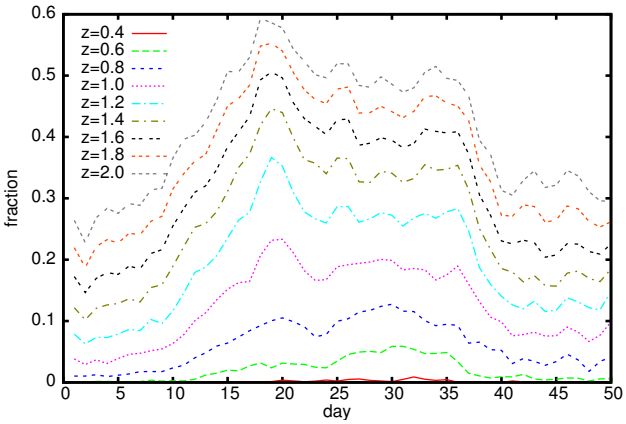

Figure 5. Effect of bias on a 50 member iZ-type ensembles with an initial $\sigma = 2 \times 10^{-5}$. The graphs show the fraction of stream function coordinates not captured with forecast lead time given the specified bias.

by the actual bias vector used. We have not investigated this phenomenon further, for example, to to see how the ensembles compare for different initial states x_0.

6.2 Imperfect model scenario

We now consider ensemble forecasting with imperfect models. It has been argued (Judd and Smith, 2004) that if a model is imperfect, then forecast ensembles might be better generated by a stochastic model, even though the system is deterministic. In this section we demonstrate the validity of this assertion in the QG model.

The experiments use the same QG model as the previous subsection. Here the T21 QG equations are used as the *system* and the T20 QG equations as the *model*, with the aim being to construct ensemble

Copyright © 0000 Royal Meteorological Society

forecasts of the T21 state using only an imperfect T20 model. Throughout, the quasi-geostrophic stream function on the grid points is used as the state, because then both the T21 system and the T20 model will have the same state space.

We will assume that initially we have been fortunate enough to have obtained the exact state of the stream function at every grid point. This is somewhat artificial, but suppressing observational error more clearly reveals the effect of model error. Later we will see that the effect of model error is so significant that ignoring model error will result in poor ensemble forecasts even when moderate observational error is present. Given the exact state, a T21 model would give a perfect forecast, but a forecast with the imperfect T20 model will diverge from the target T21 behaviour. When given the exact state x_0, then an iS-type or iZ-type with $z = 0$ will be unbiased, so a relatively small ensemble ought to capture the target at initialization time and for short-term forecasts, however, for longer term forecasts model error introduces a systematic divergence that degrades the ensemble's ability to capture the target T21 trajectory. We could attempt to accommodate for this bias (due by accumulated model error) by increasing the initial ensemble spread σ, but we will see that using a stochastic model to evolve the ensemble forecasts does better.

Our experiments begin by computing a 50 day T21 *target* trajectory $x^\star_t$, $t = 0, \ldots, 50$, that is, a trajectory of targets at one day intervals. The initial condition $x^\star_0$, was the final state of a 1000 day spin-up trajectory from a random near-zero state. For later comparison we also compute from $x^\star_0$ a 50 day T20 *control* trajectory.

We first investigated how well iS-type and iZ-type forecast ensembles capture the control and target trajectories. These 50 member ensembles were created as described in the previous sub-section with $x_0 = x^\star_0$ and $z = 0$. In the experiments we consider ensembles with a range of σ around 2×10^{-5}. We evolve the ensembles forward by computing the trajectory of each ensemble member using the T20 model.

All the iS-type and iZ-type ensembles, at all spreads examined, capture the T20 control trajectory out to around day 25. Ensembles with larger spread begin to fail, but still capture 99.8% of coordinates out to day 50. The failure for larger spreads could be occurring because these have a larger initial random bias (that is, the mean of the ensemble is not x_0), which is amplified as the ensemble evolves, or it could be a result of the larger spread being more strongly affected by nonlinear effects, which result in greater distortion of any finite ensemble's bounding box.

Figures 6 and 7 show how iS-type and iZ-type ensembles perform at capturing the T21 target trajectory for various σ. As might be expected, the ensembles with larger initial spread σ are doing better, but even the widest initial spreads we considered only captured the target for around 15 days. Comparing figures 6 and 7 we note that for a given fraction of coordinates captured the σ for the iS-type ensemble were around twice as large as those of the iZ-type ensemble. (This is

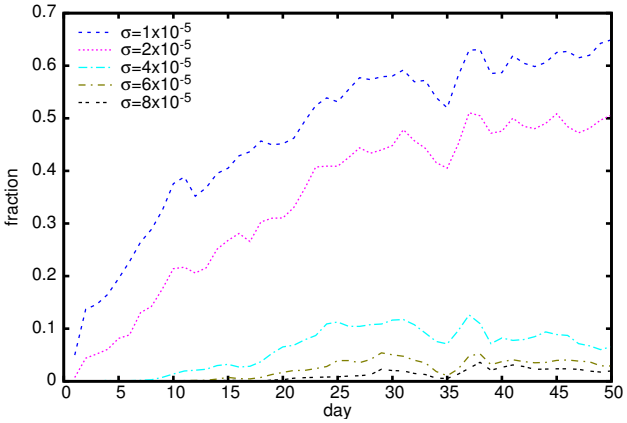

Figure 6. The fraction of T21 target coordinates outside a 50 member iS-type ensemble bounding box on each forecast day, for ensembles with different initial spreads σ.

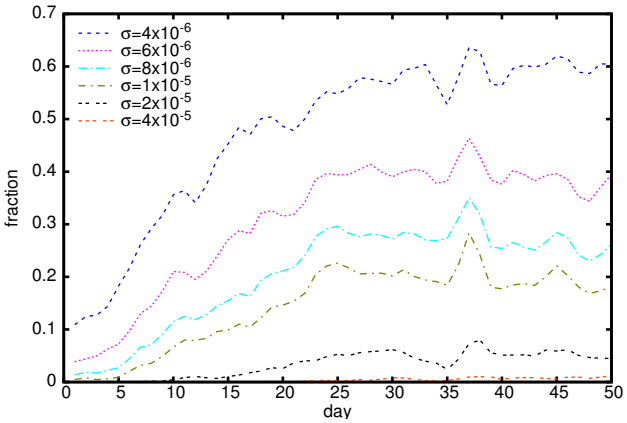

Figure 7. The fraction of T21 target coordinates outside a 50 member iZ-type ensemble bounding box on each forecast day, for ensembles with different initial spreads σ.

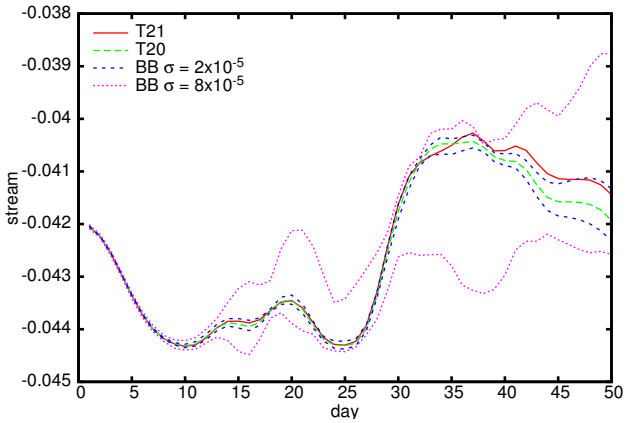

Figure 8. The evolution of a typical coordinate (lower layer mid-latitude) over the 50 day period for the T21 target, T20 control and 50 member iS-type T20 forecast ensemble bounding box with the stated initial spreads.

Copyright © 0000 Royal Meteorological Society

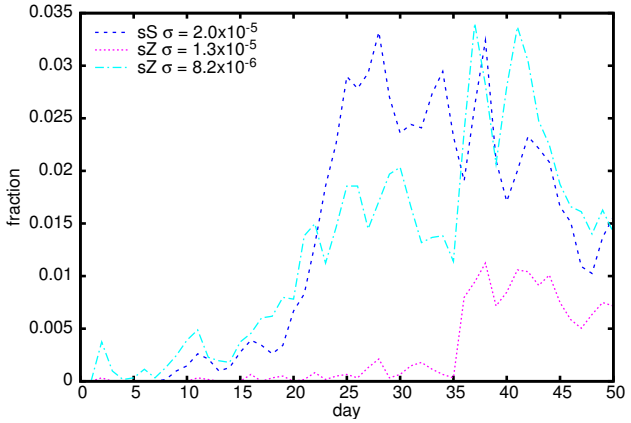

Figure 9. The fraction of T21 target coordinates outside 50 member sS-type and sZ-type ensemble bounding boxes on each forecast day.

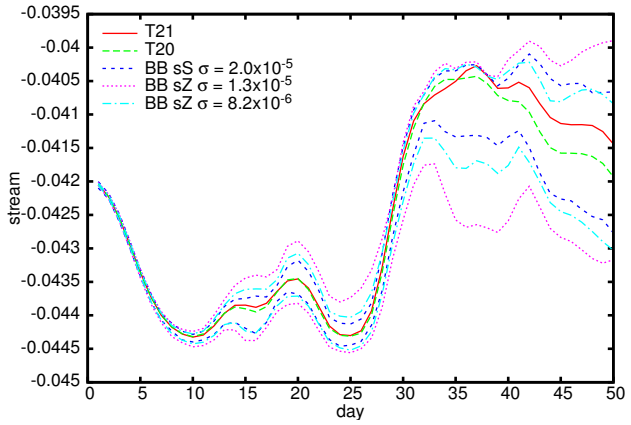

Figure 10. As figure 8 but using 50 member sS-type and sZ-type ensembles.

a curious fact that we will return to in the next paragraph and then again later when discussing a stochastic model for ensemble forecasts.)

The success of the ensembles with wider initial spreads at capturing the target come at significant cost. Figure 8 shows the evolution of a typical middle-latitude stream-function coordinate of the target, control and bounding box of the iS-type ensemble, for two initial spreads of $\sigma = 2 \times 10^{-5}$ and $\sigma = 8 \times 10^{-5}$. Note that the larger initial spread captures the target at the expense of losing specificity on the bounds on the target. The results for the iZ-type ensemble are almost identical, except that iZ-type ensembles require approximately half the initial spread σ to that of the iS-type ensembles to attain a similar capture faction and spread at a given lead time.

Judd and Smith argue (Judd and Smith, 2004) that when using an imperfect model it is essential to take model error into account when estimating states and making forecasts, because failure to do so can markedly degrade both. When making forecasts, one of the simplest methods to account for model error is to make forecasts using a stochastic model, even though the system is believed to be deterministic and a simulation model would otherwise have been deterministic. For example, when evolving a forecast ensemble forward, rather than use the deterministic dynamics $x_{t+1} = f(x_t)$, where x_t is an ensemble member state at time t and f is the model, one instead uses stochastic dynamics $x_{t+1} = f(x_t) + \omega_t$, where ω_t is a random variate that simulates the effects of the unknown model error (imperfection error). This is not the only means of accounting for model error, and certainly not the best, but it is simple, and we will see that is is quite effective.

As in the last sub-section there are a variety of possible perturbations ω_t that can be used in stochastic model forecasts. Ideally one would attempt to mimic the distribution of model errors as best as one could. Here we consider two simple stochastic model ensembles: (sS-type) the ω_t are perturbations like those used as the initial perturbations of iS-type ensembles; (sZ-type) the ω_t like those used in iZ-type ensembles. That is, instead of making one initial perturbation, we make a perturbation at the beginning of each day's integration. Once again we can discuss the size of these perturbations in terms of their standard deviation σ.

We now investigate how sS-type and sZ-type stochastic model ensembles improve the capture of the T21 target using the imperfect T20 model. Since we have access to the T21 system and the T20 model we can compute (for the purposes of comparison, which cannot be done operationally) an empirical distribution of the model errors. We found from a 50 day run that in stream function coordinates the model error had a mean -1.2451×10^{-7} and standard deviation 1.2928×10^{-5}. One might assume that stochastic perturbations ω_t for sS-type and sZ-type should be scaled to achieve stream function perturbations of similar magnitudes, but we will see this is not necesarily the best choice. Also, one may wish to (or need to) take into account spatio-temporal correlation of model errors, because model errors are not independent random variates. We chose to use perturbations in the sS-type ensemble with $z = 0$ and $\sigma = 2 \times 10^{-5}$; where $\sigma > 1.2928 \times 10^{-5}$ as an *ad hoc* means of dealing with model error dependence and correlation. (This is a common practice in time series modelling, for example.) For sZ-type ensembles we find that when $\sigma = 1.3 \times 10^{-5}$ there is too much dispersion of the ensemble. Smaller perturbations giving $\sigma \approx 8 \times 10^{-6}$ are required to give a similar dispersion and capture characteristic of the sS-type ensemble with $\sigma = 2 \times 10^{-5}$, see figure 9. This need to reduce σ is similar to that seen for iS-type and iZ-type ensembles, and the size of the reduction is about the same.

Figure 9 shows the fraction of coordinates of the T21 target outside the bounding box of T20 stochastic model ensemble forecasts. Comparing with figures 6 and 7 we see that the stochastic model ensemble forecasts have capture as good as iS-type and iZ-type ensembles with three to four times larger initial

Copyright © 0000 Royal Meteorological Society

perturbations. Comparing figure 10 to figure 8 shows that the stochastic model ensemble forecasts achieve this without excessive dispersion of the ensemble.

Our analysis of sS-type and sZ-type ensembles using bounding boxes reveals an interesting fact. We observed that in attaining similar capture of the T21 target the sZ-type ensembles did so with smaller perturbations of the stream function, however, we observe that the sS-type ensemble perturbations are closer in magnitude to actual computed stream function model errors. This would imply that sS-type perturbations are more similar to the actual model errors. This may seem surprising, because the sS-type perturbations, compared to the sZ-type perturbations, are large, spatially uncorrelated and have significantly larger projections on to stable modes. Our more detailed analysis suggests the following. Firstly, the large stable component of sS-type perturbations is of little consequence because most of it decays rapidly. Secondly, the model error when forecasting a T21 system with a T20 model has its most significant effect by exciting small-scale unstable features, which may later grow to larger scale features. Thirdly, we find that the unstable component of sS-type perturbations preferentially excite small-scale unstable features, because the perturbations tend to average out over larger scale features. On the other hand, the unstable component of sZ-type perturbations preferentially excite large-scale unstable features, because most of the energy is in the large wave numbers. Hence, sZ-type perturbations cause much faster large-scale dispersion of the ensemble. In practice, these results imply that inappropriate perturbations can be effective at capturing a target, but may be misleading about the actual size and nature of the model errors.

In our experiments we made the assumption of perfect observation, so the T20 model could be initialized directly from the observed T21 state, $x_0 = x_0^\star$, but note that since the T20 model is imperfect, there is no "*true*" state (Judd and Smith, 2004). Even though the model error in this case is quite small, its effect is very significant. When there is observational errors, one still has to take into account model error. This example illustrates that in order to capture the target observations with a deterministic forecast at a lead time of say 30 or so days, one could have estimated that the initial ensemble spread had to be up to four times larger than necessary and resulted in forecast ensembles having bounding boxes two to four times broader. That is, one could easily over-estimate "initialization error" to compensate for model error, and consequently degrade ensemble forecast accuracy, needlessly.

6.3 DEMETER experiments

In order to show how bounding box statistics can be useful in the development and evaluation of an ensemble prediction system we take a very brief look at some preliminary investigations of the DEMETER multi-model experiments (Palmer et al., 2004). These experiments employ seven different state-of-the-art coupled ocean-atmosphere models to re-forecast the earth's climate for the years 1951–2001. Multi-model ensemble forecasts are obtained is initializing each of the 7 models with the re-analyses and 9 different ocean states obtained by perturbation of the wind stress and surface temperture. Ensembles are initialized four times each year and forecasts run for 6 months with certain key fields recorded daily. The 7 different models and 9 initial ocean states, give a 63 member multi-model ensemble. The authors are currently studying how well the multi-model ensemble captures the ERA-40 re-analysis as target over various forecast lead times. Here we present, as an illustration of bounding boxes, a very simple qualitative analysis of some preliminary results. For a detailed analysis see Weisheimer et al. (2005).

Figure 11 shows the fraction of days during a month that the target 2 metre temperature re-analysis was outside the bounding box of the multi-model ensemble at each grid-point for various initialization days and forecast months. The comments made here are based on a number of plots like this one. One immediately observes that the multi-model ensemble does not capture the target sufficiently often, and that the worst failures to capture are localized on ocean regions. This immediately suggests that the ensemble needs more variability, in a way that increases ensemble spread over oceans regions. Comparing panels (A) and (B) shows that the ensemble initialized 3 months before the forecast month generally captures the target better than the ensemble initialized on the first day of the month. This seems to indicate that the ensemble has too narrow spread in the first month after initialization. The localized areas where ensembles (A) and (B) do very poorly have a good deal of correlation, except the 3 month lead time ensemble also does poorly in two regions in the Pacific and west of Africa. This suggests the 3–4 month forecast ensemble has considerable skill, tends to fail in localized ocean regions, and often fails in regions where the 0–1 month forecast ensemble also fails. Comparison of panels (A) and (B) with panel (C) shows that there is also significant correlation of capture for forecasts of the previous month. This suggests that capture in a region for given month is a good predictor of capture in the following month. Further analysis should reveal how persistent failure to capture is and whether this indicates a persistent model error.

7 Conclusions

We have investigated the use of bounding boxes as a means to assess how good an ensemble prediction system is. We have seen that if the ensemble is not biased, that is, the target lies at the median of the ensemble's distribution, then the size of ensemble required to capture the target increases with the logarithm of the dimension of the state space. If the ensembles are biased, then the probability of capturing

Copyright © 0000 Royal Meteorological Society

the target falls rapidly with increasing bias; a bias of up to one standard deviation can be accommodated with modestly sized ensembles. On the other hand, if the bias is due to centring on a primary forecast that has error, and the ensemble spread has been tuned to match primary forecast's error spread, then ensembles need to be about 5 times larger than what are currently used in operational centres. The increase in size is required to capture occasional large primary forecast errors.

We have illustrated the use of bounding boxes with qualitative investigation of the DEMETER experiment data, and by quantitative investigation of the capture of a target analysis using perfect and imperfect QG models. We have verified for imperfect models that a simple stochastic model ensemble forecast performs better than a simple deterministic model ensemble forecast that ignores model error. The stochastic forecast ensembles were demonstrated to capture the target more reliably and with greater specificity. We also demonstrated that even when the stochastic perturbations do not accurately represent the model errors, they still enable the stochastic model ensemble forecast to capture the target, and do better than just inflating initial perturbations.

It might seem that a criticism of bounding boxes is that there is no penalty for achieving capture by simply increasing the spread of the ensemble. But the spread of an ensemble is either predetermined by the method used to generate an ensemble, or free parameter. If the spread is chosen by other means, then bounding boxes provide a test of whether the ensemble is effective as designed. If the spread is a parameter, then bounding boxes can be used to choose the spread by taking the smallest spread that attains the desired rate of target capture, assuming, of course, that the model is sufficiently skillful to allow this capture rate.

The empirical bounding box statistics of operational ensemble systems are of direct utility to decision makers. Evaluating capture rates and typical ranges (relative to climatology) of a skillful operational system can immediately convey the utility of a raw ensemble in a user relevant context. The utility of the bounding box is distinct from that of any full probabilistic forecasts extracted (somehow) from the ensemble, and may prove of higher or lesser value for a given decision maker.

We conclude that bounding box statistics provide insights complementary to more common measures of ensemble performance. Specifically, casting ensemble prediction systems into the bounding box framework yields intuitive and useful information regarding the design of ensembles, the evaluation of ensemble performance against those design expectations and the application of operational ensemble products. We hope that these bounding box statistics, used in conjunction with existing evaluation tools, will prove useful in the difficult and important tasks of evaluating ensemble prediction systems.

Acknowledgements

This research was partly supported by EPSRC Faraday GR/R92363/01. KJ acknowledges the hospitality of NRL, Monterey, the support of ONR N00014-03-10389, and ARC DP0662841. AW was supported by an EU Marie-Curie Fellowship EV K2-CT-2001-50012. The authors also acknowledge the helpful suggestions of two anonymous reviewers, one of whom suggested the calculations that appear in Table IV, the other the comparison of Z-type enembles with S-type ensembles featured in sections 6.1 and 6.2.

References

Anderson, J. (1996). A method for producing and evaluating probabilistic forecasts from ensemble model intergrations. *Journal Climate*, 9:1518–1530.

Anderson, J. (1999). An ensemble adjustment Kalman filter for data assimilation. *Monthly Weather Review*, 129:2884–2903.

Atger (1999). Tubing: an alternative to clustering for classification of ensemble forecasts. *Weather Forecasting*, 14(5):741–757.

Bishop, C., Etherton, B., and Majumdar, S. (2001). Adaptive sampling with the ensemble transform Kalman filter. part 1: Theoretical aspects. *Monthly Weather Review*, 129:420–436.

Buizza, R. (1995). Optimal perturbation time evolution and sensitivity of ensemble prediction to perturbation amplitude. *Q. J. R. Meteorological Society*, 121:1705–1738.

Chatfield, C. (2001). Prediction intervals. In Armstrong, J. S., editor, *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Kluwer Academic.

Del Moral, P. (1995). Nonlinear linear filtering using random particles. *Theory Prob. Appl.*, 40:690–701.

Dethloff, K., Weisheimer, A., Rinke, A., Handorf, D., Kurgansky, M., Jansen, W., Maass, P., , and Hupfer, P. (1998). Climate variability in a non-linear atmosphere-like dynamical system. *J. Geophys. Res.*, 103(25):957–966.

Errico, R. and Baumhefner, D. (1987). Predictability experiments using a high-resolution limited-area model. *Monthly Weather Review*, 115:488–504.

Evensen, G. (1994). Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophyical Research*, 99(C5):10,143–10,162.

Gilmour, I., Smith, L., and Buizza, R. (2001). On the duration of the linear regime: Is 24 hours a long time in synoptic weather forecasting? *Journal of the Atmospheric Sciences*, 59:3525–3539.

Copyright © 0000 Royal Meteorological Society

Gneiting, T., Raferty, A., Westveld, A., and Goldman, T. (2004). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*.

Hamil, T. and Synder, C. (2000). A hybrid ensemble Kalman filter 3d-variational analysis scheme. *Monthly Weather Review*, 128:2905–2919.

Hamill, T. (2001). Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, 129:550–560.

Houtekamer, P., Lefaivre, L., Derone, J., Ritchie, H., and Mitchell, H. (1996). A system simulation approach to ensemble prediction. *Monthly Weather Review*, 124:1225–1242.

Houtekamer, P., Mitchell, H., Pellerin, G., Buehner, M., Charron, M., Spacek, L., and Hansen, B. (2005). Atmospheric data assimilation with an ensemble kalman filter: Results with real observations. *Monthly Weather Review*, 133(3):604–620.

Judd, K. and Smith, L. (2001). Indistinguishable states I : perfect model scenario. *Physica D*, 151:125–141.

Judd, K. and Smith, L. (2004). Indistinguishable states II : imperfect model scenarios. *Physica D*, 196:224–242.

Molteni, F., Buizza, R., Palmer, T., and Petroliagis, T. (1996). The ECMWF ensemble prediction system: Methodology and validation. *Q. J. R. Meteorological Society*, 129:73–119.

Mureau, R., Molteni, F., and Palmer, T. (1993). Ensemble prediction using dynamically conditioned perturbations. *Q. J. R. Meteorological Society*, 119:299–323.

Orrell, D., Smith, L., Barkmeijer, J., and Palmer, T. (2001). Model error in weather forecasting. *Nonlinear Processes in Geophysics*, 8:357–371.

Palmer, T., Alessandri, A., Andersen, U., Cantelaube, P., Davey, M., Déléluse, P., Déqué, M., Díez, E., Doblas-Reyes, F., Feddersen, H., Graham, R., Gualdi, S., Guérémy, J.-F., Hagedorn, R., Hoshen, M., Keenlyside, N., Latif, M., Lazar, A., Mainsonnave, E., Marletto, V., Morse, A., Orfila, B., Rogel, P., Terres, J.-M., and Thomson, M. (2004). Development of a european multi-model ensemble system for seasonal to interannual prediction. *Bull. Am. Meteorol. Soc.*, 85:853–872.

Petterssen, S. (1958). *Introduction to Meteorology*. McGraw Hill.

Reynolds, C. and Rosmond, T. (2003). Nonlinear growth of singular vector-based perturbations. *Q. J. R. Meteorol. Soc.*, 129:3059–3079.

Ridout, D. and Judd, K. (2001). Convergence properties of gradient descent noise reduction. *Physica D*, 165:27–48.

Roulston, M. and Smith, L. (2003). Combining dynamical and statistical ensembles. *Tellus*, 55A:16–30.

Silverman, B. W. (1988). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

Smith, L. (1995). Accountability and error in ensemble forecasting. In *Predictability*, volume 1 of *ECMWF Seminar Proceedings*, pages 351–369, Shinfield Park, Reading, Berkshire, RG29ax. ECMWF.

Smith, L. (1997). The maintenance of uncertainty. In *Proc International School of Physics "Enrico Fermi"*, volume Course CXXXIII, pages 177–246, Bologna, Italy. Società Italiana di Fisica.

Smith, L. (2000). Disentangling uncertainty and error: On the predictability of nonlinear systems. In Mees, A. I., editor, *Nonlinear Dynamics and Statistics*, pages 31–64. Birkhauser, Boston.

Smith, L. and Hansen, J. (2004). Extending the limits of forecast verification with the mst. *Mon. Weather Rev.*, 132(6):1522–1528.

Talagrand, O., Vantard, R., and Strauss, B. (1997). Evaluation of probabilistic prediction systems. In *Proceedings of a workshop on predicability*, volume CXXXIII, pages 1–25, ECMWF, Shinfield Park, Reading RG2 9AX.

Toth, Z. and Kalnay, E. (1997). Ensemble forecasting at NCEP and the breeding vector method. *Mon. Weather Rev.*, 125:3297–3319.

Tribbia, J. and Baumhefner, D. (2003). Scale interactions and atmospheric predictability : an updated perspective. *Monthly Weather Review*, 132:703–713.

Weisheimer, A., Kurgansky, M., Dethloff, K., , and Handorf, D. (2003). Extratropical low-frequency variability in a three-level quasi-geostrophic atmospheric model with different spectral resolution. *J. of Geophysical Research*, 108(4171):doi:10.1029/2001JD001282.

Weisheimer, A., Smith, L., and Judd, K. (2005). A new view of forecast skill : bounding boxes from the DEMETER ensemble seasonal forecasts. *Tellus*, 57:265–279.

Wilks, D. (1995). *Statistical Methods in the Atmospheric Sciences : An Introduction*, volume 59 of *International Geophysics Series*. Academic Press, London.

Wilks, D. (2004). The minimum spanning tree histogram as a verification tool for multidimensional ensemble forecasts. *Monthly Weather Review*, 132(6):1329–1340.

Copyright © 0000 Royal Meteorological Society

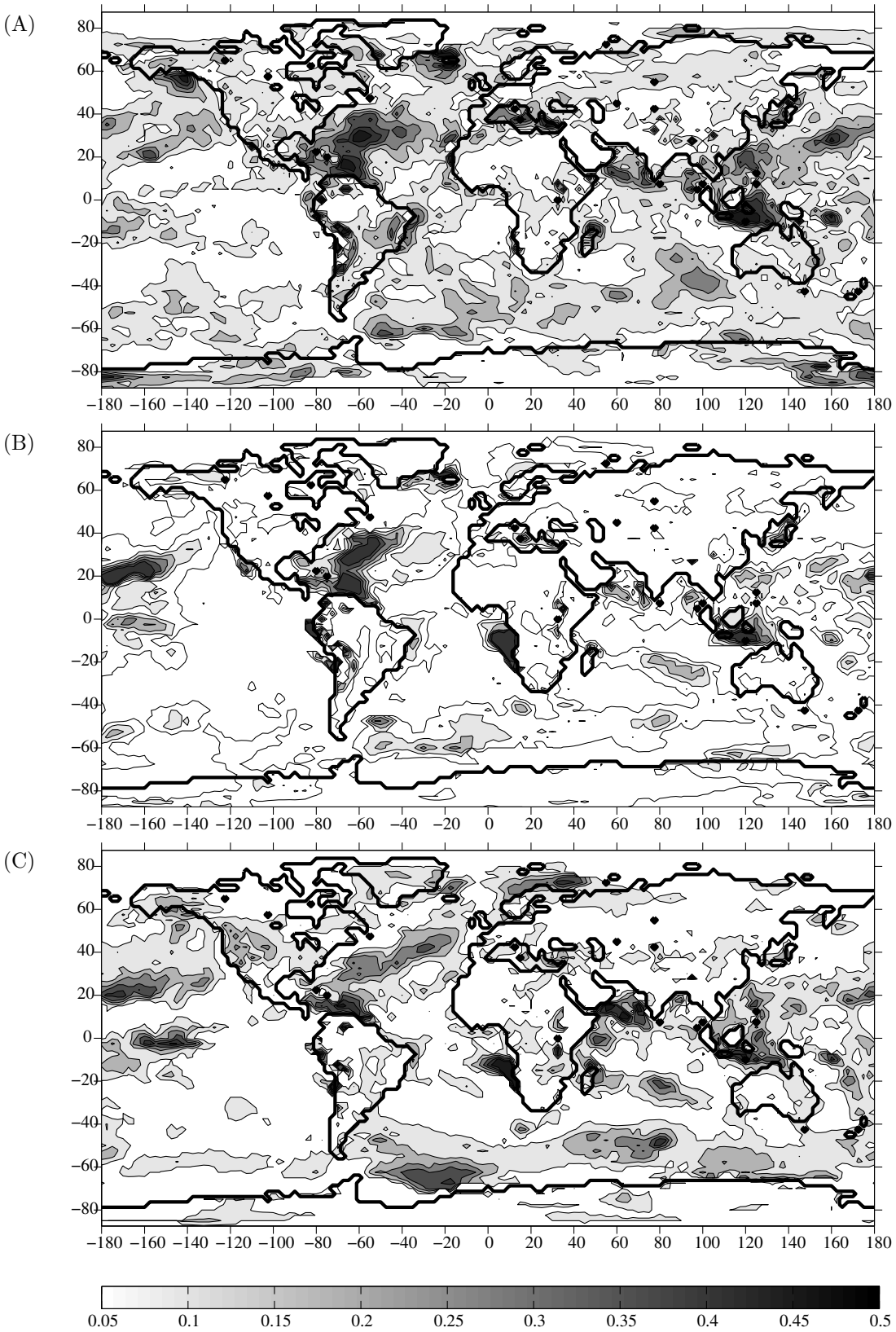

Figure 11. Each panel shows the fraction of days during the forecast month that the target ERA-40 2 metre temperature is outside the DEMETER multi-model ensemble bounding box at that grid-point. (A) The forecast month is May 1994, the ensemble initialized on 1st May 1994. (B) The forecast month is May 1994, the ensemble initialized on 1st February 1994. (C) The forecast month is April 1994, the ensemble initialized on 1st February 1994. White areas indicate capture of the target. Plots like these vary with the forecast month.

Copyright © 0000 Royal Meteorological Society